Conception graphique de l'affiche de la conférence : Cédric Lopez, Caroline Imbert(LIRMM). La couverture de cet ouvrage a été réalisée par Cédric Lopez (LIRMM). Composition et mise en page des actes : Mathieu Lafourcade (LIRMM). Impression : AVL Diffusion (Montpellier).

TALN 2011

RECITAL 2011

du 27 juin au 1er juillet 2011, Montpellier, France

Actes de la 18e conférence
sur le

Traitement Automatique des Langues Naturelles

Actes de la 15e

Rencontre des Étudiants Chercheurs en Informatique pour le Traitement automatique des Langues

---

Éditeurs
Mathieu Lafourcade et Violaine Prince

Organisation de la conférence
Équipe TAL LIRMM (UMR 5506)

Sous l'égide de l'ATALA
(Association pour le Traitement Automatique des langues)

Volume 1/2
(Articles longs)

Table des matières

Préface

Cette dix-huitième édition de TALN a un certain nombre de particularités. Elle est bien sûr assortie de la quinzième édition de RECITAL, mais elle s'est trouvée également enrichie des nouveaux ateliers apparus plus récemment, DEFT (sixième édition) qui a fait des infidélités à TALN pour revenir ensuite dans son giron, DEGELS, apparu l'an dernier, et auquel on souhaite longue vie, et enfin DISH, doctorants en informatique pour les sciences humaines, dont c'est aussi la deuxième édition, et que nous accueillons avec plaisir. Cependant, avant tout, l'univers de TALN s'est enrichi cette année d'un jumelage cher au cœur de certains d'entre nous : nous avons le plaisir d'accueillir LACL, grâce à Philippe Blache, fondateur de TALN et grand promoteur de LACL. Notons que des fondateurs de RECITAL, de DEFT se trouvent actuellement dans le comité d'organisation de cette édition de TALN. Le fait de suivre à nouveau les trajectoires des ateliers auxquels ils ont consacré leur énergie créatrice sous la houlette d'autres équipes qui se les sont appropriés, est pour eux une satisfaction certaine...

Outre le fait que nous avons souhaité un mélange des communautés TALN et LACL, l'une anglophone par nécessité, et l'autre majoritairement francophone par choix, en proposant des inscriptions conjointes, nous nous sommes dits que nous pouvions également remettre la démarche paradigmatique de l'une à l'honneur dans l'univers pluraliste de l'autre. C'est pourquoi nous avons opté pour des conférenciers invités dont l'univers des travaux relève des modèles symboliques et logiques. Ce n'est pas que ce soit l'orientation unique de notre communauté, au contraire, mais nous avons pensé que ce serait bien de mettre en exergue les intersections des communautés. Claire Gardent et Nicholas Asher sont très connus des deux communautés. Vladimir Fomichev est peu connu, et nous avons souhaité ici faire une ouverture vers la Russie, en lui proposant ce tremplin. Mathématicien d'origine, féru de sciences cognitives, grand admirateur de la langue française, il a écrit une monographie éditée chez Springer que nous avons lue avec plaisir, et qui nous a paru compléter avec qualité les travaux de Nicholas Asher et de Claire Gardent.

Cette édition de TALN et de ses conférences associées a été bien perçue de la communauté. 94 soumissions de papiers dit "longs" (et cette année, nous avons proposé 12 pages, au lieu des 10 habituelles, afin de permettre à beaucoup de parachever leur article) pour TALN et 8 soumissions pour RECITAL, ce qui fait un total de 102 soumissions. Sur ces soumissions, le taux de sélection a été comparable à celui des éditions précédentes de TALN. En effet, 38 papiers longs ont été acceptés pour TALN (taux d'acceptation = 40%) et 5 pour RECITAL. On pourrait peut-être percevoir une désaffection progressive de RECITAL (la première édition avait reçu 42 soumissions), et éventuellement, on pourrait considérer qu'une conférence "jeunes chercheurs" dans notre domaine, a fait son temps. En termes d'organisation des publications, nous avons souhaité être dans la continuité de ce qui s'est fait lors des éditions précédentes de TALN. En particulier, l'idée des papiers courts, proche de la notion de "position paper", nous a semblé très prometteuse. La communauté ne s'y est pas trompée, puisque 95 soumissions de papiers courts nous sont arrivées pour TALN (et 6 pour RECITAL). Ces papiers courts ont bénéficié du même processus de relecture consciencieux que les papiers longs, et 52 papiers courts ont été sélectionnés pour TALN, 4 pour RECITAL.

La présentation orale des papiers courts se fera sous forme de présentation affichée (poster) et "booster", les deux minutes pour convaincre les auditeurs de TALN de venir discuter avec les auteurs de la présentation.

Cette année, nous avons souhaité privilégier les démonstrations de logiciels. Qu'elles soient académiques ou industrielles, nous pensons que le TAL a beaucoup à offrir au grand public, en termes de produits, de services et d'aides diverses. Nous n'avons pas voulu sélectionner ces contributions, car nous pensons que toutes les démonstrations sont intéressantes, et qu'en particulier les industriels doivent avoir auprès de nous une tribune ouverte. 20 démonstrations, pour lesquelles nous avons demandé un résumé d'une page, ont été proposées par différents contributeurs.

Nous souhaitons, dans cette préface à ces actes volumineux, aux lecteurs de la dix-huitième édition de TALN (15ème RECITAL) une lecture agréable et passionnée. La communauté TAL a réalisé de belles choses, et nous espérons que notre édition s'en est fait l'écho. Nous voulons en particulier remercier l'ATALA, qui porte cette conférence depuis tant d'années, son comité permanent qui assure la continuité de la qualité et de la forme au fil du temps, le comité scientifique, qui s'est fortement investi dans le processus de relecture, le comité de sélection, et qui est garant de la qualité, et enfin le comité de lecture qui n'a pas ménagé ses efforts et son temps pour permettre à ce processus de converger.

Nous, équipe TAL du LIRMM (changeant de nom pour devenir TEXTE), avons été très honorés d'organiser cette conférence, d'éditer ces actes. Nous espérons que la manifestation aura le succès qu'elle mérite, et dont les prémices se profilent d'ores et déjà dans les articles qui composent cette publication.

Mathieu Lafourcade et Violaine Prince

Comités

Comité d'organisation de TALN 2011 et RECITAL 2011

Présidents

Mathieu Lafourcade (LIRMM, Université Montpellier 2)
Violaine Prince (LIRMM, Université Montpellier 2)

Membres

Nicolas Béchet (LIRMM, Université Montpellier 2)
Jacques Chauché (LIRMM, Université Montpellier 2)
Elisabeth Grêverie (LIRMM, Université Montpellier 2)
Caroline Imbert (LIRMM, Université Montpellier 2)
Alain Joubert (LIRMM, Université Montpellier 2)
Alexandre Labadié (LIRMM, Université Montpellier 2)
Stéphanie Leon (LIRMM, Université Montpellier 2)
Cédric Lopez (LIRMM, Université Montpellier 2)
Anne Preller (LIRMM, Université Montpellier 2)
Jean-Philippe Prost (LIRMM, Université Montpellier 2)
Stéphane Riou (LIRMM, Université Montpellier 2)
Mathieu Roche (LIRMM, Université Montpellier 2)
Johan Segura (LIRMM, Université Montpellier 2)
Richard Terrat (LIRMM, Université Montpellier 2)

Comité de programme de TALN 2011

Présidents

Mathieu Lafourcade (LIRMM, Université Montpellier 2)
Violaine Prince (LIRMM, Université Montpellier 2)

Membres

Nick Asher (IRIT)
Frédéric Béchet (Aix-Marseille Université)
Yves Bestgen (UCL)
Philippe Blache (Université d'Aix-en-Provence)
Hervé Blanchon (LIG, Grenoble)
Christian Boitet (LIG, Grenoble)
Malek Boualem (France Telecom Orange Labs)
Yllias Chali (University of Lethbridge)

Laurence Danlos (Université Paris 7)
Mark Dras (Macquarie University)
Denys Duchier (Orléans)
Marc Dymetman (Xerox Research Centre Europe)
Cédrick Fairon (Université catholique de Louvain)
Olivier Ferret (CEA LIST)
Michel Gagnon (Polytechnique Montréal)
Claire Gardent (CNRS/LORIA et INRIA Nancy Grand Est)
Nabil Hathout (Université Toulouse-le Mirail)
Sylvain Kahane (Université Paris Ouest - Nanterre)
Philippe Langlais (Université de Montréal)
Cédric Lopez (LIRMM, Université Montpellier 2)
Emmanuel Morin (Université de Nantes)
Adeline Nazarenko (Université Paris 13)
Luka Nerima (Université de Genève)
Alain Polguère (Nancy-Université & ATILF CNRS)
Laurent Prévot (LPL , Aix-Marseille Université)
Jean-Philippe Prost (LIRMM, Université Montpellier 2)
Christian Retoré (LaBRI)
Mathieu Roche (LIRMM, Université Montpellier 2)
Pascale Sébillot (IRISA, INSA de Rennes)
Gilles Sérasset (LIG, Grenoble)
Anne Vilnat (LIMSI-CNRS et Université Paris-Sud 11)
François Yvon (LIMSI-CNRS et Université Paris-Sud 11)
Pierre Zweigenbaum (LIMSI-CNRS)

Comité scientifique de TALN 2011

Présidents

Mathieu Lafourcade (LIRMM, Université Montpellier 2)
Violaine Prince (LIRMM, Université Montpellier 2)

Membres

Stergos Afantenos
Salah Ait-Mokhtar
Maxime Amblard
Jean-Yves Antoine
Veronique Auberge
Delphine Battistelli
Denis Bechet
Núria Bel
Patrice Bellot
Delphine Bernhard
Laurent Besacier
Romaric Besançon
Yves Bestgen
Brigitte Bigi
Philippe Blache
Hervé Blanchon
Christian Boitet
Malek Boualem
Pierrette Bouillon
Julien Bourdaillet
Francis Brunet-Manquat
Marie Candito
Jacques Chauché
Jean-Pierre Chevallet
Vincent Claveau
Nathalie Colineau
Beatrice Daille
Laurence Danlos
Eric De La Clergerie
Claude De Loupy
Pascal Denis
Mark Dras
Patrick Drouin
Denys Duchier
Iris Eshkol
Yannick Estève
Jacquey Evelyne
Cécile Fabre
Benoit Favre
Olivier Ferret
Dominic Forest
Karen Fort
George Foster
Nuria Gala
Bruno Gaume
Kim Gerdes
Jérôme Goulian

Benoit Habert
Thierry Hamon
Nabil Hathout
Stéphane Huet
Marie-Paule Jacques
Guillaume Jacquet
Christine Jacquin
Adel Jebali
Alain Joubert
Daniel Kayser
L. Kosseim
Olivier Kraif
Mathieu Lafourcade
Eric Laporte
Monceaux Laura
Anne Laurent
Thomas Lebarbe
Yves Lepage
Joseph Leroux
Cédric Lopez
Elliott Macklovitch
Denis Maurel
Aurélien Max
Jean-Guy Meunier
Jasmina Milicevic
Richard Moot
Erwan Moreau
Fabienne Moreau
Véronique Moriceau
Emmanuel Morin
Philippe Muller
Alexis Nasr
Adeline Nazarenko
Luka Nerima
Aurélie Névéol
Jian-Yun Nie
Yannick Parmentier
Guy Perrier
Sophie Piron
Sylvain Pogodalla
Thierry Poibeau
Claude Ponton
Andrei Popescu-Belis
Fred Popowich
Laurent Prévot
Violaine Prince
Jean-Philippe Prost
Christian Retore
Mathieu Roche
Sophie Rosset
Antoine Rozenknop
Benoît Sagot
Didier Schwab
Pascale Sebillot
Djamé Seddah
Gilles Serasset
Laurianne Sitbon
Kamel Smaili
Marina Sokolova
Xavier Tannier
Isabelle Tellier
Juan-Manuel Torres-Moreno
François Trouilleux
Mathieu Valette
Lonneke Van Der Plas
Tristan Vanrullen
Fabienne Venant
Jacques Vergne
Evelyne Viegas
Anne Vilnat
Eric Wehrli
Guillaume Wisniewski
François Yvon
Virginie Zampa
Imed Zitouni
Michael Zock
Amal Zouaq
Mounir Zrigui
Pierre Zweigenbaum

Comité de programme et de lecture de RECITAL 2011

Président

Cédric Lopez (LIRMM, Université Montpellier 2)

Membres

Jean-Yves Antoine (Université François-Rabelais)
Nicolas Béchet (INRIA)
Sadok Ben Yahia (Faculté des Sciences de Tunis)
Florian Boudin (Université de Avignon)
Narjès Boufaden (Keatext)
Julien Bourdaillet (RALI, Université de Montréal)
Sandra Bringay (LIRMM, Université Montpellier 2)
Marie Candito (Université Paris Diderot)
Vincent Claveau (IRISA)
Patrick Drouin (Université de Montréal)
Marie Dupuch (SPIM, Université Paris 5)
Cécile Fabre (Université de Toulouse / CLLE-ERSS)
Philippe Gambette (LIRMM, Université Montpellier 2)
Marie-Laure Guénot (Chercheur indépendant)
Thierry Hamon (LIM & BIO, Université Paris 13)
Alain Joubert (LIRMM, Université Montpellier 2)
Leila Kosseim (Université Concordia)
Jean-Guy Meunier (Université du Québec À Montréal)
Thibault Mondary (Université Paris 13)
Philippe Muller (Université Paul Sabatier)
Alexandre Patry (Keatext)
Aurélie Picton (Université de Toulouse)
Johan Segura (LIRMM, Université Montpellier 2)
Charles Tesseidre (Université Paris 10)
Jean Quirion (University of Ottawa)
Amal Zouaq (Royal Military College of Canada)

Orateurs invités

Vladimir A. Fomichov
National Research University "Higher School of Economics", Moscou (Russie)

The prospects revealed by the theory of K-representations
for bioinformatics and Semantic Web

Claire Gardent
CNRS/LORIA et INRIA Nancy Grand Est (France)

Sentence Generation : Input, Algorithms and Applications

Nick Asher
LILaC, IRIT, Université Paul Sabatier, Toulouse (France)

Théorie et Praxis
Une optique sur les travaux en TAL sur le discours et le dialogue

Vladimir A. Fomichov
(Fomichev)

Ph.D., Docteur des Sciences Techniques (Moscou, Russie)

Depuis 2007, V. Fomichov est Professeur d'informatique, à l'Université Nationale des Recherches *l'Ecole Supérieure de l'Economie*, Faculté de l'Informatique Commerciale, Département des Innovations et Commerce dans la Sphère des Technologies Informationnelles, à Moscou.

De mars 2000 à juin 2009 , V. Fomichov a été professeur *en ligne* au Centre de l'Instruction de Distance "SNHU Online Education", l'Université Sud de la Province New Hampshire (SNHU), 2500 North River Road, Manchester, New Hampshire, U.S.A. Pendant des années 2000 – 2009, il a enseigné à plus que 1000 étudiants des Etats-Unis, Canada, Israël, l'Allemagne, l'Italie, Autriche, Japon - 5 disciplines mathématiques en anglais de Moscou par l'aide de Réseau (Web).

Publications

V. Fomichov est l'auteur et co-auteur plus que des 145 publications scientifiques, en particulier, plus de 80 travaux en anglais qui sont publiés en Angleterre, Autriche, Allemagne, Belgique, Bulgarie, Canada, Espagne, aux Etats-Unis, Finlande, France, Hollande, Norvège, Slovénie, Suisse, Zélande Nouvelle et des trois travaux en français qui sont publiés en France et Suisse. Auteur de quatre monographies, et en particulier, de :

> Fomichov, V.A. *Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms.* IFSR International Series on Systems Science and Engineering , Vol. 27. Springer: New York, Dordrecht, Heidelberg, London, 2010. - 354 p.; http://www.springer.com/math/applications/book/978-0-387-72924-4

Ses travaux se rapportent aux Mathématiques Discrètes, la Linguistique Mathématique et Calculatoire, l'Intelligence Artificielle, Réseau Sémantique, les Systèmes Multi-Agents, le Commerce Electronique, la Cognitonique et la Théorie d'Instruction.

The prospects revealed by the theory of K-representations for bioinformatics and Semantic Web

Vladimir A. Fomichov

Department of Innovations and Business in the Sphere of Informational Technologies
Faculty of Business Informatics, National Research University "Higher School of Economics"
Kirpichnaya str. 33, 105679 Moscow, Russia
vfomichov@hse.ru and vfomichov@gmail.com

Résumé

L'article décrit la structure et les applications possibles de la théorie des K-représentations (représentation des connaissances) dans la bioinformatique afin de développer un Réseau Sémantique d'une génération nouvelle. La théorie des K-répresentations est une théorie originale du développement des analyseurs sémantico–syntactiques avec l'utilisation large des moyens formels pour décrire les données d'entrée, intermédiaires et de sortie. Cette théorie est décrit dans la monographie de V. Fomichov (Springer, 2010). La première partie de la théorie est un modèle formel d'un système qui est composé de dix opérations sur les structures conceptuelles. Ce modèle définit une classe nouvelle des langages formels – la classe des SK-langages. Les possibilités larges de construire des répresentations sémantiques des discours compliqués en rapport à la biologie sont manifestes. Une approche formelle nouvelle de l'élaboration des analysateurs multilinguistiques sémantico-syntactiques est décrite. Cet approche a été implémentée sous la forme d'un programme en langage PYTHON.

Abstract

The paper describes the structure and possible applications of the theory of K-representations (knowledge representations) in bioinformatics and in the development of a Semantic Web of a new generation. It is an original theory of designing semantic-syntactic analyzers of natural language (NL) texts with the broad use of formal means for representing input, intermediary, and output data. The current version of the theory is set forth in a monograph by V. Fomichov (Springer, 2010). The first part of the theory is a formal model describing a system consisting of ten operations on conceptual structures. This model defines a new class of formal languages – the class of SK-languages. The broad possibilities of constructing semantic representations of complex discourses pertaining to biology are shown. A new formal approach to developing multilingual algorithms of semantic-syntactic analysis of NL-texts is outlined. This approach is realized by means of a program in the language PYTHON.

Mots-clés : dialogue homme-machine en langage naturel, algorithme de l'analyse sémantico-syntactique, sémantique intégrale formelle, théorie des K-représentations, SK-langues, représentation sémantique, bases de données linguistiques, réseau sémantique d'une génération nouvelle, réseau sémantique multilingue, bioinformatique

Keywords: man-machine natural language dialogue, algorithm of semantic-syntactic analysis, integral formal semantics, theory of K-representations, SK-languages, semantic representation, text meaning representation, linguistic database, Semantic Web of a new generation, multilingual Semantic Web, bioinformatics

1 Introduction

Many years ago, before the birth of the World Wide Web and bioinformatics, the author of this paper came across the following idea : the progress of computers and informational technologies will reach the point when the continuation of this progress will require the applied computer systems with the well-developed abilities of processing natural language (NL): written texts and spoken speech. The analysis carried out at that time distinguished three significant problems : (a) NL-interaction with applied intelligent systems (AIS) ; (b) the construction of question-answering systems dealing with free texts ; (c) automatic extraction of information from NL-texts for updating knowledge bases of AIS (Fomitchov, 1984; Fomichov, 1992, 1993a).

An important precondition of solving these problems seemed to be the formal means allowing for representing the structured meanings (SMs), or semantic structures, of arbitrary NL-texts pertaining to economy, technology, medicine, and other fields of human professional activity. This idea underlay the birth of Integral Formal Semantics (IFS) of NL – an original branch of mathematical and computational linguistics (see, in particular, (Fomichov, 1992 - 1994) and Chapter 2 of (Fomichov, 2010a)).

The beginning of the XXIst century appears to be just the time point when the progress of computers and the Internet demands powerful and flexible technologies of NL processing for applying them in numerous thematic domains. One has been able to observe in the 2000s in different parts of the world the permanent growth of interest in designing NL-interfaces to applied intelligent systems and other kinds of natural language processing systems (NLPS), or linguistic processors. In particular, a number of projects being useful for practice is described in the publications (Cimiano et al., 2007; Duke, Glover, Davies, 2007; Frank et al., 2007; Popescu, Etzioni, Kautz, 2003).

On the one hand, the first version of a NLPS usually is not an ideal one. Additional work is necessary in order to expand the input language of the system (being a sublanguage of NL) and enhance the intelligent capabilities of the system. On the other hand, when a useful for practice system has been designed, the proposals to adapt this system to the utilisation in a different domain may be received. That is why it is important to have a collection of formal tools enabling the designers to fix the assumptions about semantic structures, linguistic databases, input and output data structures, and about intermediate data structures being outputs of some subsystems and the inputs of other subsystems.

One of the most acute and large-scale problems is to endow the existing Web with the ability of extracting information from numerous sources in various natural languages (of cross-language information retrieval) and of constructing NL-interfaces to a number of knowledge repositories recently developed under the framework of the Semantic Web project (Wilks, Brewster, 2006 ; Fomichov, 2005, 2009a – 2010b).

As far as in the middle of the 1960s, the researchers had practically the only formal approach to describing structured meanings (SMs) of NL-texts : the first-order predicates logic (FOPL). Due to numerous restrictions of FOPL, the search for more powerful and flexible formal means for decribing SMs of NL-texts was started in the second half of the 1960s. As a result, a number of new theories have been developed, first of all, the Theory of Generalized Quantifiers (TGQ), Discourse Representation Theory (DRT), Theory of Semantic Nets (TSN), Theory of Conceptual Graphs (TCG), Episodic Logic (EL), and Theory of K-representations. The latter theory being now the central component of IFS is an original theory of designing semantic-syntactic analyzers of NL-texts with the broad use of formal means for representing input, intermediary, and output data.

In order to understand the principal distinction of the theory of K-representations from other mentioned approaches to formalizing semantics of NL, let's consider an analogy. Bionics studies the peculiarities of the structure and functionning of the living beings in order to discover the new ways of solving certain technical problems. Such theories as TGQ, DRT, TSN, TCG, EL and several other theories were elaborated on the way of expanding the expressive mechanisms of FOPL. To the contrary, the theory of K-represenations was developed as a consequence of analysing the basic expressive mechanisms of NL and putting forward a conjecture about a system of partial operations on conceptual structures underpinning these expressive mechanisms. Of course, the idea was to develop a formal model of this system being compatible with FOPL.

This paper aims at attracting the attention of the researchers to new prospects revealed by the theory of K-representations for the design of semantics-oriented NLPSs (first of all, in the field of bioinformatics) and for developing a Semantic Web of a new generation (SW-2), where its principal distinguished feature will be the well-developed mechanisms for conceptual processing of texts and spoken speech in many natural languages. So

SW-2 can be also called a Meanings Understanding Web or a Multilingual Semantic Web (Fomichov, 2009a – 2010b).

The first subject of this paper is the demonstration (on the examples of complex biological discourse and definition) of some precious features of a mathematical model introduced in (Fomichov, 2010a) and describing a system of 10 partial operations on conceptual structures for building semantic representations (in other terms, text – meaning representations) of, most likely, arbitrary sentences and discourses in French, English, German, Russian, and other natural languages (texts pertaining to arbitrary spheres of professional activity). This model is the kernel of the theory of K-representations.

The second subject is the arguments in favour of employing the theory of K-representations as a foundation of an original strategy of transforming the existing Web into a Semantic Web of a new generation.

The third subject of this paper is the description of a new method of designing multilingual semantics-oriented NLPS with the help of formal means for representing intermediary, output, and a part of input data. A multilingual algorithm of semantic-syntactic analysis of NL-texts called *SemSynt1* and introduced in the second part of the monograph (Fomichov, 2010a) was developed in accordance with this new approach and was implemented by means of the language of Web programming PYTHON.

2 Two basic principles of designing linguistic processors

Most often, semantics-oriented natural language processing systems, or linguistic processors, are complex computer systems, their design requires a considerable time, and its cost is rather high. That is why usually, as it was mentioned above, it is necessary to elaborate a series of NLPSs, step by step expanding the input sublanguage of NL and satisfying the requirements of the end users. On the other hand, the same regularities of NL are manifested in the texts pertaining to various thematic domains.

That is why, in order to diminish the total expenses of designing a family of NLPSs by one research centre or group during a certain several-year time interval and in order to minimize the duration of designing each particular system from this family of NLPSs, it seems to be reasonable to pay more attention to: (a) the search for best typical design solutions concerning the key subsystems of NLPSs with the aim to use these solutions in different domains of employing NLPSs; (b) the elaboration of formal means for describing the main data structures and principal procedures of algorithms implemented in semantic-syntactic analyzers of NL-texts or in the synthesizers of NL-texts.

That is why it appears that the adherence to the following two principles in the design of semantics-oriented NLPSs by one research centre or a group will contribute, in the middle-term perspective, to reducing the total cost of designing a family of NLPS and to minimizing the duration of constructing each particular system from this family:

the ***Principle of Stability*** of the used language of semantic representations in the context of various tasks, various domains and various software environments (stability is understood as the employment of a unified collection of rules for building the semantic structures as well as domain- and task-specific variable set of primitive informational units);

the ***Principle of Succession*** of the algorithms of NLPS based on using one or more compatible formal models of a linguistic database and unified formal means for representing the intermediate and final results of semantic-syntactic analysis of natural-language texts in the context of various tasks, various domains and various software environments (the succession means that the algorithms implemented in basic subsystems of NLPS are repeatedly used by different linguistic processors).

3 Formalization of basic assumptions about primary items of conceptual level

The monograph (Fomichov, 2010a) sets forth a current version of the theory of K-representations (knowledge representations). It is an original theory of designing multilingual semantic-syntactic analyzers of NL-texts (sentences and discourses) with the broad use of formal means for representing input, intermediary, and output data. Let's start to consider the structure of this theory.

The *first basic constituent* of the theory of K-representations is the theory of SK-languages (standard knowledge languages). The kernel of the theory of SK-languages is a mathematical model describing a system of such 10 partial operations on structured meanings (SMs) of natural language texts (NL-texts) that, using primitive conceptual items as "blocks", we are able to build SMs of arbitrary NL-texts (including articles, textbooks, etc.) and arbitrary pieces of knowledge about the world. The analysis of the scientific literature on artificial intelligence theory, mathematical and computational linguistics shows that today the class of SK-languages opens the broadest prospects for building semantic representations (SRs) of NL-texts (i.e., for representing meanings of NL-texts in a formal way).

The first part of the theory of SK-languages is a mathematical model describing a system of primary conceptual units used by an applied intelligent system, in particular, by a NL processing system. This model defines (with the help of a rather long sequence of auxiliary steps) a new class of formal objects called *conceptual bases (c.b.),* where each concrete c.b. is constructed for a certain group of application domains. Each c.b. *B* is equivalent to a system of the form (c_1, ... , c_{15}) with the components c_1,..., c_{15} being mainly finite or countable sets of symbols and distinguished elements of such sets. In particular, $c_1 = St$ is a finite set of symbols called *sorts* and designating the most general considered notions (concepts); $c_5 = X = X(B)$ is a countable set of strings used as elementary blocks for building knowledge modules and semantic representations (SRs) of texts; *X* is called a primary informational universe; $c_6 = V$ is a countable set of variables; $c_8 = F$ is a subset of *X* whose elements are called functional symbols.

The set of sorts *St* can include, in particular, the elements *spatial.object, physical.object, dynamic.physical.object, intelligent.system, organization, moment, situation, event,* etc. The set of sorts *St* is a subset of the set *X*. For instance, *X* may include the elements *book, ship, firm, 12, green, Height, Weight, Authors, Part-of, Cheeper, printing, uploading*. The elements of the set *V* are used either as the marks of the entities of various kinds or jointly with the universal and existential quantifiers. The set *F* consists of the designations of functions and is a subset of the set *X = X(B).* The set *F* may include, for instance, the elements *Height, Weight, Authors.*

4 About a model of a system consisting of ten operations on conceptual structures

Each c.b. *B* determines three classes of formulas, the first class *Ls(B)* being considered as the principal one and being called *the SK-language (standard knowledge language) in the basis B*. Its strings (they are called K-strings) are convenient for building SRs of NL-texts. We'll consider below only the formulas from the first class *Ls(B).* If *Expr* is an expression in natural language and a K-string *Semrepr* can be interpreted as a semantic representation of *Expr*, then *Semrepr* will be called a K-representation (KR) of the expression *Expr*.

For determining for arbitrary c.b. *B* three classes of formulas, a collection of inference rules P[0], P[1], ... , P[10] is defined. The rule P[0] provides an initial stock of formulas from the first class. E.g., there is such c.b. B_1 that, according to P[0], $Ls(B_1)$ includes the elements *car1, green, city1, fin-set, India, 14, 14/cm, all, any, Height, Distance, Staff, Suppliers, Quantity, x1, x5*.

For arbitrary c.b. *B*, let *Degr(B)* be the union of all Cartesian m-degrees of *Ls(B),* where $m \geq 1$. Then the meaning of the rules of constructing well-formed formulas P[1], ..., P[10] can be explained as follows: for each k from 1 to 10, the rule P[k] determines a partial unary operation *Op[k]* on the set *Degr(B)* with the value being an element of *Ls(B).*

Example. There is a conceptual basis B possessing the following properties. The primary informational universe *X = X(B)* includes the conceptual items *prophase, prometaphase, metaphase, nanaphase, telophase* describing five distinct stages of mitosis (the process of somatic cell division, during which the nucleus also divides) and the conceptual items *China, India, Sri_Lanka*. Hence the value of the partial operation *Op[7]* (it governs the use of logical connectives $\wedge$ - AND and $\vee$ - OR) on the six-tuple

$$< \wedge, prophase, prometaphase, metaphase, nanaphase, telophase >$$

is the string *Semexpr1* of the form

$$(prophase \wedge prometaphase \wedge metaphase \wedge nanaphase \wedge telophase),$$

and the value on the four-tuple $< \vee, China, India, Sri\text{-}Lanka >$ is the K-string $(China \vee India \vee Sri\text{-}Lanka)$.

Let *X(B)* also include the item *mitosis* and the designation of a binary relation *Stages-relation*. Then the K-string

Stages-relation (mitosis, Semexpr1)

is the result of applying the partial operation P[4] to the operands *Stages-relation, mitosis, and Semexpr1.*

Besides, let *X(B)* include the items *article1* (a paper), *article2* (a manufactured article), and *h1 = article2, h2 = Kind1(certn article2, ceramics), h3 = (Country1(certn article2) ≡ (China ∨ India ∨ Sri-Lanka)), h4 = article2 * (Kind1, ceramics) (Country1, (China ∨ India ∨ Sri_Lanka))* are the elements of *Ls(B).* Then the K-string *h4* is the result of applying the partial operation P[8] to the operands *h1, h2, h3*.

Ls(B) includes the string *h5* of the form *certn h4*, being the result of applying the operation P[1] to the operands *certn* and *h4.* The item *certn* denotes the meaning of the expression "a certain", and the string *h5* is interpreted as a designation of a manufactured article being a kind of ceramics and produced in China, India, or Sri-Lanka.

Let h6 be the string of the form *(Height(h5) ≡14/cm).* Then *h6* belongs to *Ls(B)* and is the result of applying the partial operation P[3] to the operands *Height(h5)* and *14/cm.*Thus, the essence of the basic model of the theory of SK-languages is as follows: this model determines a partial algebra of the form (*Degr(B), Operations(B)*) , where *Degr(B)* is the carrier of the partial algebra, *Operations(B)* is the set consisting of the partial unary operations *Op[1], ..., Op[10]* on *Degr(B).*

The volume of the complete description in (Fomichov, 2010a) of the mathematical model introducing, in essence, the operations *Op[1], ..., Op[10]* on *Degr(B)* and, as a consequence, determining the class of SK-languages considerably exceeds the volume of this paper. That is why, due to objective reasons, this model can't be included in this paper.

5 Building Semantic Representations of Complex Biomedical Discourses

The theoretical results stated in chapters 1 - 6 of the monograph (Fomichov, 2010a) provide a framework for following-up the principle of stability of the used language of semantic representations. According to the hypothesis formulated in Chapter 6, the definition of the class of SK-languages enables us to build semantic representations of NL-texts in arbitrary application domains.

During several last years, the significance of natural language processing (NLP) technologies for informatics dealing with the problems of biology and medicine has been broadly recognized. As a consequence, the term BioNLP interpreted as the abbreviation for Natural Language Processing in Biology and Medicine was born. The formalization of natural language semantics is a very acute problem of BioNLP. The attention of many researchers in this field is now attracted by the phenomena of the semantics of sentences and discourses (Prince, Roche, 2009). That is why let's illustrate the new expressive possibilities provided by SK-languages on the example of building a semantic representation of a rather complex discourse pertaining to genetics.

It is known that each individual possesses two genes being responsible for a particular characteristic (e.g., the height) in case of almost all characteristics (or traits). The genes responsible for the contrasting values of a characteristic (for instance, the values "tall" and "short" for the trait "height") are referred to as *allelomorphs*, or *alleles* for short. Some genes have more than two allelic forms, i.e. multiple alleles. In the case of the ABO blood group system, there are at least four alleles (A_1, A_2, B and O). An individual can possess any two of these alleles, which can be the same or different (AO, A_2B, OO, and so on).

With respect to this context, let's consider the discourse D1 = "Alleles are carried on homological chromosomes and therefore a person transmits only one allele for a certain trait to any particular offspring. For example, if a person has the genotype AB, he will transmit to any particular offspring either the A allele or the B allele, but never both or neither" (Turnpenny, Ellard, 2005, p. 198).

Let S1 = "Alleles are carried on homological chromosomes", S2 = "therefore a person transmits only one allele for a certain trait to any particular offspring.", S3 = S1 and S2, S4 = "For example, if a person has the genotype AB, he will transmit to any particular offspring either the A allele or the B allele, but never both or neither".

First of all, we'll construct a possible K-representation (KR) of the sentence S1 as the following string *Semrepr1*:

*(Entails((Alleles-relation (certn gene * (Part, certn person : y1) : x1, certn gene * (Part, y1) : x2) ∧ Location(x1, x3) ∧ Location(x2, x4) ∧ Semantic-descr((x3 ∧ x4), chromosome * (Part, y1))), Homologous(x3, x4)) : P1 ∧ Correspondent-situation(P1, e1)).*

The K-string *Semrepr1* illustrates the following new properties of the theory of SK-languages: the possibilities (a) to construct the compound designations of the notions and of the objects qualified by these notions, (b) to use the logical connective ∧ (AND) for joining not only the semantic representations of the statements but also the designations of the objects, as in case of the substring *(x3 ∧ x4)*, (c) to associate the mark of a situation with the mark of the meaning of sentence describing this situation, as in case of the substring *Correspondent-situation(P1, e1).*

As for the sentence S2, its possible KR will be the string *Semrepr2* of the form

*(Cause(e1, e2) ∧ Correspondent-situation(P2, e2) ∧ (P2 ≡ ∀ y2(person) ∀ y3(person * (Offspring-rel, y2)) ∀ x5(trait1 * (Possessed-by, y2)) ∃ x6 (gene * (Element, Alleles-function(x5))) Situation(e3, transmission1 * (Source1, y2)(Recipient1, y3)(Object-transmitted, x6)) ∧ ¬ ∃ x7 (gene * (Element, Alleles-function(x5))) (Situation(e4, transmission1 * (Source1, y2)(Recipient1, y3)(Object-transmitted, x7)) ∧ ¬ (x7 ≡ x6))))).*

The symbols ∀ and ∃ in the K-string *Semrepr2* are the universal quantifier and of the existential quantifier. We can see here that SK-languages allow for restricting the domain of a logical quantifier with the help of the expressions like *(person * (Offspring-rel, y2)), (trait1 * (Possessed-by, y2)), (gene * (Element, Alleles-function(x5)))*, and so on.

At this point of our analysis we have the appropriate building blocks *Semrepr1* and *Semrepr2* for constructing a possible KR of the sentence S3 as the string *Semrepr3* of the form

(Semrepr1∧ Semrepr2) : P3.

Now let's build a K-representation of the final sentence S4 in the context of the sentence S3. We see that the word combination "For example" from S4 encodes the reference to the meaning of the sentence S3. The system of ten partial operations on conceptual structures proposed by the theory of K-representations contains the operation Op[5] to be used just in such cases. This operation allows for constructing the formulas of the kind *form : var*, where the first operand *form* is a semantic description of an object (in particular, a SR of a statement), and *var* is a variable.

This operation was used for constructing the subformulas *certn gene * (Part, certn person : y1) : x1* and *certn gene * (Part, y1) : x2* of the formula *Semrepr1*; besides, for building the formula *Semrepr 3* from the operands *(Semrepr1∧ Semrepr2)* and *P3.*

Now we can use the variable *P3* as a mark of the meaning of the sentence S3 in the following K-representation *Semrepr4* of the sentence S4:

*Example(P3, Entails(Situation(e4, posessing1 * (Owner1, arbitr person : y4)(Object1, certn genotype * (Designation, 'AB') : x7), Situation(e5, transmission1 * (Source1, y4)(Recipient1, arbitr person * (Offspring, y4) : y5)(Object-transmitted, (certn allele * (Designation, "A") : x8 ∨ certn allele * (Designation, "B") : x9)) ∧ Situation(e6, ¬ transmission1 * (Source1, y4)(Recipient1, y5))(Object-transmitted, (x8 ∧ x9))) ∧ Situation(e7, ¬ transmission1 * (Source1, y4)(Recipient1, y5))(Object-transmitted, NIL))))).*

Here *NIL* is the constant reflecting the meaning of the word "nothing".

Actually, we build a K-representation of the discourse D1 as a string of the form *((Semrepr1 ∧ Semrepr2) : P3 ∧ Semrepr4).*

To sum up, SK-languages allow for describing semantic structure of the sentences with direct and indirect speech and of the discourses with the references to the meanings of phrases and larger parts of a discourse, for constructing compound designations of the notions, sets, and sequences. As far as one can judge on the available scientific literature, now only the theory of K-representations explains the regularities of structured meanings of, likely, arbitrary sentences and discourses pertaining to biomedicine and other fields of professional activity.

6 K-representations of complex biomedical definitions of notions

The analysis shows that the SK-languages possess a number of interrelated expressive mechanisms making them a convenient formal tool for building arbitrarily complex definitions of notions.

Example. Let T1 = "A flock is a large number of birds or mammals (e.g. sheep or goats), usually gathered together for a definite purpose, such as feeding, migration, or defence". T1 may have the K-representation *Expr1* of the form

*Definition1 (flock, dynamic-group * (Qualitative-composition, (bird ∨ mammal * (Examples, (sheep ∧ goal)))), S1, (Estimation1(Quantity(S1), high) ∧ Goal-of-forming (S1, certain purpose * (Examples, (feeding ∨ migration ∨ defence)))).*

The analysis of this formula enables us to conclude that it is convenient to use for constructing semantic representations (SRs) of NL-texts: (1) the designation of a 5-ary relationship *Definition1*, (2) compound designations of concepts (in this example the expressions *mammal * (Examples, (sheep ∧ goal))* and *dynamic-group * (Qualitative-composition, (bird ∨ mammal * (Examples, (sheep ∧ goal))))* were used), (3) the names of functions with the arguments and/or values being sets (in the example, the name of an unary function *Quantity* was used, its value is the quantity of elements in the set being an argument of this function), (4) compound designations of intentions, goals; in this example it is the expression *certain purpose * (Examples, (feeding ∨ migration ∨ defence))* . The structure of the constructed K-representation *Expr1* to a considerable extent reflects the structure of the definition T1.

7 Related approaches to describing semantic structure of NL-texts

The advantages of the theory of SK-languages in comparison with first-order predicates logic, Discourse Representation Theory (DRT) and Episodic Logic (EL) are, in particular, the possibilities: (1) to distinguish in a formal way objects (physical things, events, etc.) and concepts qualifying these objects; (2) to build compound representations of concepts; (3) to distinguish in a formal manner objects and sets of objects, concepts and sets of concepts; (4) to build complex representations of sets, sets of sets, etc.; (5) to describe set-theoretical relationships; (6) to effectively describe structured meanings (SMs) of discourses with references to the meanings of phrases and larger parts of discourses; (7) to describe SMs of sentences with the words "concept", "notion"; (8) to describe SMs of sentences where the logical connective "and" or "or" joins not the expressions-assertions but designations of things, sets, or concepts; (9) to build complex designations of objects and sets; (10) to consider non-traditional functions with arguments or/and values being sets of objects, of concepts, of texts' semantic representations, etc.; (11) to construct formal analogues of the meanings of infinitives with dependent words and, as a consequence, to represent proposals, goals, obligations, commitments.

The items (3) - (8), (10), (11) in the list above indicate the principal advantages of the theory of SK-languages in comparison with the Theory of Conceptual Graphs (TCG). Besides, the expressive possibilities of the new theory are much higher than the possibilities of TCG as concerns the items (1), (2), (9).

The global advantage of the theory of K-representations is that it puts forward a hypothesis about a system of partial operations on conceptual structures being sufficient and convenient for constructing semantic representations (or text meaning representations) of sentences and discourses in NL pertaining to arbitrary fields of humans' professional activity.

8 A strategy of developing a Semantic Web of a new generation

It seems that the Principle of Stability of the used language of semantic representations has much broader sphere of application than the professional activity of any concrete research group or research centre dealing with NLP. There are reasons to believe that following-up this principle can considerably speed-up the progress of the studies bridging a gap between the Semantic Web and NLP. The process of endowing the existing Web with the ability of understanding many natural languages is an objective ongoing process (Wilks, Brewster, 2006). It is a decentralized process, because the research centres in different countries mainly independently develop the translators from particular natural languages to semantic representations (or text meaning representations) and the applied computer systems extracting the meanings from texts in particular natural languages or producing summaries of the collections of texts in particular languages.

The analysis has shown that there is a way to increase the total successfulness, effectiveness of this global decentralized process. In particular, it would be important with respect to the need of cross-language conceptual information retrieval and question - answering. The proposed way is a possible new paradigm for the mainly decentralized process of endowing the existing Web with the ability of processing many natural languages.

The principal idea of a new paradigm is as follows. There is *a common thing* for the various texts in different natural languages. This common thing is the fact that *the NL-texts have the meanings.* The meanings are

associated not only with NL-texts but also with the visual images (stored in multimedia databases) and with the pieces of knowledge from the ontologies.

That is why the great advantages are promised by the realization of the situation when a unified formal environment is being used in different projects throughout the world for reflecting structured meanings of the texts in various natural languages, for representing knowledge about application domains, for constructing semantic annotations of informational sources and for building high-level conceptual descriptions of visual images.

The analysis of the expressive power of SK-languages (see the chapters 3 – 6 of (Fomichov, 2010a)) shows that the SK-languages can be used as a unified formal environment of the kind. It is a direct consequence of the following hypothesis put forward by the author in (Fomichov, 2005, 2007, 2010a, 2010b): SK-languages are a convenient tool of building semantic representations of arbitrarily complex NL-texts (sentences and discourses) pertaining to arbitrary field of professional activity.

This central idea underlies an original strategy of transforming step by step the existing Web into a Semantic Web of a new generation, where its principal distinguished feature would be the well-developed ability of NL processing; it can be also qualified as a Meanings Understanding Web or as a Multilingual Semantic Web. The versions of this strategy are published in (Fomichov, 2009b – 2010b).

9 A new method of designing multilingual semantics-oriented natural language processing systems

The theory of K-representations proposes a collection of formal tools being useful for the design of arbitrarily complex NLPS. Let's consider the basic steps of a new method of designing multilingual semantics-oriented NLPS with the help of formal means for representing intermediary, output, and a part of input data. A multilingual algorithm of semantic-syntactic analysis of NL-texts called *SemSynt1* and introduced in the second part of the monograph (Fomichov, 2010a) was developed in accordance with this new approach and was implemented by means of the language of Web programming PYTHON. The rationale for using PYTHON can be found in (Bird et al., 2009). An explicit description of this approach is given below for the first time.

9.1 Step 1: formalization of additional assumptions about primary items of conceptual level

The content of Step 1 is to introduce additional assumptions about some components of the considered conceptual basis *B*. For instance, NL-texts often include the compound designations of the sets. Hence it would be reasonable to introduce the following assumptions: (a) the component $X = X(B)$ includes a subset *Nat* consisting of all strings of the form $d[1], ..., d[n]$, where $n \geq 1$, for $k = 1,..., n$, $d[k]$ is a digit from the set {'0', '1', '2', ..., '9'} ; (b) the subset $F(B)$ includes the element *Quantity* interpreted as the name of the function «Quantity of the elements of a set » ; (c) the set X(B) includes the elements *Quality-composition* and *Thing-composition* in order to construct, for example, the formulas *Quality-composition(S3, container1 * (Weight, 3/tonna))* and *Thing-composition(S4, (c1 ∧ c2 ∧ c3))*, where *c1, c2, c3* are the marks of the concrete containers with ceramics from India.

Chapter 5 of the monograph (Fomichov, 2010a) can be used as a good introduction to the ways of fixing the additional assumptions about the used system of primary conceptual items.

9.2 Step 2: selecting the form of text meaning representations

The expressive power of the class of SK-languages is very high. SK-languages enable us to build semantic representations of natural language texts in arbitrary application domains. That is why it is necessary to select such collection of the expressive mechanisms of SK-languages that it is useful and convenient to employ these expressive mechanisms for constructing semantic (or text meaning) representations of the input NL-texts. It is the content of the Step 2 of the proposed method of designing multilingual, semantics-oriented NLPS.

Let's consider the examples illustrating the correspondence between the sentences in English, Russian (in Latin transcription), and German and their semantic representations (SR) being the expressions of a certain SK-language, that is, being the K-representations of the input texts. In these examples, the SR of the input text T will be the value of the string variable *Semrepr* (Semantic representation). The considered examples illustrate the

correspondence between the inputs and outputs of the developed algorithm *SemSynt1* , see Chapters 9 and 10 of (Fomichov, 2010a).

Example 1. Let T1eng = "Find a description of the programming language PYTHON on the Web-site http://docs.python.org", T1rus = “Naydite opisanie yazyka programmirovaniya PYTHON na veb-sayte http://docs.python.org”, T1germ = “Finden eine Beschreibung der Programmiersprache PYTHON auf dem Site http://docs.python.org”. Then

*Semrepr = (Command (#Operator#, #Executor#, #now#, e1) ∧ Target (e1, finding1 * (Object-file, certn file1 * (Inf-content, certn description1t * (Focus-object, certn progr-lang * (Name1, “PYTHON”) : x3) : x2))(Web-source, http://docs.python.org))).*

Example 2. Let T2eng = "The international scientific conference “DEXA-2009” took place in Linz, Austria, during August 31 – September 4, 2009”, T2rus = “Mezhdunarodnaya nauchnaya konferentsiya “DEXA-2009” prokhodila v gorode Linz, Avstriya s 31 avgusta po 4 sentyabrya 2009 goda”, T2germ = “Die internationale wissenschaftliche Konferenz “DEXA-2009” war in Linz, Oesterreich waehrend 31. August – 4. September 2009 stattgefunden”. Suppose that the used basic semantic items are constructed with respect to the spelling of English expressions corresponding to these items. For instance, the English words “city” and “town”, the Russian word “gorod”, and the German word group “die Stadt” will be associated with the semantic item *city1*. From the formal standpoint, it means that the elements of the used conceptual basis are built on the basis of English expressions. If this condition is satisfied, the algorithm builds the K-representation

*Semrepr = Situation(e1, taking-place * (Event1, certn conference1 * (Kind-geogr, international) (Kind-focus, science) : x1)(Place1, certn city1 * (Name1, “Linz”)(Belongs-to-country, certn country1 * (Name1, “Austria”) : x3) : x2) (Time-interval, <31.08.2009, 04.09.2009>)) .*

Example 3. Let T3eng = "Did the international scientific conference "DEXA" take place in Hungary?", T3rus = “Prokhodila li mezhdunarodnaya nauchnaya konferentsiya “DEXA” v Vengrii?”, T3germ = “War die internationale wissenschaftliche Konferenz “DEXA” in Ungarn stattgefunden?”. Then

*Semrepr = Question (x1, (x1 ≡ Truth-value (Situation (e1, taking_place * (Time, certn moment * (Earlier, #now#) : t1) (Event1, certn conference * (Type1, international) (Type2, scientific) (Name1, “DEXA”) : x2) (Place, certn country1 * (Name1, “Hungary”) : x3))))).*

Example 4. Let T4eng = "What English scientist discovered penicillin?", T3rus = “Kakoy angliysky uchony otkryl penicillin?”, T3germ = “Welcher English Wissenschaftler hat Penizillin entdeckt?”. Then

*Semrepr = Question (x1, Situation (e1, discovering1 * (Time, certn moment * (Earlier, #now#) : t1) (Agent1, certn scientist * (Country1, England) : x1) (New-object, certn medicine1 * (Name1, “penicillin”) : x2))).*

Example 5. Let T5eng = "What European companies the firm “Rainbow” is cooperating with?", T5rus = “S kakimi evropeyskimi kompaniyami sotrudnichaet firma “Rainbow”, T5germ = “Mit welchen europaeischen Kompanien die Firma “Rainbow” kooperiert?”. Then

*Semrepr = Question (S1, (Qualitative-composition (S1, company1 * (Location, Europe)) ∧ Description(arbitrary company1* (Element, S1) : y1, Situation (e1, cooperation * (Time, #now#) (Agent2, certn company1 * (Name1, “Rainbow) : x1) (Cooper-partner, y1))))).*

Example 6. Let T6 = "Who produces the medicine "Zinnat"?". Then

*Semrepr = Question (x1, Situation (e1, production1 * (Time, #now#) (Agent2, x1) (Product2, certn medicine1 * (Name1, “Zinnat”) : x2))).*

Example 7. Let T7eng = "When and where did Dr. Erik Stein arrive to Zuerich from?", T7rus = “Kogda i otkuda doktor Erik Stein priekhal v Zurikh?”, T7germ = “Wann und woher hat Dr. Erik Stein nach Zuerich gekommen?”. Then

*Semrepr = Question ((x4 ∧ x1), (Situation (e1, arrival *(Time, certn moment * (Earlier, #now#) : t1)*

*(Start-location, x1)(Agent1, certn person *(Qualif, Ph.D.)(Name, “Erik”)(Surname, “Stein”) : x2) (Final-location, certn city1 * (Name1, “Zuerich”) : x3)) ∧ (x4 ≡ t1)).*

Example 8. Let T8eng = "How many countries did participate in the Olympic Games - 2008?", T7rus = "Skolko stran uchastvovalo v Olimpiyskikh Egrakh – 2008", T7germ = "Wieviel Laender haben an den Olympischen Spielen – 2008 teilgenommen?". Then

*Semrepr = Question (x1, ((x1 ≡ Numb(S1)) ∧ Qualitative-composition (S1, country1) ∧ Description (certn country1 * (Element, S1) : y1, Situation (e1, participation1 * (Time, certn moment * (Earlier, #now#) : t1) (Agent1, y1) (Time, 2008/year)(Event1, certn olymp-game : x2))).*

Example 9. Let T9eng = "How many times did Professor Bill Jones visit France?", T7rus = "Skolko raz professor Bill Jones posetil Frantsiu", T7germ = "Wieviel Mal hat Herr Professor Bill Jones Frankreich besucht?". Then

*Semrepr = Question (x1, ((x1 ≡ Numb (S1)) ∧ Qualitative-composition (S1, sit) ∧ Description (arbitrary sit * (Element, S1) : e1, Situation (e1, visiting * (Time, certn moment * (Earlier, #now#) : t1) (Agent1, certn person * (Qualif, professor)(Name, "Bill")(Surname, "Jones") : x2) (Place2, certn country * (Name1, "France") : x3)))).*

9.3 Step 3 of the new approach: formation of semantic-syntactic components of a linguistic database

Chapter 7 of the monograph (Fomichov, 2010a) contains an original, broadly applicable mathematical model of a linguistic database (LDB). This model formalizes the structure of a linguistic database allowing for setting up various conceptual relations, e.g. 'Verb + Preposition + Noun', 'Verb + Noun', 'Noun1 + Preposition + Noun2', 'Numeral + Noun', 'Adjective + Noun', 'Noun1 + Noun2', 'Participle + Noun', 'Participle + Preposition + Noun', 'Interrogative pronoun + Verb', 'Preposition + Interrogative pronoun + Verb', 'Interrogative Adverb + Verb', 'Verb + Numerical Value Representation' (a number representation + a unit of measurement representation). The model defines a class of formal objects called *linguistic bases* (l.b.). Each l.b. *Lingb* is a mathematical representation of a morphological datatabase, of some functions corresponding to the subsystems of a morphological analyzer, and of semantic-syntactic components of the LDB.

The content of the considered Step 3 is to form several semantic-syntactic components of a LDB. The first component *Lsdic* is the set of finite sequences of the form

(i, lec, pt, sem, st[1], …, st[k], comment),

where $i \geq 1$ is the ordered number of the k+5-tuple (we need it to organize the loops in the algorithms of processing NL-texts), and the rest of the components are interpreted in the following way: *lec* is an element of the set of basic lexical units *Lecs* for the considered morphological basis; *pt* is a designation of the part of speech for the basic lexical unit *lec*; the component *sem* is a string that denotes one of the possible meanings of the basic lexical unit *lec*.

For instance, the verb "to enter" has, in particular, the following two meanings: (1) entering a learning institution (in the sense "becoming a student of this learning institution"); (2) entering a space object ("Yves has entered the room", etc.). So one system from a possible lexico-semantic dictionary will have, as the beginning, the sequence *i1, enter, verb, entering1,* and the other will have, as the beginning, the sequence *i2, enter, verb, entering2.*

The component *sem* can be a complex string being an expression of the SK-language Ls(B) for a certain conceptual basis.

Example. If *lec* = 'France', then *sem* can be the K-string *certn country * (Name1, 'France');* if *lec* = 'green', then *sem* can be the K-string *Colour (z, green),* where *z* is a variable denoting an entity with the property "green".

The number *k* is the semantic dimension of the considered sort system, that is, *k* is the maximal number of the different "semantic axes" used to describe one entity in the considered application domain.

Example. Let us consider the concepts "a firm" and "a university". We can distinguish three semantic contexts of word usage associated with these concepts. Firstly, a firm or a university can develop a tool, a technology etc., so the sentences with these words can realize the semantic coordinate "intelligent system". Secondly, we can say: "This firm is situated near the metro station "Taganskaya," and then this phrase realizes the semantic coordinate "spatial object". Finally, the firms and institutes have the directors. We can say, for example: "The director of this firm is Alexander Semenov." This phrase realizes the semantic coordinate "organization". In the considered examples, we'll presume that semantic dimension of the considered sort systems equals is equal to four or three.

The elements *st[1], …, st[k]* are the different semantic coordinates of the entities characterized by the concept *sem*. For example, if *sem = firm*, then *st[1] = ints* (intelligent system), *st[2] = space.ob* (space object), *st[3] = org* (organization), *k =3*. The compo-nent *comment* is either a natural language description of a meaning associated with the concept *sem* or an empty element *nil.*

The second semantic-syntactic component of a LDB is called the *dictionary of verbal-prepositional frames*, it contains such templates (in other terms, frames) that enable us to represent the necessary conditions of realizing a specific thematic role in the combination "Verbal form + Preposition + Dependent word group". An example in the subsection 9.4.3 illustrates the structure of such templates.

The third semantic-syntactic component of a LDB is called the *dictionary of prepositional frames*, it contains the templates allowing for representing the necessary conditions of realizing a specific relation in the combination of the form "Noun 1 + Preposition + Noun 2" or of the form "Noun 1 + Noun 2".

Example. Let us assume that *Expr* is the expression "an article by Professor Novikov", and a linguistic database includes a template representing the sequence of the form

(*k1, 'by', inf.ob, ints, 1, Authors, "a poem by H. Heine'*),

were *ints* is the sort "intelligent system", 1 is the code of common case in English. We may connect the sorts *ints* and *dyn.phys.ob* (dynamic physical object) with the basic lexical unit "professor". We see that the expression *Expr* is compatible with this template having the number *k1*.

9.4 Step 4: development of an algorithm transforming the input texts into their matrix semantic – syntactic representations

9.4.1 Step 4-1: Building morphological representation of an input text

Skipping mathematical details, we'll suppose that a morphological representation (MR) of a text T with the length *nt* is a two-dimensional array *Rm* with the names of columns *base* and *morph* (more exactly, *morph* is the designation of a group of colums), where the elements of the array rows are interpreted in the following way. Let *nmr* be the number of the rows in the array *Rm* that was constructed for the text T, and *k* be the number of a row from the array Rm, i.e. $1 \le k \le nmr$. Then *Rm[k, base]* is the basic lexical unit (the lexeme) corresponding to the word in the position p from the text T. Under the same assumptions, *Rm[k, morph]* is a sequence of the collections of the values of morphological characteristics (or features) corresponding to the word in the position *p*.

Example. Let T1 be the question "Did the management board of the firm "Rainbow" change in May?", and T1germ be the same question in German "Hat der Verwaltungsrat der Firma "Rainbow" in Mai veraendernt sich?". Then the morphological representation Rm1 of T1 consists of the rows *(change, md[1]), (management-board, md[2]), (of, md[3]), (firm, md[4]), (in, md[5]), (May, md[6])*, where *md[1], …, md[6]* are the sequences of the values of morphological properties associated with the corresponding lexical units from T1. Similarly, the morphological representation Rm2 of T1germ consists of the rows *(sich-veraendern, mdg[1]), (Verwaltungsrat, mdg[2]), (Firma, mdg[3]), (in, mdg[4]), (Mai, mdg[5])*, where *mdg[1], …, mdg[5]* are the sequences of the values of morphological properties associated with the corresponding lexical units from T1germ.

9.4.2 Step 4-2: Building classifying representation of an input text

Classifying representation. From informal point of view, we will say that a classifying representation (CR) of the text T coordinated with the morphological representation *Rm* of the text T is a two-dimensional array *Rc* with the number of the rows *nt* and the column with the indices *unit, tclass, subclass, mcoord*, in which its elements are interpreted in the following way. Let *k* be the number of any row in the array *Rc* i.e. $1 \le k \le nt$. Then *Rc[k, unit]* is one of elementary meaningful units of the text T, i.e. if $T = t_1 \dots t_{nt}$, then such position *p*, where $1 \le p \le nt$, can be found that $Rc[k, unit] = t_p$. If *Rc[k, unit]* is a word, then *Rc[k, tclass], Rc[k, subclass], Rc[k, mcoord]* are correspondingly a part of

speech, a subclass of the part of speech, the sequences of the values of morphological properties. If *Rc[k, unit]* is a construct (i.e. a value of a numeric parameter), then *Rc[k, tclass]* is the string *constr*, *Rc[k, subclass]* is the designation of the subclass of informational units corresponding to this construct, *Rc[k, mcoord] = 0.*

Example. Let T1 = "Did the management board of the firm "Rainbow" change in May?". Then a classifying representation *Rc1* of the text T1 coordinated with the morphological representation *Rm1* of T1 may be the following array:

unit	tclass	Subclass	mcoord
did-change	verb	verb-in-indic-mood	1
the management-board	noun	common-noun	2
of	prep	nil	3
the-firm	noun	common-noun	4
"Rainbow"	artif-name	nil	0
in	prep	nil	5
May	noun	proper-noun	6
?	marker	nil	0

If T1germ ="Hat der Verwaltungsrat der Firma "Rainbow" in Mai veraendernt sich?", then a classifying representation *Rc2* of the text T1germ coordinated with the MR *Rm2* of T1 may have the following form:

unit	tclass	subclass	mcoord
hat-veraendernt-sich	verb	verb-in-indic-mood	1
der-Verwaltungsrat	noun	common-noun	2
der-Firma	noun	common-noun	3
"Rainbow"	artif-name	nil	0
in	prep	nil	4
Mai	noun	proper-noun	5
?	marker	nil	0

9.4.3 Step 4-3: Building the projections of the components of a linguistic basis on the input text

Let *Lingb* be a linguistic basis (see Chapter 7 of (Fomichov, 2010a)), and *Dic* be one of the following components of *Lingb*: the lexico-semantic dictionary *Lsdic*, the dictionary of verbal-prepositional semantic-syntactic frames *Vfr*, the dictionary of prepositional semantic-syntactic frames *Frp* (see Chapter 8 of (Fomichov, 2010a)). Then the projection of the dictionary *Dic* on the input text T is a two-dimensional array whose rows represent all data from *Dic* linked with the lexical units from T .

Let's introduce the following denotations: *Arls* is the projection of the lexico-semantic dictionary *Lsdic* on the input text T; *Arvfr* is the projection of the dictionary of verbal-prepositional frames Vfr on the input text T ; *Arfrp* is the projection of the dictionary of prepositional frames *Frp* on the input text T.

Example. Let T1 = "Did the management board of the firm "Rainbow" change in May?". Then the projection of the lexico-semantic dictionary *Lsdic* on the input text T1 may be the following two-dimensional array:

ord	sem	st1	st2	st3	comment
1	change1	event	nil	nil	Yves has changed 700 franks
1	change2	event	nil	nil	The city has changed very much in the 1990s - 2000s
2	manag-board	org	ints	phys.ob	Management board of a company
4	Company1	org	ints	phys.ob	The firm IBM

5	"Rainbow"	artif-name	nil	nil	nil
7	May	month-value	nil	nil	nil

Here the elements of the column *ord* are the numbers of the corresponding rows of the classifying representation *Rc1*; the sorts *org, ints, phys.ob* are interpreted as the designations of the notions "an organization", "an intelligent system", and "a physical object". The sorts *ints* and *phys.ob* characterize from different standpoints the elements (people) of any firms and management boards of the firms.

The verb "to change" has more than two meanings. That is why for real computer applications this array will be a subarray of the projection of the lexico-semantic dictionary *Lsdic* on the input text T1.

Example. If T1 = "Did the management board of the firm "Rainbow" change in May?", the projection of the dictionary of verbal-prepositional semantic-syntactic frames *Vfr* on the input text T1 *Arvfr1* may include the following subarray *Arvfr1fragm*:

nb	semsit	lang	fm	refl	vc	trole	sprep	grc	str	expl
1	change1	eng	indic	nrf	actv	Money-sum	nil	1	money-value	ex1
1	change1	eng	indic	nrf	actv	Location	nil	1	space-ob	ex2
1	change1	eng	indic	nrf	actv	Time	on	0	moment	ex3
1	change2	eng	indic	nrf	actv	Focus-object	nil	0	phys.ob	ex4
1	change2	eng	indic	nrf	actv	Start-time	since	0	moment	ex5
1	change2	eng	indic	nrf	actv	Time-interval	during	0	moment	ex6

Here the elements *eng, indic, nrf, actv* are interpreted as the values *English, indicative-mood, non-reflexive, active-voice* of the properties *language, form-of-verb, reflexivity, voice*; the elements *Money-sum, Location, Time, Focus-object, Start-time, Time-interval* are the designations of thematic roles (or conceptual cases); ex1 = "(Yves) has changed 700 franks", ex2 = "(Yves) has changed (700 franks) in the exchange office No. 14", ex3 = "(Yves) has changed (700 franks in the exchange office No. 14) on the 4th of March", ex4 = "Mary has changed (very much since last summer)"; ex5 = "(Mary) has changed (very much) since last summer"; ex6 = "The town has changed very much during the 2000s)". The fragments outside the parentheses are just the fragments where the considered thematic role (in other terms, a conceptual case) is realized. The fragments inside the parentheses only complement the fragments of the first kind in order to form a sentence.

9.4.4 Step 4-4: Constructing matrix semantic-syntactic representation of the input text

Following (Fomichov, 2010a), let's consider a new data structure called *a matrix semantic-syntactic representation (MSSR)* of a natural language input text T. This data structure will be used for representing the intermediate results of semantic-syntactic analysis on a NL-text. A MSSR of a NL-text T is a string-numerical matrix *Matr* with the indices of columns or the groups of columns

locunit, nval, prep, posdir, reldir, mark, qt, nattr ,

it is used for discovering the conceptual (or semantic) relations between the meanings of the fragments of the text T, proceeding from the information about linguistically correct short word combinations. Besides, a MSSR of a NL-text allows for selecting one among several possible meanings of an elementary lexical unit. The number of the rows of the matrix *Matr* equals to *nt* - the number of the rows in the classifying representation *Rc*, i.e. it equals to the number of elementary meaningful text units in T.

Let's suppose that *k* is the number of arbitrary row from MSSR *Matr*. Then the element *Matr[k, locunit]*, i.e. the element on the intersection of the row k and the column with the index *locunit* is the least number of a row from the array *Arls* (it is the projection of the lexico-semantic dictionary *Lsdic* on the input text T) corresponding to the elementary meaningful lexical unit *Rc[k, unit]*. It is possible to say that the value *Matr[k, locunit]* for the *k*-th elementary meaningful lexical unit from T is the coordinate of the entry into the array *Arls* corresponding to this lexical unit .

The column *nval* of *Matr* is used as follows. If *k* is the ordered number of arbitrary row in *Rc* and *Matr* corresponding to the elementary meaningful lexical unit, then the initial value of *Matr[k, nval]* is equal to the quantity of all rows from *Arls* corresponding to this lexical unit; that is, corresponding to different meanings of

this lexical unit. When the construction of *Matr* is finished, the situation is to be different for all lexical units with several possible meanings: for each row of *Matr* with the ordered number *k* corresponding to a lexical unit, *Matr[k, nval]* = 1. because a certain meaning was selected for each elementary meaningful lexical unit.

For each row of *Matr* with the ordered number *k* associated with a noun or an adjective, the element in the column *prep* (preposition) specifies the preposition (possibly, the void, or empty, preposition *nil*) relating to the lexical unit corresponding to the k-th row.

Let's consider the purpose of introducing the column group

$$posdir\ (posdir_1,\ posdir_2,\ \ldots,\ posdir_n),$$

where *n* is a constant between 1 and 10 depending on the program implementation. Let $1 \le d \le n$. Then we will use the designation *Matr[k, posdir, d]* for an element located at the intersection of the *k*-th row and the d-th column in the group *posdir*. If $1 \le k \le nt,\ 1 \le d \le n,$ then *Matr[k, posdir, d] = m*, where *m* is either 0 or the ordered number of the *d*-th lexical unit *wd* from the input text T, where *wd* governs the text unit with the ordered number k.

There are no governing lexical units for the verbs in the principal clauses of the sentences, that is why for the row with the ordered number *m* associated with a verb, *Matr[m, posdir, d] = 0* for any *d* from 1 to *n*. Let's agree that the nouns govern the adjectives as well as govern the designations of the numbers (e.g. "5 scientific articles"), cardinal numerals, and ordinal numerals. The group of the columns *reldir* consists of semantic relations whose existence is reflected in the columns of the group *posdir*. For filling in these columns, the templates (or frames) from the arrays *Arls*, *Arvfr*, *Arfrp* are to be used; the method can be grasped from the analysis of the algorithm *BuildMatr1* constructing a matrix semantic-syntactic representation of an input NL-text stated in (Fomichov, 2010a).

The column with the index *mark* is to be used for storing the variables denoting the different entities mentioned in the input text (including the events indicated by verbs, participles, gerunds, verbal nouns). The column *qt* (quantity) equals either to 0 or to the designation of the number situated in the text before a noun and connected to a noun. The column *nattr* (number of attributes) equals either to 0 or to the quantity of adjectives related to a noun presented by the *k*-th row, if we suppose that *Rc[k, unit]* is a noun.

10 Step 5 of the new method: development of an algorithm of semantic assembly

The content of the Step 5 is to use a matrix semantic-syntactic representation of a NL-text T as an intermediary data structure for constructing a semantic representation of T being an expression of a certain SK-language (that is, being a K-representation of T). The algorithm of semantic assembly *BuildSem1* described in Chapter 10 of (Fomichov, 2010a) gives an example of realizing this step of designing a NLPS.

Example. Let T1 be the question "Did the management board of the firm “Rainbow” change in May?", and T1germ be the same question in German “Hat der Verwaltungsrat der Firma “Rainbow” in Mai veraendernt sich?”. Then it is possible to associate both with T1 and with T1germ the same K-representation *Semrepr* of the form

*Question (x1, (x1 ≡ Truth-value(Situation(e1, change2 * (Focus-object,*
*certn manag-board * (Assoc-company, certn company1 * (Name1, “Rainbow”) : x3) : x2)*
(Time, Last-month(May, current-year))))) .

11 Conclusions

The theory of K-representations was developed as a tool for dealing with numerous questions of studying semantics of arbitrarily complex natural language texts: both sentences and discourses. Grasping the main ideas and methods of this theory requires considerably more time than it is necessary for starting to construct the formulas of the first-order predicates logic. However, the efforts aimed at studying the foundations of the theory of K-representations would be highly rewarded. Independently on an application domain, a designer of a NLPS will have a convenient tool for solving various problems.

A new method of developing the algorithms of semantic-syntactic analysis of NL-texts was described above. The method has a number of significant advantages in comparison with other known methods of developing the algorithms of the kind. *Firstly*, the method was used for developing the algorithm *SemSynt1* described in

(Fomichov, 2010a). The explicitness and fullness of the description of the algorithm *SemSynt1* is many times higher than it is typical for the scientific publications on this problem (see, e.g., the paper (Popescu et al., 2003)). *Secondly*, the approach doesn't foresee the construction of a pure syntactic representation of the analyzed NL-text: it is oriented at discovering the semantic relations between the elementary meaningful units of a text.

Thirdly, the algorithm *SemSynt1* is multilingual in the following sense. This algorithm allows for using the same semantic-syntactic part of a linguistic database for English, German, and Russian languages. The algorithm *SemSynt1* contains the fragments meaning the calls of language-dependent auxiliary procedures. These procedures find several parts of a compound verbal form and join them into one elementary meaningful text unit, associate a preposition with a noun, etc. However, the discovery of possible semantic relations between the elementary meaningful text units is language-independent, and this promises economic advantages in case when the significant information may be obtained from the sources in several natural languages.

It seems that the method stated above together with the algorithm *SemSynt1* (as a substantial example of using this method) can be used as a framework for designing multilingual semantics-oriented analyzers of NL-texts and for obtaining much more detailed documentation of the algorithms as it is usually done.

References

BIRD S., KLEIN E., LOPER E. (2009). *Natural Language Processing with Python.* O'Reily.

CIMIANO P., HAASE P., HEIZMANN J., MANTEL M. (2007). ORAKEL: A portable natural language interface to knowledge Bases. *Technical report*, Institute AIFB, University of Karlsruhe, Germany.

DUKE A., GLOVER T., DAVIES J. (2007). Squirrel: An advanced semantic search and browse facility. In: *Proc. of the 4th European Semantic Web Conference. Innsbruck, Austria.*

FOMITCHOV V. (1984). Formal systems for natural language man-machine interaction modeling. In: *Ponomaryov, V.M. (ed.) Artificial Intelligence. Proc. of the IFAC Symposium, Leningrad, USSR, 4-6 October 1983 (IFAC Proc. Series, 1984, No. 9)*. Oxford, UK; New York: Pergamon Press, 203-209.

FOMICHOV V.A. (1992). Mathematical models of natural-language-processing systems as cybernetic models of a new kind. *Cybernetica (Belgium),* Vol. XXXV, 63-91.

FOMICHOV V.A. (1993a). Towards a mathematical theory of natural-language communication. *Informatica. An Intern. J. of Computing and Informatics (Slovenia)*, Vol.17, 21-34.

FOMICHOV V.A. (1993b). K-calculuses and K-languages as powerful formal means to design intelligent systems processing medical texts. *Cybernetica (Belgium),* Vol. XXXVI, 161-182.

FOMICHOV V.A. (1994). Integral Formal Semantics and the design of legal full-text databases. *Cybernetica (Belgium),* Vol. XXXVII, 145-177.

FOMICHOV V.A. (2005). *The Formalization of Designing Natural Language Processing Systems.* Moscow: MAX Press (in Russian).

FOMICHOV V.A. (2007). *Mathematical Foundations of Representing the Content of Messages Sent by Computer Intelligent Agents.* Moscow: State University – Higher School of Economics, Publishing House "TEIS" (in Russian).

FOMICHOV V.A. (2009a) Theory of K-representations as a Source of an Advanced Language Platform for Semantic Web of a New Generation. *Web Science Overlay J. On-line Proceedings of the First International Conference on Web Science, Athens, Greece, March 18-20, 2009*; available at http://journal.webscience.org/221/1/websci09_submission_128.pdf.

FOMICHOV V.A. (2009b). A Scheme and Formal Tools for Transforming the Existing Web into Semantic Web of a New Generation. In: *Pre-Conference Proceedings of the Focus Symposium on Knowledge Management Systems (August 4, 2009, Focus Symposia Chair: Jens Pohl) in conjunction with InterSymp-2009, 21st International Conference on Systems Research, Informatics and Cybernetics, August 3 – 7, 2009, Baden-Baden, Germany),* Collaborative Agent Design Research Center, California Polytechnic State University, San Luis Obispo, CA, USA, 39-50.

FOMICHOV V.A. (2010a). *Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms.* New York, Dordrecht, Heidelberg, London : Springer.

FOMICHOV V.A. (2010b). Theory of K-representations as a Comprehensive Formal Framework for Developing a Multilingual Semantic Web. *Informatica. An International Journal of Computing and Informatics (Slovenia),* Vol. 34, No. 3, 387-396.

FRANK A., KRIEGER H.-U., XU F., USZKOREIT H., CRYSMANN B., JRG B., SCHAEFER U. (2007). Question answering from structured knowledge sources. *J. of Applied Logic*, 5 (1), 20-48.

POPESCU A.-M., ETZIONI O., KAUTZ H. (2003). Towards a theory of natural language interfaces to databases. In: *Proc. of the 8th International Conference on Intelligent User Interfaces,* Miami, FL, 149-157.

PRINCE V., ROCHE M., eds (2009). *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration.* IGI Global.

TURNPENNY P.D., ELLARD S. (2005). *Emery's Elements of Medical Genetics. Twelfth Edition*. Edinburgh, London, New York, Oxford, Sydney, Toronto: Elsevier Limited.

WILKS Y., BREWSTER C. (2006). *Natural Language Processing as a Foundation of the Semantic Web. Foundations and Trends in Web Science,* Vol. 1, No. 3. Hanover, MA; Delft: now Publishers Inc.

Sentence Generation: Input, Algorithms and Applications

Claire Gardent

CNRS/LORIA, Nancy (France)

Abstract

(Joint work with Paul Bedaride, Eric Kow, Shashi Narayan and Laura Perez-Beltrachini)

Sentence Generation maps abstract linguistic representations into sentences. A necessary part of any natural language generation system, sentence generation has also recently received increasing attention in applications such as transfer based machine translation (cf. the LOGON project) and natural language interfaces to knowledge bases (e.g., to verbalise, to author and/or to query ontologies).

One outstanding issue in Sentence Generation is what it starts from. What is the abstract linguistic representation it generates from? In my talk, I will explore sentence generation from two main input formats (flat semantic formulae and dependency structures) and discuss their impact on efficiency, algorithms and applications.

I will start by describing an algorithm that generates from flat semantic formulae, explain why it is computationally intractable and presenting ways of optimising it to make it usable in practice. I will then show how this algorithm can be used to generate paraphrases; to support error mining and to generate teaching material for language learners from an ontology.

In the second part of the talk, I will focus on generation from dependency structures. Based on the input data recently made available by the Generation Challenges Surface Realisation Shared Task, I will show how the algorithm previously used to generate from flat semantic formulae can be adapted to generate from dependency structures. I will moreover discuss various issues raised by the GenChal data such as, missing lexical entries and mismatches between dependency and grammar structures.

Bio of Claire Gardent

Claire Gardent is a senior researcher at the French National Center for Scientific Research (CNRS). She graduated in linguistics at the University of Toulouse in 1986, received an MSc in Artificial Intelligence from the University of Essex in 1987 and defended a PhD in Cognitive Science at the University of Edinburgh in 1991. From 1991 to 2000, she worked as a reseacher at the Universities of Utrecht and Amsterdam (The Netherlands), Clermont-Ferrand and Sarrebruecken (Germany). Since 2001 she has been working for the CNRS at the Lorraine Laboratory for Research in Computer Science (LORIA) in Nancy, France.

Claire Gardent's research focuses on the computational treatment of natural language meaning. She has worked on the automatic acquisition of lexical resources for French, on syntactic parsing and semantic role labelling and on text generation. Recently, she has become interested in in exploring the interaction between virtual worlds and natural language processing.

Claire Gardent has published a textbook on analysis and generation (with Karine Baschung) and about 100 articles in journals and conference proceedings. She has been nominated Chair of the European Chapter for the Association of Computational Linguistics (EACL), editor in chief of the french journal "Traitement Automatique des Langues" and member of the editorial board of the journals "Computational Linguistics", "Journal of Semantics". Each year she is on the programme committee of half a dozen international conferences or workshops, she also acted as scientific chair for various international conferences (EACL), workshops (TAG+, ENLG, DIALOR, SIGDIAL) and summer schools (ESSLLI).

Theorie et Praxis
Une optique sur les travaux en TAL sur le discours et le dialogue

Nick Asher

LILac, IRIT, Université Paul Sabatier

Abstract

Discourse parsing is a relatively new field and it differs from parsing in syntax in its pedegree. Parsing and computational models of syntax have the benefit of 50 years of research in generative syntax and reactions to it. Discourse parsing has on the other hand little conceptual help from linguistics or philosophy. Though impressive gains have been registered in discourse parsing with superficial features, theoretical not really come to grips with the theoretical underpinnings of text interpretation, and its interaction especially with lexical semantics, a rather neglected branch of formal semantics. In my talk I will assess the interaction between theoretical linguistics, formal methods, and experimental work on discourse structure and interpretation. Sounding a note of optimism, I will then turn to assessing the situation for the computational analysis of dialogue. I will argue that the view that we are saddled with from Grice and the philosophy of the seventies is inadequate and is great need of revision from work on communication from economics and theoretical computer science

Bio of Nicholas Asher

Nicholas Asher est directeur de recherche au CNRS depuis 2006. Il a eu son doctorat en philosophie à Yale University en 1982 et puis a été Professeur à l'University of Texas at Austin pendant 24 ans. Il a travaillé longtemps en sémantique et pragmatique formelle et s'interesse surtout au discours et dialogue. Il a développé une théorie de l'interprétation du discours basée sur la sémantique dynamique qui s'appelle la "Segmented Discourse Representation Theory" ou SDRT, sur lequel il a écrit deux livres, *Reference to Abstract Objects in Discourse* (Kluwer, 1993) et *Logics of Conversation* (avec Alex Lascarides, Cambridge 2003). Il a aussi publié une trentaine d'articles sur la SDRT dans des revues internationales. Un autre thème de recherche est la sémantique lexicale et la composition de sens, sur lequel il vient de publier un livre, *Lexical Meaning in Context*, avec Cambridge University Press. Il s'interesse aussi à la validation empirique des théories linguistiques et aux travaux sur corpus ainsi qu'aux techniques d'apprentissage sur les données structurées. Vétéran d'un projet d'annotation discursive sur des textes en français, ANNODIS, il se lance maintenant sur un projet ERC sur la conversation stratégique et une révision des fondements de la vision Gricéenne de la communication humaine.

Fouille de textes et applications

Patrons de phrase, raccourcis pour apprendre rapidement à parler une nouvelle langue

Michael Zock, Guy Lapalme

(1) CNRS – LIF (Aix-Marseille II)
Laboratoire d'Informatique Fondamentale
Case 901, 163 avenue de Luminy,
F-13288 Marseille Cedex 9

(2) RALI-DIRO
Université de Montréal
CP 6128, Succ. Centre-Ville
Montréal, QC Canada H3C 3J7
michael.zock@lif.univ-mrs.fr, lapalme@iro.umontreal.ca

Résumé

Nous décrivons la création d'un environnement web pour aider des apprenants (adolescents ou adultes) à acquérir les automatismes nécessaires pour produire à un débit "normal" les structures fondamentales d'une langue. Notre point de départ est une base de données de phrases, glanées sur le web ou issues de livres scolaires ou de livres de phrases. Ces phrases ont été généralisées (remplacement de mots par des variables) et indexées en termes de buts pour former une arborescence de patrons. Ces deux astuces permettent de motiver l'usage des patrons et de créeer des phrases structurellement identiques à celles rencontrées, tout en étant sémantiquement différentes. Si les notions de 'patrons' ou de 'phrases à trou implicitement typées' ne sont pas nouvelles, le fait de les avoir portées sur ordinateur pour apprendre des langues l'est. Le système étant conçu pour être ouvert, il permet aux utilisateurs, concepteurs ou apprenants, des changements sur de nombreux points importants : le nom des variables, leurs valeurs, le laps de temps entre une question et sa réponse, etc. La version initiale a été développée pour l'anglais et le japonais. Pour tester la généricité de notre approche nous y avons ajouté relativement facilement le français et le chinois.

Abstract

We describe a web application to assist language learners (teenagers or adults) to acquire the needed skills to produce at a 'normal' rate the fundamental structures of a new language, the scope being the survival level. The starting point is a database of sentences gleaned in textbooks, phrasebooks, or the web. We propose to extend the applicability of these structures by generalizing them: concrete sentences becoming productive sentence patterns. In order to produce such generic structures (schemata), we index the sentences in terms of goals, replacing specific elements (words) of the chain by more general terms (variables). This allows the user not only to acquire these structures, but also to express his/her own thoughts. Starting from a communicative goal, he instantiates the variables of the associated schema with words of his choice. We have developed a prototype for English and Japanese, adding Chinese and French without too many problems.

Mots-clés : apprentissage de langues, production de langage, livres de phrases, patrons, schéma de phrase, structures fondamentales

Keywords: foreign language learning, language production, phrasebook, sentence patterns, basic structure

1 Introduction

Produire du langage consiste, schématiquement parlant, à faire dans l'ordre les trois choses suivantes : concevoir un message, le traduire en langue, communiquer ce résultat sous forme graphique ou orale. Ceci semble simple, car tout le monde parle au moins une langue, et deux tiers de la population sur cette planète est bilingue. Pourtant, il n'y a pas l'ombre d'un doute : s'exprimer spontanément et à un débit normal en langue est une tâche difficile. Étant donné une intention de communication ('inviter quelqu'un au restaurant' ; 'raconter un film'), on doit concevoir un, voire plusieurs messages, (conceptualisation), trouver des mots convenables (lexicalisation) insérer ces éléments au bon endroit d'un schéma de phrase à déterminer également (syntaxe), effectuer des flexions et accords (morphologie), prononcer ce résultat (articulation) tout en commençant à planifier le segment suivant (idée).

La tâche n'est donc pas aussi aisée que cela en avait l'air et la difficulté tient essentiellement à trois facteurs : la limitation des ressources du système de traitement (cerveau, surcharge cognitive), la complexité du processus (parallélisme, multitâches) et le volume de données hétérogènes à unifier. En effet, la tâche exige un très grand nombre de choix en très peu de temps. Les informations à traiter sont distribuées à travers plusieurs niveaux. Elles sont donc de nature différente (conceptuel, linguistique, moteur). Les choix peuvent avoir des conséquences multiples, imprévisibles et interdépendantes. Enfin, les éléments à utiliser (faits, mots) doivent être localisés dans un énorme réservoir : base de faits/connaissances (encyclopédie), dictionnaire mental. Si jamais une de ces étapes tarde, ou pire, si la recherche échoue, on assiste à des pauses plus ou moins prononcées, pouvant aller jusqu'au silence total. Ceci peut facilement arriver dans le cas du mot sur le bout de la langue. Imaginez un instant comment trouver un mot particulier parmi les, disons, 30 à 60 000 mots stockés (les chiffres avancés variant selon les auteurs). C'est chercher la fameuse aiguille dans une meule de foin. La performance est impressionnante, équivalant à la consultation manuelle d'un dictionnaire comme *Le Grand Robert* trois fois par seconde pendant plusieurs heures.[1]

Si parler est déjà difficile en langue maternelle, s'exprimer couramment en langue étrangère est une véritable prouesse. Bien entendu, ce n'est pas quelque chose d'inné. Cela a été appris. La question est de savoir comment aider quelqu'un ayant cet objectif. Voici le but de notre travail.

Concernant la méthode, nous poursuivons actuellement deux directions. D'un côté, nous sommes en train de construire une grande *base de phrases multilingues* — environ 40.000 phrases, soit en anglais-japonais (A-J), soit français-japonais (F-J)— de l'autre, nous construisons une *base de phrases d'exercices*, destinée à des apprenants cherchant à acquérir l'habileté nécessaire pour s'exprimer à un débit 'normal' dans une nouvelle langue. Si les phrases de la première base sont assez variées (leur lien étant essentiellement du type 'thématique'), celles de la seconde varient sur très peu d'éléments, ce qui est normal, dans la mesure où leur fonction est de montrer un invariant ou une régularité de la langue.

Les phrases de ces deux bases viennent pour la plupart de livres scolaires, de sites comme Tatoeba (http://tatoeba.org/fre) et de *phrasebooks*[2]. Comme chaque couple de langue contient des phrases différentes (les phrases du couple A-J sont différentes de celles du couple F-J), il faut les traduire pour les autres langues, pour permettre ensuite l'accès à partir de n'importe laquelle de ces langues. C'est ce que nous avons commencé à faire, en ajoutant une 4ème langue (Chinois). La traduction faite, on pourra donc non seulement travailler sur chacune des langues en faisant des exercices (travail décrit plus bas), mais également voir la traduction des phrases dans n'importe laquelle de ces langues.

A l'avenir nous aimerions étendre ces deux bases en les enrichissant (plus ou moins) automatiquement, puis établir un pont entre les deux, pour que l'ensemble des phrases puisse être utilisé par l'apprenant pour s'exercer soit en mode traduction, soit en mode production de phrase. Nous présentons ici un générateur d'exercices dont la vocation est d'aider des adolescents ou des adultes à apprendre à s'exprimer à un débit 'normal' en langue étrangère. Le niveau visé étant celui de la survie, nous envisageons un vocabulaire et

[1] Si jamais le chiffre avancé vous paraît élevé, il est bon de savoir que le "Lexique anglais/français des sports olympiques", destiné aux journalistes ayant couvert les jeux de Sidney (2000), contenait déjà pratiquement 14 000 mots, avec 1000 entrées rien que pour les sports aquatiques (rubrique natation).

[2] Un 'phrasebook' est une collection de phrases traduites et organisées par thèmes. Ce type de recueil existe depuis fort longtemps en version papier, et plus récemment sous forme électronique (Fafiotte et al. 2009, Boitet et al. 2007). Ceux intéressés par une version commerciale consulteront http://speak.econtrader.com/

des structures en nombre et complexité limités. Avant de décrire notre proposition, voyons deux modèles très influents, caractérisant la production de langage.

2 Arrière plan théorique et motivation de notre approche

Comme nous intéressons à la production de langage, ou plus précisément à l'acquisition de cette habileté par des adultes, nous allons présenter les deux principales approches pour les comparer avec la nôtre. Malgré les nombreuses propositions on distingue deux grandes approches : celles proposés par des psychologues (Garret, 1980; Levelt, 1989, Bock, 1995, Ferrand, 2002)[3], limités généralement à la phrase, et celles venant de la part de linguistes informaticiens (Reiter & Dale, 2000), visant généralement le texte. Bien entendu, les objectifs de ces deux communautés sont assez différents. Les uns s'intéressent au traitement par le cerveau (production de phrases en temps réel)[4], et les autres s'intéressent au traitement par la machine (TAL). Si les psychologues visent des compromis (traitements imparfaits aux différents niveaux) et la souplesse, les informaticiens visent l'économie et la perfection (production sans fautes).

2.1 Le modèle de Garrett

La proposition de Garrett est à la base de pratiquement tous les modèles de production utilisés en psychologie, y compris celui de Levelt (1989). Il consiste principalement en un *conceptualiseur* (message), un *formulateur* (structure linguistique) et un *synthétiseur* de la parole (articulation). A noter qu'on ne passe pas directement du message aux sons, on passe par un module intermédiaire, le composant linguistique, qui joue un rôle de médiateur. C'est d'ailleurs surtout ce module qui a retenu l'attention de Garrett, car les traitements linguistiques laissent des traces. Garrett s'est donc appuyé sur une grande base de données d'erreurs pour construire son modèle.

La tâche du **conceptualiseur** consiste à élaborer un *message* (conceptualisation) afin de réaliser un but ou une intention de communication. Cette structure ou forme de représentation est plus ou moins élaborée, et elle est uniquement conceptuelle. C'est sur elle que s'effectueront les opérations linguistiques qui préciseront alors progressivement cette structure sous-spécifiée. Il y a des bonnes raisons de croire que cette structure est largement sous-spécifiée (Zock 1996). Des contraintes d'espace (mémoire de travail) et de temps (pression de production, manque de temps) sont des facteurs suffisamment contraignants pour dissuader le producteur d'encoder trop en détail le message, car s'il prend trop tôt des engagements forts il s'enferme, réduisant considérablement les options précieuses, utilisables ultérieurement. D'ailleurs, une des astuces rendant possible la production en temps réel est justement de partir d'une structure plus ou moins vide, coquille qu'on enrichira ensuite progressivement.

Le **formulateur** prend en charge des aspects fonctionnels, positionnels et phonologiques des éléments utilisés pour communiquer le message. Le niveau fonctionnel est responsable de l'*encodage grammatical* : les concepts seront remplacés par des mots, ou plus précisément par des lemmes, auxquels on assigne le rôle qu'ils doivent jouer au sein de la phrase. Ainsi faisant on produit une *représentation fonctionnelle* de la phrase. A l'étape suivante (*encodage phonologique*) on détermine la *représentation positionnelle*, c'est-à-dire, on récupère la forme phonologique, les caractéristiques segmentales et prosodiques des lemmes (qui, du coup deviennent des lexèmes) et on spécifie l'ordre des mots en les intégrant dans la structure spécifiée à l'étape précédente. L'**articulateur** doit transformer les symboles du module précédent en sons, afin d'évoquer chez l'auditeur des idées correspondantes à celles ayant donné naissance aux paroles du locuteur.

2.2 Le modèle de Reiter et Dale

Le modèle de Reiter et Dale (2000) se décompose également en trois étapes : planification globale, planification locale et formulation. Bien qu'existant dans certains systèmes, la synthèse de la parole, dernier élément de la chaîne, est rarement implémentée.

---

[3] Pour voir une comparaison des différentes approches on consultera Fromkin (1993), pour des propositions pour un modèle de production bilingue, voir (de Bot, 2000 ; Marini et Fabbro, 2007).

[4] Ce qui implique tout ce qu'on sait des imperfections liées à la performance : surcharge de la mémoire de travail, incapacités d'accéder à une information (latences, 'mot sur le bout de la langue'), interférences (erreurs), incohérences discursives, etc.

La **planification globale** (macroplanning) s'occupe du *choix de contenus* et de la *structuration du document*. Le premier décide des informations à communiquer explicitement dans le texte, en tenant compte des objectifs, connaissances, intérêts et croyances de l'interlocuteur. Le second traite le *groupement* (clustering) et l'*ordonnancement* des messages pour produire un ensemble cohérent, tout en évitant des déductions malheureuses (« Elle est tombée enceinte, ils se sont mariés. » vs. « Ils se sont mariés. Elle est tombée enceinte. »). L'ajout de connecteurs peut s'avérer nécessaire afin de révéler le rôle rhétorique des différents fragments (cause vs. concession) : « Il est arrivé juste à temps (car / en dépit) il y avait énormément de circulation. » Les techniques les plus utilisées pour choisir les contenus et déterminer leur organisation sont les 'schémas' (McKeown, 1980),— qui, bien que formellement différentes, sont fonctionnellement équivalentes aux nôtres,— et la 'RST' (Mann/Thomson, 1988).

La **planification locale** (microplanning) comporte la production d'*expressions référentielles* (pronoms), le *choix de mots* (lexicalisation) et l'*agrégation* (harmonisation). La génération d'expressions référentielles consiste à nommer ou à décrire l'objet visé de manière à permettre sa discrimination parmi un ensemble d'alternatives (la voiture, la voiture jaune, celle-là). La lexicalisation consiste à remplacer des concepts par des mots (CHIEN : canin, chiot), et enfin l'agrégation consiste à couper l'espace sémantique (réseau sémantique, représentant l'ensemble des messages à transmettre) pour permettre l'intégration des fragments conceptuels dans le cadre d'un paragraphe ou d'une phrase sans produire une structure trop déséquilibrée. Cette étape peut impliquer l'élimination d'éléments redondants. Les deux ressources les plus importantes à ce stade sont le dictionnaire et la grammaire, la première pour convertir les concepts en mots, et la seconde pour unifier les fragments en phrases.

La **formulation** consiste à convertir des représentations abstraites de phrases en texte concret, à la fois au niveau linguistique (réalisation linguistique) et au niveau de la mise en page (structure de réalisation): des fragments abstraits de texte (sections, paragraphes) sont signalés par des symboles de balisage.

2.3 Notre méthode, une approche hybride : les patrons ou schémas de phrases

Comme nous l'avons montré, produire du langage en temps réel est une tâche hautement complexe. Nous présentons ci-dessous une approche, montrant comment l'acquisition d'une telle performance peut néanmoins être rendue possible. Pour voir comment elle se situe par rapport aux travaux mentionnés ci-dessus, nous avons essayé de l'intégrer dans le cadre de Reiter & Dale.

Avant de présenter notre approche, nous soulignons qu'il est hautement improbable que les locuteurs ou apprenants passent par toutes les étapes décrites, appliquant une à une les règles ou contraintes décrites par des linguistes dans leurs grammaires formelles. Il y a au moins trois raisons qui nous poussent à douter de cela :

- raison liée à la *mémoire* : les gens n'ont pas stocké dans leur mémoire l'ensemble des règles décrites par les grammairiens. Essayez donc d'évoquer une des ces règles hormis celles concernant les accords. D'ailleurs, même si on avait stocké ces règles, on ne pourrait pas les utiliser séquentiellement, car leur ordre variera en fonction des informations conceptuelles qui nous viennent à l'esprit dans un ordre quelconque. La mémoire de travail (Baddeley, 1992) est déjà très sollicitée par d'autres tâches, notamment, l'encodage du message;
- raison d'*attention* : on ne peut se concentrer que sur un petit nombre de tâches (ou d'objets) à la fois. Les capacités de traitements parallèles sont sûrement bien moindres en cas d'apprentissage d'une nouvelle langue, comparées à la langue maternelle pour laquelle les mécanismes sont déjà bien rodés.
- raison de *temps* : la conception de message et sa traduction en langue, sont des processus extrêmement rapides. Tous les locuteurs le savent bien, une idée non exprimée à temps risque de retomber dans l'oubli, d'où une course effrénée entre les idées et leur expression. Les locuteurs doivent donc rapidement traduire leurs messages ; chercher les règles à appliquer et les appliquer serait beaucoup trop long.

Les linguistes décrivent généralement les langues en termes de règles, mais la plupart des gens n'apprennent pour ainsi dire jamais de telles descriptions. Il est encore moins probable qu'ils les appliquent toutes, encore moins pendant les phases initiales de l'apprentissage d'une nouvelle langue. En revanche, les gens apprennent des modèles conformes à des règles. Et s'ils utilisent des règles, c'est essentiellement

localement (morphologie), sachant que le gros du travail a été fait au préalable par le choix de schémas de phrases. Ces derniers peuvent d'ailleurs être vus comme une prise de vue instantanée d'un processus dérivationnel.

Les modèles sont des invariants, c'est-à-dire des abstractions faites à partir des formes que sont les phrases concrètes. Leurs composants (mots) peuvent être caractérisés en divers termes (syntaxiques, sémantiques, les deux). Autrement dit, il y a plusieurs manières de caractériser la même chaîne. Il n'y a pas de caractérisation absolue. Tout dépend du point de vue et du niveau d'abstraction. Les modèles ou schémas de phrase sont des structures, susceptibles d'être construites dynamiquement via des règles. Cependant, on peut également les concevoir comme des unités, structures holistiques, rencontrées, stockées ou récupérées telles quelles. Un locuteur performant possèderait donc une grande librairie de modèles et un bon index lui permettant de localiser rapidement le schéma de phrase nécessaire. Au fond, c'est un peu comme s'il cherchait un mot dans un dictionnaire ou un thésaurus. Autrement dit, les dictionnaires de mots et les mémoires de (schémas de) phrases se ressemblent. Ce sont des bases de données indexées, espaces balisés, dans lesquels on navigue pour récupérer l'élément nécessaire.

Récupérer d'un seul coup toute une phrase (ou presque) nous épargne des efforts de calcul tout en nous faisant gagner du temps. C'est d'ailleurs probablement la raison pour laquelle tant de gens s'en servent en langue (apprenants, interprètes, journalistes, etc.) ou dans d'autres domaines (musique, programmation, jeu d'échecs, etc.). Bien sûr, il y a un prix à payer : les modèles doivent non seulement être accessibles (voir ci-dessous), ils doivent également être adéquats et compatibles avec l'idée à exprimer. Si les modèles ont des qualités, ils ont aussi des faiblesses : ils sont rigides, et ils occupent de la place mémoire. Il faut donc procéder parcimonieusement. Toute variation linguistique ne justifie pas forcément qu'on en fasse une abstraction. Imaginez les variations morphologiques (temps). Elles ne méritent guère qu'on en tienne compte dans des schémas de phrase. Prenez, par exemple, les deux phrases suivantes et leurs schémas respectifs : (A) Je vais à New York cet été [je vais < LIEU > < TEMPS >]. --- (B) J'ai été à Madrid la semaine dernière [j'ai été < LIEU > < TEMPS >].

Vue la similitude des deux phrases il n'est guère justifiable d'abstraire deux schémas différents. Il serait beaucoup plus raisonnable d'avoir un modèle rendant compte de la *structure globale* [PERSONNE aller < LIEU > < TEMPS >] et un ensemble de paramètres (règles) prenant en charge des ajustements locaux : accords (singulier/pluriel), flexion de verbes (passé, présent), etc.

Tout comme les schémas, les règles ont certains inconvénients. Certes, elles ont la puissance nécessaire pour rendre compte de l'expressivité de la langue (ensemble de variations légalement possibles), mais, vu leur granularité, leur nombre devient prohibitif, empêchant le locuteur de faire son travail à temps. C'est pourquoi nous suggérons une approche hybride, approche à deux vitesses : des *schémas* pour les structures globales et des *règles* pour les ajustements morphologiques (niveau local). Cette combinaison offre le meilleur compromis. D'une part elle permet de minimiser les besoins de calcul (allégeant du coup la mémoire et l'attention), d'autre part elle maximise la puissance (rapidité pour faire le gros du travail) et la flexibilité. Cette dernière est requise pour effectuer des ajustements (accords) ou des restructurations locales.

Lorsqu'on apprend une nouvelle langue, on apprend généralement une liste de mots (vocabulaire) et un mécanisme de construction de phrase (grammaire). Spécifiant les combinaisons légales, la grammaire fournit les structures possibles, dans lesquelles le locuteur va insérer les mots choisis pour exprimer (une partie de) ses idées (concepts). Ceci dit, la structure syntaxique peut être obtenue de différentes manières : par le biais d'une construction incrémentale (unification progressive des éléments) ou par une recherche dans la mémoire, auquel cas on la récupère en un seul bloc (schémas).

Bien qu'il s'agisse d'un continuum, on peut imaginer trois grandes approches : (a) l'unité du traitement est le concept ou sa traduction, le *mot* ; (b) l'unité est un *segment de phrase* (phrases lexicales) [5] ou (c) l'unité est toute la *phrase*. La première solution est la plus risquée et la plus coûteuse, car produire des phrases à partir d'unités aussi petites implique une vision très réduite et de nombreux calculs (opérations d'unification). La dernière approche est la plus rapide, souvent la plus sûre, mais, à terme, aussi la plus limitée. Réutilisant une forme telle quelle (imitation d'une phrase), et n'ayant fait aucune abstraction, on

[5] De tels segments sont assez évidents dans des formules ('Veuillez agréer, cher Monsieur, l'expression de mes sentiments distingués'; 'je vous en prie') ou dans des expressions comme : 'dans la mesure où', ou 'Qui aurait pu croire que <phrase> ?', etc. (Nattinger et Decarrico, 1992, Becker, 1975).

reste prisonnier des formes rencontrées. On ne peut exprimer rien d'autre. La stratégie (b) est à terme la plus utile. Bien qu'elle ne soit pas parfaite, la spécificité se payant au prix de la généralité, elle offre néanmoins un excellent compromis entre la vitesse, la puissance et la souplesse. En effet, elle permet d'exprimer rapidement des idées très variées, sans obliger le locuteur à effectuer de nombreux calculs. Ayant récupéré des grands blocs ('chunks' conceptuels lexicalisés) il les insère dans un schéma plus large. Cette stratégie est très utilisée par des interprètes de conférence, car ils travaillent constamment à la limite de la surcharge cognitive. Aussi, au lieu d'attendre la fin de la phrase, ils ont tendance à commencer la traduction le plus tôt possible, opérant sur des fragments plutôt que sur l'ensemble des éléments de la phrase. Cela permet de minimiser la charge mémorielle tout en augmentant le temps disponible pour la partie à venir, partie encore à traduire. Un interprète essaie donc à tout prix de se 'débarrasser' le plus vite possible d'une partie du message, pour avoir le maximum de temps pour la partie restante. D'ailleurs, les interprètes craignent généralement moins les locuteurs à grand débit que ceux au débit lent, ou ceux utilisant des structures emboitées, car dans les deux cas on a du mal à faire de bonnes prédictions concernant le rôle joué par certains éléments à traduire.

Ceci dit, si la stratégie basée sur les *phrases lexicales* est séduisante, elle est trop ambitieuse, parce que trop difficile, pour un débutant. Ce dernier doit avoir rapidement du succès pour devenir confiant. C'est pour cette raison que nous proposons de travailler avec des phrases plus ou moins toutes faites. Certes, il ne s'agit pas d'apprendre par cœur ces phrases, mais plutôt d'abstraire le schéma sous-jacent, pour pouvoir produire des phrases analogues ou similaires.[7] Autrement dit, nous visons la productivité de la langue. Mais pour y arriver nous essayons de réduire la charge cognitive, tout en visant l'augmentation du contrôle : on ne se concentre que sur un élément à la fois.

Pour voir comment notre modèle se situe par rapport aux autres modèles de production, nous présentons les différentes étapes en termes de ces architectures. Ayant indexé les structures à apprendre en termes de but, l'utilisateur part de celui-ci pour communiquer son intention de communication (1°). Le système répond avec un, voire plusieurs schémas de phrase, parmi lesquels l'utilisateur choisit (2°). Cette étape correspond au *niveau macro* du modèle de Reiter et Dale. En effectuant des choix lexicaux (3°) on est désormais au *niveau micro*, nommé *formulateur* dans le modèle de Garrett. On notera que le choix lexical peut être fait par le système, et c'est comme ça que les choses se passent dans les fameux 'pattern drills' (Besse, 1975, Le Rouzo, 1975).

NIVEAUX	ENTRÉES CONCEPTUELLES	SORTIES LINGUISTIQUES
NIVEAU GLOBAL (macro)	1° *but* de communication : comparaison	Ensemble de schémas de phrase (a) < OBJET$_1$> être plus < ATTRIBUT > que < OBJET$_2$> (b) < OBJET$_2$> être moins < ATTRIBUT > que < OBJET$_1$ >
	2° *cadre syntaxique* : (a)	< OBJET$_1$> être plus < ATTRIBUT > que < OBJET$_2$>
NIVEAU LOCAL (micro)	3° valeurs *lexicales* : OBJET$_1$ = eau ; OBJET$_2$ = vin ATTRIBUT = cher	Structure lexicalement spécifiée : eau être plus cher que vin.
	4° valeurs *morphologiques* : OBJET$_1$ = singulier OBJET$_2$ = singulier	Structure conceptuelle et linguistique complète L'eau est plus chère que le vin

Tableau 1 : L'entrée conceptuelle, un processus en quatre étapes

Ceci dit, si l'on souhaite un système ouvert, donc sensible aux besoins de l'utilisateur, on laisse le choix à ce dernier. La structure choisie en (2°) sera instanciée par la valeur lexicale (3°), puis (4°) complétée par des valeurs morphologiques (nombre, temps, etc.). Désormais le système a tout ce dont il a besoin pour produire la phrase. On notera que l'entrée conceptuelle est répartie sur trois niveaux : au niveau le plus

[7] L'analogie est un principe d'apprentissage bien connu, utilisé par des enfants pour apprendre les régularités d'une langue (Berko, 1958). Elle a été proposé comme base dans des 'exercices structuraux' ('pattern drills' ou 'patrons de phrases'), et même en traduction automatique (Nagao, 1984), bien qu'il s'agissait là en fait plutôt de similarités.

profond elle consiste à choisir un schéma global (schéma de phrase) via un but. Aux niveaux suivants, on spécifie respectivement les valeurs lexicales et morphologiques (nombre, temps). Ainsi faisant on affine peu à peu un message sous-spécifié. Ce type de décomposition a plusieurs avantages. Les informations ne sont demandées que lorsqu'elles sont pertinentes ou nécessaires, ce qui réduit la complexité du traitement, tout en augmentant son contrôle. Deuxièmement, cette méthode est bien plus rapide pour transmettre un message que de naviguer dans une immense ontologie lexicale ou conceptuelle, tel que cela a été suggéré ailleurs (Power et al, 1998 ; Zock, 1991; Zock et al 2009).

3 Construction de la ressource

Étant donné le processus décrit dans le tableau 1, nous avons séparé la description des *buts* de celle des méthodes permettant de les atteindre, les *patrons* ou structures de phrases. Certains éléments des phrases ont été généralisés via des variables syntaxiques ou lexicales. Par exemple, l'objectif 'se présenter' peut être atteint via une des deux structures suivantes : «On m'appelle X" ou "Mon nom est X" auquel cas la variable X peut être instanciée par le nom, prénom, ou petit nom de l'orateur. Les valeurs autorisées peuvent être définies dans un dictionnaire. Pour un même objectif, on peut imaginer un grand nombre de modèles et de types de variables, et ceci pour différentes langues.

Étant donné la nature hiérarchique des objectifs et des patrons de phrase, nous avons décidé d'utiliser une structure arborescente codée en XML pour conserver l'information. Avec un éditeur approprié, un linguiste peut facilement ajouter de nouveaux objectifs, des modèles de phrases associés, des variables et des entrées lexicales sans avoir à s'occuper de la complexité du programme informatique qui affiche le résultat à l'utilisateur. Le programme de traitement lit cette structure et l'interprète à la volée pour générer des patrons et des phrases qui sont ensuite présentées à l'utilisateur et comparées avec ce que celui-ci dit ou écrit.

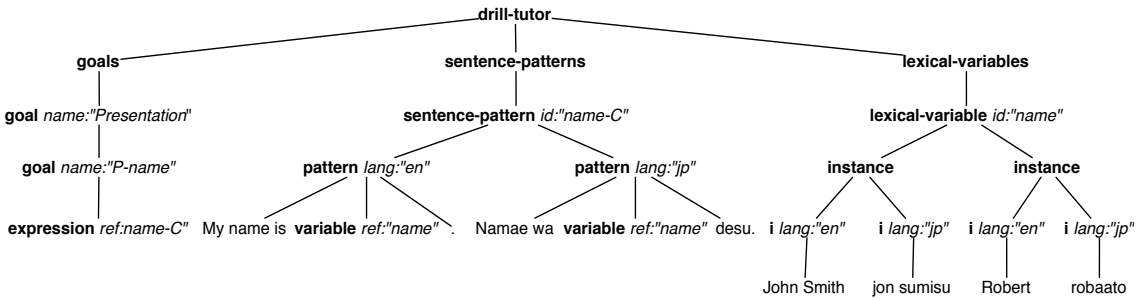


Figure 1: Définitions de *buts*, *schémas de phrases* et *entrées lexicales* sous forme d'arbre.

4 Aspects techniques et l'état actuel du prototype

Pour réaliser notre 'entraîneur' de langue (DrillTutor), nous avons décidé d'utiliser un navigateur Web à cause de sa capacité d'affichage de plusieurs jeux de caractères et de sa disponibilité sur toutes les plateformes. Pour créer les documents nous utilisons PHP, car il fournit un moyen assez commode pour créer dynamiquement des pages web et pour accéder aux fichiers XML. Pour afficher la prononciation des mots japonais, nous avons utilisé deux scripts : le romaji (une forme de représentation phonétique proche de l'anglais) et l'hiragana, le syllabaire principal du Japonais. La conversion est faite automatiquement via une fonction PHP qui ensuite affiche le résultat dans le browser.

Une fois l'objectif sélectionné, le programme engendre une phrase en remplaçant dans le modèle les valeurs des variables lexicales. Une variable peut être utilisée plusieurs fois à condition d'avoir été définie. Sa valeur sera alors propagée à travers les différents modèles de phrase. Comme une même valeur lexicale peut correspondre à différentes variables, ce lexème recevra des valeurs distinctes. Cette fonction est utilisée dans les modèles comme : « Êtes-vous Chinois? Non, je ne suis pas Chinois, mais Japonais » où Chinois et Japonais sont considérés comme des variables dans le modèle. Les deux premières occurrences auraient le même nom, alors que la dernière serait nommée différemment.

Côté serveur, on garde dans un fichier des statistiques concernant la performance des utilisateurs. Celle-ci

peut être utilisée pour déterminer la sélection des variables d'un modèle et la manière d'afficher des buts. Ceci dit, à l'heure actuelle, on choisit au hasard les variables des structures de phrases et on affiche tous les buts.

Nous nous servons de Javascript pour obtenir une interaction locale à l'intérieur des pages. Par exemple, comme on voit sur la figure 2b, la page web contient initialement tous les buts et modèles de phrase. Cependant l'affichage est contrôlé localement via Javascript qui lui gère les clics de la souris et le contenu du champ texte pour afficher un menu de mots-clés suggérés pour trouver des buts. L'interface du système (à savoir, les titres et incitations du système) se trouvent dans un autre fichier XML qui stocke les chaînes de message dans différentes langues (actuellement Français et Anglais).

La séparation de la description des objectifs et des patterns permettra à terme l'ajout par l'utilisateur de nouveaux buts et schémas de phrase. Cette fonctionnalité n'ayant pas encore été mise en œuvre, les objectifs et les schémas doivent être saisis à l'aide d'un éditeur de texte XML. Un schéma XML permet de valider que l'information sur les buts, les patrons et les variables est codée dans le bon format et que toutes les références à des variables et les schémas sont bel et bien des éléments présents dans le fichier XML. Initialement, nous avions défini les schémas en trois langues (anglais, français et japonais). Depuis lors des buts, des schémas et des valeurs lexicales ont été ajoutés en chinois, et ceci en très peu de temps. Le système a une quinzaine de buts et de schémas et 150 éléments lexicaux, en quatre langues dont deux langues d'interface. Bien que ces chiffres puissent sembler très petits, il faut garder à l'esprit les points suivants. Un des objectifs de notre travail était de vérifier les difficultés que présenterait l'ajout d'une nouvelle langue. Il s'est avéré que même l'ajout d'une langue typologiquement aussi différente que le chinois, ne présentait pas un obstacle majeur. En fait, le nombre de schémas et la taille du vocabulaire ne sont pas vraiment des facteurs qui comptent actuellement. L'accent a été mis sur la réalisation d'un éditeur conçu pour créer, modifier et utiliser une base de données. Cette dernière peut être facilement étendue. L'architecture que nous avons définie permet de gérer autant de buts, de schémas et d'éléments lexicaux que l'on veut. Il suffit pour cela d'éditer des fichiers XML en suivant un schéma bien défini (structure), et en veillant à la bonne formation, ce qui peut être fait via un éditeur de schéma XML.

Nous prévoyons de développer les composants suivants dans notre entraîneur de langue, les parties 2 et 4-8 n'étant pas encore implantées:

1. Librairie de schémas et scénarios indexés en termes de buts ;
2. Dictionnaires multilingues ;
3. Translitérateurs romaji et hiragana ;
4. Composants morphologiques : générateur de formes et d'accord ;
5. Synthèse de la parole ;
6. Vérification de cohérence et de bonne formation ;
7. Induction de schémas via une ressource comme WordNet ;
8. Extraction d'exemples via un corpus ;
9. Paramètres concernant les choix et les performances de l'utilisateur.

5 Exemple d'utilisation

Le système est disponible à l'adresse : http://agil.lif.univ-mrs.fr/DrillTutor. L'utilisateur choisit le but de la communication. Les objectifs étant indexés et ordonnés de manière hiérarchique, l'utilisateur peut les atteindre soit en écrivant leur nom dans une case réservée à cet effet, soit en naviguant dans l'arborescence (figure 2a). Pour un objectif donné il suffit de cliquer sur son nom pour voir se développer soit les sous-buts, soit la structure associée, les variables étant indiquées en gras. On est alors arrivé au niveau des feuilles. Dans la figure 2a, l'utilisateur a demandé le schéma permettant d'atteindre le but "origine". Après avoir choisi dans quelle langue il aimerait faire l'exercice, en l'occurrence le Japonais, il passe par les deux étapes suivantes. D'abord il spécifie la valeur de la variable 'origine' (2b) puis il vérifie si son résultat correspond à celui de la machine (2c).

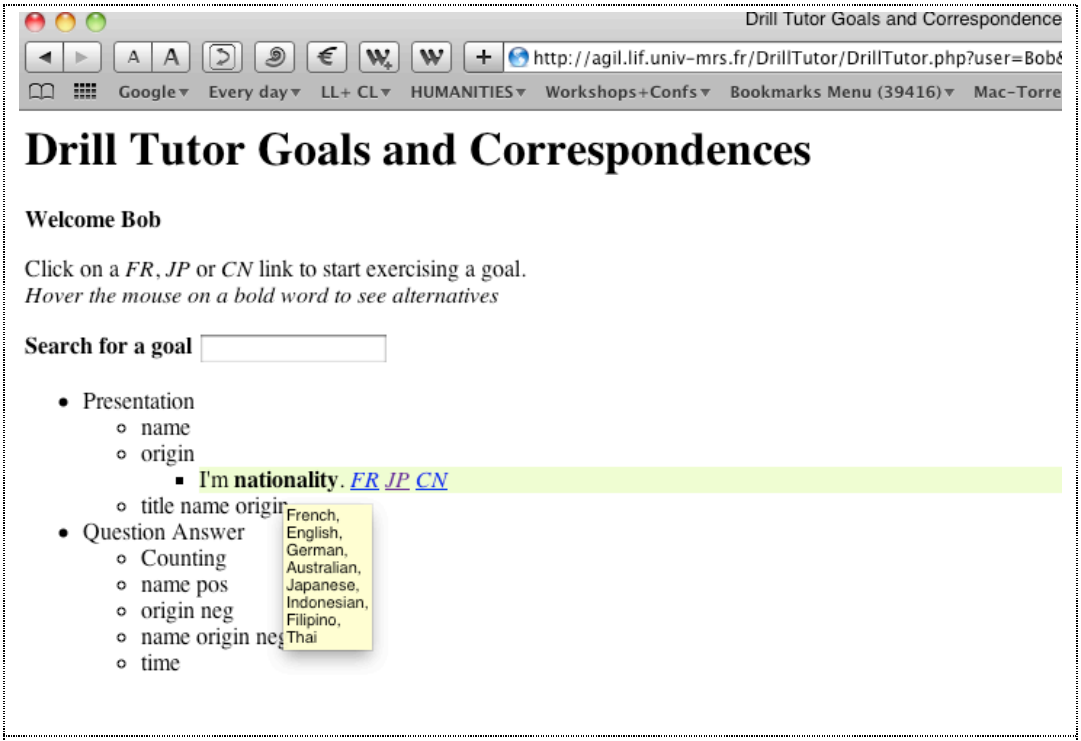


Figure 2a : **Choix du *but*** (ici, 'origine'), d'une structure lorsqu'il y en a plusieurs et de la *langue* à *apprendre* (**FR** : français, **JP** : japonais; **CN** : chinois), ici, japonais

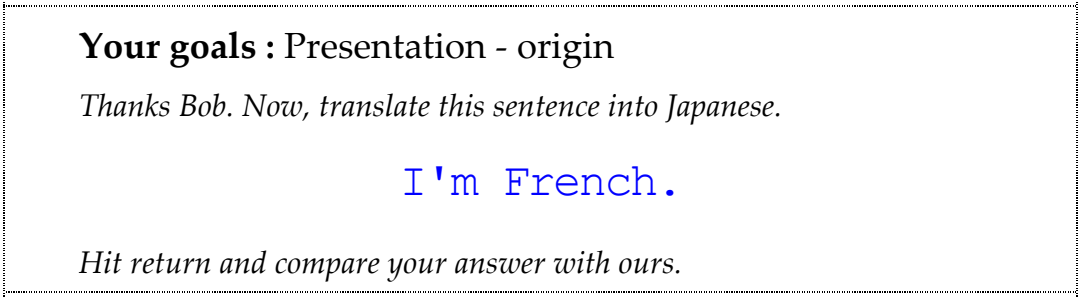


Figure 2b : **Entrée**, valeur de la variable *nationalité* = française

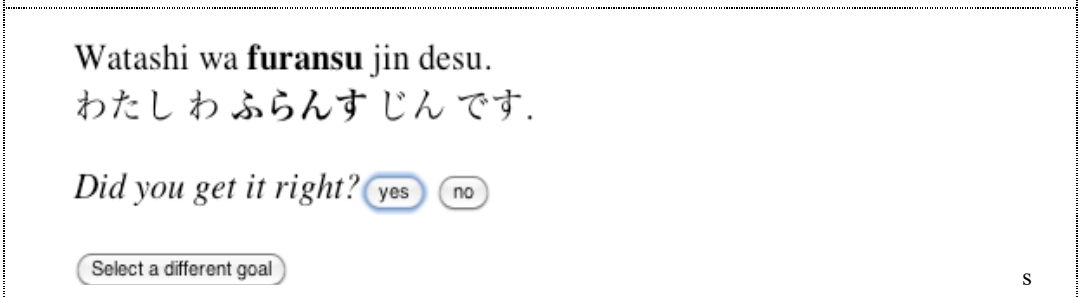


Figure 2c : **Résultat** en lettres romanes et en script japonais [8]

[8] A noter, que la *particule* 'wa', tout juste après 'watashi' ('je') indique le rôle (thème) du pronom personnel. Elle devrait être écrite 'は', équivalent de la syllabe 'ho'. Mais comme on prononce les deux syllabes de la même manière ('**wa**tashi **wa**'), notre transliterateur produit dans les deux cas le même caractère ('わ'). Ceci devrait être corrigé, mais cela demande un programme capable de reconnaître les différentes fonctions d'un mot ou d'une syllabe. A noter également, que les noms propres (noms de personnes et de villes) devraient être écrits en kanji ou en katakana. Comme il s'agit d'acquérir un niveau de base à l'oral, nous avons ignoré ces deux subtilités, car elles rendraient notre tâche et celle des apprenants inutilement difficiles.

6 Discussion

S'agissant d'un travail en cours, notre proposition est forcément incomplète. Nous n'avons pas fait référence à une théorie d'apprentissage particulière, car nous ne cherchons pas à enseigner tous les aspects de la langue, ou ceux de son traitement. Nous nous sommes limités à un seul aspect, la production de phrases simples. Nous sommes convaincus que l'exercice est indispensable et qu'il doit avoir certaines caractéristiques — limitation des éléments à faire varier à l'intérieur des structures à apprendre — pour permettre à l'apprenant d'atteindre le niveau suivant, à savoir, la construction spontanée de phrases complexes (voir également Levelt, 1970; Krashen, 1981). Ceci veut dire, mener en parallèle et de manière incrémentale l'encodage du message et son expression (forme linguistique). Malgré le grand nombre de travaux en psycholinguistique, en linguistique informatique et en didactique —(Chapelle, 2001; 2003, Crystal; 2001, Warschauer, 1998, Gethin et Gunnemark, 1995; Holland, Kaplan et Sam, 1995, Swartz et Yazdani, 1991, Brown, 1987, Novak et Gowin, 1984, Wilkins, 1972;)— il n'y a pratiquement rien qui permette leur transposition pour apprendre à passer d'un *but* et d'un *message* vers son *expression*. Ceci vaut à la fois pour les prototypes issus des laboratoires de recherche et pour les produits industriels (Pimsleur, TellMeMore, Rosetta Stone, etc.).

Il est intéressant de noter que l'utilisateur 'apprend' pratiquement les mêmes connaissances qui nous ont guidés à construire notre générateur. Celui-ci est simple, voire rudimentaire, mais assez facile à mettre en oeuvre et à s'approprier ou apprendre. En suivant la démarche proposée par notre programme on apprend donc non seulement à engendrer des phrases, mais aussi comment procéder : la nature et l'ordre des informations à fournir et à traiter sont suggérés par le programme. Bien entendu, on peut imaginer d'autres situations et d'autres stratégies. Nous considérons justement la nôtre comme une étape préliminaire, préparant l'apprenant à celle d'une génération plus souple, mais demandant plus de connaissances. C'est elle qu'on utilise normalement en discours spontané (imaginez que vous deviez décrire un film que vous venez de voir, tâche très exigeante, notamment au niveau conceptuel). D'ailleurs, rien n'interdit de commencer par le choix des mots avant de trouver leur place dans un cadre syntaxique. Mais ceci suppose quelque chose ressemblant à une grammaire d'unification en arrière-plan, type de connaissance qu'on n'acquiert que plus tard et progressivement, lorsqu'on a rencontré un très grand nombre de phrases.

Nous sommes limités à un accès simple aux phrases à partir des buts. Or toutes les phrases ne peuvent pas être indexées, et bon nombre d'entre elles peuvent être indexées selon différents points de vue. Ceci dit, étant donné notre objectif (niveau de survie), on peut raisonnablement supposer que les phrases puissent jouer le rôle que nous leur avons assigné.

Plusieurs problèmes d'organisation des schémas se posent, à savoir, quels éléments de la phrase généraliser et en quels termes. Quant au nom de la variable utilisée pour remplacer un terme (mot, expression), c'est à la fois un problème de métalangage (est-ce que 'mouton', issu de 'dessine-moi un mouton', doit être remplacé par 'animal', 'groupe nominal' ou par quel autre terme?) et un problème d'inclusion. Plus le terme choisi est large, plus le nombre de valeurs possibles est grand ('voici un <objet>' vs 'voici un <fruit>'). Nous avons induit les schémas manuellement, mais on pourrait imaginer d'automatiser l'induction de schémas, en remplaçant les mots par leur hyperonyme via une ressource comme WordNet.

Un des aspects le plus prometteurs, mais aussi des plus difficiles, est la perspective de créer ou d'étendre (semi-)automatiquement la base de phrases et celle de schémas. On allégerait ainsi le travail du créateur, on augmenterait l'étendue ou l'expérience linguistique de l'apprenant tant par la variété pour combattre la monotonie que l'authenticité des exemples. Le défi étant de trouver des exemples qui répondent aux besoins pédagogiques et cognitifs de l'apprenant.

Notre prototype ne comporte pour l'instant ni dictionnaire (mono- ou multilingue), ni générateur de morphologie. Ayant été motivés par la construction rapide d'un prototype, nous avons négligé cet aspect comme celui de l'évaluation qui reste nécessaire, mais reste tributaire d'une implantation plus complète que ne l'est notre prototype actuel.

7 Conclusion

Nous avons présenté une base de phrases multilingue convertie en générateur d'exercices pour aider les utilisateurs à acquérir une certaine maîtrise orale en langue étrangère. L'apprentissage des mots et de leurs combinaisons est indispensable pour s'exprimer à un débit 'normal' dans une nouvelle langue. L'acquisition de ces automatismes nous paraît capitale pour accéder à l'étape suivante : savoir produire spontanément des phrases plus élaborées. Vu la vitesse avec laquelle il faut effectuer un très grand nombre d'opérations, il est souhaitable d'automatiser celles qui sont les plus stables, à savoir, les opérations linguistiques (syntaxe, morphologie), car le sens peut varier à l'infini.

Notre approche peut être caractérisée par son ouverture, sa généricité et son extensibilité : de nouvelles informations peuvent être ajoutées à tout moment, et d'autres langues peuvent être mises en œuvre très rapidement. En effet, nous avons testé l'extensibilité de l'outil, constatant qu'il était assez facile d'ajouter de nouvelles informations, ou même une langue typologiquement très différente. Considérant que la production de phrases seulement via des règles était trop lourde, et que les modèles seuls étaient trop rigides et trop gourmands en mémoire (stockage, accès), nous avons opté pour un compromis (processus à deux vitesses) : utiliser des patrons au niveau global et des règles pour les ajustements locaux. Ceci augmente considérablement la vitesse de production de phrases, tout en minimisant le besoin de stockage (patrons, règles).

Bien que la couverture actuelle soit encore très faible et bien que le système n'a pas été évalué, nous pensons que l'approche est viable. Bien entendu, le juge ultime sera l'utilisateur. Va-t-il utiliser un tel système ? Ce dernier lui permettra-t-il d'atteindre les objectifs fixés ? Voici des questions auxquelles nous devrons répondre.

Références

BADDELEY, A. (1992). Working memory. *Science*, 255, 556–559

BECKER, J. (1975). The Phrasal Lexicon. In *Proceedings Interdisciplinary Workshop on Theoretical Issues in Natural Language Processing*. Cambridge, Massachusets June 70-73.

BERKO, J. (1958). The Child's Learning of English Morphology. Word, 14, 150 177

BESSE, H. (1975) De la pratique aux théories des exercices structuraux. *Etudes de Linguistique Appliquée*, 20, 8-30. Paris, Didier

BOCK, J.K. (1995). *Sentence production: From mind to mouth*. In J. L. Miller & P.D. Eimas (Ed.), Handbook of perception and cognition. Vol. 11: Speech, language and communication. Orlando, FL: Academic Press.

BOITET, C., BHATTACHARYYA, P., BLANC, E., MEENA, BOUDDH, S., FAFIOTTE, G. & FALAISE, A. & V. VACCHANI (2007). Building Hindi-French-English-UNL resources for SurviTra-CIFLI, a linguistic survival system under construction, actes de *SNLP*, Pattaya, Thaïlande

BROWN, H. 1987 Principles of Language Learning and Language Teaching. Englewood Cliffs, NJ: Prentice-Hall.

CHAPELLE, C. (2001). Computer applications in second langage acquisition: Foundations for teaching, testing and research. CUP.

CRYSTAL, D. (2001). Language and the Internet. CUP.

de BOT, K. (2000). A bilingual production model: Levelt´s 'speaking' model adapted. In L. Wei (Ed.), The bilingualism reader (pp. 420-442). London; New York: Routledge. (Reprinted from Applied Linguistics, 13, 1992, 1-24.)

FAFIOTTE, G. FALAISE, A. & J. GOULIAN. (2009). CIFLI-SurviTra, deux facettes : démonstrateur de composants de TA fondée sur UNL, et phrasebook multilingue, actes de *TALN*, Senlis

FERRAND, L. (2002). Les modèles de la production de la parole. In M. Fayol (Ed.), *Production du langage*. Traité des Sciences Cognitives (pp. 27-44). Paris: Hermès.

FROMKIN, V. (1993). *Speech Production*. In Psycholinguistics. J. Berko Gleason & N. Bernstein Ratner, *Eds*. Fort Worth, TX: Harcourt, Brace, Jovanovich.

GARRETT. M. F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), Language production (pp. 177-220). London: Academic Press

GETHIN, A. & E. GUNNEMARK. 1995. The Art and Science of Learning Languages. Intellect books, Headington, Oxford.

HOLLAND, M., J. KAPLAN & M. SAMS (Eds.).1995.Intelligent Language Tutors. Hillsdale, NJ.: LawrenceErlbaum Associates.

KRASHEN, S. (1981). http://www.sdkrashen.com/SL_Acquisition_and_Learning/index.html, *Second Language Acquisition and Second Language Learning*.

LE ROUZO, M. L. (1975). Y a-t-il une justification psychologique à la pratique des exercices structuraux ? *Etudes de Linguistique Appliquée*, 20, 37-51. Paris, Didier

LEVELT, W. (1970). Skill theory and language teaching. *Studies in Second Language Acquisition, 1*(1), 53–70.

LEVELT, W. (1989). *Speaking*. MIT Press, Cambridge, Mass.

Lexique anglais/français des sports olympiques: jeux d'été, *Insep Publications*, Paris (2000).

MANN, W. C. & THOMPSON, S. A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization', *Text* 8(3), 243--281.

MARINI, A. and FABBRO, F. (2007) "Psycholinguistic models of speech production in Bilingualism and Multilingualism". In Ardila, A. and Ramos, E. (eds.) "Speech and language disorders in Bilinguals". Nova Science Publishers Inc. New York, NY, pp. 47-67s

MCKEOWN, K. (1985).Discourse Strategies for Generating Natural-Language Text. *Artificial Intelligence*, 27, 1-41

NAGAO, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. In A. Elithorn & R. Banerji (Eds.), *Artificial and Human Intelligence* (pp. 173–180). Amsterdam: Elsevier.

NATTINGER, J. et S. DECARRICO, J. (1992) Lexical Phrases and Language Teaching , Oxford, Oxford University Press

NOVAK, J. & D. GOWIN. 1984. Learning how to Learn.Cambridge. Cambridge University Press.

POWER, R., SCOTT, D., & EVANS, R. (1998). What you see is what you meant: direct knowledge editings with natural language feedback. In H. Prade (Ed.), *13th European conference on artificial intelligence* (ECAI'98) (pp. 677–681). Chichester: Wiley.

REITER, E., & DALE, R. (2000). *Building natural language generation systems*. Cambridge: Cambridge University Press.

SWARTZ, M. et M. YAZDANI (Eds.). (1991). Intelligent Tutoring Systems for Foreign Language Learning: The Bridge to International Communication. Springer Verlag, Berlin.

WARSCHAUER, M. & HEALEY, D. (1998). Computers and Language Learning: An overview. *Language Teaching*, 31. 57–71. http://www.lll.hawaii.edu/web/faculty/markw/overview.htm

WILKINS D. (1972): *Linguistics and Language Teaching*. London: Edward Arnold.

ZOCK, M. (1991). Swim or sink: the problem of communicating thought. In M. Swartz & M. Yazdani (Eds.), *Intelligent tutoring systems for foreign language learning* (pp. 235–247). New York: Springer.

ZOCK, M. (1996). The Power of Words in Message Planning, *16th International Conference on Computational Linguistics*, Copenhagen, pp. 990-995

ZOCK, M., SABATIER, P. and L. JAKUBIEC. Message Composition Based on Concepts and Goals. *International Journal of Speech Technology*, 11(3-4):181–193, 2008.

Approche de construction automatique de titres courts par des méthodes de Fouille du Web

Cédric Lopez[1] Mathieu Roche[1]
(1) LIRMM, 161, rue ADA 34392 Montpellier Cedex 5
{lopez,mroche}@lirmm.fr

Résumé. Le titrage automatique de documents textuels est une tâche essentielle pour plusieurs applications (titrage de mails, génération automatique de sommaires, synthèse de documents, etc.). Cette étude présente une méthode de construction de titres courts appliquée à un corpus d'articles journalistiques via des méthodes de Fouille du Web. Il s'agit d'une première étape cruciale dans le but de proposer une méthode de construction de titres plus complexes. Dans cet article, nous présentons une méthode proposant des titres tenant compte de leur cohérence par rapport au texte, par rapport au Web, ainsi que de leur contexte dynamique. L'évaluation de notre approche indique que nos titres construits automatiquement sont informatifs et/ou accrocheurs.

Abstract. The automatic titling of text documents is an essential task for several applications (automatic titling of e-mails, summarization, and so forth). This study presents a method of generation of short titles applied to a corpus of journalistic articles using methods of Web Mining. It is a first crucial stage with the aim of proposing a method of generation of more complex titles. In this article, we present a method that proposes titles taking into account their coherence in connection with the text and the Web, as well as their dynamic context. The evaluation of our approach indicates that our titles generated automatically are informative and/or catchy.

Mots-clés : Traitement Automatique du Langage Naturel, Fouille du Web, Titrage automatique.

Keywords: Natural Language Processing, Web Mining, Automatic Titling.

1 Introduction

Le titre est un élément important du document textuel. Dans la littérature, deux définitions complémentaires apparaissent. D'une part, le titre peut être défini en tant qu'objet textuel nettement mis en valeur par rapport au contenu qui le suit, faisant varier des paramètres tels que sa taille, sa police de caractère, ou encore sa couleur. D'autre part, le titre peut être défini en tant qu'objet sémantique ayant trois fonctions (Ho-Dac *et al.*, 2004) : intéresser/captiver le lecteur, informer le lecteur, introduire le sujet de l'article. D'un point de vue syntaxique, un titre est une métadonnée dont la structure peut être un mot, un groupe de mots, une expression, une phrase, servant à désigner un écrit ou une de ses parties, à en donner le sujet.

Le sous-titre est une spécialisation du titre, en ce sens qu'il possède les mêmes fonctions que le titre. Néanmoins, il est attribué à un segment du texte auquel il doit s'adapter, notamment en fonction de sa taille (nombre de mots le composant). Le titre et les sous-titres peuvent être sémantiquement indépendants, en particulier s'il y a utilisation d'expression ou de tournure humoristique dans leur constitution.

L'objectif du titrage automatique est de proposer des titres respectant les contraintes mentionnées ci-dessus. Les méthodes de TALN [1] seront exploitées dans le but de respecter les contraintes qu'un titre doit être un groupe de mots bien formé et qu'il désigne le sujet traité. Le titrage de page Web est un des domaines clés de l'accessibilité des pages web. Côté lecteur, l'objectif est d'augmenter la lisibilité des pages tout venant obtenues à partir d'une recherche sur mot-clé et dont la pertinence est souvent faible, décourageant les lecteurs devant fournir de grands efforts cognitifs. Côté producteur de site Web, l'objectif est d'améliorer l'indexation des pages pour une recherche plus pertinente.

De nombreuses applications liées au titrage automatique sont envisageables. Une des applications immédiates du titrage automatique est de proposer un titre pour les documents textuels qui n'en possèdent pas (par exemple, les mails " no objects "), l'intérêt étant de faire gagner du temps à l'utilisateur. Une autre application est le titrage automatique de texte tout venant, au préalable structuré par une tâche de segmentation thématique (par exemple (Prince & Labadié, 2007)). La segmentation de texte et le titrage étant des tâches automatiques, le sommaire du document serait donc généré automatiquement. Appliqué aux contenus textuels de sessions de conversation de chat, le titrage automatique permettrait à l'utilisateur de retrouver une information pertinente noyée dans cette masse textuelle. Dernier exemple, les journaux en ligne se développent et publient de nombreux articles chaque jour. Par exemple, Le Monde publie en moyenne un article chaque 15 minutes. Un outil de titrage automatique permettrait un réel gain de temps aux journalistes en proposant des titres informatifs et accrocheurs auxquels ils n'auraient peut-être pas pensé. Enfin, une application de titrage de pages Web permettrait de respecter un des critère de la norme W3C.

Dans cet article, nous proposons une approche de construction automatique de titres courts (TC) français par des méthodes de Fouille du Web. À partir de patrons syntaxiques issus de nos analyses statistiques portées sur les titres réels (section 3.1), nous formons des TC candidats (section 3.2). Le principal problème rencontré est que plusieurs TC peuvent être pertinents pour un même texte (ou section de texte). Ils peuvent varier en fonction de leur taille (en nombre de mots), de leur forme ou bien du sujet mis en avant. Les TC candidats seront donc soumis à une validation en deux phases : (1) cohérence des candidats par rapport au texte (section 3.3.1), (2) cohérence des candidats par rapport au web (section 3.3.2). Les candidats font ensuite l'objet d'une contextualisation dynamique (section 3.4), indiquant ainsi le titre candidat le plus pertinent pour la partie de texte traitée. L'évaluation (section 4) indique que les TC déterminés par notre approche sont pertinents.

2 Travaux antérieurs

Les titres ont fait l'objet de nombreuses études linguistiques et sont vus de différentes manières (Peñalver Vicea, 2003). Ces différences d'appréciation induisent que plusieurs titres peuvent être pertinents pour un même texte. Le titrage a pour objectif de représenter pertinemment le contenu des documents en quelques mots. Il peut utiliser des métaphores, l'humour, des jeux de mots [2] ou encore des reformulations.

Le titre doit être différencié du résumé, qui est une forme condensée (abrégée, sommaire) d'un texte. Alors que

1. Traitement Automatique du Langage Naturel
2. Exemple : « À Montpellier, Ségolène fait un retour royal », Midi Libre n°23332

le résumé doit donner un aperçu du contenu du texte, le titre doit désigner le sujet traité dans le texte sans pour autant dévoiler le contenu. Le processus de résumé peut faire appel au titre, par exemple dans (Minel *et al.*, 2001; Pessiot *et al.*, 2008) où les titres sont utilisés pour la construction de résumés, démontrant ainsi leur importance. Les résumés automatiques fournissent un ensemble de données pertinentes extraites du texte, mais toujours sous forme de phrase(s). Or, un titre n'est que très rarement une phrase. Il faut aussi distinguer le titrage automatique de la compression de texte classique (par exemple (Yousfi-Monod & Prince, 2006)), puisqu'un titre peut utiliser des reformulations du contenu du texte.

De même, le titre doit être différencié de l'index car ce premier ne contient pas toujours les termes clés du texte. Effectivement, le titre peut présenter une reformulation partielle ou totale du texte, ce qui n'est pas envisageable pour un index. Le rôle de l'index est de permettre une recherche facilitée pour l'utilisateur. Encore une fois, la construction d'index peut se servir des titres présents dans le document. Ainsi, si nous parvenons à déterminer des titres pertinents, la qualité de l'index sera grandement améliorée.

Finalement, le titre et le sous-titre sont des entités à part entière, possédant leurs propres fonctions et se distinguant nettement des tâches de résumé et d'index.

Il est admis que les éléments apparaissant dans le titre sont souvent présents dans le corps du texte (Baxendale, 1958; Vinet, 1993). Les récents travaux de (Lopez *et al.*, 2010b) et (Jacques & Rebeyrolle, 2004) viennent appuyer cette idée et montrent que la proportion de recouvrement des mots de titres est très importante dans le texte. Ainsi, une grande partie de l'information permettant la détermination d'un titre se trouve dans le document.

Une approche s'appuyant sur l'extraction de syntagmes nominaux (SN) pertinents pour leur utilisation en tant que titre, propose un processus efficace permettant de faire émerger l'information (Lopez *et al.*, 2010a). L'avantage de cette approche est que des titres longs peuvent être proposés. Le principal inconvénient est qu'elle ne peut proposer de titres originaux, utilisant une tournure humouristique par exemple, à moins que celle-ci apparaissent déjà dans le texte, ce qui est rare. Par ailleurs, l'efficacité de cette approche est limitée par l'absence (ou faible présence) de SN pertinents dans le texte à titrer. En effet, si aucun SN pertinent apparaît dans le texte (qui peut parfois être de courte taille en nombre de mots), cette approche ne peut proposer de titre.

Pour remédier à ces problèmes, cette étude propose une approche utilisant le Web, permettant de construire des titres courts à partir d'éléments présents dans le texte (dans un premier temps). Cette tâche est beaucoup plus complexe que l'extraction de syntagmes. Les titres construits doivent être cohérents, en rapport avec le texte, informatifs et accrocheurs [3]. Le Web apparaît comme une immense base textuelle appartenant à tous les domaines de la connaissance et constitue un corpus en évolution permanente (Duclaye *et al.*, 2006). L'utilisation du Web permet ainsi de traiter les sujets rares qui ne se retrouveraient peut-être pas dans un corpus figé. Ainsi, nous utiliserons notamment des techniques de Fouille du Web pour la construction de titres courts, s'appuyant sur une fonction de rang fondée sur des données statistiques acquises au moyen de l'interrogation d'un moteur de recherche sur Internet, similairement à (Turney, 2001).

3 Construction automatique de titres courts

L'objectif de la construction automatique de titres est de proposer des titres pertinents, en relation avec le contenu sémantique du texte à titrer. Dans cet article, nous nous intéressons à la construction de titres courts (par exemple utilisés en tant que sous-titres d'articles). En utilisant des méthodes de Fouille du Web, nous proposons un processus global composé de trois étapes principales (cf. Figure 1) :

1. Formation des titres candidats (section 3.2) : Un ensemble de titres candidats est proposé automatiquement à partir des données extraites du texte et respectant les patrons syntaxiques déterminés lors de nos études préliminares (section 3.1).
2. Cohérence des titres candidats (section 3.3) : Parmi les titres candidats formés à l'étape précédente, nous nous intéressons à leur cohérence par rapport au texte à titrer ainsi qu'à leur cohérence par rapport au Web (via Google).
3. Contextualisation dynamique des titres candidats (section 3.4) : Le contexte du texte et le contexte web de chaque titre candidat sont comparés afin de sélectionner le plus pertinent.

3. Nous définissons ces critères dans la section 4.2

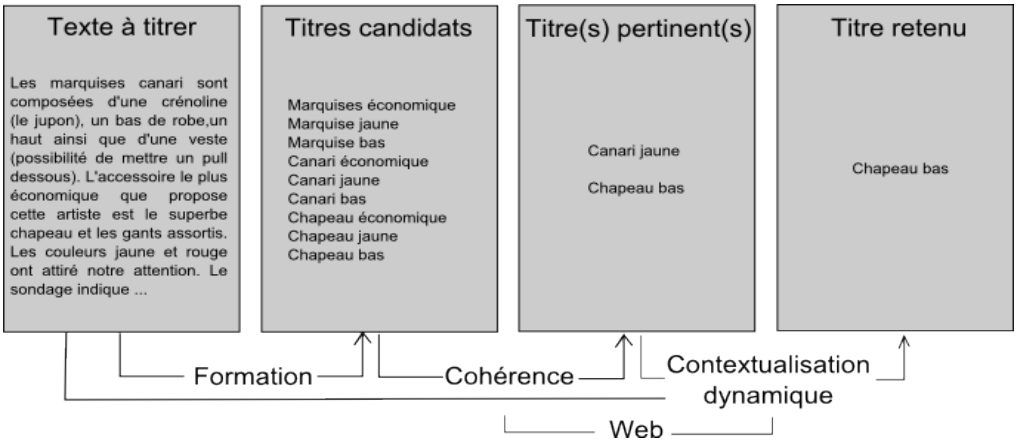


FIGURE 1 – Processus global de titrage automatique

Le premier travail consiste donc à analyser la structure morpho-syntaxique de titres courts.

3.1 Analyses préliminaires

Les articles journalistiques contiennent des sous-titres pouvant être simplement informatifs, mais aussi utilisant la présence de tournures humoristiques, l'emploi d'expressions, de citations. Nous considérons les sous-titres d'articles journalistiques comme des titres courts. Ainsi, notre étude statistique est réalisée sur les sous-titres d'articles journalistiques afin de déterminer ces patrons.

La base de données Factiva rassemble le texte intégral de plus de 8000 sources parmi lesquelles Le Monde est à disposition. Notre corpus d'étude a été constitué à partir de Factiva, sélectionnant 200 articles journalistiques français issus du quotidien Le Monde (novembre 2010) et contenant au moins un sous-titre. Afin que les résultats ne soient pas biaisés par les éventuelles erreurs induites par le choix d'un étiqueteur morpho-syntaxique, les sous-titres ont été analysés manuellement, selon 4 patrons morphosyntaxiques contenant des noms communs (NC), adjectifs (ADJ) et mots outils (MO : articles, déterminants, prépositions, etc.).

– 12% des sous-titres sont de la forme "NC" (ex. : " Objectifs ")
– 43% des sous-titres sont de la forme "$NC\ \ ADJ$" ou "$ADJ\ \ NC$" (ex. : "Paramètres sociopolitiques")
– 14% des sous-titres sont de la forme "$NC\ \ MO\ \ NC$" (ex. : "Hausse du budget")
– 26% des sous-titres contiennent quatre mots ou plus (ex. : "Les villepinistes s'élèvent contre la décision")

Compte tenu de ces résultats, nous décidons de nous intéresser plus particulièrement à la construction automatique de titres de la forme "$NC\ \ ADJ$" et "$ADJ\ \ NC$", qui couvrent 43% des sous-titres (ST) d'articles journalistiques issus de Le Monde. La section suivante consiste à construire des titres candidats de la forme "$NC\ \ ADJ$" et "$ADJ\ \ NC$".

3.2 Formation des titres candidats

La formation des titres candidats s'appuie sur le score TF-IDF (Salton & Buckley, 1988). Le TF-IDF est une mesure souvent utilisée en Recherche d'Information (RI) et Extraction d'Information (EI). Cette mesure est utilisée pour évaluer la pertinence d'un terme, en tenant compte de sa fréquence d'apparition au sein du texte et au sein du corpus. Un terme sera considéré pertinent s'il apparaît souvent dans le texte, et assez rarement dans le corpus.

La fréquence d'un terme (Term Frequency ou TF) est le nombre d'occurrences de ce terme dans le document considéré, normalisé par la somme des nombres d'occurrences de tous les termes du document. Ce nombre d'occurrence peut rendre compte de "l'importance" d'un terme dans un texte.

La fréquence inverse de document (Inverse Document Frequency ou IDF) permet de mesurer l'importance du

terme dans l'ensemble du corpus. Elle a pour intérêt de donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants.

Le TF-IDF est le produit de $TF_{i,j}$ par IDF_i. Notons que si un nouvel article est inséré dans le corpus, le TF-IDF est recalculé. Dans la suite, on notera $TF - IDF_X$ la valeur du TF-IDF obtenue pour X.

L'objectif est d'extraire les noms communs (NC) et adjectifs (ADJ) pertinents du texte à titrer. Après étiquettage du texte non lemmatisé [4] via le TreeTagger (Schmid, 1994), à chaque nom commun extrait est attribué un score correspondant au TF-IDF (noté $TF - IDF_{NC}$), permettant de classer les noms communs (NC) par ordre de pertinence, de "saillance". En revanche, à chaque adjectif (ADJ) extrait est attribué un score correspondant au TF simple (TF_{ADJ}). En effet, moins l'adjectif est spécifique, et plus la probablilité qu'il puisse être le qualificatif d'un nom commun est élevée.

Dans le texte à titrer, les trois noms communs de plus haut TF-IDF et les dix adjectifs de plus haut TF sont extraits. Cette limite est due au nombre de requêtes limitées sur les moteurs de recherche (Keller & Lapata, 2003).

Soit i le nombre de NC retenus, $1 <= i <= 3$ et j le nombre de ADJ retenus, $1 <= j <= 10$, tous les couples "$NC_i \ \ ADJ_j$" sont construits, i.e. maximum 30 titres candidats au total. Parmi eux, tous ne sont pas cohérents, en particulier concernant la grammaticalité (ex. : "chapeau belle"). La section suivante permet de déterminer la cohérence des titres candidats.

3.3 Cohérence des titres candidats

Alors que des couples potentiellement pertinents ont été construits dans la section précédente, il s'agit dans cette section de déterminer lesquels sont cohérents, à la fois grammaticalement et sémantiquement, pour leur utilisation en tant que titre. Cette cohérence est évaluée par rapport au texte (section 3.3.1), puis par rapport au Web (section 3.3.2).

3.3.1 Par rapport au texte

La cohérence des termes composant chaque titre candidat par rapport au texte est assurée par l'utilisation du TF-IDF lors de leur formation (cf. section 3.2). De cette façon, les noms communs et adjectifs les plus pertinents pour le titrage sont extraits.

Nous utilisons un autre critère de cohérence des titres candidats par rapport au texte, qui est la distance (en nombre de mots) entre les NC et les ADJ. Cette distance, notée $Dist_{NC-ADJ}$, est calculée pour chaque candidat puis utilisée dans le calcul du coefficient de distance [3].

$$Coef_{Dist} = \frac{1}{1 + Dist_{NC-ADJ}} \quad (1)$$

Si dans le texte, le candidat "$NC \ \ ADJ$" apparaît, on aura $Dist_{NC-ADJ} = 0$ et $Coef_{Dist}$ atteindra son maximum. Le candidat "$NC \ \ ADJ$" sera donc privilégié pour son utilisation en tant que titre. Cette distance est appliquée en tant que coefficient au score défini pour chaque candidat dans la suite de l'article.

3.3.2 Par rapport au Web

Un critère de cohérence par rapport au Web permet de valider la cohérence des titres candidats (TC) en se fondant sur le Web. Comme (Keller & Lapata, 2003; Béchet, 2009), nous utilisons la fréquence d'apparition de bigrammes sur le Web. Cette méthode permet notamment de mesurer la dépendance entre le nom commun et l'adjectif composant un titre candidat, d'où l'intérêt que ces derniers ne soient pas lemmatisés. On privilégie ainsi automatiquement un couple "$NC \ \ ADJ$" bien construit (ex. : "chapeau bas") par rapport à un couple mal construit (ex. : "chapeau basse"), cette dépendance entre nom et adjectif sur le Web étant largement induite par les accords en genre et en nombre entre ces termes.

4. Nous verrons dans la suite qu'il est primordial de ne pas lemmatiser dans notre cas

Dans l'objectif de mesurer cette dépendance le plus efficacement possible, nous comparons des mesures habituellement utilisées en Fouille du Web afin de déterminer laquelle est la plus adaptée à notre approche.

Soit $nb(X)$ la fonction retournant le nombre de pages renvoyées par le moteur de recherche (nous utiliserons Google) en réponse à la requête X et NC (resp. ADJ) un terme dont la nature est un nom commun (resp. un adjectif). Ainsi, $nb(NC)$ retourne le nombre de pages trouvées pour $X = NC$, ceci reflétant la popularité du terme NC sur le Web. De même, $nb(NC, ADJ)$ retourne le nombre de pages trouvées pour $X =$"$NC\ \ ADJ$".

Une des mesures les plus couramment utilisées en recherche d'information afin d'établir un classement est l'Information Mutuelle (IM) (Turney, 2001) définie comme suit [2] :

$$IM(x, y) = log_2 \frac{P(x, y)}{P(x) \times P(y)} \tag{2}$$

P(x,y) peut alors être vu comme la probabilité des réponses retournées par le moteur de recherche pour la requête $X =$"$NC\ \ ADJ$". Cette mesure vise à faire ressortir les co-occurrences les plus rares et les plus spécifiques (Daille, 1996; Thanopoulos *et al.*, 2002). Appliquée au contexte de la validation des bigrammes de la forme "$NC\ \ ADJ$", la formule [2] devient [3].

$$IM(NC, ADJ) = log_2 \frac{nb(NC, ADJ)}{nb(NC) \times nb(ADJ)} \tag{3}$$

L'information mutuelle au cube (IM^3) est une information empirique fondée sur l'information mutuelle, qui accentue l'impact des co-occurrences fréquentes, ce qui n'est pas le cas avec l'information mutuelle originale (Daille, 1994). Adaptée à la mesure de la cohérence des couples "$NC\ \ ADJ$", on obtient [4].

$$IM^3(NC, ADJ) = log_2 \frac{nb(NC, ADJ)^3}{nb(NC) \times nb(ADJ)} \tag{4}$$

Une mesure également intéressante en terme d'évaluation de qualité est le coefficient de Dice (Smadja *et al.*, 1996)[5].

$$DICE(x, y) = 2 \times \frac{P(x, y)}{P(x) + P(y)} \tag{5}$$

Adaptée, elle devient [6].

$$DICE(NC, ADJ) = 2 \times \frac{nb(NC, ADJ)}{nb(NC) + nb(ADJ)} \tag{6}$$

Ces différentes mesures statistiques adaptées à notre approche, permettent d'obtenir un classement tenant compte de la cohérence des titres candidats en fonction de leur pertinence sur le Web. Notons que $DICE$ et IM^3 privilégient les co-occurences (i.e. le numérateur) fréquentes par rapport à l'IM (Roche & Prince, 2008).

La comparaison de ces trois mesures est effectuée sur 20 articles journalistiques issus de Le Monde. Pour chaque article, 30 titres candidats de la forme "$NC\ \ ADJ$" ont été formés. Les scores indiqués dans le tableau 1 correspondent au nombre de titre(s) candidat(s) à la fois pertinent(s) par rapport au texte et grammaticalement corrects, parmi les cinq de plus haut score (selon $DICE$, IM, IM^3 et la simple prise en compte du nombre de pages nb retournés). Au total, ce sont 400 titres qui ont été manuellement expertisés.

Mesures	$DICE$	IM	IM^3	nb
Total	42	36	41	32

TABLE 1 – Évaluation de la cohérence des résultats selon différentes mesures.

Cette évaluation [5] indique que $DICE$, IM et IM^3 obtiennent des résultats similaires avec toutefois un meilleur résultat pour $DICE$ et IM^3. La simple utilisation du nombre de résultats bruts retournés par Google est la moins performante par rapport à notre application. Compte tenu de ces résultats, nous choisissons la mesure de $DICE$ dans la suite de notre travail.

Afin de prendre en compte les titres de la forme "$ADJ\ \ NC$" (cf. section 3.1), nous retenons la valeur maximum obtenue entre $DICE(ADJ, NC)$ et $DICE(NC, ADJ)$. Par exemple, on retiendra "beau chapeau" plutôt que "chapeau beau", le premier obtenant un score plus élevé que le second.

Pour chaque texte à titrer, 73 requêtes [6] sont nécessaires pour la formation des trente titres candidats.

Finalement, plusieurs candidats cohérents par rapport au texte et par rapport au Web, peuvent arriver en tête du classement. Parmi ces titres candidats nous devons déterminer quel est le plus pertinent pour son utilisation en tant que titre, en tenant compte du contexte de chacun d'entre eux.

3.4 Contextualisation dynamique

Pour un même document, plusieurs titres candidats peuvent être proposés. Afin de déterminer le titre le plus pertinent, nous comparons le contexte du texte à titrer avec le contexte dans lequel se retrouvent ces candidats sur le Web. Suite à la soumission d'une requête (via une API Google), le moteur de recherche Google présente les résultats sous forme d'une liste de sites Web. Pour chacun de ces sites, un aperçu du contenu de la page web est présenté (entre 10 et 30 mots), justifiant le résultat retourné en mettant en gras les termes initialement présents dans la requête. Le document utilisé pour la détermination du contexte Web de chaque titre candidat est la concaténation des 10 premiers aperçus (limite imposée par Google) d'une requête donnée. En ce qui concerne le contexte du texte, il est déterminé à partir du texte à titrer.

Pour déterminer le contexte Web et le contexte du texte, nous utilisons le modèle vectoriel de Salton (Salton *et al.*, 1975). Pour chaque nom commun et adjectif des documents (texte et document web), on détermine le TF qui constitue les coordonnées du vecteur contextuel (VCT pour le texte et VCW pour le Web). Finalement, à chaque titre candidat est associé un VCW. Si le vocabulaire associé à un contexte de titre candidat (VCW) est proche du vocabulaire du texte à titrer (VCT), alors nous privilégions ce candidat.

Pour chaque titre candidat, la similarité cosinus (ou mesure cosinus) est utilisée entre deux vecteurs couvrant tous les couples possibles de la forme (VCT_{Texte}, VCW_{Cand}). Ainsi, les candidats retenus sont ceux dont le contexte textuel est le plus "proche" du contexte Web.

Dans la section suivante, nous proposons une mesure globale réunissant la notion de cohérence des titres candidats et de contextualisation.

3.5 Mesure globale

En s'appuyant sur les méthodes précédemment définies, nous mettons en place une mesure globale, nommée $CATIT$ (Construction Automatique de TITres), permettant de mettre en avant les titres pertinents par rapport aux titres non pertinents, tenant compte à la fois de la cohérence des titres candidats par rapport au Web et au texte, ainsi que de leur contexte (dynamique). Cette mesure globale permet de fournir une fonction de rang globale prenant en compte tous les concepts.

Soit TI_{Cand} la fonction appliquée à un titre candidat, qui est le produit du TF-IDF du nom commun (NC) et du TF-IDF de l'adjectif (ADJ)[7].

$$TI_{Cand} = TF.IDF_{NC_{Cand}} \times TF.IDF_{ADJ_{Cand}} \tag{7}$$

La prise en considération de TI_{Cand} dans $CATIT$ permet de tenir compte de la pertinence de l'information contenu dans les termes composant les titres candidats.

5. Le détail des résultats est disponible sur `http://www.lirmm.fr/~lopez/Titrage_general/annexesTALN2011.pdf`, Table 1.

6. 3 requêtes pour NC + 10 requêtes pour ADJ + 30 requêtes pour "$NC\ \ ADJ$" + 30 requêtes pour "$ADJ\ \ NC$"

$$CATIT(Cand) = \begin{cases} Coef_{Dist} \times TI_{Cand} \times log_2(1 + cos(VCT_{Texte}, VCW_{Cand})), & si\ DICE(Cand) > K \\ Coef_{Dist} \times TI_{Cand} \times log_2(DICE(Cand)), & sinon. \end{cases}$$

$DICE(Cand)$ est toujours compris entre 0 et 1. Ainsi, avec l'utilisation de la fonction logarithme, les titres incohérents (inférieurs au seuil $K \in \mathbb{R}$ comparé à la mesure de DICE) seront toujours négatifs. Par ailleurs, le classement des candidats (via $DICE$) des titres négatifs sera aussi maintenu, grâce au log_2 qui est une fonction strictement croissante ($Coef_{Dist}$ et TI_{Cand} sont toujours positifs).
Au contraire, les titres cohérents (supérieurs au seuil K) seront toujours positifs, grâce au 1 qui correspond au maximum de la valeur de DICE possible.

Enfin, le classement par proximité contextuelle ($cos(VCT_{Texte}, VCW_{Cand})$) respecte l'ordre établi par le cosinus. Nous faisons intervenir la distance $Coef_{Dist}$ permettant de privilégier, parmi les candidats cohérents et contextuellement pertinents, ceux qui sont constitués de termes proches dans le texte (cf. section 3.3.1). Finalement, les titres candidats obtenant un résultat positif sont jugés pertinents par notre mesure. Le candidat obtenant le plus haut score est retenu pour son utilisation en tant que titre.

Le choix du seuil de pertinence K est crucial. Dans la section suivante, nous proposons une valeur de K puis évaluons notre mesure $CATIT$.

4 Évaluation

Cette section est dédiée à l'évaluation des titres construits par notre approche selon plusieurs critères. Les évaluations permettant de déterminer le seuil K puis la pertinence de notre approche $CATIT$ ont été effectuées par le premier auteur de ce papier.

4.1 Détermination du seuil K

Les résultats apportés par la mesure $CATIT$ dépendent fortement du seuil de pertinence K. Le comportement de ce seuil est analysé à partir des 10 premiers articles parus le 1er janvier 1994 dans le quotidien Le Monde, soient 900 titres évalués manuellement [7]. On ne cherchera pas à juger l'acceptabilité des trente candidats (cf. section 3.2) mais seulement leur grammaticalité. Différents seuils K_N sont testés (avec $N \in \{1, 10, 100\}$), fondés sur la moyenne des valeurs retournées par la mesure de Dice [8].

Cette détermination de K s'appuie sur la précision et le rappel, méthodes classiques d'évaluation en fouille de textes. Dans le cadre de ces mesures, un titre acceptable est un titre grammaticalement correct. Les résultats [8] sont présentés à la Figure 2.

$$K_N = \frac{moy(DICE(Cand))}{N} \tag{8}$$

L'utilisation du seuil K_1 n'est pas pertinente pour notre mesure car son utilisation entraînerait un élagage prématuré de nombreux candidats pouvant se révéler pertinents (précision élevée mais rappel faible). De même, l'utilisation du seuil K_{100} n'est pas pertinente pour notre mesure car de nombreux candidats incohérents sont conservés (précision faible mais rappel élevé). Finalement, les résultats (Figure 2) indiquent que le meilleur compromis entre précision et rappel est atteint avec K_{10}. Dans la suite de l'article, nous utiliserons donc le seuil K_{10}, que nous appliquerons lors de l'évaluation de $CATIT$.

7. 30 titres candidats × 10 articles × 3 seuils K
8. Le détail des résultats est disponible sur `http://www.lirmm.fr/~lopez/Titrage_general/annexesTALN2011.pdf`

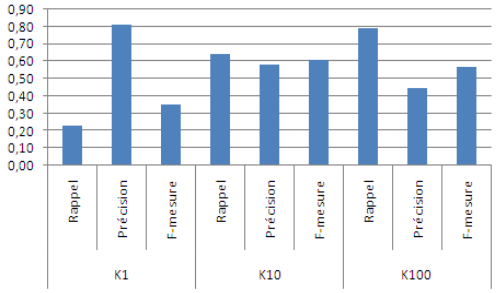


FIGURE 2 – Détermination de K

4.2 Évaluation de $CATIT$

Les titres construits automatiquement doivent répondre aux mêmes caractéristiques que les titres réels, définies dans la section 2. Le premier critère concerne l'information transmise par le titre, qui doit être en relation avec le texte traité. Si ce critère est constaté, nous conclurons que le titre est informatif (noté I). Le second critère concerne l'accroche. Un titre sera jugé accrocheur (noté A) s'il contient une tournure humoristique, une expression ou autre construction surprenant le lecteur, grammaticalement correct et informatif (en relation avec le texte). En effet, il ne sera pas convenable de juger un titre accrocheur s'il n'est pas en relation avec le texte. Par exemple, le titre "Chapeau bas" peut être considéré comme étant informatif (dans cet exemple, le texte rend hommage à un couturier qui propose entre autre des chapeaux) et accrocheur (emploi d'expression). Si le texte ne traitait pas de chapeaux et qu'il n'avait rien à voir avec l'expression "chapeau bas", on ne pourrait pas considérer le titre "chapeau bas" comme étant accrocheur, bien qu'il s'agisse d'une expression. Finalement, un objectif de cette évaluation est de détecter si nos titres sont "pertinemment accrocheurs".

Afin de tenir compte de ces critères dans l'évaluation, nous utilisons les méthodes classiques d'évaluation en fouille de textes (précision et rappel) adaptés aux critères A et I prédéfinis [9,10].

$$Rappel_{I(resp.A)} = \frac{Nb\ de\ titres\ I\ (resp.\ A)\ retenus}{Nb\ total\ de\ titres\ I(resp.A)} \tag{9}$$

$$Précision_{I(resp.A)} = \frac{Nb\ de\ titres\ I(resp.\ A)\ retenus}{Nombre\ total\ de\ titres\ retenus} \tag{10}$$

Enfin, les résultats compteront sur une mesure populaire qui combine la précision et le rappel, la F_{mesure} [9] [11].

$$F_{mesure} = 2 \times \frac{Précision \times Rappel}{Précision + Rappel} \tag{11}$$

L'évaluation est effectuée à partir d'articles journalistiques issus du journal quotidien Le Monde. Nous avons retenu les 20 premiers articles publiés le 1er janvier 1994. Ainsi, ce sont 600 titres issus de notre méthode $CATIT$, utilisant le seuil K_{10} (cf. section 4.1) qui ont été évalués manuellement en fonction de I et A (soit 1200 expertises au total). 1460 requêtes sur le moteur de recherche ont été nécessaires.

Les résultats de cette évaluation concernant la précision et le rappel sont présentés en Figure 3. En plus, pour chaque article, le titre de plus haut score retourné par $CATIT$, noté T1, est évalué (cf. Table 4). Nous notons "oui" lorsque le critère est respecté et "non" sinon. La présence du symbole "ensemble vide" indique qu'aucun titre parmi les 30 titres candidats correspond au critère demandé. Par exemple, parmi les 30 candidats construits à partir de l'article 1, aucun est informatif ou pertinent.

En ce qui concerne les titres informatifs, ils obtiennent une précision de 0,40 compensée par un rappel de 0,82 (cf. Table 3). Puisque les titres T1 sont informatifs dans 75% des cas (cf. Table 4), nous pouvons en déduire que le seuil

9. Nous utilisons la formule générique avec $\beta = 1$

K doit être affiné afin de retenir moins de titres candidats. Le moteur de recherche Google ne tenant pas compte de la présence de ponctuation dans les requêtes, un taux élevé de candidats constituent un bruit non négligeable. Un exemple directement lié à ce problème est le titre T1 de l'article 14 qui est mal construit. Notons que pour cet article, le deuxième titre de plus haut score est "Peines symboliques" qui est informatif et accrocheur. Par ailleurs, une erreur de la part de l'étiqueteur impacte fortement les résultats, surtout s'il s'agit d'une erreur concernant la détermination des trois noms communs (qui se répercute alors sur 10 titres candidats).

Du côté des titres accrocheurs, les mêmes difficultés sont rencontrées. De plus, nous avons constaté que très peu de candidats accrocheurs (maximum deux par article dans cette évaluation) sont construits, problème lié au nombre limité de noms communs et adjectifs retenus à la première étape de notre approche (cf. section 3.2). Ceci explique une précision faible et un rappel élevé. Notons tout de même que, malgré la relative rareté des titres candidats accrocheurs, 30% des titres construits par notre méthode sont accrocheurs.

Enfin, l'évaluation indique que 75% des titres T1 construits automatiquement par $CATIT$ sont informatifs et 30% sont accrocheurs (cf. Table 4). Ainsi, parmi les titres informatifs proposés, 40% sont accrocheurs, ce qui constitue un point positif pour notre approche. Finalement, nous comparons les titres T1 déterminés selon la méthode de titrage par Extraction de Syntagmes Nominaux (ESN) (Lopez *et al.*, 2010b). L'évaluation de ESN indique que seulement 60% des titres de la forme "nom adjectif" ou "adjectif nom" sont informatifs et 5% sont accrocheurs (voir Table 2).

Approche	ESN		CATIT	
Critères	I	A	I	A
Total	60%	5%	75%	30%

TABLE 2 – ESN versus CATIT

	T1		Rappel		Précision		F-mesure	
	I	A	I	A	I	A	I	A
Article 1	non	non	∅	∅	∅	∅	∅	∅
Article 2	oui	oui	0,75	0,50	0,50	0,33	0,60	0,40
Article 3	oui	oui	1,00	1,00	0,21	0,14	0,35	0,25
Article 4	oui	non	1,00	1,00	0,31	0,31	0,48	0,47
Article 5	oui	∅	0,86	∅	0,50	∅	0,63	∅
Article 6	oui	oui	0,83	1,00	0,50	0,40	0,63	0,57
Article 7	oui	x	0,80	∅	0,22	∅	0,35	∅
Article 8	oui	oui	0,67	1,00	0,57	0,17	0,62	0,29
Article 9	oui	∅	1,00	∅	0,38	∅	0,55	∅
Article 10	oui	∅	0,89	∅	0,47	∅	0,62	∅
Article 11	non	non	0,89	∅	0,53	∅	0,67	∅
Article 12	oui	non	1,00	∅	0,33	∅	0,50	∅
Article 13	oui	oui	0,83	1,00	1,00	0,20	0,91	0,33
Article 14	non	non	0,75	0,50	0,33	0,11	0,46	0,18
Article 15	oui	non	0,75	∅	0,21	∅	0,33	∅
Article 16	non	non	∅	∅	∅	∅	∅	∅
Article 17	non	non	0,50	∅	0,10	∅	0,17	∅
Article 18	oui	oui	0,50	1,00	0,25	0,13	0,33	0,22
Article 19	oui	non	0,80	1,00	0,44	0,11	0,57	0,20
Article 20	oui	non	1,00	∅	0,27	∅	0,43	∅
Total	75%	30%	0,82	0,89	0,40	0,21	0,51	0,32

TABLE 3 – Evaluation de $CATIT$

5 Conclusions et perspectives

La construction automatique de titres est une tâche complexe car des titres à la fois cohérents, grammaticalement corrects, informatifs et accrocheurs doivent être construits puis choisis parmi une liste de titres ne respectant pas ces critères. Dans cet article, nous avons proposé une approche permettant la construction automatique de titres courts. Même si les expertises ont été menées sur un corpus français (plus aisé à évaluer en terme d'information mais surtour d'accroche), la méthodologie décrite est intégralement reproductible dans d'autres langues, en particulier l'anglais. Après avoir sélectionné les candidats cohérents par des méthodes de Fouille du Web, les titres

	I	A	T1 (titre retenu)
Article 1	non	non	Sexuel destinées
Article 2	oui	non	Intérêt national
Article 3	oui	oui	Terre ennemie
Article 4	oui	non	Protection publique
Article 5	oui	∅	Service public
Article 6	oui	oui	Vieille lune
Article 7	oui	∅	Enseignement libre
Article 8	non	oui	Ottoman ottoman
Article 9	oui	∅	École laïque
Article 10	oui	∅	Avis défavorables
Article 11	non	non	Pays occidentaux
Article 12	oui	non	Économie espagnole
Article 13	oui	oui	Immigration économique
Article 14	non	non	Conditionnelle peines
Article 15	oui	oui	Grandes innondations
Article 16	non	non	Montée électrique
Article 17	non	non	Potable traitement
Article 18	oui	oui	Radioactivité nucléaire
Article 19	oui	non	Établissements hospitaliers
Article 20	oui	non	Gouvernement iranien
Total	75%	30%	

TABLE 4 – Evaluation des titres T1 avec $CATIT$

informatifs et accrocheurs sont choisis grâce à la mesure $CATIT$. L'évaluation indique que notre approche permet de titrer 75% des articles journalistiques de notre corpus de manière pertinente. Malgré une étape de construction des titres tenant compte de la cohérence grammaticale, certains titres contiennent des fautes d'orthographe. Un module de correction orthographique pourrait donc améliorer les résultats.

Il s'agit ici d'un premier travail d'évaluation, par introspection, donnant un aperçu des résultats obtenus avec notre méthode $CATIT$. Ces premiers résultats étant encourageants, le procédé d'évaluation sera développé dans nos prochains travaux, notamment par un jugement effectué selon plusieurs experts dans le but de consolider les résultats obtenus [10]. Avec un tel protocole d'évaluation, le coefficient Kappa (Cohen, 1960), qui propose de chiffrer l'intensité ou la qualité de l'accord réel entre des jugements qualitatifs appariés, pourra être utilisé.

La contextualisation est une étape importante de notre approche. Réalisée dynamiquement, elle permet de déterminer un titre contextuellement proche du texte et du Web. Un futur travail consistera à prendre en compte un contexte défini par l'utilisateur. Par exemple, les titres construits pourraient dépendre d'un contexte politique "de gauche" ou "de droite" selon le choix de l'utilisateur. De plus, une proposition de contexte "étendu", déterminé automatiquement à partir du contexte proposé par l'utilisateur pourrait permettre d'affiner le contexte, ceci supposant que le contexte fourni par l'utilisateur n'est que rarement pertinent.

L'approche présentée dans cet article utilise les termes du texte pour construire un titre. Cependant, les titres réels ne sont parfois pas composés de termes présents dans le texte référé. Dans l'objectif de construire des titres ne contenant pas de termes issus du document, un module sur la première étape de notre approche sera ajouté. Il consistera à enrichir la liste de noms communs et d'adjectifs en utilisant le réseau lexical populaire JeuxdeMots (Lafourcade & Joubert, 2009). De plus, le comportement de la construction automatique de titres avec des patrons syntaxiques plus complexes sera étudié. Dans ce cas, des processus de reformulation ou de nominalisation des verbes sont aussi à envisager.

Références

BAXENDALE B. (1958). Man-made index for technical literature - an experiment. *IBM Journal of Research and Development.*, p. 354–361.

10. Dans ce cadre, nous utiliserons un formulaire Web, développé dans le cadre de nos travaux sur l'extraction de syntagmes candidats au titrage (Lopez *et al.*, 2010b)

BÉCHET N. (2009). *Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes*. PhD thesis, Université Montpellier II.

COHEN J. (1960). A coefficient of agreement for nominal scales. In *Educ. Psychol. Meas.*, p. 27–46.

DAILLE B. (1994). Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. *Ph. D. thesis, Université Paris 7*.

DAILLE B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. *The Balancing Act : Combining Symbolic and Statistical Approaches to language.*, p. 29–36.

DUCLAYE F., COLLIN O. & PÉTRIER E. (2006). Fouille du web pour la collecte de données linguistiques : avantages et inconvénients d'un corpus horsnormes. In *6èmes journées francophones "Extraction et Gestion des Connaissances"*, p. 53–64.

HO-DAC L.-M., JACQUES M.-P. & REBEYROLLE J. (2004). Sur la fonction discursive des titres. *S. Porhiel and D. Klingler (Eds). L'unité texte, Pleyben, Perspectives.*, p. 125–152.

JACQUES M. & REBEYROLLE J. (2004). Titres et structuration des documents. In *Actes International Symposium : Discourse and Document.*, p. 125–152.

KELLER F. & LAPATA M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational linguistics*, **29**(3), 459–484.

LAFOURCADE M. & JOUBERT A. (2009). Similitude entre les sens dusage dun terme dans un réseau lexical. *Traitement Automatique des Langues*, **50**, 177–200.

LOPEZ C., PRINCE V. & ROCHE M. (2010a). Automatic titling of electronic documents by noun phrase extraction. In *Proceedings of Soft Computing and Pattern Recognition*, p. 168–171.

LOPEZ C., PRINCE V. & ROCHE M. (2010b). Titrage automatique de documents électroniques par extraction de syntagmes nominaux. In *Acte des 21èmes Journées Francophones d'Ingénierie des Connaissances*, p. 17–28.

MINEL J.-L., DESCLÉS J.-P., CARTIER E., CRISPINO G., BEN HAZEZ S. & JACKIEWICZ A. (2001). Résumé automatique par filtrage sémantique d'informations dans des textes. *Revue Techniques et Sciences Informatiques*.

PEÑALVER VICEA M. (2003). Le titre est-il un désignateur rigide ? *Dialnet, Vol. 2*, p. 251–258.

PESSIOT J., KIM Y., AMINI M., USUNIER N. & GALLINARI P. (2008). Une méthode contextuelle d'extension de requête avec des groupements de mots pour le résumé automatique. *Proceedings of the 5th Conférence en Recherche d'Information et Applications*.

PRINCE V. & LABADIÉ A. (2007). Text segmentation based on document understanding for information retrieval. In *Natural Language Processing and Information Systems*, p. 295–304 : Springer.

ROCHE M. & PRINCE V. (2008). Managing the acronym/expansion identification process for text-mining applications. *International Journal of Software and Informatics*, **2**(2), 163–179.

SALTON G. & BUCKLEY C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24*, p. 513 à 523.

SALTON G., WONG A. & YANG C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, **18**(11), 613–620.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, p. 44–49.

SMADJA F., MCKEOWN K. R. & HATZIVASSILOGLOU V. (1996). Translating collocations for bilingual lexicons : A statistical approach. *Computational linguistics*, **22**(1), 1–38.

THANOPOULOS A., FAKOTAKIS N. & KOKKIANAKIS G. (2002). Comparative evaluation of collocation extraction metrics. In *LREC'02*, volume 2, p. 620–625.

TURNEY P. (2001). Mining the web for synonyms : Pmi-ir versus lsa on toefl. In *Proceedings of ECML, LNCS*, p. 491–502.

VINET M.-T. (1993). L'aspet et la copule vide dans la grammaire des titres. *Persee*, **100**, 83–101.

YOUSFI-MONOD M. & PRINCE V. (2006). Compression de phrases par élagage d' arbre morpho-syntaxique. *TSI : Technique et Science Informatiques 25, 4*, p. 447–456.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Une approche faiblement supervisée pour l'extraction de relations à large échelle

Ludovic Jean-Louis Romaric Besançon Olivier Ferret Adrien Durand
CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
Fontenay-aux-Roses, F-92265, France.
{ludovic.jean-louis,romaric.besancon,olivier.ferret,adrien.durand}@cea.fr

Résumé. Les systèmes d'extraction d'information traditionnels se focalisent sur un domaine spécifique et un nombre limité de relations. Les travaux récents dans ce domaine ont cependant vu émerger la problématique des systèmes d'extraction d'information à large échelle. À l'instar des systèmes de question-réponse en domaine ouvert, ces systèmes se caractérisent à la fois par le traitement d'un grand nombre de relations et par une absence de restriction quant aux domaines abordés. Dans cet article, nous présentons un système d'extraction d'information à large échelle fondé sur un apprentissage faiblement supervisé de patrons d'extraction de relations. Cet apprentissage repose sur la donnée de couples d'entités en relation dont la projection dans un corpus de référence permet de constituer la base d'exemples de relations support de l'induction des patrons d'extraction. Nous présentons également les résultats de l'application de cette approche dans le cadre d'évaluation défini par la tâche KBP de l'évaluation TAC 2010.

Abstract. Standard Information Extraction (IE) systems are designed for a specific domain and a limited number of relations. Recent work has been undertaken to deal with large-scale IE systems. Such systems are characterized by a large number of relations and no restriction on the domain, which makes difficult the definition of manual resources or the use of supervised techniques. In this paper, we present a large-scale IE system based on a weakly supervised method of pattern learning. This method uses pairs of entities known to be in relation to automatically extract example sentences from which the patterns are learned. We present the results of this system on the data from the KBP task of the TAC 2010 evaluation campaign.

Mots-clés : extraction d'information, extraction de relations.

Keywords: information extraction, relation extraction.

1 Introduction

Dans le cadre de l'extraction d'information, l'extraction de relations est un processus dont l'objectif est de déterminer l'existence d'un lien sémantique entre deux entités et lorsque cela est possible, de caractériser la nature de ce lien. Nous nous intéressons plus particulièrement dans cette étude à l'extraction de relations entre entités nommées en vue de la collecte et de la construction d'une base de connaissances à large échelle. En effet, on trouve dans des sources d'informations ouvertes, en particulier dans le contexte du Web sémantique, un grand nombre d'informations disponibles sous forme semi-structurée : par exemple, l'encyclopédie Wikipédia contient des informations qui peuvent être structurées sous forme d'une base de données, comme le montre le projet DBpedia[1] (Bizer *et al.*, 2009). Cette structuration première des informations semi-structurées peut alors être complétée par l'extraction automatique de relations entre entités à partir de texte brut.

Les travaux ayant pour objet l'extraction de relations peuvent être considérés selon l'angle du degré de supervision qu'ils requièrent. Au degré le plus faible, que l'on qualifie d'approche non supervisée, le type des relations à extraire n'est pas défini *a priori*, que ce soit par le biais d'exemples ou d'un modèle. Tout au plus peuvent être fixées certaines contraintes sur les entités reliées, comme leur type par exemple. Le type des relations extraites est quant à lui défini *a posteriori*, en regroupant les relations jugées similaires. Une telle approche est mise en œuvre dans (Shinyama & Sekine, 2006) ou dans (Banko & Etzioni, 2008) par exemple. À l'autre extrême de cette échelle, le type des relations visées mais aussi les moyens de les extraire à partir des textes sont définis *a priori*. Cette approche dite supervisée se caractérise soit par la donnée d'un modèle élaboré manuellement, typiquement sous la forme de règles, soit par l'association d'un ensemble d'exemples de relations en contexte issus de l'annotation d'un corpus et d'un algorithme d'apprentissage permettant d'en construire automatiquement un modèle. Cette seconde option est dominée par les modèles d'apprentissage statistique, qui se focalisent sur la prise en compte d'un large spectre de caractéristiques de différents types (lexicales, syntaxiques, sémantiques ...) (Zhou *et al.*, 2005) et sur l'élaboration de fonctions noyaux permettant de prendre en compte ces caractéristiques, en particulier lorsqu'elles ont des structures complexes comme celles produites par l'analyse syntaxique (Zhou *et al.*, 2007).

Entre ces deux pôles se trouvent les approches dites faiblement supervisées, vocable recouvrant l'idée que des exemples ou un modèle sont fournis pour le développement du système d'extraction de relations mais que cette seule contribution n'est pas suffisante pour la réalisation d'un système pleinement opérationnel. De ce fait, elle doit être étendue de manière automatique, généralement en exploitant un corpus non annoté. Les travaux existant en la matière font apparaître deux cas de sous-détermination de la contribution initiale, cas pouvant être éventuellement associés :

– une sous-détermination liée au volume de cette contribution. Seul un petit ensemble de relations exemples ou un modèle incomplet sont fournis ;
– une sous-détermination liée à la nature de la contribution initiale, ce qui se produit lorsque les exemples ou le modèle doivent être instanciés pour être utilisés.

Le premier cas de figure est typiquement traité suivant la méthodologie initiée par Hearst (1992) grâce à un mécanisme d'amorçage exploitant le petit ensemble initial d'exemples de relations ou de règles d'extraction pour acquérir de nouveaux exemples à partir d'un corpus et venir ainsi enrichir progressivement le modèle des relations visées au fil de cycles successifs d'application de ces deux étapes. (Agichtein & Gravano, 2000) en est un représentant typique pour les relations entre entités nommées. Bien qu'opérant dans un champ différent – l'extraction de structures qualia – (Claveau & Sébillot, 2004) offre un autre exemple d'amorçage pour l'induction de patrons linguistiques en combinant deux systèmes aux caractéristiques différentes.

Le second cas de figure est quant à lui illustré par la notion récente de « Distant supervision », introduite formellement par (Mintz *et al.*, 2009) mais déjà présente dans certains travaux sur l'amorçage. Les exemples sont ici donnés sous une forme sous-déterminée puisque réduite à un couple d'entités : ils sont donc à la fois privés de contexte et de caractérisation linguistiques. Le développement de ce type d'approches est favorisé par la mise à disposition de larges bases de connaissances extraites de ressources telles que Wikipédia.

Dans cet article, nous présentons un système d'extraction d'information à large échelle fondé sur un apprentissage faiblement supervisé de patrons d'extraction de relations reposant sur des exemples sous la forme de couples d'entités. Ces couples sont projetés dans un corpus de référence pour constituer la base d'exemples de relations à partir

[1] `http://dbpedia.org/About`

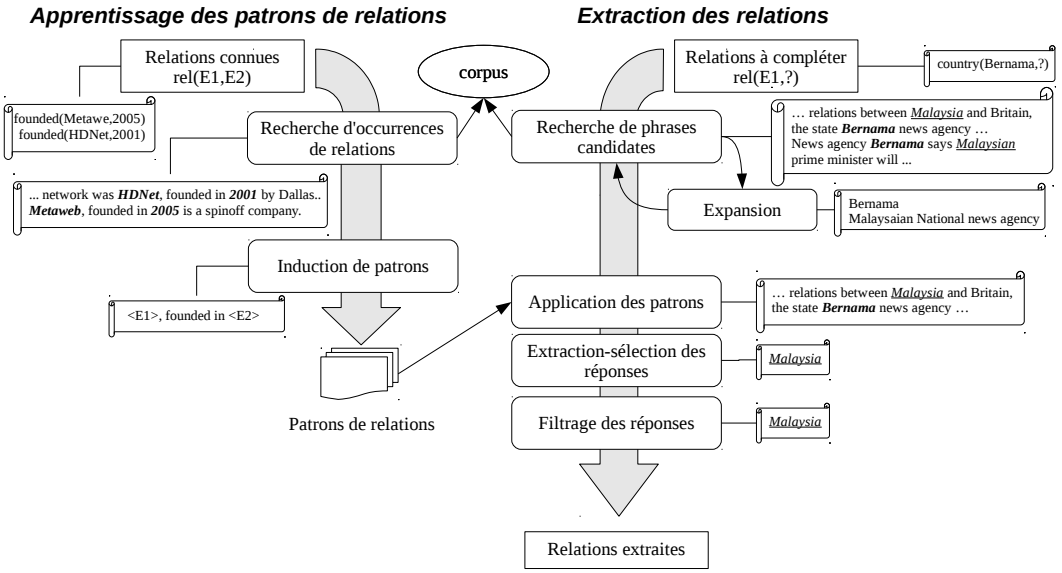


FIG. 1 – Architecture générale du système

de laquelle les patrons d'extraction sont appris. Ce travail se rattache donc au concept de « Distant supervision ». Nous présentons également les résultats de l'application d'une telle approche dans le cadre d'évaluation défini par la tâche KBP (Knowledge Based Population) de l'évaluation TAC 2010 (Text Analysis Conference).

2 Présentation de l'approche

Nous nous concentrons dans notre approche sur l'extraction de relations à large échelle en supposant la préexistence d'une base de connaissances partiellement remplie, extraite automatiquement à partir de données semi-structurées. Nous nous limitons ici aux relations entre entités nommées étant donné que, n'intervenant pas en domaine de spécialité où la recherche des entités peut être guidée par une terminologie existante, nous avons volontairement choisi de nous focaliser sur des entités aisément identifiables. La notion de « large échelle » se décline quant à elle selon plusieurs dimensions. La première réside dans le grand nombre de types de relations différents considérés, induisant une mise en œuvre difficile pour une approche à bases de règles écrites manuellement. La deuxième est liée à la prise en compte initiale d'un grand nombre de relations existantes (c'est-à-dire l'association de deux valeurs d'entités à un type de relation) ; ces relations fournissent un bon ensemble de départ pour l'apprentissage automatique d'un modèle de ces types de relations. Enfin, le corpus dans lequel de nouvelles relations sont recherchées est lui-même important, ce qui implique l'utilisation de techniques d'indexation et de recherche pour extraire des bons candidats (on ne peut pas envisager l'application directe de patrons sur toutes les phrases du corpus). Cette approche, illustrée par la figure 1, s'articule en deux phases : une phase d'*apprentissage de patrons* à partir d'occurrences de relations connues et une phase d'*extraction de relations* pour la découverte de nouvelles relations. La première phase part des relations connues *R(E1,E2)* pour trouver des occurrences de ces relations dans un corpus, c'est-à-dire les différentes expressions de cette relation dans les textes et utiliser ces occurrences pour induire des patrons de reconnaissance pour le type de relation concerné. La seconde phase part de relations incomplètes *R(E1,x)*, où l'entité source *E1* est connue et l'entité cible *x* est à trouver, cherche des occurrences de relations impliquant *E1* dans un corpus, puis extrait l'entité *x* en utilisant les patrons induits dans la première phase. Ces deux phases sont détaillées dans les sections suivantes.

2.1 Apprentissage des patrons

L'apprentissage des patrons de relations repose sur l'induction (ou généralisation) de patrons lexicaux à partir de phrases exemples contenant des occurrences des relations considérées. L'objectif de cet apprentissage est de

capturer les différentes expressions d'une relation sémantique entre deux entités. Par exemple, les deux extraits de phrases ci-dessous contiennent des occurrences de relations pour le type *founded_by*, instancié pour les couples d'entités (Charles Revson ; Revlon Cosmetics) et (Mayer Lehman ; Lehman Brothers investment).

The glamourous cabaret chanteuse reportedly had had a romantic liaison with <source>Charles Revson</source>, the founder of <cible>Revlon Cosmetics</cible> ... – Lehman was a great-grandson of <source>Mayer Lehman</source>, a founder of the <cible>Lehman Brothers investment</cible> house ...

Plusieurs travaux présentent des algorithmes de généralisation de patrons lexicaux (Ravichandran, 2005; Schlaefer *et al.*, 2006; Ruiz-Casado *et al.*, 2007). Notre approche est similaire à celle de (Pantel *et al.*, 2004) et reprend plus directement encore la méthode de (Embarek & Ferret, 2008). L'idée générale de l'approche est de trouver, dans le contexte entre les entités cible et source, des points communs entre deux phrases exprimant la relation que l'on veut capturer. Ici, nous cherchons ces points communs parmi trois niveaux d'information linguistique : forme de surface, lemme et catégorie morpho-syntaxique. Ces informations linguistiques sont mises en évidence grâce à l'outil OpenNLP[2], qui est plus globalement également utilisé pour la reconnaissance des entités nommées. La présence de ces trois niveaux d'information donne une plus grande expressivité aux patrons construits et permet ainsi de trouver un compromis intéressant en termes de niveau de généralisation entre la spécificité des éléments lexicalisés et le caractère plus général des catégories morpho-syntaxiques.

L'induction d'un patron à partir de deux occurrences de relation est plus précisément composée des trois étapes suivantes :

– le calcul de la distance d'édition entre les deux phrases exemples, c'est-à-dire le nombre minimal d'opérations d'éditions (insertion, suppression, substitution) à effectuer pour passer d'une phrase à l'autre. Toutes les opérations ont ici le même poids ;
– l'alignement optimal des phrases exemples à partir de la matrice des distances entre sous-séquences issue du calcul de la distance d'édition. L'algorithme classique pour trouver un tel alignement est ici étendu en permettant la mise en correspondance de deux mots lors d'une substitution selon les trois niveaux d'information possibles ;
– construction des patrons en complétant si nécessaire les alignements par des opérateurs *jokers (*s*)*, représentant 0 ou 1 mot quelconque, et *(*g*)*, représentant exactement un mot quelconque.

Le tableau 1 montre un exemple d'induction de patron pour le type de relation *founded_by* à partir des deux extraits de phrases ci-dessus. On peut noter la présence de la catégorie *DET* (déterminant) comme généralisation pour *(a|the)*, ce qui rend le patron pertinent pour d'autres extraits tels que *"Charles Kettering, another founder of DELCO ..."*.

Charles Revson	,	the	founder	of		Revlon Cosmetics
Mayer Lehman	,	a	founder	of	the	Lehman Brothers investment
<source>	**,**	***DET***	***founder***	***of***	***(*s*)***	*<cible>*

TAB. 1 – Exemple d'induction de patron de relation

Cet exemple illustre également le fait que la généralisation peut aller jusqu'à l'utilisation de jokers pouvant se substituer à n'importe quel mot. Comme il est toujours possible de généraliser deux phrases en un patron ne contenant que des jokers, il est nécessaire de fixer une limite supérieure au nombre de jokers pouvant être introduits dans une opération de généralisation pour conserver un niveau de spécificité raisonnable des patrons. Par ailleurs, travaillant en domaine ouvert et avec des entités nommées assez générales, nous souhaitons plutôt induire un nombre important de patrons spécifiques qu'un ensemble restreint de patrons très généraux, ceci afin de privilégier la précision. C'est également pour cette raison que nous ne cherchons pas à généraliser les patrons en leur réappliquant la procédure d'induction décrite. Dans l'évaluation présentée en section 3, le nombre maximal de jokers dans un patron est donc fixé à 1.

Dans le contexte de supervision distante dans lequel nous nous plaçons, les phrases exemples ne sont pas directement fournies en tant que telles mais résultent de la projection dans un corpus de relations se présentant sous la forme de couples d'entités (par exemple le couple (Ray Charles, Albany) pour le type de relation *city_of_birth*). Plus concrètement dans notre cas, elles sont récupérées en soumettant à un moteur de recherche des requêtes conte-

[2] `http://opennlp.sourceforge.net/index.html`

nant des couples d'entités pour un type de relations donné et en restreignant les résultats du moteur aux phrases contenant effectivement les deux valeurs des entités. On peut souligner que la nature des restrictions appliquées a un impact direct sur la quantité et la précision des patrons induits. Plus on impose de contraintes, moins on obtient de phrases exemples, mais meilleurs seront les patrons induits. Par exemple, les auteurs de (Agirre *et al.*, 2009) ne retiennent que les phrases exemples dans lesquelles les paires d'entités apparaissent dans un voisinage de zéro à dix mots.

Il est important de noter que le processus d'induction de patrons s'effectue en comparant les phrases exemples deux à deux. Il peut donc être coûteux (en temps de calcul) lorsque le nombre de phrases exemples est important : pour 10 000 exemples, on a environ 50 millions de couples distincts de phrases à comparer ($n(n-1)/2$ exactement). Pour traiter ce problème, la solution immédiate consiste à réduire de façon drastique le nombre de phrases exemples en amont du processus d'induction, la conséquence étant une réduction de la couverture des différentes forme d'expression des types de relations. Une autre solution consiste à faire une réduction sélective du nombre de couples de phrases exemples à généraliser en évitant de considérer les couples de phrases dont la distance est visiblement trop grande pour induire des patrons intéressants. Même si la distance utilisée pour cette induction est une distance d'édition, donc tenant compte de l'ordre des mots, il est évident qu'un faible recoupement des phrases en termes de mots conduira à une valeur élevée de la distance d'édition. Le filtrage *a priori* des couples de phrases peut donc se fonder sur une mesure s'appliquant à une représentation de type « sac de mots », telle que la mesure *cosinus*, en fixant une valeur minimale en dessous de laquelle la généralisation des couples de phrases n'est pas réalisée. Or, la mesure *cosinus* peut être évaluée de manière efficace, soit avec une bonne approximation, comme dans le cas du *Local Sensitive Hashing* (Gionis *et al.*, 1999), soit de manière exacte mais en fixant un seuil de similarité minimale, ce qui correspond à notre cas de figure. Nous avons donc retenu pour notre filtrage l'algorithme *All Pairs Similarity Search* (APSS), proposé dans (Bayardo *et al.*, 2007), qui calcule la mesure *cosinus* pour les seules paires d'objets considérés – ici, les phrases exemples – dont la similarité est supérieure ou égale à un seuil fixé *a priori*. Cet algorithme se fonde plus précisément sur une série d'optimisations dans l'indexation des objets tenant compte des informations recueillies sur leurs caractéristiques et d'un tri appliqué à ces objets en fonction de ces mêmes caractéristiques.

Notons que lors de l'induction de patrons à partir d'un grand volume de phrases exemples, on retrouve de nombreux doublons, soit parce que la même phrase exemple se trouve dans plusieurs documents, soit parce que l'on retrouve la même forme d'expression d'un type de relations avec des valeurs différentes (*Obama's height is 1.87m* ; *Sarkozy's height is 1.65m*). Ainsi, nous proposons de filtrer les phrases exemples à deux niveaux : d'abord avec un seuil de similarité fort afin d'identifier et éliminer les phrases identiques ; puis avec un seuil de similarité faible pour s'assurer d'un niveau minimal de similarité entre les phrases en vue du processus d'induction.

2.2 Extraction des relations

L'extraction de nouvelles relations se fait à partir des types de relations existants et d'entités connues : on cherche à compléter une base de connaissances existante en complétant les informations concernant les entités qu'elle contient. La première étape de l'extraction de relations est la recherche de phrases candidates pouvant contenir l'expression d'une relation. Elle prend comme point de départ des requêtes contenant une entité nommée associée à son type et le type de l'information recherchée. La recherche proprement dite est réalisée, comme dans le cas de l'apprentissage de patrons, grâce à un moteur de recherche ayant préalablement indexé le corpus cible pour l'extraction des relations. Nous nous sommes appuyés dans notre cas sur le moteur Lucène[3], avec une indexation adaptée aux caractéristiques de notre recherche : les documents initiaux sont découpés en unités d'indexation de petite taille, trois phrases, grâce à une fenêtre glissante et au sein de ces unités, sont indexés les mots pleins sous leur forme normalisée et les entités nommées, avec leur type. L'interrogation du corpus présente en outre la particularité d'inclure une phase d'expansion de l'entité source. En effet, on retrouve souvent dans les documents des formes plus ou moins développées des entités nommées : par exemple *Bill Clinton* est généralement utilisé au lieu de *William Jefferson Blythe III Clinton*. Il est donc intéressant de savoir que ces deux mentions d'entités sont équivalentes et associées à la même entité, en particulier lors de la recherche de documents. Nous utilisons donc une étape d'expansion des entités visant à associer à une entité donnée les formes alternatives lui faisant référence. Pour l'entité "Barack Obama", on a ainsi : *{B. Hussein Obama, Barack H. Obama Junior, Barack Obama Jr, Barack Hussein Obama Jr., etc.}*. L'intérêt est de pouvoir augmenter les chances de retrouver des phrases candidates

[3] `http://lucene.apache.org/java/docs/index.html`

liées à l'entité puisque l'on considère tous les documents dans lesquels apparaissent ses différentes expressions. Une base d'expansion des entités a été constituée de façon automatique à partir du corpus Wikipédia[4] en collectant pour chaque entité les formulations extraites des pages de redirection de Wikipédia vers cette entité. Au total, la base d'expansion contient des formes étendues pour environ 2,4 millions d'entrées.

Nous appliquons ensuite sur les phrases candidates sélectionnées les patrons induits lors de la phase d'apprentissage. Les entités cibles extraites par ces patrons sont cumulées pour ne retenir finalement que les plus fréquentes : notre hypothèse est que les entités cibles les plus pertinentes apparaissent plus souvent dans les documents que les moins pertinentes. Pour les relations mono-valuées (ex. : date de naissance), une seule valeur est conservée. Pour les relations multi-valuées (ex. : lieux de résidence), un nombre arbitraire de trois valeurs sont conservées à défaut de connaissances fournies *a priori* ou extraites des textes quant à ce point. Enfin, un dernier filtre est appliqué sur les entités cibles pour vérifier la compatibilité des valeurs obtenues avec les contraintes relatives au type d'information recherché qu'elle représentent, définies par des listes de valeurs ou d'expressions régulières : on vérifie par exemple que le pays de naissance d'une personne fasse bien partie de la liste des pays connus.

3 Évaluation

Nous présentons dans cette section les résultats de l'évaluation de notre système en utilisant les données de la tâche *Slot Filling* de la campagne d'évaluation TAC-KBP 2010 (TAC-KBP, 2010). Nos expérimentations ont donc été réalisées pour l'anglais. La tâche *Slot Filling* correspond aux spécifications de notre contexte de travail telles que nous les avons définies à la section 2 : son objectif est d'extraire à partir d'un vaste corpus l'entité cible d'une relation ayant comme source une entité présente dans une base de connaissances abritant un ensemble important d'exemples du type de relation visé. Les types de relations considérés dans ce cadre sont au nombre de 42, répartis en 16 relations pour des entités de type ORGANISATION (ORG) et 26 relations pour les entités de type PERSONNE (PERS). La liste des types de relations traités est présentée dans le tableau 2. Nous précisons que les expériences ont été réalisées sur un cluster de 24 nœuds (4 processeurs/nœud) avec une parallélisation par type de relations.

3.1 Cadre d'évaluation

Les données d'évaluation issues de TAC-KBP sont les suivantes :

- un corpus de textes composé d'environ 1,8 millions de documents (1 780 980 exactement) répartis en 0,04% de transcriptions (conversations téléphoniques, journaux radio, conversations radio), 72,24% d'articles de presse et 27,72% de pages Web ;
- une base de connaissances *(KB)* reposant sur une image de Wikipédia d'octobre 2008. Un identifiant unique et un type d'entité sont attribués à chaque page contenant des *infobox*. Le type d'entité *personne*, *organisation*, *entité géopolitique* ou *inconnu* est associé à chaque page en fonction des champs contenus dans les infobox. Typiquement, les infobox *Infobox_Actor* sont ainsi liées à des personnes. Au final 818 741 entités ont été retenues pour former la KB, chacune d'elles étant associée à un ensemble de propriétés (champs des infobox) ainsi qu'à un texte la décrivant. Ainsi les relations sont représentées dans la KB par des tuples (identifiant, type infobox, nom, type, propriété, valeurs), ex. : (E0000437 ; Infobox_Actor ; Julia Roberts ; PER ; birthplace ; Atlanta) ;
- une table de correspondance entre les propriétés issues de Wikipédia et les types de relations retenus pour l'évaluation. Par exemple, *Infobox_Actor*:*birthplace* est convertie en *per*:*city_of_birth*. Cette correspondance permet de prendre en compte une certaine hétérogénéité de désignation des propriétés dans Wikipédia ;
- une liste de 100 entités sources pour lesquelles on cherche à extraire toutes les entités en relation pour tous les types de relations considérés. On dénombre parmi ces entités 15 entités présentes dans la KB et 85 inconnues de la KB. Par ailleurs, toutes les relations considérées ne trouvent pas d'entités cibles dans le corpus pour ces 100 entités. Dans le cadre de cette étude, nous nous focalisons uniquement sur les relations pour lesquelles il existe une entité cible dans le corpus[5], ce qui représente au total 2069 relations. Le détail par type de relations est présenté dans la colonne *Nb Ref.* du tableau 2.

[4] Plus précisément, la version mise à disposition par l'université de New York : `http://nlp.cs.nyu.edu/wikipedia-data`

[5] Les entités cibles existantes dans le corpus sont établies par la référence fournie par les organisateurs de la campagne, construite à partir des résultats des participants.

Types de relations	Type de cible	Couv. Doc.	Couv. Rel.	Nb Appr.	Nb Test	Nb Induc.	Nb Patrons	Couv. Patrons	Nb Ref.
org:alternate_names	ORG	89,17%	33,33%	20 013	10 006	214	6 007	66,10%	120
org:city_of_headquarters	LOC + liste	90,12%	59,26%	6 847	3 423	4 553	2 010 749	65,52%	81
org:country_of_headquarters	LOC + liste	91,04%	55,22%	18 401	9 200	2 110	185 158	69,56%	67
org:dissolved	DATE	100%	25%	532	266	87	775	0%	4
org:founded_by	ORG/PER	95,45%	31,82%	1 954	977	197	4 385	77,87%	28
org:founded	DATE	92,86%	53,57%	13 688	6 844	127	22 482	77,34%	22
org:member_of	ORG	100%	100%	7 951	3 976	102	103	70%	2
org:members	ORG	77,78%	11,11%	531	265	183	552	86%	9
org:number_of_employees_members	regexp + liste	90,48%	23,81%	7 173	3 586	216	3 109	100%	21
org:parents	ORG	96,67%	43,33%	22 361	11 181	3 013	485 947	69,04%	30
org:political_religious_affiliation	ORG	78,57%	64,29%	3 427	1 713	406	3 250	55,36%	14
org:shareholders	ORG/PER	66,67%	33,33%	3	2	0	0	0%	3
org:stateorprovince_of_headquarters	LOC + liste	92,65%	63,24%	9 672	4 836	1 422	148 610	69,93%	68
org:subsidiaries	ORG	82,69%	28,85%	5 588	2 794	498	3 764	56,48%	52
org:top_members_employees	PER	91,48%	37,22%	40 929	20 464	108	1 010	70,57%	223
org:website	regexp	78,26%	30,43%	30 813	15 407	32	28	0%	23
per:age	regexp + liste	85,32%	32,11%	157	79	3	1	0%	109
per:alternate_names	PER	61,63%	11,63%	18 115	9 057	68	2 818	82,58%	86
per:cause_of_death	liste	100%	0%	1	1	0	0	0%	2
per:charges	liste	61,54%	0%	184	92	0	0	0%	13
per:children	PER	72%	16%	2 010	1 005	147	238	0%	25
per:cities_of_residence	LOC + liste	77,59%	34,48%	3 631	1 815	722	14 297	77,88%	58
per:city_of_birth	LOC + liste	69,23%	15,38%	4 745	2 373	2 252	62 455	63,34%	13
per:city_of_death	LOC + liste	100%	100%	1 631	816	505	2 860	70,27%	1
per:countries_of_residence	LOC + liste	73,53%	20,59%	8 098	4 049	2 181	205 344	80,08%	34
per:country_of_birth	LOC + liste	82,35%	5,88%	11 085	5 542	11 192	9 145 385	65,02%	17
per:country_of_death	LOC + liste			2 873	1 436	1 068	22 374	62,89%	0
per:date_of_birth	DATE	90%	20%	11 689	5 845	30	22	0%	20
per:date_of_death	DATE	100%	0%	4 692	2 346	54	63	33,33%	1
per:employee_of	ORG	84,21%	29,32%	24 762	12 381	2 435	704 833	71,13%	133
per:member_of	ORG	82,42%	36,26%	27 523	13 761	3 901	740 999	57,25%	91
per:origin	liste	81,58%	42,11%	37 626	18 813	2 710	276 653	74,41%	76
per:other_family	PER	86,67%	33,33%	4	2	0	0	0%	30
per:parents	PER	78,13%	9,38%	1 314	657	37	604	77,78%	64
per:religion	liste	85,71%	57,14%	1 468	734	515	1 575	80%	7
per:schools_attended	ORG + liste	87,50%	37,50%	2 246	1 123	67	170	4,17%	16
per:siblings	PER	78,26%	20,29%	4	2	0	0	0%	69
per:spouse	PER	80%	35,56%	5 385	2 693	3 094	314 329	80%	45
per:stateorprovince_of_birth	LOC + liste	80%	50%	7 047	3 523	2 097	60 782	75,42%	10
per:stateorprovince_of_death	LOC + liste	100%	100%	1 616	808	278	911	66,67%	1
per:states_or_provinces_of_residence	LOC + liste	84,21%	50%	4 980	2 490	1 166	115 418	77,90%	38
per:title	liste	84,55%	52,77%	31 574	15 787	8 797	1 573 512	49,07%	343

TAB. 2 – Résultats des différentes étapes, pour tous les types de relations

Type de cible : mécanisme utilisé pour retrouver l'entité cible. *Couv. Doc.* : couverture des documents de référence dans les résultats de la recherche de phrases. *Couv. Rel.* : couverture des phrases candidates de référence. *Nb Appr.* : nombre de relations pour l'apprentissage des patrons. *Nb Test* : nombre de relations pour l'évaluation des patrons. *Nb Induc.* : nombre de phrases contenant des occurrences de relations pour l'induction des patrons. *Nb Patrons* : nombre de patrons induits à partir des occurrences de relations. *Couv. Patrons* : couverture des patrons induits. *Nb Ref.* : nombre de relations de référence.

3.2 Évaluation de l'apprentissage des patrons

Les patrons servent à confirmer/infirmer la présence d'une relation entre deux entités. Il est donc important de vérifier que les patrons appris aient une couverture suffisamment large pour retrouver le plus possible de variantes parmi les occurrences de relations. Pour évaluer la qualité des patrons, nous avons séparé les relations connues en deux ensembles : un ensemble d'apprentissage (2/3 des relations) et un ensemble de test (1/3 des relations). Nous mesurons la qualité de la couverture des patrons en calculant le pourcentage des occurrences de relations de l'ensemble de test que l'on retrouve en appliquant les patrons appris à partir des occurrences de relations de l'ensemble d'apprentissage. Le corpus utilisé pour réaliser cette évaluation est le corpus TAC-KBP 2010 décrit ci-dessus. Précisons que l'utilisation de ce corpus pour évaluer l'extraction des relations n'empêche pas son utilisation pour l'apprentissage des patrons, les relations étant différentes pour les deux tâches.

Nous indiquons dans le tableau 2 le nombre de relations de l'ensemble d'apprentissage et de l'ensemble de test respectivement dans les colonnes *Nb. Appr* et *Nb. Test*. Le nombre de phrases trouvées contenant des occurrences des relations du corpus d'entraînement, et qui ont donc servi pour l'induction des patrons, est indiqué dans la colonne *Nb. Induc*. Le nombre de patrons générés à partir de ces phrases candidates est indiqué dans la colonne *Nb. Patrons* de ce même tableau.

Par exemple, pour le type de relation *org:alternate_names*, à partir des 20 013 relations de l'ensemble d'apprentissage, seules 214 phrases candidates contenant l'expression d'une de ces relations sont sélectionnées. Ces 214 phrases servent à générer 6 007 patrons, qui ont une couverture de 66,10% (*i.e.* on retrouve 66,10% des phrases contenant des occurrences des 10 006 relations de test). L'écart conséquent entre les 20 013 relations et les 214 phrases trouvées est dû à deux facteurs :

– une contrainte réductrice imposée lors de la sélection des phrases candidates. Seules les phrases dont tous les mots des entités nommées sont correctement identifiés sont en effet conservées. Or, les entités peuvent être partiellement (ou mal) reconnues lors des traitements linguistiques ;
– la nature des documents du corpus : 72% des documents sont des articles de presse édités entre janvier 2008 et août 2009, ce qui explique le peu de documents, voir aucun, concernant certaines personnes ou organisations pourtant présentes dans la KB.

Les résultats de la couverture des patrons sont présentés dans le tableau 2 pour chaque type de relations dans la colonne *Couv. Patrons*. À titre indicatif, le temps d'induction des patrons pour le type de relations *per:country_of_birth* (11 192 phrases exemples à comparer) passe de 690mn et 5s pour la version sans filtrage à 0mn et 30s pour la version avec filtrage[6], ce qui illustre l'intérêt de celui-ci en termes de temps de calcul.

3.3 Évaluation de l'extraction des relations

L'extraction des relations comprenant plusieurs étapes, chacune d'entre elles peut influer sur le résultat global. Nous proposons donc de faire une évaluation séparée de la recherche des phrases candidates et de l'extraction des relations proprement dite.

3.3.1 Recherche des phrases candidates

Une condition nécessaire pour des extraire relations pertinentes est de s'assurer que le moteur de recherche renvoie suffisamment de documents pertinents pour nous permettre de retrouver des entités cibles. Nous avons donc mesuré la couverture en documents de notre recherche de phrases candidates, à savoir le pourcentage de documents renvoyés par l'index que l'on retrouve effectivement dans la référence. Nous avons testé de ce point de vue différentes stratégies en faisant varier des paramètres comme le nombre de résultats retournés et l'utilisation ou non de l'expansion pour la requête. Les résultats de cette évaluation nous ont ainsi conduit à utiliser les entités sources et leurs formes étendues pour interrogation de l'index et prendre en compte les 1000 premiers résultats retournés : ces paramètres permettent de retrouver 84,24% des documents de référence. Le résultat détaillé par type de relations est donné par la colonne *Couv. Doc* du tableau 2.

À partir des documents ainsi sélectionnés, les phrases candidates à l'extraction d'une relation pour un type donné

[6]La version avec filtrage étant parallélisée, le temps donné est une somme des temps comptabilisés au niveau de chaque processeur.

sont extraites en retenant les phrases contenant à la fois l'entité source et le type de l'entité cible. La qualité et la quantité des phrases candidates sont largement influencées par la qualité de la reconnaissance des entités nommées. Comme nous ne disposons pas d'annotation de référence pour les entités nommées du corpus, il n'est pas possible de mesurer les pertes causées par la mauvaise reconnaissance des entités. En revanche, nous avons évalué la proportion de documents de référence dans lesquels nous retrouvons des phrases candidates. Cette donnée permet de fixer une borne maximale pour le pourcentage de relations qu'il serait possible d'extraire si les étapes à la suite se déroulaient idéalement. Nous obtenons au total une couverture de 37,55% des phrases appartenant aux documents de référence. Le détail par type de relations est présenté à la colonne *Couv. Rel* du tableau 2.

3.3.2 Extraction de relations

Pour évaluer les relations extraites, nous avons réutilisé les mesures et les outils d'évaluation fournis par la campagne TAC-KBP[7] sans nous limiter aux seuls documents présents dans la référence pour accepter une relation correcte[8]. Le tableau 3 fournit les résultats de cette évaluation en agglomérant tous les types de relations et en caractérisant l'impact du filtrage *a posteriori* des entités cibles sur les relations extraites en termes de rappel *(R.)*, précision *(P.)* et f1-mesure *(F1.)*. Pour mémoire, ce filtrage consiste à s'assurer que l'entité cible valide des expressions régulières et/ou une liste fermée de valeurs. Nous indiquons dans la colonne *Type de cible* du tableau 2 le mécanisme utilisé pour chaque type de relations.

Les résultats du tableau 3 montrent d'une part, que ce filtrage améliore les performances (en moyenne +2,74% de f1-mesure) et d'autre part, valident l'hypothèse que les patrons induits à partir de l'APSS sont aussi pertinents que ceux induits en considérant tous les exemples de relations deux à deux (dans ce cas, il y a même une amélioration de +1,72% de la f1-mesure en moyenne).

	Avant filtrage			Après filtrage		
	R. (%)	P. (%)	F1. (%)	R. (%)	P. (%)	F1. (%)
Tous les couples d'entités	16,28	11,20	13,26	18,07	13,66	15,56
APSS	16,90	12,76	14,54	18,67	16,87	17,72

TAB. 3 – Évaluation de l'impact du filtrage des réponses

Le tableau 4 présente les résultats de différents systèmes sur deux corpus très similaires, les corpus KBP 2009 et KBP 2010, ce dernier ajoutant au premier des documents Web et des transcriptions, *a priori* plus difficiles. Bien que ces chiffres ne portent que sur les relations effectivement présentes dans le corpus, ils intègrent la contrainte pour les systèmes ayant participé à la tâche *Slot Filling* de devoir décider si la relation existe ou non dans le corpus, ce que notre système, développé en dehors du contexte de ces campagnes, ne fait pas. Dans ce tableau, les colonnes *2009* et *2010* désignent les scores des trois systèmes les plus et les moins performants de KBP 2009 et 2010. Ji *et al.* (2010) ont montré que sur 492 relations de référence, 60,4% se trouvaient dans la même phrase tandis que les 39,6% restantes dépassaient l'espace phrastique dans leur expression et nécessitaient pour leur extraction la résolution de coréférences ou l'application de mécanismes d'inférence impliquant par exemple la composition de plusieurs relations ou l'utilisation de connaissances *a priori* sur les types de relations. De ce fait, nous avons distingué dans la colonne *2010 (a)* du tableau 4 les scores des systèmes qui nous sont les plus directement comparables, c'est-à-dire ceux se limitant à l'extraction de relations au niveau phrastique.

On peut noter que le meilleur système de KBP 2010 (Chada *et al.*, 2010) se détache très nettement : +36,63% par rapport au deuxième et +4,68% par rapport à un annotateur humain. Cette prédominance s'appuie à la fois sur l'utilisation d'un corpus annoté manuellement (différent du corpus KBP) de 3 millions de documents et la présence de plusieurs mécanismes d'extraction de relations au niveau inter-phrastique : coréférence pronominale, métonymie entre entités, résolution de dépendances sémantiques entre les mots et les entités, etc. L'utilisation du corpus supplémentaire semble être l'élément déterminant par rapport aux systèmes venant à la suite immédiate, ceux-ci se distinguant de systèmes plus médians par la prise en compte des relations inter-phrastiques. Les plus mauvais résultats, plus faibles en 2010, sont dûs pour une bonne part à des systèmes en cours de développement.

[7] `http://nlp.cs.qc.cuny.edu/kbp/2010/scoring.html`

[8] La référence du point de vue des documents n'étant constituée qu'à partir des résultats des participants à l'évaluation TAC-KBP, elle n'est en effet pas complète.

Concernant notre système, le tableau 4 permet de situer nos résultats dans la moyenne des résultats obtenus par les participants de l'évaluation KBP 2010 et parmi les trois premiers systèmes pour les approches faisant de l'extraction de relations au niveau de la phrase. Dans ce dernier cas, l'approche la plus performante (29,15% de f1-mesure) (Byrne & Dunnion, 2010) utilise des règles construites manuellement permettant d'atteindre un score de précision (66,55%) équivalent au meilleur score de la campagne (66,80%) et un score de rappel (18,67%) se situant dans la moyenne de la campagne (15,33%). Ce fort déséquilibre entre précision et rappel est d'ailleurs assez symptomatique des approches manuelles.

Systèmes TAC KBP	2009	2010	2010 (a)
Nb. soumissions (N) / participants	N=16 / 8	N=31 / 15	N=18
Annotateur humain	58,99%	61,10%	61,10%
1^{er} score	34,35%	65,78%	29,15%
$2^{ème}$ score	25,05%	29,15%	14,22%
$3^{ème}$ score	18%	28,29%	14,13%
(N-2)^{ème} score	5,90%	0,55%	0,55%
(N-1)^{ème} score	2,60%	0,19%	0,19%
N^{ème} score	1,75%	0,08%	0,08%
Notre système	–	17,72%	17,72%
Moyenne	13,43%	17,49%	9,71%
Médiane	13,93%	14,13%	12,27%

TAB. 4 – Résultats sur les données TAC-KBP (f1-mesure)

4 Travaux associés

L'extraction de relations à large échelle, au sens où nous l'avons définie à la section 2, est une problématique encore récente. Néanmoins, au travers notamment des évaluations TAC-KBP, elle a été l'objet d'un certain nombre de travaux proposant différentes approches. Concernant spécifiquement l'extraction des relations, les travaux se répartissent entre l'utilisation de l'apprentissage statistique (Agirre *et al.*, 2009; Li *et al.*, 2009b; Chen *et al.*, 2010b), l'induction de patrons lexicaux (Li *et al.*, 2009a; de Pablo-Sánchez *et al.*, 2009; McNamee *et al.*, 2009; Chen *et al.*, 2010b) et enfin, l'adaptation de systèmes existants pour la détection de relations (Schone *et al.*, 2009; Bikel *et al.*, 2009). On note pour KBP 2010 l'introduction d'approches à base de règles, par exemple (Byrne & Dunnion, 2010), et d'approches reposant sur le principe de « Distant supervision » à partir de classifieurs, dont celle de (Surdeanu *et al.*, 2010). Notre approche relève de l'induction de patrons lexicaux et fait l'hypothèse, comme (Mintz *et al.*, 2009), que la seule présence d'un couple d'entités dans une phrase est suffisante pour marquer la présence effective d'une relation entre ces entités. Ce n'est cependant pas toujours le cas et nous pensons ainsi qu'il est important de filtrer en amont les exemples utilisés pour l'induction des patrons, à l'instar de ce que propose (Riedel *et al.*, 2010).

Comme notre système, ceux élaborés pour KBP 2009 n'exploitent pas les liens de dépendance entre les types de relations, à l'image du lien entre la date de naissance et l'âge par exemple. Dans (Chen *et al.*, 2010a), les auteurs montrent que les résultats obtenus dans (Li *et al.*, 2009a) (31,96% de f1-mesure) peuvent être améliorés (ils obtiennent 34,81% de f1-mesure) par l'intégration des dépendances entre les relations en utilisant des règles d'inférence fondées sur une extension de la logique du premier ordre. Plus généralement, Chada *et al.* (2010) ont montré dans le cadre de KBP 2010 une augmentation très significative des performances en intégrant des mécanismes permettant d'extraire des relations au-delà de la phrase. Sur un autre plan, Li *et al.* (2009a) se distinguent dans KBP 2009 en utilisant deux étapes d'extraction de relations : la première vise à retrouver dans les documents du corpus des entités cibles potentielles en utilisant des patrons de relations ; la seconde applique le même processus à une version récente de Wikipédia pour trouver des entités cibles potentielles supplémentaires qui n'auraient pas été identifiées lors de la première étape. Les entités cibles ainsi acquises ne sont finalement conservées que si elles sont retrouvées dans un document du corpus. Cette récupération d'entités améliore les performances de façon significative (+9% de f1-mesure par rapport à (Bikel *et al.*, 2009)) mais ajoute l'utilisation d'un corpus externe que l'on peut considérer comme trop lié à la KB. Les résultats sur KBP 2010 ont d'ailleurs montré que

les performances globales pouvaient être améliorées sans cette ressource supplémentaire et que son impact sur les résultats est plus limité que pour KBP 2009 (une baisse des résultats a même été observée).

5 Conclusion et perspectives

Dans cet article, nous avons présenté un système d'extraction d'information à large échelle permettant d'extraire des relations de type attributive entre entités nommées. Le qualificatif « à large échelle » recouvre à la fois la prise en compte d'un grand nombre de types de relations et la recherche de ces relations dans un large corpus. Ce système se fonde sur une approche faiblement supervisée dans laquelle les exemples se limitent à des couples d'entités en relation. L'extraction des relations s'effectue par l'application de patrons lexico-syntaxiques caractéristiques des types de relations considérés et appris à partir de phrases issues de la projection des couples d'entités exemples dans un corpus. Nous avons évalué les résultats de cette approche en utilisant le cadre d'évaluation offert par la tâche *Slot Filling* de l'évaluation KBP en nous concentrant sur la problématique de l'extraction des relations proprement dite, sans nous attacher à la détection de l'absence d'une relation dans un corpus. Les résultats obtenus dans ce contexte se situent dans la moyenne des résultats obtenus par les participants de l'édition 2010, ce que nous pouvons considérer comme un point de départ intéressant dans la mesure où notre système repose sur une approche volontairement générique et n'exploite que très faiblement les spécificités des types de relations traités. Nous avons aussi pu montrer que des techniques permettant de prendre en compte certains aspects d'un passage à une « large échelle », comme le filtrage des couples de phrases exemples à généraliser par l'utilisation de l'APSS, ne dégradent pas les performances et peuvent même contribuer à les améliorer.

Nous travaillons par ailleurs sur l'amélioration de notre système en conservant l'idée de garder une certaine généricité par rapport au type des relations considérées. Pour ce faire, nous nous focalisons particulièrement sur l'apprentissage des patrons d'extraction. Un premier pas dans cette direction vise à disposer à la fois d'un nombre plus important d'exemples mais également d'exemples de meilleure qualité. Ces deux points sont liés dans la mesure où l'obtention d'un ensemble plus large d'exemples passe par le relâchement des contraintes touchant la sélection des phrases exemples. Or, si l'on peut espérer qu'un tel relâchement permettra l'obtention de nouveaux bons exemples, il sera aussi source de nouveaux mauvais exemples. Nous souhaitons donc coupler un tel relâchement avec l'utilisation d'un module de filtrage de relations qui, à l'instar de (Banko & Etzioni, 2008), est capable de déterminer si une phrase contient une relation entre deux entités sans *a priori* sur la nature de cette relation.

Références

AGICHTEIN E. & GRAVANO L. (2000). Snowball : Extracting relations from large plain-text collections. In *5^{th} ACM International Conference on Digital Libraries*, p. 85–94, San Antonio, Texas, USA.

AGIRRE E., CHANG A., JURAFSKY D., MANNING C., SPITKOVSKY V. & YEH E. (2009). Stanford-UBC at TAC-KBP. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

BANKO M. & ETZIONI O. (2008). The tradeoffs between open and traditional relation extraction. In *ACL-08 : HLT*, p. 28–36, Columbus, Ohio.

BAYARDO R., MA Y. & SRIKANT R. (2007). Scaling up all pairs similarity search. In *16^{th} International Conference on World Wide Web (WWW'07)*, p. 131–140, Banff, Alberta, Canada.

BIKEL D., CASTELLI V., RADU F. & JUNG HAN D. (2009). Entity Linking and Slot Filling through Statistical Processing and Inference Rules. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

BIZER C., LEHMANN J., KOBILAROV G., AUER S., BECKER C., CYGANIAK R. & HELLMANN S. (2009). DBpedia - A crystallization point for the Web of Data. *Journal of Web Semantics*, **7**, 154–165.

BYRNE L. & DUNNION J. (2010). UCD IIRG at TAC 2010 KBP Slot Filling Task. In *Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

CHADA D., ARANHA C. & MONTE C. (2010). An Analysis of The Cortex Method at TAC 2010 KBP Slot-Filling. In *Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

CHEN Z., TAMANG S., LEE A., LI X., PASSANTINO M. & JI H. (2010a). Top-down and Bottom-up : A Combined Approach to Slot Filling. In *6th Asia Information Retrieval Symposium on Information Retrieval Technology*, Gaithersburg, Maryland, USA : Springer-Verlag.

CHEN Z., TAMANG S., LEE A., LI X., SNOVER M., PASSANTINO M., LIN W.-P. & JI H. (2010b). CUNY-BLENDER TAC-KBP2010 Slot Filling System Description. In *Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

CLAVEAU V. & SÉBILLOT P. (2004). From efficiency to portability : acquisition of semantic relations by semi-supervised machine learning. In *20th International Conference on Computational Linguistics (COLING 2004)*, p. 261–267, Geneva, Switzerland.

DE PABLO-SÁNCHEZ C., PEREA J., SEGURA-BEDMAR I. & MARTÍNEZ P. (2009). The UC3M team at the Knowledge Base Population task. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

EMBAREK M. & FERRET O. (2008). Learning patterns for building resources about semantic relations in the medical domain. In *6th Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

GIONIS A., INDYK P. & MOTWANI R. (1999). Similarity search in high dimensions via hashing. In *25th International Conference on Very Large Data Bases (VLDB'99)*, p. 518–529, Edinburgh, Scotland, UK.

HEARST M. (1992). Automatic acquisition of hyponyms from large text corpora. In *14th International Conference on Computational linguistics (COLING'92)*, p. 539–545, Nantes, France.

JI H., GRISHMAN R. & TRANG DANG H. (2010). Overview of the TAC 2010 Knowledge Base Population Track. In *Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

LI F., ZHENG Z., BU F., TANG Y., ZHU X. & HUANG M. (2009a). THU QUANTA at TAC 2009 KBP and RTE Track. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

LI S., GAO S., ZHANG Z., LI X., GUAN J., XU W. & GUO J. (2009b). PRIS at TAC 2009 : Experiments in KBP Track. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

MCNAMEE P., DREDZE M., GERBER A., GARERA N., FININ T., MAYFIELD J., PIATKO C., RAO D., YAROWSKY D. & DREYER M. (2009). HLTCOE Approaches to Knowledge Base Population at TAC 2009. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

MINTZ M., BILLS S., SNOW R. & JURAFSKY D. (2009). Distant supervision for relation extraction without labeled data. In *ACL-IJCNLP'09*, p. 1003–1011, Suntec, Singapore.

PANTEL P., RAVICHANDRAN D. & HOVY E. (2004). Towards terascale knowledge acquisition. In *20th International Conference on Computational Linguistics (COLING'04)*, p. 771–777, Geneva, Switzerland.

RAVICHANDRAN D. (2005). *Terascale knowledge acquisition*. PhD thesis, Faculty of the Graduate School University of Southern California, Los Angeles, CA, USA.

RIEDEL S., YAO L. & MCCALLUM A. (2010). Modeling relations and their mentions without labeled text. In J. BALCÁZAR, F. BONCHI, A. GIONIS & M. SEBAG, Eds., *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *Lecture Notes in Computer Science*, p. 148–163. Springer Berlin / Heidelberg.

RUIZ-CASADO M., ALFONSECA E. & CASTELLS P. (2007). Automatising the learning of lexical patterns : An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. *Data Knowledge Engineering*, **61**, 484–499.

SCHLAEFER N., GIESELMANN P., SCHAAF T. & WAIBEL A. (2006). A pattern learning approach to question answering within the ephyra framework. In P. SOJKA, I. KOPECEK & K. PALA, Eds., *Text, Speech and Dialogue*, volume 4188 of *Lecture Notes in Computer Science*, p. 687–694. Springer Berlin / Heidelberg.

SCHONE P., GOLDSCHEN A., LANGLEY C., LEWIS S., ONYSHKEVYCH B., CUTTS R., DAWSON B., MACBRIDE J., MATRANGOLA G., MCDONOUGH C., PFEIFER C. & URSIAK M. (2009). TCAR at TAC-KBP 2009. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

SHINYAMA Y. & SEKINE S. (2006). Preemptive information extraction using unrestricted relation discovery. In *HLT-NAACL 2006*, p. 304–311, New York City, USA.

SURDEANU M., MCCLOSKY D., TIBSHIRANI J., BAUER J., CHANG A., SPITKOVSKY V. & MANNING C. (2010). A Simple Distant Supervision Approach for the TAC-KBP Slot Filling Task. In *Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

TAC-KBP (2010). Preliminary task description for knowledge-base population at TAC 2010.

ZHOU G., SU J., ZHANG J. & ZHANG M. (2005). Exploring various knowledge in relation extraction. In *43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, p. 427–434, Ann Arbor, USA.

ZHOU G., ZHANG M., JI D. & ZHU Q. (2007). Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *EMNLP - CoNLL'07*, p. 728–736, Prague, Czech Republic.

RECITAL 2011, Montpellier, 27 juin – 1er juillet 2011

Analyse de l'ambiguïté des requêtes utilisateurs par catégorisation thématique

Fanny Lalleman[1, 2]
(1) CLLE & CNRS, 5, allées Antonio Machado 31058 Toulouse Cedex 9
(2) Orange Labs, 2, Avenue Pierre Marzin 22307 Lannion Cedex
fanny.lalleman@univ-tlse2.fr

Résumé. Dans cet article, nous cherchons à identifier la nature de l'ambiguïté des requêtes utilisateurs issues d'un moteur de recherche dédié à l'actualité, 2424actu.fr, en utilisant une tâche de catégorisation. Dans un premier temps, nous verrons les différentes formes de l'ambiguïté des requêtes déjà décrites dans les travaux de TAL. Nous confrontons la vision lexicographique de l'ambiguïté à celle décrite par les techniques de classification appliquées à la recherche d'information. Dans un deuxième temps, nous appliquons une méthode de catégorisation thématique afin d'explorer l'ambiguïté des requêtes, celle-ci nous permet de conduire une analyse sémantique de ces requêtes, en intégrant la dimension temporelle propre au contexte des news. Nous proposons une typologie des phénomènes d'ambiguïté basée sur notre analyse sémantique. Enfin, nous comparons l'exploration par catégorisation à une ressource comme Wikipédia, montrant concrètement les divergences des deux approches.

Abstract. In this paper, we try to identify the nature of ambiguity of user queries from a search engine dedicated to news, 2424actu.fr, using a categorization task. At first, we see different forms of ambiguity queries already described in the works of NLP. We confront lexicographical vision of the ambiguity to that described by classification techniques applied to information retrieval. In a second step, we apply a method of categorizing themes to explore the ambiguity of queries, it allow us to conduct a semantic analysis of these applications by integrating temporal context-specific news. We propose a typology of phenomena of ambiguity based on our semantic analysis. Finally, we compare the exploration by categorization with a resource as Wikipedia, showing concretely the differences between these two approaches.

Mots-clés : recherche d'information, ambiguïté, classification de requêtes.

Keywords: Information retrieval, ambiguity, classification queries.

1 Introduction

La désambiguïsation lexicale a été appliquée à la recherche d'information avec un succès variable. Le précurseur sur cette thématique (Weiss, 1973) a été suivi par un certain nombre de travaux (Krovetz & Croft, 1992; Stokoe *et al.*, 2003; Sanderson, 2000), où la désambiguïsation était focalisée sur les mots dont les sens étaient répertoriés dans des dictionnaires (Voorhees, 1993). L'enjeu était de réperer de la polysémie, c'est-à-dire différents sens pour une même forme. Les performances des systèmes n'ont pas été à la hauteur des espérances aboutissant à un questionnement sur l'adéquation des dictionnaires pour une telle tâche, et la question de la nature de l'ambiguïté. Les dictionnaires sont peu adaptés au traitement des requêtes, contenant peu de noms propres, ou de termes complexes et la nature de l'ambiguïté des requêtes se limite aux relations sémantiques contenues dans les dictionnaires (homonymie et polysémie). Ces difficultés ont rendu le traitement de l'ambiguïté secondaire en recherche d'information. Mais le développement du web et des moteurs de recherche grand public ont remis en avant la question de l'ambiguïté des requêtes, en particulier le CLIR (Cross-Language Information Retrieval) (Darwish & Oard, 2003). En effet, la traduction est très dépendante de la désambiguïsation, d'autant plus quand le contexte est absent. La taille des requêtes entre également en jeu dans les performances des systèmes de désambiguïsation. Ils se trouvent être relativement performants sur les requêtes dites « longues » provenant des campagnes d'évaluation comme TREC ou CLEF, mais en grande difficulté face des requêtes « courtes » provenant de moteurs de recherche. Actuellement, l'accent est mis sur les données réelles et l'intéractivité avec l'utilisateur. L'enjeu n'est plus de construire un système de désambiguïsation, mais de savoir identifier une requête ambigüe soit pour pouvoir proposer par exemple une présentation adaptée à l'utilisateur, comme une présentation hiérarchique (Hearst, 2009), soit pour améliorer les performances du système en lui-même. Les solutions actuelles préfèrent en présence de données réelles, se servir des logs de requêtes contenant les pages réponses choisies par les utilisateurs pour identifier les requêtes ambigües (Clough *et al.*, 2009). Nous nous situons dans un contexte applicatif spécifique, un moteur de recherche dédié à l'actualité, où l'ambiguïté des requêtes doit être abordée différemment. La base documentaire évolue avec l'actualité et ne conserve pas les documents anciens, il faut donc envisager une méthode différente qui ne se base pas sur les choix des utilisateurs pour essayer d'identifier l'ambiguïté.

Dans ce contexte, nous cherchons à identifier la nature de l'ambiguïté présente dans nos requêtes en les examinant grâce à une tâche de catégorisation. Cette approche doit nous permettre d'étudier la forme que peut prendre l'ambiguïté des requêtes produites dans un contexte applicatif. Dans un premier temps, nous allons passer en revue les différentes formes de l'ambiguïté déjà décrites dans les travaux de TAL, en confrontant la vision « traditionnelle » de l'ambiguïté des requêtes, lexicographique, à une vision produite par des techniques de classification et de clustering appliquées à la recherche d'information. Dans un deuxième temps, nous appliquerons ces techniques de classification pour explorer l'ambiguïté des requêtes dans notre contexte applicatif 2424actu.fr afin d'effectuer une analyse sémantique des requêtes. Cette analyse débouchera sur une typologie de l'ambiguïté dans ce cadre applicatif. Enfin, nous confronterons l'exploration par catégorisation à une ressource comme Wikipédia, en montrant les divergences de ces deux approches et l'importance d'une démarche d'analyse reflétant les particularités du contexte et de la base de texte à traiter.

2 L'ambiguïté dans les requêtes utilisateurs

2.1 L'ambiguïté des requêtes vue à travers les dictionnaires

L'ambiguïté des requêtes est souvent analysée à la lumière des dictionnaires et autres ressources (WordNet, Wikipédia) (Sanderson, 2000, 2008; Santamaría *et al.*, 2010). Le principe est de rechercher les termes des requêtes dans divers dictionnaires afin de répérer les mots ayant plusieurs sens. Comme en désambiguisation classique, WordNet est la principale source de comparaison, mais il lui est reproché d'être peu représentatif de la diversité présente dans les requêtes des utilisateurs de moteur de recherche. Il est donc utilisé en complément d'une autre ressource, en l'occurrence Wikipédia (Santamaría *et al.*, 2010; Sanderson, 2008; Clough *et al.*, 2009). Par exemple, (Sanderson, 2008) propose d'examiner l'ambiguïté en s'appuyant conjointement sur WordNet et sur les pages de Wikipédia, il collecte dans cette ressource les pages identifiant une forme ambigüe et les différents sujets qui portent le nom de cette forme. Grâce à ces deux types de ressources, Sanderson évalue l'ambiguïté présente dans deux corpus de requêtes provenant de deux moteurs de recherche (1 million de requêtes Live Search et 500 000 requêtes de UK's Press Association). 16% des requêtes Live Search sont estimées ambigües à la lu-

mière des ressources, et 23,6% pour les requêtes du moteur UK's Press Association. Il confirme l'importance du phénomène et suggère que l'ambiguïté varie selon les requêtes étudiées. Par ailleurs, on peut se demander si l'ambiguïté présente en recherche d'information est de la polysémie décrite dans les dictionnaires (*hôte* peut désigner la personne qui reçoit ou celle qui est reçue) ou bien de l'homonymie comme Wikipédia le propose (*éruption* comme nom ou *Eruption* groupe de musique).

D'autre part, il y a deux problèmes liés à cette approche de l'ambiguïté des requêtes. Le premier problème est lié aux ressources. Wikipédia ne couvre que 60% des sens présents dans un moteur de recherche (en anglais) et WordNet seulement 32 %(Santamaría *et al.*, 2010). Le deuxième problème est l'absence de prise en compte de l'environnement spécifique à la tâche de recherche d'information. En effet, l'ambiguïté d'une requête ne peut être envisagée en dehors de la base textuelle interrogée, l'utilisateur cherchant une information présente dans la base qu'il consulte et qu'il anticipe. L'ambiguïté typée dans des ressources comme des dictionnaires et des thésaurus généralistes n'est donc pas forcément adaptée pour caractériser l'ambiguïté effective des requêtes utilisateurs pour une application donnée.

2.2 L'ambiguïté « non classique »

Rompant avec la vision lexicographique de l'ambiguïté des requêtes, certains travaux proposent d'analyser d'autres formes d'ambiguïté. Ils avancent que des requêtes peuvent être ambigues si elles renvoient à plusieurs « sous-domaines » (Zhai *et al.*, 2003) ou à plusieurs « facettes »(Hearst, 2006). Ce type de distinction est à mettre en rapport avec l'utilisation de techniques de classification et de clustering en recherche d'information. Ces techniques permettent de faire apparaître des rapprochements pertinents sans poser d'a priori sur ce qui est recherché. On retrouve ce type de distinction vis à vis de l'ambiguïté des requêtes dans (Song *et al.*, 2009). Ils utilisent une typologie à trois éléments qu'ils ont définie dans le but d'annoter manuellement des requêtes :

1. Type A (requête ambigüe) : requête qui a plus d'un seul sens : « giant » > plusieurs référents Giant Company, Giant (film), San Francisco Giant (équipe de basket ball)
2. Type B (requête « large ») : requête qui couvre plusieurs sujets ou thématiques : « songs » > « song lyrics », « love songs », « download songs »
3. Type C (requête non ambigüe) : requête qui a un sens spécifique et un référent facilement identifiable : « Billie Holiday » (chanteuse jazz)

Les critères ont utilisé pour discriminer les requêtes de type A et C ne sont pas clairement établis. En effet, potentiellement l'exemple donné comme requête non ambigüe, « Billie Holiday » peut être discuté puisqu'il s'avère qu'il existe un album eponyme de Billie Holiday. La distinction entre « ambigüe » et « non ambigu » est donc difficile à établir, laissant apparaître des requêtes « entre les deux ». (Song *et al.*, 2009) les distinguent en créant une catégorie de requêtes contenant un terme générique (requête B). Le phénomène de « généralité » existe déjà dans les conceptions de l'ambiguïté sémantique (Aarts & MacMahon, 2006). Il décrit un phénomène d'ambiguïté différent de la polysémie et de l'homonymie dont le nom varie selon les auteurs : sous-spécification, sens vague, indétermination ou généralité. Cette forme d'ambiguïté décrit un sens général ou inclusif qui peut avoir différentes significations selon le contexte, par exemple « vache » va pourvoir prendre un sens général « animal ruminant » ou un sens spécifique « femelle du taureau ». Il se pose donc la question de la présence de ce type d'ambiguïté « non classique » en recherche d'information.

2.3 La question des entités nommées « ambigües »

L'utilisation de Wikipédia comme ressource pour capter l'ambiguïté des requêtes ou bien désambiguïser (Rahurkar *et al.*, 2008) est principalement due à la possibilité d'exploiter les entités nommées (ENs) qu'elle contient. Or, l'étude des requêtes utilisateurs de AllTheWeb et Altavista avaient mis en évidence que 11-17% des requêtes étaient composées d'un nom propre désignant une personne (Spink *et al.*, 2004). Les entités nommées désignant des lieux sont aussi très présentes. Gan et al. (Gan *et al.*, 2008) montrent dans une étude des logs de requêtes de AOL (2006) qu'il y a environ 38% des requêtes qui contiennent des termes de type « géographique » comme « New York » ou « Kentucky Fried Chicken ». Pour des moteurs de même type comme AOL et Altavista (moteurs généralistes), on arriverait environ à plus de 50% d'entités nommées. L'utilisation d'entitées nommées en nombre par les utilisateurs de moteurs de recherche renforce les phénomènes d'ambiguïté des requêtes, à tel point que

cela a justifié une tâche dédiée à la désambiguïsation des noms de personnes dans un contexte de recherche d'information nommée WePS (Artiles *et al.*, 2007, 2010). En effet, il est courant qu'un même nom propre désigne un grand nombre de personnes. (Artiles *et al.*, 2007) donnaient l'exemple du bureau de recensement américain qui rapporte 90 000 noms portés par 100 millions de personnes. L'homonymie n'est pas la seule cause d'ambiguïté des ENs. Elles forment un ensemble hétérogène composé d'une collection d'expressions linguistiques diverses, réunies sur la base de caractéristiques référentielles communes (Erhmann, 2008), c'est-à-dire que sont rassemblées sous la même étiquette différentes expressions comme les noms de lieux, de personnes ou des dénominations d'organisation, qui par nature, ne manifestent pas les mêmes propriétés. Par conséquent, l'ambiguïté apparaît sous différentes formes : métonymie (« la France a gagné 3-0 »), facettes référentielles (Carla Bruni, chanteuse, ancienne mannequin, épouse du président de la République) ou comme on l'a vu homonymie. Face à une telle variabilité, l'utilisation de ressources pour traiter l'ambiguïté des Entités Nommées a des limites. C'est pourquoi les techniques de clustering se sont développées pour tenter de désambiguïser les ENs (Santamaría *et al.*, 2010; Bernardini *et al.*, 2009).

2.4 La question de l'évaluation de l'ambiguïté

Actuellement, la question de la forme que prend l'ambiguïté des requêtes utilisateurs reste ouverte. En effet, considérer uniquement l'ambiguïté à travers des ressources semble non satisfaisant (Hearst, 2006, 2009). L'utilisation de méthodes inductives reste dans les travaux récents un complément pour augmenter la couverture de Wikipédia. En effet, la question de l'évaluation de l'ambiguïté des requêtes reste entière. D'une part, l'accès aux réponses choisies par l'utilisateur à une requête donnée est difficile, ce sont des informations périssables et détenues par les moteurs de recherche. D'autre part, il n'y a que très peu de tests d'évaluation et il ne sont pas conçus pour des requêtes « réelles » (TREC 7). Cependant, le repérage et l'évaluation de l'ambiguïté ne peuvent être effectués sans savoir au préalable quelle forme prend le phénomène que l'on cherche à contrôler. Il est donc nécessaire de mieux comprendre comment se manifeste l'ambiguïté, en tentant de voir si les ambiguïtés dites non classiques sont bien présentes. Cette tâche implique de prendre en compte un certain nombre de paramètres variables et subjectifs comme les connaissances de l'utilisateur, son intention lorsqu'il effectue une requête. Il faut également considérer le cadre dans lequel il effectue cette recherche, le type de moteur et la base documentaire associée, un moteur généraliste ne semble pas générer le même type d'ambiguïté qu'un moteur dit vertical, dédié à la recherche d'information sur un domaine particulier comme la vidéo ou les publications scientifiques (Sanderson, 2008).

3 Faire émerger l'ambiguïté des requêtes automatiquement à partir de corpus

Le cadre applicatif de ce travail est un plateforme d'actualités vidéos développée par Orange Labs, 2424 actu.fr [1]. Ce site permet de consulter l'actualité française en temps réel, et propose différents types d'accès à l'information : par clustering, par thématiques et par barre de recherche traditionnelle. Le moteur de recherche à la disposition des utilisateurs ne propose pas de présentation de résultats par clusters ou par thématiques, mais par format des sources (vidéo, texte, audio).

La question de l'ambiguïté des requêtes dans un cadre applicatif est une question importante pour plusieurs secteurs de la recherche d'information comme la complexité des requêtes et la prédiction de la difficulté d'une requête (Clough *et al.*, 2009). L'ambiguïté est également un aspect important pour la personnalisation des moteurs de recherche vis à vis des utilisateurs. Notre contexte renforce cet intérêt, l'actualité est un usage quotidien pour beaucoup d'utilisateurs, mais elle a la particularité d'être évolutive. Nous connaissons également des contraintes propres au contexte applicatif [2] qui demandent d'aborder l'ambiguïté des requêtes sous un angle différent. Dans ce but, nous cherchons à identifier l'ambiguïté à partir d'une tâche de catégorisation.

1. http ://www.2424actu.fr

2. Le site est une plateforme qui reccueille temporairement les productions des partenaires, elle ne possède donc pas les contenus et ne peut les conserver pour un usage commercial.

3.1 L'ambiguïté examinée par une tâche de catégorisation

Le but de cette expérimentation n'est pas de reproduire un processus de recherche d'information, mais d'étudier la forme que peut prendre l'ambiguïté des requêtes produites dans un contexte applicatif. L'utilisation d'une ressource nous paraît peu appropriée pour cette tâche. En effet, comme nous l'avons vu dans (2.2), il existe des ambiguïtés que l'on peut qualifier de non-classiques et qui ne sont pas recensées dans une ressource existante. L'enjeu est donc de faire émerger à la fois l'ambiguïté « classique » telle que la polysémie ou l'homonymie, et les ambiguïtés « non classiques » et de pouvoir observer leurs manifestations.

Cette étude préliminaire utilise les moyens disponibles dans le cadre de cette application, une double catégorisation des documents « cibles » des requêtes. Notre but est de parvenir à transposer une catégorisation externe à une catégorisation interne. Pour cela, nous nous servons de documents qui ont la particularité d'être classifiés à deux niveaux : chaque document (ou news) appartient à un cluster et à une catégorie thématique. Les catégories thématiques sont héritées de l'AFP et elles sont au nombre de six : *économie* (questions économiques), *international* (actualités hors de France), *société*, *politique*, *cultures* (musique, sciences, art, people) et *sport*. Elles forment un étiquetage que nous allons utilisés. L'hypothèse est que les catégories, en nombre réduit et correspondant à un classement adapté pour les news, vont donner une approximation des domaines présents dans l'actualité et un premier point d'entrée sur le comportement des requêtes. Nous nous appuyons donc sur le corpus que nous avons constitué au fil des jours à partir des actualités quotidiennes du site (tableau 1). Le corpus de documents est la collection de documents sous forme textuelle disponible pour les utilisateurs du site d'actualité pour une période donnée en français. Les sources sont hétérogènes : audio retranscrit, dépêches AFP, articles de journaux, retranscription de journaux télévisés. Les documents proviennent des différents partenaires du site (AFP, Le Monde, Le Point, L'Express, France Télévision, Paris Match, etc.). La notion temporelle structure le corpus, il est partitionné en huit sous-corpus.

L'expérimentation consiste à catégoriser les requêtes pour pouvoir observer leur distribution sur plusieurs domaines. Les requêtes d'une période temporelle sont projetées sur la base textuelle correspondant à la même période temporelle, et si elles apparaissent dans un document, elles héritent de la catégorie thématique du document. Les catégories sont pondérées selon la fréquence d'apparition. Nous effectuons également un filtrage des catégories attribuées : une catégorie est considérée seulement si elle représente plus de 10% des textes où la requête apparaît. Le filtre permet de limiter l'apparition de catégories résiduelles, il a été choisi de façon arbitraire. Le corpus de requêtes utilisé contient les requêtes utilisateurs du site 2424actu des huit derniers mois (de Mai 2010 à Décembre 2010). Il totalise 487 231 requêtes non dédoublonnées (pour environ 30 700 requêtes uniques). Ce corpus de requêtes est découpé en partition temporelle, chaque partition est organisée en fonction de la fréquence des requêtes [3]. Nous utiliserons dans cette expérimentation les 49 requêtes les plus fréquentes de chaque sous-partition du corpus c'est-à-dire ayant une fréquence supérieure à 100 pour la période considérée, ce qui fait au total 391 requêtes. Préalablement, les requêtes contenant plusieurs mots et ne formant pas des termes sont volontairement exclues de la catégorisation comme par exemple « boue hongrie », alors qu'une requête comme « festival de cannes » va être catégorisée. Le parti pris est que ces requêtes contenant plusieurs termes sont moins affectées par l'ambiguïté (Sanderson, 2008).

Mai2010	Juin2010	Juillet2010	Août2010	Sept2010	Octobre2010	Nov2010	Déc2010
23 521	26 782	15 773	19 543	17 634	22 822	16 015	11 096

TABLE 1 – Corpus de documents (news 2424actu)

3.2 Analyse de l'ambiguïté des requêtes

Nous cherchons à savoir si la catégorisation thématique est un bon procédé pour explorer la manifestation de l'ambiguïté et donc si une pluri-catégorisation est synonyme d'ambiguïté. Pour étudier cette question, nous allons procéder tout d'abord à une analyse des requêtes qui n'ont été rattachées qu'à une seule catégorie ce qui va permettre de regarder si l'ambiguïté est présente malgré un rattachement unique. Puis dans un deuxième temps, l'analyse se focalisera sur les requêtes rattachées plusieurs catégories thématiques.

3. A noter, que les sous-partitions pour les mois de septembre et novembre ne sont pas complètes.

Nous avons effectué une évaluation sur le corpus de requêtes les plus fréquentes (mai à décembre), soit 391 requêtes. Ce corpus contient 70% de requêtes contenant une EN ou étant une EN. Parmi ces 391 requêtes, 35% ne sont pas catégorisées, leur recherche ne donnant pas de résultats dans la base textuelle. La répartition entre les requêtes mono-catégorielles et pluri-catégorielles est la suivante : 54% sont mono-catégorielles et 46% pluri-catégorielles (au moins deux catégories).

3.2.1 Les requêtes rattachées à une seule catégorie

La répartition des requêtes qui donnent lieu à un classement thématique unique varie selon les corpus-tests (environ 67% pour le sous-corpus de décembre contre 27% dans le sous-corpus de mai). Ces requêtes sont à 80% des entitées nommées, ce qui ne représente pas l'ensemble des requêtes contenant des ENs (40% environ sont rattachées à plusieurs catégories). Par exemple, « miss france » va être catégorisée exclusivement en *cultures* tout comme les requêtes « prince william » ou « audrey pulvar », « nicolas dupont-aignan » sera catégorisé en *économie*. Ces requêtes contiennent des noms propres complets ce qui aide à l'identification, mais il y a également des requêtes mono-mots qui n'ont qu'une seule catégorie comme « vogica » en *international*. Les requêtes mono-catégorielles ne contenant pas d'entitée nommée comme « neiges » en *société* ou « agriculture » en *économie* sont moins nombreuses (environ 20%). Nous observons également qu'une grande partie de ces classements uniques s'opère dans les catégories *international* (environ 40% des requêtes mono-catégorisées) et *cultures* (environ 20%). Ces catégories sont vastes et englobent énormément de sujets.

3.2.2 Les requêtes rattachées à plusieurs catégories

Nous avons effectué une analyse basée sur les 115 requêtes pluri-catégorisées, nous cherchons à faire émerger des phénomènes identifiables afin de construire une typologie des requêtes manifestant de l'ambiguïté. Les requêtes renvoyant à plusieurs catégories thématiques sont majoritairement des entitées nommées comme pour les requêtes mono-catégorielles, mais on trouve également 40% de requêtes ne contenant pas d'EN.

Parmi les requêtes pluri-catégorielles, il existe des requêtes porteuses d'ambiguïté de type classique (polysémie et homonymie). Elles renvoient à plusieurs catégories qui présentent des pondérations similaires comme par exemple « royal » (tableau 2). C'est une requête ambigüe et la catégorisation fait ressortir deux aspects intéressants, d'une part une catégorisation en *politique* et d'autre part une catégorisation en *sport*. La catégorisation en *politique* de « royal » renvoie à « Ségolène Royal », contrairement à la catégorisation en *sport* qui renvoie l'adjectif « royal » présent dans un certain nombre de stades sud-africains utilisés lors de la Coupe du monde de football, par exemple « stade Royal Bafokeng de Rustenburg ». L'ambiguïté lexicale dans nos requêtes est essentiellement de l'homonymie. En effet, le seul exemple de polysémie trouvé dans notre corpus est la requête « tabac » (tableau 3). Elle peut désigner plusieurs choses (le lieu où l'on vend du tabac, le tabac comme plante, ou comme cigarettes) et c'est aussi un nom que l'on retrouve fréquemment dans l'expression « passer à tabac ». Le contexte d'actualité contient plusieurs sens de « tabac » : lieu de vente, tabac comme cause de maladie, cigarettes, tabac comme plante cultivée ou encore passage à tabac. La polysémie est extrêment rare dans notre corpus.

Thématiques	int	pol	spr
royal	27	**63**	**40**

TABLE 2 – La requête « royal » catégorisée en juin

Thématiques	int	eco	soc
tabac	11	**27**	**25**

TABLE 3 – La requête « tabac » au mois de novembre

Nous observons un autre type de requêtes pluri-catégorielles avec des fréquences d'apparition supérieures aux requêtes précédentes (requêtes homonymiques), comme par exemple « société » (tableau 4), « obama » ou « sarkozy ». Ces requêtes sont difficiles à interpréter même à l'aide du contexte des documents cibles parce qu'elles balaient un champ très large. Ainsi on voit par exemple qu'une requête comme « sarkozy » (tableau 5), va ouvrir vers un

grand nombre de catégories thématiques. Cette requête est ambigüe car elle a besoin d'être spécifiée, d'être complétée. L'observation ne permet pas d'identifier un lien interprétable en termes d'ambiguïtés entre la requête et une catégorie. Nous les rapprochons du type de requêtes dites « larges » (type B) décrites par (Song *et al.*, 2009), ce sont des requêtes génériques.

Thématiques	eco	int	pol	soc
société	**114**	**64**	**45**	**84**

TABLE 4 – La requête "société" (mai)

Thématiques	int	pol	soc
sarkozy	196	**544**	182

TABLE 5 – La requête « sarkozy » (juin)

Enfin, nous distinguons un troisième type de requêtes pluri-catégorielles, relativement proches des requêtes « génériques » du point de vue des catégorisations. Elles sont cependant différentes car elles sont porteuses d'ambiguïté référentielle. Par ambiguïté référentielle, nous désignons une requête qui n'a pas de référent fixe et qui potentiellement peut désigner deux ou plusieurs référents. Par exemple, « intempéries » paraît peu ambigu dans un contexte d'actualités, pourtant la catégorisation fait ressortir deux catégories (*international* et *société*). De nombreuses requêtes comme « otages » ou « ministre » (tableau 6) sont ambigües d'un point de vue référentiel, « ministre » désigne une fonction mais aussi un grand nombre de personnes occupant cette fonction. Ces requêtes sont fortement dépendantes du contexte général et par conséquent sujettes à des variations de sens et de référents, dans des périodes temporelles plus ou moins longues.

Thématiques	eco	int	pol	soc
ministre	**311**	**474**	**571**	**415**

TABLE 6 – La requête « ministre » en novembre

3.2.3 Analyse de la dimension diachronique de l'ambiguïté des requêtes

L'analyse de requêtes a fait apparaître un facteur de variation : la dimension diachronique. En effet, certaines requêtes changent de catégorisation selon les périodes temporelles comme « éruption » ou « cannabis ». Le cas de la requête « éruption » est intéressant car la requête désigne un même phénomène, mais à chaque fois un volcan différent, ce qui va donner la catégorie *société* au mois de décembre pour l'éruption du Piton de la Fournaise, l'éruption du mois de novembre du volcan Merapi en Indonésie sera classée comme *international,* alors que l'éruption islandaise en mai est catégorisée en *économie*. La mono-catégorisation variable s'applique également à certaines entités nommées comme la requête « delarue » consécutivement catégorisée en *cultures* puis en *société,* ce qui permet de réperer un changement, et potentiellement une ambiguïté si deux catégories co-occurrent dans la même période temporelle. On observe que sur les 50 requêtes les plus fréquentes chaque mois, seules quelques unes persistent au cours des huit mois : « afghanistan », « haïti », « israël » ou encore « pakistan ». Elles ont pour caractéristique principale d'être pour la plupart des entitées nommées (environ 70% des requêtes « durables », ce qui est comparable à la proportion dans l'ensemble du corpus). Ces requêtes désignent toutes des pays et s'avèrent être pluri-catégorisées. Nous allons analyser plus en détail deux requêtes à l'actualité très riche « pakistan » et « haïti », pour illustrer la manifestation de l'ambiguïté dans le temps.

La requête « pakistan » (tableau 7) a une catégorisation très variable. La catégorie *international* est majoritaire sur l'ensemble du corpus. Mais nous observons plusieurs variations au niveau des catégories majoritaires selon le sous-corpus considéré. La première variation en juin (catégorisation en *politique*), correspond dans notre corpus à l'apparition d'une affaire impliquant le Pakistan en tant qu'Etat : la vente de sous-marins au Pakistan par la France aurait servi de moyens de financements à un ancien premier ministre. La catégorie *international*, correspond à un emploi du Pakistan comme lieu, servant à localiser principalement des attentats et des opérations de la CIA contre les talibans. La catégorisation ne permet malheureusement pas d'identifier strictement les deux emplois de ce mot. La deuxième variation en novembre fait ressortir une catégorisation en *société*, celle-ci renvoyant à un

emploi de Pakistan comme lieu, en l'occurrence lieu qui produit des filières de combattants pour Al-Quaïda et des « potentiels auteurs d'attentats ». On identifie donc plusieurs significations pour la requête « Pakistan », à la fois lieu géographique, état-nation et lieu de formation de combattants. On observe que la requête « haïti » (tableau 8) a le même type de comportement. La pluri-catégorisation montre deux emplois possibles du mot « haïti » : comme référence au tremblement de terre (catégorisée en *cultures* et *international*) ou comme lieu. A noter cependant, la requête au mois d'octobre qui est mono-catégorielle (*international*). Il s'avère que cela est du à une focalisation sur un évènement unique en effet pendant cette période, Haïti a été touché par une épidémie de choléra. La requête a pris une signification différente. Ces deux exemples manifestent deux types particuliers d'ambiguïté, décrit par (Lecolle, 2007) sous le terme de « polysignifiance ». La polysignifiance désigne le fait qu'un nom de lieu habité présente des valeurs sémantico-référentielles différentes, renvoyant à la fois au lieu, mais aussi aux habitants et à l'institution qui le gouverne. Cette propriété des noms de lieux leur permet de pouvoir porter différentes significations, comme par exemple « Outreau » qui a pris la valeur d'erreur judiciaire en supplément de sa valeur locative, ou « Tchernobyl » qui désigne désormais une catastrophe nucléaire. Cette maléabilité du nom de lieu décrite par (Lecolle, 2007) ouvre une gamme large de possibilités, mais aussi de problèmes évidents si les différentes valeurs ne peuvent être discriminées et qu'elles apparaissent dans des contextes identiques. La polysignifiance semble difficile à appréhender par le biais de ressources lexicographiques ou de bases de connaissances, les différentes valeurs ne sont pas prises en compte, seule la fonction de localisation est retenue dans le cas des noms de lieu. La catégorisation thématique nous permet d'observer une partie de ce phénomène, en mettant en évidence les différentes significations de requêtes comme « haïti » ou « pakistan ».

Thématiques	clt	eco	int	pol	soc
Mai		9	**22**	14	14
Juin			8	**24**	
Juillet			**67**		
Aout			**474**		
Sept			**82**		
Oct			**91**	10	
Nov			31		**68**
Dec			**47**		

TABLE 7 – Catégorisation de la requête « pakistan »

Thématiques	clt	eco	int	soc	spr
Mai	**1**		**1**	**1**	
Juin		1	**3**	2	
Juillet	**17**				5
Aout			**50**		
Sept			**9**	1	
Oct			**102**		
Nov			**146**		
Dec			**55**	34	

TABLE 8 – Catégorisation de la requête « haïti »

3.3 Vers une typologie de l'ambiguïté des requêtes

L'ensemble de ces descriptions et analyses nous conduisent à proposer une synthèse sous forme de typologie, afin de rassembler les différentes formes que peut prendre l'ambiguïté dans notre contexte applicatif. Nous distinguons au final deux types de requêtes « ambigues » à la suite de notre analyse : les requêtes qui manifestent de l'ambiguïté « classique » et celles qui manifestent de l'ambiguïté « non classique » (tableau 9). Les requêtes du premier type contiennent de l'homonymie. La polysémie est quasi-absente de notre corpus de requêtes. Les requêtes du deuxième type contiennent trois types de manifestations de l'ambiguïté : les requêtes « polysignifiantes », les requêtes « pluri-référentielles » et les requêtes « génériques ». Nous avons en effet considéré les requêtes « polysignifiantes » comme étant des requêtes ambigües car la propriété de polysignifiance vaut potentiellement pour les noms de

lieux habités dans un contexte d'actualité. Les noms de lieux sont très présents dans notre corpus de requêtes en particulier parmi les plus fréquentes (18,6% des requêtes de notre corpus test contiennent un nom de lieu). L'ambiguïté référentielle est avant tout observée sur des noms et non pas des entités nommées. Les utilisateurs en formulant ce type de requêtes semblent faire confiance au contexte immédiat (par exemple « inondations »), mais lorsque l'actualité comporte plusieurs évènements auxquels peut référer cette requête, l'ambiguïté naît. Les requêtes que nous avons qualifiées de « générique » réfèrent à un objet, un évènement ou un domaine vaste et riche, contenant différentes « facettes »(Hearst, 2006). Ainsi, la requête « roman polanski » illustre ce cas de requêtes manifestant plusieurs facettes lors de la catégorisation, celle-ci fait apparaître deux catégories *cultures* et *international*, mettant en avant ses différents rôles dans l'actualité. Ces observations questionnent les conceptions qui considèrent le nom propre en tant que « désignateur rigide » (Kripke, 1980), parce qu'il désigne le même objet dans tous les mondes où cet objet est présent, étant alors univoque. Or, les ressources crées pour contenir diverses informations sur les entités nommées sont construites sur ce modèle, les variations contextuelles ne sont pas prises en compte. C'est pourquoi nous allons à présent regarder si ces formes d'ambiguïté sont présentes dans une ressource comme Wikipédia, l'encyclopédie en ligne, utilisée pour construire de nombreuses bases de connaissances comme DBPédia [4].

Requêtes avec ambiguïté « classique »	Requêtes avec ambiguïté « non classique »		
Homonymie	Polysignifiance	Ambiguïté référentielle	Généricité
voile, younes, corée	*haïti, pakistan, irak*	*éruption, otages,*	*sport, sarkozy*
royal, tabac	*afghanistan, xynthia*	*ministres, inondations*	*obama, gouvernement*

TABLE 9 – Typologie de l'ambiguïté des requêtes 2424 actu

3.4 Comparaison avec Wikipédia

Nous avons mis au jour de formes d'ambiguïté qui ne sont pas décrites dans les dictionnaires « traditionnels » comme la polysignifiance. Nous proposons alors de comparer notre classification à une ressource, afin de montrer la différence entre une procédure exploratoire comme la nôtre et une procédure de comparaison, entre une ressource et des requêtes. Nous avons comparé manuellement une partie de notre corpus (98 requêtes) avec l'encyclopédie en ligne Wikipédia. Deux aspects sont examinés :

– est-ce que la requête considérée est présente dans Wikipédia ?
– si la requête est une entrée de page d'encyclopédie, est-ce qu'elle renvoie vers une page d'homonymie ou de désambiguisation ?

Une page d'homonymie ou de désambiguïsation dans Wikipédia est simplement une page qui répertorie les différents sujets et articles partageant un même nom. Par exemple « éruption » renvoie vers une page qui recense un certain nombre d'« éruption » :

– *une **éruption** volcanique en géologie ;*
– *une **éruption** cutanée ou rash en médecine*
– *une **éruption** solaire, un phénomène très énergétique se produisant à la surface du Soleil.*
– *un morceau de guitare électrique du groupe américain Van Halen, **Eruption**.*
– *un groupe disco **Eruption***

Nous avons comparé l'annotation à partir de Wikipédia à la classification thématique, pour mesurer l'éventuel décalage. L'annotation a été effectuée sur les 97 requêtes du corpus test et seulement 61 requêtes sont annotées et comparées. La comparaison porte donc sur un nombre réduit de requêtes (tableau 10). 23 requêtes sont considérées comme non ambigües selon Wikipédia et mono-catégorielles, mais 7 requêtes sont mono-catégorielles et considérées comme ambigües par l'encyclopédie. En effet, Wikipédia recense parfois plus de sens qu'il n'existe dans notre contexte spécialisé, par exemple, Johnny Hallyday peut aussi être un cascadeur selon l'encyclopédie, l'actualité ne connaît que le chanteur. La situation s'inverse lorsqu'on considère l'accord entre la catégorisation et Wikipédia pour la pluri-catégorisation comme le montre le tableau (10). Un certain nombre de requêtes est identifié comme n'étant pas ambigües par Wikipédia, et pluri-catégorisées dans notre contexte (19 requêtes). Ces requêtes n'ont pas de pages de désambiguïsation. Parmi ces requêtes on retrouve « Haïti », « Afghanistan », « facebook », « gouvernement », « Carla Bruni », etc. Le problème étant que Wikipédia ne recense pas des ambiguïtés aussi fines de ce type, il se limite aux ambiguïtés de type homonymie. Des aspects comme la polysignifiance

4. http ://dbpedia.org/About

ou des variations propres au contexte ne figureront pas dans l'encyclopédie aussi riche soit-elle. La question de l'inadéquation d'une ressource pour capter des phénomènes qui créent de l'ambiguïté mais qui ne relèvent pas de la polysémie ou de l'homonymie se pose.

Comparaison	Mono-catégories	Pluri-catégories	**Total**
Un seul sens dans Wikipédia	23	12	**35**
Désambiguïsation dans Wikipédia	7	19	**26**

TABLE 10 – Comparaison de la catégorisation avec Wikipédia

4 Conclusion et Perspectives

L'expérimentation montre que l'ambiguïté peut se manifester à différents niveaux et sous différentes formes. La catégorisation thématique des requêtes a été un premier pas pour observer la diversité de l'ambiguïté et la complexité d'un contexte applicatif comme le nôtre. L'actualité est un domaine riche et continuellement en mouvement, c'est également un enjeu important d'améliorer l'accès au contenu des news. La constitution des corpus a été un challenge, il a fallu réunir à la fois les requêtes utilisateurs et les corpus de news où les utilisateurs avaient effectué leurs recherches. Bien qu'il ne soit pas possible d'assimiler la pluri-catégorisation à de l'ambiguïté de manière équivalente, la catégorisation nous a permis de mettre en évidence différentes manifestations de l'ambiguïté dans nos requêtes, à différents niveaux (durables ou variables). Les requêtes ambigües ne relevant pas de la polysémie ou de l'homonymie s'avèrent très mouvantes thématiquement, faisant apparaître une granularité plus fine, illustrant des facettes ou des évènements. Par ailleurs, la typologie proposée a fait apparaître certaines phénomènes qui ne peuvent être repérés par des ressources classiques. Les dictionnaires ne recensent pas les variations de points de vue et de signification ponctuelles. On voit donc l'intérêt d'utiliser des méthodes de type catégorisation ou clustering dans le repérage de l'ambiguïté. Cela pose la question dans un deuxième temps de l'intérêt de l'utilisateur, que va-t-il gagner à percevoir ces différents facettes ? Du point de vue de la méthode utilisée dans cette expérimentation, il apparaît qu'il serait intéressant de la compléter et de la comparer à d'autres types de méthodes automatiques de classification. La catégorisation par thématique a été utilisée parce que celle-ci offrait une lisibilité des résultats obtenus, c'est un outil grossier mais éclairant. Mais l'ambiguïté référentielle ou l'ambiguïté causée par trop de généralité, gagnerait à être observée à travers une représentation de l'information différente. La double catégorisation de la base textuelle, catégorisation thématique et clustering, nous permet de débuter l'analyse de requêtes classifiées, et de mettre en relation une requête avec un ou plusieurs clusters de documents dégageant ainsi des groupements pertinents.

Dans la perspective d'un repérage de l'ambiguïté des requêtes, nous pensons que la combinaison des moyens de classification avec d'autres indices pourrait être fructueuse, comme par exemple la reformulation de requêtes. Mais c'est une mesure à ne pas dissocier du comportement utilisateur, le cadre applicatif joue sur la reformulation ainsi que le mode de consommation (dans le cas de l'actu un « mur » interactif permet de consommer de l'actualité sans faire de requêtes). Cette expérimentation mérite donc d'être complétée. Elle constitue un préalable à un potentiel repérage de l'ambiguïté et à la mise en place d'aide pour l'utilisateur. Le repérage est un outil pour avancer dans la connaissance de la complexité des requêtes, ou pour assurer une meilleure expansion de requête. Le traitement de l'ambiguïté en elle-même peut passer par la proposition d'aide à l'utilisateur (Hearst, 2009). Notre contexte nous permet de développer une réponse de ce type à la manifestation de l'ambiguïté. La présentation des résultats peut gagner à l'injection de techniques de clustering et de classification, mais elle doit être maîtrisée et faire sens pour l'utilisateur. De ce fait, la mise en place de tests utilisateurs est indispensable.

Références

AARTS B. & MACMAHON A. (2006). *The Handbook of English Linguistics*. Oxford : Blackwell.

ARTILES J., BORTHWICK A., GONZALO J., SEKINE S. & AMIGÓ E. (2010). Weps-3 evaluation campaign : Overview of the web people search clustering and attribute extraction tasks. In *CLEF (Notebook Papers/LABs/Workshops)*.

ARTILES J., GONZALO J. & SEKINE S. (2007). The semeval-2007 weps evaluation : Establishing a benchmark for the web people search task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (Semeval-2007)*.

BERNARDINI A., CARPINETO C. & D'AMICO M. (2009). Full-subtopic retrieval with keyphrase-based search results clustering. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '09, p. 206–213, Washington, DC, USA : IEEE Computer Society.

CLOUGH P., SANDERSON M., ABOUAMMOH M., NAVARRO S. & PARAMITA M. L. (2009). Multiple approaches to analysing query diversity. In *SIGIR*, p. 734–735.

DARWISH K. & OARD D. W. (2003). Probabilistic structured query methods. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, p. 338–344, New York, NY, USA : ACM.

ERHMANN M. (2008). *Les Entités Nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation.* PhD thesis, Paris VII.

GAN Q., ATTENBERG J., MARKOWETZ A. & SUEL T. (2008). Analysis of geographic queries in a search engine log. In *Proceedings of the first international workshop on Location and the web*, LOCWEB '08, p. 49–56, New York, NY, USA : ACM.

HEARST M. A. (2006). Clustering versus faceted categories for information exploration. *Communications of the ACM*, **49**, 59–61.

HEARST M. A. (2009). *Search User Interfaces*. Cambridge University Press.

KRIPKE S. A. (1980). *Naming and Necessity*. Harvard University Press.

KROVETZ R. & CROFT W. B. (1992). Lexical ambiguity and information retrieval. *ACM Trans. Inf. Syst.*, **10**, 115–141.

LECOLLE M. (2007). Polysignifiance du toponyme, historicité du sens et interprétation en corpus. Le cas Outreau. *Corpus*, (6), 101–125.

RAHURKAR M. A., ROTH D. & HUANG T. S. (2008). Which "Apple" are you talking about ? In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, p. 1197–1198, New York, NY, USA : ACM.

SANDERSON M. (2000). Retrieving with good sense. *Information Retrieval*, **2**(1), 45–65.

SANDERSON M. (2008). Ambiguous queries : test collections need more sense. In *SIGIR*, p. 499–506.

SANTAMARÍA C., GONZALO J. & ARTILES J. (2010). Wikipedia as sense inventory to improve diversity in web search results. In *ACL*, p. 1357–1366.

SONG R., LUO Z., NIE J.-Y., YU Y. & HON H.-W. (2009). Identification of ambiguous queries in web search. *Information Processing and Management*, **45**(2), 216–229.

SPINK A., JANSEN B. J. & PEDERSEN J. (2004). Searching for people on web search engine. *Journal of Documentation*, **60**(3), 266–278.

STOKOE C., OAKES M. P. & TAIT J. (2003). Word sense disambiguation in information retrieval revisited. In *SIGIR*, p. 159–166.

VOORHEES E. M. (1993). Using wordnet to disambiguate word senses for text retrieval. In *SIGIR*, p. 171–180.

WEISS S. F. (1973). Learning to disambiguate. *Information Storage and Retrieval*, **9**(1), 33–41.

ZHAI C. X., COHEN W. W. & LAFFERTY J. (2003). Beyond independent relevance : methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, p. 10–17.

Extraction Automatique d'Informations Pédagogiques Pertinentes à partir de Documents Textuels

Boutheina Smine1,2 Rim Faiz2 Jean-Pierre Desclés1

(1) LaLIC, Université Paris-Sorbonne, 28 rue Serpente, 75006 Paris, France.
Boutheina.Smine@etudiants.univ-paris4.fr , Jean-pierre.Descles@paris4.sorbonne.fr
(2) LaRODEC, IHEC de Carthage, 2016 Carthage Présidence, Tunisie. Rim.Faiz@ihec.rnu.tn

RÉSUMÉ. Plusieurs utilisateurs ont souvent besoin d'informations pédagogiques pour les intégrer dans leurs ressources pédagogiques, ou pour les utiliser dans un processus d'apprentissage. Une indexation de ces informations s'avère donc utile en vue d'une extraction des informations pédagogiques pertinentes en réponse à une requête utilisateur. La plupart des systèmes d'extraction d'informations pédagogiques existants proposent une indexation basée sur une annotation manuelle ou semi-automatique des informations pédagogiques, tâche qui n'est pas préférée par les utilisateurs. Dans cet article, nous proposons une approche d'indexation d'objets pédagogiques (Définition, Exemple, Exercice, etc.) basée sur une annotation sémantique par Exploration Contextuelle des documents. L'index généré servira à une extraction des objets pertinents répondant à une requête utilisateur sémantique. Nous procédons, ensuite, à un classement des objets extraits selon leur pertinence en utilisant l'algorithme Rocchio. Notre objectif est de mettre en valeur une indexation à partir de contextes sémantiques et non pas à partir de seuls termes linguistiques.

ABSTRACT. Different users need pedagogical information in order to use them in their resources or in a learning process. Indexing this information is therefore useful for extracting relevant pedagogical information in response to a user request. Several searching systems of pedagogical information propose manual or semi-automatic annotations to index documents, which is a complex task for users. In this article, we propose an approach to index pedagogical objects (Definition, Exercise, Example, etc.) based on automatic annotation of documents using Contextual Exploration. Then, we use the index to extract relevant pedagogical objects as response to the user's requests. We proceed to sort the extracted objects according to their relevance. Our objective is to reach the relevant objects using a contextual semantic analysis of the text.

MOTS-CLÉS : extraction d'informations, objets pédagogiques, carte sémantique, exploration contextuelle, algorithme Rocchio

KEYWORDS : Information retrieval, pedagogical objects, semantic map, Contextual Exploration, Rocchio algorithm

1 Introduction

Avec la croissance rapide de la quantité d'information disponible en ligne et dans les bases de données, les moteurs de recherche jouent un rôle important dans l'apprentissage en ligne, car ils peuvent soutenir l'apprenant dans sa recherche d'informations nécessaires à son apprentissage, à sa formation, etc. Toutefois, ces systèmes de recherche d'information sont basés sur l'indexation des termes sans tenir compte de la sémantique du contenu pédagogique (Dehors et al., 2005), (Buffa et al., 2005). Une meilleure alternative est de proposer une approche d'indexation basée sur l'annotation sémantique des informations pédagogiques qui sont attestées dans les documents. Par une telle indexation, les informations pédagogiques présentées par l'auteur d'un document sont capturées et le processus d'apprentissage ou d'enseignement pour l'élève ou l'enseignant respectivement est facilité.

Nous proposons, dans cet article, une approche d'indexation automatique d'informations pédagogiques à partir de documents. Notre travail consiste d'abord à annoter les segments textuels (objets) reflétant un contenu pédagogique (Définition, Exemple, Exercice, etc.). Ensuite, nous procédons à une indexation de ces objets annotés pour extraire ceux qui sont pertinents par rapport à une requête utilisateur. Enfin, nous procédons à un classement de ces objets en utilisant l'algorithme de classification Rocchio.

Dans la section 2, nous positionnons cette contribution par rapport aux travaux existants. Nous consacrons la section 3 à la définition de la notion d'objet pédagogique. Une description détaillée de notre approche d'indexation d'informations pédagogiques est le sujet de la quatrième section. Avant de conclure, nous illustrons les résultats des expérimentations de notre approche dans la cinquième section.

2 Etat des lieux autour de la recherche d'informations pédagogiques

Nous détaillons ici divers points de l'état de l'art liés à notre approche d'indexation d'objets pédagogiques, à savoir l'annotation, l'indexation, et l'extraction d'informations pédagogiques à partir de documents textuels.

L'annotation comme technique d'indexation est appliquée dans plusieurs systèmes comme le système QBLS (Dehors et al., 2005) qui est une partie de la plateforme TRIAL SOLUTION (Buffa et al., 2005). Dans cette dernière, les utilisateurs annotent les livres manuellement selon le rôle pédagogique de leur contenu, les sujets abordés dans leur contenu (mots clés) et leurs relations avec d'autres ressources (référence, prérequis, etc.). Le système QBLS a pour but de structurer le cours en se référant à une ontologie pédagogique constituée de fiches (définition, exemple, énoncé, procédure, solution, etc.) et de ressources pédagogiques abstraites (cours, thème, notion, question). Il existe aussi le système SYFAX (Smei et al., 2005) qui annote semi-automatiquement le document pédagogique selon plusieurs critères (type du document, point de vue de l'utilisateur sur le document, etc.).

En vue d'indexer les documents, les annotations proposées par les différents systèmes cités ci-dessus sont stockées dans un entrepôt de connaissances pédagogiques. Par la suite, les réponses aux requêtes sont extraites à partir de cet entrepôt grâce à un moteur de recherche (*Corese* pour le système QBLS). Le système SYFAX applique un processus de raffinement de la requête basé sur une ontologie des types de documents pédagogiques et une autre ontologie des domaines des documents informatiques. Ceci permet d'extraire les documents pertinents par rapport à la requête.

Pour tous les systèmes présentés ci-dessus, une intervention humaine est requise afin d'enrichir les documents par des métadonnées. Cependant, la plupart des producteurs de ressources pédagogiques ne s'intéressent probablement pas au retour aux documents pour annoter leurs propres travaux. Notre travail se place dans cette perspective tout en procédant à l'automatisation du processus d'annotation.

D'autres travaux se sont intéressés à la recherche de ressources pédagogiques à partir du web (Thomson et al., 2003). Toutefois, le but de leur travail est limité à une extraction de métadonnées (Travaux Dirigés, Programme, Travaux Pratiques) relatives au document en entier en vue de les annoter et de les classifier. Toujours dans la même perspective, (Hassen et al., 2009) comparent l'efficacité des algorithmes Naïve Bayes et SVM dans la classification des ressources pédagogiques basée sur un ensemble de propriétés (catégorie du contenu, titre du cours, année, auteur, etc.).

A notre connaissance, ces travaux de recherche portant sur l'indexation de documents pédagogiques se sont intéressés à une indexation du document en l'annotant par un ensemble de métadonnées relatives à la totalité du document. D'autres approches basées sur des patrons linguistiques ont été appliquées dans plusieurs travaux pour extraire les définitions à partir de ressources pédagogiques afin de constituer un glossaire (Westerhout et al., 2008) ou encore pour répondre à divers types de questions (Greenwood et al., 2003). Cependant, les patrons sont appliqués la plupart du temps à extraire des objets pédagogiques de type "Définition" en raison de l'accessibilité des structures langagières relatives à ce type que ce soit sur le web (wikipédia, dictionnaires, etc.) ou dans d'autres sources comme les rapports, les manuels d'utilisation, etc.

Dans cet article, nous proposons une annotation automatique des informations pédagogiques avec des métadonnées sémantiques (Définition, Exemple, Exercice, etc.). Ce qui nous permettrait d'indexer ces informations en vue d'une extraction des informations pertinentes par rapport à une requête utilisateur.

3 Notion d'objets pédagogiques

Un utilisateur "extracteur" d'informations pédagogiques pertinentes est guidé dans sa lecture par certains passages, des segments textuels (phrases ou de paragraphes). L'hypothèse générale utilisée ici est d'essayer de reproduire ce que fait un humain, en particulier l'apprenant, en soulignant certains segments textuels reflétant un contenu pédagogique. Ces segments de type pédagogique, appelés objets pédagogiques, existent, généralement, dans les documents pédagogiques sous forme de définitions, exemples, exercices, plan, questions et réponses, etc. Un objet pédagogique peut être défini comme une entité numérique ou non (Flory, 2004) qui peut être utilisée ou citée dans un apprentissage. Dans notre cas, un objet pédagogique correspond à un segment textuel reflétant un contenu pédagogique.

Un apprenant pourrait être intéressé par une définition en formulant une requête, par exemple: trouver les documents qui contiennent "La définition du langage SQL". Un autre utilisateur recherche, en explorant de nombreux textes (encyclopédies spécialisées, manuels, articles), des exemples sur un concept (par exemple, «l'inflation» dans l'économie, «polysémique» en linguistique, ..) pour l'intégrer à ses ressources pédagogiques. Un autre utilisateur peut être intéressé, à la pratique des exercices sur un concept. L'objectif de ces types d'objets pédagogiques (Définition, Exemple, Exercice) est une annotation possible des segments textuels pédagogiques qui correspondent à une recherche guidée afin d'en extraire des objets pédagogiques à partir de textes.
Chaque type pédagogique, comme nous l'avons mentionné ci-dessus, est explicitement indiqué par les marqueurs linguistiques identifiables dans les textes. Notre hypothèse est que chaque type d'objet pédagogique laisse des traces discursives dans le document texte. Les types d'objets pédagogiques sont décrits comme suit:

- D'une part, une relation complexe entre les concepts dans une structure «carte sémantique» (Figure 1) et d'autre part un ensemble de classes et sous-classes d'unités linguistiques (indicateurs et indices).

- Un ensemble de règles communautaires où chaque règle concerne une classe d'indicateurs avec des indices différents.

La carte sémantique (Figure 1) est une organisation des types d'objets pédagogiques. Elle peut être conçue aussi comme une ontologie des types d'objets pédagogiques indépendamment des différents domaines d'application. En effet, les expressions de la carte sémantique pour un type d'objet sont les mêmes dans différents domaines comme l'informatique, mathématiques, gestion, ... car ces expressions sont utilisées par l'auteur pour exprimer une information pédagogique. Dans certains types de textes (textes narratifs, articles de presse,) les expressions pédagogiques ne sont pas présentes mais dans d'autres (support de cours, devoirs, travaux dirigés, ..), ces expressions organisent le texte et donnent des informations sur l'intention de l'auteur.

Le premier niveau de la carte sémantique (Figure 1) présente 6 types d'objets pédagogiques : (i) Cours, (ii) Plan, (iii) Exercice, (iv) Exemple, (v) Définition, (vi) Caractéristique. Par exemple, les règles du type d'objet "Définition" sont déclenchés par la présence de noms ou de verbes définitoires (par exemple: "*est défini*", et l'annotation sémantique est attribuée si des indices linguistiques, comme les prépositions (l'indice de l'exemple précédent est "*par*"), sont trouvés dans le contexte de l'indicateur.

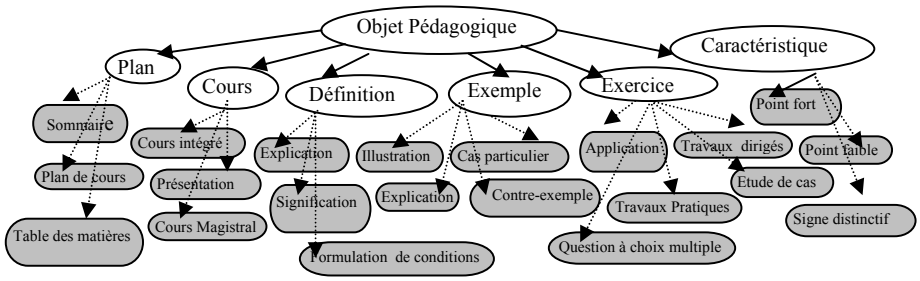


Figure 1 : Carte sémantique des types d'objets pédagogiques

4 Approche proposée pour la recherche d'informations pédagogiques à partir de documents

L'approche que nous proposons se décompose en deux principales parties: dans la première partie, nous procédons à une annotation sémantique des segments textuels représentant des objets pédagogiques (Smine et al., 2010). La deuxième partie exploite les annotations générées par la première partie pour créer un index qui est capable de localiser les segments textuels pertinents par rapport à des requêtes associées aux types pédagogiques (Définition, Exemple, Exercice, etc.). Pour classer les réponses selon leurs pertinences, nous appliquons l'algorithme de classification Rocchio sur les objets pédagogiques extraits.

4.1 Annotation des objets pédagogiques

4.1.1 Segmentation

La segmentation est la détermination des limites des unités linguistiques (unités comme proposition, phrase, paragraphe, etc.). La segmentation des textes en petites unités (phrases) reste encore une tâche difficile à réaliser, vu qu'un point suivi d'une majuscule ne peut pas déterminer le début ou la fin d'un segment. Il est nécessaire de prendre en compte tous les marqueurs typographiques. Il existe des travaux qui considèrent l'aspect multilangue dans leur segmentation comme le travail de (Mourad, 2002) qui propose de définir un segment textuel en se basant sur une étude systématique des marques de ponctuation. Nous avons effectué la segmentation de nos documents en intégrant les règles linguistiques développées par Mourad. Pour chaque document segmenté, le résultat obtenu est un fichier XML balisé par des balises <Section>, <Paragraphe>, <Phrase>.

4.1.2 Annotation des objets pédagogiques

Pour annoter les objets, nous explorons la technique d'Exploration Contextuelle 'EC' (Desclés, 1997). C'est une technique de traitement linguistique et sémantique du langage, qui fait appel à des marqueurs discursifs explicites (morphèmes, mot, expression, etc.) caractéristiques d'une intention pragmatique de l'auteur. 'EC' consiste à appliquer des règles dans un contexte déterminé par des indices. Elle a l'avantage d'être indépendante d'un domaine particulier, car les règles décrivant les structures linguistiques sont indépendantes d'un domaine particulier. C'est une méthode qui a été validée par les travaux de (Djioua et al., 2006) et (Elkhlifi et al., 2010). En plus, 'EC' ne nécessite pas une analyse morphosyntaxique du texte, ce qui réduit considérablement le temps d'exécution pendant l'implémentation de la méthode.

Par l'exploration contextuelle du contenu des documents, nous pouvons repérer et annoter les objets pédagogiques contenu dans ces documents, par exemple, « des exemples de requêtes SQL», « des exercices sur le langage UML », « les définitions d'une ou de plusieurs notions », etc. Ces objets sont exprimés par des structures langagières comme « …se définit par… », « est défini par… » pour le type *Définition* ou « Exercices sur… », « Travaux dirigés » pour le type *Exercice*. Ils sont explicitement indiqués par des indicateurs linguistiques identifiables dans les textes (verbes, noms, adjectifs). Ces indicateurs sont parfois polysémiques, ils ont besoin d'indices linguistiques pour clarifier l'indétermination. Les relations reliant les

indicateurs aux indices sont définis dans le cadre des règles. Une règle (IdR) se déclenche au moment de l'identification de l'un de ses indicateurs (Indicateur) ensuite elle essaye de localiser des indices linguistiques dans le contexte gauche (CL_1, CL_2) et/ou droite (CR_1, CR_2) de l'indicateur ce qui confirme ou non la valeur sémantique exprimée par l'indicateur. A chaque type d'objet pédagogique correspond un ensemble de règles. Des exemples de règles sont présentés dans le tableau suivant (Tableau 1).

IdR	CL_1	CL_2	Indicateur	CR_1	CR_2	Type|Sous-type de l'objet pédagogique
RD1	est|sont		défini|définie| définis	par		Définition|Explication
RD2			est | sont	le|la|un|une |des|les		Définition|Explication
RC1	La|Les Des|Une		caractéristique| caractéristiques	du|de|des	est| sont	Caractéristique|Signes distinctifs
RE1	Voici	un|l'|les |des	exemple| exemples	du|de|des		Exemple|Illustration

Tableau 1 : Des exemples de règles

Nous avons ajouté un composant à chaque règle qui représente l'emplacement du terme de la requête à rechercher dans le cadre du segment exprimant l'objet pédagogique. Le besoin d'ajouter ce composant est né de la variation de l'emplacement du terme à rechercher avec la variation des structures langagières exprimant les objets pédagogiques. Ceci permet d'identifier les segments textuels exprimant le type d'objet pédagogique ainsi que le concept demandés par l'utilisateur. Par exemple, pour le même type d'objet pédagogique "Définition" : le terme à rechercher "*Maintenance*" peut exister au début du segment "*La maintenance est définie comme l'ensemble des activités destinés à maintenir ou à rétablir un bien dans un état de sûreté de fonctionnement*" ou au milieu du segment pour le cas "*L'AFNOR a défini la maintenance comme étant l'ensemble des activités de remise en état de fonctionnement d'un système*". Sans la considération de ce paramètre, le système peut ne pas extraire l'objet demandé par l'utilisateur comme par exemple, pour le type *Cours*, la plupart de ses règles d'EC exigent un emplacement du terme de la requête au niveau du Titre du document. Au cas où le terme est recherché ailleurs du titre, le résultat de la recherche sera erroné.

De ce fait, l'emplacement du terme est un paramètre qui diffère d'une règle à une autre selon la structure langagière exprimée par cette dernière. Nous avons désigné cet emplacement par une étiquette, qui prendra une valeur parmi un ensemble fini de valeurs désignant l'emplacement du terme par rapport aux indicateurs et indices. Par exemple, GIND indique le terme se place à gauche de l'indicateur ou TITRE indique que l'emplacement du terme est au niveau du titre du document. En fait, dans plusieurs cas, le titre peut nous révéler des connaissances sur le contenu du document.

Pour chaque type d'objet de la carte sémantique (cf. Figure 1), nous avons défini un ensemble de règles qui couvrent toutes les formes linguistiques possibles de l'objet pédagogique. Nous avons commencé par un exemple textuel relatif à chaque type pour généraliser toutes les structures langagières. Cette méthode permet de définir de manière incrémentale une base solide de règles. Nous avons développé en totalité environ 200 règles. L'ensemble des règles développées, ainsi que la carte sémantique forment les ressources linguistiques utilisées dans notre approche.

Nous prenons un extrait de texte à partir d'un document pédagogique

Chapitre 1

Présentation de SQL

SQL est un langage complet de gestion de bases de données relationnelles. Il n'est pas un langage conceptuel. Il a été conçu, dans les années 70, par IBM. Il est devenu le langage standard des systèmes de gestion de base de données (SGBD).

Pour le type d'objet pédagogique "*Définition*", la règle RD2 (cf. Tableau 1), appliquée à l'exemple ci-dessus, permet d'annoter la phrase " *SQL est un langage complet de gestion de bases de données relationnelles*". Le type d'objet pédagogique est détecté grâce à l'expression "*est* " qui est une occurrence Ii de l'indicateur du type "*Définition*" et l'indice droit CR1 "*un*".

Pour le type "*Cours*", le repérage de l'occurrence Ii au niveau du titre est suffisant pour annoter le document comme un cours. L'indicateur nominal de l'objet pédagogique est le mot "Cours", et d'autres noms comme "Chapitre, "Notes de cours". A part le titre, l'existence de l'indicateur "Cours" n'implique pas l'annotation du document comme un cours.

Notons que la phrase "*Il n'est pas un langage conceptuel*" illustre le cas des indices négatifs. En effet, la présence de l'expression "n'….pas" annule l'annotation du segment comme Définition, malgré la présence de l'indicateur "*est*" et l'indice droit CR1 "*un*".

Afin d'annoter le segment " *Il a été conçu, dans les années 70, par IBM*" comme une "*Caractéristique*", nous détectons en premier lieu l'expression "*a été conçu*" ensuite nous cherchons, dans le contexte droit de l'indicateur, le CR1 "*par*". En cas où les deux éléments (Ii et CR1) sont présents, alors le système annote le segment comme une caractéristique.

Concernant le type d'objet "Exercice", l'indicateur peut être verbal (a) ou nominal (b), par exemple :

- (a) "*Formulez une clause SQL.…..…*" a comme indicateur verbal "*Formulez*"
- (b) "*Exercices sur requêtes SQL*", son indicateur est le nom "*Exercices*"

4.2 Génération de l'index

Notre objectif, par l'annotation, est de générer un index sémantique contenant à la fois des objets pédagogiques annotés selon leur type, en utilisant la méthode d'annotation détaillée ci-dessus, et l'emplacement du terme de la requête spécifié par la règle appliquée pour annoter l'objet. Cet index servira à extraire les objets répondant à la requête utilisateur. Les métadonnées générées par les annotations des différents objets sont stockés dans une base de données. Pour chaque objet pédagogique annoté, les métadonnées suivantes sont introduites dans l'index : (1) L'objet pédagogique annoté (OBJECT), (2) Chemin du document analysé (PATH), (3) Type de l'objet annoté (TYPE), (4) Identifiant de la règle appliquée pour annoter le segment (IDRule) et (5) L'emplacement du terme de la requête (TERMEMP). La figure suivante (Figure 3) montre deux exemples d'objets annotés.

EDIT	OBJECT	PATH	TYPE	IDRULE	TERMEMP
	La production est une transformation de ressources appartenant à un système productif et conduisant à la création de biens ou de services.	C:\Documents and Settings\Boutheina SMINE\Mes documents\Evaluation\Gestion de Production.txt	Définition:Explication	RD2	GIND
	2) Exprimer en algèbre relationnel les requêtes suivantes et donner ses résultats : checkbld Nom des immeubles ayant plus de 10 étages. checkbld Qui habite le « Koudalou » ? checkbld Nom et Profession des personnes ayant emménagé avant 1994. checkbld Gérant des immeubles ayant un appartement de plus de 150 m². checkbld Dans quel immeuble habite un acteur ? checkbld Age et profession des occupants de l'immeuble géré par « Ross » ? checkbld Qui n'habite pas un appartement géré par « Ross » ?	C:\Documents and Settings\Boutheina SMINE\Mes documents\Evaluation\Base de donnée.txt	Exercice:Travaux Dirigés	RE5	TITRE

Figure 3 : **Deux exemples d'objets annotés et indexés**

Afin de pouvoir extraire les objets pédagogiques qui contiennent des termes de la requête, nous avons utilisé la base de synonymes WOLF (qui représente la partie traduite en Français du dictionnaire WordNet) permettant d'enrichir la requête en prenant en compte tous les termes équivalents au terme de la requête. Ce

dernier est remplacé par la liste de ses synonymes. Ceci permet d'étendre le champ de la recherche. La requête est ainsi composée des termes à rechercher (par exemple "Langage SQL") et du type d'objets pédagogiques requis par l'utilisateur (par exemple : Exercice).

Grâce à un moteur de recherche (implémenté sous la plateforme *Lucene*), le système se connecte à l'index généré et retient les documents contenant des objets pédagogiques de même type que celui énoncé dans la requête (Exercice). Ensuite, le moteur procède à une recherche des termes de la requête (Langage SQL ainsi que ses synonymes) à partir des objets annotés et indexés. Cette recherche s'effectue dans l'emplacement désigné par la règle avec laquelle est annoté l'objet pédagogique. Par exemple, si l'emplacement du terme spécifié par la règle est DIND, le terme de la requête est recherché à droite de l'indicateur de la règle appliquée (Dans ce cas règle de type Exercice). Dans le cas où la requête est composée du type pédagogique "Exercice" et le terme «Langage SQL", le moteur de recherche procède comme suit:

- il extrait tous les objets pédagogiques trouvés dans l'index associé à l'annotation «Exercice»

- Pour chaque objet extrait, il recherche le terme "langage SQL» et ses synonymes dans l'emplacement spécifié par la règle d'annotation.

- Sélection, à partir des objets pédagogiques extraits, les objets comportant une occurrence du terme «langage SQL» ou ses synonymes dans le bon emplacement.

- Afficher toutes les informations présentes dans l'index associé à chaque objet pédagogique sélectionné.

4.3 Classement des objets pédagogiques

Après l'extraction des objets pédagogiques répondant à la requête utilisateur, une autre étape suit pour classer les réponses dans un ordre croissant selon leur similarité avec la requête. Pour classer ces objets, nous avons utilisé l'algorithme de Rocchio (Rocchio, 1971), adapté à la classification des textes (Ittner et al., 1995). L'utilisateur choisit un concept pour le correspondre au terme de sa requête, parmi une liste de 15 concepts appartenant à différents domaines (domaine de l'informatique, économie, génie mécanique, biologie, etc.). Ce sont des concepts auxquels appartient l'ensemble des documents du corpus d'annotation et d'indexation. Le concept choisi représente la classe C_{user} par rapport à laquelle les objets seront classés selon leur pertinence. Rappelons que nous considérons un objet pédagogique comme un segment textuel ayant différentes tailles (Phrase, paragraphe, document, etc.) selon le type de l'objet.

Nous représentons les données (les objets d'apprentissage et de test) par des vecteurs de poids numériques. Le vecteur de poids pour le *m* ième objet pédagogique est $V^m = (p_1^m, p_2^m, \ldots\ldots, p_l^m)$, où *l* est le nombre de termes index utilisés. Nous utilisons comme termes des mots singuliers et composés. Nous adoptons la mesure de poids TF-IDF (Salton, 1991) et nous définissons le poids p_k^m comme suit :

$$p_k^m = \frac{f_k^m \log(N/n_k)}{\sum_{j=1}^{l} f_j^m \log(N/n_j)}$$

Avec *N* est le nombre total d'objets, n_k est le nombre d'objets dans lesquels le terme index *k* apparaît, et f_k^m est :

$$f_k^m = \begin{cases} 0 & q = 0 \\ \log(q) + 1 & Sinon \end{cases}$$

Avec *q* est le nombre d'occurrences du terme index *k* dans l'objet *m*. Dans l'algorithme de Rocchio, un prototype est produit pour chaque classe *C*. Ce prototype est représenté par un vecteur singulier $\vec{c}_j$ de même dimension que le vecteur de poids original $v^1, \ldots\ldots, v^N$. Pour chaque classe *C*, the *k ième* terme dans son prototype est défini comme

$$\vec{c}_j = \frac{\alpha}{|C_j|} \sum_{m \in C_j} p_k^m - \frac{\beta}{|N - C_j|} \sum_{m \notin C_j} p_k^m$$

Avec *Cj* est l'ensemble de documents appartenant à la classe *C*. Les paramètres α et β contrôlent la contribution des exemples positifs et négatifs par rapport au vecteur prototype. Nous utilisons les valeurs standards α =4 et β = 16 (Buckley et al., 1994).

Une fois l'apprentissage achevé, nous classons les nouveaux objets fournis comme réponses à la requête utilisateur. Ce classement se fait selon leur pertinence par rapport à la classe C_{user} choisie par l'utilisateur. Les objets à classer sont tout d'abord convertis en vecteurs de poids, et puis comparés aux vecteurs de poids prototypes des différentes classes en utilisant la mesure de similarité cosinus.

La mesure de similarité entre l'objet de vecteur $\vec{O}$ et la classe C_{user} de vecteur $\vec{C}_{user}$ est définie comme :

$$\cos(\vec{C}_{user}, \vec{O}) = \frac{\vec{C}_{user} * \vec{O}}{\left|\vec{C}_{userl}\right\|\vec{O}|}$$

Les objets ayant une valeur de similarité avec la classe C_{user} supérieure à un seuil θ sont sélectionnés, ensuite classés dans un ordre croissant selon la valeur de leurs similarités par rapport à la classe C_{user}. La valeur du seuil θ varie selon le type d'objet pédagogique. Par exemple, un objet annoté par le type "*Cours*" contient plus de termes significatifs qu'un objet annoté par le type "*Exercice*" ($\theta_{Course} < \theta_{Exercice}$). Nous ne prenons en compte que les valeurs positives de la mesure de similarité. Les objets sélectionnés sont alors affichés pour constituer la fiche pédagogique demandée par l'utilisateur. Une fiche pédagogique rassemble les objets pédagogiques de type celui exprimée par l'utilisateur dans sa requête et correspondant au même concept que celui recherché par l'utilisateur. Cette fiche permet une accessibilité aux objets directement sans avoir accès au document en entier.

5 Expérimentations et Résultats

L'objectif de cette étape est d'évaluer les performances des différents modules. Un des indicateurs importants est donc le nombre des réponses pertinentes par rapport au nombre de documents indexés. Pour valider notre approche d'indexation d'objets pédagogiques, nous avons développé le système SRIDoP (**S**ystème de **R**echerche d'**I**nformations à partir de **Do**cuments **P**édagogiques) en utilisant le langage Java sous l'environnement Eclipse et le système de gestion de base de données Oracle. SRIDoP comporte les trois modules suivants : Annotation et indexation des objets pédagogiques selon leurs types, Appariement entre la requête utilisateur et les objets pédagogiques indexés, et Classement des objets pédagogiques.

Notre corpus d'apprentissage ainsi que celui du test est le même pour toutes les étapes d'annotation, d'indexation et de classification. Pour le corpus d'apprentissage, nous avons collecté un ensemble de documents couvrant 15 concepts (ceux utilisés dans la génération de fiches pédagogiques). En fait, pour chacun de ces concepts, une requête a était formulée et exécutée sur le moteur de recherche Google. Les 20 premiers résultats sont collectés. Notons que le sens de quelques termes peut être ambigu, par exemple "Base" ou "Enregistrement". Pour désambiguïser la requête, nous ajoutons le terme "Données". Pour faire disparaitre l'ambiguïté, nous misons sur le type pédagogique des documents retournés en réponse. Les documents collectés sont constitués de 60 supports de cours, 65 Travaux Dirigés, 83 Présentation PowwerPoint, 30 Travaux Pratiques, et quelques documents de différentes natures (articles de Presse, articles scientifiques, etc.). La longueur moyenne de ces documents constituant le corpus d'apprentissage est 23 pages.

Notre corpus de test est composé de 1000 documents, principalement de nature pédagogique : des Supports de cours, des Travaux Dirigés, des présentations PowerPoint, des Travaux Pratiques, des manuels d'utilisation, et d'autres documents de différentes natures. La longueur moyenne des documents est 53.6 pages. Les documents ont différents formats (DOC, PDF, HTML, PPT, etc.).

5.1 Première étape : Annotation des objets pédagogiques

Pour évaluer le processus d'annotation, le corpus de test a été annoté par deux experts : pour chaque objet pédagogique repéré, ils précisent son type. Les résultats du processus d'annotation effectué par notre système SRIDoP sont illustrés dans le Tableau 2.

Type de l'objet pédagogique	NOA	NOAC	NOMAC	Précision (%)	Rappel (%)	F-Mesure (%)
Plan	88	85	98	96,59	86,73	91,40
Cours	72	60	85	83,33	70,59	76,43
Définition	228	140	266	61,40	52,63	56,68
Caractéristique	139	124	156	89,21	79,49	84,07
Exemple	357	349	376	97,76	92,82	95,23
Exercice	760	705	776	92,76	90,85	91,80

Tableau 2 : Les résultats de l'étape Annotation

$$\text{Précision} = \frac{NOAC}{NOA} \qquad Rappel = \frac{NOAC}{NOMAC} \qquad F - Mesure = 2 * \frac{\text{Précision} * Rappel}{\text{Précision} + Rappel}$$

Avec : **NOA** : Nombre d'objets annotés, **NOAC** : Nombre d'objets annotés correctement, **NOMAC**: Nombre d'objets annotés par les experts.

Nous remarquons que la précision de l'annotation dépasse les 85% pour la plupart des types d'objets (*Exemple, Exercice, Plan, etc.*). Notons que, pour le type « Définition », cette précision est moyenne. Ceci dérive du fait que certaines règles peuvent annoter à la fois des énoncés définitoires ou non. Tel le cas de la règle « R2 » ayant comme indicateur l'occurrence «**est un**». Cet indicateur peut identifier un segment de nature définitoire (exemple : « *UML **est un** langage de modélisation conceptuelle orienté objet* ») ou un autre segment de nature non définitoire (exemple : « *Le facteur temps **est un** des plus importants dans la réalisation d'un projet* »). Pendant la phase d'expérimentation, nous avons pu constater aussi que la qualité de l'annotation est étroitement liée à la qualité de la segmentation du document.

5.2 Deuxième étape : Indexation des objets pédagogiques

A travers une interface de recherche d'informations, l'utilisateur saisit les termes à rechercher, et choisit le type (et sous-type) de l'objet pédagogique relatif au terme à rechercher. Les réponses aux requêtes sont affichées sous forme de liens permettant d'accéder à l'objet pédagogique répondant au besoin de l'utilisateur. Pour tester ce module de recherche d'objets pédagogiques, nous avons formulé les mêmes 25 requêtes pour chacun des types d'objets pédagogiques. Ces requêtes appartiennent aux différents domaines du corpus. Pour chaque type d'objet, nous avons illustré le nombre de réponses ramenées et le nombre de réponses jugées pertinentes compte tenu de l'ensemble des requêtes formulées. Les résultats sont résumés dans le tableau suivant (Tableau 3).

Type de l'objet pédagogique exprimé par la requête	NR	NRP	NRRU	Précision (%)	Rappel (%)	F-Mesure (%)
Plan	72	66	77	91,67	85,71	88,59
Cours	43	35	54	81,40	64,81	72,16
Définition	156	112	193	71,79	58,03	64,18
Caracteristique	94	86	112	91,49	76,79	83,50
Exemple	213	198	230	92,96	86,09	89,39
Exercice	517	465	520	89,94	89,42	89,68

Tableau 3 : Les résultats de l'étape d'appariement Documents-Requête

$$\text{Précision} = \frac{NRP}{NR} \qquad Rappel = \frac{NRP}{NRRU} \qquad F - Mesure = 2 * \frac{\text{Précision} * Rappel}{\text{Précision} + Rappel}$$

Avec : **NR** : Nombre d'objets (réponses) retournés à l'utilisateur, **NRP** : Nombre d'objets (réponses) pertinents retournées à l'utilisateur, **NRRU**: Nombre d'objets pertinents.

A l'issue de ces expérimentations, nous remarquons que les résultats de l'indexation d'informations pédagogiques sont étroitement liés aux résultats de l'annotation (cf. Figure 4). La valeur de "F-Mesure" de l'extraction évolue avec la valeur de "F-Mesure" de l'annotation. Ceci s'explique par le fait, que l'extraction est effectuée à partir d'objets pédagogiques annotés et indexés. La qualité de la recherche s'améliore en améliorant celle de l'annotation. Cette dernière est elle-même dépendante de la qualité de segmentation comme nous l'avons déjà mentionné.

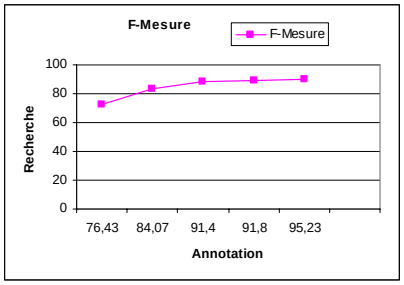


Figure 4 : Evolution des résultats de la recherche par rapport à celles de l'annotation

5.3 Troisième étape : Classement des objets pertinents

Après une extraction des objets pédagogiques, nous classons ces objets selon leur similarité avec la classe C_{user}. Suite à plusieurs expérimentations, nous avons fixé la valeur du seuil θ :

- 0.1 pour les types "Cours" et "Définition",
- 0.3 pour les types "Plan" et "Exemple",
- 0.45 pour les types "Caractéristique" et "Exercice".

Notons que d'un côté, diminuer la valeur de θ réduit l'ensemble des objets pertinents retournés à l'utilisateur. D'un autre côté, augmenter la valeur de θ amène à une sélection des objets non pertinents.

Nous avons assigné chaque objet à l'une de ces trois catégories : **A** (objets classés comme pertinents), **B** (objets classés correctement comme pertinents), **C** (objets pertinents). Les valeurs de précision, de rappel et de F-Mesure sont calculées pour chaque type d'objet pédagogique comme suit :

$$\text{Pr}écision = \frac{B}{A} \qquad Rappel = \frac{B}{C} \qquad F - Mesure = 2 * \frac{\text{Pr}écision * Rappel}{\text{Pr}écision + Rappel}$$

Nous illustrons ces valeurs relatives à chacun des types d'objets dans la Figure 5.

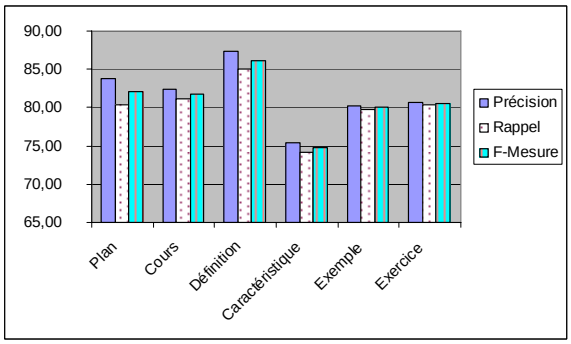


Figure 5 : Précision, Rappel et F-Mesure de l'étape de classement des objets

La figure ci-dessus présente, pour chaque type d'objet (représenté sur l'axe des abscisses), sa valeur de précision représentée en bleu, sa valeur de rappel en pointillé et sa valeur de F-Mesure représentée en rayures. Nous constatons que les valeurs de précision sont comprises entre 75% et 87% et que celles du rappel entre 74% et 85%. Notons que l'étape de classement ne dépend pas strictement de celles de l'annotation et d'appariement mais plutôt d'autres paramètres comme le corpus d'apprentissage, le choix des termes index, etc.

6 Conclusion et Perspectives

Dans cet article, nous avons proposé une approche d'indexation d'objets pédagogiques basée sur une annotation sémantique du texte par exploration contextuelle en vue d'une extraction des objets pédagogiques pertinents. Actuellement, notre travail présente un intérêt important dans plusieurs domaines d'application comme l'apprentissage en ligne, l'enseignement à distance (e-learning), l'éducation, etc. Pour évaluer notre approche, nous avons développé le système SRIDoP qui comprend les modules d'annotation, d'indexation, et de classement des objets selon leur pertinence. Nous remarquons, à travers les résultats d'évaluation, que notre approche permet d'avoir accès aux connaissances qui sont exprimées dans les textes selon un type donné, et de ramener des énoncés qu'un système de recherche d'informations classique ne parvient à capter par son approche d'indexation par mots clés.

L'un des travaux futurs que nous envisageons est l'extension de la carte sémantique des types d'objets pédagogiques par d'autres types comme Méthode, Auteur, Date, etc. Nous pensons aussi à la proposition d'une fonction score qui fusionne les résultats des deux modules d'annotation et de classement en vue de sélectionner les résultats pertinents.

Bibliographie

BUCKLEY C., SALTON G., ALLAN J. (1994). The effect of adding relevance information in a relevance feedback environment. Actes de *International ACM SIGIR Conference*, 292-300.

BUFFA M., DEHORS S., FARON-ZUCKER C., SANDER P. (2005). Vers une approche Web Sémantique dans la conception d'un système d'apprentissage. Revue du projet TRIAL SOLUTION, *AFIA*.

DEHORS S., FARON-ZUCKER C., STROMBONI J.P., GIBOIN A. (2005). Des annotations Sémantiques pour apprendre : l'Expérimentation QBLS. *WebLearn*.

DESCLES J.P. (1997). Système d'exploration Contextuelle. *Co-texte et calcul du sens*, Caen, 215-232.

DJIOUA B., FLORES, J.G, BLAIS A., DESCLES J.P., GUIBERT G., JACKIEWIEZ A., LE PRIOL F., NAIT BAHA L., SAUZAY B. (2006) Excom: an automatic annotation engine for semantic information. Dans *Proc. FLAIRS*, AAAI Press, Florida, 285-290.

ELKHLIFI A., FAIZ R. (2009). Automatic Annotation Approach of Events in News Articles. *International Journal of Computing & Information Sciences*, 19-28.

Elkhlifi A., FAIZ, R. (2010). French-Written Event Extraction Based on Contextual Exploration. Dans *Proc. FLAIRS*, AAAI Press, Florida.

FLORY L. (2004). Les caractéristiques d'une ressource pédagogique et les besoins d'indexation qui en résultent. *Journée d'étude sur l'Indexation des ressources pédagogiques numériques*, Ennsib, Villeurbanne.

GREENWOOD M.A., SAGGION H. (2004). A Pattern Based Approach to Answering Factoid, List and Definition Questions. Dans *Proc. RIAO 2004*, Avignon, France.

HASSAN S., MIHALCEA R. (2009). Learning to identify educational materials. Dans *Proc. RANLP*, Bulgaria.

ITTNER D.J., LEWIS D.D., AHN D. D. (1995). Text categorization of low quality images. Actes de *SDAIR*, Las Vegas, US, 301-315.

MOURAD G. (2002). La segmentation de textes par Explration Contextuelle automatiques, présentation du module SegATex. Dans *Inscription Spatiale du Langage : structure et processus ISLsp*, Toulouse.

ROCCHIO J. (1971). Relevance feedback information retrieval. *In Gerard Salton editor, The Smart retrieval system experiments in automatic document processing*, Prentice-Hall, Englewood Cliffs, NJ, 313-323.

SALTON G. (1991). Developments in automatic text retrieval. *Science*, 253 (5023), 974-980.

SMEI H., BEN HAMADOU A. (2005). Un système à base de métadonnées pour la création d'un cache communautaire-Cas de la communauté pédagogique. Dans *Proc. IEBC*, Hammamet, Tunisie.

SMINE B., FAIZ R., DESCLES J.P. (2010). Analyse de documents pédagogiques en vue de leur annotation. *Revue des Nouvelles Technologies de l'Information (RNTI)*, E-19, Ed. Cépaduès, 429-434.

THOMPSON C., SMARR J., NGUYEN H., MANNING C. (2003). Finding educational resources on the web : Exploiting automatic extraction of metadata. *Proc. ECML, Workshop on Adaptive Text Extraction and Mining*.

WESTERHOUT E., MONACHESI P. (2008). Creating glossaries using pattern-based and machine learning techniques. Dans *Proceedings of Map of Language Resources, Technologies and Evaluation*.

Des outils de TAL en support aux experts de sûreté industrielle pour l'exploitation de bases de données de retour d'expérience

Nikola TULECHKI
CLLE-ERSS, Université de Toulouse-Le Mirail, CNRS
nikola.tulechki@univ-tlse2.fr
Conseil en Facteurs Humains
http://www.cfh-ergonomie-linguistique.com/

Résumé. Cet article présente des applications d'outils et méthodes du traitement automatique des langues (TAL) à la maîtrise du risque industriel grâce à l'analyse de données textuelles issues de volumineuses bases de retour d'expérience (REX). Il explicite d'abord le domaine de la gestion de la sûreté, ses aspects politiques et sociaux ainsi que l'activité des experts en sûreté et les besoins qu'ils expriment. Dans un deuxième temps il présente une série de techniques, comme la classification automatique de documents, le repérage de subjectivité, et le clustering, adaptées aux données REX visant à répondre à ces besoins présents et à venir, sous forme d'outils, en support à l'activité des experts.

Abstract. This article presents a series of natural language processing (NLP) techniques, applied to the domain of industrial risk management and the analysis of large collections of textual feedback data. First we describe the socio-political aspects of the risk mangement domain, the activity of the investigators working with this data. We then present present applications of NLP techniques like automatic text classification, clustering and opinion extraction, responding to different needs stated by the investigators.

Mots-clés : REX, rapport d'incident, risque, sûreté industrielle, signaux faibles, classification automatique, clustering, recherche d'information, similarité, subjectivité.

Keywords: risk management, incident report, industrial safety, weak signals, automatic classification, information retrieval, similarity, clustering, subjectivity.

1 Introduction

Dans toute industrie hautement technologique, un incident peut avoir des conséquences désastreuses, provoquer des pertes matérielles considérables, des dégâts environnementaux ou, pire, coûter des vies humaines. La complexité de chaque opération, leurs intrications et la multiplicité des facteurs différents intervenant dans le fonctionnement de ces industries rendent les risques toujours présents et amènent les acteurs (opérateurs) à développer des stratégies de gestion de la sûreté des opérations. Avoir une vision d'ensemble sur l'état du système à tout moment est crucial pour toute démarche de maîtrise du risque et, lorsqu'il est question de macrosystèmes techniques de l'échelle d'une compagnie aérienne ou pétrolière, d'une centrale nucléaire, ou encore, à un niveau supérieur, d'un *secteur d'activité* tel que le transport aérien, acquérir des informations venant du plus près des opérations devient une tâche fondamentale et très difficile. Les politiques de retour d'expérience (REX) mises en place dans les secteurs à risque témoignent de ce besoin vital. Les REX visent précisément ce recueil systématique d'information, le plus souvent sous forme de compte rendus écrits, et sa (plus ou moins) libre transmission à toute la hiérarchie organisationnelle.

Une fois recueilli, le REX doit être correctement exploité, afin d'identifier les sources de risques. Ceci est le rôle des experts en sûreté, nos principaux interlocuteurs dans le cadre de cette recherche. Leur travail consiste à analyser des événements anormaux survenus dans un secteur d'activité donné et relatés, des incidents, quasi-accidents et accidents et, en se basant sur ces événements d'émettre des recommandations adéquates, afin que ces mêmes événements ne se reproduisent plus dans le futur. Or souvent, compte tenu de l'échelle des opérations, des politiques de recueil de REX de plus en plus développées et de la multiplications des canaux de partage d'informations liés à la sûreté entre institutions, les experts se trouvent face à une quantité de données hétérogènes qui deviennent difficilement maîtrisables de façon traditionnelle (codage manuel et statistiques classiques).

De plus, actuellement nous assistons à une évolution dans le concept même de gestion de la sûreté ; les acteurs sont incités à adopter une stratégie *proactive*, autrement dit à s'affranchir de l'analyse *a posteriori* (post accidentelle) et à identifier des risques latents avant qu'ils ne mènent à un accident majeur. Cette démarche de prévention met l'accent sur l'importance des événements mineurs qui peuvent contenir des indications sur une catastrophe à venir. « On le savait. C'était dans nos bases. Nous sommes juste passés à coté » entend-on dire les experts,le plus souvent sous anonymat, à la suite d'un drame industriel.

Le but de nos recherches, associant ergonomie et traitement automatique des langues (TAL) est donc de proposer des outils permettant d'abord un accès facilité aux contenus des bases de REX et, dans un deuxième temps des méthodes automatiques d'identification de risques émergents et de précurseurs de situations à risque. Ce projet pluridisciplinaire doit donc dans un premier temps identifier les besoins précis exprimés par les experts en sûreté, expliciter le contexte dans lequel s'inscrit leur activité, notamment les flux d'information et les contraintes politiques et sociales qui lui sont associés. Dans un deuxième temps, ces besoins seront traduits en une série de propositions opérationnels, des méthodes d'analyse automatique, ainsi que des traitements et algorithmes. À terme l'aboutissement sera une série d'outils destinés à venir en support à l'analyse de bases de REX dans une perspective d'une meilleure maîtrise du risque.

Cet article est organisé comme suit : Dans un premier temps nous ferons un tour rapide sur le concept de risque industriel en nous concentrant notamment sur les dernières évolutions dans le domaine qui placent de plus en plus l'accent sur le rôle de l'organisation dans son ensemble. Parallèlement nous mentionnerons les évolutions politiques et sociales, intervenues récemment dans certains secteurs d'activité, qui ont un impact direct sur la nature de notre objet d'étude, le REX.

Dans un deuxième temps nous décrirons le travail des experts en sûreté et leur rapport avec l'information du REX. Ayant ainsi établi le contexte général nous allons nous tourner vers les sciences de gestion et le concept de *signal faible* que nous adapterons à notre problématique.

Dans la deuxième partie de cet article, nous présenterons un éventail de méthodes et techniques de TAL, que nous adaptons à notre matériau textuel et aux besoins exprimés. Ces recherches, venant tout juste de commencer sont encore pour la plupart à un stade inachevé et fortement exploratoires, stade où nous cherchons encore à valider la pertinence des techniques par rapport aux besoins des experts. Nous commencerons par l'activité la plus aboutie à ce jour - la catégorisation automatique d'évènements. Ensuite nous présenterons l'approche d'*analyse de similarité*, encore en travaux, mais dont les premiers résultats sont encourageants. Dans un troisième temps nous développerons les pistes que nous explorons actuellement visant à exploiter davantage la notion de *similarité*

en l'associant à la fois à des méthodes de détection d'anomalie afin de repérer des événements anormaux ainsi qu'à des techniques de *clustering* afin de procéder à des regroupements d'événements similaires que nous pouvons caractériser de différents manières en fonction de leur comportement dans le temps. Enfin nous explorerons un axe de recherche différent, qui consiste à effectuer des analyses linguistiques fines sur le contenu textuel afin de repérer des variations stylistiques, afin de repérer les états émotionnels des rédacteurs de ces documents.

Chacune de ces techniques fera l'objet de publications détaillées dans le futur. De plus, étant donné la nouveauté du domaine, et le manque de protocoles d'évaluation adéquates ou de standards préétablis (contrairement aux domaines « classiques », comme le RI ou EI) nous ne sommes pas encore en mesure de proposer une évaluation chiffrée dans cet article, dont la vocation est avant tout d'introduire la problématique générale de nos recherches. Une thèse est en cours depuis le mois de janvier 2011 au laboratoire CLLE-ERSS à l'Université de Toulouse 2 - Le Mirail en étroite collaboration avec la société de conseil en ergonomie industrielle « Conseil en Facteurs Humains ».

2 REX et sûreté industrielle

2.1 Fondements du REX

Aujourd'hui, il existe un consensus total sur la nécessité de tirer des leçons d'événements passés, de dysfonctionnements, accidents, incidents ou tout autres écarts au fonctionnement normal. Le REX, que nous pouvons définir comme « toute formalisation d'un événement passé » remplit ce rôle de vecteur d'informations. A une petite échelle, lorsque peu d'acteurs sont impliqués ce processus est trivial [1], mais à l'échelle d'un macrosystème technique, tel que, par exemple, l'aviation civile au niveau européen, impliquant des centaines de milliers d'individus, des centaines de compagnies aériennes et une vingtaine de gouvernements, utiliser et faire circuler l'information devient une entreprise monumentale, mais nécessaire si l'on veut prendre en considération tous les facteurs pouvant intervenir dans la gestion du risque et adopter une approche globale envers sa maîtrise. La figure n° 1, extraite de (Rasmussen, 1997) illustre la complexité d'un macrosystème technique à risque et la multitude de forces, ayant un impact sur la sûreté, impliquées à différents niveaux, allant des opérateurs interagissant avec des machines (techniciens, pilotes etc..) passant par les syndicats, le management, les différents organismes régulateurs jusqu'aux gouvernements.

Particulièrement intéressants pour nous sont les flux d'information dans cette hiérarchie. L'un, évident, est le flux « descendant » ; législation, recommandations, formations et manuels d'utilisation visent à contraindre les opérateurs à un comportement standardisé, réputé « plus sûr » afin d'améliorer le niveau global de sûreté du système.

Le flux inverse encore moins évident. Faire remonter des informations du terrain jusqu'aux instances régulatrices nécessite un ensemble de mesures, des méthodes et un cadre juridique adéquat. Le secteur aéronautique est pionnier dans cette politique du REX global, grâce à la réglementation obligeant son recueil systématique et le partage avec des instances régulatrices au niveau national, tout comme au niveau européen. Il est utile de noter le conflit majeur suscité par le REX : la tension entre sécurité et responsabilité. Afin que le REX soit efficace, on doit pouvoir faire part des erreurs commises lors des opérations. Or « avouer » une erreur remet en cause l'opérateur qui l'a commise et peut l'exposer, dans des organisations « traditionnelles » à des sanctions éventuelles. Nous laisserons de coté le débat actuel sur le statut de « l'erreur humaine » [2] vis-à-vis de la sécurité, pour dire qu'afin de réduire le silence généré par la crainte de sanctions, et améliorer la qualité du REX, de nombreuses industries ont mis en place des politiques de *non-punition* et/ou d'anonymisation afin de favoriser le REX volontaire de la part des opérateurs. Ce nouveau canal d'information est amplement utilisé dans le secteur aéronautique.

1. Prenons un exemple de tous les jours : Une famille acquiert une nouvelle friteuse (nouvel équipement). Lors de la première utilisation, le mari (opérateur), n'ayant pas lu le manuel d'utilisation, introduit brusquement les pommes de terre fraîchement coupées dans l'huile très chaude (opération). L'accident survient immédiatement. L'huile bout et s'échappe de l'appareil (événements redouté). Après avoir nettoyé sa cuisine (récupération), il fait part de son expérience aux autres membres de la famille (REX), en les incitant à ne pas introduire les pommes de terre rapidement dans l'huile trop chaude (recommandation).

2. Sinon pour dire que le consensus est qu'il n'existe pas d'activité humaine sans erreurs, la plupart du temps récupérées par l'opérateur lui même, ces collègues ou des automates.

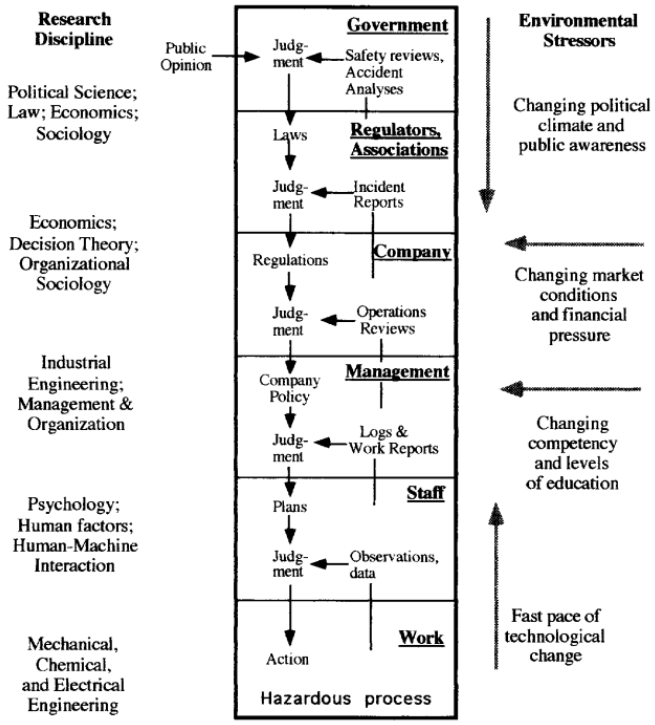


FIGURE 1 – Composantes d'une activité humaine intervenant dans le fonctionnement d'un processus dangereux

2.2 Utilisation du REX

Avoir accès aux informations ne signifie pas qu'elles seront automatiquement mises à contribution à l'amélioration de la sûreté. Nous somme rapidement arrivés à une étape, où la quantité d'information accessible dépasse les capacités d'analyse humaine. De plus puisque l'information provient le plus souvent de sources différentes, chacune avec sa propre culture vis-à-vis du recueil du REX, les informations peuvent être très hétérogènes du point de vue de leur format (bases de données, fichiers MS Excel, MS Word etc). Afin de servir de support d'analyse ces données doivent être converties en un format commun. Dans le cas de l'aviation, ou le récit d'un événement est accompagné d'un vaste ensemble de (méta)données relatives, par exemple au type d'appareil, aux conditions météorologiques, données géographiques etc, un effort de standardisation est en cours en Europe piloté par la Commission Européenne et visant à établir un tel format et un support logiciel pour son exploitation : l'environnement ECCAIRS [3] en est le fruit. Véritable « boite à outils » destinée à l'expert en sûreté, ECCAIRS propose en plus d'un format facilement échangeable, un ensemble de fonctionnalités comme un langage de requête, un navigateur spécialisé etc.

Un des étapes-clés lors de l'analyse d'un événement nouveau est sa *codification*, un procédé visant à attribuer à un évènement un ensemble de codes correspondant à ses différents facettes ; le type de l'événement (e.g sortie de piste, choc avec un oiseau, etc.) des facteurs d'environnement décrivant l'événement (conditions météo défavorables, panne d'un équipement, etc.), des facteurs contribuant à l'explication de l'événement (fatigue de l'opérateur, méconnaissance d'une procédure, etc.) En tout une liste [4] de plusieurs milliers de valeurs, organisées en une série de taxonomies, établies par l'organisation de l'aviation civile internationale (OACI), lors d'un effort de standardisation de l'analyse d'incidents en aéronautique. Une fois codifié un rapport peut être stocké dans une base et est réutilisable par la suite. Cependant puisque le contexte est en perpétuelle évolution, le schéma de codification doit être constamment mis à jour [5]. Des méthodes de classification automatique peuvent venir en aide à

3. European Coordination Centre for Accident Incident Reporting Systems : `http://eccairsportal.jrc.ec.europa.eu/`
4. Le schéma de codification ADREP actuelle est disponible à cette adresse : `http://www.icao.int/anb/aig/Taxonomy/`
5. Les perturbations du transport aérien en 2010 dues à l'éruption d'un volcan en Islande ont naturellement amenés au rajout d'une série

ces procédés de codification (voir infra).

2.3 Visée proactive, contexte dynamique et signaux faibles

Lors des investigations de véritables accidents, les experts cherchent à identifier les causes dites « primaires » de ces derniers. Il s'agit dans la plupart des cas d'une configuration particulière d'événements ou d'états, souvent clairement identifiables et signalés bien avant que l'accident lui même se produise [6].

Partant de ce constat, au cours de la dernière décennie, d'importants efforts ont été faits afin de dépasser la gestion du risque *a posteriori* et de se positionner dans une véritable démarche « proactive ». L'attention est portée non pas sur un accident qui s'est produit, mais un état des choses potentiellement dangereux, une catastrophe future que l'on peut éviter à la lumière des informations que nous avons aujourd'hui. Ainsi, les experts en sécurité sont de plus en plus sollicités pour traiter des gros volumes d'informations relatives à la sécurité, traitant de faits pouvant paraître peu importants chacun isolé et en se basant sur leur connaissance du domaine, faire des rapprochements entre ces faits et déceler des risques cachés. Parallèlement, les acteurs (pilotes, mécaniciens, opérateurs, etc.) sont incités à signaler tout évènement anormal relatif à la sécurité, ainsi qu'à s'exprimer dès qu'ils jugent qu'il y a un risque quelconque. Ce basculement vers une visée proactive dans la gestion du risque amène donc d'une part une augmentation importante du volume d'information disponible [7], puisque des faits de moins en moins éloignés de la norme prescrite sont signalés (Macrae, 2010) et d'autre part un changement du statut de cette information vis-à-vis de l'expert en sécurité, qui est amené à se concentrer sur des faits où le risque est de moins en moins explicite et, en faisant appel à son intuition et à son expériences, à chercher à aller au-delà de ce qui est réellement signalé et déceler les risques cachés.

Une telle intégration de la gestion de la sécurité dans les opérations même d'une entreprise est au cœur du modèle SGS (Systèmes de Gestion de la Sécurité) vers lequel sont amenés à s'orienter de plus en plus d'industries « à risque ». Défini comme « une façon de gérer la sécurité sous une optique commerciale » (TC, 2001), un SGS part de l'hypothèse que la sécurité au sein de l'entreprise doit devenir « l'affaire de tout le monde », amenant ainsi à une encore plus grande diversification des types d'informations relatives à la sécurité ainsi que ses sources. Ainsi l'identification d'une information importante risque de devenir de plus en plus difficile, ne serait-ce que du fait du volume et de la diversité des bases textuelles. Littéralement « noyés dans la masse » [8], l'accès à ces informations sera un générateur de frustration pour les experts en sécurité, qui doivent ne rien laisser de coté.

Les problèmes inhérents à la gestion de la sécurité dans un SGS, à savoir la diversité des sources d'information, la redondance, la nécessité absolue d'interprétation de cette information par un expert, et la disproportion entre sa fréquence et son importance, sont depuis longtemps connus des sciences de gestion. Partant d'une toute autre problématique, celle de la nécessité d'adaptation constante à un environnement commercial et concurrentiel en constante mutation, toute entreprise est amenée à mettre en place des procédés de *veille stratégique*, autrement dit d'être constamment « à l'écoute » de son environnement, pour toute information pouvant indiquer un changement futur de ce dernier.

Voulant systématiser ce processus d'écoute et d'adaptation, les chercheurs en gestion stratégique des entreprises ont forgé la notion de « signal faible » (Ansoff, 1975). Loin d'être une théorie à proprement parler, ce concept est plutôt une façon particulière et originale de voir l'information. Défini comme un « signe d'alerte précoce », le signal faible est une « information dont l'interprétation suggère qu'un événement susceptible d'être important pour l'avenir d'une firme pourrait s'amorcer » (Lesca & Blanco, 2002). L'hypothèse de base est que tout changement dans le contexte suffisamment important pour pouvoir influencer le bon fonctionnement d'une entreprise est forcement signalé bien avant qu'il ne produise des conséquences visibles par tous. De plus, dans la période

de codes en rapport avec les cendres volcaniques.

6. L'exemple de l'explosion d'une colonne dans la raffinerie de BP à Texas City au mois de mars 2005 est parlant. Ce désastre, provoquant la mort de 15 personnes, est survenu lorsqu'un nuage de vapeur, formé suite à une erreur dans la quantité de pétrole versé dans une colonne, s'est échappé et en contact avec une étincelle, s'est enflammé. Lors de l'investigation qui a suivi, six autres cas quasiment identiques, impliquant à la fois la même procédure et le même équipement, survenus au cours des dix dernières années, étaient mis au jour, aucun n'ayant tourné au cauchemar uniquement du fait de l'absence d'une source de flamme à proximité. Tous les six étaient pourtant dûment documentés, mais ce n'est que lors de l'investigation qu'un lien entre ces six occurrences est identifié.

7. Dans une grande compagnie aérienne, le nombre de nouveaux rapports d'incidents est aux alentours de 600 par mois.

8. La politique actuelle de signalement d'évènements relatifs à la sécurité actuellement mise en place préconise le signalement de *tout évènement potentiellement dangereux*. Or, on note que dans les faits, sont signalés une multitude d'évènements de routine, des « dérapages » de tous les jours, (comme par exemple des chocs avec des oiseaux au décollage pour les pilotes d'avion) qui finalement sont de peu d'intérêt pour l'expert en sécurité.

relativement longue, entre le moment du premier signalement d'un changement et le jour où ce changement devient réalité au point de menacer l'activité de l'entreprise, on peut observer une amplification de l'intensité du signalement. Inversement, la marge de manœuvre dont dispose l'entreprise diminue au fur et à mesure de cette période [9].

2.4 Caractéristiques des donnés

Les données sur lesquelles nous travaillons proviennent de bases de données différentes mises à notre disposition par des instances régulatrices de l'aviation civile, nationales et européennes, ainsi que par divers industriels dans des secteurs à risque (transports, industrie chimique, etc.). A l'heure actuelle notre corpus contient plusieurs dizaines de milliers de documents, écrits en anglais et en français et croît constamment.

La plupart des documents sont écrits dans un langage très technique, propre au secteur d'activité. Abondant d'acronymes, de termes techniques, de chiffres, de mesures, ces textes présentent, en règle générale des caractéristiques comme une variation lexicale relativement faible, peu de polysémie et une absence de constructions syntaxiques élaborés. Il s'agit de documents courts, la plupart ne dépassent pas les 500 mots.

Nous sommes en train de développer une grille de catégorisation fine de ces textes, qui sera présentée dans une prochaine publication.

3 Analyses automatiques de bases de REX

Dans cette deuxième partie nous présenterons quelques différents applications de méthodes issues du TAL aux données REX.

3.1 Catégorisation automatique d'événements

Nous avons vu que la *codification* des événements est une étape cruciale de leur analyse et permet leur réutilisabilité par la suite. Or, dans la réalité, cette tâche est effectuée de manière insatisfaisante pour plusieurs raisons. Vu la complexité des schémas de codification, contenant plusieurs centaines de classes, les codeurs attribuent souvent la classe la plus probable sans véritablement rentrer dans les détail. Les efforts de standardisation des bases de REX étant relativement récents, nous disposons de vastes quantités d'événements passés qui n'ont jamais été codés, mais qui présentent un intérêt pour des campagnes d'analyse d'aujourd'hui. Parallèlement, puisque les schémas de codification évoluent, et ce en règle générale après qu'un certain nombre d'événements de type nouveau soient survenus, de façon à justifier cette évolution, il est nécessaire de les identifier à posteriori et de leur attribuer les nouveaux codes.

Afin de répondre à ce besoin nous employons des techniques d'apprentissage automatique supervisé et plus précisément de classification automatique. Cette tâche consiste à attribuer automatiquement une classe à un individu en se basant sur les valeurs d'un ensemble de variables. Dans notre cas l'individu est un texte à classer, les variables sont les fréquences des termes dans le texte et la classe à prédire est le code de la taxonomie ADREP (voir supra). Voici le processus en détail. Nous partons d'un ensemble suffisamment large de documents déjà codifiés. Compte tenu des spécificités des textes, comme par exemple l'abondance de mesures et de noms géographiques, nous appliquerons une série de prétraitements qui réduisent ces termes à des *tokens* génériques (*mesure*, *pays*, etc). Ensuite nous procédons à une analyse morphosyntaxique en utilisant l'analyseur TreeTagger [10], suivie d'une analyse syntaxique en dépendances. Enfin en se basant sur la structure syntaxique, nous effectuons une extraction de séquences de mots en suivant les liens syntaxiques.

9. Illustrons ce propos par un exemple : prenons une entreprise spécialisée dans la fabrication de cassettes audio vierges. L'avancement de la technologie et plus précisément l'invention du CD-ROM, rend son produit obsolète et l'oblige à s'adapter en conséquence. Or le fait que le CD-ROM deviendra le support de référence est signalé bien avant que ceci ne devienne réalité. Au début, on peut imaginer des publications scientifiques qui décrivent la possibilité de stockage d'information sur un support optique. Ensuite un brevet est déposé pour ce nouveau support. Encore plus tard on commence à repérer des publications dans la presse spécialisée parlant d'un nouveau support qui vient d'être inventé, suivies de publications dans les médias généralistes, et ainsi de suite. Une progression dans la visibilité du signal est clairement perceptible et l'entreprise doit en tenir compte afin d'éviter toute surprise.

10. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

Ainsi la phrase :

```
Vers 200ft, déviation de l avion vers la gauche de la piste puis retour
                    sensible, à 200ft alarme autoland.
```

Est représentée par des séquences comme :

```
<mesure>, avion, gauche, piste, déviation, déviation avion, déviation
gauche, déviation avion gauche, déviation piste, déviation avion piste
```

Un classificateur est ensuite entraîné sur la base des corrélations entre les séquences extraites et les catégories à suggérer (voir (Hermann *et al.*, 2008) pour une explication an détail) qu'ont ces termes à apparaître dans un rapport codé dans une classe donnée.

Cette activité de codification automatique est, à l'heure actuelle opérationnelle dans le cadre de la collaboration de la société CFH avec les organismes régulateurs nationaux et européens.

3.2 Analyses de similarité et paramètre temporel

Une autre piste de recherche que nous avons entreprise, en utilisant de méthodes issues de la recherche d'information (RI), consiste à identifier automatiquement des événements similaires et d'étudier leur comportement dans le temps.

Les exemples ci dessus sont issus d'une base d'analyses d'accidents aéronautiques, de 1943 à nos jours, accessible au public [11]. Après un tri sur la longueur des textes, afin de ne pas inclure des rapports sans texte ou avec très peu de contenu, le corpus contient environ 14000 documents écrits en anglais.

La première étape est de calculer automatiquement un *score de similarité* pour une paire de documents donnée. En se basant sur les termes que les documents partagent, nous utilisons la similarité cosinus, métrique classique en RI pour attribuer un score compris entre 0 et 1 à chaque paire de documents dans la collection. Un score de 0 signifie une absence de termes en commun et un score de 1 - une identité complète. Ce score est obtenu en calculant le cosinus entre deux vecteurs dans un espace à *n* dimensions déterminées par le nombre de termes dans la collection. Chaque document est représenté par un vecteur en fonction des termes qu'il contient. Voici le processus en détail :

D'abord, afin de réduire la variation morphologique, nous procédons à une lemmatisation par Tree Tagger. Dans un deuxième temps nous construisons un espace termes en prenant les lemmes des noms, adjectifs et verbes contenus dans le texte. Ensuite nous construisons une matrice terme/document ou chaque ligne est un vecteur correspondant à un document de la collection dont les composants sont les importances dans son contenu de ses termes, calculées en utilisant la méthode de pondération TF/IDF (Jones, 2004). Enfin nous calculons le cosinus entre les deux vecteurs documents *A* et *B* en avec leur produit scalaire et leur norme.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

Afin d'explorer le comportement dans le temps des ensembles de documents similaires, nous avons mis en place un outil [12] qui projette ces documents sur un axe temporel. La copie d'écran ci dessous s'interprète de la manière suivante. Chaque point représente un document par rapport à un document source. les documents sont ordonnés chronologiquement sur l'abscisse et classés par similarité sur l'ordonnée. Un seuil de (arbitrairement fixé à 0.1) est appliqué afin de ne pas surcharger le graphique. Plus un point est à droite, plus il est récent et plus il est haut, plus il est similaire au document source. Ici est représenté l'ensemble d'évènements similaires au document source suivant, datant du 06/01/2003.

11. L'Aviation Safety Network, disponible à l'adresse suivante :`http://aviation-safety.net/` collecte les rapports traitant d'accidents aéronautiques sérieux.

12. Une démonstration de cet outil sur des données d'incidents aéronautiques est disponible à l'adresse suivante :`http://slow-start.org/safetyDataDemos/timePlotASN/main.cgi`

The captain's failure to attain a proper touchdown on runway, and his subsequent failure to perform a go-around, both of which resulted in a runway overrun. Factors were the company's inadequate dispatch procedures with their failure to provide all NOTAMS for the airport to the flight crew, and the snow covered runway

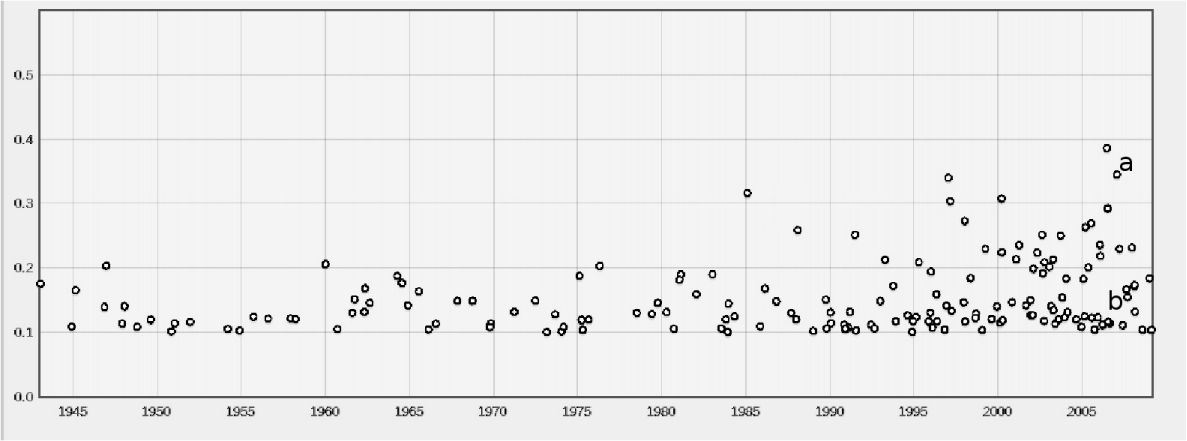


FIGURE 2 – Événements similaires sur un axe temporelle

Comparons deux documents au document source, l'un (datant du 24/01/2007, indiqué par a) relativement similaire (0.35) :

(en gras, nous avons mis les mots partagés) :

The copilot's **failure** to maintain the **proper** airspeed, and **failure** to obtain the **proper touchdown** point, and the pilot-in-command's **inadequate** supervision, which **resulted** in an **overrun.** Contributing to the accident was the PIC's **failure** to activate the speed brake upon **touchdown** and the **snow** contaminated **runway**

et le deuxième (datant du 14/09/2007, indiqué par b) un peu moins (0.15) :

The pilot's **failure** to initiate a missed **approach** and his **failure** to obtain the **proper touchdown** point while landing in the rain. Contributing to the accident were the operator's lack of standard operating **procedures** and the **inadequate** maintenance of the windshield

Nous pouvons voir comment la notion de similarité traduit un degré de ressemblance entre les deux événements ; comme le document source, les deux traitent d'atterrissage ratés, mais uniquement l'événement plus similaire mentionne des pistes enneigées.

Une entrave potentielle à l'utilisabilité de cette approche par similarité est le besoin de faire une *requête*, autrement dit de sélectionner un document,une manifestation du paradoxe de Méno [13], dont les experts en sûreté, craignant de biaiser leur analyse en faisant des présupposés, nous ont fait part. De ce fait nous envisageons de faire évoluer ces approches en employant des techniques d'apprentissage non supervisé, afin de faire émerger des regroupements naturels d'événements similaires. Ces techniques, dites de *clustering* (voir (Srivastava & Zane-Ulman, 2005) pour un exemple sur des données textuelles de type REX) permettent de s'affranchir de la requête basée sur le contenu des documents regroupés comme mode d'accès et ouvrent la voie à établir d'autres types de requêtes. Une piste que nous envisageons, peu explorée jusqu'à maintenant est celle du *profilage chronologique* (voir (Matthews *et al.*, 2010) pour un exemple d'un outil semblable destiné à des archives de presse). Une fois obtenus, les *clusters* de

13. Meno demande à Socrate comment quelqu'un peut rechercher quelque chose quand il n'a aucune idée de ce qu'est cette chose.

documents sont projetés sur un axe temporel et leur distribution chronologique est calculée. Si l'on considère les textes de l'exemple ci dessus comme les membres d'un *cluster*, ce *cluster* aurait un profil émergeant, puisque les documents récents sont beaucoup plus fréquents que les documents anciens et est susceptible d'indiquer un risque nouveau prenant de l'ampleur. A l'inverse la fréquence diminuant dans le temps d'un groupe de documents similaires, peut indiquer l'efficacité d'une recommandation nouvellement émise [14]. D'autres profils chronologiques sont aussi intéressants, comme des événements survenant à des rythmes particuliers (hebdomadaire, mensuel, annuel), par exemple.

3.3 Détection d'anomalie et événements anormaux

Les bases de REX contiennent un grand nombre d'événements similaires. Quotidiennement les avions heurtent des volatiles et ratent des atterrissages à cause d'un vent latéral. Cependant un petit nombre d'événements anormaux jamais vus jusqu'ici surviennent et il est possible de les identifier de manière entièrement automatique en utilisant des techniques de *repérage d'anomalie*, un ensemble de techniques statistiques, entièrement basées sur les données, visant à identifier dans une population les individus exceptionnels, bizarres, les *outliers* (Chandola *et al.*, 2009).

Puisqu'il s'agit de méthodes quantitatives, la principale difficulté porte sur la transformation des donnés textuelles (symboliques et qualitatives) en une série de scores numériques. Nous nous baserons pour cela sur les travaux en recherche d'information, domaine qui rencontre les mêmes difficultés. Parmi eux, certains comme (Arampatzis *et al.*, 2000) proposent des méthodes automatiques linguistiquement motivées qui visent à maîtriser la variation inhérente au langage naturel (variation lexicale, morphologique, syntaxique voire sémantique) afin d'atteindre un niveau supérieur d'abstraction, et par conséquent de produire des scores de similarité plus pertinents, scores précisément sur lesquels se basent la majorité des techniques de détection d'anomalies. Lors de quelques expériences autour du *clustering* (classification hiérarchique ascendante (CHA) (McQuitty, 1966), plus précisément), inspirés des travaux de (Ah-Pine *et al.*, 2005) nous avons pu valider l'intérêt et la faisabilité de ces techniques pour le repérage des événements « anormaux ».

Voici une esquisse de cette méthode. Partant d'une matrice de similarité, un algorithme regroupe progressivement les documents les plus similaires pour former une hiérarchie de partitions binaires incluses les unes dans les autres. Plus on monte dans la hiérarchie, plus les regroupements sont générales, plus on descend plus le critère implicite de regroupement est spécifique. Les événements anormaux, n'étant, par définition, pas similaires avec les autres, ont une tendance de former des branches hautes dans la hiérarchie (près de la racine). Nous avons expérimenté avec cette méthode utilisant un sous corpus de 110 documents, choisis manuellement afin de simuler une base avec une répartition inégale entre beaucoup de documents traitant d'événements semblables et peu de documents traitant d'événements variés. Ce corpus de test comprend 50 documents traitant des *collisions avec des oiseaux*, 50 traitant des *remises de gaz* [15] et 10 documents divers, pris au hasard. Nous avons calculé leur matrice de similarité en utilisant la méthode décrite ci-dessus avant de procéder à une CHA avec la fonction `hclust` de l'environnement d'analyse statistique *R* [16]. La figure n° 3 représente le dendrogramme produit par la CHA.

Nous nous intéresserons en particulier aux regroupements se situant haut dans la hiérarchie. Les clusters `f` et `g` contiennent la majorité des documents traitant respectivement de *collisions avec des oiseaux* et de *remises de gaz*. Plus près de la racine, les clusters `a`, `b` et `c`, contiennent les 10 documents choisies au hasard, c'est à dire les événements anormaux que nous cherchons à faire émerger. Le cluster `e` est aussi intéressant, car il contient six documents traitant à la fois de *collisions avec des oiseaux* et de *remises de gaz*, autrement dit d'événements combinant plusieurs facteurs et par ce fait intéressants pour une analyse approfondie. Cette expérience à petite échelle validant la faisabilité de la méthode, nous sommes actuellement en train d'explorer davantage cette voie en vue d'un passage à l'échelle en traitant des bases entières.

14. Nous avons récemment rencontré ce cas de figure, lors de la démonstration de ces outils aux industrielles dans une grande usine chimique. On voyait clairement les événements concernant le incidents dus aux projections reçues dans les yeux en forte baisse depuis 2007, ce qui, d'après notre interlocuteur reflétait l'effet positif de la campagne de sensibilisation au port de lunettes de protection, entreprise cette même année.

15. Une remise de gaz est une procédure d'urgence très courante lors de laquelle les pilotes décident au dernier moment d'avorter un atterrissage et de refaire un tour de l'aérodrome et une deuxième tentative d'atterrissage.

16. `http://www.r-project.org/`

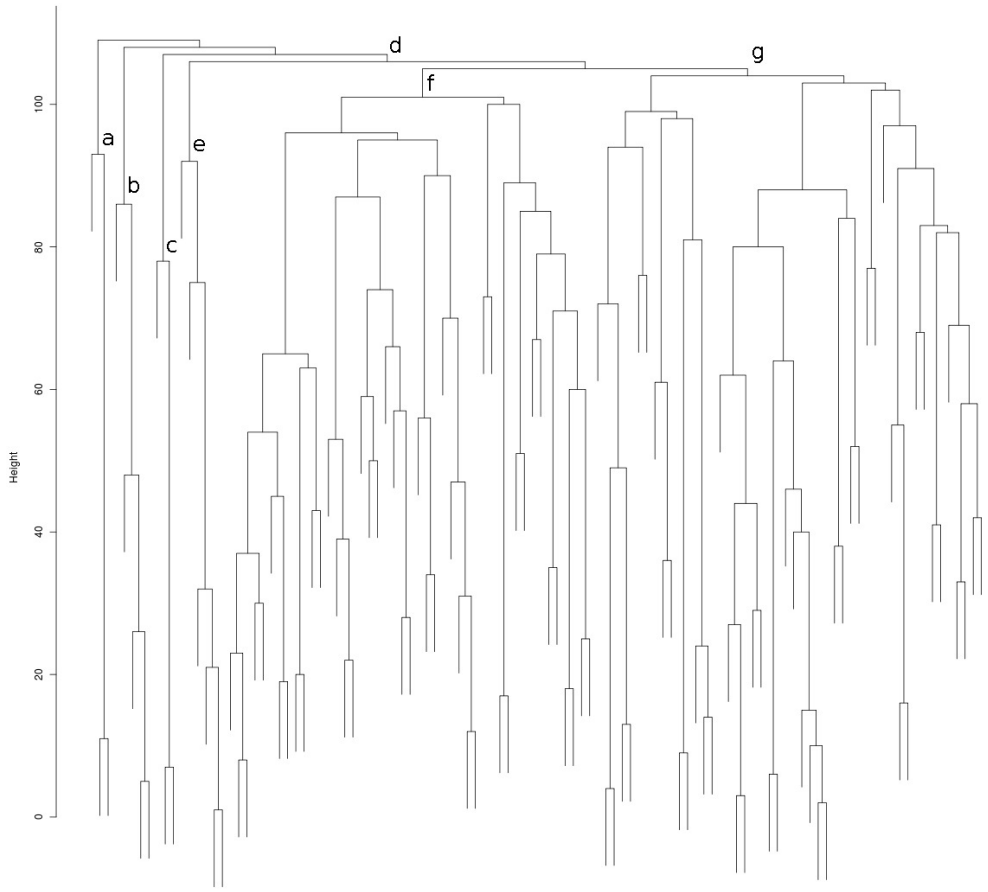


FIGURE 3 – Dendrograme des résultats de la CHA

3.4 Analyses linguistiques fines et précurseurs de situations à risque

Parallèlement aux méthodes basées sur l'ensemble des données, mentionnées ci-dessus, nous nous sommes intéressé à une autre particularité de certains donnés REX, ceux rédigés par les opérateurs eux mêmes ; le statut particulier de la place de l'auteur dans le texte. Nous avons remarqué que certains textes étaient plus « émotionnellement chargés » que d'autres. Au-delà de produire un simple récit neutre d'un événement, certains auteurs expriment des états émotionnels tels que le stress, le doute, la colère et la peur. Clairement identifiables d'un point de vue linguistique, ces états sont de véritables indicateurs de situations potentiellement à risque [17]. De plus, ayant identifié un risque récurrent et frustrés par sa non prise en compte, les acteurs manifestent souvent leur mécontentement dans leurs écrits.

Actuellement, en particulier grâce au développement du web 2.0, les travaux sur le thème du repérage automatique d'opinion et d'états émotionnels connaissent un développement spectaculaire et de nombreuses techniques innovantes voient le jour (Pang & Lee, 2008).

S'inspirant de certains travaux sur la subjectivité, nous employons une variété de traits lexicaux (adverbes axiologiques, pronoms à la première personne etc..) syntaxiques (emploi du conditionnel) et typographiques [18] que nous

17. Les pilotes, par exemple, font parfois part d'un doute ou encore d'une incompréhension d'une situation dans laquelle ils se sont trouvés. Or la maîtrise totale de la situation par les pilotes est d'une importance capitale pour la sécurité du vol et toute source de doute est à prendre au sérieux, car elle peut indiquer soit une lacune dans la formation soit un problème d'ordre organisationnel.

18. En étudiant les textes en question nous avons remarqué des pratiques tels que l'usage des majuscules ou des répétitions de point

projetons sur les textes afin de les classer par *degré de subjectivité*. Les deux textes suivants, de la même longueur et issues de la même base textuelle, illustrent cet axe, le premier faisant part de jugements personnels et écrit à la première personne, contraste fortement avec le second, beaucoup plus technique et impersonnel.

Rapport d'incident exprimant un niveau de subjectivité élevé :

LIAISON CASQUE ASSISTANT **DEFICIENTE** . [REPORT] . A l'arrivée , l'assistant est **inaudible** , **je** lui demande de changer de casque avant le départ dans 1h30 . Lors des pleins , il est toujours **inaudible** , le mécano X, présent au poste **est étonné** car à l'arrivée c'était bon . **Je** lui explique que ce n'est pas le cas et que **j'** avais déjà demandé l'échange . Liaison parfaite avec mécano X pour le litrage **faisant penser** à un autre casque . Au départ , de nouveau **inaudible** , alors que **j'** informe l'assistant de la situation , il lève l'avion (le BEACON est sur OFF **!!**) . Il demande s'il **peut pousser** . **Je** redemande un changement de casque . Attente 7 ' , rien n'est fait , avec une réception à 1/ 5 , nous poussons . A la fin **je** suis obligé de demander 3 fois qu'il se débranche et me fasse signe . **Manifestement** , la liaison est **défaillante** dans les deux sens , alors que le casque du mécano X fonctionnait parfaitement . -FIN- . [ASR] . En cas de **problème** lors du P/ B , l'équipage n'a aucune chance de comprendre ce qu'il se passe **!!! Pourquoi** cette inertie : on se contente de me dire « yes , ok , ... « et rien ne se passe ... **Faut -il** un **accident** pour que l'escale de Y se conforme au référentiel . Quant à la **qualité** de matériels ...

Cet exemple montre plusieurs indices reflétant un état émotionnel du rédacteur, sur lesquels nous nous basons pour classer ce document comme étant *subjectif* :
– emploi de la première personne
– emploi de constructions à verbe modal (peut pousser)
– emploi de mots évaluatifs négatifs [19] (défaillante, inaudible, problème, qualité, etc.)
– emploi de certains verbes reflétant des *états cognitifs* (penser, étonner)
– emploi de signes de ponctuation répétés (!!, ...)
Cette catégorisation unidimensionnelle, sur l'axe subjectif/objectif, n'est que la première étape de cette facette de nos recherches. Par la suite nous chercherons à établir un schéma de catégorisation plus fine et être capables de repérer séparément des états tels que le stress, le doute, mais aussi des cas de figure comme des erreurs de compréhension, de perception ou des lacunes dans les compétences mentionnés par les rédacteurs.

4 Conclusion

Nous venons de présenter nos travaux sur l'analyse automatique de bases de REX. Le domaine de la sûreté industrielle, l'analyse et l'exploitation des données textuelles issues de ces bases représentent un champ, qui, à notre connaissance, n'a jamais bénéficié de solutions utilisant des méthodes du TAL, méthodes que, nous venons de démontrer, peuvent réponde à une série de besoins exprimés dans ce secteur. De plus, nous sommes convaincus que, compte tenu de la dynamique actuelle, incitant d'un coté l'accroissement de la production de données tout autant que leur partage entre institutions les besoins d'outils spécialement adaptés se sentiront davantage.

Pour le TAL, un nombre de nouveaux défis se présentent. Un large éventail de techniques déjà connues devra être adapté à ce nouveau matériau bien particulier. Des aspects comme le flux constant de nouveaux documents et le langage très spécialisé, mais aucunement contraint, dans lequel sont rédigés la plupart d'eux, doivent être pris en compte. Enfin, le paramètre temporel, étant essentiel pour la maîtrise du risque dans un contexte dynamique, doit également occuper une place centrale dans toute approche visant à automatiser une partie de ce processus.

d'exclamation ou d'interrogation.

19. nous utilisons le « lexique de l'évaluation », développé par l'équipe TALN au laboratoire LINA à l'université de Nantes, que nous avons adapté à nos besoins notamment en rajoutant certains mots comme « qualité » dont nous avons vérifié le comportement axiologique dans notre corpus.

Références

AH-PINE J., LEMOINE J. & BENHADDA H. (2005). Un nouvel outil de classification non supervisée de documents pour la découverte de connaissances et la détection de signaux faibles : Rares text$^{\mathrm{tm}}$. In *Journée sur les systèmes d'information élaborés*, Île Rousse.

ANSOFF I. (1975). Managing strategic surprise by response to weak signals. *California Management Review*, **18**(2), 21–33.

ARAMPATZIS A., VAN DER WEIDE T. P., KOSTER C. H. A. & VAN BOMMEL P. (2000). An evaluation of linguistically-motivated indexing schemes. In *Proceedings of the 22nd bcs-irsg colloquium on IR research*.

CHANDOLA V., BANERJEE A. & KUMAR V. (2009). Anomaly detection : A survey. *ACM Computing Surveys (CSUR)*, **41**(3), 15.

HERMANN E., LEBLOIS S., MAZEAU M., BOURIGAULT D., FABRE C., TRAVADEL S., DURGEAT P. & NOUVEL D. (2008). Outils de Traitement Automatique des Langues appliqués aux comptes rendus d'incidents et d'accidents. In *16e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Avignon*.

JONES K. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, **60**(5), 493–502.

LESCA H. & BLANCO S. (2002). Contribution à la capacité d'anticipation des entreprises par la sensibilisation aux signaux faibles. In *Congres International Francophone sur les PME 6eme édition*, p. 10–1.

MACRAE C. (2010). Constructing near misses : Proximity, distance and the space between. *Risk&Regulation*.

MATTHEWS M., TOLCHINSKY P., BLANCO R., ATSERIAS J., MIKA P. & ZARAGOZA H. (2010). Searching through time in the New York Times. In *HCIR Challenge 2010*.

MCQUITTY L. (1966). Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, **26**(4), 825.

PANG B. & LEE L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135.

RASMUSSEN J. (1997). Risk management in a dynamic society : a modelling problem. *Safety science*, **27**(2-3), 183–213.

SRIVASTAVA A. N. & ZANE-ULMAN B. (2005). Discovering recurring anomalies in text reports regarding complex space systems. In *Proceedings of the 2005 IEEE Aerospace Conference*.

TC (2001). An introduction to safety management systems. *Transport Canada*.

Utilisation d'un score de qualité de traduction pour le résumé multi-document cross-lingue

Stéphane Huet[1] Florian Boudin[1] Juan-Manuel Torres-Moreno[1,2,3]
(1) LIA, Université d'Avignon, France
(2) École Polytechnique de Montréal, Canada
(3) GIL-IINGEN, Universidad Nacional Autónoma de México, Mexique
{stephane.huet,florian.boudin,juan-manuel.torres}@univ-avignon.fr

Résumé. Le résumé automatique cross-lingue consiste à générer un résumé rédigé dans une langue différente de celle utilisée dans les documents sources. Dans cet article, nous proposons une approche de résumé automatique multi-document, basée sur une représentation par graphe, qui prend en compte des scores de qualité de traduction lors du processus de sélection des phrases. Nous évaluons notre méthode sur un sous-ensemble manuellement traduit des données utilisées lors de la campagne d'évaluation internationale DUC 2004. Les résultats expérimentaux indiquent que notre approche permet d'améliorer la lisibilité des résumés générés, sans pour autant dégrader leur informativité.

Abstract. Cross-language summarization is the task of generating a summary in a language different from the language of the source documents. In this paper, we propose a graph-based approach to multi-document summarization that integrates machine translation quality scores in the sentence selection process. We evaluate our method on a manually translated subset of the DUC 2004 evaluation campaign. Results indicate that our approach improves the readability of the generated summaries without degrading their informativity.

Mots-clés : Résumé cross-lingue, qualité de traduction, graphe.

Keywords: Cross-lingual summary, translation quality, graph.

1 Introduction

La multiplication des documents dans de nombreuses langues, en particulier sur le Web, a rendu nécessaire la mise au point de méthodes de recherche et d'extraction d'information cross-lingue. Le résumé automatique cross-lingue vise à donner à l'utilisateur un accès rapide à des contenus exprimés dans une ou plusieurs langues qu'il maîtrise mal ou ne connaît pas. Plus précisément, cette tâche consiste à générer un résumé dans une langue cible différente de celle utilisée dans les documents sources. Dans cette étude, nous nous intéressons au résumé automatique multi-document de l'anglais vers le français, la motivation première étant de permettre aux utilisateurs francophones d'accéder à la masse toujours croissante d'actualités disponibles à travers des sources majoritairement anglophones.

Plusieurs études récentes se sont intéressées aux modèles de graphes pour représenter l'information dans des applications de Traitement Automatique des Langues Naturelles (TALN) (Banea *et al.*, 2010). Dans ces modèles, les entités — qui peuvent être par exemple les mots, les phrases ou même les documents — sont représentées sous la forme de nœuds et les relations entre elles par des arêtes. Ce type d'approche a déjà été utilisé dans des applications TALN diverses tel que l'étiquetage en parties du discours, l'extraction d'information, l'analyse de sentiments ou le résumé automatique auquel nous nous intéressons ici.

Une méthodologie simple pour aborder le résumé automatique cross-lingue serait d'appliquer un système de traduction automatique (TA) directement sur les sorties d'un système de résumé automatique classique. Toutefois, cette approche n'est pas sans inconvénients puisqu'elle devient dépendante de la qualité du système de TA. Dans cet article, nous proposons de prendre en compte la qualité de traduction des phrases en français lors de la sélection des phrases retenues pour assembler le résumé, l'idée étant de minimiser l'impact des erreurs commises par le système de TA. Les phrases ainsi sélectionnées pour construire le résumé seront celles jugées à la fois informatives

par le système de résumé automatique et faciles à traduire par le système de TA. Pour ce faire, nous recourons à une méthode d'apprentissage supervisé pour prédire les scores de qualité de la traduction et intégrons ces scores durant la construction du graphe utilisé pour sélectionner les phrases informatives.

Dans la suite de cet article, nous commençons par présenter les travaux connexes aux nôtres. La section 3 est consacrée à la description de la méthode que nous proposons. Nous décrivons ensuite en section 4 nos résultats expérimentaux avant de conclure et de montrer quelques perspectives.

2 Travaux connexes

Dans cette section, nous présentons dans un premier temps les travaux existants sur la prédiction de la qualité de traduction automatique. Nous décrivons ensuite les approches de résumé automatique basées sur les modèles de graphes ainsi que les études sur le résumé automatique cross-lingue.

2.1 Prédiction de la qualité de traduction automatique

La traduction automatique est un composant naturel d'un système automatique de résumé cross-lingue de documents. Malheureusement, bien que des progrès importants aient été réalisés depuis une décennie, les systèmes de TA restent sujets à des erreurs qui peuvent dégrader fortement la qualité des résumés produits, en introduisant en particulier des informations erronées ou en rendant les phrases générées difficiles à lire. Afin de réduire ces effets, il est intéressant de prendre en compte un score jugeant de la qualité de la traduction pour filtrer les traductions incorrectes lors du résumé.

La prédiction de la qualité de la traduction a tout d'abord été vue comme un problème de classification binaire pour distinguer les bonnes traductions des mauvaises (Blatz *et al.*, 2003). Des études plus récentes ont estimé une valeur continue de score soit au niveau du mot (Raybaud *et al.*, 2009), soit au niveau de la phrase (Raybaud *et al.*, 2009; Specia *et al.*, 2009). Dans cet article, nous employons des scores calculés au niveau de la phrase, ceux-ci étant plus faciles à intégrer dans le processus de sélection de phrases pour le résumé.

Plusieurs classificateurs ont été construits pour estimer la qualité de traduction. Ces modèles statistiques ont été utilisés sur des traductions étiquetées manuellement comme correctes ou non (Quirk, 2004; Specia *et al.*, 2009), ou étiquetées par des métriques automatiques comme le taux d'erreur de mots (Blatz *et al.*, 2003), le score NIST (Blatz *et al.*, 2003; Specia *et al.*, 2009) ou BLEU (Raybaud *et al.*, 2009). Parmi les différentes caractéristiques utilisées pour le calcul des valeurs de qualité, on retrouve des traits linguistiques — dépendant ou non de ressources telles que des analyseurs syntaxiques ou Wordnet —, des mesures de similarité entre la phrase source et la phrase cible, et des caractéristiques internes au système de traduction utilisées — comme le nombre de traductions proposées par mots sources ou les scores de segments (phrases) des meilleures hypothèses de traduction.

2.2 Résumé automatique fondé sur les modèles de graphes

Ces dernières années, de nombreuses évaluations ont été conduites sur la tâche du résumé automatique multi-document, en particulier dans le cadre des campagnes internationales *Document Understanding Conference* [1] (DUC) et *Text Analysis Conference* [2] (TAC) organisées par le *National Institute of Standards and Technology* [3] (NIST). La quasi-totalité des approches proposées recourent à des méthodes d'extraction où il s'agit d'identifier les unités textuelles — le plus souvent des phrases — les plus importantes des documents. Les phrases contenant les concepts les plus importants sont sélectionnées puis assemblées selon leur degré de pertinence afin de générer les résumés. Ce type d'approche donne de bons résultats et permet de contourner les problématiques difficiles de compréhension sémantique du texte ou de génération de texte en langue naturelle.

Les travaux menés jusqu'à présent sur la tâche du résumé automatique multi-document sont basés, entre autres, sur l'utilisation du centroïde pour la sélection de phrases (Radev *et al.*, 2004), sur l'apprentissage supervisé des critères d'informativité (Wong *et al.*, 2008) ou sur la fusion d'information (Barzilay *et al.*, 1999). Dans cet article,

1. http://duc.nist.gov
2. http://www.nist.gov/tac/
3. http://www.nist.gov

nous employons une approche basée sur les modèles de graphes introduite dans (Mihalcea, 2004; Erkan & Radev, 2004). Les algorithmes de classement basés sur les graphes tels que PAGERANK (Page *et al.*, 1998) ont été utilisés avec succès dans les réseaux sociaux, l'analyse du nombre de citations ou l'étude de la structure du Web. Appliqué au résumé automatique, ce type d'approche suggère de représenter les documents par un graphe d'unités textuelles (phrases) inter-connectées par des relations issues de calculs de similarité. Les phrases sont ensuite sélectionnées selon des critères de centralité ou de prestige dans le graphe puis assemblées pour produire des extraits. Cette approche a deux principaux avantages. Premièrement, contrairement à la plupart des autres méthodes, elle ne nécessite pas de données d'apprentissage. Deuxièmement, du fait qu'elle se base sur des traitements linguistiques minimaux (segmentation en phrases et similarité inter-phrases), cette approche est facilement adaptable à d'autres langues (Mihalcea & Tarau, 2005).

2.3 Résumé automatique cross-lingue

Quelques études se sont récemment intéressées à la problématique du résumé automatique cross-lingue. Deux solutions simples à ce problème consistent soit à traduire les documents avant la phase d'extraction, soit à traduire les résumés générés. Cette seconde approche est généralement préférée à la première car la traduction au préalable des documents rend le processus de sélection de phrases plus risqué de part les erreurs potentiellement introduites par le système de TA. Orăsan & Chiorean (2008) ont ainsi proposé d'utiliser la méthode *Maximal Marginal Relevance* (MMR) (Carbonell & Goldstein, 1998) pour produire des résumés d'actualités exprimés en roumain et ensuite de les traduire automatiquement en anglais.

Plus récemment, Wan *et al.* (2010) se sont intéressés au résumé automatique mono-document, depuis l'anglais vers le chinois, en employant des méthodes supervisées pour estimer la qualité de traduction automatique. Leur étude a montré que la prise en compte de scores de qualité de traduction permet d'améliorer à la fois le contenu et la lisibilité des résumés générés. Dans notre article, nous utilisons une approche similaire en nous intéressant cette fois au résumé automatique multi-document. Contrairement aux travaux de Wan *et al.*, notre approche utilise un algorithme non supervisé et indépendant de la langue pour sélectionner des phrases (Mihalcea & Tarau, 2005). De plus, nous n'utilisons pas de corpus annotés manuellement selon leur qualité de traduction mais un indicateur calculé automatiquement à partir de traductions de références produites par des humains, ce type de corpus étant long et parfois délicat à construire.

3 Notre méthode pour le résumé cross-lingue

Notre approche pour résumer un ensemble de documents depuis l'anglais vers le français se fait en trois étapes. Chaque phrase est tout d'abord traduite automatiquement et la qualité de la traduction est estimée (section 3.1). Chaque phrase se trouve ensuite évaluée en fonction de son contenu informatif (section 3.2) et de son score de qualité de traduction (section 3.3). Puis, les phrases de plus haut score sont sélectionnées pour les inclure dans le résumé (section 3.4). La figure 1 présente un aperçu de l'architecture de notre méthode.

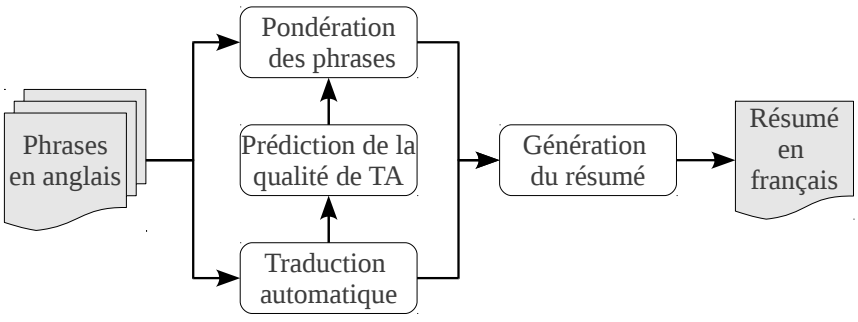


FIGURE 1 – Architecture de notre système de résumé automatique cross-lingue.

3.1 Prétraitement de documents et prédiction de la qualité de traduction

Chaque document de l'ensemble à résumer est segmenté en phrases en utilisant la méthode PUNKT de détection de changement de phrases (Kiss & Strunk, 2006) mise en œuvre dans la boîte à outils NLTK (Bird & Loper, 2004). Toutes les phrases en anglais ont été automatiquement traduites en français en utilisant le système de traduction de Google [4].

Un score de TA est calculé pour chaque phrase pour estimer la justesse et la fluidité des phrases générées en français. Pour ce faire, nous calculons pour chaque phrase 8 caractéristiques, qui donnent des informations sur la difficulté de traduction et sur la lisibilité des traductions générées :

– la longueur de la phrase source en terme de mots ;
– le ratio des longueurs des phrases source et cible ;
– le nombre de signes de ponctuation dans la phrase source ;
– la proportion des nombres et des signes de ponctuation présentes dans la phrase source qui sont retrouvées dans la phrase cible ;
– les perplexités des phrases source et cible calculées à l'aide de modèle de langue (ML) 5-grammes en avant ;
– les perplexités des phrases source et cible calculées par des ML bigrammes en arrière, *i. e.* en inversant l'ordre des mots des phrases.

Ces quatre premières caractéristiques sont parmi les traits les plus pertinents mis en exergue dans (Specia *et al.*, 2009), parmi 84 caractéristiques étudiées ; les quatre dernières ont déjà montré leur efficacité dans le calcul de mesure de confiance au niveau des mots (Raybaud *et al.*, 2009). Des ML sont construits à partir des corpus monolingues du domaine des actualités, rendus disponible pour l'atelier WMT 2010 (Callison-Burch *et al.*, 2010) et constitués respectivement de 991 et 325 millions de mots en anglais et en français. Les scores de perplexité visent à estimer la fluidité. Contrairement à d'autres études, nous nous sommes concentrés sur des caractéristiques simples ne requérant pas de ressources linguistiques comme des analyseurs syntaxiques ou des dictionnaires. En outre, nous nous sommes restreints à des scores ne dépendant par du système de traduction utilisé.

Pour prédire la qualité de la traduction, nous avons employé la méthode ϵ-SVR, qui est une extension des séparateurs à vaste marge pour faire de la régression et qui a déjà été utilisée dans le même cadre applicatif (Wan *et al.*, 2010; Raybaud *et al.*, 2009). Nous avons employé la librairie LIBSVM (Chang & Lin, 2001), en nous restreignant aux noyaux gaussiens comme recommandé par les auteurs. Le modèle de régression a deux paramètres : une erreur de coût c et le coefficient γ de la fonction noyau ; leur valeur a été optimisée par recherche par quadrillage et par validation croisée.

Le modèle ϵ-SVR devrait idéalement être appris sur un corpus étiqueté manuellement du point de vue de la qualité de traduction. Malheureusement, nous ne connaissons pas de corpus de ce genre ayant une taille suffisante pour la paire anglais-français et la production de jugements de la TA reste un processus très lent. Nous nous sommes par conséquent tournés vers un indicateur calculé automatiquement à partir de traductions de référence produites par des humains : le score NIST (Doddington, 2002). Cette métrique a en effet déjà été utilisée dans le passé dans le même objectif (Blatz *et al.*, 2003; Specia *et al.*, 2009) et s'est révélée plus corrélée avec des jugements humains au niveau de la phrase que BLEU (Blatz *et al.*, 2003). Notre corpus d'apprentissage a été obtenu à partir des traductions de référence fournies dans le domaine des actualités pour les ateliers WMT (Callison-Burch *et al.*, 2010) de 2008 à 2010, ce qui représente un ensemble de 7 112 phrases. Pour contrôler la qualité du modèle ainsi obtenu, nous avons calculé la métrique MSE (*Mean Squared Error*) : $\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$, N étant le nombre de phrases, $\hat{y}$ la prédiction estimée par le modèle et y la valeur réelle. Sur les 2 007 phrases de WMT 2007 gardées à cette fin, le MSE mesuré a été de 0,456.

3.2 Pondération des phrases

Notre système de résumé multi-document est fondé sur un graphe dirigé $G = (V, E)$ construit pour chaque ensemble de textes, V étant l'ensemble de nœuds et E les arcs (arêtes) dirigés. Un nœud est ajouté au graphe pour chaque phrase de l'ensemble de documents ; les arêtes sont définies entre ces nœuds en fonction de la mesure de similarité définie dans (Mihalcea, 2004). Cette mesure détermine le nombre de mots communs entre les représentations lexicales des deux phrases, les mots outils ayant été au préalable supprimés et les autres mots ayant été

4. http://translate.google.com

stemmés avec le *stemmeur* de Porter. Pour éviter de favoriser les phrases longues, cette valeur est normalisée par les longueurs des phrases. Si $\mathrm{freq}(w, S)$ représente la fréquence du mot w dans la phrase S, la similarité entre les phrases S_i et S_j est définie par :

$$\textit{Sim}(S_i, S_j) = \frac{\sum_{w \in S_i, S_j} \mathrm{freq}(w, S_i) + \mathrm{freq}(w, S_j)}{\log(|S_i|) + \log(|S_j|)} \quad (1)$$

Les algorithmes de classement basés sur les modèles de graphes mettent en œuvre le concept de recommandation. Les phrases sont évaluées selon des scores calculés récursivement sur l'intégralité du graphe. Dans notre étude, nous utilisons une adaptation de l'algorithme PAGERANK de Google (Page *et al.*, 1998) qui inclut les poids des arêtes :

$$p(V_i) = (1 - d) + d \times \sum_{V_j \in pred(V_i)} \frac{Sim(S_i, S_j)}{\sum_{V_k \in succ(V_i)} Sim(S_k, S_i)} \, p(V_i) \quad (2)$$

où d est un « facteur d'amortissement » (typiquement dans l'intervalle $[0.8, 0.9]$), $pred(V_i)$ représente l'ensemble des nœuds qui ont une arête en direction de V_i et *succ*(V_i) l'ensemble des nœuds connectés à V_i par une arête sortante. La méthode employée ici, décrite dans (Mihalcea, 2004), est très similaire au PAGERANK lexical, appelé LEXRANK (Erkan & Radev, 2004).

3.3 Inclusion des scores de qualité de traduction

Pour prendre en compte l'aspect cross-lingue, la mesure de similarité inter-phrases, définie dans l'équation 1 pour les phrases d'origine en anglais, est modifiée pour inclure les scores de qualité de traduction :

$$\textit{Sim}_2(S_i, S_j) = \textit{Sim}(S_i, S_j) \times \textit{Prediction}(S_i) \quad (3)$$

où *Prediction*(S_i) est le score de qualité de TA de la phrase S_i calculée comme décrit en section 3.1. Cette métrique est asymétrique, contrairement à celle définie par l'équation 1. Une phrase traduite correctement et fluide voit ainsi les poids de ses arêtes sortantes renforcés et jouera par conséquent un rôle plus central dans le graphe.

Nous avons modifié l'algorithme de classement afin de tirer profit des spécificités des documents. Comme la position d'une phrase au sein d'un document est un indicateur fort sur l'importance de son contenu — les articles de journaux présentant généralement au début une description concise du sujet — le poids des arcs sortant du nœud correspondant à la première phrase a été doublé. En outre, les phrases dupliquées ainsi que les phrases contenant moins de 5 mots ont été mises de côté.

3.4 Génération de résumé

Bien souvent, les documents regroupés sous une thématique contiennent des phrases très similaires, voire même identiques. Il est donc possible que deux phrases très redondantes se retrouvent dans un résumé, dégradant à la fois sa lisibilité et son contenu informatif. Pour pallier ce problème, Carbonell & Goldstein (1998) ont proposé la méthode d'assemblage itératif *Maximal Marginal Relevance* (MMR). Cette technique, probablement la plus utilisée, consiste à réordonner les phrases en fonction de deux critères qui sont l'importance de la phrase et la redondance par rapport aux phrases déjà sélectionnées. Le résumé est ensuite construit itérativement par l'ajout des phrases maximisant l'informativité tout en minimisant la redondance.

Dans cette étude, nous avons utilisé une approche différente. Suivant la méthode proposée dans (Mihalcea & Tarau, 2005) pour la construction de graphes, aucun arc n'est ajouté entre deux nœuds dont la similarité excède un seuil maximal. De façon à réduire la redondance, une étape supplémentaire est ajoutée lors de la génération des résumés (Genest *et al.*, 2009). Nous générons pour ce faire tous les résumés candidats à partir des combinaisons des N phrases ayant les meilleurs scores, en veillant à ce que le nombre total de caractères soit optimal (*i. e.* en dessous d'un seuil donné et qu'il soit impossible d'ajouter une autre phrase sans dépasser ce seuil). Le résumé

retenu au final est celui possédant le score global le plus élevé, ce score étant calculé comme le produit du score de la diversité du résumé — estimé par le nombre de n-grammes différents — et de la somme des scores des phrases.

Afin d'améliorer la lisibilité du résumé produit, les phrases sont triées dans l'ordre chronologique de publication des documents où ils apparaissent, ce qui maximise la cohérence temporelle ; si deux phrases sont extraites à partir d'un même document, l'ordre original du document est conservé.

4 Résultats

Cette section décrit les données utilisées, les métriques d'évaluation et les résultats de notre système.

4.1 Cadre expérimental

Nous avons employé dans notre étude les ensembles de documents mis à disposition pour l'évaluation DUC 2004, ce qui représente 50 ensembles de documents en anglais. Chaque ensemble traite d'une même thématique et comporte en moyenne 10 articles de journaux produits par *Associated Press* ou le *New York Times*. La tâche consiste à générer des résumés d'au plus 665 caractères — incluant les caractères alphanumériques, les espaces et les ponctuations — contenant l'essentiel du contenu de l'ensemble de documents correspondants. Nous avons effectué sur ces données une évaluation automatique du contenu. Une évaluation manuelle de la lisibilité a également été menée sur un échantillon constitué de 16 ensembles de documents tirés aléatoirement.

4.1.1 Évaluation automatique

La plupart des méthodes d'évaluation automatique opèrent en comparant les résumés générés avec un ou plusieurs résumés de référence. La métrique que nous avons employée ici est ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), dont on sait qu'elle est bien corrélée avec les jugements humains (Lin, 2004). ROUGE correspond en fait à plusieurs mesures, calculées à partir du nombre de n-grammes commun entre le résumé candidat et le(s) résumé(s) de référence. Nous avons calculé trois métriques au cours de nos expériences : ROUGE-1 (basée sur les unigrammes), ROUGE-2 (bigrammes) et ROUGE-SU4 (bigrammes à trou, *i. e.* des couples de deux mots contenant au plus quatre mots entre eux) [5].

Quatre résumés de référence en anglais étaient fournis pour chacun des ensembles de documents de DUC 2004. Pour évaluer notre méthode, nous avons demandé à trois annotateurs de traduire les résumés disponibles pour le sous-ensemble de 16 groupes de documents, en veillant à ce que chaque phrase du résumé soit traduite phrase par phrase sans introduire d'information supplémentaire comme la génération d'anaphores, la désambiguisation des noms propres ou la réduction des informations redondantes. 64 résumés de référence ont été ainsi produits, chaque annotateur traduisant en moyenne un résumé en 15 minutes.

L'évaluation par ROUGE étant ici réalisée dans un cadre différent des tâches habituelles — produisant des résumés en anglais à partir de documents exprimés dans la même langue —, nous avons effectué quelques modifications. Aucune contrainte stricte n'a été imposée sur la taille des résumés traduits produits en français. En revanche, nous avons fait en sorte que notre algorithme de génération construise des résumés pour lesquelles la longueur totale des phrases correspondantes en anglais respecte la contrainte imposée à DUC 2004 sur le nombre de caractères. La longueur des résumés de référence en français se trouve ainsi accrue de 25 % en moyenne par rapport aux résumés correspondants en anglais. Notons enfin que le *stemmer* de Porter utilisé dans l'évaluation ROUGE a été adapté au français.

4.1.2 Évaluation manuelle

L'évaluation de la qualité linguistique des résumés a été effectuée selon un protocole similaire à celui utilisé lors des campagnes DUC. Nous avons évalué la lisibilité des résumés sur une échelle de 1 à 5, où 5 est attribué aux résumés « faciles à lire » et 1 aux résumés « difficiles à lire ». Cinq annotateurs ont participé à cette expérience.

5. Nous avons utilisé la version 1.5.5 de ROUGE avec les paramètres par défaut indiqués pour DUC 2004.

Afin de comparer notre approche, nous avons généré deux résumés pour chaque ensemble de documents, *i. e.* pour chacune des thématiques. Le premier résumé est produit par la méthode que nous proposons tandis que le second (*baseline*) est obtenu en traduisant directement un résumé en français (obtenu par la fonction de pondération décrite en section 3.2). La tâche qui leur a été confiée était d'attribuer une note aux deux résumés d'une même thématique, l'ordre d'apparition des résumés étant aléatoire afin d'éviter tout biais.

4.2 Expériences monolingues

Les performances de notre méthode ont tout d'abord été évaluées sur une tâche de résumé monolingue. Le tableau 1 indique les scores d'évaluation automatique obtenus sur l'ensemble des données de DUC 2004 pour différentes méthodes : le plus haut score atteint lors de la campagne en 2004 (ligne 1), le score obtenu avec la méthode GRAPH-SUM décrite en section 3.2 basée sur les graphes (ligne 2) et un score calculé pour une méthode naïve prenant la première phrase des documents les plus récents de chaque ensemble à traduire. L'approche basée sur les graphes obtient de bons résultats, la différence avec le meilleur système n'étant pas statistiquement significative [6].

Système	ROUGE-1	**Rang**	ROUGE-2	**Rang**	ROUGE-SU4	**Rang**
1^{er} système	0,38244†	1	0,09218†	1	0,13323†	1
GRAPH-SUM	0,38052†	2	0,08566†	4	0,13114†	3
Méthode naïve	0,32381	26	0,06406	25	0,10291	29

TABLE 1 – Scores ROUGE moyens mesurés sur les données de DUC 2004 et rangs obtenus par rapport aux 35 participants de la campagne. Les scores indiqués par † sont statistiquement significatifs par rapport au modèle de base ($\rho < 0.001$ avec un t-test de Student).

4.3 Expériences cross-lingues

Dans cette seconde série d'expériences, nous évaluons notre approche pour le résumé automatique multi-document cross-lingue. La première partie de cette évaluation est réalisée automatiquement à l'aide des mesures ROUGE et concerne le contenu des résumés. Il s'agit d'évaluer si les résumés produits contiennent les informations les plus importantes des documents sources. Les résultats de référence sont obtenus en traduisant le résumé en anglais produit par l'approche basée sur les modèles de graphes (méthode GRAPH-SUM). Les scores ROUGE calculés avec cette méthode sont présentés à la ligne 1 du tableau 2. En utilisant notre méthode faisant intervenir la qualité des traductions automatiques (ligne 2), nous observons une légère amélioration en terme de ROUGE-2 et ROUGE-SU4. Cependant, cette évolution des scores n'est pas statistiquement significative. Ceci peut s'expliquer par le fait que notre méthode favorise les phrases dont la qualité de TA est bonne. Ainsi des phrases ayant un contenu informationnel plus faible peuvent être introduites dans le résumé, ce qui limite l'amélioration des résultats.

Système	ROUGE-1	ROUGE-2	ROUGE-SU4
Méthode de réference	0,39704	0,10249	0,13711
Notre méthode	0,39624	0,10687	0,13877

TABLE 2 – Scores ROUGE moyens calculés sur le sous-ensemble de DUC 2004 traduit en français.

La seconde partie de cette évaluation concerne la qualité linguistique des résumés générés. Il s'agit d'évaluer manuellement si les résumés produits sont lisibles mais également compréhensibles. Le tableau 3 montre les résultats de l'évaluation manuelle obtenus sur le sous-ensemble de 16 groupes de documents. Le score moyen donné par chaque annotateur est également indiqué. Tous les annotateurs ont jugé que notre méthode conduisait à une amélioration de la lisibilité des résumés produits, ce qui montre ainsi l'intérêt d'utiliser des scores de qualité de TA pour améliorer la qualité linguistique des résumés.

6. Le t-test de Student est de $\rho = 0.77$ pour ROUGE-1, $\rho = 0.17$ pour ROUGE-2 et $\rho = 0.57$ pour ROUGE-SU4.

Néanmoins, il faut noter que les scores moyens sont relativement bas. Ceci indique que les résumés générés par notre méthode, bien qu'étant meilleurs du point de vue de la lisibilité par rapport à l'approche de référence, ne sont pas encore satisfaisants. Un exemple des sorties de notre système de résumé automatique est donné en annexe. Plusieurs types d'erreurs ont été identifiées comme récurrentes. La qualité de la TA est dépendante de la difficulté de la phrase à traduire. Ainsi, par des traitements simples comme la suppression des références temporelles, la résolution des acronymes ou la normalisation des noms propres, nous espérons pouvoir réduire la difficulté des phrases sources et par conséquent réduire le nombre d'erreurs de traduction.

Annotateur	**Lisibilité**	
	Méthode de référence	Notre méthode
Annotateur 1	2,44	2,50
Annotateur 2	1,56	1,63
Annotateur 3	1,75	2,31
Annotateur 4	3,06	3,31
Annotateur 5	1,50	1,63
Moyenne	2,06	2,28

TABLE 3 – Scores moyens de lisibilité de notre méthode comparés avec une approche standard basée sur les graphes. Les scores varient selon une échelle de 1 à 5, 5 étant le plus haut score possible.

5 Conclusions et perspectives

Dans cet article, nous avons présenté une approche basée sur les modèles de graphes pour le résumé automatique multi-document cross-lingue. Nous avons proposé d'introduire des scores de qualité de traduction automatique au moment de l'étape de construction du graphe représentant les unités textuelles, un algorithme de classement par popularité étant ensuite chargé de sélectionner les phrases traduites qui sont à la fois les plus informatives mais également les plus lisibles. Cette approche a été évaluée sur un corpus de 16 ensembles de documents traduits manuellement parmi les documents mis à disposition dans le cadre de la campagne d'évaluation internationale DUC 2004. Les résultats expérimentaux montrent que notre méthode améliore sensiblement la lisibilité (*i. e.* la qualité linguistique) des résumés générés tout en maintenant un contenu informatif au niveau de l'état de l'art.

En perspectives de nos travaux, nous souhaitons dans un premier temps mener une évaluation plus complète en produisant des résumés de référence sur l'ensemble des données de la compétition DUC 2004 et en étendant l'évaluation à d'autres langues. Ceci permettra de renforcer l'importance des résultats que nous avons obtenus mais aussi d'envisager un apprentissage supervisé de la répartition entre l'informativité et à la lisibilité des phrases. Dans un deuxième temps, nous souhaitons travailler sur la réécriture des phrases sources et en étudier l'impact sur la qualité de traduction, l'idée étant de simplifier au maximum les phrases sources à l'aide de traitement linguistiques comme la résolution d'anaphores afin de faciliter le travail du système de TA. Nous souhaitons également suivre la piste de la fusion non supervisée de phrases (Filippova, 2010) afin de générer des phrases courtes et linguistiquement simples. Une dernière perspective que nous voulons étudier concerne l'utilisation de notre propre modèle de traduction automatique, ce qui permettra à la fois d'adapter ce système pour le type de documents à résumer et de prendre en compte de nouveaux indices pour prédire la qualité de la traduction.

Références

C. BANEA, A. MOSCHITTI, S. SOMASUNDARAN & F. M. ZANZOTTO, Eds. (2010). *TextGraphs-5 Workshop*. Uppsala, Suède.

BARZILAY R., MCKEOWN K. R. & ELHADAD M. (1999). Information fusion in the context of multi-document summarization. In *ACL*, College Park, MD, USA.

BIRD S. & LOPER E. (2004). NLTK : The natural language toolkit. In *ACL*, Barcelone, Espagne.

BLATZ J., FITZGERALD E., FOSTER G., GANDRABUR S., GOUTTE C., KULESZA A., SANCHIS A. & UEFFING N. (2003). *Confidence Estimation for Machine Translation*. Rapport interne, Johns Hopkins University, Batimore, MD, USA.

CALLISON-BURCH C., KOEHN P., MONZ C., PETERSON K., PRZYBOCKI M. & ZAIDAN O. (2010). Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Workshop on Statistical Machine Translation and Metrics (WMT)*, Uppsala, Suède.

CARBONELL J. & GOLDSTEIN J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, Melbourne, Australie.

CHANG C.-C. & LIN C.-J. (2001). *LIBSVM : a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

DODDINGTON G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLT*, San Diego, CA, USA.

ERKAN G. & RADEV D. (2004). LexRank : Graph-based lexical centrality as salience in text summarization. *JAIR*, **22**(1), 457–479.

FILIPPOVA K. (2010). Multi-sentence compression : Finding shortest paths in word graphs. In *Coling*, Pékin, Chine.

GENEST P., LAPALME G., NERIMA L. & WEHRLI E. (2009). A symbolic summarizer with 2 steps of sentence selection for TAC 2009. In *TAC Workshop*, Gaithersburg, MD, USA.

KISS T. & STRUNK J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, **32**(4), 485–525.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *ACL Workshop on Text Summarization Branches Out*, Barcelone, Espagne.

MIHALCEA R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *ACL*, Barcelone, Espagne.

MIHALCEA R. & TARAU P. (2005). A language independent algorithm for single and multiple document summarization. In *IJCNLP*, Jeju Island, Corée du Sud.

ORĂSAN C. & CHIOREAN O. A. (2008). Evaluation of a cross-lingual romanian-english multi-document summariser. In *LREC*.

PAGE L., BRIN S., MOTWANI R. & WINOGRAD T. (1998). *The pagerank citation ranking : Bringing order to the web*. Rapport interne, Stanford Digital Library Technologies Project.

QUIRK C. B. (2004). Training a sentence-level machine translation confidence measure. In *LREC*, Lisbonne, Portugal.

RADEV D., JING H., STY M. & TAM D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, **40**(6), 919–938.

RAYBAUD S., LANGLOIS D. & SMAÏLI K. (2009). Efficient combination of confidence measures for machine translation. In *Interspeech*, Brighton, UK.

SPECIA L., CANCEDDA N., DYMETMAN M., TURCHI M. & CRISTIANINI N. (2009). Estimating the sentence-level quality of machine translation systems. In *EAMT*, Barcelone, Espagne.

WAN X., LI H. & XIAO J. (2010). Cross-language document summarization based on machine translation quality prediction. In *ACL*, Uppsala, Suèede.

WONG K.-F., WU M. & LI W. (2008). Extractive summarization using supervised and semi-supervised learning. In *Coling*, Manchester, UK.

Annexe

Méthode de référence (Score de lisibilité moyenne de 2,6)

Leaders de l'opposition du prince Norodom Ranariddh et Sam Rainsy, invoquant des menaces de Hun Sen à l'arrestation de l'opposition, après deux tentatives présumées sur sa vie, a dit qu'ils ne pouvaient pas négocier librement au Cambodge et a appelé à des pourparlers à la résidence de Sihanouk à Pékin. (*Opposition leaders Prince Norodom Ranariddh and Sam Rainsy, citing Hun Sen's threats to arrest opposition figures after two alleged attempts on his life, said they could not negotiate freely in Cambodia and called for talks at Sihanouk's residence in Beijing.*) Le parti de Hun Sen a récemment demandé à Ranariddh pour retourner à la table des négociations et a déclaré qu'il était disposé à faire une "concession appropriées" pour sortir de l'impasse de former un gouvernement. (*Hun Sen's party recently called on Ranariddh to return to the negotiation table and said it was willing to make an "appropriate concession" to break the deadlock over forming a government.*) La semaine dernière, Hun Sen Parti du peuple cambodgien et le parti Ranariddh FUNCINPEC ont convenu de former une coalition qui laisserait Hun Sen comme Premier ministre seul et faire le prince président de l'Assemblée nationale. (*Last week, Hun Sen's Cambodian People's Party and Ranariddh's FUNCINPEC party agreed to form a coalition that would leave Hun Sen as sole prime minister and make the prince president of the National Assembly.*)

Notre méthode (Score de lisibilité moyenne de 3,2)

Le parti au pouvoir a soutenu l'action de la police dans sa déclaration, en notant que les biens publics ont été endommagés par des manifestants et que des grenades ont été lancées sur la maison de Hun Sen après Sam Rainsy a suggéré dans un discours que le gouvernement américain devrait tirer des missiles de croisière à Hun Sen. (*The ruling party supported the police action in its statement, noting that public property was damaged by protesters and that grenades were thrown at Hun Sen's home after Sam Rainsy suggested in a speech that the U.S. government should fire cruise missiles at Hun Sen.*) Politiciens cambodgiens a exprimé l'espoir lundi qu'un nouveau partenariat entre les parties de l'homme fort Hun Sen et son rival, le prince Norodom Ranariddh, dans un gouvernement de coalition ne mettrait pas fin à plus de violence. (*Cambodian politicians expressed hope Monday that a new partnership between the parties of strongman Hun Sen and his rival, Prince Norodom Ranariddh, in a coalition government would not end in more violence.*) Le roi Norodom Sihanouk a salué mardi les accords sur le Cambodge les deux principaux partis politiques précédemment amère rivaux pour former un gouvernement de coalition dirigé par l'homme fort Hun Sen. (*King Norodom Sihanouk on Tuesday praised agreements by Cambodia's top two political parties previously bitter rivals to form a coalition government led by strongman Hun Sen.*)

TABLE 4 – Example de résumés en français généré pour l'ensemble D30001T de DUC 2004 par la méthode de référence et notre approche.

Accès au contenu sémantique en langue de spécialité : extraction des prescriptions et concepts médicaux

Cyril Grouin[1] Louise Deléger[1] Bruno Cartoni[2] Sophie Rosset[1] Pierre Zweigenbaum[1]
(1) LIMSI-CNRS, BP133, 91403 Orsay Cedex, France
(2) Département de Linguistique, Université de Genève, Suisse
{cyril.grouin, louise.deleger, sophie.rosset, pierre.zweigenbaum}@limsi.fr,
bruno.cartoni@unige.ch

Résumé. Pourtant essentiel pour appréhender rapidement et globalement l'état de santé des patients, l'accès aux informations médicales liées aux prescriptions médicamenteuses et aux concepts médicaux par les outils informatiques se révèle particulièrement difficile. Ces informations sont en effet généralement rédigées en texte libre dans les comptes rendus hospitaliers et nécessitent le développement de techniques dédiées. Cet article présente les stratégies mises en œuvre pour extraire les prescriptions médicales et les concepts médicaux dans des comptes rendus hospitaliers rédigés en anglais. Nos systèmes, fondés sur des approches à base de règles et d'apprentissage automatique, obtiennent une F_1-mesure globale de 0,773 dans l'extraction des prescriptions médicales et dans le repérage et le typage des concepts médicaux.

Abstract. While essential for rapid access to patient health status, computer-based access to medical information related to prescriptions key medical expressed and concepts proves to be difficult. This information is indeed generally in free text in the clinical records and requires the development of dedicated techniques. This paper presents the strategies implemented to extract medical prescriptions and concepts in clinical records written in English language. Our systems, based upon linguistic patterns and machine-learning approaches, achieved a global F_1-measure of 0.773 for extraction of medical prescriptions, and of clinical concepts.

Mots-clés : Extraction d'information, Indexation contrôlée, Informatique médicale, Concepts médicaux, Prescriptions.

Keywords: Information extraction, Controled indexing, Medical informatics, Clinical concepts, Prescriptions.

1 Introduction

L'accès au sens présent dans les documents au moyen d'outils informatiques est indispensable, tant du point de vue de la compréhension du contenu que du développement des méthodologies informatiques facilitant cet accès. Selon le domaine de langue étudié et le format des données accessibles, la production de systèmes est loin d'être triviale. Nous avons fait le choix d'axer cette étude sur un domaine de langue particulier, le domaine médical, en travaillant sur des documents spécifiques, les comptes rendus hospitaliers. Les comptes rendus hospitaliers intègrent un nombre important d'informations sur l'état de santé des patients, tant au niveau des prescriptions médicales que des concepts médicaux utilisés. Ces informations, bien que partiellement structurées en sections (antécédents du patient, histoire de la maladie, traitement de sortie, etc.), sont rédigées en texte libre et leur appréhension par des outils informatiques, en l'absence de normalisation, se révèle difficile. Cependant, la langue employée dans les comptes rendus se caractérise par une stabilité et une formalisation élevées sur le plan syntaxique, sémantique, et même structurel (Sager, 1981; Friedman, 2000), ce qui autorise une analyse automatique.

Un accès rapide aux informations médicales contenues dans un dossier patient est essentiel pour les praticiens hospitaliers, pour résumer les antécédents du patient ou pour réaliser des études préventives. Deux types d'informations médicales émergent dans les documents cliniques : en premier lieu, les informations liées à la prise de médicaments, qu'elles concernent le médicament en lui-même ou les informations associées (dosage, fréquence, etc.) ; en second lieu, les concepts clés dans la pratique clinique, qui recouvrent les problèmes médicaux (signes, symptômes, maladies, etc.), les examens réalisés pour les diagnostiquer, et les traitements associés.

Nous présentons dans cet article un état de l'art sur l'accès au contenu sémantique dans les comptes rendus cliniques (section 2) puis les approches que nous avons développées pour accéder aux informations médicales, d'une part pour extraire les informations liées aux prescriptions médicales (section 3), d'autre part pour repérer, extraire et typer les concepts médicaux (section 4) dans le cadre de nos participations aux éditions 2009 et 2010 du challenge international i2b2 (*informatics for integrating biology to the bedside*) dont les thématiques concernaient ces aspects (Uzuner *et al.*, 2010a,b). Nous détaillons et discutons les résultats obtenus dans chacune de ces sections.

2 État de l'art

L'accès au contenu d'un document textuel peut être appréhendé de deux manières : soit par le biais d'approches à base d'apprentissage, soit par la création de patrons linguistiques faisant appel à des connaissances d'expert.

Les approches à base d'apprentissage reposent sur l'utilisation de corpus annotés avec soin, dans une volumétrie suffisante et une répartition homogène, pour permettre à un système d'apprendre les conditions dans lesquelles se rencontrent les informations à extraire. Ces approches font l'objet de nombreux travaux, en particulier dans le domaine de la reconnaissance des entités nommées médicales (Li *et al.*, 2008; Doan & Xu, 2010) ou en analyse morphologique (Claveau & Kijak, 2010), rendus possibles par la disponibilité étendue et la simplicité d'utilisation de ces outils d'apprentissage. Si ces outils permettent d'obtenir rapidement de bons résultats, ils demeurent largement dépendants des données fournies en entrée, et seules des données homogènes, de qualité et disponibles en nombre suffisant, tels les corpus des challenges médicaux i2b2, permettent d'obtenir des résultats convaincants.

À l'opposé, les techniques à base de patrons linguistiques faisant appel à des connaissances d'expert pour la production de ces patrons ne nécessitent pas de corpus annotés. Elles nécessitent une somme de travail conséquente pour produire et adapter les patrons mais proposent l'avantage de fournir de bien meilleurs résultats (Long, 2007; Hamon & Grabar, 2010), grâce aux ressources linguistiques existantes en anglais pour le domaine médical, telles que le Metathesaurus et le Specialist Lexicon de l'UMLS (Lindberg *et al.*, 1993). La généralisation de ces approches apparaît souvent délicate à opérer, du fait de la spécialisation de la langue de spécialité concernée.

La combinaison de ces deux approches permet d'accroître sensiblement la qualité des résultats produits, soit comme approches complémentaires l'une de l'autre (une technique suivie de la seconde (Tikk & Solt, 2010)), soit comme apport de l'une pour l'autre (les patrons linguistiques utilisés pour extraire des informations réutilisées comme caractéristiques lors de la construction des modèles d'apprentissage (Wang, 2009)).

Le choix de mobiliser une approche plutôt qu'une autre est souvent dicté par le type de corpus rendu disponible : une approche à base d'apprentissage en cas de corpus annoté, une approche à base de lexiques et de règles le cas échéant. Nous avons suivi cette observation dans les choix méthodologiques décrits dans les sections suivantes.

3 Accès aux prescriptions médicales

Nous avons d'abord mis au point les méthodes d'extraction de prescriptions médicales pour l'anglais, dans le cadre de notre participation à l'édition 2009 du défi i2b2 (Deléger *et al.*, 2010). Nous les avons ensuite adaptées au français. Les données étant relatives à une langue de spécialité, les techniques décrites sont en conséquence conditionnées par cette langue de spécialité.

3.1 Présentation générale

Les prescriptions médicales recouvrent le nom du médicament (qu'il s'agisse d'un nom commercial, du nom générique, ou du principe actif) et les informations associées à ce médicament. On distingue ainsi différents types d'informations. En premier lieu, les informations relatives à la posologie (dosage, fréquence, quantité, mode d'administration, durée), à la forme galénique, etc. Ces informations se présentent sous des formes relativement stables qu'il est alors possible de décrire au moyen de patrons linguistiques. Un deuxième type d'information concerne la raison de la prise de ce médicament. Ce type d'information n'apparaît pas sous une forme régulière et doit faire l'objet d'une analyse plus complexe du texte. Enfin, un troisième type d'information se situe au niveau des événements et de la temporalité relatifs à ces prescriptions médicales et nécessite une analyse des phénomènes linguistiques entrant en jeu autour des noms de médicaments.[1] Le traitement de ce dernier type d'information a été abandonné lors du déroulement du défi 2009.

3.2 Présentation du corpus

Le corpus est composé de comptes rendus hospitaliers rédigés en anglais. Les documents proviennent d'un centre médical américain spécialisé en cardiologie. Ils ont fait l'objet d'une anonymisation où les informations personnelles (noms, prénoms, etc.) ont été remplacées par d'autres informations de même type en conservant un caractère vraisemblable. Le corpus de développement intègre 696 documents, parmi lesquels 17 ont fait l'objet d'une annotation, tandis que le corpus de test intègre 547 documents. Les documents sont structurés en sections assez générales telles que histoire de la maladie, allergies, examens de laboratoire, suivi de l'hospitalisation, et prescriptions de sortie. Les textes contiennent des abréviations qui concernent les noms de médicaments ("vanc" pour vancomycin, "levo" ou "levoflox" pour levofloxacin), les symptômes médicaux ("afib" pour *atrial fibrillation*, "abd pain" pour *abdominal pain*), les fréquences ("bid" pour *bis in diem*), et les modes d'administration ("iv" pour *intravenous*, "sub" pour *sub-lingual*).

Aucune annotation de référence n'existant préalablement au lancement du défi, la référence a été constituée en deux temps, premièrement par un vote majoritaire des sorties produites par les participants, et deuxièmement via une phase d'adjudication faisant intervenir l'ensemble des participants au défi (Uzuner *et al.*, 2010b). Au final, la référence a été constituée de manière collective pour 251 documents du corpus de test. Les résultats que nous présentons dans cet article pour la partie extraction de prescriptions médicales se fondent donc sur l'évaluation opérée sur ces documents de référence.

	Développement		**Test**	
Nombre de documents	17		251	
Médicaments	749	100,0 %	8 941	100,0 %
Dosage	397	53,0 %	4 460	49,9 %
Mode d'administration	253	33,8 %	3 389	37,9 %
Fréquence	374	49,9 %	4 042	45,2 %
Durée	66	8,8 %	550	6,2 %
Raison	150	20,0 %	1 636	18,3 %

TAB. 1 – Nombre d'éléments à extraire dans les documents annotés des corpus de développement et de test.

[1] La prescription médicale est-elle en cours, ou bien doit-elle être commencée ou arrêtée ? Où se situe la prescription médicale sur l'échelle temporelle (dans le passé, le présent ou le futur) ? Comment la prescription médicale est-elle présentée au patient (le médicament doit-il être pris obligatoirement, sous certaine condition, ou s'agit-il d'une suggestion) ?

Le tableau 1 renseigne du nombre d'informations attendues dans chaque corpus. Faute de disposer d'un corpus de développement entièrement annoté, nous donnons la volumétrie pour les 17 fichiers annotés qui nous ont été fournis par les organisateurs avec le corpus de développement. Si le nombre d'informations de chaque type reste proportionnel entre les deux corpus, il apparaît d'emblée que certaines informations sont peu présentes dans l'ensemble des corpus, rendant difficile le développement d'outils robustes pour les traiter. C'est notamment le cas des informations de durée renseignées dans moins de 10 % des prescriptions. Une prescription sur cinq seulement intègre la raison pour laquelle le médicament a été prescrit. Les autres types d'information sont davantage renseignés : le mode d'administration dans une prescription sur trois, les dosage et fréquence dans des proportions équivalentes d'une prescription sur deux.

Dans l'exemple du tableau 2, nous représentons les informations à extraire en les encadrant de balises. Les deux occurrences du médicament *heparin* doivent donner lieu à deux lignes de sortie. La première ligne – relative à la première occurrence – intègrera les informations de dosage, de mode d'administration, de fréquence et de raison, alors que la seconde ligne – relative à la seconde occurrence – ne comprendra que l'information de raison, les autres informations se rapportant uniquement à la première apparition.

<raison> *Prophylaxis* </raison> , <medicament> *heparin* </medicament> <dosage> *5000 units* </dosage> <mode> *subcu* </mode> <frequence> *t.i.d.* </frequence> - *the patient has consistently refused her* <medicament> *heparin* </medicament> .

TAB. 2 – Exemple d'annotation en prescriptions médicales.

3.3 Description du système

Notre système ayant été développé dans le cadre de la participation à l'édition 2009 du défi i2b2, nous l'avons orienté vers le traitement des informations suivantes : nom du médicament, dosage, mode d'administration, fréquence, durée, raison de la prescription, et type de portion de texte dans lequel apparaît la prescription (liste ou passage narratif). Nous avons fait le choix de développer un système reposant entièrement sur des règles d'extraction et des listes, sans recourir à des outils externes tels que des étiqueteurs, lemmatiseurs ou analyseurs syntaxiques. Ce choix repose sur le fait que les informations à extraire peuvent l'être, soit par la projection de lexiques (noms de médicaments, modes d'administration), soit par l'utilisation de règles (les chiffres des dosages, fréquences, durées, etc.), ces méthodes permettant l'obtention rapide de résultats de qualité.

Les problèmes à résoudre dans cette tâche consistaient à gérer l'exhaustivité des noms de médicaments (génériques, marques, classes thérapeutiques) et l'ambiguïté intrinsèque de ces noms (distinguer la concentration du dosage, repérer les substances actives utilisées comme nom de médicament). Nous devions également calculer le rattachement des informations aux noms de médicaments, prendre en compte la factorisation des informations, et considérer les cas particuliers de reprises pronominales.

3.3.1 Lexiques

Nous avons créé trois types de lexiques. Le premier lexique concerne les noms de médicaments et existe sous deux versions : une version réduite de 8 923 noms de médicaments issus de deux sites Internet (FDA[2] et RxList[3]), et une version plus large contenant 180 089 noms correspondant aux entrées du Metathesaurus de l'UMLS[4] pour le type sémantique *Clinical drug*. Les éléments présents dans cette seconde liste sont néanmoins sujets à discussion et ne correspondent pas toujours à des noms de médicaments tels que ceux attendus (*alcool, tabac,* etc.). Le second lexique est constitué d'une liste de symptômes médicaux pour permettre l'identification de la prescription d'un médicament. Il a été créé à partir des entrées de l'UMLS classées sous le type sémantique *Sign and Symptom*. Enfin, le dernier lexique consiste en une liste d'abréviations et termes spécifiques issue des travaux de (Berman, 2004). Nous avons mis en correspondance chaque terme avec le type d'information qui lui correspond : des abréviations ou termes de types dosage (*mg, sliding scale*), mode d'administration (*iv, intramuscular*), fréquence (*qd, prn*), durée (*week*).

[2]FDA : Food and Drug Administration, `http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm`
[3]`http://www.rxlist.com/`
[4]UMLS : Unified Medical Language System.

3.3.2 Algorithme

Nous avons défini une stratégie d'extraction d'information reposant sur deux étapes principales (figure 1) : dans un premier temps, nous identifions les noms de médicaments ; à partir de cette première étape, nous recherchons les informations associées à chaque médicament.

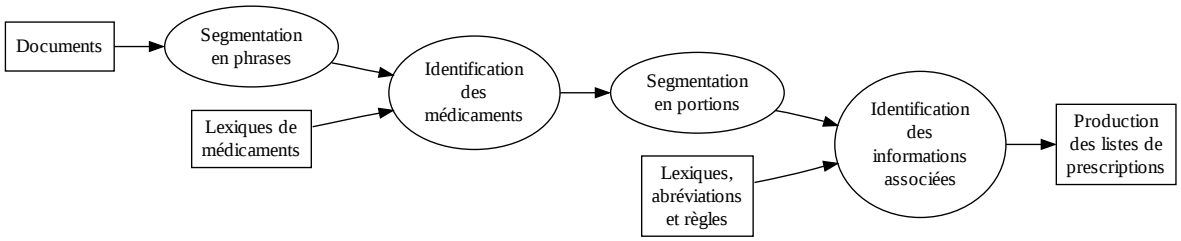


FIG. 1 – Architecture du système d'extraction de prescriptions médicales utilisé pour i2b2 2009.

L'identification des noms de médicaments repose uniquement sur un appariement exact avec le contenu des lexiques de médicaments précédemment décrits. Une fois les noms de médicaments identifiés dans un document, nous cherchons les informations qui lui sont associées. Nous avons élaboré des expressions régulières pour chaqye type d'information à traiter, à partir des guides d'annotation et d'exemples identifiés en corpus. Nous complétons l'application de ces règles par une recherche dans les listes d'abréviations et de symptômes.

Pour déterminer les informations devant être associées à chaque médicament, nous avons procédé à deux étapes de segmentation du texte. Dans un premier temps, nous segmentons le texte en phrases en nous fondant sur la mise en forme du document (lignes séparatrices et titres de section) et la ponctuation (en distinguant les points de fin de phrase des points d'abréviation ou des points mathématiques dans les décimales en anglais). Nous identifions les noms de médicaments dans ces phrases. Dans un second temps, nous procédons à une segmentation des phrases sur la base des noms de médicaments précédemment identifiés, en considérant que chaque nom de médicament constitue le début d'une portion de phrase. Nous cherchons alors les informations associées à chaque médicament à l'intérieur de ces portions, considérant que les informations associées aux prescriptions médicales suivent toujours les noms de médicaments. Pour le cas où certains types d'information n'auraient pas été trouvés à la suite du nom de médicament, nous les cherchons dans la portion qui précède.

Le système permet également de gérer les cas de doubles entrées, lorsqu'une même information s'applique à deux prescriptions différentes (deux médicaments prescrits pour soigner la même affection), ou parce qu'une seule expression factorise deux informations de même type (un dosage différent le matin et le soir). Nous avons géré ces cas au moyen de règles définies empiriquement.

Enfin, nous avons traité quelques cas particuliers de résolution des anaphores au moyen de règles dédiées : le pronom "*this*" suivi de trois syntagmes verbaux, "*was discontinued*", "*was increased*" et "*was decreased*". Dans ces cas de reprise pronominale, nous avons créé une seconde sortie pour le médicament désigné par le pronom, éventuellement complétée par les informations suivant le syntagme verbal (en cas de modification du dosage, etc.).

3.4 Résultats et discussion

Nous donnons dans le tableau 3 les résultats obtenus par notre système sur le corpus de test composé des 251 documents annotés collectivement. Comme pour toute évaluation d'un système d'extraction d'information, deux points sont ici évalués : le typage de l'élément extrait d'une part, et la portée de l'extraction d'autre part. Les résultats présentés ici exigent que la portée ait été déterminée de façon exacte (notre système peut avoir correctement typé un élément mais l'évaluation sera incorrecte du fait d'une erreur de frontière dans la portée de l'information extraite). Les informations élémentaires de type dosage, mode d'administration, fréquence, durée, et raison ne sont considérées comme pertinentes que si elles sont associées dans la référence à un médicament. Les rangées médicament, dosage, etc., évaluent chaque type d'information séparément. La rangée « niveau horizontal » demande qu'une prescription soit complètement et exactement reconnue pour être considérée comme correcte.

	Nombre	Rappel	Précision	F_1-mesure
Niveau horizontal	8 941	0.725	0.827	0.773
Médicament	8 941	0.793	0.802	0.798
Dosage	4 460	0.732	0.892	0.804
Mode d'administration	3 389	0.792	0.885	0.836
Fréquence	4 042	0.770	0.893	0.827
Durée	550	0.282	0.657	0.394
Raison	1 636	0.234	0.412	0.299

TAB. 3 – Résultats obtenus par notre système au défi i2b2 2009 (recouvrement exact).

Notre système obtient globalement de bons résultats (il a été classé 8ème sur 20 participants internationaux) avec une précision toujours supérieure au rappel, notre système générant relativement peu de bruit. Certains types d'information tels que la durée et les raisons de la prescription ont produit des résultats assez bas. Concernant les durées, le nombre restreint d'exemples dans le corpus de développement ne nous a pas permis de définir de manière précise et robuste les règles appliquées pour l'identification de ce type d'information.

Nous estimons qu'un moyen d'améliorer la détections des raisons passe par l'utilisation d'outils d'analyse syntaxique, de manière à identifier précisément les syntagmes nominaux et prépositionnels. Il semble que dans une bonne partie des situations où notre méthode n'a pas pu détecter la raison d'une prescription, cette raison était exprimée dans le contexte d'une portion de phrase relativement bien formée, où les relations grammaticales ont de bonnes chances d'être analysables automatiquement et d'aider à rattacher raison et médicament. Cependant, la variation syntaxique et l'étendue des raisons annotées dans le corpus d'entraînement témoignent de la complexité de cette tâche : les raisons "*pain*" (un seul terme), "*the previous enterococcus infection*" (un syntagme nominal), et "*had a temperature to about 101*" (un syntagme verbal) ont ainsi été associées au médicament "*vancomycin*". Un autre moyen consiste à utiliser une base de connaissances faisant le lien entre médicament et symptômes traités : si le terme "*hypercholesterolemia*" (ou une variante) est trouvé dans le voisinage des médicaments "*Zocor*" et "*simvastatin*", nous pourrions extraire la raison en accordant une importance accrue à ce terme. Une autre piste permettant l'amélioration de l'identification des informations associées consiste à mobiliser des présupposés d'expert, en adoptant une approche par inférence (déduire le mode d'administration d'un médicament à partir de sa forme galénique). Le coût de constitution d'une telle base de données associé à l'absence de normalisation des textes risquent néanmoins de limiter les apports d'une telle démarche.

4 Accès aux concepts médicaux

4.1 Présentation générale

La première piste de la campagne i2b2/VA 2010 concernait la détection et le typage de concepts médicaux dans des comptes rendus médicaux, parmi trois catégories de concepts (voir tableau 4) : les *problèmes* se rapportent aux observations faites sur l'état du patient et concernent les maladies et symptômes anormaux ou liés à une maladie existante, les *traitements* décrivent les méthodes utilisées pour résoudre le problème d'un patient (procédures, médicaments, etc.), et les *examens* se rapportent aux examens prescrits pour aider à diagnostiquer ces problèmes.

Concept	Exemples
Problème	<problem> *C5-6 disc herniation* </problem> *with* <problem> *cord compression* </problem> PRN <problem> *Shortness of Breath* </problem>
Traitement	<treatment> *bilateral lymph node dissection* </treatment> <treatment> *LISINOPRIL* </treatment> *10 MG PO DAILY*
Examen	*If* <test> *BS* </test> *is less than 125* *He was found on* <test> *physical exam* </test> *to have an asymmetric prostate*

TAB. 4 – Exemples de concepts de chaque type pour la tâche i2b2/VA 2010.

La syntaxe spécifique de la langue médicale utilisée dans les comptes rendus médicaux a notamment été décrite par (Sager *et al.*, 1994, 1995; Sager & Nhàn, 2002). Nous constatons aisi que certaines phrases peuvent être constituées presque exclusivement d'énumérations, ne comprendre qu'un seul mot ou au contraire être longues et qu'il n'y a pas eu de normalisation dans la façon de noter les éléments (voir tableau 5).

Phénomène étudié	**Exemples**
Absence de normalisation	*Supprelin La* vs *Supprelin LA* *magnetic resonance imaging of ...* vs *MRI of ...* *Thaw* vs *THAUW*
Forme des phrases	*On physical examination today , his lungs are clear to auscultation and percussion .* *Regular rhythm .* *f / u with PCP and Dr. Pump as scheduled , return to ED with worsening sob or increased cough or sputum production*

TAB. 5 – Exemples de problèmes rencontrés en langue de spécialité.

Ces différentes considérations nous ont convaincus de ne pas procéder à une analyse syntaxique des documents comme traitement de base. Du fait de la forme très variable des expressions désignant les concepts à détecter, nous avons également décidé de ne pas chercher à modéliser complètement ces expressions par une ou plusieurs grammaires locales. Par ailleurs, disposant d'un corpus d'apprentissage de taille raisonnable, nous avons opté pour une approche s'appuyant sur des champs conditionnels aléatoires (CRF) (Lafferty *et al.*, 2001), ces derniers permettant de bonnes performances pour une tâche d'étiquetage en séquence comme celle de la détection de concepts. Nous avons pour cela utilisé l'implémentation CRF++ (Kudo, 2007). Toutefois, si ces modèles permettent de bonnes performances, des expériences (Zidouni *et al.*, 2010) ont montré qu'utiliser comme attributs des informations d'ordre linguistique (POS, informations sémantiques, etc.) permettait d'améliorer les modèles. Nous avons cherché à produire des informations et des analyses partielles des expressions concernées, et à fournir au CRF des attributs encodant ces informations. L'objectif étant de produire les analyses linguistiques que l'on peut obtenir de façon fiable et de déléguer au processus d'apprentissage les décisions finales sur les frontières et type des entités.

4.2 Description du corpus

Le corpus se compose de comptes rendus cliniques provenant à part égale de trois hôpitaux nord-américains.[5] Le corpus d'entraînement se compose de 349 documents manuellement annotés[6] tandis que le corpus de test comprend 477 documents. Il n'existe pas de type de concept sur-représenté par rapport aux autres types et la distribution des types reste équivalente entre les deux corpus (voir tableau 6). Enfin, nous observons que les concepts médicaux à identifier recouvrent des formes d'expressions assez différentes à l'intérieur de chaque type. Une abréviation ou un syntagme nominal complet peuvent tous deux constituer un concept médical (tableau 4).

	Développement		**Test**	
Nombre de documents	349		477	
Concepts	27 837	100,0 %	45 009	100,0 %
Problème	11 968	43,0 %	18 550	41,2 %
Traitement	8 500	30,5 %	13 560	30,1 %
Examen	7 369	26,5 %	12 899	28,7 %

TAB. 6 – Nombre d'éléments à extraire dans les corpus de développement et de test.

[5]Beth Israel Deaconess Medical Center (Boston, MA), Partners HealthCare (Boston, MA), University of Pittsburgh Medical Center (Pittsburgh, PA). Ces instituts ont tous trois fourni des comptes rendus cliniques ; l'Université de Pittsburgh a également fourni des notes de suivi.

[6]Les organisateurs ont également fourni 827 documents non annotés avec le corpus de développement. Nous avons fait le choix de ne travailler que sur les 349 documents annotés, notre système reposant sur la construction de modèle par apprentissage (voir sous-section 4.3).

4.3 Description du système

4.3.1 Présentation générale

L'approche que nous avons développée (Minard *et al.*, 2011) repose sur un système à base d'apprentissage. Nous avons ainsi créé des modèles d'apprentissage à base de CRF en utilisant les traits habituels pour ce genre de tâche, à savoir des n-grammes et des indices typographiques (casse, ponctuation, token alphabétique ou numérique etc.). Nous avons également ajouté des traits correspondant aux résultats d'analyses linguistiques.

Afin de procéder à différents tests lors de la construction du modèle, nous avons scindé le corpus de développement en sous-corpus d'entraînement (241 documents), de développement (54 documents) et de test à blanc (54 documents). Pour la phase de test du défi, une fois trouvée la meilleure configuration, nous avons reconstruit un modèle global fondé sur l'ensemble des 349 documents.

4.3.2 Algorithme

Notre approche reposant sur l'application d'un modèle à base d'apprentissage, nous avons mobilisé plusieurs ressources pour produire les traits nécessaires à la construction du modèle (schéma 2).

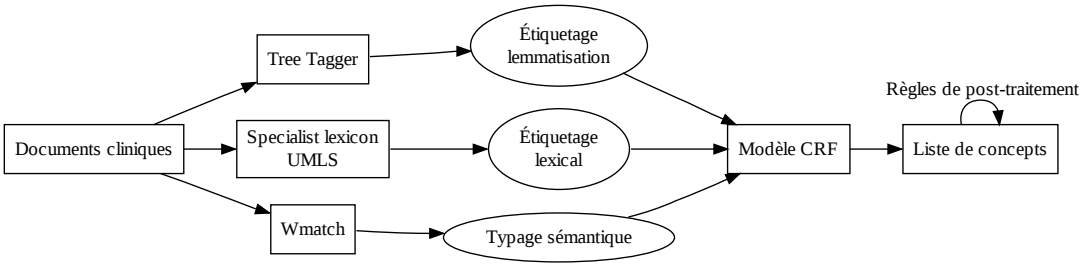


FIG. 2 – Architecture du système d'extraction de concepts médicaux utilisé pour i2b2/VA 2010.

Tous les mots des corpus ont d'abord été annotés en utilisant le Tree Tagger (Schmid, 1994) et ses modèles pour l'anglais. Ainsi chaque token a été associé avec sa partie du discours et son lemme.

Nous avons ensuite effectué un étiquetage à l'aide d'informations lexicales en utilisant les ressources fournies par le Specialist Lexicon de l'UMLS (Lindberg *et al.*, 1993). Ces ressources contiennent 62 263 adjectifs et 320 013 noms, et distinguent les adjectifs relationnels des adjectifs qualificatifs, ainsi que différents types de noms (noms propres, noms comptables et non comptables). Pour les adjectifs, cette ressource contient également des informations sur la position des adjectifs dans la phrase (*attribut* ou *post-nominal*).

Nous avons également ajouté une information sémantique en nous appuyant sur les travaux de (Sager *et al.*, 1995) et sur les données d'entraînement. Nous avons ainsi construit des lexiques spécialisés (pour les noms des parties du corps, de maladie, de médicaments) et des grammaires permettant de typer des segments en fonction de différentes catégories : les parties du corps (*sternal articular facet of third costal cartilage*), les analyses de laboratoire (*blood wbc, creatinine, hematocrit*), les différents examens (*angiography, biopsy*), des pré- et post-marqueurs d'examens (*follow-up ..., physicial ..., ... levels*), les médicaments (*Abacavir Sulfate*), les mode d'administration (*inhaler, oral, pills*), les instruments et objets médicaux (*cannula, pacemaker, stent*), les procédures (*bypass, amputation, resection*), et les dosages (*100 mg, 1 dose*). Ces différentes catégories ont paru pertinentes après examen du corpus et analyse des contextes droit et gauche des concepts ainsi que de la composition des concepts eux-mêmes. Précisons que ces catégories n'avaient pas vocation à représenter directement les concepts, mais à fournir des classes permettant de regrouper des mots ou groupes de mots sous une même appellation afin de réduire l'espace de recherche. Le tableau 7 montre des exemples de ces catégories (partie gauche du tableau) et les met en rapport avec les concepts (partie droite). Nous avons par ailleurs remarqué que ces catégories fournissent des informations structurantes qui se rapportent aux concepts. Ainsi, un test se rapporte souvent à une partie de l'anatomie et une procédure alors que certains noms ou adjectifs sont fréquemment présents en partie droite (pré-marqueur) ou gauche (post-marqueur) des concepts, en particulier pour les concepts *problème* et *examen*.

Annotation sémantique + POS	**Annotation correspondante du concept**
1)_JJ Rapid_JJ <anat> atrial_JJ </anat> <diag> fibrillation_NN </diag> with_IN demand_NN <diag> ischemia_NN </diag>	1) <problem> *Rapid atrial fibrillation* </problem> *with* <problem> *demand ischemia* </problem>
<localisation> Left_VVD </localisation> <anat> heart_NN </anat> <procedure> catheterization_NN </procedure> without_IN intervention_NN (_(**DATE[Dec_NP 16_CD 07]_NN)_) ._SENT	<test> *Left heart catheterization* </test> *without* <treatment> *intervention* </treatment> *(**DATE[Dec 16 07])* .
There_EX was_VBD no_DT <diag> diplopia_NN </diag> ,_, visual_JJ <pomark-disease> loss_NN </pomark-disease> ,_, <diag> speech_NN abnormality_NN </diag> or_CC sensory_JJ change_NN in_IN her_PP$ history_NN ._SENT	*There was no* <problem> *diplopia* </problem> *,* <problem> *visual loss* </problem> *,* <problem> *speech abnormality* </problem> *or* <problem> *sensory change* </problem> *in her history* .
<premark-disease> Significant_JJ </premark-disease> for_IN non-insulin_NN <diag> diabetes_NN mellitus_NN </diag> ,_, for_IN which_WDT he_PP takes_VVZ <medoc> Diabeta_NP </medoc> ,_, one_CD QD_NNS ;_ : <anat> right_JJ eye_NN </anat> <diag> cataract_NN </diag> ,_, operated_VVN on_IN three_CD years_NNS ago_RB ._SENT	*Significant for* <problem> *non-insulin diabetes mellitus* </problem> *, for which he takes* <treatment> *Diabeta* </treatment> *, one QD ;* <problem> *right eye cataract* </problem> *, operated on three years ago* .

TAB. 7 – Lien entre information sémantique et concepts.

Enfin, nous avons cherché à voir s'il était possible de typer sémantiquement les tokens en fonction de leurs premiers ou derniers caractères que nous appelons par commodité préfixe et suffixe. Nous avons découpé les différents mots n'appartenant pas aux dictionnaires de spécialité (médicaments ou parties du corps) puis en avons extrait les successions de caractères qui permettaient à coup sûr un début de classification. Ainsi, les suffixes de type *-stomy* renvoient fréquemment à une procédure. En tout, cinq classes sémantiques ont été utilisées (position, chiffrage, procédure, examen, diagnostic).

Les grammaires ont été construites en utilisant WMATCH, un moteur d'analyse fondé notamment sur des expressions régulières de mots (Galibert, 2009; Rosset *et al.*, 2008). L'analyseur a été construit de manière automatique à partir des données d'entraînement et des différents lexiques de spécialité à notre disposition. Ceux-ci étaient au nombre de trois : *anatomie* (145 199 mots ou expressions complexes), *médicaments* (27 518 mots ou expressions complexes), et *maladies* (175 645 mots ou expressions complexes). Nous avons d'autre part collecté les collocations des concepts et créé, en nous appuyant sur la fréquence et la distribution non ambiguë des termes, des lexiques spécifiques à la tâche (modes d'administration, procédures, outils médicaux, localisations sur le corps du patient – souvent en rapport avec une partie du corps –, examens, et pré et post-marqueurs, tant pour les examens que pour les maladies). Ces lexiques ont été utilisés pour l'acquisition des règles d'analyse au format WMATCH. Le tableau 8 présente des exemples de catégorisation de mot fondé sur le suffixe (extrait), de règle contextuelle pour la détection de la catégorie *mode* et d'appel à un lexique. La règle de catégorisation indique que les mots se terminant par "asty" sont une procédure. La règle contextuelle contient deux applications possibles (séparées par le symbole "|") : les mots détectés par la macro *&modes* (un ensemble de règles contextuelles) et suivis éventuellement de *load* sont annotés comme étant un *_mode* ; il en est de même pour le mot *release*, s'il est précédé d'un adjectif. L'application du lexique se fait en appelant la macro qui inclut le lexique de procédures.

Catégorisation	_procedure : [A-z]+ "omy" | [A-z]+ "asty" ... ;
Règle contextuelle	_mode : (&modes load ? | (<= _JJ| _VVD| _VVN) release) ;
Application de lexique	_procedure : (&procedure) ;

TAB. 8 – Exemples de règles.

Ces différentes informations ont constitué l'ensemble des traits qui ont alimenté l'apprentissage du modèle CRF. Ce modèle et les modules d'extraction de traits forment le système de base pour cette campagne d'évaluation.

Enfin, nous avons ajouté en sortie de ce système une phase de correction par l'ajout de règles de post-traitement. Nous avons supposé que l'hypothèse « *un sens par corpus* » (Fung, 1998) est vérifiée dans une langue de spécialité, à plus forte raison dans le typage de concepts médicaux : nous avons examiné les expressions étiquetées par des types de concepts différents dans le corpus et avons normalisé leur étiquette au type observé le plus fréquent (un token ayant pour trait la catégorie *médicament* qui n'aurait pas été typé ou l'aurait été typé différemment de *traitement* est corrigé avec le type *traitement*).

4.4 Résultats et discussion

Le tableau 9 présente les résultats obtenus par notre système sur l'identification et le typage des concepts médicaux. L'évaluation a été réalisée sur 477 documents. Les chiffres renseignés dans ce tableau reposent sur un appariement à l'identique des concepts ; les erreurs de frontière ont donc été pénalisantes.

	Nombre	**Rappel**	**Précision**	**F_1-mesure**
Global	27 837	0.726	0.826	0.773
Problème	11 968	0.742	0.799	0.769
Traitement	8 500	0.723	0.843	0.778
Examen	7 369	0.705	0.851	0.771

TAB. 9 – Résultats obtenus par notre système au défi i2b2 2010 (recouvrement exact).

Le système d'identification et de typage des concepts médicaux obtient une F_1-mesure générale de 0,773 (notre système s'est classé 12ème sur 22 participants internationaux). Pour cette tâche d'extraction de concepts médicaux, notre système obtient de nouveau une précision supérieure au rappel pour chaque type de concept. Nous notons que les performances du système se révèlent équivalentes sur les trois types de concepts médicaux à traiter, cette observation s'expliquant par la répartition équilibrée des concepts dans ces trois catégories. Les dix meilleurs systèmes du défi ont tous employé des méthodes d'apprentissage. Le meilleur système (De Bruijn *et al.*, 2010) a modélisé la tâche avec un CRF et s'en est servi pour définir les traits d'un modèle semi-markovien caché. Plusieurs autres systèmes bien classés ont utilisé comme traits le résultat de systèmes de reconnaissance d'entités médicales.

5 Conclusion

Dans le cadre de ce travail, nous avons constitué un ensemble de ressources nécessaires au traitement de la langue médicale. Nous avons ainsi dressé un inventaire exhaustif des noms de médicaments (génériques, marques et classes thérapeutiques) et créé des lexiques d'abréviations et de symptômes. Nous avons par ailleurs élaboré une méthodologie de détection des types d'entités de différentes sortes (par l'application d'expressions régulières et l'utilisation d'un lexique d'abréviations spécifiques) et de gestion de la factorisation d'information (coordination et duplication). Enfin, nous avons étudié les caractéristiques linguistiques à utiliser pour la construction de modèles d'apprentissage dédiés au traitement des concepts médicaux.

En Traitement Automatique des Langues, les systèmes à base de règles constituent une solution pertinente pour traiter des corpus non annotés porteurs d'informations stables syntaxiquement. L'application de patrons syntaxiques permet effectivement d'obtenir rapidement de bons résultats comme en témoignent ceux que nous avons obtenus sur l'extraction d'information dans les prescriptions médicales lors de l'édition 2009 du défi i2b2.

En revanche, la variation syntaxique des informations à extraire se révèle beaucoup plus difficile à traiter. L'utilisation seule de règles syntaxiques conduit à un manque de robustesse du système et doit faire l'objet d'une application complémentaire d'autres types de méthodes. À cet effet, l'utilisation de méthodes hybrides rassemblant un apprentissage supervisé et des informations linguistiques permet d'accroître les chances de traiter correctement ce type de données. C'est l'approche que nous avons suivie pour l'identification et le typage des concepts médicaux pour l'édition 2010 du défi i2b2 ; dans le cas présent, nous nous sommes servis d'informations d'ordre linguistique à deux reprises : en premier lieu pour constituer des traits sur chaque token de manière à construire un modèle pour l'apprentissage, puis dans un second temps, comme moyen d'affiner les résultats produits par l'application du modèle précédemment construit.

Dans le domaine médical, la langue de spécialité utilisée revêt un caractère particulièrement stable et formel, tant sur les plans syntaxique que sémantique, voire structurel. Ces caractéristiques nous autorisent à utiliser des approches hybrides lorsqu'existent des corpus annotés. Lorsque les annotations font défaut, les caractéristiques linguistiques de la langue médicale nous permettent néanmoins de travailler uniquement à base de patrons syntaxiques. Ces méthodes montrent leurs limites lorsque l'information à extraire se trouve rédigée en texte plus libre, à l'instar des raisons qui justifient une prescription médicale. Dans cette perspective, des traitements linguistiques plus complexes faisant intervenir une analyse en dépendances pourraient constituer une alternative intéressante.

Remerciements

Ce travail a été partiellement réalisé dans le cadre des projets Akenaton (ANR-07-TecSan-001) et Quæro (financement Oseo, agence française pour l'innovation et la recherche).

Les données médicales utilisées proviennent du consortium Informatics for Integrating Biology to the Bedside (i2b2) grâce aux financements numéros U54LM008748 de la National Library of Medicine, VA HSR HIR 08-374 du Consortium for Healthcare Informatics Research (CHIR), et VA HSR HIR 08-204 du VA Informatics and Computing Infrastructure (VINCI).

Références

BERMAN J. J. (2004). Pathology Abbreviated : A Long Review of Short Terms. *Archives of Pathology & Laboratory Medicine*, **128**(3), 347–352.

CLAVEAU V. & KIJAK E. (2010). Analyse morphologique en terminologie biomédicale par alignement et apprentissage non-supervisé. In *Actes de TALN 2010*.

DE BRUIJN B., CHERRY C., KIRITCHENKO S., MARTIN J. & ZHU X. (2010). MRC at i2b2 : one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. In *Proc. of i2b2/VA 2010*.

DELÉGER L., GROUIN C. & ZWEIGENBAUM P. (2010). Extracting Medical Information from Narrative Patient Records : the Case of Medication-related Information. *J Am Med Inform Assoc*, **17**(5), 555–558.

DOAN S. & XU H. (2010). Recognizing Medication related Entities in Hospital Discharge Summaries using Support Vector Machine. In *Coling2010 : Poster Volume*, p. 259–266.

FRIEDMAN C. (2000). A broad-coverage natural language processing system. In *AMIA Annu Symp Proc*, p. 270–274.

FUNG P. (1998). A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-parallel Corpora. In *AMTA*, p. 1–17.

GALIBERT O. (2009). *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. PhD thesis, Université Paris-Sud 11, Orsay, France.

HAMON T. & GRABAR N. (2010). Linguistic approach for identification of medication names and related information in clinical narratives. *J Am Med Inform Assoc*, **17**(5), 549–554.

KUDO T. (2007). CRF++. http ://crfpp.sourceforge.net/.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional Random Fields : Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, p. 282–289.

LI D., KIPPER-SCHULER K. & SAVOVA G. (2008). Conditionnal Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts. In *BioNLP2008 : Current Trends in Biomedical Natural Language Processing*, p. 94–95.

LINDBERG D., HUMPHREYS B. & MCCRAY A. (1993). The Unified Medical Language System. *Meth Inform Med*, **32**(4), 281–291.

LONG W. (2007). Lessons Extracting Diseases from Discharge Summaries. In *AMIA Annu Symp Proc*, p. 478–482.

MINARD A.-L., LIGOZAT A.-L., BEN ABACHA A., BERNHARD D., CARTONI B., DELÉGER L., GRAU B., ROSSET S., ZWEIGENBAUM P. & GROUIN C. (2011). Hybrid Methods for improving Information Access in Clinical Documents : Concept, Assertion, and Relation Identification. *J Am Med Inform Assoc*. À paraître.

ROSSET S., GALIBERT O., BERNARD G., BILINSKI E. & ADDA G. (2008). The LIMSI participation to the QAst track. In *Working Notes of CLEF 2008 Workshop*, Aarhus, Danemark.

SAGER N. (1981). *Natural Language Processing : A Computer Grammar of English and Its Applications*. Addison Wesley.

SAGER N., LYMAN M., NHÀN N. & TICK L. (1994). Automatic Encoding into SNOMED III : A Preliminary Investigation. In *Proc. of the 18th Annual Symposium on Computer Applications in Medical Care*, p. 230–234.

SAGER N., LYMAN M., NHÀN N.-T. & TICK L. J. (1995). Medical language processing : applications to Patient Data Representation and Automatic Encoding. *Meth Inform Med*, **34**(1–2), 140–146.

SAGER N. & NHÀN N.-T. (2002). The Computability of strings, transformations, and sublanguage. In B. E. NEVIN & S. M. JOHNSON, Eds., *The legacy of Zellig Harris – Language and information into the 21st century - volume 2 : computability of language and computer applications*, volume 2, chapter 4, p. 79–120. Amsterdam/Philadelphia : John Benjamins Publishing Company.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. of the International Conference on New Methods in Language Processing*, p. 44–49.

TIKK D. & SOLT I. (2010). Improving textual medication extraction using combined conditional random fields and rule-based systems. *J Am Med Inform Assoc*, **17**(5), 540–544.

UZUNER O., SOLTI I. & CADAG E. (2010a). Extracting medication information from clinical text. *J Am Med Inform Assoc*, **17**(5), 514–518.

UZUNER O., SOLTI I., XIA F. & CADAG E. (2010b). Community annotation experiment for ground truth generation for the i2b2 medication challenge. *J Am Med Inform Assoc*, **17**(5), 519–523.

WANG Y. (2009). Annotating and Recognising Named Entities in Clinical Notes. In *Proc. of the ACL-IJCNLP 2009 Student Research Workshop*, p. 18–26, Singapore.

ZIDOUNI A., ROSSET S. & GLOTIN H. (2010). Efficient Combined Approach for Named Entity Recognition in Spoken Language. In *Proc. of InterSpeech*, Makuhari, Japon.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Génération automatique de motifs de détection d'entités nommées en utilisant des contenus encyclopédiques.

Eric Charton[1] Michel Gagnon[1] Benoit Ozell[1]
(1) École Polytechnique, 2900 boul. Edouard Montpetit, Montréal, Canada
{eric.charton, michel.gagnon, benoit.ozell}@polymtl.ca

Résumé. Les encyclopédies numériques contiennent aujourd'hui de vastes inventaires de formes d'écritures pour des noms de personnes, de lieux, de produits ou d'organisation. Nous présentons un système hybride de détection d'entités nommées qui combine un classifieur à base de Champs Conditionnel Aléatoires avec un ensemble de motifs de détection extraits automatiquement d'un contenu encyclopédique. Nous proposons d'extraire depuis des éditions en plusieurs langues de l'encyclopédie Wikipédia de grandes quantités de formes d'écriture que nous utilisons en tant que motifs de détection des entités nommées. Nous décrivons une méthode qui nous assure de ne conserver dans cette ressources que des formes non ambiguës susceptibles de venir renforcer un système de détection d'entités nommées automatique. Nous procédons à un ensemble d'expériences qui nous permettent de comparer un système d'étiquetage à base de CRF avec un système utilisant exclusivement des motifs de détection. Puis nous fusionnons les résultats des deux systèmes et montrons qu'un gain de performances est obtenu grâce à cette proposition.

Abstract. Encyclopedic content can provide numerous samples of surface writing forms for persons, places, products or organisations names. In this paper we present an hybrid named entities recognition system based on a gazetteer automatically extracted. We propose to extract it from various language editions of Wikipedia encyclopedia. The wide amount of surface forms extracted from this encyclopedic content is then used as detection pattern of named entities. We build a labelling tool using those patterns. This labelling tool is used as simple pattern detection component, to combine with a Conditional Random Field tagger. We compare the performances of each component of our system with the results previously obtained by various systems in the French NER campaign ESTER 2. Finally, we show that the fusion of a CRF label tool with a pattern based ones, can improve the global performances of a named entity recognition system.

Mots-clés : Étiqueteur, Entités nommées, Lexiques.

Keywords: Tagger, Named entities, Gazetteer.

1 Introduction

La tâche d'*étiquetage par des entités nommées* (EEN) est un processus lors duquel chaque mot d'une phrase correspondant à une *entité nommée* (EN) (généralement un nom propre et par extension des dates ou des quantités) reçoit une étiquette de classe. Cette classe correspond à un

arbre taxonomique dans la complexité et la nature sémantique peuvent varier. La tâche d'EEN s'étend à la reconnaissance de locution nominales (au sens de suite de mots, figée par l'usage, pouvant être substituée à un nom) en regroupant plusieurs mots étiquetés (comme par exemple dans le cas de *Paris* qui est une entité de type localité tout comme *Ville Lumière*, ou encore l'acronyme *TGV* qui décrit le même produit de type véhicule que la locution nominale *Train à Grande Vitesse*). Les campagnes d'évaluation telles que MUC [1], ACE (Doddington *et al.*, 2004), CoNLL (Tjong & Meulder, 2003) et dans le contexte francophone, la tâche d'étiquetage de la campagne ESTER 2 (Galliano *et al.*, 2009), ont permis d'expérimenter des approches variées dans un contexte standardisé et de mesurer leurs performances avec des métriques communes. A la suite des ces campagnes, deux grandes familles de systèmes d'EEN ont fait la démonstration de leur potentiel : celles dérivées de la linguistique computationnelle, recourant à des règles de détection plus ou moins sophistiquées, et celles par apprentissage automatique qui consistent à entraîner un classifieur sur un corpus pré-étiqueté. Ces deux grandes familles d'approches exploitent à des degrés divers des ressources lexicales externes dont la finalité est de renforcer leur capacités de détection d'EN. L'une des caractéristiques récurrente des systèmes d'EEN à base de règles est qu'ils intègrent dans leur processus de détection d'EN des lexiques plus ou moins riches, dont la disponibilité de grand corpus numériques favorise aujourd'hui l'extraction automatisée. Dans ce contexte, nous avons souhaité chercher à évaluer dans quelle mesure un lexique de grande taille, tel que ceux implémentés dans les détecteurs à base de règles, pourrait être utilisé en tant que système rudimentaire de détection par motif pour améliorer un étiqueteur numérique. Cette communication, présente une ressource lexicale automatiquement extraite du corpus Wikipédia, que nous utilisons en tant que motifs rudimentaires de détection d'EN. Nous évaluons les capacités d'un système d'EEN reposant uniquement sur ces motifs de détection, puis nous l'hybridons avec un système d'EEN par apprentissage automatique à base de CRF.

L'article est structuré ainsi : dans la section 2, nous passons en revue les différentes méthodes d'étiquetages d'EN proposées et leur caractéristiques. Nous présentons ensuite dans la section 3 notre proposition de système d'EEN par motifs. Nous décrivons une méthode d'extraction de motifs de détection non ambigus contenus dans un corpus encyclopédique, et la ressource que nous avons obtenue. Puis, dans la section 4, nous évaluons ce système de détection à base de motifs en l'appliquant au corpus de test ESTER 2. Nous fusionnons ses résultats avec ceux obtenus par un EEN à base de CRF et discutons du gain de performance obtenu. Nous concluons dans la section 5 qu'il est possible d'élaborer une méthode peu coûteuse d'introduction de connaissance lexicale en complément des méthodes statistiques et que cette méthode permet d'améliorer la robustesse des système d'EEN.

2 Méthodes d'étiquetage d'entités nommées

Pour extraire les EN d'un texte l'utilisation, en tant que motifs de détection, de lexiques d'entités issus de corpus tels que Wikipédia est une solution applicable (Bunescu & Pasca, 2006; Nothman *et al.*, 2009; Kazama & Torisawa, 2007) mais insuffisante pour plusieurs raisons. En premier lieu, de nombreuses entités à détecter sont absentes de ces corpus de ressources,

1. Voir `http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html`.

fussent-ils aussi vaste que Wikipédia (principe des *mots hors vocabulaires* ou OOV) ; en second lieu, de nombreuses EN sont hautement ambiguës et ne permettent pas la détection directe sans recours à une analyse de leur contexte. Les lexiques apparaissent donc toujours en renfort d'un étiqueteur à règle ou en tant que ressource pour améliorer l'apprentissage d'un étiqueteur statistique.

2.1 Méthodes automatiques par apprentissage

De manière générale, la plupart des approches de reconnaissance automatique reposent sur la théorie des probabilités et peuvent être caractérisées par une méthode générative ou discriminante selon que la distribution de probabilités de la caractéristique à reconnaître est modélisée ou non. Cette différence joue un rôle important dans la tâche d'étiquetage d'EN. En effet, un classifieur discriminant est théoriquement plus précis mais moins capable d'inférer qu'un classifieur génératif et donc moins adaptable aux innombrables possibilités de représentations d'une EN. Ces deux voies possibles d'approches se retrouvent dans la tâche d'étiquetage d'EN : les méthodes génératives, comme par exemple celle reposant sur des modèles de Markov cachés (HMM) (Bikel *et al.*, 1999), et les méthodes discriminantes telles que SVM, Maximum Entropie (MaxEnt) (Borthwick *et al.*, 1998). Les Champs Conditionnels Aléatoires (CRF) (Lafferty *et al.*, 2001) occupent une place à part car ils combinent une nature générative et discriminante. Comme les modèles discriminants, ils tiennent compte, lors de la construction du modèle, des nombreuses observations issues du corpus d'apprentissage et les corrèlent entre elles lors de l'entrainement. Mais à l'instar des modèles génératifs, les CRF probabilisent les décisions en fonction de la position des séquences d'apprentissage. Ce mode de fonctionnement hybride qui favorise l'inférence, c'est-à-dire la reconnaissance par un classifieur CRF d'une entité qu'il n'a jamais observée dans le corpus d'apprentissage, mais aussi la précision en utilisant les données d'apprentissage pour discriminer, explique pourquoi des études (Raymond & Riccardi, 2007) montrent régulièrement que les systèmes à base de CRF sont plus performants que ceux à base de HMM, de SVM ou de MaxEnt pour résoudre la tâche d'étiquetage d'EN avec un système automatique. La performance d'un système CRF est aussi très largement dépendante de la formation des échantillons d'apprentissage qui vont lui être soumis (McCallum & Li, 2003). Pourtant, dans un contexte expérimental normalisé tel que celui de la campagne ESTER, quelque soit le soin apporté à la sélection des échantillons et à la formation du corpus d'apprentissage, il est régulièrement observé (voir par exemple (Raymond & Fayolle, 2010)) que les performances du CRF demeurent inférieures à celles d'un EEN à base de règles sur des données non bruitées.

2.2 Méthodes à règles et automates

Les systèmes d'EEN à base de règles utilisent des méthodes linguistiques ou des automates à états finis pour identifier les EN dans un texte. Certains de ces systèmes sont fortement inspirés par les méthodes linguistiques. Tel XIP (Brun *et al.*, 2010) qui en partant d'un ensemble de règles, identifie les syntagmes noyaux et extrait les relations de dépendance syntaxiques pour localiser des EN. L'analyse syntaxique peut d'ailleurs être plus ou moins profonde pour détecter des structures qui fiabiliseront le processus de détection des EN. D'autres systèmes

{ERIC.CHARTON, MICHEL.GAGNON, BENOIT.OZELL}@POLYMTL.CA

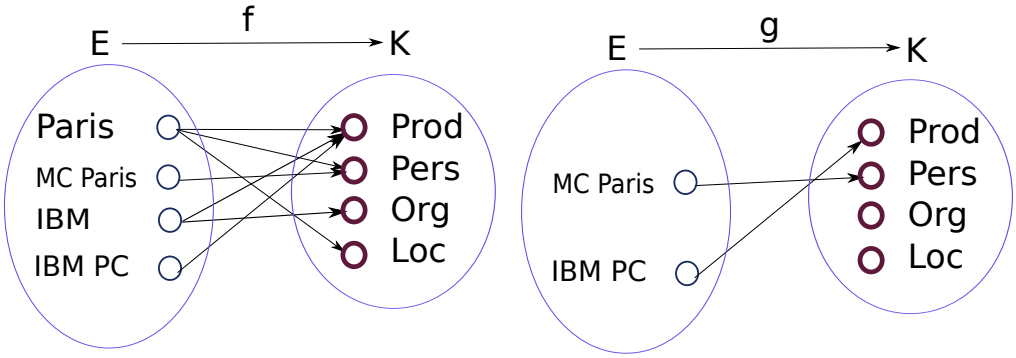


FIGURE 1 – Représentation du principe de réduction de l'ensemble de formes de surface disponibles par identifications des motifs non discriminants.

d'EEN à règles se contentent d'automates de détection plus ou moins sophistiqués. Ainsi, le système CasEN, déployé lors de la campagne ESTER 2 (Nouvel *et al.*, 2010) exploite exclusivement des transducteurs, environ 150, qui s'appliquent à reconnaitre des séquences de mots qui contiennent une EN.

La plupart des systèmes de cette famille des systèmes d'EEN non automatiques complètent le processus de détection par un ensemble de règles très spécialisées qui font appel à des ressources lexicales, des informations liées aux parties du discours, et parfois des traits lexico-sémantiques.

La littérature fait apparaitre que la famille des étiqueteurs à règles utilise quasi systématiquement des automates complémentaires pour identifier les expressions numériques, les quantités, les devises, les dates. Ces automates, lorsqu'ils sont appuyés par des ressources lexicales, peuvent être amenés à jouer un rôle plus ou moins important dans l'identification des EN de la famille des noms propres. Mais de manière générale, peu d'explications détaillées sont fournies sur le rôle et l'influence sur les performances globales des systèmes de ces ressources lexicales associées à des automates. L'un des objectifs du travail présenté dans cet article sera de contribuer à l'étude de l'influence des lexiques utilisés directement en tant qu'automates simples pour détecter des EN sans avoir à utiliser le système de règles ou le classifieur numérique.

3 Système proposé

Nous proposons d'enrichir un système d'EEN à base de CRF avec un ensemble de motifs de détection extraits automatiquement depuis l'encyclopédie Wikipédia. Notre idée est qu'il est possible d'améliorer les performances d'un système d'EEN par apprentissage automatique en lui associant un module qui détecterait les graphies non ambiguës des EN. Nous souhaitons ainsi évaluer à quel point les motifs issus de ressources lexicales qui renforcent les systèmes à base de règle influent sur les performances globales de cette famille d'étiqueteurs. Cette proposition permet d'envisager d'intégrer une connaissance lexicale rudimentaire dans un processus d'EEN par CRF pour le rapprocher des performances des systèmes linguistiques et à base de règles. Les motifs de détection que nous proposons d'extraire sont rudimentaires et ne concernent que des EN non ambiguës. Leur principe fonctionnel peut être illustré par ces exemples :

– Si nous prenons l'exemple du nom **Montréal**, celui-ci correspond à plusieurs entités distinctes (*Montréal (Québec)*, *Montréal (Ardèche)*) qui sont de classe identique, à savoir des localités (étiquette LOC). Considérons une graphie associée à une EN, que nous appelons forme de surface de cette EN. Il est possible d'exploiter une forme de surface *Montréal* en tant que motif pour détecter plusieurs entités nommées d'identités différentes mais qui sont toute de type LOC.
– La forme de surface *Paris*, en revanche, est associée à plusieurs entités de type localité telles que *Paris, France* ou *Paris, Texas* (LOC), mais aussi à des noms de personnes (PERS.HUM) *Antoine Paris*, *Paris Hilton*, de navires ou de produits (le Paquebot *Paris* (PROD.VEHICLE)) ou l'album musical *Paris* (PROD.DIV)). La forme de surface *Paris* est donc hautement ambiguë et ne peut être utilisée en tant que motif de détection susceptible d'identifier une entité et de lui attribuer une classe d'étiquetage valable.
– On pourra en revanche conserver les formes de surfaces intégrant un élément ambigu, mais plus longues - de type bi-grammes à n-grammes, si elles sont non ambiguës : ainsi les formes de surface *MC Paris* (nom de personne) ou *SS Paris* (nom de véhicule) peuvent être utilisées en tant que motifs de détection.

On peut formaliser d'après ces exemples que l'ensemble des motifs de détection non ambigus est le sous ensemble injectif constitué des relations entre l'ensemble des formes de surfaces et l'ensemble des classes qui leur sont reliées, si et seulement si tout élément de l'ensemble d'arrivée K possède au plus un antécédent par g de l'ensemble de départ E (voir figure 1).

3.1 Extraction automatique de motifs de détection

Nous souhaitons extraire les motifs de détection depuis l'encyclopédie Wikipédia. Nous avons présenté dans (Charton & Torres-Moreno, 2009) un système capable de produire, d'après un corpus encyclopédique tel que Wikipédia, une ressource multilingue de concepts que nous avons intitulé *métadonnées*. Ces *métadonnées* incluent des noms propres, des noms communs, des entités nommées, ainsi que des locutions rigoureusement classées selon la norme taxonomique de la campagne ESTER 2 et associées chacune à plusieurs formes de surface. La proportion des ensembles de fiches encyclopédiques transformées en *métadonnées* affectées à chaque classe est présentée dans la table 1. Pour chaque *métadonnée*, les formes de surfaces qui permettent d'écrire le concept encyclopédique sont collectées dans les éditions polonaise, italienne, française, anglaise, espagnole, allemande et italienne de Wikipédia. La quantité totale de formes de surfaces disponibles est indiquée dans la table 2. Un exemple d'ensemble de formes de surfaces contenu dans une *métadonnée* est montré dans la figure 2. Cet exemple [2] montre l'ambiguïté de certains motifs collectés dans le corpus encyclopédique. On peut observer dans cet exemple que la forme *Renault* est hautement ambiguë (puisqu'elle caractérise également un nom de personne dans l'encyclopédie), en revanche, des séquences telles que *Renault Nissan Group, Renault Motor* collectées depuis Wikipédia en Anglais ou encore le sigle *RNUR* collecté depuis Wikipédia en Polonais, sont des motifs de détection non ambigus. Nous obtenons ainsi un ensemble de paires composées de motifs de détections associés à une classe unique, que nous allons utiliser trivialement dans un étiqueteur d'EN à expression régulières.

2. Consultable en ligne sur `http://www.nlgbase.org/perl/display.pl?query=Renault&search=EN`

{ERIC.CHARTON, MICHEL.GAGNON, BENOIT.OZELL}@POLYMTL.CA

Qté	Contenu	Classe taxonomique
3515	Fonctions et titres	FONC
753629	Lieu	LOC
346218	Organisations	ORG
972663	Personne	PERS
411569	Produit	PROD
14294	Date	TIME
621082	Contenu encyclopédique	UNK
3122970	*métadonnées* disponibles	

TABLE 1 – Quantité de *métadonnées* disponibles pour chaque classe d'étiquetage.

3 122 970	métadonnées disponibles
8 142 183	formes de surface disponibles
5 832 730	formes de surface conservées

TABLE 2 – Formes de surfaces non ambiguës extraites depuis les *métadonnées* et utilisables en tant que motifs de détection.

3.2 Étiqueteur CRF

Nous utilisons en tant que *baseline* la première version de l'étiqueteur d'EN mis au point par le LIA pour la campagne ESTER 2 (Béchet & Charton, 2010). Nous l'intitulons CRF-V1. Cet étiqueteur a pour caractéristique d'être entraîné sur un corpus de grande taille préalablement étiqueté par un étiqueteur HMM, en utilisant une ressource lexicale issue des *métadonnées*. Des itérations successives permettent de diminuer le bruit qui subsiste sur le corpus d'entrainement.

La version que nous utilisons ici pour comparer notre système à CRF-V1 est intitulée CRF-V2 et décrite dans (Charton & Torres-Moreno, 2010). Elle complète la phase de préparation par HMM du corpus d'entrainement par un étiquetage supplémentaire utilisant les liens internes de Wikipédia. CRF-V2 est appris sur un ensemble de phrases issues du corpus d'entrainement d'ESTER 2, renforcé par 140 000 phrases étiquetées extraites depuis Wikipédia en français. CRF-V2 est légèrement plus performant que CRF-V1. Il a déjà été déployé dans le système Poly-Co du challenge GREC 2010 [3].

L'architecture complète du système est la suivante. Dans un premier temps, les deux étiqueteurs d'EN, celui à motifs et celui à CRF, sont appliqués sur le document à étiqueter. On obtient par ce moyen deux documents étiquetés que nous nommerons $doc.crf$ et $doc.rule$. L'étiquetage des EN de ces documents est soit un nom de classe k (issu de la taxonomie ESTER) soit le label indéfini UNK appliqué lorsqu'aucune étiquette n'est attribuée. Dans un second temps une fusion de $doc.crf$ et $doc.rule$ est réalisée. Le processus de fusion est trivial et consiste à comparer les étiquettes appliquées à $doc.crf$ et $doc.rule$ en donnant priorité à l'un des documents. L'algorithme de fusion donne ici priorité aux EN contenues dans $doc.crf$.

3. Voir `http://www.itri.brighton.ac.uk/research/genchal10/grec/`

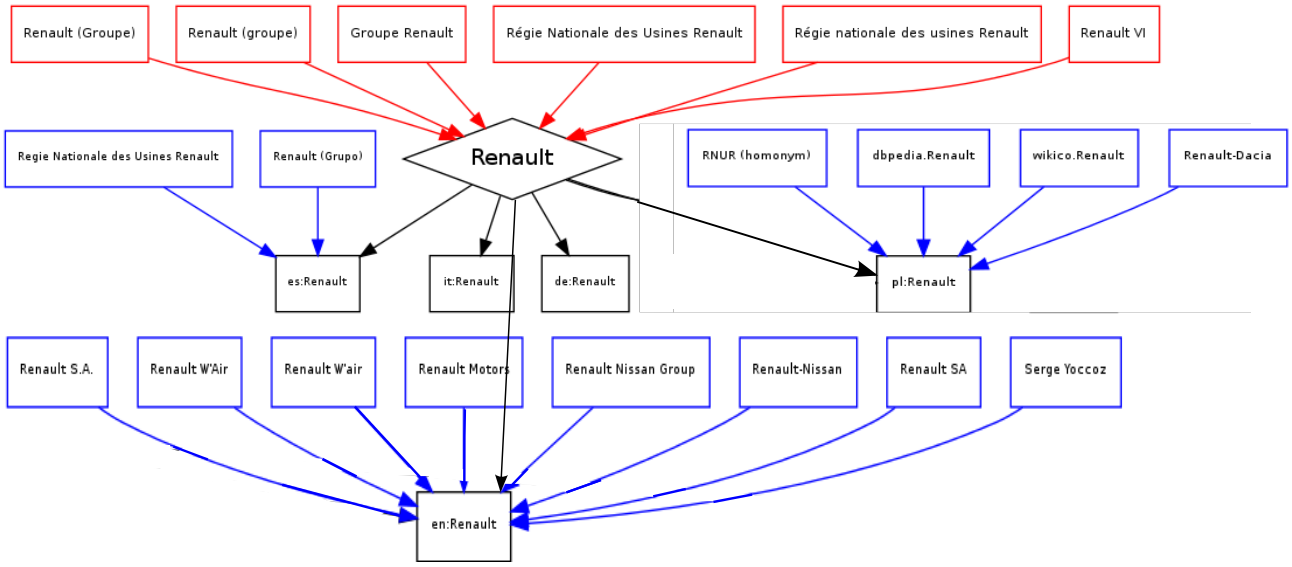


FIGURE 2 – Un exemple de formes de surface collectées pour la *métadonnées* de classe ORG correspondant à la fiche encyclopédique du constructeur automobile *Renault* dans plusieurs éditions linguistiques de Wikipédia.

4 Évaluation et résultats

Nous évaluons les capacités d'un EEN à base de motifs en le comparant aux autres méthodes d'étiquetage dont les résultats sont connus pour un corpus de référence. Puis nous évaluons les performances d'un EEN à CRF renforcé par l'EEN à motifs. Notre expérience vise à mesurer jusqu'à quel point l'introduction de motifs de détection non ambigus et collectés automatiquement peuvent améliorer les performances du CRF, et le cas échéant jusqu'à quel point il permet de rapprocher les performances des CRF de ceux à base de règles, sur des corpus non bruités. Nous utiliserons le corpus de test de la campagne ESTER 2.

4.1 Corpus et mesures de référence

Le corpus complet de la tâche de détection d'EN d'ESTER 2 se compose de 72 heures d'émissions radiophoniques francophones (France-Inter, France Info, RFI, RTM, France Culture, Radio Classique) manuellement transcrites et annotées en EN suivant les conventions des deux campagnes ESTER. La première campagne comportait un jeu de 30 types d'EN réparties en 9 catégories racines, alors que la seconde possède un jeu de 37 types d'entités nommées réparties en 7 catégories racines (personne, fonction, organisation, lieu, fabrication humaine, date et heure, quantités). Seules les catégories racines sont mesurées dans les résultats de référence. La campagne ESTER 2 prévoit plusieurs tâches de reconnaissance d'EN : la première consiste à reconnaître les EN dans la transcription manuelle du corpus de test (NE-Ref). La seconde s'applique à trois transcriptions automatiques dites ASR et dont les taux d'erreurs de reconnaissance de mots vont croissants : 12.11%, 17.83% et 26.09%. La volonté de l'organisateur est ici de tester la précision des systèmes sur NE-Ref qui est non bruité, mais aussi leur ro-

EN	AMOUNT	FONC	LOC	ORG	PERS	PROD	TIME	tous
Qté	239	196	1215	1267	1108	58	1025	5123
précision	0,85	0,61	0,77	0,79	0,93	0,53	0,91	0,86
rappel	0,56	0,559	0,81	0,63	0,75	0,12	0,60	0,718
F-Score	0,68	0,58	0,79	0,70	0,84	0,20	0,73	0,78

TABLE 3 – Résultats par entité à étiqueter du système LIA dit CRF-V1 appliqué au corpus NE-Ref lors de la campagne ESTER 2.

EN	AMOUNT	FONC	LOC	ORG	PERS	PROD	TIME	tous
Qté	239	196	1215	1267	1108	58	1025	5123
précision	0,93	0,818	0,897	0,89	0,97	100	0,95	0,93
rappel	0,86	0,899	0,88	0,83	0,95	0,42	0,95	0,91
F-Score	0,90	0,85	0,89	0,87	0,97	0,59	0,96	0,93

TABLE 4 – Résultats par entité à étiqueter du système XIP de Xerox à base de règles appliqué au corpus NE-Ref lors de la campagne ESTER 2.

bustesse dans le contexte plus difficile des corpus de test ASR bruités de manière croissante.

Les résultats de la campagne ESTER 2 (Galliano *et al.*, 2009) soulignent l'efficacité d'un système EEN à base de règle linguistique sur la transcription de référence (NE-Ref). Sur ce corpus, les deux meilleurs systèmes sont à base de règle, et le troisième est de type automatique à base de CRF. Le tableau 4 présente les résultats obtenus par le meilleur système sur transcriptions de référence (NE-Ref) en termes de Précision, Rappel, F-Score. Le tableau 3 présente les résultats du meilleur système automatique à CRF sur ce même corpus de référence. Nous considérerons les résultats du meilleur système linguistique (XIP) et du meilleur système automatique (CRF-V1) sur le corpus NE-Ref pour situer les performances obtenues par l'hybridation de l'étiqueteur à motif, que nous appellerons ici EEN-M, avec CRF-V2.

Ce plan d'expérience vise à évaluer dans quelle mesure un ensemble de motifs appris automatiquement sur un corpus encyclopédique peut améliorer les performances du système CRF et jusqu'à quel point les performances de ce système CRF amélioré peuvent se rapprocher d'un système d'EEN de nature linguistique à l'état de l'art (en l'occurrence le système XIP). Notre expérience consistera à appliquer au corpus NE-Ref d'ESTER 2 les détecteurs EEN-M et CRF-V2 et à fusionner leurs résultats puis à mesurer les performances de chaque élément de notre système.

4.2 Résultats

Le tableau 5 expose les résultats des différents composants de notre système d'EEN. Il indique pour chaque jeu d'étiquettes du corpus de test NE-Ref ESTER 2 les performances individuelles de chaque composant. Dans la section motif du tableau, qui présente les résultats de l'étiqueteur par détection de motif EEN-M, on remarque que la classe AMOUNT n'est pas traitée par ce composant d'étiquetage car non observée dans les *métadonnées* utilisées pour collecter les motifs. La classe TIME qui correspond aux dates dans le corpus ESTER 2 est pour ce qui la

	EN	AMOUNT	FONC	LOC	ORG	PERS	PROD	TIME	tous
	Qté	239	196	1215	1267	1108	58	1025	5123
EEN-M	précision	x	0,85	0,73	0,94	0,98	0.11	0,96	0,88
/ Motifs	rappel	x	0,30	0,32	0,27	0,50	0,07	0,36	0,34
	F-Score	x	0,43	0,44	0,42	0,66	0,08	0,53	0,48
CRF-V2	précision	0,90	0,99	0,77	0,92	0,94	0.38	0,97	0,88
	rappel	0,70	0,46	0,90	0,61	0,93	0,25	0,69	0,74
	F-Score	0,79	0,63	0,83	**0,73**	0,93	**0,30**	0,89	0,80
Hybride	précision	0.90	0.91	0.76	0,91	0,96	0.27	0,96	0,88
	rappel	0,70	0,55	0,92	0,60	0,93	0,25	0,83	0,78
	F-Score	**0,79**	**0,69**	**0,83**	0,72	**0,94**	0,26	**0,90**	**0,83**
CRF-V1	F-Score	0,68	0,58	0,79	0,70	0,84	0,20	0,73	0,78

TABLE 5 – Résultats détaillés du système EEN à motifs de détection, CRF (dit CRF-V2) et hybride, comparé au système CRF du LIA (dit CRF-V1) ayant obtenu les meilleures perfomances sur le corpus de test NE-Ref de la campagne ESTER 2.

concerne traitée car cette classe de contenu est représentée dans Wikipédia et donc modélisée dans les métadonnées [4]. On note que EEN-M offre une couverture de détection des EN relativement faible (rappel de 0,34) mais une précision supérieure à celle de CRF-V1. Cette précision est également supérieure à celle de l'étiqueteur à règles linguistique présenté dans le tableau 4 pour les classes FONC, ORG et TIME. Il est important de remarquer les performances inférieures de EEN-M sur la détection de la classe LOC qui sont attribuables à l'impossibilité pour EEN-M de traiter la différence entre les notions LOC.ADMI et ORG.GSP (un nom toponymique peut désigner une localité ou une organisation géo-politique dans le corpus ESTER 2) par un système à motif.

La section CRF-V2 présente les résultats de l'étiqueteur CRF amélioré tel que décrit dans la section 3. On observe que les performances de cet étiqueteur sont légèrement meilleures que celles de l'étiqueteur déployé par le LIA lors de la campagne ESTER 2, et dont les résultats sont indiqués dans la section CRF-V1 du tableau. Une comparaison plus détaillée avec le tableau 3 montre que les performances de CRF-V2 sont améliorées globalement tant pour la précision que le rappel, avec les mêmes difficultés de modélisation des séquences d'EN de type PROD. L'amélioration des performances du système CRF-V2 appliqué sur NE-Ref, par rapport à CRF-V1, n'est pas l'objet de cette communication, mais doivent être commentées ici car l'amélioration de la précision joue un rôle sur le processus de fusion. Les expériences de fusions que nous avons menées entre les résultats produits par CRF-V1 et ceux de EEN-M nous ont montré une très légère minoration des performances globales du système (les moindres performances de CRF-V1 étant compensées par l'introduction des EN détectées par EEN-M).

L'hybridation de EEN-M et CRF-V2 indiquée dans la ligne *Hybride* du tableau 5, est le résultat de la fusion entre les deux sorties de ces systèmes. Elle montre un gain de performance de 3% sur CRF-V2 seul, et de plus de 5% par rapport à CRF-V1 déployé lors de ESTER 2.

4. Voir par exemple la catégorie `http://fr.wikipedia.org/wiki/Categorie:Jour_de_septembre` et une *métadonnée* telle que `http://www.nlgbase.org/perl/display.pl?query=2septembre&search=FR`

4.3 Discussion

Il apparaît que le système de détection hybride à base de motifs et de CRF proposé améliore substantiellement les performances du système CRF-V1 déployé lors de la campagne ESTER 2 sur le corpus de référence NE-Ref.

On notera que les expériences d'hybridation de *métadonnées* et du système CRF qui avaient été employées lors de cette campagne, qui reposaient sur une détection de motifs non désambiguïsés associée à un calcul de similarité cosinus entre le contexte du motif et les métadonnées, n'avaient produit qu'un gain de 1% sur un F-Score du CRF de 0,77 (voir sur ce point (Béchet & Charton, 2010)). Le module de détection de motifs expérimenté dans cet article introduit un gain de 3% sur un F-Score de CRF de 0,80. Ce gain souligne le potentiel de la méthode. On observe par ailleurs que notre proposition réduit globalement l'écart de performance entre un système à règles et un système statistique complété par des motifs, sur une transcription de référence corrigée telle que NE-Ref de ESTER2.

En terme de précisions, les performances de notre système s'approchent pour plusieurs classes de celles obtenues par le meilleur système à règles linguistiques appliqué sur NE-Ref, lors de la campagne ESTER 2. Ces résultats permettent d'envisager qu'une augmentation de la couverture des formes de surfaces extraites des *métadonnées* (par exemple à la suite d'une augmentation de la quantité de formes de surfaces disponibles dans Wikipédia) puisse fournir d'autres gains de performance.

5 Conclusion et perspectives

Nous avons décrit une méthode d'introduction dans un système d'étiquetage d'entités nommées à base de CRF d'un composant d'identification d'EN exploitant des motifs de détection collectés automatiquement dans un corpus encyclopédique. Il était apparu lors de la campagne ESTER 2 qu'un système CRF correctement entrainé pouvait obtenir les meilleurs résultats sur un corpus bruité, mais que les systèmes d'EEN à règles linguistiques étaient plus performants sur des corpus non bruités. Nous avons donc cherché à évaluer dans quelle mesure le renforcement d'un CRF par des motifs de détection simples pouvait réduire l'écart de performances entre un étiqueteur CRF et un étiqueteur à base de règle sur un document textuel non bruité. Nous avons montré que notre proposition pouvait amener un gain de performances global important sur un système CRF, et améliorer de manière conséquente sa précision. La solution que nous proposons améliore la robustesse d'un système CRF sur des corpus non bruités et réduit l'écart avec un système d'EEN linguistique tout en conservant au CRF son faible coût de développement, l'intégralité du processus d'entrainement de notre système demeurant automatique. La précision du système que nous avons élaboré et sa facilité de déploiement nous ont permis de l'entrainer dans plusieurs versions linguistiques (Français, Espagnol et Anglais) et de l'exploiter en tant que module dans des applications qui prolongent la tâche d'étiquetage d'entités nommées. Nous travaillons en particulier sur la détection de co-références et avons à ce titre déployé cet étiqueteur dans sa version anglaise en tant que composant de l'architecture de détection de co-référence du challenge Grec 2010 ou il a obtenu des résultats satisfaisants [5].

5. Les ressources décrites sont disponibles sous forme d'API et en téléchargement sur `www.nlgbase.org`.

Références

BÉCHET F. & CHARTON E. (2010). Unsupervised knowledge acquisition for extracting named entities from speech. In *ICASSP 2010*, Dallas : ICASSP.

BIKEL D., SCHWARTZ R. & WEISCHEDEL R. (1999). An algorithm that learns whats in a name. *Machine learning*, **7**.

BORTHWICK A., STERLING J., AGICHTEIN E. & R (1998). Exploiting diverse knowledge sources via maximum entropy in named entity. *Proc. of the Sixth*, p. 152–160.

BRUN C., EHRMANN M. & MAUPERTUIS C. D. (2010). Un système de détection d'entités nommées adapté pour la campagne d'évaluation ESTER 2. In *TALN 2010*, volume 2.

BUNESCU R. & PASCA M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, volume 6.

CHARTON E. & TORRES-MORENO J. (2009). Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées. In *Taln 2009*, volume 1, p. 24–26 : TALN.

CHARTON E. & TORRES-MORENO J. (2010). NLGbAse : a free linguistic resource for Natural Language Processing systems. In LREC, Ed., *LREC 2010*, number 1, Matla : Proceedings of LREC 2010.

DODDINGTON G., MITCHELL A., PRZYBOCKI M., RAMSHAW L., STRASSEL S. & WEISCHEDEL R. (2004). The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proceedings of LREC*, volume 4, p. 837–840 : Citeseer.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. In *International Speech Communication Association conference 2009*, p. 2583–2586 : Interspeech 2010.

KAZAMA J. & TORISAWA K. (2007). Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 698–707.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, p. 282–289 : Citeseer.

MCCALLUM A. & LI W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* -, p. 188–191, Morristown, NJ, USA : Association for Computational Linguistics.

NOTHMAN J., MURPHY T. & CURRAN J. (2009). Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, number April, p. 612–620 : Association for Computational Linguistics.

NOUVEL D., ANTOINE J., FRIBURGER N. & MAUREL D. (2010). An analysis of the performances of the casen named entities recognition system in the ester2 evaluation campaign. *LREC 2010*.

{ERIC.CHARTON, MICHEL.GAGNON, BENOIT.OZELL}@POLYMTL.CA

RAYMOND C. & FAYOLLE J. (2010). Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In *Traitement Automatique des Langues Naturelles*, volume 1, p. 19–23.

RAYMOND C. & RICCARDI G. (2007). Generative and discriminative algorithms for spoken language understanding. In *Proceedings of Interspeech2007, Antwerp, Belgium*, p.$\tilde{2}$: Citeseer.

TJONG E. & MEULDER F. D. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. In *In CoNLL*.

Parole

Comparaison et combinaison d'approches pour la portabilité vers une nouvelle langue d'un système de compréhension de l'oral

Bassam Jabaian [1,2], Laurent Besacier [1], Fabrice Lefèvre [2]

(1) LIG, University Joseph Fourier, Grenoble - France
(2) LIA, University of Avignon, Avignon - France

{bassam.jabaian,laurent.besacier}@imag.fr , fabrice.lefevre@univ-avignon.fr

Résumé

Dans cet article, nous proposons plusieurs approches pour la portabilité du module de compréhension de la parole (SLU) d'un système de dialogue d'une langue vers une autre. On montre que l'utilisation des traductions automatiques statistiques (SMT) aide à réduire le temps et le cout de la portabilité d'un tel système d'une langue source vers une langue cible. Pour la tache d'étiquetage sémantique on propose d'utiliser soit les champs aléatoires conditionnels (CRF), soit l'approche à base de séquences (PH-SMT). Les résultats expérimentaux montrent l'efficacité des méthodes proposées pour une portabilité rapide du SLU vers une nouvelle langue. On propose aussi deux méthodes pour accroître la robustesse du SLU aux erreurs de traduction. Enfin on montre que la combinaison de ces approches réduit les erreurs du système. Ces travaux sont motivés par la disponibilité du corpus MEDIA français et de la traduction manuelle vers l'italien d'une sous partie de ce corpus.

Abstract

In this paper we investigate several approaches for language portability of the spoken language understanding (SLU) module of a dialogue system. We show that the use of statistical machine translation (SMT) can reduce the time and the cost of porting a system from a source to a target language. For conceptual decoding we propose to use even conditional random fields (CRF) or phrase based statistical machine translation PB-SMT). The experimental results show the efficiency of the proposed methods for a fast and low cost SLU language portability. Also we proposed two methods to increase SLU robustness to translation errors. Overall we show that the combination of all these approaches reduce the concept error rate. This work was motivated by the availability of the MEDIA French corpus and the manual translation of a subset of this corpus into Italian.

Mots-clés : Système de dialogue, compréhension de la parole, portabilité à travers les langues, traduction automatique statistique

Keywords: Spoken Dialogue Systems, Spoken Language Understanding, Language Portability, Statistical Machine Translation.

1 Introduction

La portabilité d'un système de dialogue d'une langue vers une autre est une tâche difficile qui a fait l'objet, récemment, de plusieurs recherches. Certains composants d'un système de dialogue, tel que le gestionnaire de dialogue, sont relativement indépendants de la langue, ce qui n'affectent pas le processus de portabilité. Cependant, d'autres modules tel que le module de compréhension automatique de la parole (*Spoken Language Understanding, SLU*), doivent être réadaptés pour chaque nouvelle langue cible considérée. Dans cette étude, nous nous intéressons particulièrement à la portabilité d'un système de compréhension automatique de la parole vers une nouvelle langue.

Des travaux récents ont proposé d'utiliser des méthodes stochastiques pour la compréhension automatique de la parole. Ces méthodes sont des alternatives efficaces aux méthodes à base de règles; elles réduisent les besoins en expertise humaine tout en ayant la capacité de produire efficacement des réseaux d'hypothèses ou des listes de N-meilleures (N-best) (Suenderman, Liscombe, 2009) (Hahn, Lehnen, Raymond, Ney, 2008) (Raymond, Riccardi, 2007) (Wang, Acero, 2006) (Schwartz, Miller, Stallard, Makhoul, 1996). L'apprentissage de tels modèles nécessite un corpus annoté qui représente une couverture complète de la sémantique du domaine. Le portage d'un tel modèle vers une nouvelle langue consiste à transférer les connaissances présentes dans le corpus annoté en langue source vers une nouvelle langue avec un minimum de temps et d'effort humain. Par la suite, nous nommerons «langue source » la langue d'origine du système NLU et « langue cible », la langue vers laquelle le système doit être porté.

Récemment, quelques études ont montré que l'utilisation de la traduction automatique à différents niveaux du processus de compréhension peut aider au portage d'un système SLU vers une nouvelle langue (Suenderman, Liscombe, 2009) (Servan, Camelin, Raymond, Bechet, De Mori, 2010) (Lefèvre, Mairesse, Young, 2010) (Jabaian, Besacier, Lefèvre, 2010). Par exemple, dans (Suenderman, Liscombe, 2009) les auteurs proposent de traduire automatiquement les données de la langue source vers la langue cible, puis de re-apprendre une grammaire stochastique pour effectuer l'interprétation dans la langue cible. Une autre possibilité est de considérer que la sémantique d'un domaine est indépendante de la langue. Dans ce cas, une solution est de traduire le corpus d'apprentissage vers la langue cible et d'inférer les balises sémantiques associées au corpus traduit. Un système SLU stochastique peut ensuite être re-entraîné sur ce nouveau corpus annoté en langue cible. Comme décrit dans (Servan, Camelin, Raymond, Bechet, De Mori, 2010), les phrases du corpus d'apprentissage sont composées d'un ou plusieurs segments annotés sémantiquement. Traduire le corpus d'apprentissage en conservant l'information de segmentation en segments permet alors un appariement direct des segments traduits avec les étiquettes sémantiques. Les auteurs de (Servan, Camelin, Raymond, Bechet, De Mori, 2010) montrent qu'un portage du français vers l'italien est possible en utilisant cette approche, avec une traduction manuelle ou automatique. Le portage de l'annotation est encore moins difficile lorsqu'on utilise des méthodes qui n'ont pas besoin d'annotation sémantique au niveau mot ou segment. Par exemple, dans (Lefèvre, Mairesse, Young, 2010), le modèle proposé ne nécessite pas d'informations d'alignement.

Le choix d'une approche dépend de considérations techniques et également des caractéristiques du domaine ainsi que des données disponibles. Disposer de données manuellement traduites ou annotées, disposer d'annotateurs ou d'outils spécifiques pour la langue cible, peut faire la différence quant au choix de l'approche. Dans cet article, nous proposons plusieurs approches pour le portage d'un système de compréhension automatique de la parole vers une nouvelle langue. La langue source est le français étant donné que nous travaillons sur le corpus MEDIA (Bonneau -Maynard, Rosset, Ayache, Kuhn, Mostefa, 2005) et la langue cible considérée est l'italien puisque nous disposons également, au départ, d'une partie du corpus MEDIA traduite en italien. Nous sommes conscients de la proximité des langues source et cibles dans cette étude, mais ce choix est guidé par les données disponibles au départ. Le portage vers l'arabe est aussi envisagé et fait l'objet de travaux en cours. Les approches proposées dans cet article sont complètement automatiques et sans aucune supervision humaine lors du processus de portage.

Plus précisément, le but de cet article est de :

- proposer et évaluer différentes approches utilisant la traduction automatique pour porter un système SLU vers une nouvelle langue,

- considérant ensuite la meilleure approche obtenue, accroître sa robustesse aux erreurs de traduction.

Dans ces travaux, les modèles utilisés sont les champs aléatoires conditionnels (*Conditional Random Fields, CRF*) pour la compréhension automatique et l'approche à base de séquences (*phrase-based statistical machine*

translation) pour la traduction automatique. Les CRF (Lafferty, McCallum, Pereira, 2001) sont connus pour être très performants sur des tâches d'étiquetage temporel (annotation en entités nommées, étiquetage syntaxique, etc). Ils nécessitent un corpus annoté au niveau mots. Pour porter notre système vers une nouvelle langue, nous avons proposé plusieurs méthodes qui diffèrent selon le moment où est utilisé le module de traduction. En effet, un système de compréhension peut être porté soit au niveau du test (*TestOnSource*), en conservant le système SLU en langue source et en traduisant simplement les données de test en langue cible vers la langue source. La seconde possibilité consiste à porter le système au niveau de l'apprentissage *(TrainOnTarget)* en construisant un nouvel étiqueteur sémantique dans la langue cible. Pour cela, nous traduisons automatiquement l'intégralité du corpus d'apprentissage de la langue source vers la langue cible, puis nous inférons l'annotation sémantique de ce corpus pour les données traduites. Pour ce faire, nous proposons deux méthodes différentes. La première consiste à l'aide d'un système de traduction probabiliste source-cible, à traduire chaque phrase source annotée, segmentée en segments, et d'utiliser les segments traduits, associés aux étiquettes sémantiques, pour construire un nouveau système SLU en langue cible. La seconde approche utilise les informations extraites des alignements mot-à-mot pour inférer une relation mot_cible-étiquette_sémantique (plus de détails seront données dans la section 2).

Une autre approche, complètement différente, consiste à voir le processus de compréhension comme une tâche de traduction automatique d'une chaîne de mots vers une chaîne d'étiquettes sémantiques. Dans ce cas, le modèle de compréhension est une table de traduction mots-concepts. L'approche à base de séquences (*phrase-based*) (Koehn, Och, Marcu, 2003) nécessite des données alignées au niveau phrase avant le processus d'apprentissage (dont la première étape sera un alignement automatique en mots utilisant les modèles IBM). Dans ce cas, la portabilité vers une nouvelle langue consiste simplement à traduire en langue cible la partie "mots" du corpus d'apprentissage mots-concept sans modifier les concepts.

Pour répondre au deuxième point (robustesse), et puisque nous allons voir que la méthode *TestOnSource* donne les meilleures performances, nous proposons quelques méthodes pour augmenter la robustesse aux erreurs de traduction du système porté. Pour cela, une première approche présentée consiste à re-entrainer le système de compréhension, fondé sur les CRF, sur des données bruitées représentant les erreurs potentielles de traduction. La deuxième approche consiste à utiliser une post-édition automatique statistique (*Statistical Post Edition, SPE*) dans la langue-source pour tenter de corriger automatiquement les sorties issues du système de traduction automatique, avant de les envoyer à l'étiqueteur sémantique. Pour finir, nous proposons aussi dans cet article de combiner toutes les approches proposées dans cette étude, afin de réduire le taux d'erreurs de compréhension.

Cet article est structuré de la façon suivante : la section 2 présente en détail les approches que nous proposons pour porter un système de compréhension vers une nouvelle langue. La section 3 décrit deux solutions pour améliorer la robustesse de notre meilleur système porté, aux erreurs de traduction automatique. Le corpus MEDIA et les outils utilisés sont décrits dans la section 4 tandis que la section 5 présente les résultats expérimentaux obtenus. Finalement, conclusion et perspectives sont présentées dans la section 6.

2 Différentes méthodes pour porter un système de compréhension d'une langue vers une autre

Dans un système de dialogue, le rôle du processus de compréhension est d'extraire une liste d'hypothèses d'étiquettes de concepts à partir d'une phrase en entrée. Ces concepts représentent la sémantique de l'information existant dans la phrase en entrée. Les modèles de compréhension développés dans cette étude sont entraînés sur le corpus MEDIA, annoté en concepts sémantiques (voir section 4).

La génération automatique de ces concepts à partir d'une séquence de mots par des méthodes stochastiques telle que décrite dans (Raymond, Riccardi, 2007), peut être résumée de la façon suivante :

Soit C= c_1,...,c_n une séquence d'étiquettes sémantiques qui peut être associée initialement à la séquence de mots W= w_1,…,w_n ; pour chaque concept, une séquence de mots de W est associée et une étiquette est attribuée à chaque mot. Cette étiquette correspond au concept sémantique c_i et à la position de w_i.

Plusieurs études ont proposé de comparer différentes méthodes pour entraîner un modèle de compréhension de la parole (Hahn, Lehnen, Raymond, Ney, 2008) (Raymond, Riccardi, 2007). Dans cet article, nous proposons d'utiliser et d'évaluer deux approches état-de-l'art.

La première est fondée sur les champs aléatoires conditionnels (*Conditional Random Fields, CRF*), qui ont besoin d'un corpus annoté au niveau mot pour être entraînés. La seconde utilise une approche de traduction

probabiliste fondée sur les séquences (*Phrase-Based Statistical Machine Translation, PB-SMT*) et nécessite un corpus d'apprentissage annoté au niveau des phrases.

2.1 Champs aléatoires conditionnels (CRF)

Les CRF ("Conditional Random Fields" ou "Champs Aléatoires (Markoviens) Conditionnels") sont une famille de modèles graphiques introduits récemment (Lafferty, McCallum, Pereira, 2001). Ils permettent d'apprendre à annoter des données, en se basant sur un ensemble d'exemples déjà annotés. Les CRF ont le plus souvent été utilisés dans le domaine du TAL, pour étiqueter des séquences d'unités linguistiques. Ces modèles possèdent les avantages des modèles génératifs et discriminants. En effet, comme les classifieurs discriminants, ils peuvent manipuler un grand nombre de descripteurs, et comme les modèles génératifs, ils intègrent des dépendances entre les étiquettes de sortie et prennent une décision globale sur la séquence. Par rapport aux modèles de Markov Cachés (HMMs), les CRF ont par ailleurs l'avantage de relâcher certaines hypothèses d'indépendance.

Dans notre cas, pour apprendre notre modèle CRF, les données d'apprentissage doivent être représentées selon le formalisme BIO décrit dans (Raymond, Riccardi, 2007), qui indique les frontières entre les concepts sémantiques selon l'exemple ci-dessous :

"Je voudrais réserver un hôtel à Paris "

Sera représenté par la séquence de couples (w,c):

(je, B_command-tache) (voudrais, I_command-tache) (réserver, I_command-tache) (un, B_Objet) (hôtel, I_objet) (à, B_loc-ville) (Paris, I_loc-ville)

La probabilité d'une sequence de concepts, étant donnée une séquence de mots est alors calculée par :

$$p(c_1^N|w_1^N) = \frac{1}{Z}\prod_{n=1}^{N} H\,(c_{n-1}, c_n, w_{n-2}^{n+2})$$

avec

$$H(c_{n-1}, c_n, w_{n-2}^{n+2}) = \sum_{m=1}^{M} \lambda_m . h_m\,(c_{n-1}, c_n, w_{n-2}^{n+2})$$

H est un modèle log-linéaire fondé sur des fonctions caractéristiques $h_m\,(c_{n-1}, c_n, w_{n-2}^{n+2})$ qui représentent l'information extraite du corpus d'apprentissage ; les poids λ du modèle log-linéaire sont estimés lors de l'apprentissage et *Z* est un terme de normalisation définit tel que :

$$Z = \sum_{m=1}^{M}\prod_{n=1}^{N} H\,(c_{n-1}, c_n, w_{n-2}^{n+2})$$

Afin de porter notre système de compréhension fondé sur les CRF (note SLU/CRF par la suite) d'une langue à une autre, nous avons proposé plusieurs approches qui diffèrent selon le moment où est appliqué le processus de transfert entre les langues.

2.1.1 Portage au niveau du test (*Test On Source*)

Dans cette approche, nous supposons qu'un système SLU est disponible en langue source, et nous utilisons un système de traduction automatique probabiliste pour traduire les phrases de test en langue cible vers la langue source. Ces traductions sont ensuite les entrées du système SLU original. En d'autres termes, nous portons le système « au niveau du test » sans modifier le processus d'apprentissage du système SLU. Cette technique sera dénommée *TestOnSource* dans la suite de cet article. Elle a l'avantage d'être très simple mais ses performances dépendront, bien évidemment, des performances du système de traduction automatique utilisé pour revenir de la langue cible à la langue source.

2.1.2 Portage au niveau de l'apprentissage (*Train On Target*)

Cette approche consiste à re-entraîner un système SLU en langue cible. L'idée générale est de traduire le corpus d'apprentissage de la langue source vers la langue cible et d'inférer les étiquettes sémantiques associées. Pour inférer l'annotation sémantique, nous proposons deux approches différentes :

1. Traduire avec des tags XML (*Tagged Translation*):
   Dans cette approche, le corpus d'apprentissage est traduit en prenant en compte la segmentation en « segments sémantiques » (un « segment » est composé potentiellement de plusieurs mots mais correspond à une et une seule étiquette sémantique). Pour cela, nous utilisons une option (*-xml-input*) du décodeur MOSES (Koehn, Hoang, Birch, Callisonburch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, Herbst, 2007), qui force la segmentation d'une phrase à traduire, cette segmentation étant décrite par des tags XML. Ainsi, en sortie, nous obtenons chaque phrase du corpus d'apprentissage traduite, ainsi qu'une projection des tags XML de la source vers la cible. L'exemple donné précédemment peut alors être représenté sous la forme suivante :

   <tag c=command_tache > Je voudrais réserver </tag> <tag c=objet > un hôtel </tag> <tag c=localisation_ville> à Paris </tag>

   En utilisant l'option de MOSES qui prend en compte les tags XML comme information de segmentation, nous obtenons la sortie traduite suivante :

   <tag c= command_tache > vorrei prenotare </tag> <tag c=objet> un hotel </tag> <tag c= localisation_ville> a Parigi </tag>

   Tout le corpus d'apprentissage est traduit de cette façon puis re-formaté au format BIO avant un nouvel apprentissage du modèle CRF de compréhension en langue cible.

2. Projections des concepts sémantiques d'une langue à l'autre en utilisant un alignement en mots (*Alignment*):

   L'alignement automatique en mots est une étape importante dans le processus de construction d'un modèle de traduction probabiliste. Plusieurs boites à outils existent pour cette tâche telles que GIZA++ (Och, Ney, 2000) qui utilise les modèles IBM et HMM, ou Berkeley aligner (Liang, Taskar, Klein, 2006) qui repose sur une méthode d'alignement par consensus (*alignment by agreement*).

   Pour projeter les concepts sémantiques d'une langue à l'autre, on peut utiliser les informations d'alignement bilingue en mots. Plus précisément, la première phase consiste à aligner automatiquement le corpus parallèle source-cible. Ensuite, comme le corpus source est déjà annoté sémantiquement, il est possible d'apparier les étiquettes sémantiques aux mots en langue cible en utilisant l'information d'alignement. Certains cas ambigus demeurent cependant, comme illustré dans la figure 2 (alors que la figure 1 présente un cas où la projection est évidente). Dans cet article, l'aligneur utilisé est *Berkeley Aligner*.

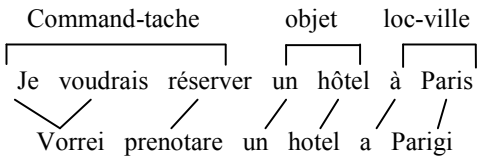


Figure 1 : Exemple de projection des tags sémantiques du français vers l'italien

Pour faire cette projection, nous avons développé un algorithme qui parcourt la phrase en langue cible et associe aux mots la bonne étiquette sémantique. Pour les cas ambigus où un mot cible est aligné avec plusieurs mots sources correspondent à deux concepts différents (voir figure 2), nous devons prendre une décision sur quel concept doit être associé au mot cible. Pour cela, notre proposition est de simplement associer le mot cible au premier concept rencontré. Par exemple, sur la figure 2, le mot italien *alla* sera associé au concept *loc-dis* et pas au concept *loc-lieu*. Cette décision, bien qu'arbitraire, a l'avantage d'être cohérente d'un bout à l'autre du corpus si le même cas est rencontré à nouveau.

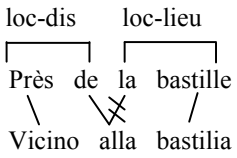


Figure 2 : Exemple de cas ambigu pour la projection des concepts sur les mots cible

2.2 Compréhension par une approche de traduction (PB-SMT)

Afin d'éviter de recourir à un corpus d'entrainement annoté au niveau mot, nous proposons d'utiliser l'approche PB-SMP qui ne nécessite qu'un alignement au niveau des phrases complètes. Dans cette approche, nous considérons que les séquences de concepts sont les traductions des séquences de mots initiales. Ainsi l'étiquetage sémantique est vu comme une tâche de traduction : la meilleure séquence C à partir des mots W est définie par :

$$\hat{C} = argmax_c\, P(C|W) = argmax_c\, P(W|C).P(C)$$

Pour résoudre cette équation sont requis : un modèle de langage de concepts *P(C)* (qui peut être appris à l'aide de SRILM (Stolcke, 2002) sur le corpus de concepts) et d'un model de traduction *P(W|C)* (qui peut être un modèle PB-SMT par exemple). Nous avons utilisé MOSES pour entraîner un tel modèle PB-SMT à partir d'un corpus « parallèle ». Les poids associés à ce modèle sont optimisés par un apprentissage à taux d'erreur minimum (MERT) qui est traditionnellement utilisé pour optimiser le score BLEU. Puis les performances de cette approche PB-SMT de base ont été améliorées en utilisant des caractéristiques de la tâche de compréhension sémantique.

D'abord, en suivant l'hypothèse raisonnable que la sémantique d'une phrase respecte l'ordre dans lequel les mots sont émis, la table de segments est re-entrainée en utilisant une contrainte de monotonie durant l'alignement automatique en mots. Puis, dans la mesure où une difficulté majeure du processus de traduction est l'alignement automatique correct d'un mot du langage source avec le mot correspondant dans le langage cible, nous avons tenté d'aider le processus d'alignement par l'utilisation du formalisme BIO. De cette façon, l'extraction de la table de segments a été obtenue sur un corpus avec un alignement de meilleure qualité. Enfin, la mesure d'évaluation du SLU étant le CER et non le score BLEU, nous avons modifié l'optimisation MERT pour optimiser le CER directement. Finalement, pour éviter les mots hors-vocabulaire (venant principalement de noms de ville absents des données d'entrainement), une liste de villes est ajoutée aux données d'apprentissage et le système *SLU/PB-SMT* est re-entrainé.

3 Accroître la robustesse du SLU aux erreurs de traduction

Nos expériences, ainsi que d'autres travaux (Lefèvre, Mairesse, Young, 2010) (Jabaian, Besacier, Lefèvre, 2010), ont montré que la meilleure méthode pour la portabilité SLU est aussi la plus simple, le *TestOnSource*. La faiblesse principale de cette méthode est que la qualité de l'étiquetage dépend grandement de la qualité de la traduction préalable. Ainsi, le système SLU doit prendre en compte des entrées bruitées par les erreurs de traduction.

De sorte à améliorer la robustesse de l'approche, nous proposons deux méthodes dans ce papier. La première prend en compte le bruit venant de la traduction durant le processus d'apprentissage des modèles SLU ; la seconde corrige automatiquement la sortie du système de traduction avant de la transférer au système SLU. Il est notable que, bien que pas encore évaluées dans ce cadre (par manque de données audio dans la langue cible), les deux méthodes seront tout à fait adaptées pour traiter aussi les erreurs dues à la reconnaissance de la parole dans une tache de compréhension réelle.

3.1 Apprentissage sur des données bruité

Le principe de cette méthode est d'entrainer un modèle SLU (dans le langage source) avec des données additionnelles provenant de la sortie d'un système de traduction automatique. En pratique, nous traduisons les données d'apprentissage disponibles entre les langues cible et source et nous inférons les concepts associés aux données bruitées (en suivant la même méthode que *TrainOnTarget*). Puis nous ajoutons les données corrompues (maintenant annotées sémantiquement) aux données originales et l'ensemble est utilisé pour entrainer le nouveau modèle SLU (dans la langue source) qui alors intégrera le bruit présent dans les données traduites.

3.2 Post-édition statistique

Plusieurs travaux récents en traduction automatique comme (Simard, Goutte, Isabelle, 2007) (Diaz de Ilarraza, Labaka, Sarasola, 2008) ont utilisé une approche basée sur un système de traduction pour post-éditer les sorties d'un autre système de traduction. Un tel système, malgré une démarche qui peut paraitre contre-intuitive, a été proposé pour améliorer la qualité des données traduites avant leur envoi à des post-éditeurs humains. Pour entrainer un tel post-éditeur, (Simard, Goutte, Isabelle, 2007) (Diaz de Ilarraza, Labaka, Sarasola, 2008) utilisent les sorties d'un système SMT avec comme données parallèles leur post-édition manuelle.

Dans notre cas, dans la mesure où la sortie du système SMT sera utilisée comme entrée du système SLU entrainé sur les données du langage source, nous proposons de post-éditer cette sortie afin de diminuer le bruit du à la traduction des entrées utilisateurs.

Pour apprendre un SPE, notre choix a été de traduire automatiquement l'ensemble de données disponibles pour la langue cible, puis d'utiliser les sorties traduites avec les parties correspondantes transcrites manuellement, comme corpus parallèle. Nous pensons que le module de post-édition permettra ainsi de réordonner quelques mots ou de retrouver des mots manquants dans un certain nombre de phrases.

4 Description du corpus et d'outils

Toutes les expériences décrites dans le papier ont été réalisées sur le corpus français MEDIA. Ce travail a été motive par la disponibilité d'une traduction manuelle en italien d'une sous-partie de ce corpus.

4.1 Le corpus MEDIA

Comme décrit dans (Bonneau -Maynard, Rosset, Ayache, Kuhn, Mostefa, 2005), ce corpus couvre un domaine lié aux réservations de chambres d'hôtels et aux informations touristiques. Le corpus est constitué de 1257 dialogues enregistrés par 250 locuteurs, collectés en situation de Wizard-of-Oz (un humain simule le système de dialogue).

Les dialogues sont regroupés en 3 parties : un ensemble d'apprentissage (environ 13k phrases), un ensemble de développement (1,3k phrases) et un ensemble d'évaluation (3k phrases). Dans nos expériences, nous ne prenons en compte que les phrases utilisateurs.

Le corpus est étiqueté avec 99 concepts différents. Ces étiquettes peuvent être simples comme les dates ou les noms de ville ou peuvent être plus complexes comme les coréférences. A titre d'illustration, voici une phrase de MEDIA :

Je voudrais une chambre double à Marseille

L'annotation sémantique de cette phrase aura la forme :

Je voudrais [null], une [nombre-chambre], chambre double [chambre-type], à Paris [localization-ville].

Cette annotation sémantique découpe chaque phrase en plusieurs segments. Chaque segment est non seulement annoté avec le nom du concept mais aussi par une valeur, une modalité et un specifieur. Les expériences présentées dans le papier prennent en compte uniquement le nom du concept et la modalité du segment.

Un sous-ensemble de l'apprentissage (environ 5.6k phrases), de même que les ensembles de test et de développement, ont été manuellement traduit en italien dans le contexte du projet européen LUNA (Servan, Camelin, Raymond, Bechet, De Mori, 2010).

4.2 Les SMTs apprit

Dans cette étude, nous utilisons deux systèmes de traduction automatique pour obtenir les traductions du français vers l'italien et de l'italien vers le français. Pour réaliser ces traductions, la boite à outils Moses (Koehn, Hoang, Birch, Callisonburch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, Herbst, 2007) est utilisée. Moses implémente l'état-de-l'art des systèmes de traduction par segments utilisant des modèles log-linéaires.

Nous utilisons la partie manuellement traduite en italien de l'ensemble d'apprentissage du corpus MEDIA comme corpus parallèle pour Moses dans les deux directions pour entrainer les modèles. Chacune des parties séparément permet l'apprentissage d'un modèle de langage. Aussi l'ensemble de développement avec sa traduction est utilisé comme corpus parallèle pour ajuster les poids du modèle log-linéaire des systèmes SMT. Finalement, nous obtenons un système de français vers l'italien avec un score BLEU de 43,62 et de l'italien vers

le français avec un score de 47,18. Ces scores sont mesurés sur l'ensemble de test de MEDIA manuellement traduit. Dans la mesure où une seule référence par phrase est utilisée pour évaluer le score BLEU, de même que l'ensemble d'apprentissage est réduit (5,6k), ces performances peuvent être considérées comme très acceptables.

Nous avons aussi re-traduit automatiquement en français la version manuelle en italien du corpus, de sorte à utiliser cette traduction en parallèle de la partie originale pour entrainer le SPE et fournir les données bruitées pour le SCTD. Du point de vue de la performance de traduction exclusivement, l'utilisation de la post-édition automatique améliore le score BLEU du système de 47,18 à 49,25.

4.3 Compléter la traduction de MEDIA

Le système de traduction français-italien est utilisé pour obtenir une traduction automatique de la partie restante (non traduite manuellement) du corpus d'apprentissage, ainsi une traduction intégrale (manuelle + automatique) est disponible. Le système italien-français est utilisé pour traduire le test italien en français, qui sera utilisé pour l'approche *TestOnSource*. Le tableau 1 donne un aperçu des ensembles disponibles pour les expériences.

MEDIA data	Train	Dev	Test
French MEDIA	13K	1,3K	3,5K
Italian manual	5,6K	1,3K	3,5K
Italian automatic	7,4K	-	-

Tableau 1: aperçu du corpus MEDIA et de sa traduction vers l'italien (# phrases).

5 Expériences et résultats

Afin d'évaluer les performances des approches proposés, la traduction manuelle en italien des données de test est utilisée. Le CER est le critère d'évaluation retenu pour cette étude. Le CER est l'équivalent du taux d'erreur en mots (WER), et peut être défini comme le ratio de la somme des concepts omis, insérés et subtilisés sur le nombre de concepts dans la référence. Premièrement nous évaluons et comparons SLU/CRF et SLU/PB-SMT, nous évaluons de même notre proposition de robustesse, puis les systèmes sont combinés. Pour finir, nous validons nos approches en utilisant les traductions obtenues par un système de traduction en ligne (ie sans utiliser de traduction manuelle).

5.1 Les stratégies de portabilité SLU/CRF

La totalité de l'ensemble d'apprentissage de MEDIA est utilisé pour apprendre un étiqueteur français de base utilisant des uni- et bi-grammes. Cette base atteint de bonnes performances (12,9% CER) et peut être considérée comme une référence pour les méthodes proposées. Pour évaluer les performances de l'approche *TestOnSource* la traduction automatique en français du test italien (comme décrit en 4) est fournie à l'étiqueteur de base.

La méthode *TrainOnTarget* décrite dans 2.1.2 a été appliquée. Nous utilisons le système français-italien (décrit en 4) pour traduire le corpus d'entrainement intégrant des balises XML correspondant aux segments conceptuels afin d'évaluer la méthode *TaggedTranslation*, et la totalité des traductions en italien de MEDIA (manuelle et automatique) avec la version française comme corpus parallèle pour obtenir l'alignement mot-a-mot. L'information d'alignement telle que défini en 2.1.2 est utilisée, pour évaluer la performance de la méthode *Alignement*. Toutes les expériences utilisent l'outil CRF++ (*https://crfpp.sourceforge.net)*. L'ensemble des résultats est regroupés dans le tableau 2.

Il apparait clairement à la lecture des résultants que la méthode *Alignement* est meilleure que *TaggedTranslation*. Ceci peut être expliqué par le fait que *Alignement* est seulement influencé par les erreurs d'alignement, tandis que *TaggedTranslation* est influencé par les erreurs de traduction automatique qui sont plus importantes. On note aussi que la méthode *TestOnSource* est plus performante que les méthodes *TrainOnTarget.*les performances de toutes les méthodes sont considérées come bonne en comparaison avec la référence française.

Model	Sub	Del	Ins	CER
FR	3,1	8,1	1,8	12,9
SLU/CRF(TestOnSource)	5,2	12,1	2,6	19,9
SLU/CRF(TaggedTranslation)	3,7	16,9	2,1	22,7
SLU/CRF(Alignment)	3,1	15,0	2,3	20,5

Tableau 2 : Evaluation (CER %) de différentes stratégies de portabilité du SLU utilisant les méthodes SLU/CRF

5.2 SLU/PB-SMT

Nos premières tentatives pour construire le modèle PB-SMT pour le SLU italien ont clairement montré des performances inférieures aux CRF (CER=28,1% après réglage MERT pour le PB-SMT comparé aux ~20% pour les CRF). Les améliorations progressives du modèle proposées en Section 2.2 sont évaluées dans le tableau 3. L'utilisation de la contrainte de monotonie durant l'alignement en mot permet une réduction de 0,6% absolu. Convertir les données selon le formalisme BIO avant la phase d'apprentissage réduit le CER de façon significative de 2,8%. Enfin optimiser le CER à la place du BLEU réduit le CER de 0,3% supplémentaire. L'ajout d'une liste de villes à l'ensemble d'apprentissage avant réapprentissage du modèle PB-SMT permet une réduction finale de 0,5%.

Les résultats montrent qu'en dépit de réglages fins de l'approche SMT, les approches à base de CRF obtiennent toujours les meilleures performances. De plus, dans une expérience parallèle, un modèle PB-SMT a été construit pour le SLU Français afin de le tester dans l'approche *TestOnSource*. Mais les performances de l'approche sont décevantes et bien en-deça de celles des autres méthodes. Elle a donc été écartée pour le reste de l'étude.

A partir d'une analyse rapide du type d'erreurs de chaque modèle, nous pouvons observer que les méthodes utilisant des CRF ont un haut niveau de suppressions comparativement aux autres types d'erreurs, tandis que la méthode PB-SMT présente un meilleur compromis entre les erreurs de suppression et d'insertion, et ce bien qu'elle abouti à un CER plus élevé.

SLU/PB-SMT	Sub	Del	Ins	CER
Initial	6,5	4,0	18,6	29,1
+ MERT (BLEU)	6,3	9,3	12,5	28,1
+ Monotone align	7,4	8,4	11,8	27,5
+ BIO format	6,5	10,6	7,7	24,7
+ MERT (CER)	6,4	10,9	7,2	24,4
+ City list	7,2	10,5	6,1	23,9

Tableau 3 : améliorations itératives de la méthode SLU/PB-SMT sur le test italien de MEDIA (CER%)

5.3 SLU/CRF *TestOnSource* robuste

Nous avons tenté d'améliorer les performances de la méthode *TestOnSource* SLU/CRF en renforçant sa robustesse aux erreurs de traduction. Premièrement nous traduisons automatiquement la partie manuelle en italien. Ensuite nous apprenons un nouvel étiqueteur CRF simultanément sur les données d'apprentissage en français et traduites (approche +SCTD, décrite en 3.1). La méthode de la section 3.2 (SPE) a aussi été évaluée, dans laquelle le test traduit post-édité a été transmis aux CRF de base (+SPE) ou aux CRF appris sur les données corrompues (+SCTD+SPE).

L'évaluation des performances de ces approches sont rapportées dans le tableau 4. Les deux méthodes, d'apprentissage sur données bruitées et SPE, améliorent les performances de l'étiqueteur sémantique. Leur mise en série donne les meilleures performances.

SLU/CRF	Sub	Del	Ins	CER
TestOnSource	5,2	12,1	2,6	19,9
+SCTD	5,9	11,4	2,3	19,6
+SPE	6,5	10,6	2,5	19,7
+SCTD +SPE	6,4	9,9	2,9	19,3

Tableau 4 : Evaluation (CER %) des approches proposées pour la robustesse des systèmes au bruit de traduction

5.4 Combinaison de systèmes

Nous proposons de combiner les trois approches principales (*TestOnTarget* et *TrainOnTarget* SLU/CRF, et SLU/PB-SMT) afin de bénéficier de leurs caractéristiques respectives pour améliorer la performance globale. La combinaison (dénotée BASIC dans le tableau 5) est simple : un réseau de confusion est construit à partir des trois hypothèses et la séquence de concept correspondant à la plus grande probabilité a posteriori est calculée. La performance est améliorée de façon significative (-1,3% CER) ce qui confirme la complémentarité des méthodes.

Finalement nous combinons toutes les méthodes proposées dans ce papier (SLU/CRF *TrainOnTarge,* SLU/CRF *TestOnSource*, +SCTD, +SPE, +SCTD+SPE, SLU/PB-SMT). Ce qui permet d'atteindre les meilleures performances rapportées sur ce test (18,2%). Afin de mesurer l'influence de la méthode SLU/PB-SMT sur les performances de la combinaison, nous avons aussi évalué les performances de la combinaison diminuée de SLU/PB-SMT. Cette expérience a montré qu'en dépit de ses mauvais résultats individuels, la méthode PB-SMT a une influence importante sur la combinaison.

Model	Sub	Del	Ins	CER
BASIC	6,2	9,7	2,7	18,6
ALL	5,4	10,5	2,3	18,2
ALL – SLU/PB-SMT	6,6	10,2	2,7	19,4

Tableau 5 : combinaison de systèmes avec et sans l'approche PB-SMT

5.5 Validation des stratégies de portabilité SLU/CRF en utilisant des traductions en ligne uniquement

Les expériences présentées dans cet article ne sont pas totalement non-supervisées, dans tous les cas nous avons utilisé des données traduites manuellement pour obtenir le système de traduction pour *TestOnSource* ou pour compléter la traduction de l'apprentissage et obtenir les informations d'alignement pour la méthode *TrainOnTraget*.

Le coût associé à cette traduction manuelle est relativement bas comparé à celui de collecter et d'annoter un nouveau corpus d'apprentissage, mais il reste non négligeable. Nous voulons vérifier qu'un tel coût est justifié par comparaison à une approche totalement non-supervisée et donc (potentiellement) « gratuite ».

Afin de répondre à cette interrogation nous proposons de reproduire les expériences en utilisant un système de traduction gratuit en ligne à la place de notre système de traduction appris sur la tâche.

Pour évaluer la méthode *TestOnSource* nous traduisons le test MEDIA en italien à l'aide d'une solution gratuite en ligne puis nous utilisons cette traduction comme entrée de l'étiqueteur CRF de base. Pour la méthode *TrainOnTarget* deux approches ont été testées. Pour permettre la comparaison avec la méthode

TaggedTranslation, nous proposons de traduire les données d'entrainement de MEDIA, segment par segment, au moyen du traducteur en ligne, puis ces traductions sont associées aux étiquettes sémantiques initiales. Dans une seconde version, les données sont traduites intégralement puis utilisées comme corpus parallèle pour l'approche *Alignement*.

Pour choisir un traducteur automatique dans le cadre de nos expériences nous avons comparé les performances de deux traducteurs réputés. Le test MEDIA et sa traduction manuelle ont été utilisé comme couple test/référence dans chacune des directions de traduction. Pour l'Italien vers le français un traducteur de score BLEU 42,58 a été sélectionné (à comparer à 47,18 obtenu par le système SMT appris sur les traductions manuelles), et pour le français vers l'italien un traducteur de score 39,75 a été retenu (à comparer avec 43,62). Les résultats de cette expérience sont reportés dans le tableau 6.

De manière attendue les performances des systèmes obtenus par cette approche non-supervisée sont inferieures à celle des systèmes semi-supervisés. Toutefois malgré la dégradation du CER pour toutes les approches son niveau absolu reste tout à fait acceptable considérant les besoins de la tache et la réduction substantielle du coût de développement.

La méthode *Alignement* est la plus perturbée et devient presque équivalente à *TaggedTranslation* en version non-supervisée. Le CER augmente de 22,7% à 26,6% (+3,9% absolu) pour *TaggedTranslation* et de 20,5% à 26,5% (+6%) pour *Alignement*. *TestOnSource* perd 3,2% mais reste la plus performante. Ces résultats nous engagent à tester de nouvelles langues pour lesquelles nous ne disposons pas de traductions manuelles.

Model	**Sub**	**Del**	**Ins**	**CER**
Semi supervisé				
SLU/CRF(TestOnSource)	5,2	12,1	2,6	19,9
SLU/CRF(TaggedTranslation)	3,7	16,9	2,1	22,7
SLU/CRF(Alignment)	3,1	15,0	2,3	20,5
Non supervisé				
SLU/CRF(TestOnSource)	6,1	14,5	2,5	23,1
SLU/CRF(TaggedTranslation)	5,5	15,4	5,7	26,6
SLU/CRF(Alignment)	6,3	14,8	5,4	26,5

Tableau 6 : Evaluation (CER %) des stratégies de portabilité SLU/CRF en utilisant des traductions en ligne

6 Conclusion

Dans cet article on a proposé et comparé plusieurs approches pour la portabilité d'un SLU a travers les langues. Les CRFs et le PB-SMT on été utilisé pour cette tache et les résultats montrent que l'utilisation d'un étiqueteur a base de CRF avec des données de test traduites donne la meilleur performance. On a aussi montré l'intérêt de l'utilisation de deux méthodes différentes pour accroitre la robustesse du SLU aux erreurs de traduction. Enfin on a montré que la combinaison de toutes les méthodes proposées augmente la performance du système.

Remerciements

Ce travail est supporté par le projet ANR PORT-MEDIA (ANR 08 CORD 026 01). Plus d'information disponible sur le site du projet : www.port-media.org

Références

Suenderman K., Liscombe J. (2009). From rule-based to statistical grammars: Continuous improvement of large-scale spoken dialog system. Actes de *ICASSP*.

Hahn S., Lehnen S., Raymond C., Ney H. (2008). A comparison of various methods for concept tagging for spoken language understanding. Actes de *LREC*.

Raymond C., Riccardi G. (2007). Generative and discriminative algorithms for spoken language understanding. Actes de *Interspeech*.

Wang Y., Acero A. (2006). Discriminative models for spoken language understanding. Actes de *ICSLP*.

Schwartz R., Miller S., Stallard D., Makhoul J. (1996). Language understanding using hidden understanding models. Actes de *ICSLP*.

Suenderman K., Liscombe J. (2009). Localization of speech recognition in spoken dialog systems: How machine translation can make our lives. Actes de *Interspeech*.

Servan C., Camelin N., Raymond C., Bechet F., De Mori R. (2010). On the use of machine translation for spoken language understanding portability. Actes de *ICASSP*.

Lefèvre F., Mairesse F., Young S. (2010). Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation. Actes de *Interspeech*.

Jabaian B., Besacier L., Lefèvre F. (2010). Investigating multiple approaches for SLU portability to a new language. Actes de *Interspeech*.

Bonneau-Maynard H., Rosset S., Ayache C., Kuhn A., Mostefa D. (2005). Semantic annotation of the French media dialog corpus. Actes de *Eurospeech*.

Lafferty J., McCallum A., Pereira F. (2001). Conditional random fields: Probabilistic models for segmenting and labelling sequence data. Actes de *ICML*.

Koehn P., Och F., Marcu D. (2003). Statistical phrase_based translation. Actes de *HLT/NAACL*.

Koehn P., Hoang H., Birch A., Callisonburch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., Dyer C., Bojar O., Constantin A., Herbst E. (2007). Moses: Open source toolkit for statistical machine translation. Actes de *ACL*.

Och F., Ney H. (2000). Improved Statistical Alignment Models. Actes de *ACL*.

Liang P., Taskar B., Klein D. (2006). Alignment by agreement. Actes de *HLT*.

Stolcke A. (2002). SRILM an extensible language modeling toolkit. Actes de *SLP*.

Simard M., Goutte C., Isabelle P. (2007). Statistical phrase-based post-editing. Actes de *NAACL*.

Diaz de Ilarraza A., Labaka G., Sarasola K. (2008). Statistical post-editing: A valuable method in domain adaptation of RBMT systems for less-resourced languages. Actes de *MATMT*.

Qui êtes-vous ? Catégoriser les questions pour déterminer le rôle des locuteurs dans des conversations orales *

Thierry Bazillon[1], Benjamin Maza[2], Mickael Rouvier[2], Frederic Bechet[1], Alexis Nasr[1]
(1) Aix Marseille Université, LIF-CNRS, Marseille, France
(2) Université d'Avignon, LIA-CERI, Avignon, France

Résumé. La fouille de données orales est un domaine de recherche visant à caractériser un flux audio contenant de la parole d'un ou plusieurs locuteurs, à l'aide de descripteurs liés à la forme et au contenu du signal. Outre la transcription automatique en mots des paroles prononcées, des informations sur le type de flux audio traité ainsi que sur le rôle et l'identité des locuteurs sont également cruciales pour permettre des requêtes complexes telles que : « chercher des débats sur le thème X », « trouver toutes les interviews de Y », etc. Dans ce cadre, et en traitant des conversations enregistrées lors d'émissions de radio ou de télévision, nous étudions la manière dont les locuteurs expriment des questions dans les conversations, en partant de l'intuition initiale que la forme des questions posées est une signature du rôle du locuteur dans la conversation (présentateur, invité, auditeur, etc.). En proposant une classification du type des questions et en utilisant ces informations en complément des descripteurs généralement utilisés dans la littérature pour classer les locuteurs par rôle, nous espérons améliorer l'étape de classification, et valider par la même occasion notre intuition initiale.

Abstract. Speech Data Mining is an area of research dedicated to characterize audio streams containing speech of one or more speakers, using descriptors related to the form and the content of the speech signal. Besides the automatic word transcription process, information about the type of audio stream and the role and identity of speakers is also crucial to allow complex queries such as : " seek debates on X ,"" find all the interviews of Y", etc. In this framework we present a study done on broadcast conversations on how speakers express questions in conversations, starting with the initial intuition that the form of the questions uttered is a signature of the role of the speakers in the conversation (anchor, guest, expert, etc.). By classifying these questions thanks to a set of labels and using this information in addition to the commonly used descriptors to classify users' role in broadcast conversations, we want to improve the role classification accuracy and validate our initial intuition.

Mots-clés : Fouille de données orales, Traitement Automatique de la Parole, Annotation de corpus oraux, Classification en rôles de locuteurs.

Keywords: Speech data mining, Automatic Speech Processing, Speech Corpus Annotation, Speaker role classification.

*. Ce travail a été effectué dans le cadre du projet ANR DECODA (2009 CORD 005 01) http://decoda.univ-avignon.fr

1 Introduction

La fouille de données orales est un domaine de recherche visant à caractériser un flux audio contenant de la parole d'un ou plusieurs locuteurs à l'aide de descripteurs liés à la forme et au contenu du signal. Parmi ces descripteurs, le plus important est bien évidemment la transcription automatique en mots des paroles prononcées. Dans le cas de la parole multi-locuteurs, telle qu'on peut en trouver dans les émissions de radio ou de télévision, ces descripteurs peuvent aussi porter sur l'identité ou le rôle du locuteur, mais aussi sur le type de conversations enregistrées : débat, interview, journal d'information, etc. Ces informations peuvent permettre de répondre à des requêtes complexes telles que : "chercher des débats sur le thème X", "trouver toutes les interviews de Y", mais aussi d'aider le processus de transcription automatique en choisissant des modèles adaptés au type d'émissions considéré.

L'étude présentée dans cet article a été réalisée sur le corpus EPAC contenant la transcription et l'annotation d'une centaine d'heures de parole essentiellement spontanée [1]. Il est notamment constitué d'émissions ou de débats radiophoniques tels que *Le Téléphone sonne*, *Quartiers d'Été*, *Sous les étoiles exactement*, *Culture vive* ou *Les Matins de France Culture*. Toutes ces émissions ont été segmentées en locuteurs et transcrites manuellement. En complément de ces annotations, d'autres annotations sur les rôles des locuteurs et les formes interrogatives utilisées ont été ajoutées sur une partie du corpus. Nous avons choisi de nous intéresser spécifiquement au domaine de l'interrogation car il nous semble que la structure même des questions, à l'intérieur d'un discours, est représentative d'un certain type d'oral. Des travaux comme ceux de (Garcia-Fernandez & Lailler, 2008), tendant à définir une "morphosyntaxe de l'interrogation", vont déjà en ce sens. Pour notre part, nous essaierons de voir si les questions peuvent être associées à des classes de locuteurs spécifiques, dans un contexte d'oral radiophonique précis. Un arbre de catégorisation des questions a donc été créé *a priori*, et une vingtaine d'heures de données issues de l'émission *Le Téléphone sonne* ont été étiquetées selon cet arbre.

Nous présentons dans la section 2 une étude descriptive des formes interrogatives et leur répartition en différents types ou catégories. Nous tentons en 3 d'établir un lien entre les types de questions et les rôles des locuteurs qui les ont posées. L'apport de la prise en compte du type des questions posées pour la segmentation automatique en rôles de locuteurs est présentée dans la section 4. Enfin la détection et la classification automatique des questions et leur intégration dans le classifieur en rôle de locuteurs est présentée dans la section 5.

2 Un corpus étiqueté en questions

À l'intérieur du corpus EPAC, l'émission *Le Téléphone sonne* représente près de 20 heures de données, réparties en 32 émissions. Dans chacune d'entre elles, chaque question a été annotée manuellement grâce à un système de balises spécifiques. Les questions ont été catégorisées selon différentes catégories et sous-catégories linguistiques. Tout d'abord, une distinction de premier niveau a été établie entre les questions directes (*comment vas-tu ?*) et les questions indirectes (*je voudrais savoir comment tu vas ?*). Ensuite, à l'intérieur de ces deux ensembles, trois sous-catégories ont été isolées : les interrogations totales, les interrogations partielles et les interrogations alternatives (*vous voulez une réponse précise ou pas ?*). Enfin, un troisième et dernier niveau d'analyse concernant les marqueurs interrogatifs a été pris en compte. Ces marqueurs peuvent être :

– des pronoms interrogatifs (*qui*, *que*)
– des adverbes interrogatifs (*quand*, *comment*, *pourquoi*)
– des déterminants interrogatifs (*quel*, *quelle*)
– des structures complexes (*qu'est-ce que*, *qu'est-ce qui*, *à qui*, *depuis quand*...)
– la forme *est-ce que*
– l'adverbe *si* (je voudrais savoir si vos intervenants sont d'accord ?)
– l'inversion du sujet
– l'intonation seule (*tu viens ?*)

La nomenclature ci-dessus est certes un peu différente de celle proposée notamment par (Stolcke *et al.*, 2000), mais les principales distinctions y sont préservées. Ainsi, notre critère « intonation seule » correspond aux questions « déclaratives » de Stolcke, et ce qu'il nomme les *wh- questions* est ici transposé en pronoms, adverbes

1. nous renvoyons le lecteur à (Estève *et al.*, 2010) pour une description détaillée des tâches de transcription et d'annotation du corpus EPAC

et déterminants interrogatifs. Enfin, les questions avec la locution *est-ce que* ou l'inversion sujet-verbe (soit les interrogations dites « totales ») correspondent aux « yes-no-question » de la terminologie de Stolcke.

Type de questions			Nombre d'occurrences	Fréquence (%)
directe	totale	est-ce que	214	13,76
		intonation	178	11,45
		inversion	154	9,9
	partielle	intonation	404	25,98
		adverbe	198	12,73
		complexe	117	7,52
		pronom	88	5,66
		déterminant	76	4,89
	alternative	inversion	10	0,64
		intonation	5	0,32
		est-ce que	1	0,06
indirecte	totale	si	46	2,96
	partielle	adverbe	26	1,67
		complexe	17	1,09
		déterminant	17	1,09
		groupe nominal	3	0,19
		pronom	1	0,06

TABLE 1 – Répartition des questions par type

Sur les 48 catégories possibles, 17 sont attestées dans le corpus du *Téléphone sonne*. La table 1 présente ces 17 catégories avec leur fréquence absolue et relative. Cette table met en relief plusieurs éléments : en premier lieu, les questions directes sont beaucoup plus nombreuses que les questions indirectes. Cet écart s'explique par le fait que *Le Téléphone sonne* est avant tout un débat, donc avec des propos essentiellement spontanés. En conséquence, les locuteurs posent la plupart de leurs questions de la façon la plus directe qui soit, c'est-à-dire avec un nombre réduit de mots. En conséquence, ils ont très peu recours aux structures telles que *je voudrais savoir si...* ou *je me demandais si....* À la place de *je voudrais savoir si vos invités sont d'accord avec ça*, on trouvera ainsi beaucoup plus souvent la forme oralisée *vos invités sont d'accord avec ça ?*.

Ce dernier exemple nous amène à une autre observation, qui peut être faite à la lecture de la table 1 : les questions avec l'intonation pour seul marqueur interrogatif sont les plus représentées (plus de 37% en cumulant interrogations totales et partielles). L'explication est directement liée à ce que nous écrivions plus haut au sujet du type de parole utilisé dans *Le Téléphone sonne*. En effet, la parole spontanée est souvent synonyme de structures interrogatives réduites, notamment dans les débats. Pour être plus en phase avec des situations discursives souvent mouvementées (parole simultanée, locuteurs qui se coupent la parole, temps d'antenne réduit), les locuteurs se doivent d'être les plus concis possibles. En conséquence, en plus d'utiliser essentiellement des structures interrogatives directes, ils font aussi abstraction des marqueurs interrogatifs traditionnels (*Est-ce que vous y croyez ?* deviendra *vous y croyez ?*, par exemple). Il en va de même pour les constructions syntaxiques avec inversion du sujet (*voulez-vous poser une autre question ?*), qui sont assez peu nombreuses dans notre relevé. Bien qu'elles ne rallongent pas à proprement parler la longueur d'une question (aucun mot supplémentaire n'est nécessaire pour passer de *vous avez une autre question ?* à *avez-vous une autre question ?*), elles ne correspondent pas au modèle canonique français sujet + verbe + complément.

3 Un corpus étiqueté en rôles

Le rôle d'un locuteur correspond à son statut et à sa fonction dans une émission donnée : présentateur principal, invité, consultant, journaliste hors studio, etc. Identifier le rôle de chaque locuteur est une étape indispensable à la compréhension d'une émission. Le format de chaque émission définit le nombre et les rôles de chaque intervenant. Pour cette étude, nous avons privilégié une segmentation en rôles relativement générique. Par exemple, les éventuels changements de rôle d'un locuteur à l'intérieur d'une même émission n'ont pas été pris en compte. Cette distinction, pertinente dans le cadre d'un découpage en actes de dialogue, ne nous a pas semblé nécessaire ici. Ainsi, dans l'émission *Le Téléphone sonne*, nous avons donc identifié 4 rôles de locuteurs :

1. le présentateur : c'est l'animateur de l'émission, il a comme rôle de distribuer la parole aux différents intervenants en animant le débat ;
2. les experts : ce sont les invités, sur le plateau ou par téléphone, ils ont pour rôle à la fois de répondre aux questions des auditeurs mais aussi de débattre entre eux sur le sujet du jour ;
3. les auditeurs : toujours au téléphone, ils sont sélectionnés avant l'émission et sont appelés pour qu'ils puissent poser leur questions, sans pour autant participer au débat ;
4. le rapporteur : il a pour rôle de lire les questions écrites des auditeurs ; ce rôle est généralement assumé par le présentateur.

Nous présentons dans les lignes suivantes une analyse du type des questions exprimées par les différents rôles de locuteurs. C'est cette étude qui nous a conduit à utiliser les questions comme indices ou marqueurs du rôle des locuteurs, comme nous le verrons dans la partie expérimentale de cet article.

3.1 Les questions comme marqueurs des rôles des locuteurs

La table 2 présente les premiers résultats de l'annotation des rôles des locuteurs, couplée avec celle des questions. Il apparaît ainsi que c'est le présentateur qui pose plus de la moitié des questions lors d'une émission comme *Le Téléphone sonne*. Cette prédominance tient essentiellement à sa fonction de *médiateur* que nous avons évoquée précédemment, et que nous expliciterons plus précisément avec d'autres chiffres dans le paragraphe suivant.

Rôle	Nombre de questions	Fréquence (%)
Présentateur	791	50,87
Expert	323	20,77
Auditeur	307	19,74
Rapporteur	134	8,62
TOTAL	1555	100

TABLE 2 – Répartition du nombre de questions en fonction du rôle des locuteurs

Il est également intéressant de constater que les auditeurs, pourtant supposés être au cœur du programme, posent en moyenne moins de questions que les experts présents en studio lors de chaque émission. Toutefois, ces chiffres doivent être nuancés par quelques précisions. Tout d'abord, les questions de la catégorie *rapporteur* peuvent être associées à celles des auditeurs, dans la mesure où ce sont eux qui les rédigent puis les envoient au standard de l'émission. Elles sont certes lues par le présentateur, mais il n'en est jamais l'auteur. Ainsi, on peut considérer que près de 30% des questions de notre corpus sont, directement ou non, posées par des auditeurs.

Ensuite, beaucoup de questions de la catégorie *expert* sont des interrogations n'attendant pas véritablement de réponses. Sans être rhétoriques (au sens linguistique du terme), elles permettent plutôt au locuteur d'étayer sa réponse en s'appuyant plus ou moins explicitement sur la question qui lui a été posée :

Quand tout cela va cesser ? Ça madame je n'en sais rien, mais ce qui sûr...
Alors est-ce qu'il doit démissionner ? Oui, bien sûr, puisque la situation actuelle du pays...

Enfin, et c'est souvent le cas dans les émissions dites de débat, le temps de parole alloué à chaque participant est loin d'être équitablement réparti. Les auditeurs, n'étant pas présents physiquement sur le plateau, sont les premiers à être coupés, interrompus ou parfois même exclus du débat faute de temps. Ils ne posent d'ailleurs qu'une seule voire deux questions, jamais plus. À l'inverse, les experts bénéficient d'une liberté d'expression très large. Celle-ci leur permet certes de répondre aux auditeurs, mais aussi d'étayer le débat en se posant mutuellement de nombreuses questions :

Monsieur XX, vous pensez vraiment que ça va changer les choses ?
Comment pouvez-vous en être sûrs, madame YY ?

Afin d'envisager une analyse plus fine, la table 3 indique, pour chaque type de questions recensées dans notre corpus, leur répartition selon le rôle des locuteurs. Ces résultats sont notamment l'occasion de revenir sur la prépondérance du rôle du présentateur, que nous évoquions précédemment. Comme on le voit, celle-ci est majoritairement due aux questions ayant l'intonation pour seul marqueur interrogatif. En effet, en tant que responsable

des débats qu'il instaure, le présentateur du *Téléphone sonne* est celui qui distribue la parole à ses invités et à ses auditeurs. Pour ce faire, il recourt abondamment à des questions elliptiques tantôt partielles (*sur la guerre en Irak, votre sentiment ?*), tantôt totales (*vous êtes d'accord avec ça, monsieur Gorce ?*). Ces deux catégories réunies représentent ainsi près de 70% des questions assimilées au rôle de présentateur. Les 30% restants concernent principalement des questions qui sont posées au début de chaque émission, lors d'un monologue servant à introduire le sujet du jour. Manifestement écrites, elles sont toujours très structurées sur le plan grammatical, que ce soit avec la locution *est-ce que*, l'inversion du sujet ou des pronoms ou adverbes interrogatifs.

Type question			Présentateurs (%)	Experts (%)	Auditeurs (%)	Rapporteurs (%)
directe	totale	est-ce que (214)	28,5	34,58	33,64	3,27
		intonation (178)	80,9	16,3	2,81	0
		inversion (154)	33,12	7,14	21,43	38,31
	partielle	intonation (404)	98,51	1,24	0,25	0
		adverbe (198)	18,69	43,94	21,72	15,66
		complexe (117)	21,37	48,72	19,66	10,26
		pronom (88)	34,1	35,23	19,32	11,36
		déterminant (76)	44,74	19,74	19,74	15,79
	alternative	inversion (10)	20	70	0	10
		intonation (5)	80	20	0	0
		est-ce que (1)	100	0	0	0
	global (1445)		54,46	21,94	14,46	9,13
indirecte	totale	si (46)	4,35	2,18	93,48	0
	partielle	adverbe (26)	0	3,85	88,46	7,69
		complexe (17)	5,88	11,76	82,35	0
		déterminant (17)	0	11,76	88,24	0
		groupe nominal (3)	0	0	100	0
		pronom (1)	100	0	0	0
	global (110)		3,64	5,45	89,1	1,82

TABLE 3 – Répartition du type de questions en fonction du rôle des locuteurs

Les experts, outre le fait d'occuper la deuxième place au niveau du nombre global de questions posées (table 2), utilisent également de nombreuses structures interrogatives différentes. S'il n'y a pas une catégorie aussi dominante que chez les présentateurs, on notera toutefois qu'ils utilisent beaucoup les questions directes à base d'adverbes, de pronoms, de *est-ce que* et de structures complexes (pronoms ou déterminants : *qu'est-ce que*, *qu'est-ce qui*, *duquel*, *lequel*, etc.). Cela témoigne d'un certain soin quant à la formulation et l'expression, puisque les questions basées sur l'intonation sont ici beaucoup moins employées. Cela est d'autant plus remarquable que dans un cadre énonciatif spontané, les structures les plus simples sont souvent les plus utilisées (*vous pensez vraiment que...* en lieu et place de *pensez-vous vraiment que...* ou *est-ce que vous pensez vraiment que...*). Mais il ne faut pas oublier que nous sommes ici dans le cadre d'une émission radiophonique écoutée, jugée et aussi soumise à des directives éditoriales précises. Les personnes qui y participent n'apportent donc sans doute pas le même soin à leurs propos dans la vie quotidienne, ni même peut-être lorsqu'elles s'expriment dans d'autres medias.

C'est sans doute ce qui explique que le constat soit sensiblement identique du côté des auditeurs, où l'on constate que les questions sans marqueur interrogatif grammatical sont moins fréquentes encore. À l'inverse, la structure *est-ce que* est fortement utilisée par les auditeurs, de même que les constructions indirectes (quel que soit leur marqueur interrogatif). Les incises telles que *j'aimerais savoir*, *j'aurais voulu savoir* ou *je voulais savoir* sont en effet presque exclusivement employées par les auditeurs, sans doute parce qu'elles permettent une sorte de transition entre l'invitation à la prise de parole du présentateur (*Posez votre question*, *nous vous écoutons*) et la question à proprement parler, qu'il serait assez brutal de formuler au style direct. Ainsi, en réponse aux deux exemples ci-dessus, on trouvera beaucoup plus souvent des formes comme *je voulais savoir si la droite avait une chance* plutôt que *est-ce que la droite a une chance ?*. De leur côté, les questions commençant par la locution *est-ce que* sont souvent précédées d'un témoignage de l'auditeur, plus ou moins long, mais qui lui permet de contextualiser sa demande (*j'ai lu dans un journal que [...] Est-ce que vos experts sont d'accord avec ça ?*). De façon plus générale, et pour en revenir à notre idée de départ, les questions posées par les auditeurs sont elles aussi particulièrement soignées sur le plan de la syntaxe, d'une part parce qu'elles sont présélectionnées par le standard du *Téléphone sonne*, et d'autre part parce que beaucoup de personnes les écrivent avant de les lire à l'antenne, craignant de les oublier ou de mal les formuler sinon.

4 Segmentation automatique en rôles de locuteurs

La segmentation automatique en rôles de locuteurs est une tâche relativement nouvelle qui vient en complément des tâches de segmentation en locuteurs effectuées en préalable de tout processus de transcription automatique de parole. En effet, la parole étant un flux, il convient de la segmenter en *tours de parole* correspondant à chaque locuteur de la conversation à transcrire, puis en segments (correspondant généralement à des groupes de souffle) sur lesquels les processus de transcription automatique sont appliqués. La segmentation en rôle correspond à une tâche d'étiquetage des tours de parole attribués aux différents locuteurs en fonction d'une liste de rôles possibles dans le document sonore à traiter.

Une des première études sur le sujet a été publiée en 2000 (Barzilay *et al.*, 2000) et de nombreuses études récentes ont popularisé cette tâche dans la communauté du traitement automatique de la parole (Bigot *et al.*, 2010; Hutchinson *et al.*, 2010; Yaman *et al.*, 2010; Damnati & Charlet, 2011). Les approches diffèrent de par le nombre de rôles considérés (de 3 à 6), le type d'émissions (débats, reportages, interviews, etc.) et le niveau de segmentation utilisé pour l'évaluation : les segments, les tours de parole ou bien directement les locuteurs. Différents types de paramètres sont utilisés dans la phase de classification : paramètres acoustiques et prosodiques (Bigot *et al.*, 2010) ; paramètres lexicaux (Barzilay *et al.*, 2000; Hutchinson *et al.*, 2010) ou encore combinaison des deux (Damnati & Charlet, 2011).

Les études se distinguent également par le degré de supervision nécessaire à la production des paramètres utilisés dans l'étape de classification : depuis une supervision complète en utilisant les segmentations et transcriptions manuelles des émissions comme dans (Yaman *et al.*, 2010) ; ou bien sans supervision aucune avec des processus automatiques de segmentation et de transcription comme dans (Damnati & Charlet, 2011).

Par rapport aux études précédentes, nous nous proposons ici d'apporter un nouveau type de paramètres afin de caractériser les rôles de locuteurs, basé sur la catégorisation des questions présentée dans le paragraphe précédent. Le but de cette étude est de valider l'intérêt de ces paramètres grâce à des expériences contrastives que nous allons effectuer sur le corpus EPAC précédemment décrit. Les expériences sont faites sur les segmentations et les transcriptions manuelles du corpus EPAC afin de démontrer l'intérêt de notre approche indépendamment des erreurs faites durant les étapes de transcription par un système automatique. Néanmoins il nous restera dans une prochaine étude à valider ces résultats en montrant qu'ils restent pertinents même en présence d'erreurs de segmentation et de transcription.

4.1 Un classifieur pour la segmentation en rôle

Nous utilisons ici une méthode de classification supervisée basée sur un algorithme de combinaison de classifieurs simples (méthode de *boosting* dans la terminologie de l'apprentissage automatique). Ce type de classifieur discriminant a donné des résultats comparables aux approches basées sur les machines à vecteur de support (SVM) sur un grand nombre de tâches de classification tout en apportant un certain nombre d'avantages : d'une part une grande liberté est donnée dans la définition des classifieurs simples, ce qui permet de prendre en compte très facilement des paramètres hétérogènes tels que des symboles, des séquences de symboles ou bien encore des valeurs numériques ; d'autre part il est possible de connaître facilement quels classifieurs simples ont été choisis comme étant les plus discriminants pour la classification durant la phase d'apprentissage, et quel est leur poids dans le modèle final.

Dans toutes nos expériences l'implémentation ICSIBOOST (Favre *et al.*, 2007) de l'algorithme AdaBoost a été choisie comme méthode d'apprentissage et de classification. Ce classifieur est appliqué à chaque tour de parole de notre corpus en utilisant la segmentation en locuteur de référence produite manuellement. Il calcule un score pour chaque rôle possible pour chaque segment. En choisissant l'hypothèse ayant reçu le score maximum, nous obtenons une classification en rôle des différents tours de parole des locuteurs, chaque tour étant classé indépendamment des autres tours de parole du même locuteur.

Nous avons testé 3 types de paramètres pour cette phase de classification :

– la durée du tour de parole : ce paramètre est relatif à la structure de la conversation, il est pertinent car les différents rôles sont souvent caractérisés par des temps de parole très différents ;

– les 2-gram de mots : les choix lexicaux sont bien évidemment des paramètres majeurs dans l'attribution des rôles aux locuteurs ; nous considérons ici toutes les séquences de 2-gram de mots ;

– les labels des questions présentes dans le tour de parole ; ces labels représentent à la fois le nombre de questions se trouvant dans le tour, mais aussi leurs caractéristiques (directe/indirecte, totale/partielle, type).

Il est à noter que nous n'utilisons pas ici d'optimisation globale de la segmentation en rôles sur toute l'émission, ni d'informations connues *a priori* sur la structure de cette émission. notre but est d'effectuer une expérience contrastive de classification en rôle, sans pour autant chercher à obtenir les meilleurs taux possibles de classification sur ce corpus. En effet on peut grandement améliorer les résultats en utilisant des connaissances *a priori* sur le format de l'émission telles que : *les auditeurs sont toujours au téléphone* ; *le présentateur est toujours celui qui parle en premier* ; *après un auditeur il y a toujours une reprise de la parole du présentateur* ; etc. Ces informations relatives à la structure connue de l'émission *Le Téléphone Sonne* sont ignorées dans nos expériences.

4.2 Protocole expérimental et premiers résultats

Etant donné le nombre limité d'émissions *Le Téléphone Sonne* dans le corpus EPAC, nous avons utilisé un protocole expérimental basé sur la validation croisée par la méthode du *Leave-One-Out*. Ce protocole consiste, sur un jeu C de n exemples à classer, à retirer un exemple e du jeu d'exemples, à apprendre un classifieur B sur l'ensemble $C - \{e\}$, puis à tester B sur l'exemple e pour obtenir l'hypothèse e' que l'on ajoute à l'ensemble C', initalement vide. À l'issue de n itérations de cet algorithme, l'ensemble C' contient tous les exemples de C avec les hypothèses prédites par les n classifieurs. En comparant les hypothèses prédites dans C' aux hypothèses de référence de C nous obtenons une estimation de la qualité du processus de classification sur l'ensemble du corpus, sans le problème du biais de la sélection de corpus séparés pour l'apprentissage et le test.

Nous avons adapté le principe du *Leave-One-Out* à notre corpus d'émissions de la manière suivante :

– le corpus C contient 32 enregistrements de l'émission *Le Téléphone Sonne* : $C = \{e_1, e_2, \ldots, e_{32}\}$;
– à chaque itération i on sélectionne l'émission e_i comme étant le corpus de test T_i, l'émission e_{i+1} comme étant le corpus de développement D_i et les 30 émissions restantes $A_i = C - \{e_i, e_{i+1}\}$ constituent le corpus d'apprentissage A_i ;
– un classifieur B_i est entraîné sur les tours de parole du corpus A_i ; le nombre d'itérations de l'algorithme de boosting est choisi sur D_i et enfin les tours de parole du corpus T_i sont étiquetés automatiquement par B_i et rangés dans T'_i ;
– à l'issue des 32 itérations, le corpus $C' = \bigcup_{i=1}^{32} T'_i$ contient toutes les hypothèses de classification en rôles des tours de parole du corpus C.

En comparant les annotations manuelles de C et celles automatiques de C', nous pouvons évaluer la qualité de nos prédictions selon plusieurs métriques soit au niveau des tours de parole, soit au niveau des locuteurs. Étant donné que la répartition en rôles n'est pas uniforme (le nombre de tours de parole du présentateur est bien supérieur à celui des auditeurs ; inversement il y a bien plus d'auditeurs différents que de présentateurs), les métriques utilisées sont la précision, le rappel, la F-mesure [2] pour chaque type de rôles en complément de l'erreur totale de classification.

Notre première série d'expériences vise à conforter notre hypothèse initiale concernant la pertinence de la clas-

2. La précision, le rappel et la F-mesure sont calculées de la manière suivante :
   – Soit un échantillon $e \in C$ correspondant à un tour de parole (ou à un locuteur selon le niveau d'évaluation choisi) avec $r = ref(e)$ l'étiquette en rôle de référence contenue dans C et $r' = hyp(e)$ l'étiquette hypothèse prédites par les classifieurs B et contenue dans C'. Nous avons $r, r' \in \{$*présentateur,auditeur,expert,rapporteur*$\}$.
   – Si $r = r'$ alors *correct(r)=correct(r)*$+1$
   – Si $r \neq r'$ alors :
     – *erreur_totale=erreur_totale*$+1$
     – *suppression(r)=suppression(r)*$+1$
     – *insertion(r')=insertion(r')*$+1$
   – La mesure de précision pour l'étiquette r est : $P(r) =$ (*correct(r)* $\times$ 100) $\div$ (*correct(r)+insertion(r)*)
   – La mesure de rappel pour l'étiquette r est : $R(r) =$ (*correct(r)* $\times$ 100) $\div$ (*correct(r)+suppression(r)*)
   – La F-mesure pour l'étiquette r est : $F(r) = (P \times R \times 2) \div (P + R)$
   – La mesure d'erreur totale est définie par : $E = erreur\_totale \div |C|$

sification des questions pour caractériser les rôles des locuteurs. Pour cela nous avons effectué une expérience contrastive consistant à ajouter dans la liste des classifieurs simples utilisés par l'algorithme d'apprentissage, des paramètres liés à la présence ou non de questions dans les tours des locuteurs, puis des paramètres sur la forme et le type des questions posées. Nous obtenons 5 expériences contrastives définies de la manière suivante :

1. *durée+2-grams* : les seuls classifieurs utilisés ici sont l'absence ou la présence de bigrammes de mots dans les transcriptions des tours de parole ainsi qu'un classifieur sur la durée des tours de parole.
2. *durée+2-grams+question* : on ajoute aux classifieurs précédents un classifieur sur la présence ou l'absence de questions dans un tour de parole.
3. *durée+2-grams+question+directe/indirecte* : le label *directe/indirecte* est ajouté aux étiquettes *question*.
4. *durée+2-grams+question+directe/indirecte+totale/partielle* : même chose mais avec l'indication de portée de la question.
5. *durée+2-grams+question+directe/indirecte+totale/partielle+type* : on considère maintenant tous les types de questions, tels qu'ils sont définis dans le tableau 1.

paramètres	nb tests	durée+2-grams (1)	+question (2)	+directe/indir. (3)	+totale/part. (4)	+type (5)
F(auditeur)	*500*	66,4	67,0	66,9	65,4	**67,1**
F(expert)	*1443*	73,7	**74,6**	74,0	73,9	74,2
F(présentateur)	*1860*	81,2	81,5	81,3	81,5	**81,9**
F(rapporteur)	*163*	37,2	42,2	41,9	36,4	**57,3**
Erreur totale (*E*)	*3966*	24,7%	23,9%	24,2%	24,5%	**23,5%**

TABLE 4 – Résultats sur l'étiquetage de chaque tour de parole (annotation manuelle des questions et types de questions)

paramètres	nb tests	durée+2-grams (1)	+question (2)	+directe/indir. (3)	+totale/part. (4)	+type (5)
F(auditeur)	*220*	88,2	89,6	89,2	89,7	**90,4**
F(expert)	*118*	83,9	83,0	82,4	81,2	**84,4**
F(présentateur)	*35*	66,7	67,4	67,3	64,7	**68,8**
F(rapporteur)	*27*	45,7	45,7	36,4	36,4	**74,4**
Erreur totale (*E*)	*400*	17,8%	17,3%	18,0%	18,5%	**15,0%**

TABLE 5 – Résultats sur l'étiquetage de chaque locuteur (annotation manuelle des questions et types de questions)

Les résultats sont donnés dans le tableau 4 pour les tours de parole et dans le tableau 5 pour les locuteurs. Les résultats sur les locuteurs sont obtenus à partir de l'étiquette en rôle majoritaire de tous les tours de parole de ce même locuteur dans une émission donnée. Comme nous pouvons le voir, l'introduction du classifieur binaire *question/non question* améliore légèrement les résultats de classification en rôle, par contre l'ajout des labels *directe/indirecte* et *totale/partielle* n'améliore pas, voir dégrade les résultats.

Les meilleurs résultats sont obtenus en rajoutant le type des questions dans les paramètres de classification, ce qui conforte les analyses descriptives faites à partir de la table 3. On obtient une réduction significative de l'erreur totale, à la fois sur les tours de parole et les locuteurs, grâce à cette catégorisation. Ces résultats valident notre hypothèse initiale sur la pertinence des formes interrogatives pour caractériser les rôles des locuteurs dans des conversations.

Cependant, dans une perspective de réalisation d'un système entièrement automatique, il appartient maintenant de vérifier dans quelle mesure le type d'une question peut être déterminé automatiquement, et quel est l'impact des inévitables erreurs d'étiquetage en questions et en types de questions sur la tâche de segmentation en rôle. Comme précisé en début de paragraphe, nous utilisons dans cette étude les segmentations en locuteurs, en tours de parole ainsi que les transcriptions de référence (manuelles) de notre corpus. Nous limitons ainsi l'étude aux seules erreurs d'étiquetage en question, l'impact des erreurs de transcription et de segmentation en locuteurs est l'objet d'une étude en cours. Le paragraphe suivant présente une méthode d'étiquetage de questions dans des transcriptions de parole, utilisant à la fois des paramètres lexicaux et prosodiques. Les résultats de la segmentation en rôle utilisant ces étiquettes automatiques sont présentés dans le paragraphe 5.3.

5 Détection et classification automatique des questions

La tâche de détection automatique de questions dans des énoncés oraux a principalement été abordée dans des corpus de parole conversationnelles (Yuan & Jurafsky, 2005) et des enregistrements de réunions (Boakye *et al.*, 2009). Dans les deux cas les études se basent sur une segmentation *a priori* des énoncés, effectuée manuellement sur les transcriptions de référence. La tâche revient à une classification binaire des segments de parole : segment interrogatif ou affirmatif. Elle constitue ainsi une sous-tâche d'un étiquetage plus général des conversations en *actes de dialogue* qui consiste à segmenter un dialogue en unités discursives telles que : affirmation, question, appréciation, confirmation, négation, etc. Différentes listes d'actes de dialogue ont été proposées, comme par exemple la liste *DAMSL* (Core & Allen, 1997). Les paramètres utilisés sont principalement des indices lexicaux, prosodiques, également couplés à une analyse syntaxique dans (Boakye *et al.*, 2009).

Nous allons enrichir cette tâche dans nos expériences en rajoutant à cette classification binaire la classification en types de questions, en considérant les 8 types de questions suivants : *adverbe*, *complexe*, *déterminant*, *est-ce-que*, *inversion*, *pronom*, *si* et *intonation*. Les marqueurs des 7 premiers types de question sont des marqueurs "syntaxiques" dans la mesure où c'est la structure syntaxique des énoncés qui permet de les considérer comme des questions. Pour le dernier type, *intonation*, ce sont uniquement des marqueurs prosodiques qui permettent de qualifier les énoncés. Deux types de traitement ont donc été mis en oeuvre sur ces deux familles de questions : un classifieur basé sur des marqueurs syntaxiques, un classifieur basé sur des marqueurs acoustiques.

5.1 Caractérisation des questions avec marqueurs syntaxiques

Nous avons utilisé pour les 7 types de questions avec marqueurs syntaxiques la même méthodologie que pour la classification en rôle présentée dans le paragraphe 4. Cette fois chaque échantillon d'apprentissage ou de test correspond à un segment ou à une « phrase » manuellement annoté sur les transcriptions de référence du corpus. Est considérée comme phrase toute séquence de mots, à l'intérieur d'un tour de parole, séparée par un signe de ponctuation forte (point, point d'interrogation, point d'exclamation) ajouté par les annotateurs humains durant la phase de transcription manuelle [3]. Nous avons, sur les 32 émissions *Le Téléphone Sonne* de cette étude, un ensemble de 13224 segments dont 973 questions avec marqueurs syntaxiques et 562 questions *intonation*. Le classifieur ICSIBOOST a été entraîné sur ces échantillons en utilisant la méthodologie de validation croisée *Leave-One-Out* présentée dans le paragraphe 4.2. Les résultats sont présentés dans la table 6 en utilisant comme seuls paramètres des bigrammes de mots. Nous n'avons pas pour l'instant intégré d'informations relatives aux structures syntaxiques des énoncés, ce travail fait partie d'une étude en cours.

Classification	nb de segments	Précision	Rappel	F-mesure
segments interrogatifs	*995*	94,2	85,1	89,4
autres segments	*12229*	98,8	99,6	99,2
question type=adverbe	*223*	96,1	87,9	91,8
question type=complexe	*139*	79,0	67,6	72,9
question type=déterminant	*99*	87,6	78,8	83,0
question type=est-ce-que	*209*	96,7	97,6	97,1
question type=inversion	*159*	82,5	53,5	64,9
question type=pronom	*94*	80,9	58,5	67,9
question type=si	*45*	83,3	66,7	74,1
segments non interrogatifs	*12229*	98,4	99,7	99,1

TABLE 6 – Résultats sur l'étiquetage des segments en question et type de question

Comme nous pouvons le voir le taux de détection moyen des questions est satisfaisant (environ 90% de F-mesure), cependant de grandes disparités sont constatées selon le type de questions. De manière assez prévisible les questions de type *inversion* et *pronom* sont les plus difficiles à classer, ce qui justifie l'intérêt de disposer de paramètres liés à la structure syntaxique des énoncés et non pas seulement à leur lexicalisation. Cependant l'analyse syntaxique automatique de l'oral spontané est encore un domaine de recherche largement ouvert.

3. Bien évidemment tout symbole de ponctuation a été supprimé des transcriptions des segments dans toutes les expériences de classification

5.2 Caractérisation des questions avec intonation

Nous avons choisi de traiter les questions "purement" intonatives de notre corpus uniquement avec des paramètres prosodiques basés sur la courbe de fréquence fondamentale, ou $F0$ (Yuan & Jurafsky, 2005; Quang *et al.*, 2007). Ces paramètres sont obtenus directement à partir du signal de parole avec une fenêtre temporelle de 10 millisecondes. À partir de cette courbe nous proposons d'extraire un ensemble de 15 paramètres divisés en 3 classes : paramètres statistiques (6 paramètres), paramètres de trajectoire (5 paramètres) et paramètres de formes (4 paramètres). Voici une description rapide de ces paramètres qui sont calculés sur la fin de chaque phrase sur des périodes de 300 et 700 millisecondes :

– **Statistique** : nous avons 6 paramètres numériques sur la courbe de F0 : minimum, maximum, intervalle, moyenne, médiane et déviation standard de la F0 sur nos fenêtres de 300 et 700 millisecondes.
– **Trajectoire** : ces 5 paramètres décrivent si la courbe de fréquence fondamentale monte ou descend en fin de phrase.
– **Forme** : Les 4 paramètres de formes constituent l'une des originalités de cette étude. Ils consistent à modéliser la forme de la courbe de F0 grâce à une interpolation polynomiale Lagrangienne. Différents degrés de polynômes ont été testés et des résultats empiriques ont montré qu'un degré de 2 était satisfaisant pour la tâche. Nous utilisons donc les 3 paramètres a, b, c du polynôme $a * x^2 + b * x + c$ ainsi que l'erreur d'interpolation de la fonction approchée comme quatrième paramètre.

Une fois les 15 paramètres extraits, un classifieur est entraîné, en utilisant le même protocole que décrit précédemment, pour séparer les segments "question" des segments "autre".

question/non question	Précision	Rappel	F-Mesure
Forme+Statistique	0,62	0,37	0,46
Forme+Trajectoire	0,58	0,32	0,41
Statistique+Trajectoire	0,58	0,33	0,42
Combinaison	0,58	0,41	0,48

TABLE 7 – Combinaison des paramètres prosodiques basés sur la F0 pour la classification binaire *question/non question* de segments de parole

Une évaluation de ces paramètres est donnée dans la table 7 sur la classification binaire question/non question des segments du corpus. Comme nous pouvons le voir les meilleurs résultats sont obtenus en combinant les différents paramètres avec une F-mesure d'environ 50%. Dans notre système de classification du type des questions, ce classifieur est utilisé de la manière suivante : si un segment n'est pas considéré comme une question par le classifieur basé sur les marqueurs syntaxiques mais qu'il est classé *question* par le classifieur prosodique, alors le segment reçoit l'étiquette *question intonation*.

5.3 Évaluation sur la segmentation en rôle

niveau	*tours de parole*			*locuteurs*		
paramètres	nb tests	type question (ref)	type question (aut)	nb tests	type question (ref)	type question (aut)
F(auditeur)	*500*	65,9	66,5	*220*	90,2	90,2
F(expert)	*1443*	74,8	73,8	*118*	83,6	83,3
F(présentateur)	*1860*	82,5	81,2	*35*	77,7	71,7
F(rapporteur)	*163*	56,7	52,4	*27*	68,3	57,9
Erreur totale (E)	*3966*	23,1%	24,2%	*400*	14,5%	15,8%

TABLE 8 – Résultats sur l'étiquetage en rôle des tours de parole et des locuteurs. Comparaison annotation manuelle/automatique des questions avec leurs types

La table 8 présente les résultats obtenus sur la tâche de segmentation en rôle en comparant l'utilisation des étiquettes de type de question de référence (manuelles) à celles produites automatiquement par les deux classifieurs présentés dans ce paragraphe. Comme nous pouvons le voir, même si une dégradation est constatée dans les performances à cause des erreurs de détection et de classification des questions, les résultats restent meilleurs que

ceux obtenus sans ces paramètres : -0.5% d'erreur totale pour les tours de parole et -2% d'erreur totale pour les locuteurs.

6 Conclusion

Nous avons proposé dans cette étude une analyse du type des questions exprimées dans un corpus de conversation d'émissions de radio. Nous avons montré que le typage des questions pouvait être un indicateur du rôle du locuteur à l'intérieur de la conversation. Nous avons validé cette hypothèse sur une tâche de segmentation automatique en rôle des locuteurs de notre corpus, en constatant une amélioration significative des résultats après ajout de paramètres liés au typage des questions dans le processus de classification automatique. Enfin nous avons validé la mise en pratique de ces paramètres en montrant qu'on pouvait les obtenir de manière complètement automatique au prix d'une légère dégradation des résultats. Il nous reste cependant à nous attaquer au défi que constitue la segmentation automatique en unité, ou « pseudo-phrases » de l'oral spontané. Se baser uniquement sur les pauses ou les groupes de souffle ne permet pas de segmenter de manière cohérente les énoncés, et les résultats des systèmes de segmentation automatiques en phrases basés sur la prosodie et des indices syntaxiques, s'ils obtiennent des résultats intéressants sur de la parole lue ou préparée, sont encore très insuffisants pour être utilisé directement sur des conversations spontanées. À terme, une solution pourrait consister à utiliser des méthodes et paramètres syntaxiques qui se passeraient d'une segmentation en phrases ou « pseudo-phrases », afin de ne pas être exposé aux limites que sous-entend cette tâche sur la parole spontanée.

Références

BARZILAY R., COLLINS M., HIRSCHBERG J. & WHITTAKER S. (2000). The rules behind roles : Identifying speaker role in radio broadcasts. In *Proc. of AAAI*.

BIGOT B., PINQUIER J., FERRANÉ I. & ANDRÉ-OBRECHT R. (2010). Looking for relevant features for speaker role recognition. In *Proc. of Interspeech*.

BOAKYE K., FAVRE B. & HAKKANI-TÜR D. (2009). Any Questions ? Automatic Question Detection in Meetings. In *ASRU, Merano (Italy)*.

CORE M. & ALLEN J. (1997). Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, p. 28–35 : Citeseer.

DAMNATI G. & CHARLET D. (2011). Robust speaker turn role labeling of tv broadcast news shows. In *ICASSP'2011*.

ESTÈVE Y., BAZILLON T., ANTOINE J.-Y., BÉCHET F. & FARINAS J. (2010). The EPAC corpus : manual and automatic annotations of conversational speech in French broadcast news. In *LREC*, Malta.

FAVRE B., HAKKANI-TÜR D. & CUENDET S. (2007). Icsiboost. `http://code.google.come/p/icsiboost`.

GARCIA-FERNANDEZ A. & LAILLER C. (2008). Morphosyntaxe de l'interrogation pour le système question-réponse ritel. In *RECITAL 2008*.

HUTCHINSON B., ZHANG B. & OSTENDORF M. (2010). Unsupervised broadcast conversation speaker role labeling. In *Proc. of ICASSP*.

QUANG V., BESACIER L. & CASTELLI E. (2007). Automatic question detection : prosodic-lexical features and cross-lingual experiments. In *Proc. Interspeech*, volume 2007, p. 2257–2260.

STOLCKE A., RIES K., COCCARO N., SHRIBERG E., BATES R., JURAFSKY D., TAYLOR P., MARTIN R., ESS-DYKEMA C. & METEER M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, **26**(3), 339–373.

YAMAN S., HAKKANI-TUR D. & TUR G. (2010). Social role discovery from spoken language using dynamic bayesian networks. In *Proc. of Interspeech*.

YUAN J. & JURAFSKY D. (2005). Detection of questions in Chinese conversational speech. In *2005 IEEE Workshop on Automatic Speech Recognition and Understanding*, p. 47–52.

Sémantique

Recherche d'information et temps linguistique : une heuristique pour calculer la pertinence des expressions calendaires

Charles Teissèdre (1,2) Delphine Battistelli (3) Jean-Luc Minel (1)

(1) MoDyCo - UMR 7114 CNRS, Paris Ouest Nanterre La Défense, 200, av. de la République, 92001 Nanterre
(2) Mondeca, 3, cité Nollez, 75018 Paris
(3) STIH, Université Paris Sorbonne, 28, rue Serpente, 75006 Paris
charles.teissedre@ u-paris10.fr, delphine.battistelli@paris-sorbonne.fr, jean-luc.minel@u-paris10.fr

Résumé. A rebours de bon nombre d'applications actuelles offrant des services de recherche d'information selon des critères temporels - applications qui reposent, à y regarder de près, sur une approche consistant à filtrer les résultats en fonction de leur inclusion dans une fenêtre de temps, nous souhaitons illustrer dans cet article l'intérêt d'un service s'appuyant sur un calcul de similarité entre des expressions adverbiales calendaires. Nous décrivons une heuristique pour mesurer la pertinence d'un fragment de texte en prenant en compte la sémantique des expressions calendaires qui y sont présentes. A travers la mise en œuvre d'un système de recherche d'information, nous montrons comment il est possible de tirer profit de l'indexation d'expressions calendaires présentes dans les textes en définissant des scores de pertinence par rapport à une requête. L'objectif est de faciliter la recherche d'information en offrant la possibilité de croiser des critères de recherche thématique avec des critères temporels.

Abstract. Unlike many nowadays applications providing Information Retrieval services able to handle temporal criteria - applications which usually filter results after testing their inclusion in a time span, this paper illustrates the interest of a service based on a calculation of similarity between calendar adverbial phrases. We describe a heuristic to measure the relevance of a fragment of text by taking into account the semantics of calendar expressions. Through the implementation of an Information Retrieval system, we show how it is possible to take advantage of the indexing of calendar expressions found in texts by setting scores of relevance with respect to a query. The objective is to ease Information Retrieval by offering the possibility of crossing thematic research criteria with temporal criteria.

Mots-clés : Indexation d'informations calendaires ; Recherche d'information ; Annotation et extraction d'expressions calendaires
Keywords: Calendar information indexing ; Information Retrieval ; Annotation and extraction of calendar expressions

1 Introduction

Les systèmes de Recherche d'Information sur le Web destinés au grand public ne sont pas en mesure aujourd'hui de répondre à des requêtes exprimant des informations calendaires complexes, telles qu'on peut les exprimer en langue au travers notamment de la catégorie des adverbiaux calendaires (Klein, 1994). Dans de nombreux cas de figure, cette catégorie pourrait pourtant utilement intervenir dans le calcul de la pertinence d'un fragment de texte ou d'un document. En effet, la prise en compte de la sémantique des expressions adverbiales calendaires (soulignées ci-après) permettrait ainsi de traiter des requêtes comme celles qui suivent :

- « *Festival de musique aux alentours de la mi-août* »
- « *La France à la fin du XVII^e siècle* »
- « *Cinéma italien au début des années 60* »

Si les systèmes de bases de données spécialisés peuvent permettre de fournir des réponses à de telles requêtes en filtrant les réponses incluses dans une fenêtre de temps, pour autant, parce qu'elles s'appuient sur une représentation discrète, elles n'offrent pas de moyen d'ordonnancer les résultats selon des critères permettant de mesurer la pertinence relative des propriétés temporelles retournées par rapport à la requête. De tels systèmes ne permettent pas, par exemple, de jongler avec la granularité des expressions temporelles pour retenir en priorité celles qui partagent les mêmes caractéristiques que la requête ; en effet, le plus souvent, les réponses fournies dans ce cadre de recherche sont uniquement celles qui répondent favorablement au test d'inclusion dans la période recherchée.

Par delà le test d'inclusion. Une autre approche nous semble possible. Elle s'inspire du fonctionnement des moteurs de recherche, à savoir leur capacité à trier les documents par pertinence en évaluant la distance qui les rapproche ou les éloigne d'une requête. L'objectif est alors de fournir des critères pour calculer des scores de proximité entre les zones temporelles recherchées et des expressions calendaires présentes dans un texte ou un corpus de textes. La méthode que nous décrivons pour attribuer un score de pertinence temporelle permet ainsi de combiner des requêtes thématiques et des requêtes temporelles calendaires, en agrégeant et pondérant les scores de pertinence.

Afin de mettre en regard notre approche avec les travaux de recherche et les applications existants, l'article s'ouvre sur un état des lieux de la recherche d'information fondée sur des critères calendaires (section 2). Il se poursuit par la présentation de l'heuristique que nous proposons pour calculer des scores de pertinence temporelle (section 3). Enfin (section 4), nous illustrons l'intérêt de cette approche à travers une expérimentation qui prend corps dans un outil de recherche d'information destiné à montrer le type de résultats qu'il est possible d'obtenir en croisant des critères de recherche thématique avec des critères temporels.

2 Etat de l'art

2.1 Recherche d'Information et temps calendaire

Le traitement automatique de l'« information temporelle » exprimée dans les textes s'impose depuis quelques années comme un champ de recherche important auquel on associe des retombées dans le domaine de la recherche d'information (Alonso et al., 2007 ; Mestl et al., 2009) ; parmi les applications visées : les systèmes de questions/réponses, les systèmes de résumé automatique, les moteurs de recherche sur le web et, intégrés ou non à ces derniers, les systèmes visant à proposer en sortie une visualisation des informations sur une ligne du temps.

La caractérisation de l'« information temporelle » en tant que telle constitue un enjeu – au cœur des programmes d'annotation automatique - tant sur le plan descriptif (quelles sont les unités de la langue qui expriment une information temporelle ?) que sur le plan analytique (quels sont les niveaux de représentation et les stratégies calculatoires à mettre en œuvre pour appréhender la catégorie sémantique du temps ?). Dans le champ de la recherche d'information, par rapport auquel se situent précisément le plus souvent les dits

programmes d'annotation automatique, l'information temporelle est la plupart du temps rapportée à ce qui permettrait la résolution d'une tâche en particulier : celle du calcul de l'ancrage calendaire de situations (souvent appelées « événements ») décrites dans les textes. On pourra se reporter à (Battistelli, 2011) pour une présentation des enjeux descriptifs de la temporalité linguistique pour des systèmes de recherche d'information. Dans les applications citées ci-dessus, les expressions linguistiques référant explicitement à un calendrier (le calendrier grégorien par exemple) ont ainsi toujours constitué un champ d'investigation particulièrement exploré. C'est d'ailleurs dans le cadre des systèmes de questions/réponses qu'a été organisée pour la première fois en 2004 une tâche d'évaluation uniquement dévolue à la problématique de repérage puis de normalisation (i.e. de calcul en référence à une norme) de ce type d'expressions : *Time Expression Recognition and Normalization* (TERN) ; tâche plus largement à l'origine de la démarche visant à proposer une standardisation quant à l'annotation sémantique de ces expressions (cf. en particulier (Schilder et Habel, 2001 ; Pustejovsky et al., 2002 ; Ferro et. al., 2003 ; Saquete et. al., 2004 ; Ehrmann et Hagège, 2009 ; Bittar, 2010), avec en corollaire, l'élaboration de corpus annotés tels que ACE[1] et TimeBank[2] et d'un certain nombre de systèmes automatiques (cf. par exemple Mani et Wilson, 2000; Han et al., 2006; Ahn et al., 2007). Depuis peu, avec les avancées récentes sur le terrain de l'acquisition d'informations temporelles de type calendaire dans les textes dont les résultats commencent à être exploitables (UzZaman et Allen, 2010 ; Llorens et al., 2010), le champ des applications s'étend progressivement à d'autres initiatives originales, telles que la construction automatique de chronologies pour explorer et visualiser le contenu de corpus de presse (ainsi des travaux de Alonso et al, 2010 qui s'appuient sur l'outil de production de chronologies SIMILE timeline[3] ou encore ceux de Matthews et al., 2010). Parmi les quelques initiatives des moteurs de recherche sur le terrain de la recherche d'information selon des critères temporels, citons celle de Google qui permet de visualiser les résultats d'une recherche sur une chronologie, puis de filtrer les résultats sur une fenêtre de temps (il s'agit du service view:timeline). Pour autant, si ce service tire parti des expressions calendaires présentes dans les textes, il ne propose pas à proprement parler de formuler des requêtes temporelles. Du reste, seule une sous-partie des expressions calendaires rencontrées dans les textes est indexée et donne lieu à une analyse (peu ou prou, les expressions de la forme JJ MM AAAA, MM AAAA ou AAAA). Ces expressions sont en outre systématiquement réduites à une représentation atomique : ainsi, une expression telle que « *de 1815 à 1871* » n'est pas analysée comme formant une zone temporelle bornée à gauche et à droite, mais plutôt comme deux dates. On retrouve un comportement similaire dans des systèmes de gestion de connaissances structurées tel que le projet TimeSearch History[4] qui permet de croiser une recherche par mots-clés et une recherche temporelle : le champ dédié au filtre temporel permet uniquement de spécifier une année.

La difficulté du traitement des informations calendaires exprimées en langue tient dans ceci qu'elles font intervenir des opérations de régionalisation (du type avant, après, etc.), de focalisation (du type début, fin, milieu), des opérations de pointage (consistant à désigner une zone de référence sur le calendrier) et des grains calendaires variés (jour, mois, année, parties du jour, etc.). De là l'impossibilité d'organiser les expressions calendaires selon un ordre total : comment en effet ordonner d'après un unique critère des expressions aussi variées que « *avant 2009* », « *en mars 2009* », « *de mi 2009 à fin 2011* », « *aux alentours de 2009* », etc. ? De là également la tentation de simplifier le problème du traitement de ces expressions pour la recherche d'information, en les réduisant à une représentation atomique qu'il devient alors possible d'ordonner, quitte à perdre une grande partie de leur sémantique. Ce problème renvoie à celui de l'opposition entre duratif et ponctuel, identifié en Intelligence Artificielle comme un « problème de granularité » (Bettini et al., 2000 ; Bechet et al., 2000), dont il est pourtant possible de sortir en considérant qu'il s'agit essentiellement d'une question d'échelle.

En dépit des quelques initiatives mentionnées, les expressions calendaires dans les textes sont encore très largement aujourd'hui traitées par les moteurs de recherche grand public comme des mots-clés dont la sémantique n'est pas ou peu exploitée. Ainsi, une recherche par mots-clés sur un intervalle de temps (mettons « *de 1750 à 1800* ») ne ramène que des résultats où les termes mêmes de la recherche apparaissent (on pourra ainsi trouver des expressions telles que « *en 1750* » ou « *en 1800* », mais pas des expressions telles que « *peu après 1763* » ou « *de 1755 à 1799* »).

[1] Cf. les corpus LDC2005T07 et LDC2006T06 du catalogue LDC (http://www.ldc.upenn.edu).

[2] Cf. le corpus LDC2006T08 du catalogue LDC.

[3] SIMILE Timeline toolkit: http://simile.mit.edu/timeline/

[4] http://www.timesearch.info/

2.2 Inclusion vs. similarité temporelle

Pour comprendre l'intérêt d'un principe de pertinence temporelle que nous souhaitons appréhender, il faut souligner les limites des approches réduisant l'interrogation temporelle à l'inclusion dans une fenêtre de temps - approche retenue dans les systèmes de gestion de BDD ou dans les moteurs de recherche qui proposent d'indexer des données temporelles. En filtrant la recherche par l'inclusion, le risque est (1) de ne pas renvoyer de résultat, (2) de ne pas pouvoir ordonner les résultats par pertinence sous l'angle des propriétés temporelles, et (3) d'avoir des difficultés à jongler entre des granularités ou échelles de temps différentes. Ces limites tiennent à la méthodologie retenue, qui relève, au fond, d'une approche booléenne (inclusion vs. non inclusion). L'intérêt du calcul d'un score de similarité est d'échapper à cette approche restrictive, en permettant d'évaluer la distance/proximité entre différentes caractéristiques des expressions calendaires. Si les relations entre intervalles de temps telles que les a décrites (Allen, 1983) permettent de comparer des expressions calendaires et d'obtenir des résultats booléens (en testant l'inclusion, le recouvrement, l'intersection, etc.), elles ne permettent pas à elles seules de les hiérarchiser par pertinence, ce qui constitue le cœur de notre approche. Pour cela il faut recourir à des mesures telles que la distance temporelle entre intervalles de temps, le taux de recouvrement, les rapports de proportions entre expressions, etc.

(Le Parc-Lacayrelle et al., 2007) proposent une méthode de calcul de la pertinence temporelle par rapport aux centroïdes des intervalles de temps. La méthode décrite consiste à calculer la pertinence relative entre la requête et la partie de l'index qui entre en intersection avec la requête. Pour deux expressions incluses dans la période définie par la requête (par exemple, les expressions « *en 1804* » et « *en 1859* » pour une requête telle que « *au XIX^{e} siècle* »), on privilégie celle qui est la plus proche du centre (« *en 1859* »). Ces travaux ont ceci d'intéressant qu'ils introduisent la notion de pertinence en proposant une méthode d'ordonnancement de fragments de documents (qui renvoie aux problématiques de « ranking » ou de « scoring » familières aux moteurs de recherche). Ils se limitent toutefois à l'intersection entre une requête et des expressions calendaires présentes dans les textes (soit encore à l'inclusion dans une fenêtre de temps, bien que l'inclusion soit ici entendue en un sens moins restrictif). Ceci a pour effet, pour une requête telle que « *le 12 août 1988* », d'exclure des résultats des expressions telles que « *dans la nuit du 10 au 11 août 1988* » qui sont pourtant susceptibles de présenter un intérêt.

2.3 L'annotation des expressions calendaires

Dans le cadre du traitement automatique des informations calendaires dans les textes, il est nécessaire que les systèmes d'annotation soient en mesure de traiter des expressions parfois très complexes tout en proposant une représentation formelle manipulable par les machines et suffisamment riche pour couvrir au mieux la manière dont la langue exprime une référence au calendrier : le modèle d'annotation a donc ici toute son importance.

Quelles que soient les difficultés rencontrées par les systèmes, en particulier quant au calcul de la valeur d'une expression relative - déictique ou anaphorique - (Cailliau et al., 2008 ; Wang et Zang, 2008 ; Mazur et Dale, 2008), il reste qu'aucune démarche n'a à notre connaissance pris pour objet l'analyse des *relations* entre ces expressions en tirant parti, non seulement de leurs valeurs, mais aussi des unités linguistiques à proprement parler dans lesquelles elles entrent le plus souvent, à savoir des unités adverbiales. A dire vrai, ce ne sont d'ailleurs pas des expressions temporelles adverbiales qui font l'objet d'une annotation (si l'on s'en tient en particulier aux deux projets d'annotation majeurs à l'heure actuelle que sont TIMEX2 (Ferro et al., 2003) et TIMEX3 (Pustejovsky et al., 2003, 2005), mais uniquement leur référence au système calendaire à proprement parler, la signification des locutions prépositionnelles étant vouée à être traitée à part (via une balise nommée SIGNAL). D'autres approches, comme (Aunargue et al., 2001), (Battistelli et al., 2008), (Teissèdre et al., 2010) tiennent compte, elles, des prépositions et de l'ensemble des éléments qui interviennent dans la composition des adverbiaux temporels – éléments qui s'avèrent très utiles pour l'analyse sémantique et l'indexation d'expressions telles que « *au début du XXe siècle* », « *vers le milieu des années 1950* », « *3 mois avant la fin de l'année* ».

Le modèle d'annotation proposé dans (Battistelli et al., 2008) permet ainsi un traitement fin de la granularité, en décrivant la composition des expressions calendaires et la façon dont elles imbriquent, par-dessus une référence à un système calendaire (autrement dit, la base calendaire), des opérations de pointage (pointage déictique, anaphorique ou absolu), de focalisation (début, milieu, fin), de déplacement (ex : « *deux jours avant* ») et de régionalisation (avant, après, pendant, jusqu'à, etc.).

Enfin, on peut mentionner des travaux qui portent sur les expressions itératives (Teissèdre et al., 2010), qui requièrent également des traitements particuliers, faisant notamment appel à des outils de raisonnement pour passer de propriétés temporelles définies en intension (ex: « *tous les lundis* ») à une représentation en extension (ex : le lundi 15 mars, le lundi 22 mars, etc.). Ces traitements permettent d'appréhender des expressions complexes comme les dates et horaires d'ouverture (ex : « *ouverture du lundi au vendredi, de 9h à 19h.* »).

S'appuyant sur les informations issues des moteurs d'annotation, le système d'indexation et de recherche que nous avons développé met en regard une requête et des expressions calendaires présentes dans les documents indexés, afin de faire ressortir les parties les plus pertinentes de ces documents. La modélisation retenue par les systèmes d'annotation influe donc lourdement en aval sur la capacité à comparer entre elles les caractéristiques de différentes expressions calendaires.

3 Vers une heurististique de calcul de la pertinence temporelle

3.1 L'intérêt d'évaluer la pertinence temporelle

Afin d'illustrer l'intérêt du calcul de la pertinence temporelle, la figure 1 montre la façon dont le système que nous avons développé hiérarchise, pour deux requêtes différentes, un même ensemble d'expressions calendaires. Les résultats sont présentés à la manière d'un nuage de tags où les éléments les plus pertinents sont surlignés graphiquement : la pertinence d'un résultat joue sur la taille de la police et le niveau de gris[5].

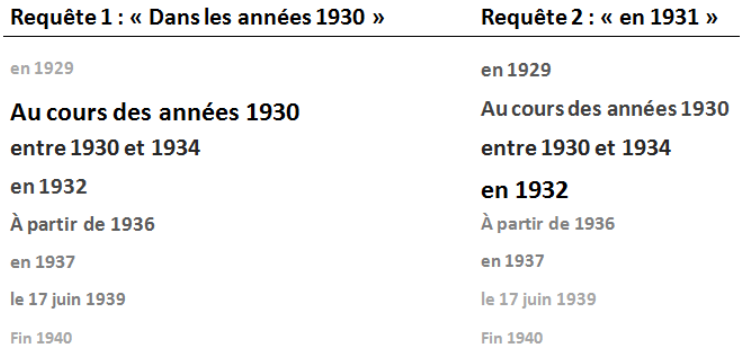


Fig. 1 : nuages d'expressions calendaires proposées pour deux requêtes

La figure 1 montre que la hiérarchie des résultats varie en fonction des caractéristiques de la requête : elle dépend de plusieurs critères, qui tiennent compte de l'inclusion, certes, mais aussi de la distance temporelle et des rapports de proportion (ou granularité) entre la requête et les expressions calendaires indexées. L'outil de navigation présenté dans la dernière section s'appuie sur cet algorithme d'ordonnancement pour présenter les portions de textes les plus pertinentes pour une requête comportant un critère temporel. La figure permet également d'illustrer l'intérêt de l'approche par rapport aux systèmes qui filtrent les résultats sur le seul critère d'inclusion : la requête 2 présente une étendue temporelle restreinte, dans laquelle aucun des résultats présentés n'est inclus. Pour autant, le système est tout de même en mesure de fournir des résultats classés par pertinence, même s'ils ne sont pas inclus dans la période définie par la requête de l'utilisateur.

3.2 Heuristique de pertinence temporelle

Les critères entrant dans le calcul de la pertinence temporelle sont : (i) le score brut d'intersection, (ii) la mesure de la distance entre la requête et les expressions calendaires apparaissant dans les documents, (iii) la mesure du rapport de proportion (iv) la mesure de la distance pondérée par la granularité et (v) la mesure du

[5] D'autres choix visuels sont à l'étude, afin de permettre de jongler entre différents types d'ordonnancement tout en laissant apparaître le degré de pertinence des expressions calendaires par rapport à une requête (tri par ordre de pertinence, tri par ordre d'apparition dans un texte, tri selon l'ordre temporel, etc.).

taux de recouvrement. Chacun de ces scores produit un résultat entre 0 et 1, où 1 est le meilleur score possible. La logique de ces scores est illustrée par différentes figures où les expressions calendaires (EC) dont on évalue la pertinence relative sont représentées au dessus de l'axe du temps : la dénomination « EC Requête » y désigne la zone temporelle définie par la requête et « EC Rép. » les zones temporelles définies par les expressions formant un ensemble possible de réponses.

- Score brut d'intersection

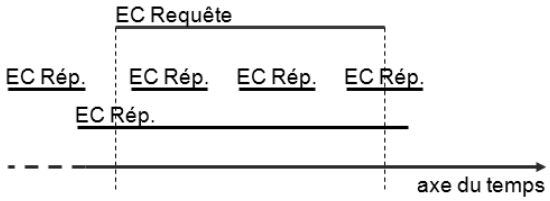


Fig. 2 : intersection

Ce score recouvre une sous partie des relations décrites dans (Allen, 1983) : la relation « during » (inclusion), la relation « overlaps » (recoupement), la relation « meets » (adjacence) et leur relation inverse. Les scores attribués diffèrent pour l'inclusion complète, partielle, ou inverse (lorsque la requête est incluse dans une expression d'étendue plus vaste) ; le score est nul pour une expression qui n'a pas d'intersection avec la requête. Ce score permet de catégoriser les documents à ordonner en cinq classes (intervalles inclus, incluant, en intersection, concomitant ou exclus). Ce critère est insuffisant à lui seul, parce que l'intersection n'est pas toujours synonyme de plus grande pertinence : ainsi, pour une requête temporelle telle que « *en 1953* », l'expression « *durant le XX*^{e} *siècle* » obtient un meilleur score d'intersection que l'expression « *en 1954* », qui est pourtant plus proche de la requête du point de vue de la granularité. La catégorisation obtenue à l'aide du score d'intersection est utile pour décider ensuite, parmi les autres critères de mesure de la pertinence ceux qui seront calculés (score de distance ou score de recouvrement), mais aussi ceux des scores intermédiaires qu'il faudra privilégier au détriment des autres.

- Score de distance

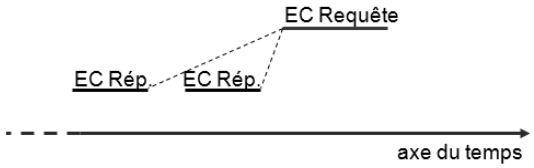


Fig. 3 : distance

La distance correspond à l'étendue de temps séparant deux expressions calendaires. Ce score est calculé pour les expressions dont l'intersection avec la requête est nulle. Pour celles où il y a intersection, le pendant du score de distance est le taux de recouvrement. La mesure de la distance porte sur la borne droite pour les expressions s'achevant avant la période désignée par la requête et sur la borne gauche pour les expressions débutant après.

- Score de proportion

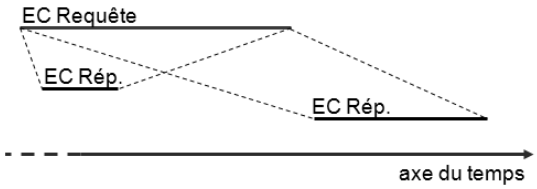


Fig. 4 : proportion

La mesure de la proportion relative permet d'évaluer la proximité des résultats avec la requête du point de vue de la granularité. Pour une requête telle que « *le 15 juillet 2007* », il semble en effet plus pertinent d'avoir en tête de liste une date comme « *le 16 juillet 2007* » plutôt que « *en 2007* » ou encore « *au XXI*^{e}

siècle ». On postule donc ici qu'un utilisateur sera vraisemblablement plus intéressé par une réponse qui n'est pas nécessairement incluse dans la période définie par sa requête, mais qui en est proche sémantiquement, au sens où la nature de la requête détermine les caractéristiques attendues dans les résultats.

- Score de distance pondéré par la granularité

Ce score vise à pondérer la mesure de l'étendue de temps séparant les deux expressions comparées en la faisant dépendre de la granularité de la requête : la distance entre deux expressions telles que « *en 1840* » et « *en 1860* » est évaluée comme étant plus faible qu'entre les expressions « *le 11 juin 1840* » et « *le 20 mars 1860* », car la granularité des premières est plus étendue.

- Score de recouvrement

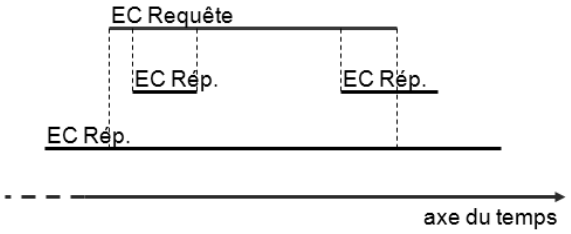


Fig. 5 : score de recouvrement

Ce score reprend ici la proposition de modèle de calcul de la pertinence formulée par (Le Parc et al., 07), à ceci près que l'on fait varier le centroïde d'une expression en fonction de la nature des opérations de régionalisation ou de focalisation qui interviennent dans la composition des expressions comparées. On privilégie ainsi la distance par rapport à la borne du début pour la plupart des expressions : ceci permet, pour une requête telle que « *dans les années 80* », de faire remonter dans la hiérarchie des résultats « *en 1981* » plutôt que « *en 1985* » ; ce qui revient à dire qu'on privilégie l'ordonnancement par rapport à l'axe du temps (de gauche à droite). Pour les expressions présentant une régionalisation de type « avant » ou « jusqu'à » (expressions définissant des intervalles infinis à gauche), la distance des intervalles comparés dépend de la borne droite. On considère en revanche le centroïde pour les expressions présentant une focalisation de type « milieu » (« *à la mi-mai* »). Ce score n'a de sens et n'est calculé qu'en cas d'intersection entre la requête et l'expression testée.

Le cas des intervalles infinis à gauche ou à droite. Les intervalles infinis à gauche ou à droite (ex. : « *depuis la fin du XVIe siècle* », « *jusqu'au 13 mars 2003*) demandent un traitement spécifique pour le calcul des proportions relatives et du taux de recouvrement, parce que leur durée est infinie. Ainsi, pour le calcul du score de proportion, on ne retient que l'étendue de la borne principale (« *1950* » pour l'expression « *jusqu'en 1950* »). On suppose donc que l'étendue de cette borne détermine la proportion des expressions susceptibles d'intéresser l'utilisateur : une requête telle que « *avant le XVIe siècle* » renverra en tête de liste des résultats de même granularité. Pour le calcul de l'intersection, on attribue une valeur minime, mais non nulle, qui place tous les résultats qui sont en intersection avec la requête au même niveau.

Combiner les modèles de pertinence. La difficulté pour établir un score global de pertinence temporelle, c'est que, pris isolément, chacun de ces critères peut produire un ordre ou un classement différent des autres : il s'agit donc de les combiner afin de produire des résultats cohérents et pertinents. Notre proposition d'heuristique consiste à agréger les scores intermédiaires, avec un facteur d'augmentation ou de réduction des scores.

$$\begin{aligned}\text{Score de pertinence temporelle} = {} & K \times \text{Score d'intersection} + \\ & L \times \text{Score de distance} + \\ & M \times \text{Score de proportion} + \\ & N \times \text{Score de distance pondéré} + \\ & O \times \text{Score de recouvrement}\end{aligned}$$

Les facteurs de maximisation d'un score et le calcul même des scores sont interdépendants. Ils peuvent ainsi varier selon que les résultats sont ou non inclus dans la période déterminée par la requête. En cas de non inclusion, plus la distance est grande, moins les rapports de proportions importent :

Score de Proportion = Score Proportion × Score Distance

De la même façon, si le score de proportion est très faible, on minimise le score de distance et le score de recoupement. A ce stade les facteurs intervenant dans le calcul de pertinence ont été déterminés empiriquement. Nous travaillons actuellement à une formalisation plus systématique, à la fois du modèle de pertinence et de la prise en compte des pôles des intervalles (début, milieu, fin), qui fera l'objet d'un prochain article (Battistelli et al., soumis).

4 Un outil de navigation temporelle dans les textes

L'outil de navigation textuelle que l'on décrit dans cette section permet à l'utilisateur/lecteur d'effectuer des recherches thématiques et temporelles dans un texte ou un ensemble de textes, pour filtrer ou surligner les passages les plus pertinents. Les fonctionnalités de recherche s'appuient sur les techniques d'indexation auxquelles ont communément recours les moteurs de recherche. De ce point de vue, l'outil développé peut tout aussi bien être présenté comme un prototype de moteur de recherche en mesure de traiter des requêtes temporelles.

Le modèle retenu pour établir la pertinence par rapport aux mots-clés est ici plutôt simpliste et repose sur les outils standards proposés dans la suite Lucene. Cette expérimentation a en effet pour fonction d'illustrer le type de résultats qu'il est possible d'obtenir en croisant une requête thématique (recherche par mots-clés) et une requête temporelle. Ceci implique, sur un plan technique, de combiner les scores de pertinence obtenue par ces différents modèles - temporel et thématique -, afin d'obtenir une liste de résultats hiérarchisés. Naturellement, les deux modèles de calcul de la pertinence entrent en contradiction, ce qui oblige à trouver un juste équilibre entre eux. Sur ce plan toutefois, il s'agit d'un problème classique que tout moteur de recherche doit gérer, puisqu'il leur faut bien croiser les modèles de pertinence par rapport aux mots-clés (« word similarity »), par rapport à la popularité des sites Web (« popularity »), par rapport à la qualité du site (« trust »), pour ne reprendre que quelques uns des modèles d'ordonnancement les plus connus, qui chacun produit un ordre différent des autres.

Le démonstrateur, un service Web accessible en ligne[6], permet à l'utilisateur de charger un texte ou un corpus de textes. Un premier pré-traitement segmente les textes en paragraphes et en phrases, dont les expressions calendaires sont alors annotées, puis indexées. Pour cette expérimentation, les expressions relatives, qui demandent un calcul particulier (les déictiques et anaphoriques tels que « *demain* », « *deux jours plus tard* »), ne sont pas traitées. Seules les expressions absolues, qui peuvent être disposées sur le calendrier sans recourir à des outils de résolutions des anaphores, sont à ce jour indexées.

4.1 Pré-traitements : des expressions calendaires aux intervalles calendaires

Les ressources pour l'annotation des expressions calendaires utilisées dans notre système consistent en un ensemble de transducteurs Unitex (Paumier, 2000) présentés dans (Teissèdre et al., 2010). Les expressions annotées par ces ressources sont ensuite transformées en intervalles calendaires. Ce passage du modèle linguistique ou symbolique vers le modèle calendaire constitue une problématique à part entière. Elle est prise en charge par un module décrit dans (Battistelli et al., 2011). Retenons donc ici que les expressions calendaires annotées sont transformées en un ou plusieurs intervalles (pour les itératifs en particulier, comme « *tous les jours sauf le dimanche* »), dont les bornes peuvent être transformées au format ISO-8601.

Exemples de transformation vers des intervalles calendaires

- « *de 1830 à 1940* » : [1830, 1940]
- « *XVIe siècle* » : [15**, 15**][7]

[6] http://client1.mondeca.com/TemporalQueryModule/

[7] La notation YY** renvoie à des siècles et la notation YYY* à des décennies. Les intervalles peuvent également s'écrire de la façon suivante [début(date1), fin(date2)], dans la mesure où les bornes des intervalles (date1 et date2) forment elles-mêmes des intervalles ayant une durée.

- « *fin du mois de juin 2010* » : [2010−06−20, 2010−06−30]
- « *depuis le milieu des années 60* » : [1965, + ∞[

Il faut souligner que l'exactitude (du reste impossible à établir) dans la transformation d'une représentation symbolique des expressions calendaires vers une représentation discrète (les intervalles calendaires) ne revêt pas une importance majeure, dans la mesure où la recherche repose sur des mesures de similarité : ainsi, savoir si « *le 18 juin 2010* » est inclus ou non dans la période définie par « *fin juin 2010* » n'est pas crucial, puisque que la « distance sémantique » entre les deux sera de toute façon évaluée comme étant faible. Dit autrement, une expression comme « *le 18 juin 2010* » sera évaluée comme étant moins prototypique (ou similaire) que « *le 30 juin 2010* » au regard des expressions attendues pour une requête portant sur la « *fin juin 2010* », elle sera toutefois considérée comme étant plus pertinente qu'une expression comme « *le 10 juin 2010* ».

4.2 Indexation avec Lucene

L'indexation repose sur l'API de Lucene[8] et les analyseurs proposés dans la librairie pour le français et l'anglais. Le paradigme standard pour le traitement des mots-clés (qui exclut les « mots-vides » ou « stop-words ») est donc repris tel quel, dans la mesure où il ne s'agit pas directement de traiter la question des recherches thématiques, mais plutôt d'illustrer les possibilités ouvertes par le traitement des requêtes calendaires.

A ce stade des développements, les objets indexés sont des fragments de documents, en l'occurrence, des phrases, afin de disposer du segment textuel contenant l'expression temporelle qui nous intéresse, mais aussi afin de s'assurer que la distance entre l'expression temporelle et les mots-clés, lorsqu'ils sont retrouvés dans le texte, ne soit pas trop importante. A ces fragments sont associés les termes du paragraphe, l'url de la page Web ou le titre du document et les différentes expressions calendaires présentes. Lorsqu'un fragment de document contient plusieurs expressions calendaires, il est indexé plusieurs fois. Pour une requête thématique, on privilégie ainsi les mots-clés présents dans la phrase, mais on tient compte également de ceux présents dans le paragraphe et l'url ou le titre. Il s'agit là d'une simplification rudimentaire du problème complexe, du point de vue de la linguistique textuelle, de la portée des expressions calendaires, c'est-à-dire de la façon dont elles prennent part à la structuration du discours (Van Reamdonck, 2001 ; Charolles et Vigier, 2005 ; Bilhaut et al., 2003).

Mettre en regard une requête temporelle et des expressions calendaires pour évaluer leur pertinence ne peut pas se faire d'emblée au niveau de l'indexation, car la mesure de la pertinence relative des résultats par rapport à la requête nécessite une comparaison deux à deux des expressions calendaires. Pour éviter des traitements coûteux qui demanderaient de balayer tout l'index pour comparer les informations temporelles stockées et la requête, le parti pris est de réduire les expressions calendaires, lors de l'indexation, à des éléments atomiques (ponctuels) sur l'axe du temps. On réduit ainsi provisoirement les expressions calendaires à une ancre calendaire sans étendue, en ne retenant des expressions que leur « point focal » qui correspond à l'ancre calendaire évaluée comme la plus pertinente. Ce point correspond le plus souvent à la borne gauche de l'intervalle temporel décrit par une expression calendaire, ce qui revient à dire qu'on privilégie l'ordre lié au sens de l'écoulement du temps : ainsi pour une requête du type « *au XX^{e} siècle* », on ordonnera par exemple une série de résultats de la façon suivante « *en 1903* », « *de 1910 à 1912* », « *vers 1920* », notamment parce que la distance des ces expressions par rapport à la borne de gauche de la requête va croissant. Toutefois, pour des expressions formant un intervalle infini à gauche (ex : « *jusqu'en 2007* ») ou une focalisation de type « fin » (ex : « *fin août 1988* »), le point d'ancrage est la borne droite. Ce point correspond en revanche au centre pour les expressions présentant une focalisation de type « milieu » (« *mi-août 1993* »). Ainsi, par exemple, l'expression « *au XVI^{e} siècle* » est réduite pour l'indexation à la date suivante au format ISO-8601, 1500-01-01T00:00:00, qui correspond au début de la borne gauche de l'intervalle [15**, 15**]. Le point d'ancrage de l'expression « *au milieu des années 50* » correspond lui au centroïde de l'intervalle calendaire, à savoir la date suivante : 1955-01-01T00:00:00.

Le processus de comparaison et d'évaluation du score de pertinence n'intervient qu'après filtrage des résultats. Le requêtage de l'index se fait ainsi en deux étapes : (1) recherche dans l'index des K plus proches voisins de la requête (par rapport au point focal de l'intervalle calendaire) ; (2) ordonnancement des

[8] http://lucene.apache.org/

résultats ainsi filtrés en fonction de leur pertinence relativement à la requête. La première étape du requêtage consiste à récupérer dans l'index des documents les K plus proches voisins de la requête, où K correspond au nombre de résultats que l'on souhaite présenter (soit les K résultats a priori les plus intéressants). L'étape suivante consiste à évaluer la pertinence des résultats issus de l'étape de filtrage, qui permet d'obtenir un ordre total des expressions calendaires sous l'angle de la pertinence relative des documents face à une requête.

4.3 Croiser une requête thématique et une requête temporelle

Les requêtes soumises au système par les utilisateurs sont annotées par le même module d'annotation que celui utilisé pour annoter les expressions calendaires dans les documents indexés. L'analyse de la requête sépare un ensemble de mots-clés (la requête thématique) et une requête temporelle, exprimée en langage naturel. Lors du parcours de l'index, les cinquante plus proches voisins de la requête sur l'axe du temps sont rassemblés, selon la méthode décrite précédemment ; cet ensemble ainsi filtré est ensuite ordonné par pertinence.

Pour cette première expérimentation, 7782 articles en français provenant de Wikipedia et relatifs à l'histoire de France ont été annotés et indexés. Le corpus contient près de 180 000 expressions calendaires « absolues ». La figure 6 est une copie d'écran des résultats renvoyés pour la requête suivante « *Jules Ferry à la fin des années 1880* ». Les premiers résultats présentés sont ceux dont l'expression calendaire est considérée comme la plus proche sémantiquement de la requête.

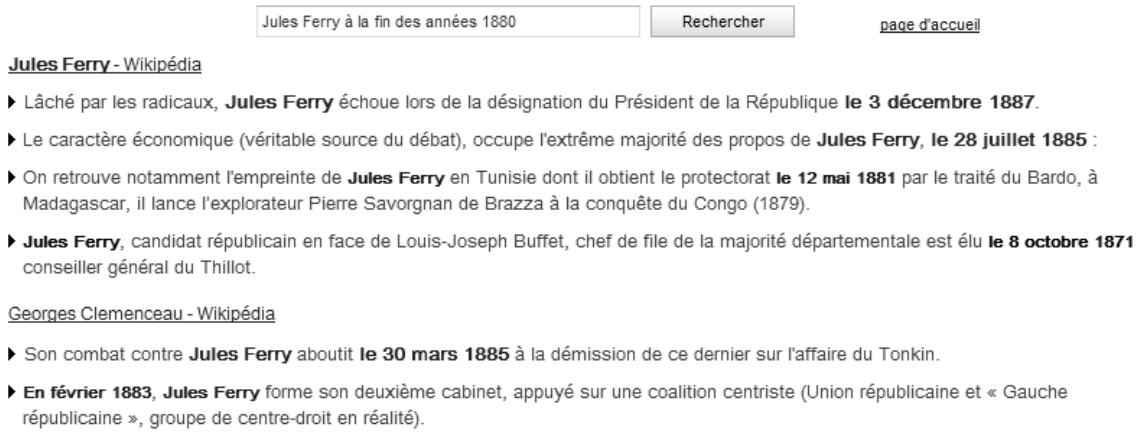


Fig. 6 : copie d'écran de l'outil de navigation temporelle dans un corpus

Les résultats présentés à l'écran (un ensemble de phrases) sont regroupés par textes source et, pour chacun des textes, par pertinence. La figure 6 montre que la sémantique de la requête temporelle est correctement interprétée : en effet, dans les résultats présentés, on trouve en tête de liste des résultats proches de la borne droite de l'intervalle correspondant à l'expression « *à la fin des années 1880* ». Les résultats sont en outre ordonnés du plus prototypique au moins pertinent par rapport à la période définie par la requête.

5 Perspectives

L'expérience montre que la méthodologie proposée pour calculer des scores de pertinence temporelle permet de traiter des requêtes temporelles et de faire remonter des résultats qui tiennent compte des caractéristiques temporelles de la requête. Elle illustre également la façon dont il est possible de croiser une requête thématique et une requête temporelle pour explorer un texte ou un corpus.

Nous commençons à travailler sur la spécification de protocoles pour permettre d'évaluer le système lui-même. Les prochains développements viseront également à proposer un modèle de pertinence pour des documents en entier (des sites Web, plutôt que des phrases), afin de disposer d'un outil de recherche d'informations au sein de corpus étendus sur le Web. En effet, l'algorithme de calcul de la pertinence temporelle peut servir tout aussi bien à la navigation au sein d'un texte ou d'un corpus de textes (l'application que l'on présente dans le cadre de cet article) qu'à la recherche d'informations sur le Web, dans la mesure où les techniques utilisées sont semblables, même s'il faut encore parvenir à établir un

modèle de pertinence pour un document. Ces travaux visent ainsi, à terme, à montrer qu'il est possible de combler le vide qu'il y a aujourd'hui entre la recherche d'informations sur le Web et la consultation d'un document proposé dans les résultats d'une recherche, en permettant à l'utilisateur de fouiller le texte d'un site Web aisément, avant d'y accéder.

Remerciements

Ce projet est partiellement financé par l'ANR (Contint) RMM2.

Références

AHN D., VAN RANTWIJK J., DE RIJKE M. (2007). A cascaded machine learning approach to interpreting temporal expressions. Actes de *NAACL-HLT'07* Rochester, NY, USA, April.

ALLEN J. F. (1983). Maintaining knowledge about temporal intervals. Actes de *ACM 26, no. 11* (November). 832-843.

ALONSO O., GERTZ M., BAEZA-YATES R. (2007). On the value of temporal information in information retrieval. Actes de *ACM SIGIR Forum 41, no. 2* (December). 35-41.

ALONSO O., BERBERICH K., BEDATHUR S., WEIKUM G. (2010). Time-Based Exploration of News Archives. In *HCIR 2010*. New Brunswick. 12-15.

AUNARGUE M., BRAS M., VIEU L., ASHER N. (2001). The syntax and semantics of locating adverbials. *Cahiers de Grammaire*, 26, 11–35.

BATTISTELLI D., CORI M., MINEL J.-L., TEISSEDRE C. (soumis). Querying calendar references in texts. 8 p.

BATTISTELLI D., CORI M., MINEL J.-L., TEISSEDRE C. (2011). Semantics of Calendar Adverbials for Information Retrieval. *ISMIS 2011* (LNCS), Warsaw, June 28-30 2011, 9 p.

BATTISTELLI D. (2011). Linguistique et recherche d'information : la problématique du temps. *Hermès Sciences*, 249 pages, coll. Traitement de l'Information, avril 2011.

BATTISTELLI D., COUTO J., MINEL J.-L., SCHWER, S. (2008). Representing and Visualizing calendar expressions in texts. Actes de *STEP'08*, Venice.

BECHET G., CLERIN-DEBARD F., ENJALBERT P. (2000). A qualitative Model for Time Granularity. *Computational Intelligence*, Vol. 16 (2), 137-175.

BETTINI C., JAJODIA S., WANG S. (2000). Time granularities in Databases, Datamining, and Temporal Reasoning. *Springer* (Eds), 2000, XI, 230 p.

BILHAUT F., HO-DAC L.M., BORILLO A., CHARNOIS T., ENJALBERT P., LE DRAOULEC A., MATHET Y., MIGUET H., PÉRY-WOODLEY M.-P., SARDA L. (2003). Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. Actes de *TALN 2003*, 315-320.

BITTAR A. (2010). Building a TimeBank for French: A Reference Corpus Annotated According to the ISO-TimeML Standard. *Thèse de doctorat*, Université Paris Diderot (Paris 7).

CAILLIAU F., GIRAUDEL A., ARNULPHY B. (2009). Tracking Out-of-date Newspaper Articles. Advances in Computational Linguistics*, Research in Computing Science 41*, 2009, 277-288.

CHAROLLES M., VIGIER D. (2005). Les adverbiaux en position préverbale : portée cadrative et organisation des discours. *Langue Française* vol 2, no. 148, 9-30.

EHRMANN M., HAGÈGE C. (2009). Proposition de caractérisation et de typage des expressions temporelles en contexte. Actes de *TALN'09*, Senlis, France

FERRO L., GERBER L., MANI I., SUNDHEIM B., WILSON G. (2003). TIDES Standard for the Annotation of Temporal Expressions, http://www.mitre.org/work/tech\-papers/tech-papers-04/ferro-tides/.

HAN B., GATES D., LEVIN L. (2006). From language to time: A temporal expression anchorer. Actes *TIME'06*, IEEE Computer Society, June, 196–203.

KLEIN W. (1994). Time in Language. London, *Routledge*.

LE PARC-LACAYRELLE A., GAIO M., SALLABERRY C. (2007). La composante temps dans l'information géographique textuelle, *Document numérique* 2/2007 (Vol. 10), 129-148.

LLORENS H., SAQUETE E., NAVARRO B. (2010). TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. Actes de *5th International Workshop on Semantic Evaluation, ACL 2010*, 284-291.

MANI I., WILSON G. (2000). Robust temporal processing of news. Actes de 38th ACL, 69–76.

MATTHEWS M., TOLCHINSKY P., MIKA P., BLANCO R., ZARAGOZA, H. (2010). Searching through time in the New York Times Categories and Subject Descriptors. Actes de *HCIR 2010* - Challenge Report, 41-44.

MAZUR P., DALE R. (2008). What's the Date? High Accuracy Interpretation of Weekday Names. Actes de *22nd International Conference on Computational Linguistics*. Manchester. 553–560.

MESTL T., CERRATO O., ØLNES J. , MYRSETH P., GUSTAVSEN I.-M. (2009). Time Challenges – Challenging Times for Future Information Search. *D-Lib Magazine*, Volume 15 Number 5/6, May/June 2009.

PAUMIER S. (2002). Manuel d'utilisation du logiciel Unitex. IGM, Université de Marne-La-Vallée.

PUSTEJOVSKY J., CASTANO J., INGRIA R., SAURÍ R., GAIZAUSKAS R., SETZER A., KATZ G. (2003). TimeML: Robust specification of event and temporal expressions in text. Actes de *IWCS-5, Fifth International Workshop on Computational Semantics*.

SAQUETE E., P. MARTÍNEZ-BARCO, R. MUÑOZ, J.L. VICEDO (2004). Splitting Complex Temporal Questions for Question Answering systems. Association for Computational Linguistics (ACL) Barcelona, SPAIN. July 2004

SCHILDER F., HABEL C. (2001). From Temporal Expres-sions to Temporal Information: Semantic Tagging of News Messages. Actes de *ACL'01, Workshop on temporal and spatial information processing*, 65 -72.

TEISSEDRE C., BATTISTELLI D., MINEL J.-L. (2010). Resources for Calendar Expressions Semantic Tagging and Temporal Navigation through Texts, Actes de *LREC 2010*, 3572-3577.

TEISSEDRE C., BATTISTELLI D., MINEL J.-L. (2010). Du texte au portail sémantique : cas d'utilisation lié à des données temporelles. In Actes d'*IC'2010*, 209-220.

UZZAMAN N., ALLEN J. F. (2010). TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text. Actes du *5th International Workshop on Semantic Evaluation, ACL 2010*. 276-283.

VAN RAEMDONCK D. (2001). Est-il pertinent de parler d'une classe d'adverbes de temps ? Actes de *CLAC* 7.

WANG R., ZHANG Y. (2008). Recognizing Textual Entailment with Temporal Expressions in Natural Language Texts. Actes de *2008 IEEE International Workshop on Semantic Computing and Applications* (IWSCA '08).

Extraction de patrons sémantiques appliquée à la classification d'Entités Nommées

Ismaïl El Maarouf (1,2) Jeanne Villaneau (2) Sophie Rosset (3)

(1) HCTI UBS-UEB, Centre de Recherche Christiaan Huygens, 56321 Lorient
(2) Valoria UBS-UEB, Rue Yves Mainguy, Campus de Tohannic 56017 Vannes cedex
(3) LIMSI-CNRS, F-91403 Orsay Cedex
ismail.el-maarouf@univ-ubs.fr, jeanne.villaneau@univ-ubs.fr, sophie.rosset@limsi.fr

Résumé **La variabilité des corpus constitue un problème majeur pour les systèmes de reconnaissance d'entités nommées. L'une des pistes possibles pour y remédier est l'utilisation d'approches linguistiques pour les adapter à de nouveaux contextes : la construction de patrons sémantiques peut permettre de désambiguïser les entités nommées en structurant leur environnement syntaxico-sémantique. Cet article présente une première réalisation sur un corpus de presse d'un système de correction. Après une étape de segmentation sur des critères discursifs de surface, le système extrait et pondère les patrons liés à une classe d'entité nommée fournie par un analyseur. Malgré des modèles encore relativement élémentaires, les résultats obtenus sont encourageants et montrent la nécessité d'un traitement plus approfondi de la classe Organisation.**

Abstract **Corpus variation is a major problem for named entity recognition systems. One possible direction to tackle this problem involves using linguistic approaches to adapt them to unseen contexts : building semantic patterns may help for their disambiguation by structuring their syntactic and semantic environment. This article presents a preliminary implementation on a press corpus of a correction system. After a segmentation step based on surface discourse clues, the system extracts and weights the patterns linked to a named entity class provided by an analyzer. Despite relatively elementary models, the results obtained are promising and point on the necessary treatment of the Organisation class.**

Mots-clés : entités nommées, patrons sémantiques, segmentation discursive de surface
Keywords: named entities, semantic patterns, surface discourse segmentation

1 Introduction

Ritel est un Système de Question-Réponse interactif à domaine ouvert (Rosset et al., 2008), permettant à un utilisateur de dialoguer et d'obtenir des réponses à ses questions. Parmi les questions auxquelles doit pouvoir répondre un tel système, certaines (factuelles, définitions) correspondent à une demande d'information sur une Entité Nommée (EN), une valeur ou une date. L'analyse détaillée des composants d'un SQR (Moldovan et al., 2003) montre que la prédiction du type de réponse dans les documents indexés est l'une des causes d'erreurs majeure, en partie due à la Reconnaissances de ces Entités Nommées (REN).

Si la REN semble être une tâche bien maitrisée lorsque l'on se réfère aux résultats des campagnes d'évaluation classiques (Grishman et al., 1995), on connaît moins leurs performances sur des corpus différents de ceux pour lesquels ils ont été développés (Grishman, 2010). La robustesse des systèmes peut être mise à l'épreuve par le degré de granularité de la typologie d'EN (Galliano et al., 2009 ; Markert et al., 2007), la qualité de retranscription du corpus (Galliano et al., 2009) et l'hétérogénéité des corpus (Mota et al. 2008). Les dégradations de performances constatées dans ces trois cas justifie la conception de systèmes de correction qui assurent leur robustesse face à des textes et à des EN inconnus.

C'est dans la perspective d'adaptation d'un analyseur linguistique intégrant la détection d'EN (Ritel-nca) utilisé par le SQR Ritel, que s'inscrit notre recherche. Nous présentons un système destiné à corriger les résultats de l'analyseur, à partir de patrons sémantiques extraits de corpus écrits. Après avoir décrit notre approche et les traitements effectués (section 2), les modèles de construction de patron sont présentés en section 3 puis évalués en section 4. Les premiers résultats obtenus ouvrent la voie à des perspectives de recherche explicitées section 5.

2 Approche et traitements

2.1 Choix de l'approche

Les systèmes de REN sont principalement divisés en deux catégories : d'une part, les systèmes à base de règles, correspondant à des automates qui utilisent des listes d'entités nommées (« gazetteers »), des mots déclencheurs et des indices de surface, et d'autre part, les modèles probabilistes utilisant différents niveaux de représentation (forme, lemme, catégorie morphosyntaxique) entrainés sur des corpus annotés (pour un état de l'art, voir Nadeau et al., 2007). Notre système s'appuie sur un corpus annoté automatiquement par un système de REN. Les EN sur lesquelles nous travaillons correspondent à la triade Personne-Lieu-Organisation car elles cristallisent de nombreux problèmes : la possibilité pour une forme de recouvrir les trois catégories rend une analyse fine du contexte indispensable (voir également 4.1).

L'approche développée dans cet article s'inscrit dans une perspective d'enrichissement linguistique de systèmes à base de règles, en créant des grammaires de patrons sémantiques sur le modèle de celles qui sont employées en Extraction d'Information (Ward et al., 1992). Elle s'inspire de la linguistique de corpus britannique (Sinclair, 1991), dont le but est de décrire les usages des mots en contexte à partir de l'exploration de grands corpus. Le sens d'un mot est défini en fonction des patrons majeurs dans lesquels il est employé. La notion de patron peut recouvrir différentes réalités, des collocations, associations significatives entre lexies (Sinclair, 1991), jusqu'à l'identification de cadres complexes, proches des structures prédicatives (Hanks, 2008).

Les associations sont automatiquement calculées sur de grands corpus et permettent de créer des réseaux sémantiques pour chaque mot. La méthode générale consiste à sélectionner les mots apparaissant de manière significative dans une fenêtre de taille arbitraire autour du mot-clé. Dans notre système, les unités du réseau correspondent aux informations syntaxiques et sémantiques contenues dans des chunks (décrits en 2.2). La fenêtre, quant à elle, peut être de taille variable, car elle correspond à la notion de segment, défini à partir d'indices discursifs de surface tels que les marques de ponctuation et les formes en « qu- » (cf. 2.3).

2.2 Chunking grammatical

Ritel utilise les résultats d'un analyseur linguistique à base de règles, Ritel-nca, dont les sorties sont présentées en structure arborescente (figure 1). Ce système a fait l'objet d'un développement particulièrement approfondi : il permet l'accès à des lexiques catégorisant plus d'un million de mots, dont une grande partie de noms propres, et près de 2000 règles sont actuellement implémentées. La taxonomie utilisée et continuellement augmentée, comprend plus de 300 types, dont les EN classiques (*Personne*, *Organisation* et *Lieu*), affinés et structurés en sous-types et en composants. La f-mesure associée à la classification d'entités classiques est de 0,8 sur l'écrit et à hauteur de l'état de l'art pour les corpus oraux (Rosset et al., 2008).

Un système comme Ritel-nca facilite l'analyse linguistique : la détection des entités classiques, comme les dates, permet de regrouper de nombreuses variantes, faisant ainsi émerger des associations nouvelles et sémantiquement pertinentes. Les mots grammaticaux (déterminants, prépositions) sont cependant rarement rattachés, ce qui nuit à la construction de patrons. Pour réduire les phrases à des groupes plus homogènes, nous avons intégré les entités dans des chunks grammaticaux (dont la version initiale est décrite dans Villaneau et al. 2007).

Les chunks grammaticaux regroupent les nœuds de l'arbre de la phrase à partir de règles exploitant des indices de formes, de type d'entité et de position. Cinq types de chunks sont définis parmi les groupes nominaux (GN, GNP) et les groupes verbaux (GV, GVP, GVADJ). Les chunks sont associés à la catégorie majeure : un verbe dans le cas des groupes verbaux, un lieu dans le cas d'un GNP regroupant les entités [préposition déterminant lieu]. La figure 1 illustre la représentation arborescente de l'exemple (1) après la passe de chunking.

(1) Patricia Highsmith est morte le 4 février 1995 dans un hôpital de Locarno

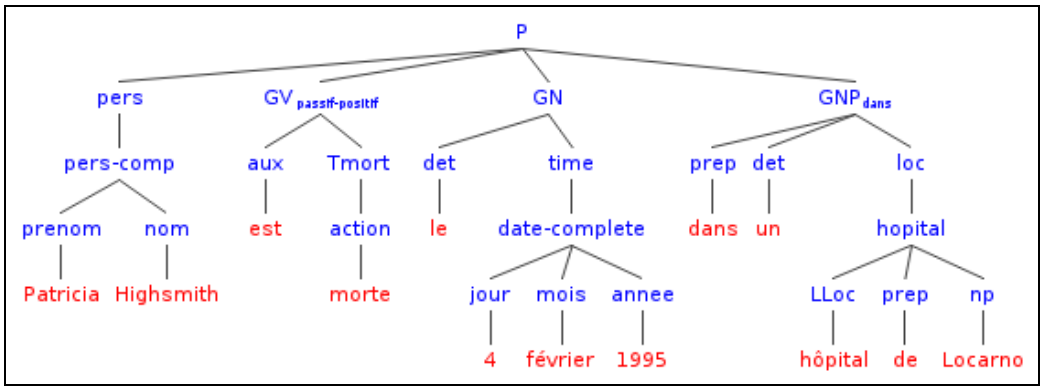


Figure 1- Représentation arborescente après chunking de l'exemple 1.

Dans cet exemple, le nombre de nœuds de l'arbre de la phrase est divisé par deux sans perte d'information : la catégorie et la forme de la tête sont conservées et il est toujours possible de parcourir les fils d'un nœud pour obtenir des informations plus détaillées. Les nœuds de chunks grammaticaux possèdent en sus des attributs concernant la voix et le mode du verbe (pronominal, passif, actif) permettant de distinguer rapidement les différentes réalisations verbales.

Nous avons manuellement cherché à extraire des relations sémantiques en projetant des cadres sémantiques du type de FrameNet (Fillmore et al., 2003) à partir de grammaires régulières de chunks. Si les résultats obtenus sont précis, l'analyse se heurte à des limites qui compliquent la création de règles : rigidité due à l'ordre de détection des éléments et sensibilité à la variation de surface, due principalement à la présence de ponctuation. Plutôt que de multiplier les règles, nous avons intégré une étape de segmentation de surface qui permet d'isoler des séquences de chunks.

2.3 Segmentation de surface

La segmentation consiste à isoler des séquences de chunks en fonction d'indices de frontière surfaciques comme la ponctuation : elle est principalement applicable à des textes écrits, quoique les travaux récents en insertion automatique de ponctuation (Favre et al., 2008) ouvrent des perspectives d'application à l'oral. Les indices de ponctuation jouent un rôle structurant dans la présentation de l'information. Leur prise en compte peut permettre la délimitation de structures appositives, comme en (2), citationnelles en (3) et parenthétiques en (4).

(2) le secrétaire d' Etat aux PME , Renaud Dutreil , s' était ainsi vu convoquer par le directeur du cabinet du premier ministre , Michel Boyon.

(3) "le navire doit être capable de prendre un mauvais coup de chien sans risque", explique Guy Ribadeau-Dumas, l' architecte naval maître d' œuvre du chantier .

(4) Eric Tanguy (né en 1968) passait alors pour un nouveau Dusapin.

Ces structures imposent des contraintes sur leur environnement : dans l'exemple (3), la citation entre guillemets précède un verbe dont le sujet est inversé. On observe également en (4) que les parenthétiques isolent un groupe d'information semi-autonome qui contient une relation sémantique (date de naissance).

Les segments que nous constituons se définissent par leurs frontières gauche et droite et leur taille. Pour première expérience, nous nous limitons à une typologie simple de frontières en distinguant les frontières fortes (point, point d'exclamation, etc.) des frontières faibles (virgules, parenthèses, etc.). Seules les frontières faibles sont franchies pour identifier des relations sémantiques (virgule dans l'exemple 2). Nous avons également considéré les connecteurs et conjonctions comme des frontières faibles pour obtenir des segments proches de propositions. L'algorithme de segmentation applique d'abord les règles associées à la détection de frontières fortes, puis, dans une seconde passe et au sein des unités déjà segmentées, celles associées aux frontières faibles (automates adaptés aux sorties du chunking). Le tableau 1 résume les fréquences des segments obtenus avec cette méthode de segmentation sur un corpus de presse (4,5 millions de mots).

Taille de Segment	Fréquence	Proportion
1	207260	0.29
2	160177	0.22
3	110760	0.15
4	78605	0.11
5	53846	0.07
6	35657	0.05
7	23103	0.03
8	14686	0.02
9	9282	0.01
10	5783	<0.01
>10	*29951*	*0.04*

Tableau 1 – Fréquence des segments en fonction de leur taille (exprimée en nombre de chunks).

En focalisant sur les relations contenues dans les segments de petite taille, on peut ainsi traiter la majorité des segments, sachant que 30% des segments (de taille 1) ne renferment pas de relation. La détection de relations sémantiques peut s'effectuer de deux manières : en analysant les associations entre chunks au sein de segments de taille supérieure à 1, et en analysant les relations entre chunks appartenant à des segments différents. Les segments de taille 1 sont donc réservés à l'analyse entre segments. L'exemple (5) montre le type de relations que l'on peut détecter au sein des segments dits « pluriels » (car contenant plus d'un chunk) : dans l'exemple, ces derniers figurent en rouge, les segments simples, en violet.

(5) A. V. Shinde, né à Goa, en Inde, décédé en 2003 à New York (à l' âge de 86 ans), avait parcouru le monde en quête des plus belles pierres pour le joaillier Harry Winston.

Nous nous concentrons dans cet article sur les relations existant entre chunks au sein des segments pluriels. Il s'agit dans l'exemple (5), d'analyser les relations entre « né » et « à Goa », ou encore entre « décédé », « en 2003 » et « à New York ». Plus précisément, nous utilisons ces chunks comme indices de désambiguïsation des EN figurant dans les mêmes segments. Le système permettant d'extraire les patrons d'association [chunk-EN] est décrit en section suivante.

3 Système et modèles

3.1 Un système multi-niveaux

Le système analyse l'arbre obtenu après la passe de chunking et y ajoute les segments. Quatre classes de règles dédiées à chaque niveau de représentation sont définies : *Forme*, *Entité*, *Chunk Grammatical* et *Segment*. Chaque niveau de représentation possède son propre lexique de contraintes (attribut-valeur), permettant d'identifier un élément ou un segment. Par exemple, un groupe prépositionnel en « à » sera défini par sa classe (Chunk Grammatical), son sous-type (Groupe Nominal Prépositionnel) et la forme de la préposition. Les règles et les lexiques sont externalisés afin de pouvoir appliquer des patrons définis manuellement ou extraits automatiquement.

Les patrons extraits au sein des segments peuvent s'appuyer sur les chunks, les entités ou les formes. À titre d'illustration, la figure 3 décline les caractéristiques internes de chaque segment de l'exemple (6) en fonction du niveau de représentation.

(6) il y a près de cinquante ans déjà , Jacques Monod rappelait au colloque de Caen que 50 pourcent du chiffre d' affaires de la société américaine Du Pont de Nemours provenait de la commercialisation de produits inconnus dix ans plus tôt .

Segment 1

Chunk	-	-	-
Entité	<_pres>	<_annee_dur>	<_adv>
Forme	il y a	50 ans	déjà

Segment 2

Chunk	-	-	GNP_au	GNP_de
Entité	<_pers>	<_action>	<_subs>	<_loc>
Forme	Jacques Monod	rappelait	colloque	Caen

Segment 3

Chunk	GN	GNP_de	-	GNP_de	-	GNP_de	GNP_de	-
Entité	<_subsnn>	<_org>	<_pers>	<_loc>	<_action>	<_subs>	<_subs>	<_time>
Forme	chiffre d'affaires	société américaine	Du Pont	Nemours	provenait	commercialisation	produits	dix ans plus tôt

Figure 3 – Tableaux des niveaux de représentation des segments de l'exemple (6).

Cette représentation nivelée de la phrase permet de concevoir des modèles qui combinent les informations de plusieurs niveaux. Trois modèles sont évalués : la combinaison des niveaux Chunk et Forme (modèle CF), des niveaux Chunk et Entité (modèle CE) et un modèle mixte (modèle CEM) qui combine les niveaux Chunk et Entité, en substituant les entités « substantif », « action » et « adjectif » par les formes correspondantes, les verbes étant lemmatisés. L'existence de ce dernier modèle est motivée par l'hypothèse que ces classes contiennent régulièrement des informations sémantiques pertinentes qui seraient autrement masquées. Le tableau 2 fait figurer les éléments extraits dans le segment 2 en fonction de chaque modèle.

	E1	*E2*	*E3*	*E4*
CF	**Jacques Monod**	**rappelait**	**GNP_au/colloque**	**GNP_de/Caen**
CE	**pers**	**action**	**GNP_au/subs**	**GNP_de/loc**
CEM	**pers**	**rappeler**	**GNP_au_colloque**	**GNP_de/loc**

Tableau 2 - Exemples de patrons extraits du segment 2 en fonction des modèles.

Alors que le modèle CEM cherche à optimiser les informations détenues par chaque élément, le modèle CE est le plus générique. Quant au modèle CF, il peut être plus précis en cas d'erreurs d'analyse des entités.

3.2 Méthode d'extraction et score

Le corpus dont nous disposons correspond à une année du Journal LeMonde de l'année 2003. Il est divisé en deux parties, un corpus de développement, à partir duquel nous avons extrait les patrons (17 millions de mots), un quart (5,5 millions de mots, 10 000 articles) ayant été conservé dans la perspective de l'annoter manuellement (cf. 4.1). Pour chaque modèle, nous avons sélectionné les segments contenant une des entités classiques (« personne », « organisation » ou « lieu ») fournies par l'analyseur Ritel-nca, et puisque notre étude porte sur les patrons trouvés à l'intérieur d'un segment, exclu les segments de taille 1.

Le système évalué s'appuie sur les patrons intra-segment observés dans le corpus de développement. Un patron correspond à un chunk identifié dans un segment contenant une EN, modélisé selon un niveau de représentation. Par exemple, le patron « GNP_de/_loc » du modèle CEM apparaît 12829 fois en corpus, 6411 fois en coocurrence avec un Lieu, 2604 fois avec une Organisation et 3814 avec une Personne. À partir de ces données, nous calculons deux scores d'association d'un patron pour chaque classe : la probabilité de cooccurrence entre un chunk et une classe d'EN donnée (PROBA) et l'information mutuelle (IM) :

$$PROBA(EN \mid Patron) = \frac{P(EN, Patron)}{P(Patron)}$$

$$IM(EN,\ Patron) = P(X=EN, Y=Patron) \times \log \frac{P(X=EN, Y=Patron)}{P(X=EN) \times P(Y=Patron)}$$

Ces scores nous permettent de prédire la classe d'EN la plus probable vis-à-vis d'un patron donné. Pour l'évaluation, nous avons calculé trois scores globaux pour prendre en compte l'ensemble des patrons contenus dans le segment : la moyenne des scores des patrons (Mean) pour tenir compte du nombre de patrons dans le segment, le score du meilleur patron (Max) et le produit des scores, pour atténuer l'importance de patrons fréquents et communs aux différentes classes. Pour exemple, le score PP (Produit de Probabilités) d'un segment pour la classe Personne, se calcule à partir du produit des probabilités de ses patrons :

$$PP(Personne) = \prod_{1}^{n} PROBA(Personne \mid Patron_i)$$

Six scores différents ont ainsi été expérimentés pour chaque type de modèle de représentation vis-à-vis de chacune des trois classes, Personne, Lieu et Organisation.

4 Évaluation

L'évaluation présentée dans cette section compare les performances des modèles CF, CE et CEM, combinés avec les scores PROBA et IM présentés précédemment. Leur performance de classification est comparée à celle du système Ritel-nca qui sert de référence.

4.1 Un corpus d'évaluation établi sur des critères contextuels

200 articles de presse ont été annotés pour obtenir plus de mille instances d'entités de chaque classe (plus exactement 1426 organisations, 1004 lieux et 1377 personnes). Les segments de taille 1 étaient pré-détectés et exclus de l'annotation. Les conventions d'annotation ont réduit la tâche de détection des EN au nom propre (avec ou sans majuscule) lorsque c'était possible, en excluant les titres, fonctions,

déterminants. Lorsque certains éléments (parfois même des entités nommées dans des cas d'imbrication) pouvaient être considérés comme constitutifs du nom d'une EN, ils ont cependant été inclus (exemple 7). En revanche, seule l'EN était prise en compte lorsque sa dénomination ne dépendait pas d'éléments englobants. Ces conventions distinguent ainsi les cas (7) et (8).

(7) l'<org> université de Poitiers </org>

(8) le maire de <loc> Poitiers </loc>

La difficulté majeure consiste à identifier des critères fiables pour résoudre les cas où le type d'une EN diffère de son rôle en contexte. Les conventions d'annotation privilégient dans ces cas l'interprétation contextuelle. Deux types de divergences ont été rencontrées : lorsque cette divergence était explicitée par un déclencheur immédiatement apposé (9) et lorsqu'elle était due à une interprétation globale de la phrase voire du contexte de l'article (10)

(9) c' est la mesure phare de la loi Perben du 9 mars sur la criminalité

(10) l' Italie s' oppose à une réforme du Conseil de sécurité de l' ONU

En (9), l'EN « Perben » a été exclue de l'annotation car elle relève du type « Loi », bien qu'elle soit nommée d'après son fondateur. L'exemple (10) est généralement décrit comme un cas de métonymie (Markert et al., 2007), pour rendre compte de la relation existant entre un lieu (« Italie ») et des individus, l'interprétation étant due au verbe avec lequel elle est employée. Cet exemple ne désigne pas une personne comme les conventions d'annotation de métonymie de la campagne Semeval7 semble l'indiquer à travers la catégorie « Loc-for-People » (Markert et al., 2007) : il s'agit d'une organisation politique, dans ce cas très probablement le gouvernement. D'autres types d'organisations répondent à ce phénomène, comme les équipes de sport (11).

(11) dans les autres rencontres disputées mercredi soir, <org>Auxerre</org> s' est imposé à <loc>Rennes</loc>

Les conventions considèrent ainsi que l'EN « Italie » peut désigner un lieu ou une organisation, comme l'EN « Florence », une personne (12) ou un lieu (13).

(12) ce n' est pas le moindre des mérites de l' essai d' Anton Brender et Florence Pisani

(13) une forte pluie commença à tomber sur la Toscane et Florence

Les lieux ont donc été annotés comme tels lorsque l'interprétation en contexte le justifiait (localisation, destination, origine, etc.), comme en (13).

4.2 Détection et classification brute des EN

Les résultats de l'évaluation du système confirment ce qui a été dit précédemment à propos de l'impact du corpus de développement ainsi que des conventions d'annotations adoptées. 980 des 3807 EN annotées n'ont pas été détectées par le système Ritel-nca, soit 25%, parmi lesquelles 515 sont étiquetées comme noms propres non catégorisés. Les « erreurs » de détection affectent principalement la catégorie *Organisation* et s'expliquent pour les raisons suivantes : EN non retenues, mots inconnus, problèmes de normalisation du texte (suppression de majuscules, encodage), etc.

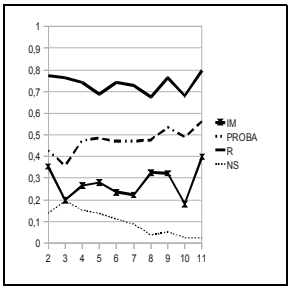
Classe	Correct	Faux Positif	Raté	PRECISION	RAPPEL	FMESURE
PERSONNE	1087	333	290	0,77	0,79	0,78
LIEU	686	508	318	0,57	0,68	0,62
ORGANISATION	523	360	903	0,59	0,37	0,45

Tableau 3 – Rappel, Précision et F-mesure de classification du système Ritel-nca.

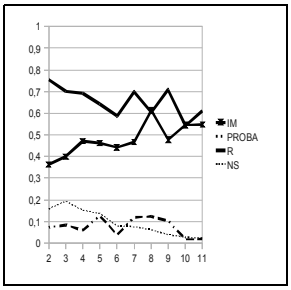
Étant donné que les patrons générés classent les EN à partir des entités fournies par le système Ritel-nca, l'évaluation a uniquement porté sur les EN détectées. Les résultats de ce dernier ont ainsi été recalculés et figurent dans le tableau 4 (modèle R). Le nombre de segments total s'élève à 1712, réduisant le nombre de segments contenant au moins une personne détectée à 943 (les lieux à 818 et les organisations à 659), 72% d'entre eux ne contenant qu'une seule entité.

4.3 Résultats

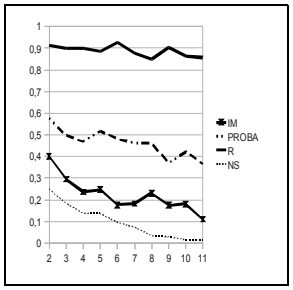
Les diagrammes 1 à 3 présentent les f-mesures des modèles en fonction de la taille des segments pour chaque classe d'EN ; le nombre de segments par taille figure également sur les diagrammes (NS), ainsi que les résultats de Ritel-nca (R), à titre comparatif. Par degré d'importance, le score d'association (IM, PROBA) est la variable qui influence le plus les résultats, suivi par le niveau de représentation (CE, CF, CEM). Quand au calcul global du score (MAX, PROD, MEAN), il n'a qu'une faible influence : le choix du meilleur score d'association (MAX) équivaut globalement à calculer la moyenne ou le produit des scores de tous les patrons. Les diagrammes font uniquement figurer les moyennes des scores en fonction de la mesure d'association (PROBA, IM).



Lieu



Organisation



Personne

Diagramme 1 à 3 : F-mesure pour les Lieux, les Organisations et les Personnes.

Comme on peut l'observer, ces modèles ne rivalisent pas avec le modèle de référence (R). Les meilleurs modèles atteignent 0,62 de F-mesure sur les Personnes, 0,55 pour les Lieux et 0,45 sur les Organisations. On peut retenir globalement que le score PROBA est plus approprié pour la classification des Lieux et des Personnes, alors que l'IM semble plus performante sur les Organisations. L'augmentation de la taille du segment semble avoir un impact négatif sur la classification des Personnes, mais elle est liée à une amélioration des modèles PROBA pour les Lieux et des modèles IM pour les Organisations.

Ces résultats nous permettent de connaître le comportement global des modèles mais ne nous renseignent pas directement sur leur utilité dans le cadre de la correction. Pour expérimenter la tâche de correction, motivés par le fait que la mesure de score global avait une influence minime sur les résultats, nous avons sélectionné tous les modèles MAX, qui, pour chaque EN, nous permet d'extraire un patron (un chunk du segment). Par exemple, le modèle CEM_IM_MAX a classé correctement 17 instances de personnes grâce au patron "expliquer", comme dans l'exemple (14) :

(14) " mon rôle est de bousculer la perception que les gens ont de Burberry ", *explique* Christopher Bailey

L'exemple (15) est un cas d'erreur que ce patron permet de corriger : « Maud » est classé en Lieu par le système de référence.

(15) une sorte de tri sélectif qui " élimine les cellules mortes et rend la peau douce et satinée ", explique *Maud*

En nous basant sur les résultats de l'évaluation, on peut alors assigner un taux de réussite à chaque patron. En ne sélectionnant que les patrons dont le score est sans appel (100%), il est alors possible de corriger les erreurs du modèle de référence, et, ce faisant, de juger de la pertinence linguistique des patrons correcteurs.

Les résultats présentés dans le tableau 4 indiquent les performances obtenues lorsque le système de référence détecte correctement une EN (R) et lorsque les patrons de correction sont appliqués. Les résultats sont organisés en fonction de l'ajout de patrons issus d'un des trois niveaux de représentation, du score d'association, ou tous modèles réunis. Lorsqu'aucun patron n'est identifié pour une instance d'EN donnée, le choix se porte sur la catégorie choisie par le système de référence (R).

Catégorie	Modèle	Précision	Rappel	F-mesure	# Corrigés
LIEU	R+TOUS	0,767	0,945	0,847	91
	R+CF	0,754	0,930	0,833	78
	R+PROBA	0,726	0,939	0,819	86
	R+CE	0,727	0,910	0,808	61
	R+IM	0,739	0,886	0,806	41
	R+CEM	0,724	0,899	0,802	52
	R	0,670	0,836	0,744	*NA*
ORGANISATION	R+TOUS	0,952	0,686	0,797	121
	R+CF	0,941	0,672	0,784	107
	R+IM	0,917	0,666	0,772	101
	R+CE	0,925	0,624	0,745	61
	R+CEM	0,918	0,623	0,742	59
	R+PROBA	0,940	0,606	0,737	45
	R	0,866	0,561	0,681	*NA*
PERSONNE	R+TOUS	0,924	0,978	0,950	31
	R+CF	0,916	0,975	0,944	27
	R+PROBA	0,909	0,977	0,942	30
	R+CE	0,898	0,970	0,932	21
	R+IM	0,897	0,967	0,930	18
	R+CEM	0,894	0,969	0,930	20
	R	0,861	0,951	0,904	*NA*

Tableau 4 - Potentiel de correction du système de référence.

Globalement, la prise en compte de tous les patrons de correction permet d'améliorer les f-mesures de 5% pour les Personnes, et de 10% pour les Lieux et les Organisations. Le niveau de représentation qui permet de corriger le plus d'instances est le niveau CF, en partie du fait qu'il génère un plus grand nombre de patrons, multipliant ainsi les possibilités de désambiguïsation. Le taux de correction par modèle ne dépend pourtant pas uniquement du nombre de patrons extraits : les niveaux CE et CEM ont un taux de correction relativement équivalent alors que le niveau CEM génère un plus grand nombre de patron. Ceci s'explique simplement par le fait qu'une large part des patrons corrects de ce dernier vient confirmer le modèle de référence. La combinaison de tous les modèles permet de corriger 66% d'erreurs pour les Lieux, 55% pour les Personnes et 28% pour les Organisations.

Ces résultats semblent corroborer le lien qui peut exister entre la performance d'un modèle et sa capacité de correction : les modèles ayant permis de corriger un grand nombre d'organisations sont basés sur le score IM, alors que le score PROBA contribue à générer plus de patrons de correction pour les lieux et les personnes. Nous avions constaté des tendances similaires sur les diagrammes 1 à 3.

Travaux similaires

Les systèmes obtenant les meilleurs performances sur la REN en français comme dans d'autres langues s'appuient généralement sur une classification supervisée qui nécessite l'établissement d'un corpus d'entrainement annoté manuellement. Notre système est entrainé sur un corpus automatiquement annoté par un système de REN, qui, par conséquent, comporte nécessairement des erreurs. Dans ce cadre, Petasis et al. (2001) ont initié un travail proche de nos objectifs : leur système est basé sur un corpus

automatiquement annoté par un premier système et les points de désaccord sont considérés comme des indices de défaillance. Leur système ne leur permet cependant pas d'extraire des patrons pour envisager une correction automatique : les erreurs sont manuellement corrigées par un expert. Plus généralement, l'inconvénient des systèmes d'apprentissage automatique est leur manque de transparence sur le lien entre les indices contextuels et la décision de classification. Les travaux rapportent au mieux l'impact de classes d'indices (« feature sets » en anglais) sur les performances : capitalisation, taille de fenêtre, prise en compte d'information syntaxique ou de ressources externes, etc. Notre méthode permet de juger directement de la pertinence d'un patron sur lequel s'appuie la décision de classification et d'en induire des règles de correction.

Conclusion et Perspectives

Cet article présente un système de correction d'EN à partir de patrons sémantiques. L'extraction s'appuie sur l'annotation de l'analyseur linguistique Ritel-nca, une phase de chunking et une segmentation de surface. Plusieurs modèles de patrons combinant différentes dimensions sont évalués : mesure d'association, score global et niveau de représentation. L'évaluation de ce système montre que la mesure d'association a une forte influence sur les performances, même si ces dernières sont en-deça de celles du système de référence. Nous mesurons le potentiel de correction du système de référence par ces modèles et obtenons des améliorations de 10% en F-mesure pour les Organisations et les Lieux et de 5% pour les Personnes. Ces améliorations nous encouragent à tester cette approche sur d'autres classes d'EN.

Les modèles employés dans l'évaluation sont relativement élémentaires : dans les travaux à venir, nous évaluerons l'apport de patrons conçus à partir des probabilités conjointes des éléments d'un segment, en commençant par exemple par la prise en compte du chunk contenant l'EN. Des modèles qui prennent en compte l'ordre (en établissant des contraintes de position droite ou gauche par exemple) méritent également d'être testés. Ces pistes seront évaluées dans le cadre d'analyses intra-segment et inter-segment telles que décrites dans cet article. Le problème majeur de notre approche que nous ne pouvons qu'évoquer ici, réside dans la sélection des patrons de correction parmi la totalité des patrons générés par chaque modèle. L'intervention humaine semble indispensable pour permettre d'y remédier mais l'utilisation de méthodes de filtrage automatique n'est pas exclue.

Remerciements

Ce travail a été partiellement réalisé dans le cadre du programme QUAERO (financement OSEO).

Références

EL MAAROUF I., VILLANEAU J., SAID F., DUHAUT D. (2009). Comparing Child and Adult Language: Exploring Semantic constraints. Actes de *WOCCI ICMI-MLMI 2009*.

EL MAAROUF I. (2009). Natural Ontologies at Work : Investigating Fairy Tales. Actes de *Corpus Linguistics Conference 2009.*

ERHMANN M. (2008). *Les Entités Nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. Thèse de Doctorat en sciences du langage, Université Paris 7.

Favre B., Grishman R., Hillard D., Ji H., Hakkani-Tür D. & Ostendorf M. (2008). Punctuating speech for information extraction. Actes de *ICASSP 2008* : 5013-5016.

FILLMORE, C. J., JOHNSON C.R., PETRUCK M.R.L. (2003). Background to Framenet. *International Journal of Lexicography*, (16.3) : 235-250.

GALIBERT O. (2009). *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. Thèse de doctorat en informatique, Université Paris-Sud 11, Orsay.

GALIBERT O., QUINTARD L., ROSSET S., ZWEIGENBAUM P., NÉDELLEC C., AUBIN S., GILLARD L., RAYSZ J.-P., POIS D., TANNIER X., DELÉGER L., LAURENT D. (2010). Named and specific entity detection in varied data: The Quaero Named Entity baseline evaluation. Actes de *LREC'10.*

GALLIANO S., GRAVIER G., CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of French radio broadcasts. Actes de *INTERSPEECH-2009*. 2583-2586.

GRISHMAN R. & SUNDHEIM, B. (1995). Design of the MUC-6 evaluation. Actes de *MUC6*.

GRISHMAN R. (2010). The Impact of Task and Corpus on Event Extraction Systems. Actes de *LREC'10*.

HANKS P. **(**2008). Lexical Patterns: From Hornby to Hunston and Beyond. *Actes d'Euralex* 2008.

MARKERT K., NISSIM M. (2007) SemEval-2007 task 08: metonymy resolution at SemEval-2007. Actes de the 4th International Workshop on Semantic Evaluations : 36-41.

MOLDOVAN D., PASCA M., HARABAGIU S., SURDEANU M. (2003). Performance Issues and Error Analysis in an Open-Domain Question Answering System. *ACM Transactions in Information Systems*.

MOTA C., GRISHMAN R. (2008). Is this NE tagger getting old. Actes de *LREC'2008*.

NADEAU D., SEKINE S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1) : 3–26.

PETASIS G., VICHOT F., WOLINSKI F., PALIOURAS G., KARKALETSIS V., SPYROPOULOS. C.D. (2001). Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. Actes de *ACL-EACL 2001* .

ROSSET S., GALIBERT O., BERNARD G., BILINSKI E., ADDA G. (2008). The LIMSI participation to the QAst track. Actes de *Working Notes of CLEF 2008 Workshop*.

SEKINE S., NOBATA C. (2004). Definition, Dictionary and Tagger for Extended Named Entities. Actes de *LREC 04*.

SINCLAIR J. (1991). *Corpus, concordance, collocation: Describing English language*. Oxford: Oxford University Press.

VILLANEAU J., ROSSET S., GALIBERT O. (2007). Semantic Relations for an Oral and Interactive Question-Answering System. Actes de *SRSL7*.

WARD W., ISSAR S., HUANG X., HON H., HWANG M., YOUNG S., MATESSA M., LIU F., STERN R. (1992). Speech Understanding In Open Tasks. Actes de *the Fifth DARPA Workshop on Speech and Natural Language.*

TALN 2011, Montpellier, 27 juin –1er juillet 2011

Désambiguïsation lexicale par propagation de mesures sémantiques locales par algorithmes à colonies de fourmis

Didier Schwab, Jérôme Goulian, Nathan Guillaume
LIG-GETALP (Laboratoire d'Informatique de Grenoble, Groupe d'Étude pour la Traduction/le Traitement Automatique des Langues et de la Parole)
Université Pierre Mendès France, Grenoble 2
{didier.schwab, jerome.goulian}@imag.fr

Résumé. Effectuer une tâche de désambiguïsation lexicale peut permettre d'améliorer de nombreuses applications du traitement automatique des langues comme l'extraction d'informations multilingues, ou la traduction automatique. Schématiquement, il s'agit de choisir quel est le sens le plus approprié pour chaque mot d'un texte. Une des approches classiques consiste à estimer la proximité sémantique qui existe entre deux sens de mots puis de l'étendre à l'ensemble du texte. La méthode la plus directe donne un score à toutes les paires de sens de mots puis choisit la chaîne de sens qui a le meilleur score. La complexité de cet algorithme est exponentielle et le contexte qu'il est calculatoirement possible d'utiliser s'en trouve réduit. Il ne s'agit donc pas d'une solution viable. Dans cet article, nous nous intéressons à une autre méthode, l'adaptation d'un algorithme à colonies de fourmis. Nous présentons ses caractéristiques et montrons qu'il permet de propager à un niveau global les résultats des algorithmes locaux et de tenir compte d'un contexte plus long et plus approprié en un temps raisonnable.

Abstract. Word sense disambiguation can lead to significant improvement in many Natural Language Processing applications as Machine Translation or Multilingual Information Retrieval. Basically, the aim is to choose for each word in a text its best sense. One of the most popular method estimates local semantic relatedness between two word senses and then extends it to the whole text. The most direct method computes a rough score for every pair of word senses and chooses the lexical chain that has the best score. The complexity of this algorithm is exponential and the context that it is computationally possible to use is reduced. Brute force is therefore not a viable solution. In this paper, we focus on another method : the adaptation of an ant colony algorithm. We present its features and show that it can spread at a global level the results of local algorithms and consider a longer and more appropriate context in a reasonable time.

Mots-clés : Désambiguïsation lexicale, Algorithmes à colonies de fourmis, Mesures sémantiques.

Keywords: Lexical Disambiguation, Ant colony algorithms, Semantic relatedness.

1 Introduction

Effectuer une tâche de désambiguïsation lexicale peut permettre d'améliorer de nombreuses applications du traitement automatique des langues comme l'extraction d'informations multilingues, le résumé automatique ou encore la traduction automatique. Schématiquement, il s'agit de choisir quel est le sens le plus approprié pour chaque mot d'un texte dans un inventaire pré-défini. Par exemple, dans "*La souris mange le fromage.*", l'animal devrait être

préféré au dispositif électronique. De nombreux travaux existent sur le sujet, que l'on sépare habituellement en approches supervisées et non-supervisées. Les premières utilisent des apprentissages réalisés grâce à des corpus manuellement annotés, les secondes n'utilisent pas de telles données. Une catégorie intermédiaire, constituée des approches semi-supervisées, utilise quelques données annotées comme, par exemple, un sens par défaut issu d'un corpus annoté lorsque l'algorithme principal échoue (Navigli & Lapata, 2010). Le lecteur pourra consulter (Ide & Véronis, 1998) pour les travaux antérieurs à 1998 et (Agirre & Edmonds, 2006) ou (Navigli, 2009) pour un état de l'art complet.

La création de données annotées est une opération compliquée puisqu'elle nécessite une importante main d'œuvre et qu'elle doit être réalisée pour chaque inventaire de sens, pour chaque langue et même pour chaque domaine spécifique (sport, finance, ...). Cette constatation, que nous partageons avec (Navigli & Lapata, 2010), nous conduit à nous intéresser plus particulièrement à des approches non-supervisées. Une de ces approches classiques consiste à estimer la proximité sémantique qui existe entre deux sens de mots puis de l'étendre à l'ensemble du texte. En d'autres termes, il s'agit de donner des scores locaux et de les propager au niveau global (phrase, paragraphe, texte, ...). La méthode la plus directe, utilisée par exemple par (Pedersen *et al.*, 2005) utilise un algorithme brutal qui donne un score à toutes les paires de sens de mots puis choisit la chaîne de sens qui a le meilleur score. La complexité de cet algorithme est exponentielle et le contexte qu'il est calculatoirement possible d'utiliser s'en trouve réduit. Ainsi, alors qu'une analyse au niveau de la phrase n'est déjà pas toujours possible, un contexte linguistiquement plus pertinent comme, par exemple, le paragraphe l'est encore moins.

Les applications que nous visons doivent pouvoir être utilisées en temps réel. Lorsque l'on recherche une image et encore plus lorsque l'on appelle quelqu'un qui parle une autre langue au téléphone, les réponses doivent être immédiates. Il ne s'agit donc pas une solution viable et nous étudions d'autres méthodes.

Dans cet article, nous nous intéressons à la propagation de mesures de proximité sémantique locales grâce à une adaptation d'un algorithme à colonies de fourmis. Nous présentons dans un premier temps les mesures locales que nous utilisons puis quelques unes des caractéristiques de notre algorithme de propagation. Enfin, à titre d'exemple, nous évaluons notre approche sur la tâche *gros grain* de la campagne d'évaluation Semeval 2007 (Navigli *et al.*, 2007). Nous comparons en particulier notre algorithme de propagation à l'algorithme exhaustif classique et montrons qu'il permet d'obtenir efficacement une meilleure F-mesure.

2 Algorithmes locaux

2.1 Mesures de proximité sémantique

Ces méthodes consistent à donner un score censé refléter la proximité des objets linguistiques (généralement des mots ou des sens de mots) comparés. Ces scores peuvent être des similarités, donc avoir une valeur entre 0 et 1, des distances, et donc respecter leurs trois propriétés (séparation, symétrie et inégalité triangulaire) ou plus généralement, être une valeur positive non bornée.

Parmi elles, on peut citer *Hirst & Saint-Honge* basée sur la distance en terme de graphe entre deux sens dans un réseau lexical ; *Rada et al.* ainsi que *Leacock and Chodorow* similaires à la précédente mais ne considérant que les liens de type hyperonymie ; les mesures ou distances entre vecteurs (LSA (Deerwester *et al.*, 1990), vecteurs conceptuels (Schwab, 2005)). On pourra consulter (Pedersen *et al.*, 2005), (Cramer *et al.*, 2010) ou (Navigli, 2009) pour un panorama plus complet.

En désambiguïsation lexicale, ces méthodes sont utilisées de façon locale entre deux sens de mots, et sont ensuite

appliquées à un niveau global. Dans cet article, nous nous concentrons sur l'algorithme global et, à des fins de comparaison, nous présentons deux algorithmes locaux basés sur l'algorithme de Lesk.

2.2 Algorithmes locaux de cette expérience

2.2.1 Des algorithmes inspirés par Lesk

Nous utilisons dans cet article deux variantes de l'algorithme de Lesk (Lesk, 1986). Proposées il y a plus de 25 ans, il se caractérise par sa simplicité. Il ne nécessite qu'un dictionnaire et aucun apprentissage. Le score donné à une paire de sens est le nombre de mots – ici simplement les suites de caractères séparées par des espaces – en commun dans leur définition, sans tenir compte ni de leur ordre, ni de sous-séquences communes (approche sac de mots), ni d'informations morphologiques ou syntaxiques. Les variantes de cet algorithme sont encore aujourd'hui parmi les meilleures sur l'anglais (Ponzetto & Navigli, 2010). Ce premier algorithme local est nommé dans la suite *Lesk*.

Nous utilisons WordNet (Fellbaum, 1998), une base lexicale pour l'anglais, dans laquelle les sens de mots (les synsets) sont reliés par des relations (hyperonymie, hyponymie, antonymie, *etc.*). Notre second algorithme local exploite ces liens. Au lieu d'utiliser uniquement la définition d'un sens, elle utilise également la définition des différents sens qui lui sont liés. Cette idée est similaire à celle de (Banerjee & Pedersen, 2002) [1]. Ce second algorithme local est nommé dans la suite *Lesk étendu*.

2.2.2 Efficacité algorithmique

L'algorithme de base pour comparer le nombre de mots communs à deux définitions a une complexité en $O(n \times m)$ avec n et m, les longueurs en mots des définitions. De plus, la comparaison de chaînes de caractères est une opération relativement chère. On pourrait penser qu'il suffirait de précalculer la matrice de similarités avec l'ensemble des définitions. Cette idée est utopique vu la taille que peuvent atteindre les dictionnaires (jusqu'à plusieurs millions de définitions) [2] mais aussi parce qu'on a toujours besoin de faire des calculs sur de nouvelles données puisque (1) les données et les sens peuvent évoluer au cours du temps comme dans (Schwab, 2005), (2) notre algorithme de propagation utilise des pseudo-définitions crées à la volée (voir partie 4.2.2).

Nous avons amélioré ce calcul en utilisant un prétraitement qui se déroule en deux étapes. Dans la première, nous affectons à chacun des mots trouvés dans le dictionnaire un nombre entier tandis que, dans la seconde, nous convertissons chacune des définitions en un vecteur de nombres correspondant aux mots qu'elle contient, triés du plus petit au plus grand. Nous appelons ces vecteurs, *vecteurs de définitions*.

Par exemple, si notre première étape a donné ‹‹kind››$= 1$; ‹‹of››$= 2$; ‹‹evergreen››$= 3$; ‹‹tree››$= 4$; ‹‹with››$= 5$ ‹‹needle-shaped››$= 6$; ‹‹leaves››$= 7$; ‹‹fruit››$= 8$; ‹‹certain››$= 9$ avec la définition A, "*kind of evergreen tree with needle-shaped leaves*", nous obtenons le vecteur $[1, 2, 3, 4, 5, 6, 7]$ et avec B, "*fruit of certain evergreen tree*", nous obtenons $[2, 3, 4, 8, 9]$.

Cette conversion a deux avantages : (1) la comparaison de nombres est bien plus efficace que la comparaison de chaînes de caractères, (2) ordonner ces nombres permet d'éviter des comparaisons inutiles et de gagner en

1. (Banerjee & Pedersen, 2002) introduit également une notion de sous-séquence identique dans les définitions. Nous n'avons pas encore testé cette variante dont la complexité algorithmique est nettement supérieure à celle de notre algorithme.
2. Une forme de cache pourrait en partie régler ce problème.

efficacité. Ainsi, avec ce prétraitement, la complexité passe de $O(n \times m)$ à $O(n)$ où n et m ($n \geq m$) sont les longueurs (en nombre de mots) des définitions.

Pour les définitions A et B, calculer cette même proximité sémantique avec l'algorithme sur les définitions brutes se fait en $7 \times 5 = 35$ opérations (qui plus est sur des chaînes de caractères) tandis que si les définitions sont converties en vecteurs, nous n'avons plus que 7 opérations.

3 Algorithmes globaux

L'algorithme global est l'algorithme qui va permettre de propager les résultats d'un ou plusieurs algorithmes locaux à l'ensemble du texte afin de pouvoir en déduire un sens pour chaque mot. La méthode la plus directe est la recherche exhaustive utilisée par exemple dans (Banerjee & Pedersen, 2002). Il s'agit de considérer les combinaisons de l'ensemble des sens des mots dans le même contexte (fenêtre de mots, phrase, texte, *etc.*), de donner un score à chacune de ces combinaisons et de choisir celle qui a le meilleur score. Le principal problème de cette méthode est la rapide explosion combinatoire qu'elle engendre. Considérons la phrase suivante tirée du corpus d'évaluation que nous utilisons dans la partie 5, "*The pictures they painted were flat, not round as a figure should be, and very often the feet did not look as if they were standing on the ground at all, but pointed downwards as if they were hanging in the air.*", ‹*picture*› a 9 sens, ‹*paint*› 4, ‹*be*› 13, ‹*flat*› 17, ‹*figure*› 13, ‹*very*› 2, ‹*often*› 2, ‹*foot*› 11, ‹*look*› 10, ‹*stand*› 12, ‹*ground*› 11, ‹*at all*› 1, ‹*point*› 13, ‹*downwards*› 1, ‹*hang*› 15 et ‹*air*› 9 sens, il y a alors 137 051 946 345 600 combinaisons de sens possibles à analyser. Ce nombre est comparable à la quantité d'opérations (et le calcul d'une combinaison nécessite des dizaines voire des centaines d'opérations) que peuvent théoriquement effectuer 3300 processeurs Core i7-990X (2,43GHz, 6 cœurs, 12 fils d'exécutions) sortis par Intel au premier trimestre 2011 en une seconde. Le calcul exhaustif est donc très compliqué à réaliser dans des conditions réelles et, surtout, rend impossible l'utilisation d'un contexte d'analyse plus important.

Pour contourner ce problème, plusieurs solutions ont été proposées. Par exemple, des approches utilisant un corpus pour diminuer le nombre de combinaisons à examiner comme la recherche des chaînes lexicales compatibles (Gale *et al.*, 1992; Vasilescu *et al.*, 2004) ou encore des approches issues de l'intelligence artificielle comme le recuit simulé (Cowie *et al.*, 1992) ou les algorithmes génétiques (Gelbukh *et al.*, 2003).

Ces méthodes ont en commun de ne pas permettre l'exploitation de façon directe et simple d'une structure linguistique sous forme de graphe que ce soit une analyse morphologique ou une analyse syntaxique. Nous utilisons, au contraire, une méthode à colonies de fourmis pour l'analyse sémantique inspirée de (Schwab & Lafourcade, 2007) afin de pouvoir à terme utiliser de telles structures [3].

4 Notre algorithme global : un algorithme à colonies de fourmis

4.1 Les algorithmes à colonies de fourmis

Les algorithmes à fourmis ont pour origine la biologie et les observations réalisées sur le comportement social des fourmis. En effet, ces insectes ont collectivement la capacité de trouver le plus court chemin entre leur fourmilière et une source d'énergie. Il a pu être démontré que la coopération au sein de la colonie est auto-organisée et résulte d'interactions entre individus autonomes. Ces interactions, souvent très simples, permettent à la colonie

3. Dans un premier temps, nous utiliserons ici une structure linguistique extrêmement simpl(ist)e.

de résoudre des problèmes compliqués. Ce phénomène est appelé intelligence en essaim (Bonabeau & Théraulaz, 2000). Il est de plus en plus utilisé en informatique où des systèmes de contrôle centralisés gagnent souvent à être remplacés par d'autres, fondés sur les interactions d'éléments simples.

En 1989, Jean-Louis Deneubourg étudie le comportement des fourmis biologiques dans le but de comprendre la méthode avec laquelle elles choisissent le plus court chemin et le retrouvent en cas d'obstacle. Il élabore ainsi le modèle stochastique dit *de Deneubourg* (Deneubourg *et al.*, 1989), conforme à ce qui est observé statistiquement sur les fourmis réelles quant à leur partage entre les chemins. Ce modèle stochastique est à l'origine des travaux sur les algorithmes à fourmis.

Le concept principal de l'intelligence en essaim est la stygmergie, c.-à-d. l'interaction entre agents par modification de l'environnement. Une des premières méthodes que l'on peut apparenter aux algorithmes à fourmis est l'écorésolution qui a montré la puissance d'une heuristique de résolution collective basée sur la perception locale, évitant tout parcours explicite de graphe d'états (Drogoul, 1993).

En 1992, Marco Dorigo et Luca Maria Gambardella conçoivent le premier algorithme basé sur ce paradigme pour le célèbre problème combinatoire du voyageur de commerce (Dorigo & Gambardella, 1997). Dans les algorithmes à base de fourmis artificielles, l'environnement est généralement représenté par un graphe et les fourmis virtuelles utilisent l'information accumulée sous la forme de chemins de phéromone déposée sur les arcs du graphe. De façon simple, une fourmi se contente de suivre les traces de phéromones déposées précédemment ou explore au hasard dans le but de trouver un chemin optimal, fonction du problème posé, dans le graphe.

Ces algorithmes offrent une bonne alternative à tout type de résolution de problèmes modélisables sous forme d'un graphe. Ils permettent un parcours rapide et efficace et offrent des résultats comparables à ceux obtenus par les différentes méthodes de résolution. Leur grand intérêt réside dans leur capacité à s'adapter à un changement de l'environnement. Le lecteur trouvera dans (Dorigo & Stützle, 2004) ou (Monmarche *et al.*, 2009) de bons états de l'art sur la question.

4.2 Algorithme à colonies de fourmis et désambiguïsation lexicale

4.2.1 Vue d'ensemble

L'environnement des fourmis est un graphe. Il peut être linguistique – morphologique comme dans (Rouquet *et al.*, 2010) ou morpho-syntaxique comme dans (Schwab & Lafourcade, 2007; Guinand & Lafourcade, 2009) – ou être simplement organisé en fonction des éléments du texte. En fonction de l'environnement choisi, les résultats de l'algorithme ne sont évidemment pas les mêmes. Des recherches sont actuellement menées à ce sujet mais, dans cet article, nous ne nous intéressons qu'à un cas de base c.-à-d. un graphe simple (voir fig.1), sans information linguistique externe, afin de mieux comprendre la mécanique de nos algorithmes.

Dans ce graphe, nous distinguons deux types de nœuds : les *fourmilières* et les *nœuds normaux*. Suivant les idées développées dans (Schwab, 2005) et (Guinand & Lafourcade, 2009), chaque sens possible d'un mot est associé à une fourmilière. Les fourmilières produisent des fourmis. Ces fourmis se déplacent dans le graphe à la recherche d'*énergie* puis la rapportent à leur fourmilière mère qui pourra alors créer de nouvelles fourmis. Pour une fourmi, un nœud peut être : (1) *la fourmilière maison* où elle est née ; (2) *une fourmilière ennemie* qui correspond à un autre sens du même mot ; (3) *une fourmilière potentiellement amie*, toutes celles qui ne sont pas ennemies ; (4) un *nœud qui n'est pas une fourmilière*, les nœuds normaux.

Par exemple, dans la figure 1, pour une fourmi née dans la fourmilière 19, le nœud 18 est un ennemi comme il a

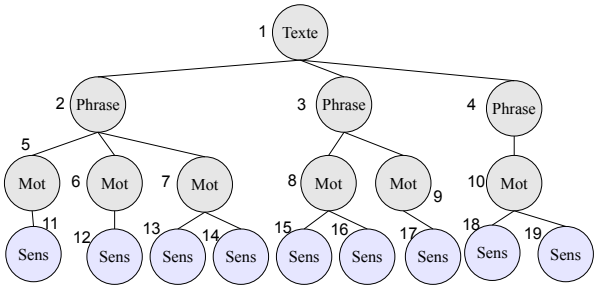


FIGURE 1 – Environnement utilisé dans cette expérience : texte, phrases et mots correspondent aux nœuds dits normaux 1 à 10, un sens de mot correspond à à une fourmilière (nœuds 11 à 19).

le même père (10), les fourmilières potentiellement amies sont les nœuds 11 à 17 et les nœuds normaux sont les nœuds 1 à 10.

Les déplacements des fourmis se déroulent en fonction des scores locaux (cf. section 2.2), de la présence d'énergie, et du passage des autres fourmis (Les fourmis laissent des traces sur les arcs où elles passent sous la forme de *phéromone*). Une fois arrivée sur la fourmilière d'un autre terme, une fourmi peut choisir de revenir directement à sa fourmilière mère. Elle établit alors, entre les deux fourmilières, un pont que les autres fourmis sont, à leur tour, susceptibles d'emprunter et de renforcer grâce à leur phéromone. Ce renforcement a lieu si les informations lexicales conduisent les autres fourmis à emprunter le pont et disparaît dans le cas inverse. Ainsi, les fourmis établissent de nombreux liens entre fourmilières de sens compatibles.

Les ponts correspondent ainsi à des interprétations de la phrase. L'émergence de tels circuits dans le graphe contribue à la monopolisation des ressources de la colonie (fourmis et énergie) et à l'épuisement des ressources associées aux autres fourmilières (ces cas correspondent donc aux sens incompatibles dans le contexte et avec les ressources considérés).

4.2.2 Détails de l'algorithme

Énergie Au début de la simulation, le système possède une certaine énergie qui est répartie équitablement sur chacun des nœuds. Les fourmilières utilisent celle qu'elles possèdent pour fabriquer des fourmis avec une probabilité fonction de cette même énergie et suivant une courbe sigmoïde (cf. fig. 2). On peut remarquer que l'utilisation de cette fonction permet aux fourmilières qui n'ont plus d'énergie de fabriquer quelques fourmis supplémentaires (et ainsi d'avoir une quantité d'énergie négative). L'idée est de leur donner une dernière chance au cas où ces fourmis, trouvant des informations lexicales pertinentes, rapportent de l'énergie et relancent la production de fourmis.

Les fourmis ont une durée de vie (nombre de cycles identique pour toutes et paramétré (cf. tableau 4.2.2)). Lorsqu'une fourmi meurt, l'énergie qu'elle porte ainsi que l'énergie utilisée par la fourmilière pour la produire est déposée sur le nœud où elle se trouve. Il n'y a donc ni perte ni apport d'énergie à aucun moment que ce soit. Si on excepte l'emprunt à la nature que peuvent faire de façon très limitée les fourmilières, le système fonctionne complètement en vase clos. La quantité d'énergie est un élément fondamental de la convergence du système vers une solution. En effet, puisque l'énergie globale est limitée, les fourmilières sont en concurrence les unes avec les

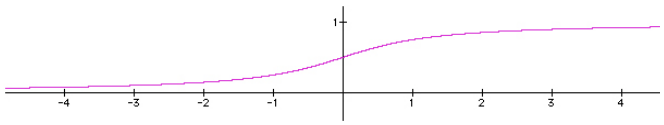


FIGURE 2 – Courbe de la fonction sigmoïde $\frac{arctan(x)}{\pi} + \frac{1}{2}$ qui permet de calculer la probabilité de la naissance d'une fourmi à partir de la quantité d'énergie présente sur le nœud.

autres et seules des alliances peuvent permettre de faire émerger des solutions.

Phéromone de passage Les fourmis ont deux types de comportement. Elles peuvent soit chercher de l'énergie, soit chercher à revenir à leur fourmilière mère. Lorsqu'elles se déplacent dans le graphe, elles laissent des traces sur les arcs où elles passent sous la forme de phéromone. La phéromone influe sur les déplacements des fourmis qui préfèrent l'éviter lorsqu'elles cherchent de l'énergie et préfèrent la suivre lorsqu'elles tentent de revenir déposer cette énergie à leur fourmilière mère.

Lors d'un déplacement, une fourmi laisse une trace en déposant sur l'arc A traversé une quantité de phéromone $\theta \in \mathbb{R}^+$. On a alors $\varphi_{t+1}(A) = \varphi_t(A) + \theta$.

À chaque cycle, il y a une légère évaporation des phéromones. Cette baisse se fait de façon linéaire jusqu'à la disparition totale de la phéromone. Nous avons ainsi, $\varphi_{c+1}(A) = \varphi_c(A) \times (1 - \delta)$ où δ est la proportion de phéromone qui s'évapore à chaque cycle.

Création, suppression et type de ponts Un pont peut être créé lorsqu'une fourmi atteint une fourmilière potentiellement amie, c.-à-d. lorsqu'elle arrive sur un nœud qui correspond à un sens d'un autre mot que celui de la fourmilière mère. Dans ce cas, la fourmi évalue non seulement les nœuds liés à cette fourmilière mais aussi le nœud correspondant à sa fourmilière mère. Si ce dernier est sélectionné, il y a création d'un pont entre les deux fourmilières. Ce pont est ensuite considéré comme un arc standard par les fourmis, c.-à-d. que les nœuds qu'il lie sont considérés comme voisins. Si le pont ne porte plus de phéromone, il disparaît.

Odeur L'odeur d'une fourmilière est la représentation vectorielle que nous avons introduite dans la partie 2.2.2. Elle correspond donc à la définition du sens sous forme de vecteur de nombres entiers. Chaque fourmi née dans cette fourmilière porte la même odeur, le même vecteur. Lors de son déplacement sur les nœuds normaux du graphe, une fourmi propage son vecteur. Le vecteur V(N) porté par un nœud normal N est modifié lors du passage d'une fourmi. La fourmi dépose une partie de son vecteur, un pourcentage des composantes prises au hasard qui remplace la même quantité d'anciennes valeurs elles aussi choisies au hasard.

Cette propagation intervient dans le déplacement des fourmis. Laisser une partie de son vecteur, c'est laisser une trace de passage. Ainsi plus un nœud est proche d'une fourmilière plus il y a de chance que les fourmis de cette fourmilière y soient passées. Ce phénomène permet aux fourmis de revenir à leur fourmilière, ou éventuellement de se tromper et de se diriger vers des fourmilières amies. Cette erreur est ainsi potentiellement bénéfique puisqu'elle

peut permettre de créer un pont entre les deux fourmilières (cf. 4.2.2). En revanche, lorsqu'une fourmi se trouve sur une fourmilière, le vecteur n'est pas modifié. Ces nœuds conservent ainsi un vecteur constant tout au long de la simulation.

La table suivante présente les paramètres, les notations et les valeurs utilisées dans l'algorithme présenté et expérimenté ici. Cet article ne présente pas les expériences réalisées pour trouver ces valeurs.

Notation	Description	Valeurs
F_A	Fourmilière correspondant au sens A	na
$V(X)$	Vecteur odeur associé à X. X est un nœud ou une fourmi	na
f_A	Fourmi née dans la fourmilière F_A	na
E_f	Énergie utilisée par une fourmilière pour produire une fourmi	na
$E(X)$	Énergie possédée par X. X est un nœud ou une fourmi	na
E_{max}	Énergie maximale que peut porter une fourmi	5
$\varphi(A)$	Quantité de phéromone sur l'arc A	na
θ	Phéromone déposée par une fourmi lors de la traversée d'un arc	1
δ	Évaporation de la phéromone entre chaque cycle	20%
$Eval_f(X)$	Évaluation de X selon la fourmi f. X est un arc ou un nœud	na
$Eval_f(N, A)$	Évaluation du nœud N en passant par l'arc A selon la fourmi f	na
	Nombre de cycles de la simulation	100
	Quantité initiale d'énergie sur chaque nœud	20
	Durée de vie d'une fourmi	10
	Énergie prise par une fourmi lorsqu'elle arrive sur un nœud	1
	Longueur du vecteur odeur	50
	Quantité du vecteur odeur modifié par une fourmi lorsqu'elle arrive sur un nœud	10%

4.2.3 Déroulement de l'algorithme

L'algorithme consiste en une itération potentiellement infinie de cycles. À tout moment, la simulation peut être interrompue et l'état courant observé. Durant un cycle, on effectue les tâches suivantes : (1) éliminer les fourmis trop vieilles (la durée de vie est un paramètre) ; (2) pour chaque fourmilière, solliciter la production d'une fourmi (une fourmi peut ou non voir le jour, de façon probabiliste) ; (3) pour chaque arc, diminuer le taux de phéromone (évaporation des traces) ; (4) pour chaque fourmi : déterminer son mode (recherche d'énergie, retour à la fourmilière, le changement est fait de manière probabiliste) et la déplacer. Créer un pont interprétatif le cas échéant ; (5) calculer les conséquences du déplacement des fourmis (sur l'activation des arcs et l'énergie des nœuds).

Les déplacements d'une fourmi sont aléatoires mais influencés par son environnement. Lorsqu'une fourmi est sur un nœud, elle estime tous les nœuds voisins et tous les arcs qui les lient. La probabilité d'emprunter un arc A_j pour aller à un nœud N_i est $P(N_i, A_j) = max(\frac{Eval_f(N_i,A_j)}{\sum_{k=1,l=1}^{k=n,l=m} Eval_f(N_k,A_l)}, \epsilon)$ où $Eval_f(N, A)$ est l'évaluation du nœud N en prenant l'arc A, c.-à-d. la somme de $Eval_f(N)$ et de $Eval_f(A)$. ϵ permet à certaines fourmis de choisir des destinations évaluées comme improbables mais qui permettraient d'atteindre des informations lexicales et des ressources qui s'avèreraient intéressantes ensuite.

Une fourmi qui vient de naître (c.-à-d. être produite par sa fourmilière) part à la recherche d'énergie. Elle est attirée par les nœuds qui portent beaucoup d'énergie ($Eval_f(N) = \frac{E(N)}{\sum_0^m E(N_i)}$) et évite les arcs qui portent beaucoup de phéromone ($Eval_f(A) = 1 - \varphi(A)$) afin de permettre l'exploration de plus de solutions. Elle continue à collecter de l'énergie jusqu'au cycle où un tirage aléatoire avec la probabilité $P(retour) = \frac{E(f)}{E_{\max}}$ la fera passer en mode retour. Dans ce mode, elle va (statistiquement) suivre les arcs avec beaucoup de phéromone ($Eval_f(A) = \varphi(A)$) et vers les nœuds dont l'odeur est proche de la leur ($Eval_f(N) = \frac{Lesk(V(N),V(f_A))}{\sum_{i=1}^{i=k} Lesk(V(N_i),V(f_A))}$).

5 Évaluation

Nous avons testé notre méthode sur le corpus de la tâche *gros grain* de la campagne d'évaluation *Semeval 2007* (Navigli *et al.*, 2007) dans laquelle les organisateurs fournissent un inventaire de sens plus grossiers que ceux de WordNet. Pour chaque terme, les sens considérés comme proches (par exemple, "*neige/précipitation*" et "*neige/couverture*" ou "*porc/animal*" et "*porc/viande*") sont groupés. Le corpus est composé de 5 textes de genres divers (journalisme, critique littéraire, voyage, informatique, biographies) dont il faut annoter les 2269 mots. Le nombre moyen de sens par mot est de 6,19 ; ramené à 3,1 pour l'inventaire de sens grossiers. Les compétiteurs étaient libres de se servir de cet inventaire (sens grossiers connus *a priori*) ou non (sens grossiers connus *a posteriori*). Dans le premier cas, le nombre de choix à faire pour chaque mot est réduit et la tâche moins compliquée. Dans le second cas, les sens annotés sont jugés corrects s'ils sont dans le bon groupement, une sorte d'erreur acceptable. Notre objectif est de tester un système en vue d'une utilisation dans un cadre applicatif réel or l'inventaire de sens grossiers n'est disponible que pour les 2269 mots utilisés dans le corpus d'évaluation, nous ne l'utilisons donc pas. Dans les expériences présentées ici, nous nous situons ainsi dans un cas de sens connus *a posteriori*. Les résultats sont analysés par les formules classiques :

$$\text{Précision } P = \frac{sens\ correctement\ annotés}{sens\ annotés} \qquad \text{Rappel } R = \frac{sens\ correctement\ annotés}{sens\ à\ annoter} \qquad \text{F-mesure } F = \frac{2\times P\times R}{P+R}$$

Dans le corpus, les mots sont annotés avec leur partie du discours (verbe, nom, adverbe, adjectif). À partir de ces informations, nous construisons l'environnement des fourmis : un nœud au niveau du texte, un nœud pour chaque phrase, un nœud pour chaque mot et une fourmilière pour chaque sens (voir fig. 1). À la fin d'un cycle, le sens sélectionné pour chaque mot correspond à la fourmilière qui a la plus grande quantité d'énergie.

5.1 Exécution de l'algorithme

L'algorithme à colonies de fourmis garantit la réalisation d'un choix entre les différentes possibilités pour chaque terme. Ainsi, 100% du corpus est annoté et P=R=F puisque les sens annotés sont égaux aux sens à annoter (P=R) et dans ce cas $F = \frac{2\times P\times P}{P+P} = \frac{2\times P^2}{2P} = P$. De plus, un algorithme à colonies de fourmis est un algorithme stochastique, il ne sélectionne donc pas exactement les mêmes sens à chaque exécution ni même à chaque cycle. Nous avons exécuté cet algorithme des centaines de fois et avons noté qu'après 70-80 cycles, les résultats restaient globalement constants comme l'illustre la figure suivante.

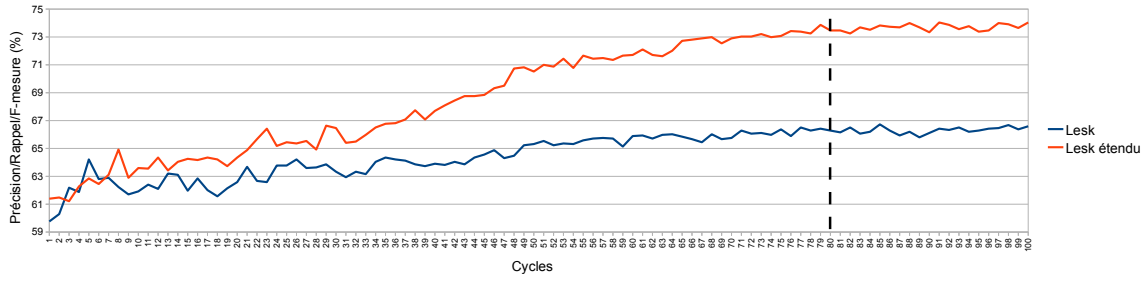


FIGURE 3 – Évolution de la précision/rappel/F-mesure dans les 100 cycles d'une exécution de l'algorithme à colonies de fourmis utilisé avec l'algorithme local Lesk et avec l'algorithme local Lesk étendu

5.2 Comparaison d'exécutions

De la même manière que les résultats évoluent entre deux cycles, les résultats peuvent être différents entre deux exécutions. Pour donner une idée de cette différence, nous avons répété notre expérience, arrêtée au bout de 100 cycles, sur chacun des algorithmes locaux, 100 fois. La table suivante en présente les résultats. Nous obtenons seulement $2,95\%$ d'écart entre le meilleur et le moins bon résultat (soit 67 termes mal annotés sur 2269) pour *Lesk étendu* et 3,39% (soit 77 termes mal annotés sur 2269) pour *Lesk*.

Algorithme local	Minimum	Maximum	Moyenne	Médiane	Étendue	Écart-type
Lesk	64,43	67,83	66,34	66,35	3,39	0,66
Lesk étendu	72,54	75,5	74,01	74,04	2,95	0,58

5.3 Comparaisons avec l'algorithme exhaustif

À titre de comparaison avec notre approche, nous présentons les résultats obtenus par l'algorithme global exhaustif (Banerjee & Pedersen, 2002). Nous avons choisi comme contexte la phrase, excluant *de facto* les phrases d'un mot (au nombre de quatre, soit moins de $0,002\%$ du corpus). Pour des raisons calculatoires, nous avons également exclu les phrases de plus de 10 milliards de combinaisons. Nous pouvons voir que seulement $77,3\%$ du corpus a

Algorithme global	Algorithme local	Étiquetés	Précision	Rappel	F-mesure	Temps
Calcul exhaustif	Lesk	77,30	69,21	53,50	60,35	$\approx$ 40h
	Lesk étendu	77,30	77,82	60,16	67,86	$\approx$ 300h
Fourmis	Lesk	100,0	64,43 - 67,83	64,43 - 67,83	64,43 - 67,83	$\approx$ 3m
	Lesk étendu	100,0	72,54 - 75,5	72,54 - 75,5	72,54 - 75,5	$\approx$ 8m

été annoté au prix d'une durée de plusieurs heures incompatible avec des applications en temps réel [4].

Pour les deux algorithmes locaux, la F-mesure est clairement supérieure à celle du calcul brut pour un temps nettement moins long (800 fois plus court pour Lesk et 2250 fois pour Lesk étendu). Le tableau suivant présente pour les mêmes exécutions les résultats sur les différentes sous-parties du corpus : A, la partie annoté par les 2 algorithmes globaux et B celle qui n'est annotée que par l'algorithme fourmis. Sur la partie A, les fourmis sont, comme on pouvait s'en douter, légèrement en dessous de l'algorithme exhaustif et leur meilleur résultat s'explique par la possibilité d'annoter la sous-partie B.

Pour conclure cette évaluation, nous avons comparé nos résultats avec les résultats obtenus par les différents systèmes qui participaient à la campagne Semeval 2007. Avec Lesk étendu, nous serions arrivés $8^{ème}$/15 en tenant compte de tous les participants, $5^{ème}$/8 sur ceux qui ne connaissent pas *a priori* les sens grossiers, 1^{er}/7 sur les approches non supervisées. Ces résultats sont très encourageants vu les temps de calcul (aucun article des participants n'aborde ce point), les possibilités d'extension qu'offrent les algorithmes à fourmis et la simplicité des algorithmes locaux envisagés ici.

4. Expériences réalisées sur des processeurs Intel Xeon X5550, 4 cœurs à 2.66Ghz (durées converties en temps monoprocesseurs).

Algorithme local	Sous-corpus	Algorithme global	Étiquetés	Rappel	Différentiel
Lesk	A + B	Exhaustif	77,30	53,50	
		Fourmis	100,0	64,43 - 67,83	+ 10,93 à + 14,33
	A	Exhaustif	100,0	69,21	
		Fourmis	100,0	65,45 - 68,99	- 3,76 à - 0,22
	B	Exhaustif	00,00	00,00	
		Fourmis	100,0	60,97 - 63,88	+ 60,97 à + 63,88
Lesk étendu	A + B	Exhaustif	77,30	60,16	
		Fourmis	100,0	72,54 - 75,5	+ 12,38 à + 15,34
	A	Exhaustif	100,0	77,82	
		Fourmis	100,0	74,69 - 77,25	- 3,13 à - 0,57
	B	Exhaustif	00,00	00,00	
		Fourmis	100,0	65,24 - 69,52	+ 65,24 à + 69,52

6 Conclusions et Perspectives

Dans cet article, nous avons présenté un algorithme à colonies de fourmis destiné à la désambiguïsation lexicale et basé sur des mesures de proximité sémantique. Cet algorithme, non supervisé, est volontairement simple puisqu'il n'utilise qu'une seule ressource lexicale (WordNet) et aucune analyse morphologique ou morpho-syntaxique. Il permet pourtant de choisir un sens, pour chaque mot d'un texte, d'une manière plus rapide que l'algorithme exhaustif et en atteignant une bonne F-mesure pour un système non supervisé. Nous considérons ces résultats comme une ligne de base (baseline) à partir de laquelle nous allons poursuivre nos recherches. Outre l'ajout d'informations morphologiques et/ou syntaxiques, nous travaillons actuellement sur la combinaison de mesures locales et l'utilisation de WordNet dans l'environnement des fourmis. Nos travaux portent également sur d'autres algorithmes locaux et leur impact sur l'utilisation dans d'autres langues notamment flexionnelles. Enfin, nous travaillons à la comparaison des algorithmes à colonies de fourmis avec d'autres algorithmes globaux comme les algorithmes génétiques ou le recuit simulé.

Références

AGIRRE E. & EDMONDS P. (2006). *Word Sense Disambiguation : Algorithms and Applications (Text, Speech and Language Technology)*. Secaucus, NJ, USA : Springer-Verlag New York, Inc.

BANERJEE S. & PEDERSEN T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *the Third International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2002*, Mexico City.

BONABEAU É. & THÉRAULAZ G. (2000). L'intelligence en essaim. *Pour la science*, (271), 66–73.

COWIE J., GUTHRIE J. & GUTHRIE L. (1992). Lexical disambiguation using simulated annealing. In *COLING 1992, International Conference on Computational Linguistics*, volume 1, p. 359–365, Nantes, France.

CRAMER I., WANDMACHER T. & WALTINGER U. (2010). *WordNet : An electronic lexical database*, chapter Modeling, Learning and Processing of Text Technological Data Structures. Springer.

DEERWESTER S. C., DUMAIS S. T., LANDAUER T. K., FURNAS G. W. & HARSHMAN R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, **41**(6).

DENEUBOURG J.-L., GROSS S., FRANKS N. & PASTEELS J.-M. (1989). The blind leading the blind : Modeling chemically mediated army ant raid patterns. *Journal of Insect Behavior*, **2**, 719–725.

DORIGO & STÜTZLE (2004). *Ant Colony Optimization*. MIT-Press.

DORIGO M. & GAMBARDELLA L. (1997). Ant colony system : A cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, **1**, 53–66.

DROGOUL A. (1993). When ants play chess (or can strategies emerge from tactical behaviors). In *Maamaw'1993*.

FELLBAUM C. (1998). *WordNet : An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

GALE W., CHURCH K. & YAROWSKY D. (1992). One sense per discourse. In *Fifth DARPA Speech and Natural Language Workshop*, p. 233–237, Harriman, New-York, États-Unis.

GELBUKH A., SIDOROV G. & HAN S. Y. (2003). Evolutionary approach to natural language word sense disambiguation through global coherence optimization. *WSEAS Transactions on Communications*, **2**(1), 11–19.

GUINAND F. & LAFOURCADE M. (2009). *Fourmis Artificielles 2. Nouvelles Directions pour une Intelligence Collective*, chapter Fourmis Artificielles et Traitement de la Langue Naturelle, p. 225–267. Lavoisier.

IDE N. & VÉRONIS J. (1998). Word sense disambiguation : the state of the art. *Computational Linguistics*, **28**(1), 1–41.

LESK M. (1986). Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, p. 24–26, New York, NY, USA : ACM.

N. MONMARCHE, F. GUINAND & P. SIARRY, Eds. (2009). *Fourmis Artificielles et Traitement de la Langue Naturelle*. Prague, Czech Republic : Lavoisier.

NAVIGLI R. (2009). Word sense disambiguation : a survey. *ACM Computing Surveys*, **41**(2), 1–69.

NAVIGLI R. & LAPATA M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 678–692.

NAVIGLI R., LITKOWSKI K. C. & HARGRAVES O. (2007). Semeval-2007 task 07 : Coarse-grained english all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, p. 30–35, Prague, Czech Republic : Association for Computational Linguistics.

PEDERSEN T., BANERJEE S. & PATWARDHAN S. (2005). *Maximizing Semantic Relatedness to Perform Word Sense Disambiguation*. Research Report UMSI 2005/25, University of Minnesota Supercomputing Institute.

PONZETTO S. P. & NAVIGLI R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, p. 1522–1531, Stroudsburg, PA, USA : Association for Computational Linguistics.

ROUQUET D., FALAISE A., SCHWAB D., BOITET C., BELLYNCK V., NGUYEN H.-T., MANGEOT M. & GUILBAUD J.-P. (2010). *Rapport final de synthèse, passage à l'échelle et implémentation : Extraction de contenu sémantique dans des masses de données textuelles multilingues*. Rapport interne, Agence Nationale de la Recherche.

SCHWAB D. (2005). *Approche hybride - lexicale et thématique - pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte.* PhD thesis, Université Montpellier 2.

SCHWAB D. & LAFOURCADE M. (2007). Lexical functions for ants based semantic analysis. In *ICAI'07- The 2007 International Conference on Artificial Intelligence*, Las Vegas, Nevada, USA.

VASILESCU F., LANGLAIS P. & LAPALME G. (2004). Evaluating variants of the lesk approach for disambiguating words. In *Proceedings of LREC 2004, the 4th International Conference On Language Resources And Evaluation*, p. 633–636, Lisbon, Portugal.

Lexique et Corpus

Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé

Benoît Sagot[1] Karën Fort[2,3] Gilles Adda[4] Joseph Mariani[4,5] Bernard Lang[6]
(1) Alpage, INRIA Paris–Rocquencourt & Université Paris 7,
Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
(2) INIST-CNRS, 2 allée de Brabois, 54500 Vandoeuvre-lès-Nancy, France
(3) LIPN, Université Paris Nord, 99 av J-B Clément, 93430 Villetaneuse, France
(4) LIMSI-CNRS, Bât. 508, rue John von Neumann, Université Paris-Sud BP 133, 91403 Orsay Cedex, France
(5) IMMI-CNRS, Bât. 508, rue John von Neumann, Université Paris-Sud BP 133, 91403 Orsay Cedex, France
(6) INRIA Paris–Rocquencourt, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
{benoit.sagot, bernard.lang}@inria.fr, karen.fort@inist.fr, {gilles.adda,joseph.mariani}@limsi.fr

Résumé. Cet article est une prise de position concernant les plate-formes de type Amazon Mechanical Turk, dont l'utilisation est en plein essor depuis quelques années dans le traitement automatique des langues. Ces plate-formes de travail en ligne permettent, selon le discours qui prévaut dans les articles du domaine, de faire développer toutes sortes de ressources linguistiques de qualité, pour un prix imbattable et en un temps très réduit, par des gens pour qui il s'agit d'un passe-temps. Nous allons ici démontrer que la situation est loin d'être aussi idéale, que ce soit sur le plan de la qualité, du prix, du statut des travailleurs ou de l'éthique. Nous rappellerons ensuite les solutions alternatives déjà existantes ou proposées. Notre but est ici double : informer les chercheurs, afin qu'ils fassent leur choix en toute connaissance de cause, et proposer des solutions pratiques et organisationnelles pour améliorer le développement de nouvelles ressources linguistiques en limitant les risques de dérives éthiques et légales, sans que cela se fasse au prix de leur coût ou de leur qualité.

Abstract. This article is a position paper concerning Amazon Mechanical Turk-like systems, the use of which has been steadily growing in natural language processing in the past few years. According to the mainstream opinion expressed in the articles of the domain, these online working platforms allow to develop very quickly all sorts of quality language resources, for a very low price, by people doing that as a hobby. We shall demonstrate here that the situation is far from being that ideal, be it from the point of view of quality, price, workers' status or ethics. We shall then bring back to mind already existing or proposed alternatives. Our goal here is twofold : to inform researchers, so that they can make their own choices with all the elements of the reflection in mind, and propose practical and organizational solutions in order to improve new language resources development, while limiting the risks of ethical and legal issues without letting go price or quality.

Mots-clés : Amazon Mechanical Turk, ressources linguistiques.

Keywords: Amazon Mechanical Turk, language resources.

1 Introduction

Le traitement des langues a grandement évolué au cours des ces vingt dernières années, tant dans le traitement de l'écrit que de la parole. Stimulé par le paradigme de l'évaluation, le rôle des ressources linguistiques dans ce développement a été et reste crucial : elles sont à la fois matière première, objet d'étude et ressource pour l'évaluation de systèmes. Nous proposons ici une critiqued'un outil nouveau de constitution de ces ressources, le *microworking* par le biais du *crowdsourcing*. *Microworking* fait référence au fait que le travail est segmenté en petites tâches, *crowdsourcing* au fait que le travail est délocalisé (*outsourced*) et est effectué par un grand nombre de personnes (*crowd*), payées ou non. Nous néologiserons crowdsourcing en « myriadisation » et microworking en « travail parcellisé », et la conjonction des deux par « myriadisation du travail parcellisé ».

Nous aborderons en détails le cas d'un système de myriadisation du travail parcellisé (m.t.p. dans la suite) qui a fait florès ces derniers temps, Amazon Mechanical Turk (MTurk), notamment pour sa capacité à produire des corpus annotés à un coût très faible. Les auteurs de cet article ont contribué, à des degrés divers, à la mise en place du paradigme de l'évaluation et au développement de nombreux outils et ressources dans le domaine du

traitement du langage. Nous sommes à ce titre conscients de l'importance du développement et de la diffusion de celles-ci et du frein que représente leur coût, souvent rédhibitoire. Cependant, nous voulons mettre en avant le fait que le coût du développement est un argument non fondé en ce qui concerne la m.t.p., tout d'abord parce qu'il masque des problèmes économiques complexes, ensuite parce qu'il met sous le boisseau le problème de la qualité des ressources ainsi obtenues, enfin parce qu'il omet la question de l'éthique et du droit du travail. Nous aborderons ici l'ensemble de ces questions, sans pour autant remettre en cause l'utilité de la m.t.p., à condition que son fonctionnement et son utilisation se fassent selon certains principes.

2 Que sont les systèmes de myriadisation ?

Le concept de myriadisation est venu de l'idée qu'un certain nombre de tâches pouvaient être effectuées par des utilisateurs d'Internet, en utilisant les atouts propres à celui-ci, c'est-à-dire pouvoir accéder à un grand nombre de personnes, de manière quasi-instantanée, partout dans le monde. La participation de ces internautes peut être bénévole ou rétribuée, suivant les tâches et les systèmes. Parmi les systèmes bénévoles, nous pouvons citer l'exemple fameux de Wikipedia et, parmi ceux avec rétribution, RentACoder (où l'on peut soumettre un projet de programmation à une communauté de programmeurs) ou LiveOps (qui est un centre d'appels virtuel, les opérateurs étant des internautes). A la suite de ces systèmes est apparu le concept de *Human computing*. Dans ce dernier cas, on ne fait plus appel à des compétences particulières d'internautes, mais on utilise deux propriétés très élémentaires : être un humain et avoir du temps libre. C'est l'application des grilles de calcul aux humains : chaque utilisateur, à la manière d'un processeur, effectue une tâche élémentaire en n'ayant accès qu'à la seule information nécessaire pour la mener à bien. Dans ce type de systèmes, seules des tâches très simples sont effectuées par les humains, soit parce qu'elles sont intrinsèquement simples (par exemple, mettre une étiquette sur une image), soit parce que la tâche est découpable en micro-tâches élémentaires. Ce sont les systèmes de myriadisation du travail parcellisé, qui sont le cœur de cet article. Dans ce concept, il y a souvent rétribution [1], mais celle-ci peut-être non monétaire, comme dans certains GWAP (*Games with a purpose*) (von Ahn, 2006; Chamberlain *et al.*, 2008). La création d'Amazon Mechanical Turk en 2005 s'inscrit dans cette dernière catégorie de systèmes de m.t.p. avec rémunération, qui a été suivie par un grand nombre d'autres systèmes (Biewald, 2010), ceux-ci n'ayant pas acquis la même importance, en particulier en raison du nombre de personnes inscrites. Comme souvent pour les nouveaux usages issus du Web, on ressent à la fois une fascination pour la potentialité des m.t.p. et une méfiance en face de ces pratiques qui ne semblent pas avoir de réelles considérations pour le droit du travail. L'apparition des systèmes de myriadisation pose de nombreux problèmes, légaux, éthiques et philosophiques, abordés par exemple dans (Zittrain, 2008). Elle soulève d'importantes questions : qu'est-ce que le travail ? qu'est-ce qu'une rétribution juste ? un être humain est-il assimilable à un ordinateur ? Ces questions essentielles débordent largement à la fois le cadre d'un article de conférence et nos compétences. C'est pourquoi nous nous limiterons, autant que possible, aux problèmes précis que pose l'introduction de MTurk comme moyen de produire des ressources linguistiques, car nous jugeons cela à la fois urgent et crucial.

3 Amazon Mechanical Turk : légendes et réalité

Amazon Mechanical Turk (MTurk) permet, selon de nombreux auteurs dont le premier est Snow *et al.* (2008), de produire à peu de frais et rapidement des ressources linguistiques de qualité. Cette découverte est d'une telle importance pour la communauté qui manque toujours cruellement de moyens pour développer lesdites ressources, qu'elle a entraîné un important effet de mode. Ce phénomène est, nous allons le voir, ni totalement justifié, ni sans conséquences pour le développement futur de telles ressources. Par ailleurs, pour de nombreux chercheurs, les Turkers [2] utilisent MTurk comme un hobby, il n'est donc pas scandaleux de très mal les rémunérer. Nous allons voir ici que la situation est loin d'être aussi simplement idéale.

1. mais pas toujours, par exemple dans le système reCAPTCHA `http://www.google.com/recaptcha/learnmore`, où les CAPTCHAs proviennent de mots mal reconnus lors de la numérisation de Google books

2. Il est d'usage d'appeler les personnes effectuant des tâches au sein du « turc mécanique » des *Turkers*, et celles qui fournissent les tâches des *Requesters*.

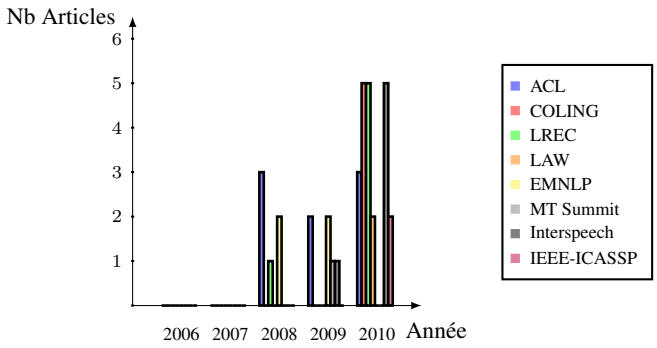


FIGURE 1 – Évolution de l'utilisation réelle de MTurk dans les publications TAL et parole

3.1 Etat des lieux

Créé en 2005, le système de m.t.p. MTurk est aujourd'hui de plus en plus utilisé pour la création ou la validation de ressources linguistiques pour le TAL (Traitement Automatique des Langues), et la plupart des conférences internationales du domaine ont vu la présentation de projets de recherche utilisant MTurk.

La figure 1, reprise de (Fort *et al.*, 2011), montre l'évolution rapide du phénomène. Elle comptabilise le nombre de publications dans les principales conférences internationales décrivant des expériences utilisant MTurk.

Afin de compléter cette étude, nous avons également réalisé une recherche plus globale, cette fois dans l'anthologie de l'ACL. [3] Cette recherche, effectuée le 5 novembre 2010, a ramené 124 résultats, dont, après filtrage manuel, 86 papiers utilisant effectivement MTurk (Fort *et al.*, 2011). Ces résultats incluent un atelier spécialisé, fournissant 35 des 86 publications, le NAACL-HLT 2010 Workshop on Amazon Mechanical Turk, dont l'existence même est le signe de l'importance grandissante de MTurk dans le domaine. Mentionnons enfin que certaines expériences relatées dans des articles ont utilisé MTurk sans le mentionner explicitement (Fort *et al.*, 2011). Ainsi, Kevin B. Cohen, co-auteur de l'article précité, a remarqué qu'un article (Biadsy *et al.*, 2008) dont il avait vu la présentation en conférence, utilisait MTurk et n'en faisait pas mention. Côté francophone, nous ne trouvons aucun article utilisant MTurk dans les actes des précédents TALN, ni dans les numéros publiés à ce jour de la revue TAL.

3.2 MTurk est un hobby pour les Turkers ?

Afin de pouvoir efficacement juger de l'éthique et de la légalité de MTurk, il est fondamental de pouvoir qualifier l'activité que mènent les Turkers lorsqu'ils effectuent des tâches dans MTurk. S'agit-il d'une activité bénévole, comme celle effectuée par les participants à Wikipedia ? Clairement non, lorsque l'on regarde la page d'accueil où MTurk met directement l'accent sur l'argent gagné. Peut-elle être assimilée à un hobby, la rétribution étant alors assimilable à un bonus ne correspondant pas à un salaire, comme cela est suggéré dans quelques articles (Novotney & Callison-Burch, 2010) ?

Un certain nombre d'études (Ross *et al.*, 2009, 2010; Ipeirotis, 2010b), fournissent, grâce à des questionnaires soumis aux Turkers *via* MTurk, des chiffres déclarés sur un certain nombre de facteurs socio-économiques (pays, âge, revenu, éducation...), sur la façon dont ils utilisent MTurk (nombre de tâches effectuées par semaine, revenu acquis, date d'entrée dans MTurk ...) [4] et dont ils qualifient leur activité. La motivation financière (déclarée) est minoritaire chez les Turkers américains (38%), mais majoritaire chez les Turkers indiens (69%). Si 60% des Turkers pensent que MTurk est un moyen utile de gagner de l'argent sur leur temps libre, ils ne sont que 30% à motiver leur participation par l'intérêt des tâches, et 20% (5% des travailleurs indiens) disent l'utiliser pour tuer le temps. Enfin, ils sont 20% (30% des Indiens) à dire que MTurk leur est nécessaire pour vivre, et à peu près le même pourcentage à dire que MTurk constitue leur principale source de revenus.

3. Association for Computational Linguistics, `http://www.aclweb.org/anthology/`

4. On pourra se reporter par exemple à (Adda & Mariani, 2010) pour un résumé de celles-ci. On y apprend par exemple (Ross *et al.*, 2010) que les Turkers en provenance d'Inde représentaient 5% à la fin de 2008, 36% fin 2009, plus de 50% en mai 2010 selon `http://blog.crowdflower.com/2010/05/amazon-mechanical-turk-survey/` et, selon (Biewald, 2010) sont responsables de plus de 60% de l'activité dans MTurk.

Un autre moyen de vérifier la nature de l'activité des Turkers est d'examiner la nature de la tâche. Certaines tâches actuellement proposées sur MTurk correspondent à de nouveaux usages (par exemple, des expériences artistiques comme `http://www.thesheepmarket.com/`), mais d'autres sont effectuées habituellement par des employés (ce qui peut donc assimiler MTurk à une forme de délocalisation sur le web, pour faire baisser les coûts de production), et constituent donc un travail. Tel est le cas des activités de transcription ou de traduction, qui sont (pour ce qui concerne les ressources les plus significatives) produites par des employés d'entités comme le LDC ou ELDA. Pour les 20% des Turkers qui passent plus de 15h par semaine sur MTurk (Adda & Mariani, 2010) et contribuent à hauteur de 80% des tâches, la durée d'activité est significative, et est assimilable à un travail.

Nous ne pouvons pas conclure de manière définitive sur la nature de l'activité de *tous* les Turkers, car la nature des tâches sur MTurk et la motivation des Turkers est composite. Cependant, nous pensons que pour les 20% des Turkers pour qui MTurk constitue la source principale de revenus, ainsi que pour les tâches assimilables à un travail (qui ont 8 chances sur 10 d'être effectuées par des Turkers travaillant plus de 15 heures par semaine), la nature de l'activité est assimilable à un travail.

3.3 MTurk permet de réduire les coûts ?

Dans la plupart des articles ayant utilisé MTurk, le faible coût de développement de la ressource est mis en avant. Il est vrai que MTurk permet de proposer des rétributions si faibles aux Turkers que le coût en est forcément réduit, par exemple 0.005$ pour transcrire un segment d'environ 5 secondes de parole téléphonique (Novotney & Callison-Burch, 2010). Il faut cependant nuancer ces chiffres. Tout d'abord, le coût effectif n'est pas toujours calculé avec rigueur. En effet, le temps de développement de l'interface et de mise en place des garde-fous est non nul (Callison-Burch & Dredze, 2010). De même, le coût de validation (Kaisser & Lowe, 2008) ou de développement (Xu & Klakow, 2010) post-MTurk permettant de compenser la mauvaise qualité des résultats (voir section 3.4) n'est généralement pas précisément évalué. Or, ces coûts supplémentaires ne sont jamais pris en compte dans le calcul final. De plus, certaines tâches peuvent se révéler plus coûteuses que prévues. Ainsi, si l'on ne trouve pas de Turkers pour faire la tâche, on peut être obligé d'augmenter la rémunération, comme Novotney & Callison-Burch (2010), qui, partant d'un coût très bas (5 dollars de l'heure transcrite), ont été obligés de le multiplier par 7 (37 dollars de l'heure) pour transcrire du coréen, par manque de Turkers qualifiés.

3.4 MTurk permet de produire une qualité équivalente ?

3.4.1 Limitations liées à la non expertise

Les Turkers étant des non experts, le Requester (fournisseur de tâches) doit découper les tâches complexes en tâches plus simples (HIT, Human Intelligence Task), afin de les rendre réalisables. Ce faisant, le chercheur est amené à faire des choix qui peuvent biaiser les résultats. Un exemple de ce type de biais est analysé dans (Cook & Stevenson, 2010), où les auteurs reconnaissent que le fait de ne proposer qu'une phrase par type d'évolution lexicale (amélioration ou péjoration) influence le résultat.

Plus grave encore que ces biais potentiels, certains chercheurs ont observé que, lorsque la complexité de la tâche augmente, la qualité produite sous MTurk est insuffisante. C'est notamment le cas dans (Bhardwaj *et al.*, 2010), qui démontre que, pour leur tâche de désambiguïsation lexicale, un petit nombre d'annotateurs bien formés produit de bien meilleurs résultats qu'un grand nombre de Turkers (le nombre étant supposé contrebalancer la non expertise). De ce point de vue, leurs résultats contredisent ceux de Snow *et al.* (2008) dont la tâche était semblable mais beaucoup plus simple.Cette même difficulté d'obtenir une qualité suffisante sur des tâches complexes apparaît dans (Gillick & Liu, 2010), qui démontre que l'évaluation par des non experts de systèmes de résumé automatique est « risquée », les Turkers n'étant pas capables d'obtenir des résultats comparables à ceux des experts. On retrouve ce problème de qualité dans de nombreux articles, dans lesquels les auteurs ont dû faire valider les résultats des Turkers par des spécialistes (des étudiants en thèse pour (Kaisser & Lowe, 2008)) ou leur faire subir un post-traitement assez lourd (Xu & Klakow, 2010). Enfin, la qualité du travail des annotateurs non experts varie considérablement (Tratz & Hovy, 2010).

Il existe également un effet « boule de neige » qui tend à surestimer la qualité signalée dans les articles : des chercheurs louent MTurk (Xu & Klakow, 2010), citant des recherches qui ont fait usage de MTurk, mais qui n'auraient pas donné de résultats utilisables sans une intervention postérieure plus ou moins lourde (Kaisser & Lowe, 2008). On pourrait en conclure que MTurk ne devrait être utilisé que pour des tâches simples, or, outre le

fait que son fonctionnement même induit d'importantes limitations (voir section 3.4.2), il est intéressant de noter que dans certains cas simples, des outils de TAL font d'ores et déjà mieux que les Turkers (Wais *et al.*, 2010).

3.4.2 Limitations liées au fonctionnement même de MTurk

Une première limitation est l'interface de MTurk. Tratz & Hovy (2010) notent ainsi que les limites de l'interface constituent « le premier et le plus important des défauts » de MTurk. Les auteurs regrettent par ailleurs l'impossibilité d'avoir la certitude que les Turkers participant à la tâche sont bien de langue maternelle anglaise. Cette impossibilité de connaître les capacités réelles des Turkers, notamment de connaître leur langue maternelle (bien que leurs adresses IP soient géolocalisables), est un problème bien réel. S'il est possible de mettre en place des tests préalables, qui, là encore, représentent un coût supplémentaire à prendre en compte, il est très facile de tricher (Callison-Burch & Dredze, 2010). Bien entendu, il est toujours possible de mettre en place des garde-fous (Callison-Burch & Dredze, 2010; Welinder *et al.*, 2010), mais, encore une fois, cela demande du temps et représente donc un coût supplémentaire que peu de Requesters sont prêts à investir. Ainsi, dans (Xu & Klakow, 2010), les auteurs ont identifié des spammeurs mais n'ont pas réussi à les éliminer. Pour certaines tâches, il peut s'avérer difficile de trouver des Turkers ayant les compétences nécessaires en raison de la complexité de la tâche (Gillick & Liu, 2010; Lambert *et al.*, 2010), ou de la langue à maîtriser (Novotney & Callison-Burch, 2010).

Par ailleurs, il ne faut pas négliger l'impact du paiement à la tâche, qui induit comme comportement logique de placer le nombre de tâches réalisées au-dessus de la qualité de la réalisation, et ce, quelle que soit la rétribution. Kochhar *et al.* (2010) sont ainsi arrivés à la conclusion qu'il valait mieux payer à l'heure (avec, bien sûr, des procédures de vérification et de justification du temps passé).

4 Quelques réflexions sur le statut de MTurk

4.1 Quel est le statut de l'activité dans MTurk ?

En obscurcissant la relation entre Turkers et Requesters, et entre les Turkers eux-mêmes, MTurk empêche de fait la possibilité de s'organiser en syndicats, de protester contre d'éventuelles pratiques douteuses des Requesters ou d'ester en justice. Au-delà des problèmes de droit du travail, il faut parler des problèmes des taxes et cotisations sociales : Amazon considère (selon l'accord de licence de MTurk) que les Turkers sont assimilables à des travailleurs indépendants,et donc qu'il leur incombe de payer toutes les taxes et charges afférant à leur activité. Étant donné la hauteur des rémunérations prises individuellement, il est parfaitement hypocrite de penser que cela est possible. Il est donc fortement probable que les Turkers ne déclarent pas ces revenus et ne cotisent pas non plus à une quelconque caisse de retraite ou de sécurité sociale. Il en va bien entendu de même pour les fournisseurs de travail. Les états sont donc privés d'un revenu légitime.

Il faut souligner également que la nature de la relation entre les trois partenaires, Turker, Requester et MTurk, vague pour le droit américain, est encore plus douteuse en regard du droit français. En effet, selon la législation française du travail, en dehors du fait que le travail à la tâche est illégal, soit il s'agit de travail salarié, qui serait, en l'occurrence, non déclaré par l'employeur, donc illégal (article 8200 et suivants du Code du Travail), soit il s'agit d'un rapport de prestation de service, dont le donneur d'ordre serait MTurk et le prestataire le Turker et, dans ce cas, le Turker doit être enregistré au registre du commerce (article 8222-1 du Code du Travail).

4.2 Le modèle économique de MTurk est-il fondé ?

Comme souligné dans la partie 2, lorsque l'on aborde pour la première fois MTurk, on est sidéré devant les conditions financières imposées aux Turkers, qui amènent à des rémunérations horaires ridiculement basses (inférieures à 2 dollars, soit 1,46 euros (Ross *et al.*, 2009; Ipeirotis, 2010b)). Ce coût fabuleusement bas correspond-il à une réalité économique saine (comme suggéré par Marge *et al.* (2010) et McGraw *et al.* (2010), qui considèrent que cette rétribution n'est qu'une sorte de bonus pour une tâche par ailleurs effectuée avec plaisir) ?

Nous l'avons vu dans la partie 3.2, l'assertion que les Turkers considèrent MTurk comme un hobby est fausse, au moins pour une partie significative d'entre eux. Dès lors, pourquoi, si cela constitue pour eux un travail, acceptent-ils un salaire horaire aussi bas ? La loi de l'offre et la demande n'est pas suffisante pour l'expliquer, tout d'abord

parce que le nombre réel de Turkers n'est pas si important (Fort *et al.*, 2011), ensuite, parce qu'il est souvent difficile de faire exécuter des tâches de grande taille en un temps limité pour un coût standard (Ipeirotis, 2010a).

Un fait peut nous mettre sur la voie d'une explication crédible : beaucoup d'articles (par exemple (Marge *et al.*, 2010)) soulignent que la qualité n'est pas liée au coût associé à chaque tâche. Cela est dû en particulier à la présence de spammeurs (c'est-à-dire de Turkers qui répondent au hasard ou en utilisant un système automatique), attirés par les tâches bien rémunérées, et qui sont en grand nombre dans le système MTurk, le système de réputation mis en place par Amazon étant notoirement facile à mettre en défaut [5]. Cela conduit à une situation semblable au « marché des tacots », décrit par le prix Nobel Georges Akerlof (Akerlof, 1970) : l'acheteur d'une voiture d'occasion prend en compte dans le prix qu'il offre le risque que le vendeur lui « fourgue » un tacot. Les vendeurs propriétaires d'une bonne voiture ne peuvent donc obtenir un bon prix et quittent le système, ce qui accroît en retour la défiance de l'acheteur, car cela augmente le risque d'acheter un tacot. La présence des spammeurs, de par le laxisme du système mis en œuvre par Amazon, conduit à une stabilisation à un prix très bas, les bons travailleurs quittant donc le système (70% des Turkers utilisent MTurk depuis moins de 6 mois (Ross *et al.*, 2009)).

De plus, il est indéniable qu'un certain nombre de Turkers utilisent MTurk comme moyen de divertissement : ceux-ci sont attirés par les tâches intéressantes, quelle que soit leur rémunération. Sur ces tâches, ils sont en concurrence avec les Turkers-travailleurs (qui, naturellement, souhaitent également faire des tâches intéressantes), ce qui conduit également à faire baisser le taux horaire moyen « acceptable », c'est-à-dire le taux horaire seuil, en dessous duquel un travailleur n'acceptera pas d'effectuer la tâche.

Dernier facteur qui tend à faire accepter un taux horaire finalement inacceptable : le travail à la tâche. Un Turker, fait bien sûr une relation entre la difficulté de la tâche et la rétribution, mais n'a pas une idée claire du salaire horaire avant de commencer à travailler. De plus, le travail à la tâche induit un comportement que l'on peut voir également dans des jeux en ligne ou à chaque fois que l'on effectue une tâche contre une rétribution absolue : la personne a tendance à regarder grossir son compteur d'argent, ou de points, et à se fixer des objectifs absolus, déconnectés d'un quelconque taux horaire : « aujourd'hui, je reste dans le système, jusqu'à ce que j'ai gagné 5 dollars ». Ce qui n'est bien sûr pas le meilleur moyen d'optimiser le taux horaire. Si le travail à la tâche est interdit en France, c'est bien pour empêcher que des travailleurs gagnent moins que le salaire minimum.

Comme le souligne (Ipeirotis, 2010c), les défauts de la plateforme MTurk remettent en cause sa viabilité à moyen terme, si elle n'évolue pas fondamentalement, en particulier sur les problèmes de rémunération et de systèmes de réputation fiables pour les Turkers et les Requesters.

4.3 Quelle est la situation par rapport à la propriété intellectuelle sur MTurk ?

Au regard du droit européen, la problématique de la propriété intellectuelle pour des données telles que l'on peut en produire via MTurk se pose le plus souvent en termes de protection associée aux bases de données, au sens de la directive européenne du 11 mars 1996. Cette directive, qui concerne les ensembles d'informations de toutes natures, offre une double protection : (1) par le droit d'auteur concernant la structure de la base, conditionné au fait qu'il y ait là une *création originale*, et (2) par un droit spécifique couvrant le contenu, droit proche de celui du droit d'auteur mais conditionné à la valeur économique des données (et non à leur originalité), au sens où ces données doivent avoir été obtenues grâce à un *investissement substantiel du point de vue qualitatif ou quantitatif*. Cette deuxième protection est indépendante du caractère public ou non des données, l'objet de la protection étant la base dans son ensemble (c'est-à-dire l'assemblage des données).

Dans le cas de MTurk, les droits semblent devoir être la propriété du Requester, soit en tant qu'auteur (pour les HIT eux-mêmes et ce qu'ils pourraient contenir, sauf lorsque sont utilisées des données elles-mêmes soumises à des droits d'auteurs propres, comme, par exemple, si l'on fait transcrire ou traduire des contenus existants), soit en tant qu'organisateur de ce qui est une *œuvre collective* [6] (pour les productions des Turkers).

Cela dit, il n'est pas clair, avec MTurk relevant des États-Unis et des Requesters et des Turkers relevant souvent d'autres pays, qu'il y ait un droit applicable, la situation ne semblant pas envisagée dans les traités internationaux.

5. http ://behind-the-enemy-lines.blogspot.com/2010/10/be-top-mechanical-turk-worker-you-need.html

6. L'œuvre collective est définie par l'article L. 113-2 alinéa 3 du Code de la propriété intellectuelle comme étant une œuvre *créée sur l'initiative d'une personne physique ou morale qui l'édite, la publie et la divulgue sous sa direction et son nom et dans laquelle la contribution personnelle des divers auteurs participant à son élaboration se fond dans l'ensemble en vue duquel elle est conçue, sans qu'il soit possible d'attribuer à chacun d'eux un droit distinct sur l'ensemble réalisé*. L'article 113-5 stipule alors que *L'œuvre collective est, sauf preuve contraire, la propriété de la personne physique ou morale sous le nom de laquelle elle est divulguée. […]* .

5 Alternatives existantes ou proposées

Comme indiqué ci-dessus, les objectifs principaux des développeurs de ressources linguistiques faisant appel à MTurk sont l'obtention de résultats de bonne qualité, à un faible coût et dans un délai très bref. Mais ces objectifs ne sont pas nécessairement faciles à atteindre avec MTurk, alors que des approches alternatives existent. Tout d'abord, bien qu'une comparaison systématique entre MTurk et des algorithmes état-de-l'art impliquant un échantillon varié de tâches liées au TAL reste à faire, il semble que certains auteurs aboutissent à la conclusion que les annotateurs automatiques déjà disponibles pour certaines tâches font aussi bien voire mieux que les Turkers (Wais *et al.*, 2010) : les outils automatiques peuvent faire mieux que les non experts. La réutilisation intelligente de ressources existantes peut également être une alternative simple et peu coûteuse à MTurk. Enfin, MTurk n'est qu'une des nombreuses possibilités de m.t.p.

5.1 Approches non supervisées et semi-supervisées pour le développement de ressources linguistiques à faible coût

La communauté du TAL s'intéresse depuis longtemps à des approches dites *non supervisées* d'apprentissage automatique, pour un large éventail de tâches parfois complexes. De la segmentation en mots à l'analyse syntaxique (Hänig, 2010) en passant par l'annotation morphosyntaxique (Goldwater & Griffiths, 2007), le développement de ressources lexicales (y compris de niveau sémantique ou pragmatique, cf. (Pak & Paroubek, 2010)) ou la catégorisation de documents, nombreuses sont les tâches pour lesquelles des techniques existent qui ne nécessitent aucune ressource préalable. Bien que les résultats obtenus soient souvent inférieurs aux résultats des approches supervisées (utilisant un corpus d'apprentissage) ou symboliques avancées (utilisant des ressources symboliques également coûteuses à développer), on peut penser que, pour certaines tâches, ils ne sont pas inférieurs à ce que l'on peut attendre de MTurk. C'est notamment le cas pour des tâches complexes comme l'analyse syntaxique.

Pour améliorer à faible coût la qualité des outils statistiques ainsi développés et/ou pour les faire correspondre à des modèles préexistants (par exemple, à un inventaire préétabli de catégories dans le cadre de l'annotation morphosyntaxique), il n'est pas forcément nécessaire de recourir à des techniques totalement supervisées. Une utilisation optimale d'un ensemble limité d'informations (annotations, ressources externes) peut donner de bons résultats : c'est le paradigme de l'*apprentissage semi-supervisé* (Abney, 2007). Dans le cas du développement de ressources linguistiques, on peut identifier deux types (non mutuellement exclusifs) de semi-supervision.

La première idée que l'on peut avoir est d'entraîner des modèles sur les quelques données annotées, puis d'annoter automatiquement les autres données : on peut alors choisir parmi les données annotées automatiquement celles pour lesquelles le modèle a un niveau de confiance optimal, et les considérer comme de nouvelles données annotées pour l'apprentissage d'un nouveau modèle, et ainsi de suite : c'est le *self-training*, utilisé en TAL depuis longtemps (Yarowsky, 1995). Cette idée peut être généralisée en utilisant deux modèles les plus différents possible, et à compléter les données d'apprentissage de l'un par les annotations automatiques les plus sûres produites par l'autre. C'est le *co-training* (Blum & Mitchell, 1998), qui cherche à éliminer au maximum les biais spécifiques à chaque modèle par la confrontation à un autre. À ce stade, on reste dans une situation où une annotation manuelle peu coûteuse sert de graine pour la construction successive, mais automatisée, de modèles qui vont en s'améliorant, jusqu'à obtenir des performances satisfaisantes. Si l'on accepte de continuer à annoter des données manuellement au cours des étapes de construction successive de modèles, on peut faire en sorte que soient choisies et présentées aux annotateurs les données telles que disposer d'une annotation de référence pour elles soit de nature à améliorer au mieux la qualité des outils. C'est l'idée au cœur de l'*active learning* (Cohn *et al.*, 1995).

La deuxième idée, combinable avec la première, consiste à utiliser au mieux des données annotées d'une façon moins complète que l'annotation visée. Par exemple, pour l'annotation morphosyntaxique, on peut disposer d'un lexique externe mais pas d'un corpus d'apprentissage : projeter le lexique sur le corpus correspond alors à une annotation ambiguë, qu'il faut désambiguïser (Smith & Eisner, 2005). Pour le développement de lexiques morphologiques, disposer d'une description formalisée de la morphologie de la langue permet l'utilisation de techniques efficaces de suggestion d'entrées lexicales (Sagot, 2005). De même, on peut chercher à exploiter un corpus partiellement parenthésé pour guider des modèles d'analyse syntaxique complets (Watson *et al.*, 2007).

5.2 Réutilisation de ressources existantes

Moins coûteux encore, la construction de ressources linguistiques peut se faire en réutilisant des ressources existantes. Considérons par exemple la tâche de détection d'entités nommées. Nothman *et al.* (2008) montrent qu'il est possible de transformer Wikipedia en une ressource annotée en entités nommées de large couverture et de très bonne qualité. De tels corpus ont pourtant été construits au moyen de MTurk, notamment sur des corpus non standard, en particulier médicaux (Yetisgen-Yildiz *et al.*, 2010), twitter (Finin *et al.*, 2010), e-mails (Lawson *et al.*, 2010). Naturellement, ces corpus sont très différents de ce que l'on peut obtenir au moyen de Wikipedia. Mais la taille des données extraites, ainsi que les caractéristiques de Wikipedia en tant que corpus, font que les détecteurs d'entités nommées entraînés sur un corpus construit à partir de Wikipedia tendent à avoir de très bons résultats lorsqu'ils sont utilisés sur d'autres types de corpus (Balasuriya *et al.*, 2009).

Il ne s'agit là que d'un exemple, mais nombreuses sont les ressources susceptibles de fournir des données de toutes natures : Wikipedia [7] et autres projets wiki, notamment wiktionary, corpus (annotés ou non, oraux ou textuels) et ressources lexicales (phonétiques, morphologiques, syntaxiques, sémantiques), pour peu qu'elles soient disponibles pour la communauté. Il s'agit ici d'un autre débat, sur lequel nous n'insisterons donc pas plus avant.

5.3 Développement collaboratif ou myriadisé de ressources linguistiques au-delà de MTurk

Toutes les méthodes alternatives décrites jusqu'à présent ont prouvé leur utilité et leur efficacité, mais elles requièrent des compétences expertes, ne serait-ce que pour concevoir et développer les outils automatiques, mais également pour effectuer, si besoin est, les tâches d'annotation optimisées. Il existe des méthodes de développement de ressources linguistiques qui ne font pas nécessairement appel à des experts, sans être pour autant touchées par tous les problèmes décrits pour MTurk. Il s'agit en particulier des approches collaboratives, des approches ludiques mais également de certaines plateformes de m.t.p., qui se sont données les moyens d'en éviter les écueils.

Les approches collaboratives de développement de ressources lexicales reposent sur la stratégie mise en place par le projet Wikipedia, les autres projets de la constellation Wikimedia, et d'autres types de wiki comme les *Semantic Wiki* (Freebase, OntoWiki…). Différents participants volontaires, experts ou non, enrichissent progressivement une même ressource, soit sous forme d'annotations soit sous forme de bases de données (lexicales, ontologiques…). C'est une première étape vers la m.t.p. : ici, le travail n'est pas parcellisé, et il n'est que faiblement myriadisé. Les annotations des uns sont « contrôlées » par les autres, et des divergences de vues entre différents participants se manifestent le plus souvent par des discussions, conduisant éventuellement à ce que l'administrateur tranche et décide. C'est ainsi qu'un très haut niveau de qualité peut être finalement atteint. Une des premières plateformes wiki dédiée au développement d'une ressource TAL est l'outil Serengeti, développé à l'Université de Bielefeld (Stürenberg *et al.*, 2007), à des fins d'annotation sémantique des textes. Cet outil est utilisé actuellement dans le cadre du projet AnaWiki (`http://www.anawiki.org`).

Toutefois, ces approches restent plus adaptées pour le développement de ressources de taille raisonnable avec une bonne qualité (*gold standard*). Elles sont moins indiquées pour le développement rapide de ressources à grande échelle. Une autre stratégie, qui repose également sur le Web, est d'attirer de grand nombres de non experts au moyen de jeux en ligne dits *ayant un but* (en anglais *games with a purpose*, ou *GWAP*). Cette idée, initiée par (von Ahn, 2006; von Ahn & Dabbish, 2008) avec le jeu en ligne ESP (`http://www.espgame.org/`) consiste à faire étiqueter des images par des joueurs qui rentrent en compétition : ceux-ci reçoivent des crédits lorsque leurs réponses coïncident avec celles d'autres joueurs [8]. ESP a connu un succès important en mobilisant 13 500 utilisateurs, créant 1,3 million d'étiquettes dans les premiers mois suivant son apparition sur la Toile. Cette idée a été par la suite déclinée pour divers types de tâches, y compris en TAL. Des exemples en sont le jeu *JeuxDeMots* (Lafourcade & Joubert, 2008, `http://www.lirmm.fr/jeuxdemots`), qui vise à collecter des relations entre mots, et son *alter ego* PtiClic (ibid., `http://www.lirmm.fr/pticlic`), qui vise à typer explicitement ces relations. Le jeu *PhraseDetective* (Chamberlain *et al.*, 2008, `http://www.phrasedetectives.org`), quant à lui, a pour objectif l'annotation de liens anaphoriques, tâche pourtant réputée complexe. L'idée est alors que l'on peut aussi utiliser le jeu pour former les utilisateurs à la tâche. *Phrase Detective* comprend ainsi une phase d'entraînement où l'on apprend la tâche au nouveau joueur, par le biais de tests de plus en plus durs basés sur un petit ensemble de données venant de corpus existants (annotés par des experts).

7. Il faut notamment citer le projet DBpedia (`http://dbpedia.org`), qui cherche à extraire des informations ontologiques structurées à partir de Wikipedia, constituant ainsi une ressource aux potentialités multiples et déjà largement utilisée.

8. Différentes procédures sont prévues pour exclure les utilisateurs malveillants, notamment le contrôle des adresses IP, la vérification aléatoires des étiquetages pour des réponses connues, etc. (von Ahn & Dabbish, 2008)

Toutefois, la frontière entre jeu de type GWAP et m.t.p. à la MTurk n'est pas nette. On ne peut pas distinguer facilement les GWAP, qui seraient plus ludiques, et MTurk, qui serait *stricto sensu* du travail : même contribuer à Wikipedia est un travail, certes bénévole. On ne peut pas non plus distinguer les GWAP de MTurk en tant qu'ils ne donneraient lieu à aucune récompense tangible : certains jeux en ligne sont des GWAP mais proposent des rémunérations non-monétaires (ainsi, PhraseDetective permet de gagner des bons à dépenser sur le site d'achat en ligne Amazon). Enfin, on ne peut pas distinguer la m.t.p. à la MTurk par le caractère « éthique » des premiers. Il existe en effet des alternatives à MTurk pour développer des ressources linguistiques dans le paradigme de la m.t.p., tout en évitant les écueils évoqués tout au long de cet article.

Pour le recueil de données langagières, en particulier pour les langues peu dotées, des alternatives à MTurk semblent être plus appropriées. Ainsi, l'utilisation d'applications sur des téléphones portables de nouvelle génération est un moyen plus efficace de pouvoir accéder à toute une population. Hughes *et al.* (2010) ont ainsi embauché des locuteurs locaux et leur ont prêté des téléphones sur lesquels tournait une application dédiée. Les auteurs ont ainsi recueilli 3 000 heures en 17 langues. Un exemple de m.t.p. éthique est Samasource, une ONG qui utilise ce type de méthode pour faire effectuer des tâches [9] à des personnes réellement nécessiteuses formées et rémunérées équitablement selon des barèmes dépendant du pays. Il s'agit là d'une alternative éthique à l'utilisation de MTurk qui permet également de tirer parti des avantages de la m.t.p.

5.4 Optimiser le coût de l'annotation manuelle : pré-annotation et interfaces dédiées

Indépendamment de la façon dont on s'en sert, l'annotation manuelle par des experts peut être considérablement accélérée voire améliorée au moyen d'outils d'annotation automatique utilisés comme pré-annotateurs. Par exemple, Fort & Sagot (2010) ont démontré que, dans le cas de l'étiquetage morphosyntaxique, une préannotation, même de piètre qualité et donc développable à faible coût, permet d'améliorer très largement le temps et la qualité des annotations manuelles. Ainsi, les auteurs ont montré que 50 phrases annotées à la main sans pré-annotation, ce qui prend environ 40 minutes [10], permettent de construire un préannotateur tel que la vitesse de l'annotation manuelle par un expert est quasiment identique à ce que l'on obtient avec un préannotateur de niveau état-de-l'art, c'est-à-dire que l'on peut construire un corpus complet de taille standard (10 000 phrases) en environ 6 000 minutes (100 heures). Des annotateurs experts et coûteux, pour peu que leur travail soit préparé puis utilisé de façon optimale, permettent donc le développement de ressources de très bonne qualité à un coût qui reste limité. À l'inverse, sur cette tâche d'apparence simple, des Turkers seraient bien en peine de suivre correctement un guide d'annotation détaillé, nécessairement complexe s'il est linguistiquement sérieux.

Par ailleurs, les remarques de Tratz & Hovy (2010) mentionnées ci-dessus concernant les limitations des interfaces déployables dans MTurk s'appliquent de manière générale. L'expérience acquise, par exemple, dans le développement de corpus annotés syntaxiquement ou sémantiquement montre que la rapidité et la qualité l'annotation, de quelque nature qu'elle soit, est fortement influencée par l'interface d'annotation elle-même (cf. par exemple (Erk *et al.*, 2003)). Il y a donc là aussi matière à accélérer et améliorer toute étape d'annotation manuelle, au point qu'une interface adaptée à une tâche donnée pourrait permettre de réduire les coûts dans des proportions comparables à celles obtenues par l'utilisation de MTurk, sans en présenter les inconvénients.

6 Conclusion et perspectives

MTurk illustre la complexité et la difficulté d'appréhender les relations (commerciales, de travail et autres) dans les nouveaux modes d'activités sur Internet. Les chercheurs qui ont utilisé MTurk l'ont fait souvent de bonne foi, par manque de moyens financiers, pour produire plus de données et les redistribuer à la communauté. Pour ceux qui ont eu des doutes sur de possibles problèmes d'éthique et de droit du travail, une recherche superficielle les a convaincus que MTurk est une sorte d'avatar de Wikipedia, et que les Turkers sont motivés surtout par le plaisir d'effectuer des tâches amusantes.

Nous pensons avoir montré que MTurk n'est pas une panacée et que d'autres solutions existent aujourd'hui pour réduire les coûts de construction de ressources linguistiques de qualité, tout en respectant ceux qui travaillent sur ces ressources et en tirant un meilleur parti de leurs compétences. Car derrière le débat autour de MTurk se trouve

9. Comme traduire des SMS en créole, lors du tremblement de terre à Haïti afin de permettre aux secours internationaux d'aller à leur secours, en liaison avec le site CrowdFlower. `http://www.samasource.org/haiti/`.

10. Les estimations proposées dans ce paragraphe, très grossières, reposent sur celles de (Fort & Sagot, 2010).

finalement la question de la considération due aux annotateurs, aux traducteurs, aux spécialistes de la transcription.

Nous aimerions, en conclusion, aller au-delà des faits actuels et mettre l'accent sur les conséquences à plus ou moins long terme de cette « mode ». En effet, sous la pression de ce type de systèmes à bas coût, les agences de moyens pourraient bientôt être plus réticentes à financer des projets de développement de ressources linguistiques à des coûts « normaux » (ou plutôt réalistes). Le coût à la MTurk deviendrait alors une norme de fait et nous n'aurions plus le choix de nos méthodes de développement.

Nous avons vu, dans la partie 5.3, qu'un système de m.t.p. peut permettre de faire produire des tâches rémunérées en préservant l'éthique, cela peut même être une chance pour des personnes qui ne peuvent se trouver sur le marché du travail, de par leur isolement, leur handicap, etc ; mais cela nécessite un encadrement légal strict afin de s'assurer que ce système n'est pas une remise en cause des droits des travailleurs. On peut penser à moyen terme au développement d'une plateforme m.t.p. opérée par les acteurs de recherche au niveau européen, et un guide des bonnes pratiques concernant l'utilisation des m.t.p., comme cela se fait dans d'autres secteurs de recherches, par exemple en sciences sociales. Mais ces solutions risquent de ne pas freiner le développement actuel de l'utilisation de MTurk, au nom du "pragmatisme" et de la concurrence avec les équipes (par exemple) outre-Atlantique ; c'est pourquoi nous proposons la création d'un label de qualité et d'éthique, qui pourrait être décerné aux ressources par les associations savantes concernées, l'ATALA [11] pour le TAL et l'AFCP [12] pour la parole. Les questions d'éthique sont dès à présent un critère de sélection pour les projets européens, ce label permettrait de préciser le statut des ressources comme critère de sélection pour l'ensemble des agences de moyens, tout en valorisant les bonnes pratiques de développement.

Remerciements

Ce travail a été réalisé en partie dans le cadre du programme Quaero, financé par OSEO, agence nationale de valorisation de la recherche, et dans celui du projet ANR EDyLex (ANR-09-CORD-008).

Références

ABNEY S. (2007). *Semisupervised Learning for Computational Linguistics*. Chapman & Hall/CRC, 1ère edition.

ADDA G. & MARIANI J. (2010). Language resources and amazon mechanical turk : legal, ethical and other issues. In *LISLR2010, "Legal Issues for Sharing Language Resources workshop", LREC2010*.

AKERLOF G. A. (1970). The market for 'lemons' : Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, **84**(3), 488–500.

BALASURIYA D., RINGLAND N., NOTHMAN J., MURPHY T. & CURRAN J. R. (2009). Named entity recognition in wikipedia. In *People's Web '09 : Proceedings of the 2009 Workshop on The People's Web Meets NLP*, p. 10–18, Morristown, NJ, USA : Association for Computational Linguistics.

BHARDWAJ V., PASSONNEAU R., SALLEB-AOUISSI A. & IDE N. (2010). Anveshan : A tool for analysis of multiple annotators' labeling behavior. In *Proceedings of The fourth linguistic annotation workshop (LAW IV)*, Uppsala, Suède.

BIADSY F., HIRSCHBERG J. & FILATOVA E. (2008). An unsupervised approach to biography production using Wikipedia. In *Proceedings of ACL 2008*, p. 807–815 : Association for Computational Linguistics.

BIEWALD L. (2010). Better crowdsourcing through automated methods for quality control. *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*.

BLUM A. & MITCHELL T. (1998). Combining labeled and unlabeled data with co-training. In *COLT : Proceedings of the Workshop on Computational Learning Theory* : Morgan Kaufmann Publishers.

CALLISON-BURCH C. & DREDZE M. (2010). Creating speech and language data with amazon's mechanical turk. In *CSLDAMT '10 : Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Morristown, NJ, USA : Association for Computational Linguistics.

CHAMBERLAIN J., POESIO M. & KRUSCHWITZ U. (2008). Phrase Detectives : a Web-based Collaborative Annotation Game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*, Graz.

11. http://www.atala.org/
12. http://www.afcp-parole.org/

COHN D. A., GHAHRAMANI Z. & JORDAN M. I. (1995). Active learning with statistical models. In G. TESAURO, D. TOURETZKY & T. LEEN, Eds., *Advances in Neural Information Processing Systems*, volume 7, p. 705–712 : The MIT Press.

COOK P. & STEVENSON S. (2010). Automatically identifying changes in the semantic orientation of words. In N. C. C. CHAIR), K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER & D. TAPIAS, Eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* : European Language Resources Association (ELRA).

ERK K., KOWALSKI A. & PADO S. (2003). The salsa annotation tool. In D. DUCHIER & G.-J. M. KRUIJFF, Eds., *Proceedings of the Workshop on Prospects and Advances in the Syntax/Semantics Interface*, Nancy, France.

FININ T., MURNANE W., KARANDIKAR A., KELLER N., MARTINEAU J. & DREDZE M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, p. 80–88, Stroudsburg, PA, USA : Association for Computational Linguistics.

FORT K., ADDA G. & COHEN K. B. (2011). Amazon mechanical turk : Gold mine or coal mine ? *Computational Linguistics (editorial)*, **37**(2).

FORT K. & SAGOT B. (2010). Influence of Pre-annotation on POS-tagged Corpus Development. In *Proc. of the Fourth ACL Linguistic Annotation Workshop*, Uppsala, Suède.

GILLICK D. & LIU Y. (2010). Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, p. 148–151, Stroudsburg, PA, USA : Association for Computational Linguistics.

GOLDWATER S. & GRIFFITHS T. (2007). A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of ACL*, Prague, République tchèque.

HÄNIG C. (2010). Improvements in unsupervised co-occurrence based parsing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, p. 1–8, Stroudsburg, PA, USA : Association for Computational Linguistics.

HUGHES T., NAKAJIMA K., HA L., VASU A., MORENO P. & LEBEAU M. (2010). Building transcribed speech corpora quickly and cheaply for many languages. In *Proceedings of Interspeech*, p. 1914–1917.

IPEIROTIS P. (2010a). Analyzing the amazon mechanical turk marketplace. CeDER Working Papers, http ://hdl.handle.net/2451/29801. CeDER-10-04.

IPEIROTIS P. (2010b). Demographics of mechanical turk. CeDER Working Papers, http ://hdl.handle.net/2451/29585. CeDER-10-01.

IPEIROTIS P. (2010c). A plea to amazon : Fix mechanical turk ! http ://behind-the-enemy-lines.blogspot.com/2010/10/plea-to-amazon-fix-mechanical-turk.html.

KAISSER M. & LOWE J. B. (2008). Creating a research collection of question answer sentence pairs with amazon's mechanical turk. In *Proceedings of the International Language Resources and Evaluation (LREC-2008)*.

KOCHHAR S., MAZZOCCHI S. & PARITOSH P. (2010). The anatomy of a large-scale human computation engine. In *Proceedings of Human Computation Workshop at the 16th ACM SIKDD Conference on Knowledge Discovery and Data Mining, KDD 2010*, Washington D.C.

LAFOURCADE M. & JOUBERT A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles*, p. 657–666.

LAMBERT B., SINGH R. & RAJ B. (2010). Creating a linguistic plausibility dataset with non-expert annotators. In *Proceedings of Interspeech*, p. 1906–1909.

LAWSON N., EUSTICE K., PERKOWITZ M. & YETISGEN-YILDIZ M. (2010). Annotating large email datasets for named entity recognition with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, p. 71–79, Stroudsburg, PA, USA : Association for Computational Linguistics.

MARGE M., BANERJEE S. & RUDNICKY A. I. (2010). Using the amazon mechanical turk for transcription of spoken language. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, p. 5270–5273, Dallas, TX.

MCGRAW I., YING LEE C., HETHERINGTON L., SENEFF S. & GLASS J. (2010). Collecting voices from the cloud. In *Proceedings of the International Language Resources and Evaluation (LREC-2010)*, p. 1576–1583.

NOTHMAN J., CURRAN J. R. & MURPHY T. (2008). Transforming Wikipedia into Named Entity Training Data. In *Proceedings of the Australian Language Technology Workshop*, p. 124–132.

NOVOTNEY S. & CALLISON-BURCH C. (2010). Cheap, fast and good enough : automatic speech recognition with non-expert transcription. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, p. 207–215, Stroudsburg, PA, USA : Association for Computational Linguistics.

PAK A. & PAROUBEK P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, La Valette, Malte : European Language Resources Association (ELRA).

ROSS J., IRANI L., SILBERMAN M. S., ZALDIVAR A. & TOMLINSON B. (2010). Who are the crowdworkers ? : shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, CHI EA '10, p. 2863–2872, New York, NY, USA : ACM.

ROSS J., ZALDIVAR A., IRANI L. & TOMLINSON B. (2009). Who are the turkers ? worker demographics in amazon mechanical turk. Social Code Report 2009-01, http ://www.ics.uci.edu/ jwross/pubs/SocialCode-2009-01.pdf.

SAGOT B. (2005). Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658 ((c) Springer-Verlag), Proceedings of TSD'05*, p. 156–163, Karlovy Vary, République tchèque.

SMITH N. & EISNER J. (2005). Contrastive estimation : Training log-linear models on unlabeled data. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 354–362, nn Arbor, Michigan, USA.

SNOW R., O'CONNOR B., JURAFSKY D. & NG. A. Y. (2008). Cheap and fast - but is it good ? evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*, p. 254–263.

STÜRENBERG M., GOECKE D., DIE-WALD N., CRAMER I. & MEHLER A. (2007). Web-based annotation of anaphoric relations and lexical chains. In *ACL Workshop on Linguistic Annotation Workshop (LAW)*, Prague, République tchèque.

TRATZ S. & HOVY E. (2010). A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 678–687, Uppsala, Suède : Association for Computational Linguistics.

VON AHN L. (2006). Games with a purpose. *IEEE Computer Magazine*, p. 96–98.

VON AHN L. & DABBISH L. (2008). General techniques for designing games with a purpose. *Communications of the ACM*, p. 58–67.

WAIS P., LINGAMNENI S., COOK D., FENNELL J., GOLDENBERG B., LUBAROV D., MARIN D. & SIMONS H. (2010). Towards building a high-qualityworkforce with mechanical turk. In *Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS)*.

WATSON R., BRISCOE T. & CARROLL J. (2007). Semi-supervised training of a statistical parser from unlabeled partially-bracketed data. In *Proceedings of the 10th International Conference on Parsing Technologies*, IWPT '07, p. 23–32, Stroudsburg, PA, USA : Association for Computational Linguistics.

WELINDER P., BRANSON S., BELONGIE S. & PERONA P. (2010). The multidimensional wisdom of crowds. In *Neural Information Processing Systems Conference (NIPS)*.

XU F. & KLAKOW D. (2010). Paragraph acquisition and selection for list question using amazon's mechanical turk. In *Proceedings of the International Language Resources and Evaluation (LREC-2010)*, p. 2340–2345, La Valette, Malte.

YAROWSKY D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, p. 189–196, Cambridge, MA.

YETISGEN-YILDIZ M., SOLTI I., XIA F. & HALGRIM S. R. (2010). Preliminary experience with amazon's mechanical turk for annotating medical named entities. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, p. 180–183, Stroudsburg, PA, USA : Association for Computational Linguistics.

ZITTRAIN J. (2008). Ubiquitous human computing. *Phil. Trans. R. Soc. A 28*, **366**(1881), 3813–3821.

TALN 2011, Montpellier, 27 *juin* –1er *juillet 2011*

Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue

Bo Li[1] Eric Gaussier[1] Emmanuel Morin[2] Amir Hazem[2]
(1) Université Grenoble I, LIG UMR 5217
(2) LINA, UMR 6241,Université de Nantes
{bo.li,eric.gaussier}@imag.fr, {emmanuel.morin,amir.hazem}@univ-nantes.fr

Résumé. Nous étudions dans cet article le problème de la comparabilité des documents composant un corpus comparable afin d'améliorer la qualité des lexiques bilingues extraits et les performances des systèmes de recherche d'information interlingue. Nous proposons une nouvelle approche qui permet de garantir un certain degré de comparabilité et d'homogénéité du corpus tout en préservant une grande part du vocabulaire du corpus d'origine. Nos expériences montrent que les lexiques bilingues que nous obtenons sont d'une meilleure qualité que ceux obtenus avec les approches précédentes, et qu'ils peuvent être utilisés pour améliorer significativement les systèmes de recherche d'information interlingue.

Abstract. We study in this paper the problem of enhancing the comparability of bilingual corpora in order to improve the quality of bilingual lexicons extracted from comparable corpora and the performance of cross-language information retrieval (CLIR) systems. We introduce a new method for enhancing corpus comparability which guarantees a certain degree of comparability and homogeneity, and still preserves most of the vocabulary of the original corpus. Our experiments illustrate the well-foundedness of this method and show that the bilingual lexicons obtained are of better quality than the lexicons obtained with previous approaches, and that they can be used to significantly improve CLIR systems

Mots-clés : Corpus comparables, comparabilité, lexiques bilingues, recherche d'information interlingue.

Keywords: Comparable corpora, comparability, bilingual lexicon, cross-language information retrieval.

1 Introduction

Les lexiques bilingues sont une ressource incontournable dans différentes applications multilingues du traitement automatique des langues comme la traduction automatique (Och & Ney, 2003) ou la recherche d'information interlingue (Ballesteros & Croft, 1997). Dans la mesure où la constitution manuelle de lexiques bilingues est une tâche coûteuse et qu'il est difficilement envisageable de développer un lexique pour chaque domaine d'étude, les recherches se sont intéressées à l'extraction automatique de ces lexiques à partir de corpus. Dans la mesure où la plupart des corpus bilingues existants sont par essence comparables, c'est-à-dire qu'ils regroupent des documents dans des langues différentes traitant du même domaine sur la même période sans être en relation de traduction, différents travaux s'intéressent à l'extraction de lexiques bilingues à partir de corpus comparables (Fung & McKeown, 1997; Fung & Yee, 1998; Rapp, 1999; Déjean *et al.*, 2002; Gaussier *et al.*, 2004; Robitaille *et al.*, 2006; Morin *et al.*, 2007; Garera *et al.*, 2009; Yu & Tsujii, 2009; Shezaf & Rappoport, 2010, entre autres). Le

socle commun à ces travaux est de reposer sur une hypothèse de distribution qui postule que les mots qui sont en correspondance de traduction sont susceptibles d'apparaître dans des contextes identiques pour des langues différentes. En s'appuyant sur cette hypothèse fondatrice, les chercheurs ont aussi cherché à identifier de meilleures représentations pour le contexte des mots de même qu'à utiliser différentes méthodes pour mettre en correspondance les mots entre différentes langues toujours en s'appuyant sur cette représentation du contexte. Ces méthodes semblent avoir atteint leur limite en termes de performance et les améliorations les plus récentes concernent plus le cadre d'évaluation des ces approches, plus contraint et limité (Yu & Tsujii, 2009), ou encore le traitement de langues spécifiques (Shezaf & Rappoport, 2010). Plus récemment, et en s'éloignant des approches traditionnelles, Li & Gaussier (2010) ont proposé une approche basée sur l'amélioration de la comparabilité des corpus comme préalable à l'extraction lexicale bilingue. Cette approche postule qu'il ne sert à rien d'essayer d'extraire des lexiques bilingues à partir d'un corpus avec un faible degré de comparabilité puisque la probabilité de trouver des traductions d'un mot donné sera faible dans une telle situation. Notre étude se situe dans la même veine que cette précédente approche et vise dans un premier temps à améliorer la comparabilité d'un corpus donné, tout en préservant une large part de son vocabulaire. Néanmoins, nous nous différencions de ce précédent travail en montrant qu'il est possible de garantir un certain degré d'*homogénéité* du corpus amélioré, et que celle-ci induit une amélioration significative de la qualité du corpus résultant et des lexiques bilingues extraits. En outre, nous montrons que les lexiques extraits avec notre approche améliorent de manière manifeste les résultats d'un système de recherche d'information interlingue, même lorsque ces lexiques sont issus d'un corpus différent de la collection interrogée.

2 Améliorer le degré de comparabilité d'un corpus

Nous commençons par donner dans cette partie la mesure de comparabilité que nous utilisons, avant de décrire un algorithme permettant d'améliorer la comparabilité d'un corpus donné. Nous fournissons également une preuve du bien-fondé de notre algorithme, ainsi qu'une approximation conduisant à une implantation efficace. Pour des raisons pratiques, notre discussion se fera sur la base du couple de langues anglais-français.

2.1 Mesure de comparabilité

Afin de mesurer le degré de comparabilité d'un corpus bilingue, nous utilisons la mesure développée dans (Li & Gaussier, 2010) : étant donné un corpus comparable $\mathcal{P}$ constitué d'une partie anglaise $\mathcal{P}_e$ et d'une partie française $\mathcal{P}_f$, le degré de comparabilité de $\mathcal{P}$ est défini comme l'espérance de trouver la traduction d'un mot du vocabulaire source (respectivement cible) dans le vocabulaire cible (respectivement source). Soit σ une fonction indiquant si une traduction de l'ensemble des traductions possibles $\mathcal{T}_w$ du mot w se trouve dans le vocabulaire $\mathcal{P}^v$ du corpus $\mathcal{P}$, c'est-à-dire :

$$\sigma(w, \mathcal{P}) = \begin{cases} 1 & \text{si } \mathcal{T}_w \cap \mathcal{P}^v \neq \emptyset \\ 0 & \text{sinon} \end{cases}$$

et soit $\mathcal{D}$ un dictionnaire bilingue dont le vocabulaire anglais (respectivement français) est noté $\mathcal{D}_e$ (respectivement $\mathcal{D}_f$). La mesure du degré de comparabilité M est définie par :

$$M(\mathcal{P}_e, \mathcal{P}_f) = \frac{\sum_{w \in \mathcal{P}_e \cap \mathcal{D}_e} \sigma(w, \mathcal{P}_f) + \sum_{w \in \mathcal{P}_f \cap \mathcal{D}_f} \sigma(w, \mathcal{P}_e)}{\#_w(\mathcal{P}_e \cap \mathcal{D}_e) + \#_w(\mathcal{P}_f \cap \mathcal{D}_f)}$$

où $\#_w(\mathcal{P})$ représente le nombre de mots différents présents dans $\mathcal{P}$. Comme on peut le voir d'après cette définition, M mesure la proportion de mots source et cible dont une traduction est présente dans le vocabulaire cible et source de $\mathcal{P}$. Pour des raisons qui deviendront claires plus tard, nous utiliserons aussi des mesures partielles où seuls les vocabulaires français ou anglais sont considérés. Ainsi, la proportion de mots anglais traduits sera notée M_{ef}, définie par : $\frac{\sum_{w \in \mathcal{P}_e \cap \mathcal{D}_e} \sigma(w, \mathcal{P}_f)}{\#_w(\mathcal{P}_e \cap \mathcal{D}_e)}$. La mesure M_{fe} est définie de la même façon.

2.2 Classer les documents pour une meilleure comparabilité

L'hypothèse distributionnelle sous-tendant l'extraction de lexiques bilingues est d'autant plus valide que les documents dans les différentes langues couvrent des thématiques proches, car les auteurs ont alors tendance à puiser dans le même vocabulaire (voir (Morin *et al.*, 2007) pour une analyse reliée). En d'autres termes, si un corpus couvre un nombre limité de thématiques, il est plus à même de contenir une information répétée et cohérente qui pourra être exploitée au mieux pour l'extraction de lexiques bilingues. Le terme *homogénéité* rend compte de ce phénomène et nous dirons, de façon informelle, qu'un corpus est homogène s'il couvre un nombre limité de thématiques. Nous conjecturons ici que si l'on peut garantir un certain degré d'homogénéité, en plus d'un certain degré de comparabilité, alors les lexiques bilingues extraits seront de meilleure qualité. Comme nous le verrons, cette conjecture sera validée par les expériences menées. De façon à garantir un certain degré d'homogénéité, nous nous appuyons sur des techniques de classification non supervisée (*clustering*). Nous utilisons ici des techniques de classification agglomérative ascendante, mais toute autre technique, pour peu qu'elle dispose d'une procédure de filtrage adaptée, peut être utilisée.

2.2.1 Algorithme de classification bilingue

L'ensemble du processus permettant de construire, à partir d'un corpus donné, un corpus plus homogène et de plus fort degré de comparabilité peut être résumé par les étapes suivantes :

1. À partir de la mesure de similarité, définie en 2.2.2 et fondée sur la mesure de comparabilité présentée ci-dessus, et de l'ensemble des documents anglais et français du corpus originel $\mathcal{P}$, construire les dendrogrammes en suivant les étapes classiques de la classification agglomérative ascendante ;
2. Filtrer les dendrogrammes en ne retenant que les classes les plus profondes (voir ci-dessous) ;
3. Fusionner les classes retenues pour former un nouveau corpus $\mathcal{P}_H$, qui contient une sous-partie homogène et fortement comparable de $\mathcal{P}$;
4. Répéter les étapes ci-dessus pour enrichir la partie restante de $\mathcal{P}$ (partie qui sera notée $\mathcal{P}_L$, $\mathcal{P}_L = \mathcal{P} \setminus \mathcal{P}_H$) avec des documents extraits d'autres corpus.

Les trois premières étapes sont détaillées dans l'algorithme 1, où CAA signifie Classification Agglomérative Ascendante. Comme on peut le remarquer, seul $\mathcal{P}$ est utilisé pour construire $\mathcal{P}_H$, à travers des étapes de classification et de filtrage. Ainsi, l'algorithme 1 vise à extraire de $\mathcal{P}$ une sous-partie fortement comparable et homogène. Une fois cela réalisé, c'est-à-dire une fois que $\mathcal{P}$ a été exploité, il est nécessaire de recourir à des ressources externes si l'on veut construire un corpus fortement comparable à partir de $\mathcal{P}_L$ (qui est la partie restante de $\mathcal{P}$). Pour cela, deux nouveaux corpus comparables sont considérés dans l'étape 4 du processus global : le premier consiste en la partie anglaise de $\mathcal{P}_L$ et la partie française d'un autre corpus $\mathcal{P}_T$; le second consiste en la partie française de $\mathcal{P}_L$ et la partie anglaise de $\mathcal{P}_T$. Les deux sous-parties fortement comparables et homogènes obtenues à partir de ces deux corpus sont alors ajoutées à $\mathcal{P}_H$ pour constituer le corpus final. L'utilisation de la classification agglomérative ascendante et du filtrage associé garantit que le corpus final est homogène. La propriété 1 que nous présentons plus

Algorithm 1: Algorithme de classification bilingue

Entrée :

Ensemble $\mathcal{U}$ de tous les documents anglais et français de $\mathcal{P}$

Réel positif θ (seuil de profondeur)

Sortie :

$\mathcal{P}_H$, sous-partie fortement comparable et homogène de $\mathcal{P}$

```
Initialiser P_H = ∅ ;
CAA(U) → ensemble S de dendrogrammes
for chaque dendrogramme T de S do
    m ← profondeur maximale de T ;
    for tous les nœuds n de T do
        if profondeur(n) ≥ m · θ then
            Ajouter tous les documents sous le nœud n à P_H ;
        end if
    end for
end for
Supprimer les doublons de P_H ;
return P_H ;
```

loin établit que ce corpus est fortement comparable. Mais avant de voir en détail cette propriété, nous introduisons la mesure de similarité utilisée.

2.2.2 Mesure de similarité

Imaginons deux classes de documents bilingues $\mathcal{C}_1$ et $\mathcal{C}_2$. Pour la tâche d'extraction de lexiques bilingues, ces deux classes sont similaires et devraient être regroupées si leur combinaison permet de compléter le contenu de chacune des classes prise isolément, ou, en d'autres termes, si la partie anglaise $\mathcal{C}_1^e$ de $\mathcal{C}_1$ et la partie française $\mathcal{C}_1^f$ de $\mathcal{C}_1$ sont comparables à leur contrepartie dans l'autre classe (respectivement la partie française $\mathcal{C}_2^f$ de $\mathcal{C}_2$ et la partie anglaise $\mathcal{C}_2^e$ de $\mathcal{C}_2$) [1]. Ceci conduit à la mesure de similarité suivante pour $\mathcal{C}_1$ et $\mathcal{C}_2$:

$$sim(\mathcal{C}_1, \mathcal{C}_2) = \beta M(\mathcal{C}_1^e, \mathcal{C}_2^f) + (1-\beta) M(\mathcal{C}_2^e, \mathcal{C}_1^f) \quad (1)$$

où $\beta\,(0 \leq \beta \leq 1)$ est un poids qui permet de contrôler l'importance de chacune des deux parties ($\mathcal{C}_1^e$, $\mathcal{C}_2^f$) et ($\mathcal{C}_2^e$, $\mathcal{C}_1^f$). De façon intuitive, on aimerait donner plus de poids dans cette combinaison à la partie la plus importante, car elle contient plus d'information. Si nous utilisons le nombre de paires de documents anglais-français pour quantifier cette information, le poids β peut être défini comme la proportion de paires de documents dans ($\mathcal{C}_1^e$, $\mathcal{C}_2^f$) sur l'ensemble des paires de documents dans le corpus fusionné :

$$\beta = \frac{\#_d(\mathcal{C}_1^e) \cdot \#_d(\mathcal{C}_2^f)}{\#_d(\mathcal{C}_1^e) \cdot \#_d(\mathcal{C}_2^f) + \#_d(\mathcal{C}_2^e) \cdot \#_d(\mathcal{C}_1^f)}$$

où $\#_d(\mathcal{C})$ représente le nombre de documents dans $\mathcal{C}$. Dans la mesure où les classes sont tout d'abord formées de documents anglais et français isolés, la mesure de similarité correspond à un score de comparabilité normalisé entre les corpus anglais et français qui forment la nouvelle classe. Cependant, cette mesure ne tient pas compte des

1. Dans la mesure où $\mathcal{C}_1$ et $\mathcal{C}_2$ sont des classes, leurs parties anglaise et française sont comparables par construction.

longueurs relatives des corpus anglais et français, qui ont pourtant un impact sur la qualité des corpus bilingues extraits. Si une contrainte de type 1-1 (c'est-à-dire imposant à chaque classe de contenir le même nombre de documents anglais et français) est trop forte, se reposer sur des classes par trop déséquilibrées n'est pas non plus souhaitable. Nous introduisons donc une nouvelle fonction ϕ qui a pour but de pénaliser les classes pour lesquelles les nombres de documents anglais et français sont trop différents :

$$\phi(\mathcal{C}) = \frac{1}{(1 + log(1 + \gamma \frac{|\#_d(\mathcal{C}^e) - \#_d(\mathcal{C}^f)|}{\min(\#_d(\mathcal{C}^e)), \#_d(\mathcal{C}^f))})} \tag{2}$$

avec $\gamma \in \mathbb{R}^+$. Cette fonction de pénalité fournit une nouvelle mesure de similarité sim_l qui est celle utilisée dans l'algorithme 1 :

$$sim_l(\mathcal{C}_1, \mathcal{C}_2) = sim(\mathcal{C}_1, \mathcal{C}_2) \cdot \phi(\mathcal{C}_1 \cup \mathcal{C}_2) \tag{3}$$

Dans la suite de cette étude, γ est fixé à 1 dans ϕ.

2.2.3 Analyse théorique

Le processus de classification utilisé dans l'algorithme 1 garantit que les documents qui portent sur la *même thématique* seront regroupés avant les documents portant sur des *thématiques différentes*. Le corpus obtenu ($\mathcal{P}_H$) sera ainsi homogène, c'est-à-dire qu'il ne couvrira qu'un nombre restreint de thématiques. De plus, le fait que le corpus comparable (que nous noterons $\mathcal{P}_F$) obtenu au travers des étapes 1 à 4 découle du corpus originel $\mathcal{P}$ indique que la plus grande partie du vocabulaire de $\mathcal{P}$ sera préservée dans $\mathcal{P}_F$. Nous verrons dans la partie expérimentale que c'est bien le cas. Ce qui semble moins évident, c'est le fait que le processus que nous avons défini garantisse un fort degré de comparabilité. La propriété suivante établit que c'est bien le cas.

Propriété 1 *Soit $\mathcal{C}_1$ et $\mathcal{C}_2$ deux classes de documents qui doivent être regroupées dans le processus de classification. Nous faisons l'hypothèse que le dictionnaire bilingue $\mathcal{D}$ a été construit indépendamment des documents traités, ce qui implique que le degré de comparabilité M_{ef} (respectivement de même pour M_{fe}) est à peu près le même pour différentes parties du corpus*[2]*. Nous faisons de plus l'hypothèse que :*

$$\textit{(I)} \quad \frac{|\mathcal{C}_1^e \cup \mathcal{C}_2^e|}{|\mathcal{C}_2^e|} = \frac{|\mathcal{C}_1^f \cup \mathcal{C}_2^f|}{|\mathcal{C}_2^f|}$$

Alors :

$$M(\mathcal{C}_1^e \cup \mathcal{C}_2^e, \mathcal{C}_1^f \cup \mathcal{C}_2^f) \geq \textit{min}(M(\mathcal{C}_1^e, \mathcal{C}_1^f), M(\mathcal{C}_2^e, \mathcal{C}_2^f))$$

Démonstration (esquisse) : Soit $V = \mathcal{C}_1^e \cap \mathcal{C}_2^e$. En utilisant le fait que $M_{ef}(\mathcal{C}_i^e, \mathcal{C}_i^f) \leq M_{ef}(\mathcal{C}_i^e, \mathcal{C}_i^{f'})$ pour tout $\mathcal{C}_i^{f'}$ tel que $\mathcal{C}_i^f \subseteq \mathcal{C}_i^{f'}$ (et de même pour la direction français vers anglais), nous avons, pour $i = 1, 2$:

$$\sum_{w \in \mathcal{C}_i^e \setminus V} \sigma(w, \mathcal{C}_1^f \cup \mathcal{C}_2^f)) \geq |\mathcal{C}_i^e \setminus V| M_{ef}(\mathcal{C}_i^e, \mathcal{C}_i^f)$$

et, pour les mots de V :

$$\sum_{w \in V} \sigma(w, \mathcal{C}_1^f \cup \mathcal{C}_2^f)) \geq |V| \max(M_{ef}(\mathcal{C}_1^e, \mathcal{C}_1^f), M_{ef}(\mathcal{C}_2^e, \mathcal{C}_2^f))$$

2. En d'autres termes, la proportion de mots anglais (respectivement français) traduits dans le corpus français (respectivement anglais) est homogène sur l'ensemble du corpus.

Alors, d'après l'hypothèse d'indépendance entre corpus et dictionnaire faite en énonçant la propriété 1 :

$$\sum_{w \in (\mathcal{C}_1^e \cup \mathcal{C}_2^e) \cap D_e} \sigma(w, \mathcal{C}_1^f \cup \mathcal{C}_2^f))$$
$$\geq |(\mathcal{C}_1^e \cup \mathcal{C}_2^e) \cap D_e| \min(M_{ef}(\mathcal{C}_1^e, \mathcal{C}_1^f), M_{ef}(\mathcal{C}_2^e, \mathcal{C}_2^f))$$

Un développement similaire sur M_{fe} et l'utilisation de la condition (I) complètent la démonstration.

La propriété précédente garantit que la classe obtenue en fusionnant deux classes existantes a un degré de comparabilité au moins égal à celui de la classe la moins comparable. Le degré de comparabilité ne peut donc décroître dans le processus de classification agglomérative. Comme l'on commence par fusionner les documents les plus comparables, on ne construit que des classes avec un bon degré de comparabilité. Enfin, la condition (I) a de grandes chances d'être réalisée car tous les corpus sont prétraités de façon à éliminer les documents trop courts ou trop longs, souvent source de bruit, et la pénalité utilisée dans la mesure de similarité fournit des classes comprenant des nombres comparables de documents dans les deux langues. Le processus global que nous avons défini permet donc d'obtenir des corpus homogènes et fortement comparables.

2.3 Considérations informatiques

Dans la mesure où les corpus comparables disponibles à l'heure actuelle comprennent en général un nombre important de documents, la classification agglomérative peut s'avérer trop coûteuse. Nous proposons ici une borne inférieure de la mesure de comparabilité qui peut être calculée efficacement ainsi qu'une mise à jour efficace de la matrice de similarité pendant le processus de classification. Le fait de se reposer sur une borne inférieure de la mesure de similarité garantit que les classes obtenues auront un bon degré de comparabilité, car seules les classes les plus similaires sont regroupées à chaque itération de l'algorithme de classification. La propriété suivante établit une telle borne inférieure, sur la base du degré de comparabilité moyen des paires de documents.

Propriété 2 *Soit $\mathcal{P}$ un corpus comparable comprenant une partie anglaise $\mathcal{P}_e$ et une partie française $\mathcal{P}_f$, et soit $\mathcal{D}$ un dictionnaire bilingue, $\mathcal{D}_e$ dénotant le vocabulaire anglais et $\mathcal{D}_f$ le vocabulaire français. Supposons que le dictionnaire est distribué de façon uniforme sur le corpus, c'est-à-dire que :*

$$\forall d_e \in \mathcal{P}_e, \frac{\#_w(d_e \cap \mathcal{D}_e)}{\#_w(d_e)} = \frac{\#_w(\mathcal{P}_e \cap \mathcal{D}_e)}{\#_w(\mathcal{P}_e)}$$

et de même pour la partie française. Supposons de plus que tous les documents, ainsi que les parties anglaise et française du corpus, ont à peu près la même longueur :

$$\forall d_e \in \mathcal{P}_e \text{ and } d_f \in \mathcal{P}_f, \frac{\#_w(d_e)}{\#_w(\mathcal{P}_e)} \simeq \frac{\#_w(d_f)}{\#_w(\mathcal{P}_f)} (= \lambda)$$

Alors :

$$M(\mathcal{P}_e, \mathcal{P}_f) \geq \frac{1}{\#_d(\mathcal{P}_e) \cdot \#_d(\mathcal{P}_f)} \sum_{d_e \in \mathcal{P}_e, d_f \in \mathcal{P}_f} M(d_e, d_f)$$

Nous ne détaillons pas ici la démonstration de cette propriété, purement technique. La première hypothèse faite semble raisonnable (et rejoint celle faite dans la propriété précédente) en l'absence de toute connaissance *a priori* sur les thématiques couvertes par le corpus et leur lien avec le dictionnaire. La seconde hypothèse est en partie garantie dans notre cas par le processus de construction que nous avons défini et la fonction de pénalité associée.

Remplacer M par la borne ci-dessus dans l'équation 1 conduit à une mesure de similarité qui peut être vue comme la valeur accumulée de toutes les connexions entre deux classes. Il est alors possible de mettre à jour la matrice de similarité de façon itérative. Supposons en effet que le processus de classification doive, à un instant donné, fusionner les classes $\mathcal{C}_1$ et $\mathcal{C}_2$ en une seule classe $\mathcal{C}_{new}$. Un nouveau score de similarité entre $\mathcal{C}_{new}$ et toutes les autres classes doit alors être calculé. La similarité entre $\mathcal{C}_{new}$ et une autre classe $\mathcal{C}_3$ peut s'écrire, à partir de l'équation 3 et de la formule de similarité :

$$sim_l(\mathcal{C}_{new}, \mathcal{C}_3) = \frac{(N_{\mathcal{C}_1} + N_{\mathcal{C}_2})\phi(\mathcal{C}_1 \cup \mathcal{C}_2)}{\#_d(\mathcal{C}^e_{new}) \cdot \#_d(\mathcal{C}^f_3) + \#_d(\mathcal{C}^e_3) \cdot \#_d(\mathcal{C}^f_{new})}$$

où $(j = 1, 2)$ et :

$$N_{\mathcal{C}_j} = \frac{(\#_d(\mathcal{C}^e_j) \cdot \#_d(\mathcal{C}^f_3) + \#_d(\mathcal{C}^e_3) \cdot \#_d(\mathcal{C}^f_j))sim_l(\mathcal{C}_j, \mathcal{C}_3)}{\phi(\mathcal{C}_j \cup \mathcal{C}_3)}$$

Dans le processus de classification, dans la mesure où $sim_l(\mathcal{C}_1, \mathcal{C}_3)$ et $sim_l(\mathcal{C}_2, \mathcal{C}_3)$ sont déjà connus avant le calcul de $sim_l(\mathcal{C}_{new}, \mathcal{C}_3)$, la matrice de similarité peut directement être mise à jour à chaque itération. En notant N_c le nombre de classes avant fusion, la complexité de cette mise à jour est de l'ordre de $\mathcal{O}(N_c)$, alors qu'elle atteint $\mathcal{O}(N_c \times \bar{C}^2)$ si l'on applique directement les équations 1 et 3 ($\bar{C}$ représentant le nombre moyen de documents par classe).

3 Expériences et résultats

Les différentes expériences que nous avons réalisées ont pour objectif d'évaluer : (i) si l'algorithme que nous avons proposé induit des corpus d'une meilleure qualité en ce qui concerne la comparabilité, (ii) si les lexiques bilingues extraits de ces corpus sont eux aussi d'une qualité plus importante, et (iii) si ces lexiques peuvent être utilisés pour améliorer les performances des systèmes de recherche d'information interlingue.

Dans nos expériences, différents corpus sont utilisés : le corpus anglais TREC [3] de l'*Associated Press* (noté *AP*) et les corpus fournis dans les tâches multilingues des campagnes CLEF [4] dont pour l'anglais le *Los Angeles Times* (*LAT94*) et le *Glasgow Herald* (*GH95*) et pour le français *Le Monde* (*MON94*), le *SDA 94* (*SDA94*) et *95* (*SDA95*). Outre ces corpus existants, deux corpus monolingues ont été extraits à partir de *Wikipédia* : le corpus anglais *Wiki-En* construit en retenant l'ensemble des articles appartenant à la catégorie *Society* pour une profondeur inférieure à 4 (soit 33 000 mots anglais distincts) et le corpus français *Wiki-Fr* toujours pour la catégorie *Société* pour une profondeur inférieure à 7 (soit 28 000 mots français distincts). Le dictionnaire bilingue bd_0 nécessaire pour la tâche d'extraction de lexiques est quant à lui construit à partir de dictionnaires en ligne. Dans toutes nos expériences, nous utilisons la méthode décrite dans le présent article complétée par celle présentée dans (Li & Gaussier, 2010). Cette dernière méthode est à notre connaissance la seule approche alternative pour améliorer la comparabilité des corpus, d'où son importance dans l'évaluation.

3.1 Comparabilité de corpus

L'algorithme de classification décrit en section 2.2.1 est utilisé pour améliorer le degré de comparabilité d'un corpus comparable. Les corpus *GH95* et *SDA95* sont utilisés pour construire le corpus comparable $\mathcal{P}^0$ (56 000

3. http://trec.nist.gov/
4. http://www.clef-campaign.org

mots pour l'anglais et 42 000 le français). En outre, nous exploitons deux corpus comparables supplémentaires pour nous assurer que l'efficacité de notre algorithme n'est pas liée à une ressource externe spécifique : i) $\mathcal{P}_T^1$ composé à partir des corpus *LAT94*, *MON94* et *SDA94* (109 000 mots pour l'anglais et 87 000 pour le français) et ii) $\mathcal{P}_T^2$ composé à partir des corpus *Wiki-En* et *Wiki-Fr* (368 000 mots pour l'anglais et 378 000 pour le français).

Après le processus de classification, nous obtenons les corpus $\mathcal{P}^1$ (pour le corpus externe $\mathcal{P}_T^1$) et $\mathcal{P}^2$ (pour le corpus externe $\mathcal{P}_T^2$). Comme nous l'avons indiqué précédemment, nous utilisons aussi la méthode décrite dans (Li & Gaussier, 2010) sur les mêmes données pour comparer nos résultats et obtenons ainsi le corpus $\mathcal{P}^{1'}$ (pour $\mathcal{P}_T^1$) et $\mathcal{P}^{2'}$ (pour $\mathcal{P}_T^2$) à partir de $\mathcal{P}^0$. Au niveau de la couverture lexicale, $\mathcal{P}^1$ couvre 97,9% du vocabulaire de $\mathcal{P}^0$, tandis que $\mathcal{P}^2$ couvre 99,0% de celui de $\mathcal{P}^0$. Nous pouvons ainsi constater qu'une très grande partie du vocabulaire du corpus d'origine a été conservé, ce qui est l'une des exigences de notre approche. En ce qui concerne les scores de comparabilité, $\mathcal{P}^1$ atteint 0,924 et $\mathcal{P}^2$ 0,939. Les deux corpus comparables ont donc bien un degré de comparabilité supérieur au corpus d'origine qui était de l'ordre de 0,881 comme cela est suggérée par la propriété 1. En outre, les corpus $\mathcal{P}^1$ et $\mathcal{P}^2$ sont plus comparables que le corpus $\mathcal{P}^{1'}$ (comparabilité de 0,912) et $\mathcal{P}^{2'}$ (comparabilité de 0,915) ce qui montre bien que l'homogénéité est un élément crucial pour évaluer la comparabilité.

3.2 Extraction de lexiques bilingues

TABLE 1 – Évaluation des lexiques bilingues extraits pour différents corpus comparables

	$\mathcal{P}^0$	$\mathcal{P}^{1'}$	$\mathcal{P}^{2'}$	$\mathcal{P}^1$	$\mathcal{P}^2$	$\mathcal{P}^1 > \mathcal{P}^0$	$\mathcal{P}^2 > \mathcal{P}^0$
Précision	0,226	0,277	0,325	0,295	0,461	0,069 30,5 %	0,235 104,0 %
Rappel	0,103	0,122	0,145	0,133	0,212	0,030 29,1 %	0,109 105,8 %

TABLE 2 – Comparaison de la précision pour différents intervalles de fréquences des mots de la liste d'évaluation

	$\mathcal{P}^0$	$\mathcal{P}^{2'}$	$\mathcal{P}^2$	$\mathcal{P}^{2'} > \mathcal{P}^0$	$\mathcal{P}^2 > \mathcal{P}^0$	$\mathcal{P}^2 > \mathcal{P}^{2'}$
W_l	0,135	0,206	0,304	0,071 52,6 %	0,169 125,2 %	0,098 47,6 %
W_m	0,256	0,390	0,564	0,134 52,3 %	0,308 120,3 %	0,174 44,6 %
W_h	0,434	0,632	0,667	0,198 45,6 %	0,233 53,7 %	0,035 5,5
,All	**0,226**	**0,325**	**0,461**	**0,099 43,8 %**	**0,235 104,0 %**	**0,136 41,8 %**

Comme les travaux antérieurs en extraction de lexiques bilingues à partir de corpus comparables exploitent des ressources différentes et opèrent des choix distincts des nôtres, il est relativement difficile de se comparer à ceux-ci (Laroche & Langlais, 2010). En outre, puisque notre approche vise à améliorer la comparabilité de corpus, elle peut être ensuite couplée à une méthode existante d'extraction de lexiques bilingues. Il est donc tout aussi intéressant de directement évaluer si un tel couplage peut conduire à des performances accrues en termes de qualité des lexiques extraits.

L'extraction de lexiques bilingues à partir de corpus comparables repose sur la méthode proposée par Fung & Yee (1998) plus connue maintenant sous le nom d'*approche standard* notamment dans les travaux de (Déjean *et al.*, 2002; Gaussier *et al.*, 2004; Yu & Tsujii, 2009). Dans cette approche, chaque mot est représenté sous la forme d'un vecteur de contexte composé des mots qui co-occurrent avec lui dans une fenêtre donné. Les vecteurs de contexte de la langue source sont ensuite traduits vers la langue cible en s'appuyant sur un dictionnaire bilingue.

Enfin, la traduction d'un mot est obtenue en comparant son vecteur de contexte traduit à l'ensemble des vecteurs de la langue cible à travers une mesure de distance ou similarité vectorielle telle que le cosinus.

3.2.1 Paramètres expérimentaux

Afin d'évaluer la qualité des lexiques bilingues extraits, nous divisons notre dictionnaire bilingue bd_0 en deux parties : 10 % des mots anglais accompagnés de leurs traductions sont choisis aléatoirement et uniquement utilisés comme liste d'évaluation, les 90 % restant sont utilisés pour assurer la traduction des vecteurs de contexte dans l'approche standard. Les mots anglais absents de $\mathcal{P}_e$ ou pour lesquels aucune traduction n'a été trouvée dans $\mathcal{P}_f$ sont retirés de la liste d'évaluation. Pour chaque mot anglais de la liste d'évaluation, tous les mots français de $\mathcal{P}_f$ sont ordonnés suivant leur similarité avec les mots anglais. Les mesures de précision et rappel sont ensuite calculées sur les N premiers candidats. Les valeurs de la précision dans ce cas correspondent à la proportion de listes contenant la traduction correcte (en cas de traductions multiples, une liste est réputée contenir la traduction correcte dès lors que l'une des traductions possibles est présente). Le rappel est quant à lui la proportion de traductions correctes trouvée dans les listes sur toutes les traductions fournies dans le corpus. Cette manière de procéder a été utilisée dans différents travaux antérieurs et peut être maintenant considérée comme un méthode d'évaluation attestée. En outre, plusieurs études ont montré qu'il est plus facile de trouver les traductions correctes pour les mots fréquents que pour les mots rares (Pekar *et al.*, 2006). Afin de prendre en compte ce phénomène, nous distinguons différents intervalles d'effectifs pour évaluer la validité de notre approche. Ainsi, les mots avec un effectif inférieur à 100 sont définis comme étant des mots de faibles fréquence (W_l), ceux avec un effectif supérieur à 400 sont définis comme étant des mots très fréquents (W_h), et enfin les mots dont l'effectif est compris entre ces deux seuils sont considérés comme des mots de fréquence intermédiaire (W_m).

3.2.2 Analyse des résultats

Dans une première série d'expériences, les lexiques bilingues sont extraits à partir des corpus obtenus ii) par notre approche ($\mathcal{P}^1$ et $\mathcal{P}^2$), ii) par la méthode décrite dans (Li & Gaussier, 2010) ($\mathcal{P}^{1'}$ and $\mathcal{P}^{2'}$) et iii) enfin avec le corpus d'origine $\mathcal{P}^0$, avec N fixé à 20. La table 1 présente les résultats obtenus. Les deux dernières colonnes "$\mathcal{P}^1 > \mathcal{P}^0$" et "$\mathcal{P}^2 > \mathcal{P}^0$" indique les différences absolue et relative, exprimées en pourcentage, par rapport à $\mathcal{P}^0$. Comme nous pouvons le constater, les meilleurs résultats sont obtenus à partir des corpus construits avec la méthode que nous avons proposée. Les lexiques extraits à partir du corpus où le degré de comparabilité a été renforcé sont d'une bien meilleure qualité que ceux obtenus à partir du corpus d'origine ou encore du corpus construit avec l'approche de (Li & Gaussier, 2010). La différence de qualité est encore plus notable avec $\mathcal{P}^2$ qui est obtenu à partir d'un corpus externe volumineux $\mathcal{P}^2_T$. Ces résultats semblent confirmer l'intuition qu'il est possible de trouver plus aisément dans des corpus volumineux des documents en relation avec un corpus donné.

Afin d'évaluer la relation entre la qualité de ces méthodes et la fréquence des mots à traduire, nous nous concentrons sur les meilleurs résultats sur $\mathcal{P}^{2'}$ pour l'approche précédente et sur ceux de $\mathcal{P}^2$ pour notre approche. La table 2 résume les résultats obtenus. On remarquera, sans véritablement de surprise, que les résultats obtenus pour les mots ayant une haute fréquence sont meilleurs que ceux obtenus pour les mots de faible fréquence. En outre, notre approche est la meilleure quel que soit l'intervalle de fréquence pris en compte. La précision globale peut être augmentée en relatif de 41,8 % (de 0,325 à 0,461). En comparant $\mathcal{P}^2$ avec le corpus d'origine $\mathcal{P}^0$, nous pouvons noter pour la précision globale, une augmentation relative de 104,0 % (de 0,226 à 0,461), ce qui est très satisfaisant dans ce contexte d'évaluation. Enfin, l'amélioration pour les mots de faible et moyenne fréquence est plus importante pour $\mathcal{P}^2$, ce qui démontre que notre approche se comporte bien mieux sur ce qui est généralement

considéré comme un problème difficile (Pekar *et al.*, 2006).

3.3 Expériences en recherche d'information interlingue

TABLE 3 – Score MAP pour la tâche de recherche d'information interlingue suivant différents dictionnaires bilingues

	mon	bd_1	bd_1+cc_0	bd_1+cc_1	bd_1+cc_2	bd_2	bd_2+cc_0	bd_2+cc_1	bd_2+cc_2
MAP	$0,422$	$0,313$	$0,327^{\bullet}$	$0,328^{\bullet}$	$0,338^{\bullet}$	$0,375$	$0,382$	$0,377$	$0,391^{\bullet}$

Dans la dernière série d'expériences, nous cherchons à évaluer l'apport des différents lexiques extraits à partir de corpus comparables pour une tâche de recherche d'information interlingue. Pour ce faire, nous exploitons les sujets des campagnes CLEF de 2001 et 2002, rassemblant environ 100 sujets distincts, comme requêtes sur une collection de 113 000 documents issus du *Los Angeles Times*. Les sujets anglais correspondants sont utilisés pour interroger la même collection (référence *mon*). Seul le titre et la partie description des sujets CLEF sont utilisés pour construire des requêtes. En outre, les mots outils et les phrases non pertinentes telles que *find documents which report about* sont supprimés des requêtes. La recherche est réalisée avec le modèle Indri du système de recherche d'information Lemur (http ://www.lemurproject.org). Une variante de l'approche introduite dans (Pirkola, 1998) et (Talvensaari *et al.*, 2007) est aussi utilisée pour transformer les sujets français en requêtes en anglais. L'idée est de borner toutes les possibilités de traduction d'un mot français dans le sujet du texte avec un opérateur WSYN. Ensuite, toutes les traductions candidates dans l'opérateur WSYN sont traitées comme des synonymes avec des poids différents.

Dans nos expériences, nous combinons deux dictionnaires bilingues de langue générale bd_1 (68 0000 traductions) et bd_2 (116 000 traductions) avec les lexiques bilingues obtenus automatiquement dans la précédente section. Nous utilisons ici les lexiques cc_0 (extrait de $\mathcal{P}^0$), cc_1 (extrait de $\mathcal{P}^{2'}$) et cc_2 (extrait de $\mathcal{P}^2$). Différentes combinaisons de ces ressources sont réalisées, y compris $bd_{1/2}$, $bd_{1/2}$+cc_0, $bd_{1/2}$+cc_1, $bd_{1/2}$+cc_2. Lorsque qu'un dictionnaire de langue générale et un lexique extrait sont combinés, plus de poids est attribué aux traductions candidat du dictionnaire de langue générale. Le poids des différents mots traduits à partir de $cc_{0/1/2}$ est quant à lui le cosinus entre les vecteurs de contexte de chaque mot (c'est-à-dire le score donné par l'approche standard précédemment évoquée). Le poids pour les traductions trouvées dans le dictionnaire bilingue est fixé empiriquement à 25. Comme il est d'usage en recherche d'information, nous utilisons la mesure MAP (Mean Average Precision) afin d'évaluer les performances des différents systèmes. L'importance des différences entre les différents systèmes est estimée par un t-test apparié de Student (p-value fixée à 0,1). Les résultats obtenus sont indiqués dans la table 3. Pour le dictionnaire de langue générale bd_1, on note toujours une amélioration significative des résultats (identifiée par la marque $\bullet$) du score MAP lorsque l'un des lexiques bilingues extraits du corpus comparables est utilisé. Lorsque bd_2, qui est beaucoup plus riche que bd_1, est utilisé, seulement le lexique bilingue cc_2 extrait avec notre méthode à partir $\mathcal{P}^2$ conduit à une amélioration significative des résultats. Cela montre que cc_2 est supérieure à cc_1 et cc_0 dans la tâche de recherche d'information interlingue, en particulier lorsque le dictionnaire de langue générale utilisé est d'une taille importante. Ces résultats semblent confirmer que notre approche basée sur de la classification est plus adaptée que l'approche gloutonne des travaux précédents de (Li & Gaussier, 2010). Enfin, la combinaison actuelle des lexiques extraits avec le système de recherche d'information interlingue est relativement simple et pourrait être certainement améliorée en exploitant d'autres modèles de combinaison.

4 Conclusion

Dans cet article, nous avons proposé une nouvelle approche pour augmenter le degré de comparabilité des documents constituant un corpus comparable afin d'améliorer la qualité des lexiques bilingues extraits de corpus comparables et les performances des systèmes de recherche d'information interlingue. Nous avons démontré théoriquement puis empiriquement que notre approche permet de garantir un certain degré de comparabilité et l'homogénéité du corpus tout en préservant une large part du vocabulaire du corpus d'origine. Enfin, nos expériences montrent que les lexiques bilingues que nous obtenons sont d'une meilleure qualité que ceux obtenus avec les approches précédentes, et que ces lexiques peuvent être utilisés pour améliorer significativement les résultats des systèmes de recherche d'information interlingue.

Les deux étapes cruciales de notre approche sont d'une part l'extraction d'un noyau fortement comparable du corpus original, et, d'autre part, l'alignement des parties du corpus orignal, non présentes dans ce noyau, avec un corpus externe. Le seuil introduit au niveau du degré de comparabilité permet de contrôler la taille et la qualité du noyau extrait dans la première étape. Si le corpus original n'est que très faiblement comparable, il est alors possible que ce noyau soit vide (ce qui est un résultat souhaitable dans ce cas). Dans tous les cas, excepté celui où le noyau correspond au corpus original, le corpus final dépend de la proximité du corpus original (en fait de la partie restante après extraction du noyau) et du corpus externe utilisé. Bien évidemment, si le corpus externe est trop différent du corpus original, l'on ne pourra pas compléter correctement le noyau. Considérer des corpus externes les plus larges possibles permet ici d'augmenter les chances de trouver des documents comparables [5] L'idéal serait bien sûr d'avoir accès à la collection la plus large possible, et le web constitue ici un excellent candidat. Il est cependant nécessaire de pouvoir, à partir d'un document donné dans une langue source, extraire du web un ensemble de documents comparables en langue cible (on peut ensuite directement utiliser notre méthode sur l'union de ces ensembles). Or nous n'avons pas réussi jusqu'à présent à réaliser correctement cette extraction. La constitution entièrement automatique de collections comparables à partir du web nous semble être un problème difficile, qui requiert d'autres attributs que ceux utilisés pour les corpus parallèles. C'est un point que nous comptons développer dans le futur.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet METRICC (`www.metricc.com`) a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-CORD-009. Enfin, nous tenons à remercier les relecteurs pour leurs commentaires précieux.

Références

BALLESTEROS L. & CROFT W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th ACM SIGIR*, p. 84–91, Philadelphia, Pennsylvania, USA.

5. C'est ce qui distingue les corpus $\mathcal{P}^1$ et $\mathcal{P}^1$ dans nos expériences, le deuxième étant obtenu à partir d'un corpus externe à plus large couverture.

DÉJEAN H., GAUSSIER E. & SADAT F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics*, p. 1–7, Taipei, Taiwan.

FUNG P. & MCKEOWN K. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, p. 192–202, Hong Kong.

FUNG P. & YEE L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics*, p. 414–420, Montreal, Quebec, Canada.

GARERA N., CALLISON-BURCH C. & YAROWSKY D. (2009). Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *CoNLL 09 : Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, p. 129–137, Boulder, Colorado.

GAUSSIER E., RENDERS J.-M., MATVEEVA I., GOUTTE C. & DÉJEAN H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, p. 526–533, Barcelona, Spain.

LAROCHE A. & LANGLAIS P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, p. 617–625, Beijing, China.

LI B. & GAUSSIER E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, p. 644–652, Beijing, China.

MORIN E., DAILLE B., TAKEUCHI K. & KAGEURA K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, p. 664–671, Prague, Czech Republic.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1), 19–51.

PEKAR V., MITKOV R., BLAGOEV D. & MULLONI A. (2006). Finding translations for low-frequency words in comparable corpora. *Machine Translation*, **20**(4), 247–266.

PIRKOLA A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, p. 55–63, Melbourne, Australia.

RAPP R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, p. 519–526, College Park, Maryland, USA.

ROBITAILLE X., SASAKI Y., TONOIKE M., SATO S. & UTSURO T. (2006). Compiling French-Japanese terminologies from the web. In *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics*, p. 225–232, Trento, Italy.

SHEZAF D. & RAPPOPORT A. (2010). Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 98–107, Uppsala, Sweden.

TALVENSAARI T., LAURIKKALA J., JÄRVELIN K., JUHOLA M. & KESKUSTALO H. (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Trans. Inf. Syst.*, **25**(1), 4.

YU K. & TSUJII J. (2009). Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *Proceedings of HLT-NAACL 2009*, p. 121–124, Boulder, Colorado, USA.

Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée

Nadja Vincze[1] Yves Bestgen[2]

(1) UCLouvain, Cental, Place Blaise Pascal, 1, B-1348 Louvain-la-Neuve, Belgique
(2) UCLouvain, CECL, B-1348 Louvain-la-Neuve, Belgique
nadja.vincze@uclouvain.be, yves.bestgen@uclouvain.be

Résumé

De nombreuses méthodes automatiques de classification de textes selon les sentiments qui y sont exprimés s'appuient sur un lexique dans lequel à chaque entrée est associée une valence. Le plus souvent, ce lexique est construit à partir d'un petit nombre de mots, choisis arbitrairement, qui servent de germes pour déterminer automatiquement la valence d'autres mots. La question de l'optimalité de ces mots germes a bien peu retenu l'attention. Sur la base de la comparaison de cinq méthodes automatiques de construction de lexiques de valence, dont une qui, à notre connaissance, n'a jamais été adaptée au français et une autre développée spécifiquement pour la présente étude, nous montrons l'importance du choix de ces mots germes et l'intérêt de les identifier au moyen d'une procédure d'apprentissage supervisée.

Abstract

Many methods of automatic sentiment classification of texts are based on a lexicon in which each entry is associated with a semantic orientation. These entries serve as seeds for automatically determining the semantic orientation of other words. Most often, this lexicon is built from a small number of words, chosen arbitrarily. The optimality of these seed words has received little attention. In this study, we compare five automatic methods to build a semantic orientation lexicon. One among them, to our knowledge, has never been adapted to French and another was developed specifically for this study. Based on them, we show that choosing good seed words is very important and identifying them with a supervised learning procedure brings a benefit.

Mots-clés : Analyse de sentiments, lexique de valence, apprentissage supervisé, analyse sémantique latente
Keywords: Sentiment analysis, semantic orientation lexicon, supervised learning, latent semantic analysis

1 Introduction

La classification de textes consiste à classer automatiquement les textes dans un ensemble prédéfini de catégories. Ce sont initialement les classifications thématiques et par genre qui ont motivé les recherches, mais, depuis une dizaine d'années, ce champ d'études s'est élargi et intègre la classification de textes en fonction des sentiments qui y sont exprimés : détection de la subjectivité, avec une classification objectif / subjectif (Wiebe et al., 2004 ;Yu, Hatzivassiloglou, 2003) et détermination de la valence des documents, avec une classification binaire positif / négatif, parfois multi-classes selon le degré de polarité (Abbasi et al., 2008 ; Pang et al., 2002). La plupart des implémentations de ces classifieurs requièrent des lexiques porteurs de valence, c'est-à-dire des lexiques où à chaque entrée est associée une polarité ou un degré de polarité. Une série d'approches attribuent une valence globale aux textes selon des statistiques sur la présence de mots subjectifs (Bestgen, 2006 ; Turney, 2002). Les approches dites symboliques intègrent la prise en compte de phénomènes syntaxiques qui viennent modifier l'orientation sémantique de mots ou de groupes de mots (Harb et al., 2008 ; Vernier et al., 2009 ; Wilson et al., 2005). Enfin, quelques tentatives d'apprentissages supervisés ont également pris en compte des mots, dont la valence est connue, comme caractéristiques de leurs vecteurs (Chesley et al., 2006). Ces lexiques constituent donc des ressources sémantiques capitales au développement de classifieurs efficaces.

Dans un premier temps, ces lexiques ont été construits manuellement par des juges (Nasukawa, Yi, 2003 ; Wiebe et al., 2005), mais le travail étant lent et couteux, des procédures automatiques ou semi-automatiques ont vu le jour et constituent aujourd'hui un sous-domaine de recherche important. Comme le souligne la présentation des travaux antérieurs (section 2), une spécificité des recherches menées dans ce champ est qu'elles portent presque exclusivement sur l'anglais, langue pour laquelle de nombreuses ressources linguistiques ont été développées comme WordNet (Miller, 1990). Un des deux objectifs principaux de notre étude est de déterminer dans quelle mesure ces méthodes sont applicables au français. Une autre spécificité des recherches menées dans ce champ est que la quasi-totalité des méthodes proposées utilise un petit nombre de mots, comme *bon*, *mauvais*, *gentil*, afin de servir de germes (*seed*) pour déterminer automatiquement la valence d'autres mots (voir par exemple Hu, Liu, 2004; Kamps, Marx, 2002 ; Turney, Littman, 2003). La question de l'optimalité de ces mots germes a bien peu retenu l'attention, le plus souvent les chercheurs reprenant ceux proposés dans des travaux antérieurs (Esuli, Sebastiani, 2006 ; Harb et al., 2008). Notre second objectif est de proposer une méthode permettant d'identifier automatiquement ces germes au moyen d'une technique d'apprentissage supervisée.

Après une brève présentation des travaux antérieurs, la section 3 décrit les différentes méthodes comparées dans le cadre de cette étude. Une série d'expériences visant à évaluer leur efficacité sont présentées dans la section 4. La section 5 rapporte les principaux résultats, dont les implications et les développements possibles sont discutés dans la conclusion.

2 Travaux antérieurs

Parmi les méthodes automatiques ou semi-automatiques proposées pour construire des lexiques porteurs de valences, on peut distinguer deux types d'approches : celles basées sur des ressources linguistiques comme WordNet et celles basées sur des corpus de textes.

Les approches qui s'appuient sur des bases de connaissances linguistiques calculent généralement la similarité entre les mots à partir de leur relation de synonymie. Une méthode de base consiste à partir de quelques mots dont la valence est connue et à lancer un algorithme d'amorçage (*bootstrapping*) qui parcourt les liens synonymiques et antonymiques de la base, en attribuant la même orientation aux mots synonymes et vice-versa (Hu, Liu, 2004 ; Kim, Hovy, 2004). Kamps et Marx (2002) ont probablement été les premiers à proposer une telle procédure en dérivant de WordNet un graphe dans lequel chaque nœud représente un terme et un lien est présent entre deux nœuds s'ils sont synonymes. À partir de ce graphe, ils calculent une valeur normalisée pour les nœuds liés aux mots g*ood* et *bad*. Esuli et Sebastiani (2006) ont étendu cette approche pour développer SentiWordNet, une ressource basée sur WordNet, qui assigne à chaque *synset* trois valeurs normalisées : une positive, une négative et une objective. La spécificité principale de leur approche est qu'elle s'appuie sur un apprentissage semi-supervisé basé sur les définitions de mots germes sélectionnés manuellement.

Ne disposant pas d'informations sur les liens synonymiques, les approches qui s'appuient sur des corpus calculent les similarités différemment. Hatzivassiloglou et McKeown (1997) ont proposé un algorithme capable de déterminer l'orientation sémantique d'adjectifs à partir de l'analyse de leurs cooccurrences avec des conjonctions. Turney et Littman (2003 ; Turney, 2002) et Bestgen (2002, 2008) ont proposé des méthodes plus générales puisqu'elles permettent d'estimer la valence de n'importe quel terme présent dans un corpus. Ils utilisent l'analyse sémantique latente (ASL, *Latent Semantic Analysis*, Deerwester et al., 1990) pour construire un espace sémantique à partir d'informations statistiques sur les cooccurrences de termes dans des textes. Turney et Littman l'emploient pour estimer la distance sémantique entre des mots et 14 mots germes, 7 positifs (*good*, *nice*, *excellent*, *positive*, *fortunate*, *correct*, *superior*) et 7 négatifs (*bad*, *nasty*, *poor*, *negative*, *unfortunate*, *wrong*, *inferior*). Un mot est d'autant plus positif qu'il est plus proche des germes positifs et plus éloigné des germes négatifs. Pour sa part, Bestgen (2002) a recours à l'ASL pour identifier les mots fréquemment associés aux mots dont il veut déterminer la valence affective. Il attribue à chaque mot la valence moyenne de ses plus proches voisins dont la valence est connue. Pour cela, il s'appuie sur un dictionnaire de 3000 mots dont la valence a été évaluée par des juges. On notera que les similarités peuvent être calculées sans passer par l'analyse sémantique latente, mais que, dans ce cas, des corpus de très grande taille semblent nécessaires (Turney, Littman, 2003; Velikovich et al., 2010), sauf si, à la manière de Harb et al. (2008), on emploie un corpus très spécifique et des règles d'associations.

Peu d'initiatives de construction automatique de lexiques ont eu lieu en français, comparé à l'effervescence dans le milieu anglophone. Nous pouvons citer Bestgen (2002) et Chardon (2010) qui a développé une méthode pour élaborer une ressource lexicale d'adjectifs d'opinion à partir d'une liste de mots germes et d'une taxinomie des mots du français. Pak et Paroubek (2010) ont proposé une méthode de construction automatique d'un lexique affectif à partir de messages disponibles sur Twitter. Leur procédure est basée sur la comparaison de la fréquence d'occurrence d'un mot dans les messages contenant une émoticône positive et dans ceux contenant une émoticône négative. Vernier et Monceaux (2010) ont proposé une méthode d'apprentissage pour enrichir automatiquement un lexique subjectif à partir d'un corpus annoté. L'apprentissage automatique se base sur des tests sémantiques, qui permettent de mesurer le degré de subjectivité des termes, ainsi que leur valence s'il s'agit d'adjectifs, et qui sont effectués à l'aide du moteur de recherche Yahoo!.

3 Méthodes évaluées pour estimer la valence de mots

Cinq méthodes pour estimer automatiquement la valence de mots ont été comparées, deux de celles-ci consistant en une transposition de méthodes efficaces pour la langue anglaise : celle de Turney et Littman (2002, 2003) et celle de Kamps et Marx (2002 ; Kamps et al., 2004). Nous avons également repris la méthode de Bestgen (2002, 2008). Ces trois méthodes serviront de référence pour évaluer deux nouvelles approches : une extension de la méthode de Kamps et Marx et une méthode d'apprentissage supervisé de mots germes. La présente section décrit les principes à la base de ces différentes méthodes. Des précisions à propos de leur implémentation et des ressources linguistiques qu'elles requièrent sont données dans la section suivante.

3.1 Niveaux de base : SO-ASL et DIC-ASL

Ces deux méthodes se basent sur l'analyse sémantique latente d'une collection de textes pour déterminer la proximité entre des mots et des germes dont la valence est connue.

- SO-ASL : il s'agit de la méthode proposée par Turney et Littman (2003) décrite ci-dessus. Elle est basée sur 14 mots germes choisis en raison de leur valence extrême sur la dimension positif-négatif. La valence d'un mot correspond à la somme des cosinus entre ce mot et les germes positifs dont on soustrait la somme des cosinus entre ce mot et les germes négatifs.

- DIC-ASL : il s'agit de la méthode proposée par Bestgen (2002) décrite ci-dessus. Pour chaque mot dont on veut déterminer la valence, on identifie les 30 plus proches voisins dont la valence est connue et on lui affecte la valence moyenne de ceux-ci.

3.2 Estimation sur la base de relations de synonymie : KA1 et KA7

Ces deux méthodes sont basées sur la fonction d'évaluation définie par Kamps et Marx (2002).

- KA1 : cette méthode est basée sur les liens synonymiques entre les adjectifs. Le principe consiste à mesurer la distance minimale, c'est-à-dire le plus court chemin, entre le mot auquel on veut attribuer une valeur et les mots germes *good* et *bad*. La valence d'un terme *t* est alors égale à sa distance relative avec les deux germes :

$$KA1(t) = \frac{d(t, mauvais) - d(t, bon)}{d(bon, mauvais)}$$

où *d (i, j)* représente la distance du plus court chemin synonymique entre les mots i et j.

- KA7 est une adaptation de KA1 dans laquelle le nombre de paires d'adjectifs de référence est multiplié par 7. Nous avons repris les 7 paires de référence de Turney et Littman (2003), que nous avons traduites comme suit : *bon, gentil, excellent, positif, heureux, correct, supérieur* et *mauvais, méchant, médiocre, négatif, malheureux, faux, inférieur*. La fonction d'évaluation adaptée reprend alors la somme des évaluations pour chaque paire :

$$KA7(t) = \frac{\sum_{k=1}^{n} d(t, i_k) - \sum_{k=1}^{n} d(t, j_k)}{\prod_{k=1}^{n} d(i_k, j_k)}$$

où i_k et j_k forment une paire d'adjectifs positif et négatif des n paires prises en compte.

3.3 Apprentissage supervisé de mots germes : ASG

Un des objectifs de cette recherche est de proposer et d'évaluer une méthode dérivée de celles de Turney et Littman (2003) et de Bestgen (2002) dans laquelle les mots germes originaux, sélectionnés arbitrairement, sont remplacés par des germes optimaux obtenus par une procédure d'apprentissage supervisée basée sur la régression. Pour ce faire, nous employons comme matériel d'apprentissage une norme lexicale pour la dimension évaluative obtenue en demandant à des juges d'évaluer un grand nombre de mots sur cette dimension. À la suite de Heise (1965), une série de normes de ce type ont été développées, principalement en psycholinguistique (Syssau, Font, 2005). La méthode proposée est composée des quatre étapes suivantes :

1. Sélectionner comme germes potentiels les mots qui sont les plus extrêmes sur la dimension positif-négatif selon une norme évaluative comme celle employée dans DIC-ASL.

2. Sur la base d'un espace sémantique obtenu par l'ASL d'une collection de textes, calculer le cosinus entre chacun de ces germes potentiels et tous les mots qui se trouvent dans la norme.

3. Utiliser une procédure de régression afin de construire un modèle prédictif basé sur les germes les plus efficaces pour prédire la valence.

4. Employer le modèle construit à l'étape précédente pour estimer la valence de termes présents dans l'espace sémantique, mais non dans la norme initiale.

Le critère de sélection des germes potentiels proposé à la première étape devrait permettre l'identification de mots germes similaires à ceux originellement choisis par Turney et Littman (2003). Toutefois, lorsqu'on considère le fait que le seuil pour sélectionner les mots les plus extrêmes est par définition arbitraire, il devient immédiatement évident que la procédure proposée n'est qu'un cas particulier d'une procédure plus générale dans laquelle les germes potentiels sont composés de l'ensemble des mots présents dans la norme. Et, d'une manière tout aussi évidente, cette première généralisation n'est, elle-même, qu'un cas particulier d'une seconde généralisation, qui emploie comme germes potentiels tous les mots pour lesquels il est possible de calculer un cosinus avec les mots qui se trouvent dans la norme, soit tous les mots présents dans l'espace sémantique, que leur valence soit connue ou non. Étant donné que les candidats germes pour l'approche la plus restrictive forment un sous-ensemble des candidats germes employés dans les approches plus générales, on doit s'attendre à ce que la qualité de la prédiction de la valence des mots du dictionnaire

initial soit d'autant meilleure que l'approche est la plus générale. Par contre, les capacités de généralisation des différents modèles pourraient être équivalentes si ceux basés sur le plus grand nombre de germes potentiels présentent un défaut de surapprentissage.

4 Expériences

4.1 Ressources linguistiques pour l'implémentation des méthodes

Les différentes méthodes proposées ci-dessus nécessitent des ressources linguistiques spécifiques comme un dictionnaire de synonymes ou une collection de textes pour extraire l'espace sémantique. Les ressources que nous avons employées sont décrites dans la présente section.

4.1.1 Dictionnaire de synonymes

L'adaptation de la méthode de Kamps et Marx (2002) au français nécessite une ressource plus ou moins équivalente au WordNet anglais. En raison de la trop faible couverture de WOLF (WordNet Libre du Français) et du WordNet français développé dans le cadre du projet EuroWordNet[1], nous avons employé le dictionnaire de synonymes développé par le laboratoire CRISCO de l'université de Caen (Manquin et al., 2004)[2]. Celui-ci a été constitué à partir de sept dictionnaires français et comprend plus de 49 000 entrées et 396 000 relations synonymiques. De manière similaire à Kamps et Marx (2002), nous avons récupéré récursivement tous les mots liés à la paire d'adjectifs *bon* et *mauvais*, avec des restrictions sur la catégorie grammaticale pour éviter de générer trop de bruit. Une petite adaptation a dû être faite pour rendre la liste des synonymes récupérés symétrique (Kamps et al., 2004 : 1115).

4.1.2 Norme de valence : Nev

La norme de valence employée pour les méthodes DIC-ASL et ASG est composée de 3252 mots évalués sur une échelle à 7 points allant de *très désagréable* (1) à *très agréable* (7) par un minimum de 30 juges (Hogenraad et al., 1995). À titre d'exemple, la liste suivante donne les valeurs attribuées à quelques mots extraits aléatoirement de ce dictionnaire : détresse = 1.4, impassible = 2.6, ambigu = 3.2, outil = 4.3, revenir = 5.0, admiratif = 5.7, doux = 6.0.

4.1.3 Constitution de l'espace sémantique

L'espace sémantique, utilisé pour calculer les cosinus entre les mots nécessaires pour SO-ASL, DIC-ASL et ASG, a été construit sur la base d'une collection de textes littéraires composée de romans, nouvelles et contes disponibles sur le Web (principalement dans les bases littéraires ABU et Frantext). Elle contient approximativement 5 300 000 mots. Chaque texte a été subdivisé en segments de 125 mots. Pour construire le tableau lexical, les prétraitements suivants ont été effectués : lemmatisation par le logiciel TreeTagger (Schmid, 1994), suppression de mots outils et suppression des mots de fréquence totale inférieure à 10. La matrice de cooccurrences des 12 285 termes dans les 40 635 segments a été soumise à une décomposition en valeurs singulières et les 300 premiers vecteurs propres ont été conservés.

4.2 Méthode pour l'évaluation

Pour évaluer l'efficacité de méthodes visant à déterminer automatiquement la valence de mots, le test classique, lorsque l'étude est réalisée en anglais, se base sur les listes de mots positifs et négatifs incluses dans le *General Inquirer* (p.e., Dragut et al., 2010 ; Kamps et al., 2004 ; Turney, Littman, 2003). Ces listes n'étant pas, à notre connaissance, disponibles en français, nous avons recherché un matériel équivalent dans

[1] Le WOLF couvre 30 % du WordNet de Princeton (Mouton & Chalendar, 2010) et, selon nos calculs, le WordNet français couvre environ 25 % des synsets de la version 1.5 de WordNet.

[2] www.crisco.unicaen.fr/cgi-bin/cherches.cgi

cette langue. La section 4.2.1 décrit les normes de valence de Syssau et Font (2005). Ces normes présentent l'avantage d'avoir été récoltées dans des conditions rigoureuses et bien documentées, alors qu'on ne dispose de pratiquement aucune information sur la procédure suivie pour constituer les deux listes du *General Inquirer*. Cependant, elles ne portent que sur 735 mots alors que les listes originales du *General Inquirer* en contiennent plusieurs milliers. À titre comparatif, nous avons réalisé une première adaptation française des listes du *General Inquirer*.

4.2.1 Valemo : V80, V50 et Vscore

Syssau et Font (2005) ont demandé à 600 juges d'évaluer 735 mots[3] sur deux échelles : une échelle nominale à trois modalités (négatif, neutre et positif) et une échelle bipolaire en 11 points allant de très négatif (-5) à très positif (+5) (voir Syssau et Font pour une discussion des avantages et inconvénients de ces deux types d'évaluation). Chaque mot a été évalué par 100 juges et un même juge n'a effectué qu'un seul des deux types d'évaluation. Les mots ont été sélectionnés sur la base de deux normes d'associations verbales de manière à constituer "un ensemble de mots suffisamment diversifié pour être représentatif de la langue française" (Syssau, Font, 2005). De la première évaluation, Syssau et Font ont dérivé deux normes catégorielles : les mots "indubitablement" positifs ou négatifs qui ont été classés dans la catégorie correspondante par au moins 80% des juges (V80) et les mots "majoritairement" positifs ou négatifs qui ont été classés ainsi par au moins 50% des juges (V50). La seconde évaluation a produit une norme valencée (Vscore) avec pour chaque entrée un score compris entre -5 et +5.

4.2.2 General Inquirer (version francisée) : GI

Le *General Inquirer* est un projet né en 1961 qui visait à développer un programme d'analyse objective de contenu (Stone et al., 1966) basé sur un dictionnaire composé de 182 catégories sémantiques. Les deux dernières catégories ajoutées sont les catégories positive et négative, qui répertorient respectivement 1915 et 2291 mots. Ces listes n'étant pas, à notre connaissance, disponibles en français, nous les avons traduites automatiquement à l'aide du traducteur en ligne Systran. Après avoir été lemmatisées avec TreeTagger, ces deux listes ont été contrôlées par deux juges. Après suppression des doublons et des mots présents dans les deux listes – problèmes présents dans la version originale, mais également dus à la traduction – , nous avons obtenu 1246 mots positifs et 1527 mots négatifs.

5 Résultats

Cinq normes ont été employées pour comparer l'efficacité des méthodes de construction automatique de lexiques dans l'estimation de la valence de mots : la norme Nev, les trois normes issues du projet Valemo (Vscore, V50 et V80) et notre traduction des listes positive et négative du *General Inquirer* (GI). Pour les deux normes qui définissent la valence comme une variable continue (Nev et Vscore), nous avons évalué la qualité de la prédiction en calculant le coefficient de corrélation de Pearson entre les valences prédites par les méthodes automatiques et les valeurs moyennes attribuées par les juges. Lorsque la variable à prédire est dichotomique (positif versus négatif : V50, V80 et GI), nous avons employé comme mesure d'efficacité le pourcentage de mots classés par les procédures automatiques dans la catégorie déterminée par la norme. Pour chacune des méthodes évaluées, un mot est considéré comme négatif lorsque sa valence prédite est inférieure à la moyenne et comme positif dans le cas contraire[4].

La principale difficulté que nous avons rencontrée lors de ces analyses trouve son origine dans le fait que les différentes méthodes testées ne donnent pas des valeurs de valence aux mêmes mots : celles dérivées de Kamps et Marx (2002) en proposent un nombre nettement plus restreint que celles qui s'appuient sur l'ASL. Ceci nous a conduits à présenter séparément les résultats de ces deux groupes de méthodes.

[3] La norme initiale portait sur 605 mots, mais elle a été ultérieurement étendue à 735 mots. Elle est disponible à l'adresse : http://www.lexique.org/

[4] Des analyses complémentaires ont montré que ce seuil était proche de la valeur optimale obtenue par régression logistique.

5.1 Approche basée sur le dictionnaire de synonymes : KA1 et KA7

Le tableau 1 présente les performances des méthodes KA1 et KA7 pour les différentes normes. Pour tous les tests, KA7, la version basée sur les 7 paires de mots germes de Turney et Litman (2003), est supérieure à KA1 qui n'emploie qu'une seule de ces paires, celle sélectionnée par Kamps et Marx (2004). Les corrélations entre la valence prédite par les méthodes et la valence moyenne selon les juges sont élevées et même très élevées pour Vscore. Pour la prédiction de la catégorie des mots, les performances sont également impressionnantes pour les trois tests. Dans leur étude sur l'anglais, Kamps et al. (2004) rapportent un pourcentage de mots bien classés par leur procédure de 67 % pour les 667 adjectifs pour lesquels ils ont pu calculer un score d'évaluation à partir de WordNet et qui se trouvent dans la liste du *General Inquirer* (Evaluation II, Table 1 dans Kamps et al., 2004). Cette valeur est nettement inférieure à celle que nous avons obtenue. S'il est difficile d'identifier précisément l'origine de l'amélioration, force est de constater que l'implémentation de la technique de Kamps et Marx sur la base d'un dictionnaire de synonymes plutôt que de WordNet est une alternative viable.

	Nev	Vscore	V80	V50	GI
N	663	76	20	43	688
KA1	0.55	0.64	90%	84%	80%
KA7	0.61	0.72	100%	88%	84%

Tableau 1 : Performances (corrélation et pourcentage de classification correcte)

5.2 Approches basées sur l'ASL

Dans cette section, nous comparons la nouvelle méthode ASG à celles de Turney et Littman (2003) et de Bestgen (2002). Quatre versions différentes de ASG ont été testées. Elles se distinguent par l'étendue des germes potentiels pris en compte : ASG0.5 limite ceux-ci aux valeurs les plus extrêmes de la norme (de 1 à 1.5 et de 6.5 à 7), ASG1.0 est moins stricte et prend en compte celles comprises entre 1 et 2 et entre 6 et 7, ASGnorme prend en compte l'ensemble des mots repris dans la norme Nev et ASGtout sélectionne les germes parmi l'ensemble des termes présents dans l'espace sémantique. Pour construire le modèle prédictif sur la base de ces ensembles de germes potentiels, nous avons employé une régression linéaire multiple[5] avec sélection des prédicteurs par la technique ascendante (*forward*) et un seuil de probabilité pour la sélection fixé à 0.01.

5.2.1 Performances pour le matériel d'apprentissage : Nev

La première ligne du tableau 2 présente les corrélations entre les valeurs données dans la norme Nev, qui a servi pour l'apprentissage, et les valeurs prédites par les différentes méthodes. Comme on pouvait s'y attendre, SO-ASL, la seule des méthodes qui ne s'appuie pas sur la norme, obtient le moins bon résultat. Tout aussi attendus sont les bénéfices apportés par l'apprentissage supervisé (ASG versus DIC-ASL) et par la possibilité de choisir les germes parmi un nombre plus important de candidats. On note néanmoins que la différence principale se situe entre ASG0.5 et ASG1.0.

5.2.2 Performances pour Vscore

L'analyse de Vscore, deuxième ligne du tableau 2, donne comme attendu, des valeurs inférieures à celles obtenues pour la norme ayant servi à l'apprentissage, mais la différence est assez faible. On note tout particulièrement que les méthodes ASG sont nettement plus performantes que SO-ASL, ce qui confirme l'hypothèse que les mots germes employés par cette dernière sont loin d'être optimaux.

[5] Toutes les analyses ont également été effectuées en employant la SVR (SVM appliqué à la régression), mais ils ne sont pas présentés, car les deux techniques ont produit des résultats très similaires.

Normes	N	SO-ASL	DIC-ASL	ASG0.5	ASG1.0	ASGnorme	ASG,tout
Nev	2685	0.38	0.60	0.60	0.65	0.66	0.67
Vscore	631	0.32	0.60	0.56	0.61	0.61	0.60

Tableau 2 : Corrélation entre les valeurs prédites par les méthodes et les normes

5.2.3 Performances pour les catégories : V80, V50 et GI

Le tableau 3 présente le pourcentage de mots bien classés pour les différentes normes catégorielles. Les performances pour V80 et V50 sont très élevées, mais il faut prendre en compte le fait que ces deux normes ne contiennent qu'un nombre réduit de mots. Pour l'adaptation française du *General Inquirer*, les performances sont moins bonnes. Elles dépassent toutefois largement la performance de SO-ASL rapportée par Turney et Littman (2003) pour le *General Inquirer* en version anglaise (65%), valeur très proche de celle que nous avons obtenue pour l'adaptation française (64%). On observe aussi que DIC-ASL fait presque aussi bien que les méthodes basées sur une procédure d'apprentissage automatique.

Test	SO-ASL	DIC-ASL	ASG0.5	ASG1.0	ASGnorme	ASG,tout
V80 (N=128)	73%	88%	83%	87%	88%	91%
V50 (N=280)	63%	82%	78%	82%	84%	82%
GI (N=1992)	64%	71%	70%	72%	73%	72%

Tableau 3 : Pourcentage de classification correcte

Dans le tableau 3, tous les mots mentionnés dans les normes sont pris en compte, même ceux qui sont présents dans la norme Nev qui a servi à l'apprentissage supervisé. Il s'ensuit qu'il est problématique de se baser sur ces données pour évaluer les capacités de généralisation de la méthode ASG à des mots qui ne font pas partie du matériel d'apprentissage. Pour cette raison, les mêmes analyses que celles rapportées ci-dessus ont été effectuées après suppression dans les normes catégorielles de tous les mots présents dans Nev. Les résultats sont présentés dans le tableau 4. Pour GI, on observe une diminution assez faible et relativement égale des performances pour toutes les méthodes, y compris celles qui n'ont pas recours à l'apprentissage supervisé. Pour V50 et surtout V80, les différences sont plus nettes et s'observent même pour SO-ASL, alors que cette méthode ne s'appuie pas sur la norme Nev. L'explication la plus probable est que les mots qui ont été supprimés sont particulièrement faciles à classer par toutes les méthodes.

Test	SO-ASL	DIC-ASL	ASG0.5	ASG1.0	ASGnorme	ASG,tout
V80 (N=25)	60%	80%	68%	72%	72%	76%
V50 (N=82)	60%	82%	72%	73%	78%	74%
GI (N=1130)	62%	68%	68%	71%	71%	70%

Tableau 4 : Pourcentage de classification correcte pour les mots non inclus dans Nev

D'une manière générale, ces tests confirment le caractère non optimal des mots germes employés dans l'approche SO-ASL, cette méthode atteignant un niveau de performance nettement inférieur à celui atteint par toutes celles basées sur l'apprentissage supervisé de germes.

5.3 Comparaison globale

Une dernière série d'analyses visent à comparer le plus rigoureusement possible les performances de toutes les procédures testées, y compris KA1 et KA7, sur une même tâche afin de les rendre comparables. On a donc calculé le pourcentage de termes bien classés pour les mots de GI traités par toutes les méthodes. Le tableau 5, qui présente ces résultats, souligne la supériorité de KA7 sur toutes les autres méthodes. Il faut toutefois garder à l'esprit que KA7 propose au maximum des valeurs pour 688 mots du GI alors que les méthodes basées sur l'ASL traitent 1992 mots de cette même liste. De plus, nous n'avons employé qu'un seul espace sémantique d'un genre très spécifique (voir discussion). Les mêmes analyses ont été réalisées en supprimant, en plus, les mots qui sont dans la norme NEV, sans que les conclusions ne soient modifiées (différences plus petites ou égales à 2%).

N	SO-ASL	DIC-ASL	ASG0.5	ASG1.0	ASGnorme	ASGtout	KA1	KA7
550	64%	70%	75%	75%	76%	75%	80%	83%

Tableau 5 : Pourcentage de classification correcte pour les mots de GI traités par toutes les méthodes

5.4 Mots germes les plus importants pour prédire la valence

Si la méthode ASG n'apparait pas comme nettement supérieure à DIC-ASL, elle présente un avantage potentiellement très important en termes d'identification de mots germes. Alors que DIC-ASL sélectionne les germes localement puisqu'un ensemble différent de germes est employé pour chaque mot, ASG sélectionne les germes globalement : un seul et même ensemble de germes est employé pour prédire la valence de tous les mots. Il reste cependant à montrer que les germes choisis par ASG sont bien pertinents.

Une première manière de répondre à cette question consiste à s'intéresser au modèle prédictif construit par la régression multiple. Faute de place, il n'est pas possible de reprendre ici tous les mots germes sélectionnés par les différentes versions de ASG. La liste suivante présente l'ensemble des germes sélectionnés par ASG1.0, suivant l'ordre dans lequel ils ont été introduits dans le modèle (chaque fois suivi par la valence selon la norme Nev) : *épouvantable* (1.8), *délicieux* (6.2), *irriter* (1.9), *admiration* (6.1), *affectueux* (6.2), *atroce* (1.5), *heureux* (6.5), *monstrueux* (1.4), *magnifique* (6.5), *embrasser* (6.4), *lugubre* (1.8), *rêver* (6.3), *libre* (6.3), *savourer* (6.0), *ennui* (1.7), *intéressant* (6.0), *indifférence* (2.0), *espoir* (6.1), *pire* (1.4), *fidèlement* (6.1), *gaieté* (6.4), *rat* (1.9), *insulte* (1.6), *maladie* (1.5), *laideur* (1.6), *enlacer* (6.4), *enfant* (6.3), *crasse* (1.8), *voyage* (6.2), *malchance* (1.6), *admirable* (6.1).

L'analyse qui précède repose sur le modèle prédictif construit par la régression multiple. Celui-ci correspond à la meilleure combinaison possible de mots germes pour prédire la norme et non aux mots germes qui apportent individuellement la contribution la plus importante à la prédiction de celle-ci. Tout particulièrement, la régression multiple ne sélectionnera qu'un seul de deux mots sémantiquement très liés, même si tous les deux sont d'excellents prédicteurs (cf. *rage* et *colère* dans le tableau 6). Or, comme notre objectif prioritaire est d'identifier des mots germes spécifiques qui pourraient être ensuite employés dans d'autres méthodes, comme celle de Kamps et Marx (2002), il semble préférable de s'intéresser à ces derniers et donc à ceux dont le vecteur de cosinus (avec les mots présents dans la norme) est le plus corrélé avec la valence de ces mots. Le tableau 6 présente, à titre d'exemple, une petite fraction des germes les plus importants pour prédire la valence, classés par ordre d'efficacité, lorsqu'on prend en compte l'ensemble des mots présents dans l'espace sémantique. La partie gauche reprend les 30 germes les plus corrélés négativement avec la valence et la partie droite les 30 germes les plus corrélés positivement. La quasi-totalité des germes négatifs mentionnés dans ce tableau correspond à ce qu'on entend habituellement par mots germes pour la valence[6]. La grande majorité des germes positifs sont aussi pertinents et plus de la moitié d'entre eux ne se trouve pas dans la norme ayant servi à l'apprentissage (signalé par un "-" à la place du score de valence). Cette observation souligne la valeur heuristique de la méthode proposée. On y trouve néanmoins quelques mots spécifiques à la collection de textes employée pour l'ASL (*mythologique*, *nymphe*, *pampre*). Il est à noter que les germes qui suivent, par ordre d'importance, ceux présentés dans le

[6] Il n'est pas possible, à ce stade de l'analyse, de déterminer le nombre de cas dans lesquels *débattre* correspond à *se débattre*. Il s'agit là d'une limite évidente des prétraitements effectués avant l'extraction de l'espace sémantique

tableau semblent tout aussi pertinents. À titre d'exemple, on trouve de 10 en 10 pour l'orientation négative : 31. *brute*, 41. *monstrueux*, 51. *exécration*, 61. *exaspération*, 71. *désespérer*, 81. *sourd*, 91. *égorgement*, 101. *râle*.

	Négatif	Nev		Négatif	Nev		Positif	Nev		Positif	Nev
1	rage	2.1	16	infamie	-	1	charmant	5.7	16	charme	6.1
2	colère	2.2	17	imprécation	-	2	charmer	5.8	17	description	-
3	épouvantable	1.8	18	tourmenteur	-	3	ravissant	6.4	18	modeste	-
4	fureur	2.8	19	lâche	1.1	4	délicieux	6.2	19	ravir	5.6
5	atroce	1.5	20	menaçant	-	5	gracieux	5.9	20	admirable	6.1
6	horrible	1.8	21	menace	-	6	merveille	6.1	21	romance	4.4
7	abominable	1.9	22	épouvanter	2.0	7	magnifique	6.5	22	nymphe	-
8	écraser	1.9	23	saigner	-	8	brillant	-	23	exquis	-
9	horreur	2.1	24	cracher	-	9	aimable	5.9	24	distingué	-
10	crachat	1.3	25	débattre	-	10	harmoniser	-	25	pampre	-
11	exaspérer	2.4	26	effrayant	2.4	11	élégant	-	26	enchanter	-
12	misérable	1.8	27	plainte	-	12	riant	-	27	exotique	-
13	étrangler	-	28	crever	1.7	13	splendide	-	28	raffoler	-
14	affreux	1.9	29	meurtre	1.4	14	mythologique	-	29	modestement	-
15	assassin	-	30	injurier	1.9	15	composer	-	30	fraîcheur	-

Tableau 6 : Mots germes sélectionnés par la méthode ASG

6 Conclusion

Pour conclure, nous avons transposé au français deux méthodes de construction automatique de lexiques porteurs de valences bien établies dans le monde anglo-saxon : celles de Turney et Littman (2003) et de Kamps et Marx (2002). Cette dernière montrant des résultats encourageants, nous l'avons étendue en augmentant le nombre de paires de mots germes. Cette modification nous a permis d'obtenir les meilleurs résultats, avec plus de 80 % de termes bien classés. Ce pourcentage doit cependant être relativisé dans la mesure où il est calculé sur un nombre restreint de mots. Nous avons également développé une méthode qui sélectionne les mots germes par apprentissage supervisé. Avec une efficacité d'environ 75 %, elle surpasse nettement la méthode SO-ASL dont elle est dérivée. Il est, hélas, impossible de déterminer si les valeurs obtenues reflètent un niveau de performance proche de celui atteint par des annotateurs parce qu'on ne dispose pas d'information à propos du degré d'accord entre ceux-ci. L'analyse des mots apportant la plus grande contribution individuelle à la prédiction de la valence souligne l'intérêt de cette méthode pour l'identification de mots germes. Un des principaux développements envisagés est d'utiliser ces mots germes dans des méthodes comme celles de Kamps et Marx (2002) ou d'Esuli et Sebastiani (2006). Des adaptations seront nécessaires puisque, dans la version actuelle, les mots germes identifiés ne forment pas des couples comme requis par la méthode de Kamps et Marx. Il sera tout particulièrement intéressant de déterminer si la méthode proposée, qui ne requiert pas WordNet, est plus efficace que celle développée par Esuli et Sebastiani et, surtout, si l'emploi dans leur méthode des mots germes identifiés par ASG améliore encore les performances. Enfin, il sera nécessaire d'évaluer les bénéfices apportés par l'apprentissage supervisé de germes pour l'objectif principal de ce genre d'études : déterminer l'orientation de textes (Harb et al., 2008).

Cette étude comporte plusieurs limitations qui sont autant de pistes pour des recherches futures. Tout d'abord, un seul espace sémantique, extrait de textes littéraires, a été exploité. Les implications de cette limitation sont particulièrement mises en évidence par la sélection de mots germes spécifiques à ce genre de textes. Il serait intéressant d'effectuer ces analyses sur un corpus plus diversifié ou, séparément, sur des corpus de genres différents. Dans ce dernier cas, il devrait être possible d'attribuer aux mots germes un indice qui traduit leur degré de généralité. Ensuite, les germes identifiés par la méthode ASG consistent en des formes (lemmes) isolées, ce qui réduit fortement la qualité linguistique de l'analyse (voir *débattre*). La prise en compte de mots composés ou d'expressions figées serait également un développement intéressant (Vernier, Monceaux, 2010). D'autres méthodes pour mesurer les proximités sémantiques devraient également être testées. Il est en effet loin d'être évident que le passage par l'ASL améliore l'efficacité (Bestgen, 2006). Enfin, notre traduction des listes du *General Inquirer* pourrait sans aucun doute être améliorée afin de récupérer un certain nombre de mots perdus. Cependant, on peut s'interroger sur l'utilité d'un tel travail, étant donné le peu d'information disponible sur la procédure de construction de ces listes. Il

nous semble plus intéressant pour la communauté scientifique d'étendre les normes V50 et V80, dont la rigueur et les procédés de construction sont bien établis.

Remerciements

Yves Bestgen est chercheur qualifié du F.R.S-FNRS. Les auteurs remercient vivement A. Syssau pour les explications complémentaires à propos de la norme *Valemo* et l'équipe du CRISCO pour l'autorisation d'extraction des informations incluses dans le dictionnaire de synonymes.

Références

ABBSASI, A., CHEN, H., SALEM, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems* 26.

BESTGEN, Y. (2002). Détermination de la valence affective de termes dans de grands corpus de textes. Actes de *CIFT*'02, 81-94.

BESTGEN, Y. (2006). Déterminer automatiquement la valence affective de phrases : Amélioration de l'approche lexicale. Actes des *JADT 2006*, 179-188.

BESTGEN, Y. (2008). Building affective lexicons from specific corpora for automatic sentiment analysis. Proceedings of *LREC 2008*, 496-500.

CHARDON, B. (2010). Catégorisation automatique d'adjectifs d'opinion à partir d'une ressource linguistique générique, Actes de *RECITAL 2010*.

CHESLEY, P., VINCENT, B., XU, L., SRIHARI, R.K. (2006). Using verbs and adjectives to automatically classify blog sentiment. Proceedings of *AAAI-CAAW-06*, 27-29.

DEERWESTER, S., DUMAIS, S.T., FURNAS, G.W., LANDAUER, T.K., HARSHMAN, R. (1990). Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science* 41, 391-407.

DRAGUT, E.C., YU, C., SISTLA, P., MENG, W. (2010). Construction of a sentimental word dictionary. Proceedings of *ACM ICIKM*, 1761-1764**.**

ESULI, A., SEBASTIANI, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. Proceedings of *LREC'06*, 417–422,.

HARB, A., PLANTIE, M., ROCHE, M., DRAY, G., TROUSSET, F., PONCELET, P. (2008). Détection d'opinion. Comment déterminer les adjectifs d'opinion d'un domaine donné? *Document numérique* 11, 37-61.

HATZIVASSILOGLOU, V., MCKEOWN, K.R. (1997). Predicting the semantic orientation of adjectives. Proceedings of *EACL 1997*, 174-181.

HEISE, D.R. (1965). Semantic differential profiles for 1000 most frequent english words. *Psychological Monographs* 79, 1-31.

HOGENRAAD, R., BESTGEN, Y., NYSTEN, J.L. (1995). Terrorist Rhetoric : Texture and Architecture, In Nissan et Schmidt (Eds.), *From Information to Knowledge*,48-59, Intellect

HU, M., LIU, B. (2004). Mining Opinion Features in Customer Reviews. Proceedings of *AAAI*, 755-760.

KAMPS, J., MARX, M. (2002). Words with Attitude. Proceedings of *the 1st Interational Conference on Global WordNet,* 332-341.

KAMPS, J., MARX, M., MOKKEN, R.J., DE RIJKE, M. (2004). Using WordNet To Measure Semantic Orientations Of Adjectives. Proceedings of *LREC 2004*, 1115-1118.

KIM, S.M., HOVY, E. (2004). Determining the sentiment of opinions. Proceedings of *COLING*, 1367-1373.

MANQUIN, J.L., FRANÇOIS, J., EUFE, R., FESENMEIER, L., OZOUF, C., SENECHAL, M. (2004). Le dictionnaire électronique des synonymes du CRISCO : un mode d'emploi à trois niveaux. *Les Cahiers du CRISCO* 17, 1-64.

MILLER, G.A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography* 3, 235–312.

MOUTON, C., CHALENDAR, G. (2010). JAWS : Just AnotherWordNet Subset. Actes de *TALN 2010*.

NASUKAWA, T., YI, J. (2003). Sentiment analysis: capturing favorability using natural language processing. Proceedings of *the 2nd international conference on Knowledge capture (K-CAP)*, 70-77.

PAK, A., PAROUBEK, P. (2010). Construction d'un lexique affectif pour le français à partir de Twitter. *Actes de TALN 2010*.

PANG, B., LEE, L., VAITHYANATHAN, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 79-86.

SCHMID, H., (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of *the International Conference on New Methods in Language Processing*, 44-49.

STONE, P.J., DUNPHY, D.C., SMITH, M.S., OGILVIE, D.M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge : MIT Press.

SYSSAU, A., FONT, N. (2005). Evaluations des caractéristiques émotionnelles d'un corpus de 604 mots. *Bulletin de Psychologie* 58, 361-367.

TURNEY, P.D., (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proceedings of *the 40th Annual ACL Meeting*, 417-424.

TURNEY, P.D., LITTMAN, M. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *Technical Report*, National Research Council Canada.

TURNEY, P.D., LITTMAN, M. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems* 21, pp. 315--346

VELIKOVICH, L., BLAIR-GOLDENSOHN, S., HANNAN, K., MCDONALD, R. (2010). The Viability of Web-derived Polarity Lexicons. Proceedings of *NAACL 2010*, 777-785.

VERNIER, M., MONCEAUX, L. (2010). Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques. *Traitement automatique des langues* 51, 125-149.

VERNIER, M., MONCEAUX, L., DAILLE, B., DUBREIL, E. (2009). Catégorisation sémantico-discursives des évaluations exprimées dans la blogosphère. Actes de *TALN 2009*.

WIEBE, J., WILSON, T., BRUCE, R., BELL, M., MARTIN, M. (2004). Learning subjective language. *Computational Linguistics* 30, 277-308.

WIEBE, J., WILSON, T., CARDIE, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39, 165-210.

WILSON, T., WIEBE, J., HOFFMANN, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of *HLT-EMNLP 2005*, 347-354.

YU, H., HATZIVASSILOGLOU, V. (2003). Toward answering opinion questions : Separating facts from opinions and identifying the polarity of opinion sentences. Proceedings of *EMNLP 2003*, 129-136.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Comparaison d'une approche miroir et d'une approche distributionnelle pour l'extraction de mots sémantiquement reliés

Philippe Muller[1, 2] Philippe Langlais[3]
(1) IRIT, Université Paul Sabatier
(2) Alpage, INRIA Paris-Rocquencourt
(3) RALI / DIRO / Université de Montréal
muller@irit.fr, felipe@iro.umontreal.ca

Résumé. Dans (Muller & Langlais, 2010), nous avons comparé une approche distributionnelle et une variante de l'approche miroir proposée par Dyvik (2002) sur une tâche d'extraction de synonymes à partir d'un corpus en français. Nous présentons ici une analyse plus fine des relations extraites automatiquement en nous intéressant cette fois-ci à la langue anglaise pour laquelle de plus amples ressources sont disponibles. Différentes façons d'évaluer notre approche corroborent le fait que l'approche miroir se comporte globalement mieux que l'approche distributionnelle décrite dans (Lin, 1998), une approche de référence dans le domaine.

Abstract. In (Muller & Langlais, 2010), we compared a distributional approach to a variant of the mirror approach described by Dyvik (2002) on a task of synonym extraction. This was conducted on a corpus of the French language. In the present work, we propose a more precise and systematic evaluation of the relations extracted by a mirror and a distributional approaches. This evaluation is conducted on the English language for which widespread resources are available. All the evaluations we conducted in this study concur to the observation that our mirror approach globally outperforms the distributional one described by Lin (1998), which we believe to be a fair reference in the domain.

Mots-clés : Sémantique lexicale, similarité distributionnelle, similarité traductionnelle.

Keywords: Lexical Semantics, distributional similarity, mirror approach.

1 Introduction

Collecter les relations entre les entités lexicales en vue de construire ou de consolider un thésaurus est une activité qui possède une longue tradition en traitement des langues. Les efforts les plus importants ont été dédiés à la recherche de synonymes, ou plus exactement des "quasi-synonymes" (Edmonds & Hirst, 2002), c'est-à-dire des entrées lexicales ayant un sens similaire dans un contexte donné. D'autres relations comme l'antonymie, l'hyperonymie, l'hyponymie, la méronymie ou l'holonymie ont également été étudiées. Certains thésaurus, comme Moby que nous utilisons ici, listent de plus des relations qui sont difficiles à caractériser.

De nombreuses ressources ont été utilisées pour parvenir à acquérir de tels thésaurus. Les dictionnaires électroniques ont tout d'abord été investis, soit pour en extraire des relations sémantiques au niveau lexical (Michiels & Noel, 1982), soit pour définir des mesures de similarité sémantiques entre les entités lexicales (Kozima & Furugori, 1993). L'analyse distributionnelle, qui compare les mots à travers leur contexte d'usage, est également une ressource populaire pour la réalisation d'une mesure de similarité sémantique (Niwa & Nitta, 1994; Lin, 1998).

Plusieurs approches ont montré l'intérêt d'utiliser des corpus dans plusieurs langues et plus particulièrement des corpus parallèles. Dans ces travaux, les entrées lexicales sont dites similaires lorsqu'elles sont alignées avec les mêmes traductions dans une autre langue (van der Plas & Tiedemann, 2006; Wu & Zhou, 2003). Une variante de ce principe proposée par Dyvik (2002) considère comme sémantiquement reliés les mots d'une langue qui sont traduction d'un même mot dans une autre langue ; ces mots sont appelés par l'auteur des *traductions miroir*. Des variantes de cette approche ont été étudiées pour l'acquisition de paraphrases, qui porte sur des associations d'expressions de plusieurs mots : voir par exemple (Bannard & Callison-Burch, 2005) et (Max & Zock, 2008).

Les évaluations des travaux à base de similarité lexicale sont souvent décevantes : différents types de relations lexicales sont typiquement identifiés, qu'il est difficile de distinguer automatiquement. Des travaux comme ceux

de Zhitomirsky-Geffet & Dagan (2009) tentent dans une étape de post-traitement de sélectionner les relations les plus pertinentes qui caractérisent des paires de mots similaires. D'autres, comme Wu & Zhou (2003) tentent de combiner le résultat de différents processus d'extraction de mots reliés (approche distributionnelle, dictionnaires, etc.).

Notre étude s'inscrit dans ce dernier courant. Nous poursuivons l'étude amorcée par Muller & Langlais (2010) où une variante de Dyvik, faisant usage de modèles de traduction statistiques entraînés sur de grands volumes de données, est combinée à une approche distributionnelle. Contrairement à ce travail, nous nous intéressons ici à la langue anglaise pour laquelle des ressources sont disponibles en plus grand nombre. Ceci nous permet de mener une évaluation à l'état de l'art de l'approche miroir, que nous comparons au thésaurus produit par l'approche décrite dans (Lin, 1998) et que Lin tient à la disposition de la communauté. Nous montrons que l'approche miroir se comporte favorablement par rapport à l'approche distributionnelle, et ce, selon différentes évaluations que nous avons menées.

Dans la suite de cet article, nous présentons les ressources mises à profit en section 2 et notre protocole expérimental en section 3. Nous analysons nos résultats en section 4 et discutons les travaux reliés en section 5. Nous concluons cette étude et en dressons les perspectives en section 6.

2 Ressources

Nous avons utilisé deux bases lexicales dans ce travail :

– La base lexicale WordNet [1] que nous interrogeons à travers l'API de NLTK [2]. WordNet encode les relations de synonymie (`gain` / `acquire`), d'antonymie (`gain` / `lose`), d'hyperonymie/hyponymie (`odor` / `stench`) et d'holonymie/méronymie (`wood` / `tree`). Chaque entrée lexicale dans WordNet possède une moyenne de 5 à 6 synonymes et de 8 à 10 termes reliés, toutes relations confondues.
– Le thésaurus Moby [3] est une ressource plus étoffée que WordNet : chaque mot dispose en effet d'environ 80 mots reliés en moyenne. La nature des relations n'est cependant pas annotée.

Afin de comparer les approches miroir et distributionnelle, nous avons sélectionné de manière aléatoire deux ensembles de 1000 mots, un pour les noms et un pour les verbes. Nous appelons ces mots les "cibles" dans la suite. Nous avons imposé arbitrairement un seuil minimal de fréquence sur les mots cibles (> 1000). La fréquence des mots a été calculée à l'aide du corpus libre de droit Wacky [4], qui compte 2 milliards de mots. Les caractéristiques des deux ensembles de cibles ainsi construits sont décrites en table 1.

			nombre d'associations			
Pos	fréquence médiane	référence	moyen	médian	min	max
Noms	3 538	WordNet syns	3,63	2	1	36
Noms	3 538	Moby	73,87	57	3	509
Verbes	11 136	WordNet syns	5,57	4	1	47
Verbes	11 136	Moby	113,23	90	6	499

TABLE 1 – Caractéristiques des deux ensembles de cibles (noms et verbes) : fréquence médiane dans Wacky, nombre moyen de termes associés selon la référence spécifiée, nombre médian, minimum et maximum.

3 Protocole

Nous comparons les termes similaires produits soit par l'approche des miroirs (section 3.1), soit par l'approche distributionnelle (section 3.2). Chaque approche produit un ensemble de termes associés ou *candidats*, classés selon leur degré de similarité. Ces candidats ordonnés sont alors évalués au regard d'une ressource de référence

1. `wordnet.princeton.edu/wordnet`
2. `www.nltk.org`
3. `www.gutenberg.org/dirs/etext02/mthes10.zip`
4. `http://wacky.sslmit.unibo.it/doku.php`

(WordNet ou Moby), soit en gardant les n-meilleurs candidats, soit en gardant ceux dont le score de similarité dépasse un certain seuil (voir les détails plus loin).

À titre d'exemple, la figure 1 montre les candidats proposés par les deux approches pour le mot cible choisi aléatoirement `groundwork`. On observe la grande différence de couverture de WordNet et de Moby.

Candidats Miroir	Candidats Lin	WordNet	Moby
base	**preparation**	base	arrangement
basis	framework	basis	**base**
foundation	timetable	cornerstone	basement
land	rationale	foot	**basis**
ground	impetus	fundament	bed
job	modality	foundation	bedding
field	**foundation**	substructure	bedrock
plan	**prerequisite**	understructure	bottom
force	precondition		briefing
development	blueprint		cornerstone
			... [47 de plus]

FIGURE 1 – Dix premiers candidats proposés par les approches miroirs et distributionnelles pour le mot cible `groundwork`. Les synonymes selon WordNet ainsi qu'un sous ensemble des mots reliés selon Moby sont indiqués. Les candidats soulignés appartiennent à WordNet, tandis que ceux en gras sont présents dans Moby.

3.1 Approche miroir

L'approche miroir est fondée sur l'hypothèse que des mots d'une langue $\mathcal{E}$ qui sont fréquemment alignés avec le même mot dans une autre langue $\mathcal{F}$ sont sémantiquement proches. Dans l'exemple de la figure 2, les mots français `manger` et `consommer` sont tous les deux alignés avec le mot anglais `eat` et sont donc candidats à l'appariement sémantique.

Un	bébé	mange	toutes	les deux heures.
	Babies	eat	every	two hours.
	Canadians	eat	too much	poutine.
Les	Canadiens	consomment	trop de	poutine.

FIGURE 2 – Exemple de traductions miroir.

Notre variante de l'approche miroir repose sur la consultation de deux modèles de traduction statistique p_{e2f} et p_{f2e} qui donnent respectivement la probabilité qu'un mot français soit la traduction d'un mot anglais et la probabilité inverse. Nous calculons la vraisemblance qu'un mot anglais s (pour synonyme) soit relié sémantiquement à un mot anglais w, soit $p(\mathrm{s}|\mathrm{w})$:

$$p(\mathrm{s}|\mathrm{w}) \approx \sum_{\mathrm{f} \in \tau_{e2f}(\mathrm{w})} p_{e2f}^{\delta_1}(\mathrm{f}|\mathrm{w}) \times p_{f2e}^{\delta_2}(\mathrm{s}|\mathrm{f}) \qquad \tau_{e2f}(\mathrm{w}) = \{\mathrm{f} \,:\, p_{e2f}(\mathrm{f}|\mathrm{w}) > 0\}$$

Ici $\tau_{e2f}(\mathrm{w})$ désigne l'ensemble des mots français associés par le modèle p_{e2f} au mot anglais w. En pratique, les distributions lexicales utilisées étant bruitées, nous appliquons deux seuils δ_1 et δ_2 (fixés à 0.001 dans cette expérience) qui filtrent les associations peu probables d'un modèle :

$$p_\bullet^\delta(t|s) = \begin{cases} p_\bullet(\mathrm{t}|\mathrm{s}) & \text{si } p_\bullet(\mathrm{t}|\mathrm{s}) \geq \delta \\ 0 & \text{sinon} \end{cases}$$

D'autres façons de filtrer les tables de transfert pourraient être déployées. Nous pourrions par exemple utiliser un test de significativité afin de retenir les associations les plus pertinentes. Notre approche au filtrage est certainement perfectible mais présente l'avantage d'être particulièrement simple à mettre en œuvre.

Les modèles lexicaux p_{e2f} et p_{f2e} ont été entraînés sur un bitexte anglais-français de 8,3 millions de paires de phrases extraites des transcriptions des débats parlementaires canadiens (Hansard). Ce bitexte est exploité par le concordancier bilingue `TSRali` [5]. Nous avons lemmatisé les phrases anglaises et françaises du corpus à l'aide de `TreeTager` [6] avant d'entraîner dans les deux directions [7] (anglais→français et français→anglais) des modèles IBM 4 à l'aide de `Giza++` [8] utilisé dans sa configuration par défaut.

Dans l'évaluation qui suit, nous avons considéré les 200 premiers lemmes associés à chaque mot cible par cette approche car c'est le nombre de candidats que produit l'approche distributionnelle que nous avons testée (voir la section suivante).

3.2 Similarité distributionnelle

L'approche distributionnelle que nous utilisons est celle décrite par Lin (1998). Elle représente selon nous une approche de référence dans le domaine. Un thésaurus calculé par l'auteur à l'aide de cette méthode est disponible gratuitement [9].

Pour l'obtenir, Lin a fait usage d'un analyseur grammatical en dépendance afin de comptabiliser les occurrences de triplets `(lemme_de_tête, relation, lemme_dépendant)` où `relation` est une relation (syntaxique) de dépendance. À chaque lemme w est associé un vecteur de compte pour l'ensemble $F(w)$ des traits `(rel, autre_lemme)` où `autre_lemme` est soit un dépendant de w, soit un gouverneur de w.

Par exemple, le verbe `eat` est caractérisé par un ensemble de traits $F(eat)$ contenant `(has_subj,man)`, `(has_obj,fries)`, `(has_obj,pie)`, etc qui correspondent aux contextes syntaxiques de `eat`. Appelons c la fonction de comptage d'occurrence d'un triplet (w, rel, w') et V l'ensemble du vocabulaire, on pose :

$$c(\_, rel, w) = \sum_{w' \in V} c(w', rel, w) \qquad I(w, rel, w') = \log \frac{c(w, rel, w') \times c(\_, rel, \_)}{c(w, rel, \_) \times c(\_, rel, w')}$$

$$c(w, rel, \_) = \sum_{w' \in V} c(w, rel, w')$$

$$c(\_, rel, \_) = \sum_{w' \in V} c(\_, rel, w') \qquad ||w|| = \sum_{(r, w') \in F(w)} I(w, r, w')$$

$I(w, rel, w')$ est alors la spécificité d'une relation (w, rel, w'), définie comme l'information mutuelle entre les éléments du triplet (Lin, 1998). On note $||w||$ la quantité d'information totale associée à w. La similarité entre deux lemmes w_1 et w_2 mesure alors à quel point ils partagent des contextes syntaxiques spécifiques, en utilisant la quantité d'information des contextes qu'ils partagent, normalisée par la quantité d'information totale qu'on peut leur associer séparément.

$$sim(w_1, w_2) = \frac{\sum_{(r,w) \in F(w_1) \cap F(w_2)} [I(w_1, r, w) + I(w_2, r, w)]}{||w_1|| + ||w_2||}$$

D'après (Lin *et al.*, 2003), le corpus utilisé pour obtenir le thésaurus que nous avons utilisé ici serait de 3 milliards de mots, c'est-à-dire plus de 10 fois la taille du corpus que nous avons utilisé pour développer l'approche miroir. Il nous est apparu préférable de prendre ce thésaurus plutôt que de tester notre implémentation de l'approche que nous venons de décrire en particulier parce que nous ne disposons pas d'information sur les réglages des paramètres utilisés par les auteurs pour optimiser leurs sorties. En fait, notre implémentation de l'approche est de moins bonne qualité sur les jeux de tests que nous présentons que ceux obtenus à l'aide du thésaurus compilé par les auteurs. À tout le moins, nous soulignons que la comparaison de l'approche miroir avec l'approche distributionnelle n'est pas biaisée en faveur de l'approche miroir.

Le thésaurus calculé par Lin présente pour chaque mot cible les 200 lemmes les plus proches au sens de cette mesure de similarité.

5. `http://www.tsrali.com/`
6. `www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
7. Les modèles IBM ne sont pas symétriques.
8. `fjoch.com/GIZA++.html`
9. `webdocs.cs.ualberta.ca/~lindek/Downloads/sim.tgz`

4 Expériences

En suivant le protocole décrit plus haut, nous avons évalué la sortie des deux similarités (miroir et distributionnelle) en considérant soit les n-meilleurs candidats de chaque approche, soit en considérant ceux dont le score de similarité dépasse un seuil donné (que nous faisons varier). Nous avons séparé notre jeu de test en deux ensembles de manière à mesurer les différences entre les noms et les verbes : comme le montre la table 1, le nombre de synonymes et autres entités lexicales reliées varie fortement en fonction de la catégorie morpho-syntaxique.

Nous avons considéré lors de l'évaluation les seuls items communs à la référence et au lexique de la ressource utilisée pour le développement d'une mesure de similarité. Par exemple, WordNet contient des synonymes qui ne sont pas présents dans les Hansards que nous avons mis à profit pour développer l'approche miroir : ils sont simplement écartés de notre évaluation.

Les deux approches que nous comparons sont sensibles à la fréquence des mots cibles considérés. Dans les deux approches décrites, tous les sens d'un mot sont regroupés lors des calculs de la similarité et il est vraisemblable que les usages les plus fréquents dominent les autres dans ces calculs. Sachant qu'un mot fréquent a plus de chance d'être polysémique qu'un autre, nous souhaitions prendre en considération dans notre évaluation l'influence de la fréquence des mots cibles étudiés. Nous filtrons à cet effet les candidats dont la fréquence (telle que calculée à l'aide de Wacky) est inférieure à un seuil donné, pour un ensemble de ces valeurs seuils.

Il a été montré (Weeds, 2003) que la plupart des méthodes de similarité lexicale se comportent de façon très différente par rapport à ce critère, sélectionnant selon les réglages plutôt des mots fréquents ou plutôt des mots rares.

Nous avons remarqué la tendance de l'approche miroir à souvent proposer des mots "vides". Cela s'explique par le fait que ces mots sont souvent biens notés par les distributions lexicales que nous utilisons. Ce phénomène a été analysé notamment dans (Moore, 2004). Nous avons arbitrairement éliminé des listes de candidats miroirs les termes apparus dans plus de 25% des listes (ce seuil pourrait être ajusté à l'aide d'un corpus de développement). Ce filtre élimine des noms courants comme `thing` ou `way`, des verbes comme `have`, `be` ou `come`, ainsi que des mots sur-représentés dans les Hansards (ex. : `house`).

Au final, nous avons combiné les listes candidates produites par les deux approches en prenant l'intersection des deux listes. D'autres schémas de combinaison seront étudiés dans des travaux ultérieurs.

Deux aspects nous intéressent plus particulièrement dans cette expérience : la quantité de mots reliés dans la référence que nous sommes capables d'identifier par l'une des approches et la fiabilité avec laquelle ils sont identifiés. En d'autres termes, nous voulons que la tête de liste des candidats soit la meilleure possible au regard d'une liste de référence. Nous évaluons donc les deux approches à l'aide des taux de précision et de rappel [10] que nous mesurons en différents points. Nous résumons ces taux à l'aide des taux MAP (Mean Average Precision) et MRR (Mean Reciprocal Rank) couramment employés en recherche d'information. MAP calcule la précision en chaque point de la liste où un candidat pertinent est identifié ; MRR est calculé comme la moyenne de l'inverse des rangs du premier terme pertinent dans la liste.

Enfin, nous avons également calculé la précision de chaque approche en faisant l'hypothèse d'un "oracle" qui indique le nombre exact de candidats à proposer pour chaque mot cible (il s'agit dans notre cas du nombre de mots reliés dans la référence). Cette mesure est semblable à ce que l'on appelle la R-précision. Par exemple, les 10 candidats de la méthode des miroirs de la figure 1, évalués à l'aide de la référence WordNet, reçoivent une précision de 3/10, un rappel de 3/5 (et pas 3/8 car les mots `understructure`, `substructure` et `fundament` sont absents des Hansards). La R-précision est également de 3/5 car tous les candidats corrects sont proposés à un rang inférieur au nombre de mots reliés dans la référence (5 synonymes). La précision au rang 1 est de 1, alors que la précision au rang 5 est de 3/5. Finalement, le taux MAP est de 0,63 = 6,29/10 = (1/1 + 2/2 + 3/3 + 3/4 + . . .+ 3/10) / 10 alors que MRR est de 1 car le premier candidat est correct ; il serait de 1/2 si seulement le second candidat avait été correct, etc.

Il est apparu empiriquement qu'il était préférable de couper une liste candidate à un rang donné que d'essayer de seuiller en fonction d'un score de similarité, et nous détaillons donc uniquement par la suite les résultats avec la première méthode, en faisant varier le rang.

10. Que nous présentons sous forme de pourcentage pour plus de lisibilité.

4.1 WordNet

La table 2 montre les résultats pour les noms évalués selon les synonymes de WordNet. Pour chaque approche, nous indiquons les précisions aux rangs n=1, ..., 100 dans les listes candidates, les taux MAP, MRR, la R-précision, le nombre de synonymes dans la référence ($||ref||$) et le rappel global, pour les 200 premiers candidats de chaque méthode [11]. Nous rapportons également l'influence de différents filtres de fréquence. La ligne f>1000, par exemple, indique que nous retenons des listes candidates et de la référence les seuls mots dont la fréquence (dans Wacky) est supérieure à 1000.

n-meilleur(s)		P1	P5	P10	P20	P100	MAP	MRR	R-prec	$\|ref\|$	rappel
Miroir	f>1	16,4	5,1	3,8	2,7	1,3	11,9	15,1	16,6	2342	50,0
	f>5000	19,1	5,4	3,8	2,6	1,2	11,3	13,2	17,5	1570	54,8
	f>20000	22,1	5,7	3,9	2,5	1,1	9,8	11,4	22,7	1052	60,6
Lin	f>1	17,4	5,2	3,5	2,5	1,5	11,7	14,3	14,7	2342	35,9
	f>5000	16,5	5,0	3,5	2,5	1,6	9,2	10,8	16,7	1570	36,6
	f>20000	17,5	4,5	3,3	2,5	1,6	7,3	8,4	20,1	1052	36,9
M/L	f>1	25,8	7,5	5,7	4,4	3,8	15,9	17,6	22,0	2342	29,3
	f>5000	27,4	7,4	5,5	4,3	3,8	12,7	13,6	24,6	1570	31,1
	f>20000	26,1	6,4	4,7	3,5	2,6	9,7	10,4	28,9	1052	32,7

TABLE 2 – Résultats pour les noms, micro-moyennés, avec les synonymes de WordNet pour référence.

Comme WordNet répertorie peu de synonymes, les précisions à faible rang (1 et 5) sont les plus pertinentes, ainsi que la R-précision : les autres sont nécessairement très basses. Les autres mesures sont données à des fins de comparaison car elles sont plus pertinentes pour la référence Moby. Ceci étant noté, la table 2 amène plusieurs commentaires [12].

Premièrement, nous observons que la précision de l'approche miroir au rang 1 culmine à 22% alors que le rappel plafonne à un peu plus de 60% : un bien meilleur résultat combiné que l'approche distributionnelle que nous avons testée (moins de 18% de précision au rang 1 et moins de 37% de rappel). Deuxièmement, il apparaît clairement que filtrer les candidats les moins fréquents est beaucoup plus payant pour l'approche miroir [13]. C'est sans doute la conséquence d'un corpus de départ plus petit, pour lequel les occurrences rares de mots peu fréquents entraînent des probabilités d'alignement peu fiables. Troisièmement, nous observons que notre combinaison des deux approches, aussi simpliste soit-elle, s'accompagne d'une augmentation significative de la précision, notamment la R-précision (au détriment cependant du rappel).

Enfin, les résultats sur les verbes sont similaires à ceux présentés ici pour les noms, avec cependant une meilleure précision à rang faible et un compromis sur la fréquence de coupure plus élevé, et ce, même si la précision oracle est globalement la même pour toutes les configurations. Combiner les deux méthodes améliore la précision de manière similaire à ce que nous observons sur les noms, avec une précision oracle qui varie cette fois entre 20% et 27%. Les différences sont toutes significatives sauf cette fois sur P1 à fréquence élevée.

Nous ne montrons pas le détail des scores si on ajoute dans la référence les relations issues de toutes les fonctions lexicales de WordNet, mais on peut noter que les résultats sont très proches entre les deux méthodes sur les noms (R-prec≈13% et P@1=23% quand f>1). En revanche, les traductions miroirs sont légèrement meilleures sur les verbes en R-précision (16% contre 18% pour f>1 et 17% contre 21% pour f>20000), et inférieures en terme de précision au rang 1 (41% contre 37% pour f>1 à 33% contre 34%).

Nous montrerons en section 4.3 que la différence semble se jouer essentiellement sur les relations de synonymie et d'hyperonymie (et hyponymie).

11. Les candidats de Lin étant limités à ce nombre.

12. Sauf précision contraire, les différences entre méthodes discutées plus bas sont tous significatives à p<0.05. Pour toutes les mesures sauf P1 et le rappel global, nous avons utilisé le test de Wilcoxon sur les résultats mot par mot. Dans le cas de P1 qui donne un résultat binaire par cible, nous avons fait un test binomial.

13. Les différences significatives de précision et MAP entre les deux méthodes n'apparaissent que pour les valeurs de fréquence élevée.

4.2 Moby

La table 3 résume les résultats des deux approches pour les noms, en prenant cette fois-ci Moby pour référence. Les relations listées dans ce thésaurus étant du tout venant, nous nous attendions à ce que la référence soit plus proche des sorties produites par une approche distributionnelle. Nous observons que c'est bien le cas sur les noms : la précision de l'approche de Lin est systématiquement supérieure à celle des miroirs, avec presque 10 points de plus au rang 1 ; et ce, même si le rappel est légèrement en faveur de l'approche miroir. Sur les verbes, cependant, les deux approches se comportent de manière comparable. En observant la différence de scores de précision entre

n-meilleur(s)		P1	P5	P10	P20	P100	MAP	MRR	R-prec	$\|ref\|$	rappel
Miroir	f>1	33,7	15,8	13,3	11,0	7,0	18,5	40,1	11,0	60774	18,1
	f>5000	32,7	14,5	12,1	9,8	6,1	18,7	38,1	11,8	43294	21,6
	f>20000	30,3	13,2	10,7	8,6	5,3	18,1	34,9	12,8	28488	26,7
Lin	f>1	44,8	19,9	16,4	13,4	9,5	26,6	46,8	14,7	60774	15,4
	f>5000	40,7	18,5	15,0	12,5	9,3	25,6	41,6	15,0	43294	16,3
	f>20000	39,4	16,1	13,5	11,2	8,4	23,3	35,2	16,8	28488	16,8
M/L	f>1	53,1	25,1	21,4	18,1	35,2	46,6	22,9	25,0	60774	9,4
	f>5000	52,4	23,0	19,3	16,6	13,7	30,7	41,2	23,4	43294	10,9
	f>20000	45,9	19,4	16,5	14,0	11,2	24,6	32,6	21,6	28488	12,5

TABLE 3 – Résultats pour les noms, micro-moyennés, avec les mots reliés de Moby pour référence.

la table 3 et la table 2, il semble que les approches miroir et distributionnelle ramènent bien d'autres entités que des synonymes dans leurs meilleurs candidats. Cela peut sembler un peu surprenant pour l'approche miroir puisque cette approche capitalise à priori sur des relations de traduction. Nous devons analyser cela de façon plus précise afin de savoir si nous sommes en présence de bruit dans les modèles de traduction (ce qui est très probable) ou si Moby contient plus de synonymes que WordNet, ou les deux.

Nous observons également que le rappel de l'approche miroir est plus grand que celui de l'approche distributionnelle, une observation en accord avec notre évaluation sur WordNet, et qui est peut-être due à la différence de performance des deux approches sur les synonymes (qui sont nombreux dans Moby).

Le filtre des entités lexicales offre un rendement mitigé : la précision de la variante f>20000 est légèrement inférieure à celle de la variante f>1 pour les deux approches. Le rappel de l'approche miroir augmente cependant de manière notable et consistante dans les cas où l'on s'intéresse aux mots très fréquents.

4.3 Analyse des erreurs produites par l'approche miroir

Les expériences que nous venons de décrire possèdent quelques limites. La référence WordNet que nous utilisons pour la synonymie possède un nombre relativement restreint de synonymes par mot candidat, ce qui ne permet pas de rendre compte avec précision de la pertinence des autres candidats proposés. Le fait d'utiliser un thésaurus plus vaste comme Moby ne résout que partiellement le problème car la nature des relations encodées dans Moby n'est pas étiquetée et certaines relations présentes dans cette ressource ne correspondent pas à des relations lexicales typiques (ex : `raging` / `abandoned`).

On peut mener une première analyse à l'aide de WordNet afin d'évaluer la présence de termes reliés par d'autres relations que la synonymie. Si l'on regarde les premiers candidats produits pour chaque cible (voir la table 4), on constate que pour les verbes, 19% sont recensés comme hyperonymes et 6% comme hyponymes, les proportions étant de 7% et 4% pour les noms. Les autres fonctions (holonymes, méronymes et antonymes) apparaissent de façon marginale. En regardant les 5 premiers candidats produits par les deux approches, on observe que ceux produits pour les verbes correspondent davantage à des relations présentes dans WordNet. En grande majorité, les candidats identifiés ne correspondent pas à une relation étiquetée dans WordNet. Un peu moins de la moitié des cibles ne reçoit d'ailleurs aucun candidat validé par WordNet (Ø).

Les problèmes de couverture de WordNet se posent malheureusement pour toutes les fonctions lexicales et ceci ne peut être qu'indicatif. Nous avons donc conduit une évaluation manuelle de la sortie produite par l'approche

		top 5							top 1						
		Ø	S	He	Ho	Hl	A	M	Ø	S	He	Ho	Hl	A	M
nom	Miroir	3146	181	175	98	13	5	1	570	64	58	28	5	1	1
	Lin	3078	186	161	123	12	11	2	565	65	49	31	6	4	1
verbe	Miroir	2807	406	428	216	0	7	0	466	95	139	42	0	3	0
	Lin	2882	414	272	212	0	20	0	444	140	106	50	0	5	0

TABLE 4 – Nombre de fonctions lexicales correspondant aux 5 (colonne de gauche) ou 1 (colonne de droite) premiers candidats proposés par chaque approche sans filtre de fréquence, selon WordNet. Ø signifie qu'aucune relation selon WordNet n'est associée à un candidat. Ces occurrences sont comptabilisées pour les cibles traitées par les deux approches, soit 724 noms et 743 verbes. S=synonymes, He=hyperonymes, Ho=hyponymes, Hl=holonymes, A=antonymes, M=méronymes.

miroir en sélectionnant aléatoirement 100 paires de mots `cible` / `candidat` où le candidat est le premier proposé par l'approche miroir, bien qu'il ne soit pas validé comme synonyme par WordNet. Nous avons observé les phénomènes suivants :

– 25% des mots candidats constituent une partie d'une unité composée de plusieurs mots, comme le mot `sea` dans la paire `sea` / `urchin` ;
– 18% des mots candidats non validés par WordNet sont en fait des synonymes selon d'autres thésaurus que nous avons consulté manuellement [14]. C'est par exemple le cas de la paire `torso` / `chest` ;
– 13% des candidats sont en fait des hyperonymes listés dans WordNet ou dans `www.thesaurus.com`, comme la paire `twitch` / `movement` ;
– 6% des paires mettent en relation des mots morphologiquement reliés, comme `accountant` / `accounting`, probablement en raison d'un problème d'étiquetage en partie du discours dans la langue pivot où un mot français comme ici `comptable` peut être aussi bien être un nom qu'un adjectif.

Parmi les erreurs (au sens de WordNet) fréquentes restantes, certaines sont dues à la polysémie d'une traduction pivot, comme par exemple le mot anglais `aplomb` traduit en français par `assurance` qui veut également dire `insurance` en anglais. Ce type de problème est cependant difficile à analyser sans retracer méticuleusement les nombreuses associations utilisées par les modèles de traduction dans notre approche.

D'autres erreurs sont plutôt imputables à des termes peu fréquents dans le corpus des Hansards que nous avons utilisé et que nous aurions dû filtrer au préalable.

Cette analyse suggère que tous les candidats rejetés ne sont pas nécessairement mauvais et qu'il y a donc place à amélioration. La polysémie demeure le problème le plus difficile à résoudre, que ce soit pour notre approche miroir ou pour l'approche distributionnelle.

Nous avons également regardé de manière très informelle les mots cibles pour lesquels WordNet ne propose aucun synonyme alors que l'approche miroir propose des candidats. Dans une proportion non négligeable de cas, les traductions miroir sont pertinentes comme `whopper` / `lie`. Une analyse plus fine est cependant requise pour quantifier plus précisément cette observation.

4.4 Tests de synonymie

Comme évaluation secondaire, plusieurs auteurs utilisent, pour évaluer la pertinence d'une mesure de similarité sémantique, des tests de synonymie semblables à ceux posés dans les examens du TOEFL (Turney, 2008) où la tâche consiste à distinguer parmi quatre candidats, le synonyme d'un mot dans un contexte donné. On peut voir cet exercice comme une version simplifiée d'une tâche de désambiguïsation, où le but est de reconnaître le bon terme dans un ensemble de *distracteurs* (termes à priori sans rapport), au lieu de distinguer les sens d'un même mot. On teste alors la similarité sémantique en prenant celui des candidats qui a le score de similarité le plus élevé avec la cible.

14. Comme par exemple le Roget's 21st Century Thesaurus, `http://www.thesaurus.com`

Les données TOEFL ne sont pas librement disponibles, aussi avons nous utilisé ici un test généré artificiellement à partir des données de WordNet par Freitag *et al.* (2005) [15]. Les auteurs le considèrent comme plus difficile que les équivalents du TOEFL. Ce test est notamment utilisé par Ferret (2010) afin d'ordonner différentes mesures de similarité entre vecteurs de cooccurrences. Le meilleur score qu'il obtient sur ce test est de 71.6% de réponses correctes, ce qui est proche du score de 72% d'exactitude obtenu par (Freitag *et al.*, 2005) à l'aide d'autres méthodes distributionnelles. Les systèmes répondent à toutes les questions.

Les instances de test sont de la forme `house: family obstacle filing surgeon` le premier terme étant la cible, le second un synonyme d'un des sens de la cible selon WordNet, les trois autres sont des distracteurs. Quelques restrictions sont ajoutées pour ne pas rendre le test trop facile : les synonymes dont la forme est proche de la cible (`group`/`grouping`) sont éliminés. Par ailleurs les cibles sont choisies avec une fréquence minimale. Les distracteurs sont choisis complètement au hasard, mais les termes associés sont choisis parmi les synsets de WordNet et privilégient donc les termes polysémiques.

Nous avons appliqué ce test à nos deux approches en choisissant, comme les autres travaux mentionnés, celui des candidats ayant le meilleur rang dans la liste de similarités. Si aucun des candidats du test n'est présent dans les candidats d'une méthode, le système ne répond rien. Nous pouvons donc évaluer l'aptitude d'une mesure de similarité à identifier le bon terme, ce que nous mesurons en terme de précision, rappel et F-score.

La table 5 résume les résultats pour les 200 premiers candidats de chaque méthode. Dans les cas où les candidats sont tous absents de la réponse du système, (Ferret, 2010) renvoie une réponse au hasard, mais cela arrive rarement vue la couverture de son système. Nous avons ici fait le choix de considérer que le système ne répond pas faute de données fiables, car c'est un cas beaucoup plus courant ici. Ceci a pour effet de faire baisser le rappel, et la précision évalue réellement la méthode des miroirs.

Comme nous disposons pour l'approche miroir d'une liste de candidats plus étendue, nous avons aussi évalué cette méthode sans coupure. On constate un nombre important de non-réponses également, cette fois due sans doute à une couverture lexicale limitée de la ressource de départ (les Hansards). Noms et verbes regroupés, cette variante obtient un F-score de 0,73, avec 3908/17285 cibles sans réponse (22%), majoritairement des noms.

À nombres de candidats égaux, on constate donc que la méthode miroir a une précision équivalente à celle de Lin mais un rappel bien supérieur. Sans limite de candidats, elle atteint un F-score comparable aux meilleures méthodes distributionnelles testées dans les travaux susmentionnés.

Noms	F1	P	R	sans réponse
Lin[200]	0,55	0,95	0,38	5885/9887
Miroir[200]	0,63	0,95	0,47	4975/9887
Miroir	0,72	0,87	0,61	2995/9887

Verbes	F1	P	R	sans réponse
Lin[200]	0,55	0,87	0,40	3983/7398
Miroir[200]	0,69	0,89	0,56	2694/7398
Miroir	0,74	0,79	0,70	913/7398

TABLE 5 – Évaluation pour le test de synonymie basé sur WordNet de (Freitag *et al.*, 2005)

Faute de place, nous ne ferons que décrire brièvement une autre façon d'analyser le test effectué, proposée par (Freitag *et al.*, 2005), consistant à mettre les résultats en rapport avec le niveau de polysémie des termes cibles, et qui semblait mettre en évidence que les cibles polysémiques étaient les plus dures à résoudre. Nous n'avons pas constaté ce phénomène ici, la précision reste constante pour les verbes et ne baisse que très légèrement pour les noms, quelle que soit la polysémie des cibles, alors que le rappel augmente, sans doute parce que les miroirs sont plus susceptibles d'avoir une réponse à fournir sur les mots plus fréquents.

5 Travaux reliés

Plusieurs types de travaux peuvent être comparés à la présente étude, ayant des objectifs, des données en entrée et des méthodologies d'évaluation plus ou moins variés. L'extraction de paraphrases partage certains de nos objectifs et ressources, même si elle concerne le rapprochement de termes comportant plus d'une unité lexicale. L'extraction de synonymes, la construction de thésaurus recouvrent aussi nos buts et peuvent être évalués de façon similaire.

15. Ce test et est librement disponible à l'URL `http://www.cs.cmu.edu/~dayne/wbst-nanews.tar.gz`

De façon plus large, les nombreux travaux récents sur la conception et la réalisation de mesures de similarité sémantique peuvent être rapprochés naturellement de la méthode présentée, même si les objectifs sont différents.

L'évaluation de l'acquisition de paraphrases est souvent évaluée par des jugements humains d'acceptabilité des substitutions en contexte, ce qui limite à des petits jeux de test. Barzilay & McKeown (2001) rapportent que 90% des paraphrases extraites par patrons (sur un corpus monolingue de traductions littéraires) sont acceptables, mélangeant synonymes, hyperonymes et termes coordonnés, sans bien sûr pouvoir donner une idée de la couverture d'une telle méthode. En se fondant sur des alignements bilingues et une méthode similaire à la nôtre mais sur plusieurs unités lexicales, Bannard & Callison-Burch (2005) estiment que les meilleurs paraphrases extraites pour chaque cible sont valides dans 75% des cas avec un alignement parfait (48% avec un alignement automatique). De même Lin *et al.* (2003) ou Curran & Moens (2002) évaluent précisément la présence de synonymes dans des listes de similarité dans des petits ensembles de paires de synonymes ou antonymes, ce qui rend difficile une extrapolation sur le genre de données que nous utilisons afin d'atteindre une large couverture.

Plus proche de la méthodologie que nous avons suivie, on trouve des études qui évaluent la classification de paires de mots en synonymes ou non-synonymes. Cela peut être fait directement sur les candidats sélectionnés pour un ensemble de cibles, comme dans l'étude présente, ou sur des ensembles de test rééchantillonnés pour augmenter artificiellement la présence de paires positives et pouvoir appliquer des techniques standards de classification avec une fiabilité raisonnable. Ne pas rééchantillonner est plus réaliste mais donne des scores assez bas, comme nous l'avons constaté : (van der Plas & Tiedemann, 2006) partent de vecteurs d'alignement à la place des vecteurs d'arguments syntaxiques de Lin, en définissant la même similarité et atteignent 12% de F-score par rapport à leur référence ; (Wu & Zhou, 2003) fait de même, ajoutant aussi une distance calculée dans un graphe lexical issu d'un dictionnaire, et apprend des régressions linéaires des différents scores de similarité, tout en restreignant la fréquence des cibles jusqu'à obtenir un maximum de 23% sur les noms et 30% sur les verbes. On peut aussi mentionner (Heylen *et al.*, 2008), qui analysent la répartition des fonctions lexicales dans des listes de similarités de mot en néerlandais. La seconde option, où l'on rééchantillone à l'entraînement et au test, est pertinente seulement si l'on connaît un moyen de présélectionner naturellement les candidats pour atteindre la proportion supposée, ce qui n'est pas le cas pour les études existantes (Hagiwara *et al.*, 2009). Notre étude peut en fait être considérée comme une entrée pour des expériences de ce genre.

L'étude des similarités distributionnelles faite par (Ferret, 2010), qui utilise de la cooccurrence simple, montre des résultats proches de ce que l'on obtient avec les miroirs, plus bas sur WordNet et comparables ou meilleurs sur Moby. Il opère sur un jeu de test beaucoup plus large, sans distinguer les parties de discours, et le jeu de test est découpé différemment par rapport aux fréquences lexicales puisqu'il sélectionne les cibles et les candidats. Sur WordNet il obtient au mieux 11% de R-précision et 17% pour la meilleure P@1 (sur les mots les plus fréquents). Sur Moby, la meilleure R-précision est de 10% et la meilleure P@1 est de 41%, P@5 de 28%. Le rappel est systématiquement inférieur (25% sur WordNet et 10% sur Moby), mais seuls 100 candidats sont gardés par cible. Les résultats sont plus bas que ce que l'on obtient ici avec les données de Lin, et nous pouvons donc supposer que la comparaison que nous faisons est représentative de l'approche distributionnelle dans ce contexte. La combinaison miroir et distribution est par contre supérieure sur tous les scores sauf le rappel.

Avec assez peu de réglage, on voit donc que l'approche des miroirs atteint des résultats comparables ou meilleurs que les similarités d'alignement ou distributionnelle pour isoler des synonymes dans certaines configurations. Les approches distributionnelles peuvent sans doute être améliorées mais l'approche choisie semble représentative. Il faut noter que le calcul qui sous-tend les traductions miroirs est computationnellement bien plus simple que les calculs de similarité entre $n \times n$ vecteurs d'alignement ou de cooccurrences, où n est la taille du vocabulaire.

On peut estimer que l'on peut atteindre un niveau de filtrage des candidats qui rend possible de tenter ensuite la classification des paires restantes. D'après (Hagiwara *et al.*, 2009) [16] , on peut associer (par classification) une fonction lexicale de façon fiable à des paires de mots si la proportion de candidats effectivement à relier par rapport à ceux qui n'ont aucun lien peut atteindre un ratio de 1 pour 6. Une conclusion similaire est tirée par (Piasecki *et al.*, 2008) dans le cas de la détection d'hyperonymie. Nos résultats nous encouragent à penser que l'on peut atteindre une proportion de 1 pour 4 ou 5. L'étude citée ne précise pas la proportion de paires de mots synonymes/non synonymes de départ, mais si on prend les chiffres de (Ferret, 2010), il y a 30000 paires de synonymes sur un vocabulaire de référence de 10000 mots, donc pour environ 50M de paires possibles, soit une proportion de 0,06% de paires de synonymes ou un ratio de 1 pour 1600.

16. (Hagiwara *et al.*, 2009) utilise comme descripteurs des schémas syntaxiques et des vecteurs de cooccurrence.

Une autre façon de juger de la pertinence des mesures de similarité sémantique entre mots dérive des données collectées par (Miller & Charles, 1991) où on demande à des sujets de juger la similarité ou le lien entre des items lexicaux, sur une échelle numérique. C'est une façon intéressante de fournir une évaluation intrinsèque de ces associations, mais le jeu de test ne peut couvrir qu'une part très limitée du vocabulaire (300 mots environ, avec 2 ou 3 associations par mot au plus).

6 Conclusion

Nos expériences confirment la variété des relations lexicales que l'on peut récupérer en appliquant ce que l'on a usage d'appeler des mesures de similarité sémantique. Les deux approches que nous avons étudiées ici semblent corrélées aux ressources de référence que nous avons considérées.

En ce qui concerne les synonymes, nos expériences indiquent que les traductions miroir offrent des candidats plus pertinents que l'approche distributionnelle de Lin (1998). Dans la mesure où les approches miroir ne sont pas aussi prisées que les approches distributionnelles, nous espérons que cette étude contribuera à en montrer l'intérêt pour l'acquisition de relations lexicales. Nous soulignons de plus que l'approche miroir est beaucoup moins coûteuse à développer et à appliquer, pour autant que l'on soit en mesure de trouver des bitextes de taille suffisante mettant en jeu la langue d'intérêt. Patry & Langlais (2011) dressent un portrait des bitextes existants qui indique que de telles ressources sont de plus en plus disponibles pour de nombreuses paires de langues.

L'approche miroir que nous avons mise en place ne tire pas profit du fait que plusieurs bitextes mettant en jeu la langue d'intérêt sont disponibles. C'est par exemple le cas pour la langue française pour laquelle les bitextes des débats parlementaires européens sont disponibles en plus du bitexte que nous avons mis profit ici. L'ajout de telles ressources devrait être en mesure d'augmenter les performances (et la précision en particulier) de notre approche.

La complémentarité des approches testées dans cette étude amène à nous interroger sur la manière optimale de les combiner. La simple intersection que nous avons étudiée ici améliore nettement la précision des listes candidates. Une approche plus originale consisterait à combiner ces approches avec une approche par patron telle que celle de (Barzilay & McKeown, 2001). Le problème de la polysémie discuté en section 4.3 demeure un problème pour toutes les approches dont nous avons discuté, en particulier lorsque deux sens d'une entité lexicale sont fréquents en corpus. Il semble souhaitable d'intégrer l'apport de méthodes qui vise à repérer des groupes de sens équivalents multilingues, comme par exemple dans (Apidianaki, 2008).

Il n'en reste pas moins que notre objectif à moyen terme est de distinguer automatiquement la nature des différentes relations lexicales identifiées. Cette information est pertinente dans bon nombre d'applications (paraphrasage, choix d'une traduction, etc.). Des travaux comme ceux de (Hagiwara *et al.*, 2009) tendent à indiquer qu'il est envisageable d'entraîner de manière supervisée un classificateur à reconnaître certaines fonctions lexicales, pour autant que la proportion de candidats d'une classe particulière soit plus équilibrée que dans les distributions naturelles. Ceci indique qu'il faut être en mesure d'affiner la liste de candidats, ce que notre approche par filtrage ou combinaison réalise.

Remerciements

Nous remercions les relecteurs pour la pertinence de leurs commentaires.

Références

APIDIANAKI M. (2008). Translation-oriented Word Sense Induction Based on Parallel Corpora. In *Actes de LREC Language Resources and Evaluation (LREC)*, p. 3269–3275, Marrakech Maroc.

BANNARD C. & CALLISON-BURCH C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 597–604.

BARZILAY R. & MCKEOWN K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*.

CURRAN J. R. & MOENS M. (2002). Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, p. 59–66.

DYVIK H. (2002). Translations as semantic mirrors : From parallel corpus to wordnet. In *The Theory and Use of English Language Corpora, ICAME 2002*. http ://www.hf.uib.no/i/LiLi/SLF/Dyvik/ICAMEpaper.pdf.

EDMONDS P. & HIRST G. (2002). Near-Synonymy and lexical choice. *Computational Linguistics*, **28**(2), 105–144.

FERRET O. (2010). Testing semantic similarity measures for extracting synonyms from a corpus. In *Proceedings of LREC 2010*.

FREITAG D., BLUME M., BYRNES J., CHOW E., KAPADIA S., ROHWER R. & WANG Z. (2005). New experiments in distributional representations of synonymy. In *Proceedings of CoNLL*, p. 25–32.

HAGIWARA M., OGAWA Y. & TOYAMA K. (2009). Supervised synonym acquisition using distributional features and syntactic patterns. *Journal of Natural Language Processing*, **16**(2), 59–83.

HEYLEN K., PEIRSMAN Y., GEERAERTS D. & SPEELMAN D. (2008). Modelling Word Similarity. An Evaluation of Automatic Synonymy Extraction Algorithms. In *Proceedings of LREC 2008*, p. 3243–3249 : ELRA.

KOZIMA H. & FURUGORI T. (1993). Similarity between words computed by spreading activation on an english dictionary. In *Proceedings of the conference of the European chapter of the ACL*, p. 232–239.

LIN D. (1998). Automatic retrieval and clustering of similar words. In *COLING/ACL98*, volume 2, p. 768–774, Montreal.

LIN D., ZHAO S., QIN L. & ZHOU M. (2003). Identifying synonyms among distributionally similar words. In *Proceedings of IJCAI'03*, p. 1492–1493.

MAX A. & ZOCK M. (2008). Looking up phrase rephrasings via a pivot language. In *Coling 2008 : Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*, p. 77–85.

MICHIELS A. & NOEL J. (1982). Approaches to thesaurus production. In *Proceedings of Coling'82*.

MILLER G. & CHARLES W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, **6**(1), 1–28.

MOORE R. C. (2004). Improving IBM word alignment model 1. In *42nd Meeting of the Association for Computational Linguistics (ACL)*, p. 518–525.

MULLER P. & LANGLAIS P. (2010). Comparaison de ressources lexicales pour l'extraction de synonymes. In *Article court au 17e TALN*, Montréal, Canada.

NIWA Y. & NITTA Y. (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of Coling 1994*.

PATRY A. & LANGLAIS P. (2011). PARADOCS : l'entremetteur de documents parallèles indépendant de la langue. *TAL*, **51-2**, pp. 41-63.

PIASECKI M., SZPAKOWICZ S., MARCIŃCZUK M. & BRODA B. (2008). Classification-based filtering of semantic relatedness in hypernymy extraction. In A. RANTA & B. NORDSTRÖM, Eds., *GoTAL 2008*, number 5221 in LNAI, p. 393–404 : Springer.

TURNEY P. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*.

VAN DER PLAS L. & TIEDEMANN J. (2006). Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, p. 866–873.

WEEDS J. E. (2003). *Measures and Applications of Lexical Distributional Similarity*. PhD thesis, University of Sussex.

WU H. & ZHOU M. (2003). Optimizing synonyms extraction with mono and bilingual resources. In *Proceedings of the Second International Workshop on Paraphrasing*.

ZHITOMIRSKY-GEFFET M. & DAGAN I. (2009). Bootstrapping distributional feature vector quality. *Computational Linguistics*, **35**(3), 435–461.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs

Yann Mathet[1] Antoine Widlöcher[1]
(1) GREYC, UMR CNRS 6072, Université de Caen, 14032 Caen Cedex
{prenom.nom}@unicaen.fr

Résumé. L'alignement et la mesure d'accord sur des textes multi-annotés sont des enjeux majeurs pour la constitution de corpus de référence. Nous défendons dans cet article l'idée que ces deux tâches sont par essence interdépendantes, la mesure d'accord nécessitant de s'appuyer sur des annotations alignées, tandis que les choix d'alignements ne peuvent se faire qu'à l'aune de la mesure qu'ils induisent. Nous proposons des principes formels relevant cette gageure, qui s'appuient notamment sur la notion de désordre du système constitué par l'ensemble des jeux d'annotations d'un texte. Nous posons que le meilleur alignement est celui qui minimise ce désordre, et que la valeur de désordre obtenue rend compte simultanément du taux d'accord. Cette approche, qualifiée d'holiste car prenant en compte l'intégralité du système pour opérer, est algorithmiquement lourde, mais nous sommes parvenus à produire une implémentation d'une version légèrement dégradée de cette dernière, et l'avons intégrée à la plate-forme d'annotation Glozz.

Abstract. Building reference corpora makes it necessary to align annotations and to measure agreement among annotators, in order to test the reliability of the annotated ressources. In this paper, we argue that alignment and agreement measure are interrelated : agreement measure applies to pre-aligned data and alignment assumes a prior agreement measure. We describe here a formal and computational framework which takes this interrelation into account, and relies on the notion of disorder of annotation sets available for a text. In this framework, the best alignment is the one which has the minimal disorder, and this disorder reflects an agreement measure of these data. This approach is said to be holistic insofar as alignment and measure depend on the system as a whole and cannot be locally determined. This holism introduces a computational cost which has been reduced by a heuristic strategy, implemented within the Glozz annotation platform.

Mots-clés : Alignement d'annotations, mesure d'accord inter-annotateurs, linguistique de corpus.

Keywords: Alignment, inter-coder agreement measure, corpus linguistics.

1 Contexte

La multiplication des travaux sur corpus, en linguistique computationnelle et en TAL conduit naturellement à la multiplication des campagnes d'annotation et rend nécessaire la mise en place de méthodes et d'outils permettant d'interpréter le fruit de ces campagnes. Pour établir des corpus annotés de référence, ou simplement pour mieux comprendre les phénomènes linguistiques que ces campagnes prennent pour objets, il est notamment nécessaire de mettre en correspondance (d'aligner) les annotations produites par différents annotateurs (humains ou automatiques), sur un même jeu de données, et de prendre la mesure de leurs accords et désaccords.

Dans cet article, nous nous intéressons aux questions d'alignement et d'accord inter-annotateurs, en nous limitant à des annotations de textes consistant, de façon très générale, à délimiter et à catégoriser des unités. Il est important de noter que la méthode que nous cherchons à définir doit permettre d'aligner et de comparer des objets textuels relativement variés, distribués dans le texte de manières elles aussi variées, et qu'à ce titre, nous devons nous écarter de nombreux travaux eux aussi consacrés à l'alignement et à la mesure d'accord (*cf.* section 2).

Nous cherchons à aligner et à comparer des *unités*, segments de texte commençant et s'achevant en des positions déterminées. Insistons sur le fait que la segmentation du texte, *i.e.* le positionnement des unités, n'est pas considérée comme acquise. En effet, dans certains cas, les annotateurs n'auront pas exclusivement à caractériser des données déjà délimitées, mais devront également déterminer leur position dans le texte et leur taille. Concernant ce

positionnement des unités, précisons de plus qu'il ne conduit pas nécessairement à un pavage complet du texte, la sporadicité des phénomènes étant même parfois assez grande. Concernant leur taille, ajoutons que celle-ci pourra varier fortement d'une unité à l'autre, et cela, éventuellement, pour un même type d'objet linguistique. Pour le positionnement relatif des unités, nous souhaitons de plus offrir une grande souplesse : les unités pourront se succéder, s'inclure, se chevaucher. Chaque unité possède par ailleurs une *catégorie* choisie parmi un ensemble prédéfini pour une campagne d'annotation donnée. Ajoutons que les couches d'annotations correspondant aux différentes catégories ne doivent pas être regardées comme indépendantes, l'attribution d'une mauvaise catégorie à un objet pouvant être parfois, dans une certaine mesure, acceptable.

Les raisons pour lesquelles nous devons privilégier la tolérance tiennent dans une large mesure à la nature des objets linguistiques sur lesquels nous travaillons par ailleurs. En effet, nous opérons souvent dans le champ disciplinaire de l'analyse du discours et explorons des structures textuelles variées, telles qu'envisagées par des approches aussi hétérogènes que l'*Argumentative Zoning* (Teufel *et al.*, 1999), l'*encadrement du discours* (Charolles, 1997) ou encore la SDRT (*Segmented Discourse Relation Theory*) (Asher, 1993). À titre d'exemple, précisons que ce travail prend place dans la continuité du projet ANR Annodis (Péry-Woodley *et al.*, 2009), qui vise la mise en place d'un corpus de référence pour le français, en analyse de discours. Comme on le verra, il entretient par ailleurs de nombreuses relations avec la plate-forme d'annotation et d'exploration de corpus Glozz (Widlöcher & Mathet, 2009), plate-forme permettant de produire des annotations hétérogènes exigeant cette tolérance.

La méthode que nous proposons ici n'est néanmoins pas dédiée à l'évaluation d'annotations discursives. Elle se veut aussi générique que possible et nous pouvons résumer ainsi son objectif : nous recherchons à la fois un alignement et une mesure d'accord multi-annotateurs portant sur des annotations composées d'unités marquées par leur possible variété de grain, leur possible variété catégorielle, leur possible sporadicité et la souplesse de leurs distributions relatives.

2 État de l'art

Parmi les travaux dans la continuité desquels notre étude prend position, nous pouvons distinguer ceux qui portent leur attention sur la question de l'attribution de catégories à des unités prédéfinies (la caractérisation) d'une part et ceux qui privilégient la question de la segmentation des unités d'autre part.

Pour les premiers, l'accord entre annotateurs concerne principalement l'affectation, par chacun, d'une catégorie choisie parmi un ensemble défini pour une campagne d'annotation donnée, à des unités dont la délimitation est considérée comme non problématique, souvent le mot. Dans cette perspective, de nombreux travaux se réfèrent notamment aux coefficients que sont π (Scott, 1955) et κ (Cohen, 1960) ainsi qu'à des variantes multi-annotateurs tel le K de (Siegel & Castellan, 1988) et aux coefficients pondérés α (Krippendorff, 1980) et κ_w (Cohen, 1968)[1]. Non spécifiquement issues du TAL ou de la linguistique, ces différentes approches de la mesure d'accord font l'objet de travaux qui visent à étudier leur pertinence et leurs limites dans ces domaines d'accueil et à en comparer les retombées. Nous pensons ici en particulier à l'excellente synthèse de (Artstein & Poesio, 2008), sur laquelle nous nous appuyons fortement ici. Si la présente étude apporte, comme on le verra, au problème de la caractérisation une réponse provisoire relativement légère, la fréquentation de ces travaux nous ouvre toutefois d'ores et déjà des perspectives essentielles, dont l'influence sera encore accrue dans nos travaux futurs. En particulier, l'importance que ces travaux accordent à la confrontation entre les accords observés et ceux que le seul hasard peut engendrer est tout à fait éclairante, de même que l'est leur réflexion sur l'obtention du « meilleur hasard possible », qui tiendra compte des propriétés particulières de la campagne engagée et de son corpus, ainsi, le cas échéant, que des tendances des annotateurs. Une autre avancée importante concerne la proposition de solutions, intimement liée à la réflexion sur les coefficients pondérés, permettant de rendre compte du fait important que tous les désaccords ne se valent pas. Ce point sera évidemment au cœur de la question de la segmentation, mais, dans la continuité de ces travaux, nous y accorderons aussi une large place en ce qui concerne la caractérisation des unités. Enfin, mentionnons la place méritée que ces travaux accordent à la délicate question de l'interprétation qualitative des résultats quantitatifs. Dans le prolongement de leur effort, nous serons aussi amenés à envisager des « grilles » permettant l'interprétation des mesures que nous proposons.

Naturellement, la principale limite de ces travaux, du point de vue qui nous occupe, est le fait que la segmentation y

[1] Comme le note (Artstein & Poesio, 2008), certains flottements de dénomination perturbent souvent les discussions relatives à ces coefficients. Nous retenons la clarification qu'ils proposent.

soit globalement considérée comme acquise. Il convient toutefois de remarquer, et (Artstein & Poesio, 2008) nous y invitent, que ces approches peuvent fournir un cadre pour l'estimation de l'accord sur des tâches de segmentation. Ainsi, (Teufel *et al.*, 1999) envisagent par exemple l'accord obtenu sur l'attribution de rôles argumentatifs à des phrases, en utilisant le coefficient de (Siegel & Castellan, 1988), phrases dont l'ajacence conduit à l'émergence de segments. Les mesures d'accord sur l'attribution de catégories peuvent encore être utilisées, comme c'est le cas dans (Hearst, 1997), non plus sur le contenu des segments, mais pour mesurer l'accord sur l'identification des bornes, c'est-à-dire sur l'attribution d'une catégorie *borne*. Toutefois, l'utilisation de ces approches pour des tâches de segmentation se heurte à la difficulté majeure suivante : une délimitation d'unité n'est regardée comme faisant consensus que si les annotateurs sont parfaitement d'accord sur le positionnement exact des bornes. Or en la matière, et en particulier aux échelles discursives, une plus grande souplesse est nécessaire, pour que de légers désaccords dans le positionnement des bornes soient moins lourdement pénalisés.

La méthode proposée par (Grouin *et al.*, 2011), adossée à la mesure d'erreur *slot error rate* (Makhoul *et al.*, 1999), permet de combiner alignement et mesure d'erreur et d'aborder simultanément positionnement des unités et attribution de catégories. Certes, elle permet d'aligner des unités dont les positions ne sont pas identiques, mais les différentes raisons suivantes la rendent peu adaptée à notre perspective : tous les écarts sont sanctionnés de manière identique ; elle est prévue pour comparer seulement deux annotations, dont l'une fait office de référence ; enfin, elle opère à l'échelle de la phrase et non du texte.

D'autres travaux visent à prendre spécifiquement en charge les problèmes de segmentation, en particulier dans le domaine de la segmentation thématique. Dans ce domaine, un consensus s'est établi autour de la mesure WindowDiff (Pevzner & Hearst, 2002), qui consiste à déplacer une fenêtre glissante le long du texte, et à comparer le nombre de ruptures présentes dans une annotation considérée comme référence et dans une annotation évaluée. Aux limites de cette méthode évoquées par exemple par (Lamprier *et al.*, 2007) et (Bestgen, 2009) (difficulté à interpréter les résultats, dépendance à l'égard de la taille de la fenêtre glissante, erreurs pénalisées différemment selon leur position dans le texte, erreurs légères parfois trop pénalisées...) s'ajoute dans notre perspective le fait que cette méthode ne fournit pas à proprement d'alignement, limite qui s'applique également aux aménagements de WindowDiff proposés par (Lamprier *et al.*, 2007). (Bestgen, 2009) préconise pour sa part le recours à la distance de Hamming généralisée (DHG) (Bookstein *et al.*, 2002), distance d'édition entre deux annotations, qui ajoute à la distance de Hamming l'opération de déplacement qui permet de donner la souplesse nécessaire à la prise en compte d'erreurs légères dans le positionnement des bornes. Offrant un résultat plus facile à interpréter que d'autres indices, cette méthode souffre selon nous de limites qui s'appliquent du reste également à WindowDiff. Pensées (dans le cas de WindowDiff) ou détournées (dans le cas de DHG) pour l'évaluation de la segmentation thématique, ces méthodes sont intimement liées d'une part à l'idée de pavage complet du texte (ce qui enfreint notre contrainte d'éventuelle sporadicité) et d'autre part à l'unicité du phénomène envisagé, *i.e.* le phénomène de rupture thématique (ce qui enfreint notre contrainte de prise en charge d'annotation multi-catégorielles). Ajoutons que ces méthodes n'intègrent pas de correction par le hasard. La solution α_U proposée par (Krippendorff, 1995), qui repose sur la mesure du recouvrement entre les annotations de différents annotateurs répond à beaucoup des exigences que nous avons fixées et nous devrons l'évaluer davantage. Elle impose toutefois, comme l'indiquent (Artstein & Poesio, 2008), que les annotations d'un même annotateur ne se recouvrent pas, ce qui contredit la contrainte de souplesse positionnelle que nous nous sommes fixée.

3 Quelle mesure d'accord ?

3.1 Difficulté majeure : interdépendance de la mesure d'accord et de l'alignement

Pour un « même phénomène » repéré par plusieurs annotateurs, il est nécessaire de prévoir une mesure d'accord suffisamment souple pour pouvoir rendre compte d'une double divergence, la première portant sur le choix de catégorie attribuée au phénomène, la seconde portant sur son positionnement. Il n'est pas rare, en particulier, que le positionnement diffère de façon substantielle sur l'une, l'autre, ou même les deux bornes. Du fait de ces divergences de positionnement, la mesure d'accord est assujettie à la détermination d'un alignement inter-annotateurs, un tel alignement consistant à déterminer quelle unité de tel annotateur correspond à telle autre de tel autre annotateur. Si l'on dispose d'un alignement complet des annotations de l'ensemble des annotateurs, il est possible, pour chaque unité repérée, de déterminer dans quelle mesure les annotateurs se sont entendus sur son positionnement et sur sa qualification. Cette quantification sera établie au moyen d'une mesure de « dissimilarité » entre

unités annotées : plus les unités seront considérées comme « proches », plus cette mesure devra être faible. Des propositions relatives à l'établissement de telles mesures seront faites ci-après.

Mais comment obtenir un tel alignement ? Aligner une unité u_a de l'annotateur A avec une unité u_b de l'annotateur B consiste à considérer que les propriétés (catégorie, position) de u_a et de u_b sont suffisamment « proches » pour pouvoir être assimilées : l'annotateur A et l'annotateur B ont rendu compte d'un « même phénomène », bien que de façon éventuellement (et légèrement) différente. La méthode d'alignement doit donc pour sa part s'appuyer sur une « distance » entre unités pour pouvoir opérer.

Dès lors, mesure d'accord et alignement sont inter-dépendants : on ne peut mesurer sans disposer d'un alignement, ni l'on ne peut aligner sans disposer d'une mesure, si bien que ces deux processus ne peuvent constituer deux étapes successives. Cette interdépendance renvoie simplement à l'unicité de l'objectif effectivement posé : établir dans quelle mesure des éléments éventuellement différents peuvent malgré tout être considérés comme semblables, soit pour quantifier ces différences (dans le cas de la mesure), soit pour assimiler des unités « similaires » (dans le cas de l'alignement). Il est donc nécessaire de disposer d'une méthode unifiée pour la mesure et l'alignement.

3.2 Le désaccord comme créateur de désordre

Considérons un ensemble de n annotateurs travaillant sur un même texte et une même tâche d'annotation. Idéalement, si la tâche d'annotation était rigoureusement établie, et si elle relevait de phénomènes ne prêtant pas à confusion, les n annotateurs devraient délivrer le même ensemble d'éléments annotés. C'est cependant bien entendu un constat que nous ne faisons jamais. Pour autant, les différentes propositions des annotateurs devraient en grande partie converger, à défaut de quoi la tâche proposée devrait être considérée comme un échec (tâche trop peu spécifiée, phénomènes étudiés ne permettant aucun consensus...). Ainsi, pour une campagne d'annotation donnée, on constatera un « taux d'accord » inter-annotateurs situé entre l'idéal constitué par une annotation unique (les n annotateurs ont annoté exactement le même ensemble d'unités) et le cas le pire constitué par n générateurs aléatoires d'annotations. L'enjeu de la mesure d'accord est de situer ce jeu d'annotations entre ces deux extrêmes.

Notre proposition est de considérer que l'annotation multiple est potentiellement génératrice de désordre. Le cas idéal (dans lequel tous les annotateurs ont délivré exactement le même jeu d'annotations) peut être considéré comme parfaitement ordonné : l'information portée par les annotations d'un annotateur donné est parfaitement confirmée par les annotations de chacun des autres annotateurs. Par rapport à cette situation idéale, opérons un ensemble de transformations élémentaires sur un certain nombre d'unités : déplacement de l'une des deux bornes d'une unité, requalification de sa catégorie, ou encore, suppression pure et simple. Chacune de ces transformations va engendrer un certain désordre au sein de ce système. Le désordre total obtenu pour un ensemble de transformations élémentaires sera la résultante de l'ensemble des désordres élémentaires ainsi créés. Nous définirons ci-après un cadre formel et une méthode de calcul de ce désordre, et poserons que le taux d'accord inter-annotateurs correspond au niveau d'ordre du système relativement au niveau d'ordre d'un système construit aléatoirement.

4 Dissimilarité, alignement, entropie et accord

4.1 Définitions : unité, annotateur, jeu d'annotations

Nous définissons tout d'abord $\mathcal{A}$ l'ensemble des annotateurs, $\mathcal{T}$ l'ensemble des textes et $\mathcal{U}$ l'ensemble des unités.

Unité : une unité u possède une catégorie notée $cat(u)$, et une position donnée par ses deux bornes, correspondant chacune à un indice de caractère du texte, notées respectivement $start(u)$ et $end(u)$, $start$ et end étant donc des fonctions de $\mathcal{U}$ vers $\mathbb{N}^+$. Nous définissons l'égalité entre deux unités comme suit :

$$\forall (u,v) \in \mathcal{U}^2, u = v \Leftrightarrow ((cat(u) = cat(v)) \wedge (start(u) = start(v)) \wedge (end(u) = end(v)))$$

Une unité est produite par un annotateur donné, et est relative à un texte donné (dans le cadre d'une campagne donnée). L'unité émanant de l'annotateur a et de rang i est notée u_a^i.

Jeu d'Annotations : un jeu d'annotations j est un ensemble d'unités relatives à un même texte et produites par un ensemble donné d'annotateurs. Un tel jeu est dit aléatoire quand ses annotateurs sont des processus aléatoires.

4.2 Dissimilarité entre deux unités

Une dissimilarité est une fonction $d : \mathcal{U}^2 \rightarrow \mathbb{R}^+$, telle que :

$$\forall (u,v) \in \mathcal{U}^2, \begin{cases} d(u,v) = d(v,u) \text{ (d est symétrique)} \\ d(u,v) = 0 \Leftrightarrow u = v \end{cases}$$

Une dissimilarité n'est pas nécessairement une distance au sens mathématique dans la mesure où l'inégalité triangulaire n'est pas imposée. Nous verrons pourquoi.

4.2.1 Dissimilarité positionnelle d_{pos}

Il est possible de proposer différentes mesures de dissimilarités positionnelles pour différents paradigmes d'annotation. Nous nous en tiendrons ici à la dissimilarité suivante, bien adaptée à des annotations sporadiques :

$$d_{pos-sporadique}(u,v) = \left(\frac{|start(u) - start(v)| + |end(u) - end(v)|}{\left(\frac{(end(u)-start(u))+(end(v)-start(v))}{2} \right)} \right)^2 \quad (1)$$

Cette dissimilarité rend compte des différences entre les bornes gauches des deux unités ainsi qu'entre leurs bornes droites. Sa croissance est quadratique par rapport à la somme de ces différences, si bien que l'on pénalise d'autant plus les écarts importants. Elle ne respecte pas l'inégalité triangulaire pour cette raison. Par ailleurs, le fait de diviser les différences par la moyenne des deux longueurs des unités (*cf.* dénominateur) permet de rendre la dissimilarité insensible aux changements d'échelle. C'est un choix qui peut être discuté selon la campagne d'annotation envisagée. A titre d'exemple, une seconde dissimilarité positionnelle est actuellement expérimentée pour émuler la distance de Hamming généralisée, basée sur la longueur moyenne des unités notée k :

$$d_{pos-Hamming}(u,v) = \frac{|end(u) - end(v)|}{k/2} \quad (2)$$

4.2.2 Dissimilarité catégorielle d_{cat}

Soit C l'ensemble des catégories. Pour une campagne d'annotation donnée, n catégories distinctes sont définies.

Nous définissons tout d'abord la distance catégorielle entre catégories $dist_{cat}$ au moyen d'une matrice carrée de taille n, prenant l'ensemble des catégories à la fois sur les lignes et sur les colonnes. Chaque case indique la distance entre deux catégories par une valeur située dans l'intervalle $[0, 1]$. La valeur 0 signifie l'égalité des catégories (du fait des propriétés d'une distance), tandis que la valeur 1, maximale, signifie que les deux catégories sont incompatibles (l'une ne peut en aucun cas se substituer à l'autre). Une telle matrice est nécessairement symétrique et possède une diagonale nulle, du fait, là encore, des propriétés d'une distance. Voici un exemple de matrice rendant compte d'un ensemble de 3 catégories. Elle permet une correspondance possible entre une unité de type cat_1 avec une unité de type cat_2, avec un coût de 0.5 (qui sera à mettre en balance avec les coûts issus des dissimilarités positionnelles), et elle interdit les autres correspondances :

	cat_1	cat_2	cat_3
cat_1	0	0.5	1
cat_2	0.5	0	1
cat_3	1	1	0

TAB. 1 – Exemple de matrice pour 3 catégories

On définit alors la dissimilarité catégorielle entre deux unités par :

$$d_{cat}(u,v) = dist_{cat}(cat(u), cat(v)).\Delta_{\emptyset} \tag{3}$$

$\Delta_{\emptyset}$ est une constante qui sera définie ultérieurement et qui assure ici notamment que deux unités de catégories distinctes ne seront jamais alignées.

4.2.3 Dissimilarités combinée d_{combi}

Soient deux dissimilarités d_1 et d_2 données. On définit $d_{combi(d_1,d_2,\alpha,\beta)}$ par :

$$d_{combi(d_1,d_2,\alpha,\beta)}(u,v) = \alpha.d_1(u,v) + \beta.d_2(u,v) \tag{4}$$

Cette combinaison linéaire de dissimilarités est elle-même une dissimilarité. Elle permet notamment, dans le cas où $\alpha = 0.5$ et $\beta = 0.5$, de donner un poids égal à deux dissimilarités (par ex. positionnelle et catégorielle).

4.3 Alignement

4.3.1 Alignement unitaire $\dot{a}$

Un alignement unitaire $\dot{a}$ est un i-uplet, i étant compris entre 1 et n, n étant le nombre d'annotateurs, contenant au plus une unité de chaque annotateur. Pour des raisons d'homogénéité facilitant notamment son implémentation informatique, nous créons une unité fictive vide, notée $u_{\emptyset}$, correspondant à la réification du fait qu'un alignement unitaire ne contienne aucune unité pour un annotateur donné. Nous ferons comme si cet alignement contenait cette unité fictive pour cet annotateur là, si bien que tout alignement unitaire devient finalement, dans tous les cas, un n-uplet, contenant au moins une unité non vide, et, pour chaque annotateur, soit l'une de ses unités, soit $u_{\emptyset}$. Pour n annotateurs numérotés de 1 à n, et ayant respectivement annoté $card_i$ unités, le nombre d'alignement unitaires qu'il est possible de générer est de $(\prod_{i=1}^{n} card_i) - 1$ (en retirant l'alignement ne contenant que des $u_{\emptyset}$).

4.3.2 Alignement $\bar{a}$

Pour un jeu d'annotations donné, un alignement $\bar{a}$ est défini comme un[2] ensemble d'alignements unitaires tel que chaque unité de chaque annotateur apparaît dans un et un seul de ses alignements unitaires.

4.4 Alignement et entropie[3]

4.4.1 Entropie d'un alignement unitaire

L'entropie d'un alignement unitaire $\dot{a}$, notée $\dot{e}(\dot{a})$, est définie pour une dissimilarité d_x donnée comme la valeur moyenne des dissimilarités deux à deux de ses unités constituantes :

$$\dot{e}(\dot{a}) = \frac{1}{C_n^2} \cdot \sum_{(u,v)\in\dot{a}^2} d_x(u,v) \tag{5}$$

Cependant, étant donné qu'un alignement unitaire peut comporter des unités fictives $u_{\emptyset}$, il est nécessaire de définir la dissimilarité entre une unité réelle et l'unité fictive $u_{\emptyset}$.

Pour toute dissimilarité d_x, pour toute unité u, $d_x(u,u_{\emptyset}) = d_x(u_{\emptyset},u) = \Delta_{\emptyset}$, constante qui est à définir pour une campagne donnée. En effet, cette valeur indique jusqu'à quel seuil de dissimilarité il convient de préférer aligner une unité avec une autre plutôt qu'avec $u_{\emptyset}$. Par exemple, si on choisit comme dissimilarité positionnelle

[2]Notons que, pour n annotateurs qui auraient chacun créé le même nombre p d'unités, le nombre d'alignements qu'il est possible de générer est supérieur à $(p!)^{n-1}$, ce qui dépasse très rapidement les capacités de stockage et de traitement des machines.

[3]Le terme entropie est ici quelque peu usurpé, pris seulement pour évoquer la notion de désordre.

$d_{pos-sporadique}$, et que l'on souhaite que deux unités de même longueur soient alignables tant qu'elles se touchent (et en faisant abstraction de la dissimilarité catégorielle éventuelle), on calcule la dissemblance d'une telle configuration (unité u positionnée de x à $x+l$ et v positionnée de $x+l$ à $x+2l$, soit $d_{pos-sporadique}(u,v) = ((l+l)/l)^2 = 2^2 = 4$) et on pose donc $\Delta_\emptyset = 4$. Par ailleurs, concernant la dissimilarité catégorielle, la formule (3) montre qu'une valeur de 1 dans la matrice des distances fait systématiquement préférer le choix de $u_\emptyset$ (valeur $\Delta_\emptyset$) à une unité de la catégorie concernée (valeur $\Delta^+_\emptyset$), même si la dissimilarité positionnelle est nulle. Bien sûr, lorsque l'on combine d_{pos} et d_{cat}, les écart positionnels et catégoriels s'ajoutant, on en vient d'autant plus rapidement à dépasser la valeur $\Delta_\emptyset$ et à préférer $u_\emptyset$ à une unité réelle.

Enfin, le choix de la valeur moyenne des dissimilarité plutôt que leur somme permet de s'abstraire du nombre d'annotateurs.

4.4.2 Entropie d'un alignement

L'entropie d'un alignement $\bar{a}$, notée $\bar{e}(\bar{a})$, est la valeur moyenne de l'entropie de ses alignements unitaires :

$$\bar{e}(\bar{a}) = \frac{1}{|\bar{a}|} \cdot \sum_{i=1}^{|\bar{a}|} \dot{e}(\dot{a}_i) \tag{6}$$

Nous faisons le choix de considérer la valeur moyenne des entropies unitaires plutôt que leur somme afin par exemple que l'entropie d'un jeu multiple d'annotations qui serait constitué de la duplication d'un jeu donné possède la même entropie que ce dernier et non pas son double.

4.5 Alignement idéal et mesure d'accord

Alignement idéal $\hat{a}$. Un alignement $\bar{a}$ d'un jeu d'annotation j est considéré comme idéal vis-à-vis d'une fonction de dissimilarité d_x donnée s'il minimise son entropie parmi tous les alignements possibles de j. Il est alors noté $\hat{a}$.

Entropie d'un jeu d'annotations $e(j)$. L'entropie d'un jeu d'annotations j, notée $e(j)$, pour une fonction de dissimilarité d_x donnée, est définie comme l'entropie de son ou de ses alignements idéaux $\bar{e}(\hat{a})$. Par prudence, nous sommes contraints de parler de « ses alignements idéaux » et non pas de son alignement idéal car, même si c'est peu probable, plusieurs alignements distincts peuvent minimiser l'entropie d'un jeu d'annotations.

Nous venons d'établir les deux définitions cruciales de notre approche, qui rendent compte en particulier de son caractère unifié. En effet, le choix de l'alignement idéal se fait sur la base de l'entropie, donc de la mesure d'accord (*cf.* ci-dessous) entre annotateurs, et, réciproquement et parallèlement, la mesure d'accord se fait sur la base de l'alignement idéal.

Corpus : un corpus c est un ensemble donné de textes et l'ensemble des jeux d'annotations relatifs à ces textes.

Entropie aléatoire $e_{aleatoire}$. Pour un corpus c donné, soit P l'ensemble des processus aléatoires d'annotation actuellement disponibles[4]. $\forall p \in P$, soit $eAvg(p)$ la moyenne des entropies obtenues sur un ensemble significatif de jeux d'annotations produits par p. L'entropie aléatoire de ce corpus, notée $e_{aleatoire}(c)$, est définie comme $min(\{eAvg(p)/p \in P\})$. C'est une valeur qui sera susceptible de s'améliorer (diminuer) au fil des avancées en termes de génération aléatoire astucieuse.

Mesure d'accord. La mesure d'accord inter-annotateurs d'un jeu d'annotations est alors donnée par :

$$\forall j \in c, accord(j) = \frac{e_{aleatoire}(c) - e(j)}{e_{aleatoire}(c)} \tag{7}$$

Si les annotateurs sont parfaitement d'accord, comme dans le cas idéal évoqué au début de cet article, l'entropie résultante est nulle, si bien que la mesure d'accord est égale à 1. Au contraire, si les annotateurs ne font pas mieux que le hasard, leur taux d'accord est nul, voire négatif.

[4]Nous proposons deux tels processus de génération aléatoire en section suivante.

La méthode proposée peut être qualifiée d'holiste dans la mesure où c'est la considération de l'ensemble des annotations qui permet de déterminer les alignements unitaires. Il est impossible de partir d'alignements unitaires « sûrs » pour constituer, de façon ascendante, l'alignement idéal complet.

5 Opérationalisation : vers une méthode d'implémentation de l'alignement et de la mesure d'accord holistes

Pour toute la suite de cet article, nous allons utiliser la dissimilarité positionnelle $d_{pos-sporarique}$, et une dissimilarité catégorielle définie par une matrice remplie de 1, à l'exception de la diagonale qui est, comme toujours, nulle. Son rôle est ici limité à l'interdiction des couplages entre unités de catégories distinctes, afin de limiter les phénomènes entrant en jeu dans le cadre de cet article. Enfin, nous combinons ces deux dissimilarités en les sommant via $d_{combi(d_{pos-sporadique},d_{cat},1,1)}$.

Les définitions que nous venons de poser ont valeur d'un point de vue théorique, mais leur implémentation informatique pose un important problème de complexité, en raison du caractère holiste de la méthode proposée. Un parcours de toutes les possibilités est en effet inenvisageable, l'espace de recherche étant minoré par $(p!)^{n-1}$. A titre d'illustration, mentionnons que pour 5 annotateurs ayant simplement annoté chacun 5 unités, cette minoration est de $(5!)^4 = 120^4$, soit plus de 207 millions.

Nous allons établir des principes permettant de réduire cet espace de recherche de façon à obtenir une méthode utilisable avec des jeux de données réels, tels que 4 annotateurs ayant chacun annoté une centaine d'unités.

5.1 Une réduction de l'espace de recherche

Parmi les innombrables possibilités d'alignements qu'offre l'espace de recherche, une immense majorité reposent sur des alignements unitaires improbables. Nous allons démontrer qu'il est possible d'éliminer un grand nombre d'entre eux sans écarter l'alignement idéal.

En effet, considérons l'alignement idéal $\hat{a}$, de cardinalité m. Soit $\dot{a}$ l'un quelconque de ses alignements unitaires. Par commodité, nous lui donnons l'indice 1 ($\dot{a} = \dot{a_1}$), les autres ayant donc les indices de 2 à m. Cet alignement unitaire $\dot{a}$ contient n unités (réelles ou $u_\emptyset$). Pour chacune de ces unités u_i (avec $1 \leq i \leq n$), créons l'alignement unitaire $\dot{a}_{m+i} = (u_i, u_\emptyset, ..., u_\emptyset)$ de cardinalité n. Il est possible de créer un alignement $\bar{a}$ constitué de l'ensemble des alignements unitaires de $\hat{a} \setminus \{\dot{a}\}$, auquel on ajoute les alignements unitaires $\dot{a}_{m+1}$ à $\dot{a}_{m+n}$ que l'on vient de créer[5]. Il est de cardinalité $m + n - 1$. On a, du fait que $\hat{a}$ minimise l'entropie :

$$\bar{e}(\hat{a}) \leq \bar{e}(\bar{a}) \Rightarrow \frac{1}{m}\sum_{i=1}^{m}\dot{e}(\dot{a}_i) \leq \frac{1}{m+n-1}\sum_{i=2}^{m+n}\dot{e}(\dot{a}_i) \Rightarrow \sum_{i=1}^{m}\dot{e}(\dot{a}_i) \leq \frac{m}{m+n-1}\sum_{i=2}^{m+n}\dot{e}(\dot{a}_i) \leq \sum_{i=2}^{m+n}\dot{e}(\dot{a}_i)$$
$$\Rightarrow \dot{e}(\dot{a_1}) \leq \sum_{i=m+1}^{m+n}\dot{e}(\dot{a}_i)$$

et comme $\forall i > m, \dot{a}_i = \Delta_\emptyset$, et que l'on a posé $\dot{a} = \dot{a_1}$,

$$\Rightarrow \dot{e}(\dot{a}) \leq n \cdot \Delta_\emptyset \tag{8}$$

On a donc majoré l'entropie unitaire de tout alignement unitaire candidat à l'alignement idéal. Ainsi, sur exemple réel, pour 3 annotateurs créant chacun 25 unités, on passe d'environ 19000 à 1000 alignements unitaires.

5.2 Un algorithme rapide pour l'obtention d'une solution approchée

Une fois l'ensemble des alignements unitaires (restreint aux seuls alignements susceptibles d'appartenir à la solution, *cf.* point précédent) généré, trions ce dernier, et obtenons ainsi la liste $L_{initial}$. L'algorithme 1 permet

[5] $\bar{a}$ est bien un alignement, puisque chacune des unités apparaît dans un et un seul alignement unitaire.

d'obtenir une solution approchée de la recherche d'un alignement idéal. Sa complexité observée est telle qu'il est utilisable en des temps raisonnables, comme le montre le tableau 2.

Algorithm 1 Algorithme rapide pour une solution approchée

Require: $L_{initial}, L, L^-$ des listes, $i \in \mathbb{N}$, $\dot{a}$ un alignement unitaire

1: $L \leftarrow L_{initial}$
2: $i \leftarrow 0$
3: **while** $i < size(L) - 1$ **do**
4: $\quad \dot{a} \leftarrow L[i]$
5: $\quad L^- \leftarrow L[i+1, (size(L) - 1)]$
6: $\quad$ Retirer de L^- tous les alignements contenant (au moins) l'une des unités de $\dot{a}$
7: $\quad i \leftarrow i + 1$
8: **end while**

	Nb. d'annotateurs	Nb. d'unités par annotateur	Espace de recherche (alignements unitaires après filtrage)	Temps d'exécution
Cas 1	3	25	1145	Instantané
Cas 2	3	100	4347	$<$ 1 sec.
Cas 3	4	100	38624	5 sec.
Cas 4	4	200	69994	16 sec.
Cas 5	5	25	96794	9 sec.

TAB. 2 – Temps d'exécution de l'algorithme 1

Les alignements résultant de cette méthode ont été évalués qualitativement, par observation graphique de sorties telles que sur la figure 1. Si l'on observe localement quelques croisements, on notera toutefois qu'ils sont rares et de faible amplitude. Nous considérons que le degré de précision obtenu sur le calcul d'entropie qui en résulte suffira aux expériences menées dans la présente étude.

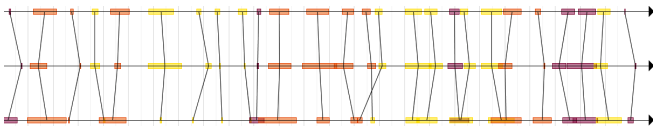


FIG. 1 – Exemple de résultat d'alignement avec l'algorithme 1 (un annotateur par ligne)

5.3 Vers une valeur minimale de $e_{aleatoire}$

En l'absence provisoire de l'alignement idéal, nous nous contenterons, pour la suite de cette étude, de l'approximation obtenue ci-dessus. Il reste à calculer la mesure d'accord, c'est-à-dire à situer l'entropie des annotations observées par rapport à ce que pourrait produire le hasard. Pour ce faire, nous chercherons à simuler le travail d'un observateur judicieux qui annoterait en aveugle un nouveau texte (en connaissant sa taille, mais sans pouvoir le lire), en s'inspirant des annotations faites par d'autres sur d'autres textes et en cherchant minimiser l'entropie.

Une première stratégie (*random1*) de génération aléatoire d'annotations a consisté, tout simplement, à observer des données réelles, en tenant compte du nombre (plages) d'annotations par annotateur, puis à générer des annotations conformes à ces observations, et de taille aléatoire. Il nous a toutefois semblé qu'un hasard bien entendu devrait prendre en compte des régularités plus fines observables sur un corpus donné, telles que (et de façon non limitative) : telle catégorie conduit à un pavage complet du texte, telle autre donne lieu à des tailles relativement stables, les textes commencent ou se terminent toujours par tel type particulier... La stratégie *random2* vise à prendre en compte de telles régularités. Elle consiste à assembler, au sein d'un même texte virtuel, des annotations émanant de textes réels différents, en respectant la règle suivante : pour générer une annotation virtuelle de n annotateurs, on utilise n textes différents dans chacun desquels on puise les annotations d'un et d'un seul annotateur[6]. Nous obtenons ainsi un jeu d'annotations qui respecte par construction la distribution des unités dans

[6]Une méthode permettant de pallier les différences de taille entre les différents textes a été mise en place.

les textes (régularité de position, de quantité...). Il est d'autre part effectivement aléatoire, puisque les annotations sont assemblées en aveugle : elles sont issues de textes sans rapport entre eux et ni avec le texte virtuel[7]. L'expérience rapportée ci-dessous confirme la supériorité de cette stratégie *random2* : nous obtenons en effet des tirages aléatoires ayant une entropie plus faible que celle provenant d'un tirage réalisé selon *random1* (3.48 au lieu de 3.67 pour un maximum possible de 4, soit 57 % de désordre en moins).

6 Observations expérimentales

6.1 Expérimentations sur un jeu de données factices

Afin de disposer d'une grille de lecture permettant de savoir à quoi correspondent les valeurs d'accord situées entre 0 et 1 (1 étant l'accord parfait, et 0 ce qu'est capable de faire le hasard), nous avons procédé à l'établissement empirique de deux courbes correspondant à deux modes de lecture parallèles. En premier lieu, nous cherchons à savoir comment fluctue le taux d'accord d'un jeu de données à partir d'un état parfait vers un état de plus en plus dégradé du seul point de vue du placement des unités. Pour ce faire, un algorithme crée aléatoirement les annotations d'un premier annotateur (25 unités réparties aléatoirement et de tailles aléatoires), puis crée des annotations pour deux autres annotateurs par dégradation du premier jeu d'annotation selon un facteur k, selon le principe suivant : chaque unité de l'annotateur 1 est dupliquée pour chacun des annotateurs 2 et 3, et modifiée aléatoirement par translation de chacune de ses deux bornes d'un vecteur compris entre $-\frac{k}{2}$ et $+\frac{k}{2}$ fois la longueur de l'unité dupliquée. En second lieu, nous nous intéressons aux fluctuations relatives à un jeu de données possédant de plus en plus de faux négatifs. Cette fois, l'algorithme crée les annotations des annotateurs 2 et 3 en dupliquant chacune des unités de l'annotateur 1 avec une probabilité p d'oubli (faux négatif). Avec $p = 0.5$, il y aura en moyenne une unité sur deux non présente dans les jeux 2 et 3 par rapport au jeu 1. Par contre, les entités conservées restent parfaitement alignées avec les originales. Les fluctuations obtenues sont reportées dans les graphes suivants.

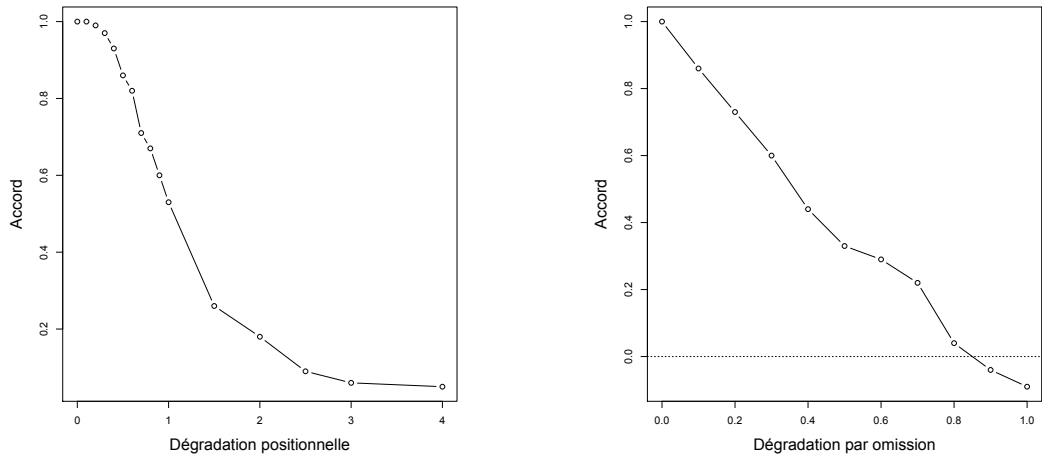


FIG. 2 – Évolution du taux d'accord selon la dégradation appliquée

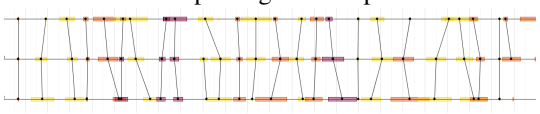
Il est ainsi possible de voir quel est l'équivalent d'une valeur d'accord donnée soit en termes de dégradation positionnelle, soit en termes d'oublis (faux négatifs). À titre d'illustration, la figure 3 donne deux exemples issus de ces deux paradigmes d'interprétation correspondant à la même valeur d'accord de 0.5.

6.2 Expérimentation sur un jeu de données réelles

Une première expérimentation sur des données réelles a été réalisée sur la base du corpus établi par (Labadié *et al.*, 2010), dont les annotations portent sur la segmentation thématique des textes, comportent plusieurs catégories, et comprennent des unités superposées au sein même des annotations d'un annotateur.

[7] Ajoutons que cette méthode permet de générer à moindre coût des données aléatoires en quantité importante. Pour a annotateurs et t textes ($t \geq a$), on peut en effet générer jusqu'à $C_t^a \cdot a^a$ combinaisons différentes.

Accord de 0.5 par dégradation positionnelle



Accord de 0.5 par dégradation par omissions

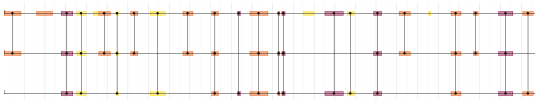


FIG. 3 – Illustrations d'un accord de 0.5

	Nb. annot.	e_{obs}	$e_{random1}$	a_1	$e_{random2}$	a_2
Texte 31	3	3.44	3.76	0.09	3.48	0.01
Texte 121	3	3.20	3.76	0.15	3.48	0.08
Texte 20	4	2.81	3.72	0.24	3.34	0.16
Texte 14	3	2.79	3.76	0.25	3.48	0.20
Texte 1	3	2.67	3.76	0.29	3.48	0.23
Texte 13	3	2.37	3.76	0.36	3.48	0.32
Texte 6	3	2.18	3.76	0.42	3.48	0.37

TAB. 3 – Résultats obtenus sur données réelles

Le tableau 3 rapporte les résultats obtenus sur 7 de ces textes (en utilisant la même dissimilarité que précédemment), en donnant pour chacun son entropie (e_{obs}), celles obtenues par $random1$ ($e_{random1}$) et par $random2$ ($e_{random2}$) pour le corpus correspondant, et les mesures d'accord respectives qui en résultent.

On constate que $random2$ est sensiblement plus efficace que $random1$, faisant passer le meilleur accord de 0.42 à 0.37, et que la baseline ainsi obtenue est relativement sévère dans la mesure où dans le cas le pire, l'accord obtenu n'est que de 0.01, soit d'un niveau équivalent à $random2$. Une observation visuelle, donc qualitative, des 7 textes annotés, confirme l'ordre obtenu par le calcul. Mentionnons que le texte 6 donnant lieu au meilleur accord $a_2 = 0.37$ correspond respectivement à une valeur $k = 1.2$ (translation des bornes jusqu'à ± 0.6 fois la largeur des unités associées, mais aucune omission) et $p = 0.46$ (omission d'une unité dans 46% des cas, mais positionnements parfaits) de nos deux grilles de lecture.

7 Conclusions et perspectives

Nous avons proposé une méthode réalisant alignement et mesure d'accord dans un processus commun. Très générale, cette méthode autorise les variations positionnelles et catégorielles, et ne dépend pas de la taille des entités annotées. Elle peut être configurée pour chaque paradigme d'annotation, par le choix des formules de calcul de la dissimilarité positionnelle et par l'indication des similarités entre catégories. Nous la disons holiste dans la mesure où elle s'appuie sur l'intégralité des données pour faire des choix, là où des méthodes fenêtrées comme WindowDiff opèrent localement. Par ailleurs, elle n'exige pas de jeu d'annotations de référence pour fonctionner ou s'amorcer. Nous travaillons actuellement à la mise en place de méthodes permettant de calculer automatiquement les valeurs optimales de $\Delta_\emptyset$ et des coefficients des matrices catégorielles (pour ce second point, voir aussi (Fort *et al.*, 2010)). Un algorithme permettant d'approcher la solution optimale a été défini et nous avons procédé à une implémentation complète du système, avec rendu graphique des alignements effectués. Cette implémentation sera distribuée avec la prochaine version de la plate-forme Glozz (1.1.0), devenant ainsi publiquement accessible, et directement exploitable par les utilisateurs de cette plate-forme. Un travail est actuellement mené, en partenariat avec Jean-Philippe Métivier (GREYC), qui vise à calculer dans un temps raisonnable la solution optimale. À cette fin, un système à base de CSP (*Constraint Satisfaction Problem*) prend en entrée l'ensemble des alignements unitaires candidats à la solution. Nous pourrons ainsi d'une part quantifier la différence d'entropie entre la solution approchée et la solution idéale, et d'autre part proposer, mais dans un temps éventuellement beaucoup plus long, la solution idéale.

Remerciements

Nous tenons à remercier Jérôme Chauveau qui a récemment rejoint l'équipe de développement de Glozz et contribué aux développements nécessaires aux expérimentations dont il est fait mention ici.

Références

ARTSTEIN R. & POESIO M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–596.

ASHER N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer.

BESTGEN Y. (2009). Quel indice pour mesurer l'efficacité en segmentation de textes ? In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

BOOKSTEIN A., KULYUKIN V. A. & RAITA T. (2002). Generalized Hamming distance. *Information Retrieval*, (5), 353–375.

CHAROLLES M. (1997). L'encadrement du discours : Univers, champs, domaines et espaces. *Cahier de Recherche Linguistique*, (6).

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

COHEN J. (1968). Weighted kappa : Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**(4), 213–220.

FORT K., FRANÇOIS C. & GHRIBI M. (2010). Evaluer des annotations manuelles dispersées : les coefficients sont-ils suffisants pour estimer l'accord inter-annotateurs ? In *Traitement Automatique des Langues Naturelles (TALN) Traitement Automatique des Langues Naturelles (TALN)*, p.0̃, Montréal France. Quaero.

GROUIN C., GALIBERT O., ROSSET S., QUINTARD L. & ZWEIGENBAUM P. (2011). Mesures d'évaluation pour entités nommées structurées. In *Évaluation des méthodes d'Extraction de Connaissances dans les Données*, Brest, France.

HEARST M. (1997). Texttiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, **23**(1), 33–64.

KRIPPENDORFF K. (1980). *Content Analysis : An Introduction to Its Methodology*, chapter 12. Sage : Beverly Hills, CA.

KRIPPENDORFF K. (1995). On the reliability of unitizing contiguous data. *Sociological Methodology*, (25), 47–76.

LABADIÉ A., ENJALBERT P., MATHET Y. & WIDLÖCHER A. (2010). Discourse structure annotation : Creating reference corpora. In *Workshop on Language Resource and Language Technology Standards - state of the art, emerging needs, and future developments*, La Valetta, Malta : Conference LREC 2010.

LAMPRIER S., AMGHAR T., LEVRAT B. & SAUBION F. (2007). On evaluation methodologies for text segmentation algorithms. In , Ed., *Proceedings of ICTAI 2007*, p. 19–26 : .

MAKHOUL J., KUBALA F., SCHWARTZ R. & WEISCHEDEL R. (1999). Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop*, p. 249–252.

PÉRY-WOODLEY M.-P., ASHER N., ENJALBERT P., BENAMARA F., BRAS M., FABRE C., FERRARI S., HO-DAC L.-M., LE DRAOULEC A., MATHET Y., MULLER P., PRÉVOT L., REBEYROLLE J., TANGUY L., VERGEZ-COURET M., VIEU L. & WIDLÖCHER A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. In *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN'09), session posters*, Senlis, France.

PEVZNER L. & HEARST M. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, **28**(1), 19–36.

SCOTT W. (1955). Reliability of content analysis : The case of nominal scale coding. *Public Opinion Quarterly*, **19**(3), 321–325.

SIEGEL S. & CASTELLAN N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, 2nd edition.

TEUFEL S., CARLETTA J. & MOENS M. (1999). An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of Ninth Conference of the EACL*, p. 110–117, Bergen.

WIDLÖCHER A. & MATHET Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis, France : ATALA LIPN.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

French TimeBank : un corpus de référence sur la temporalité en français

André Bittar[1] Pascal Amsili[2] Pascal Denis[3]
(1) Xerox Research Centre Europe
(2) LLF, Université Paris Diderot, UMR CNRS 7110
(3) EPI Alpage, INRIA Rocquencourt et Université Paris Diderot
andre.bittar@xrce.xerox.com,
pascal.amsili@linguist.jussieu.fr,
pascal.denis@inria.fr

Résumé. Cet article a un double objectif : d'une part, il s'agit de présenter à la communauté un corpus récemment rendu public, le French Time Bank (FTiB), qui consiste en une collection de textes journalistiques annotés pour les temps et les événements selon la norme ISO-TimeML ; d'autre part, nous souhaitons livrer les résultats et réflexions méthodologiques que nous avons pu tirer de la réalisation de ce corpus de référence, avec l'idée que notre expérience pourra s'avérer profitable au-delà de la communauté intéressée par le traitement de la temporalité.

Abstract. This article has two objectives. Firstly, it presents the French TimeBank (FTiB) corpus, which has recently been made public. The corpus consists of a collection of news texts annotated for times and events according to the ISO-TimeML standard. Secondly, we wish to present the results and methodological conclusions that we have drawn from the creation of this reference corpus, with the hope that our experience may also prove useful to others outside the community of those interested in temporal processing.

Mots-clés : Annotation temporelle, corpus, ISO-TimeML.

Keywords: Temporal annotation, corpus, ISO-TimeML.

1 Introduction

Le repérage des entités temporelles comme les événements et les dates, ainsi que le calcul des relations entre ces entités (précédence, inclusion...), est un aspect important de la compréhension des textes en langue naturelle. Plus spécifiquement, la détermination automatique de ces entités et de leurs relations est clairement susceptible d'apporter un plus aussi bien au niveau de diverses tâches du TAL (résumé automatique, résolution des anaphores...) qu'au niveau d'applications générales (extraction d'information, systèmes de question-réponse...). Durant les dernières années, de nombreux progrès ont été enregistrés dans le traitement automatique de ces phénomènes, mais la plupart de ces progrès concernent l'anglais. Ces améliorations ont été en large part dues au développement de la norme ISO-TimeML (ISO, 2008) et à la mise à disposition des corpus TimeBank (Pustejovsky *et al.*, 2003, 2006). Il s'agit de corpus de référence annotés pour les événements, les expressions temporelles et leur relations. Dans cet article, nous présentons le French TimeBank (FTiB) (Bittar, 2010a), qui comme son nom l'indique, est un corpus annoté du français et se base également sur la norme ISO-TimeML. Au-delà de la ressource elle-même, que nous présentons brièvement, nous mentionnons également les points principaux de notre méthodologie, qui, nous semble-t-il, sont partiellement transférables à d'autres tâches d'annotation. En particulier, nous avons tenté de mesurer de manière systématique l'impact d'une phase de pré-annotation automatique sur la qualité finale du corpus et sur le temps d'annotation.

L'article est organisé de la manière suivante. Dans une première section, nous présentons la norme ISO-TimeML, non sans lui apporter un certain nombre de modifications, certaines liées à l'adaptation au français, mais d'autres ayant une portée plus générale (section 3). Est ensuite décrite, en section 4, la méthodologie mise en œuvre : celle-ci se fonde sur une phase de pré-annotation automatique, suivie par une phase de correction manuelle. La section 5 est consacrée à la description des caractéristiques quantitatives et qualitatives du corpus produit, avant de revenir en conclusion sur les leçons à tirer de notre expérience, et les perspectives ouvertes par notre travail.

2 ISO-TimeML

ISO-TimeML est un langage d'annotation des informations temporelles pour les textes en langue naturelle. Il permet de baliser, avec un point de vue surfacique, les événements et les expressions temporelles (ou « marquables »), ainsi que les différentes relations qui existent entre ceux-ci. Le langage comporte six balises : deux pour les marquables, trois pour les relations, et enfin une pour les marqueurs (ou "signaux") de relations. Celles-ci sont brièvement décrites ci-dessous[1] et illustrées par l'exemple suivant :

```
Jean est <EVENT id="e1" class="OCCURRENCE" pos="VERB" tense="PAST" vform="PASTPART">né
</EVENT> <SIGNAL id="s1">avant</SIGNAL> l'<EVENT id="e2" class="OCCURRENCE"
pos="NOUN">introduction</EVENT> de l'euro.
<TLINK id="l1" eventID="e1" relatedToEvent="e2" signalID="s1" relType="BEFORE"/>
```

Le premier type de marquable utilisé en ISO-TimeML est **<EVENT>**. La notion d'événement correspond ici à la notion élargie d'*éventualité* de Bach (1986) et recouvre tous les types de situations (états, activités, achèvements, *etc*.). Cette balise comporte un ensemble d'attributs pour les traits morpho-syntaxiques et sémantiques (classe sémantique, temps, aspect, mode, modalité, polarité, *etc.*) de l'événement annoté. Les événements annotés peuvent correspondre à des catégories syntaxiques variées (en particulier nom, verbe, adjectif), et le choix fait dans ISO-TimeML est de placer la balise sur la tête du groupe (ou du chunk) événementiel, en excluant les auxiliaires, les modifieurs, les adverbes de négation, les clitiques, *etc.* Le second type de marquable est **<TIMEX3>** et correspond aux expressions temporelles dans le texte. Cette balise comporte des attributs pour le type de l'expression (date, heure, durée ou fréquence) et sa "valeur" normalisée[2]. Tout l'empan est marqué.

Les trois types de relations annotés en ISO-TimeML sont les **<ALINK>**, **<SLINK>**, et **<TLINK>**. Les ALINKS indiquent une relation aspectuelle entre deux événements. Par exemple, cette relation intervient entre un verbe aspectuel (*commencer*, *cesser*, *continuer*...) et son complément événementiel. Les SLINKS servent à marquer les relations de subordination (modale) entre deux événements. Typiquement, on l'utilisera pour marquer la relation qui existe entre un verbe modal (*falloir*, *devoir*...) ou de perception (*voir*, *entendre*...) et son complément événementiel. Enfin, les TLINKS marquent les relations (strictement) temporelles entre marquables. Comme pour les deux autres balises de relation, il existe différents sous-types, indiqués au moyen de l'attribut `relType`. L'attribut `signalID` permet de spécifier l'identifiant du marqueur qui réalise la relation dans le texte, s'il y en a un. Dans ce cas, c'est la balise **<SIGNAL>** qui est utilisée. Elle étiquette les marqueurs de relation lexicalisés dans les textes, comme, typiquement, les prépositions temporelles (*avant, après* et *pendant*).

Soulignons à nouveau le caractère résolument surfacique de cette norme : l'idée n'est pas d'annoter le sens en soi, mais de fournir une normalisation des formes linguistiques qui expriment la temporalité dans les textes, en limitant autant que possible l'engagement théorique. Mais cette position de principe (qui conduit par exemple à éviter de désambiguïser l'annotation) n'est pas toujours suivie à la lettre, et certaines annotations reposent parfois sur des informations qui ne relèvent pas uniquement des formes de surface. Par exemple, l'annotation de la subordination modale d'un événement (ex. *Jean croit que Léa est allée au Japon*) nécessite une connaissance de la structure syntaxique de la construction en question (ISO, 2008, p. 12). On notera aussi que, par définition, l'annotation des relations n'est pas strictement surfacique, puisque les relations sont le plus souvent implicites dans les textes, et ne correspondent à un élément visible que dans les cas où un marquable de type SIGNAL est présent.

3 Modifications d'ISO-TimeML

La norme ISO-TimeML est une norme récente, élaborée d'abord pour l'anglais, puis adaptée à d'autres langues, telles que l'italien (Caselli, 2008), le chinois et le coréen (ISO, 2008). Il n'est donc pas étonnant que cette norme soit encore sujette à des adaptations et changements éventuels. Nous proposons ici deux types de modifications : certaines indépendantes de la langue (§ 3.1), d'autres plus spécifiques au français (§ 3.2). Les modification que nous proposons concernent deux des balises ISO-TimeML : la balise `<EVENT>`, avec ses attributs de classe, temps, aspect, mode, modalité, polarité, *etc.*, et la balise `<ALINK>` qui sert à réaliser les relations aspectuelles entre éventualités. Certaines de ces modifications sont en cours d'adoption dans la norme ISO-TimeML. On trouvera

[1] Pour plus de détails sur la norme dans son état actuel (et sur son historique), voir p.ex. (ISO, 2008).
[2] La norme adoptée est une extention de la norme ISO 8601 pour la représentation internationale des dates et heures.

dans (Bittar, 2010b) l'ensemble des consignes d'annotation qui ont été élaborées.

3.1 Modifications indépendantes de la langue

Les modifications proposées correspondent à l'annotation d'expressions qui n'étaient jusqu'alors pas présentes dans le schéma ISO-TimeML ; nous pensons qu'elles sont utiles à annoter, en restant dans l'esprit surfacique du projet ISO-TimeML : même si elles ne dénotent pas en elles-mêmes des événements ou des temps, ces expressions contiennent des informations exploitables pour raffiner les informations aspectuelles ou de localisation temporelle des autres marquables du texte.

Conteneurs événementiels La notion de *conteneur (événementiel)*, introduite par Vendler (1967), désigne les prédicats (verbes ou adjectifs) qui sélectionnent un événement comme argument : p.ex. *se passer, se produire, avoir lieu*. Ils servent à établir l'existence d'un événement (1-a) dans le temps, mais aussi à relier un événement à des éléments de modalité et de polarité, ainsi qu'à des adverbiaux temporels (1-b). Afin de permettre la prise en compte de ces cas de portée, nous proposons d'ajouter une nouvelle classe d'événement au schéma ISO-TimeML afin de traiter ces contextes : la classe `EVENT_CONTAINER`.

(1) a. *La cérémonie a eu lieu.*
b. *La cérémonie ne devrait pas se tenir aujourd'hui.*

Constructions à verbe support On parle de verbe support (ou *light verb*) dans le cas de verbes ayant une contribution sémantique faible mais participant à une prédication complexe avec un nom (ou un autre verbe, un adjectif, etc). On s'intéresse aux constructions qui font intervenir un nom dénotant une éventualité :

(2) a. *Ce politicien a mené une attaque contre le libéralisme.*
b. *Ce politicien a lancé une attaque contre le libéralisme.*

Dans (2-a), le verbe *mener* a une lecture aspectuelle "neutre" (sans aucune précision aspectuelle sur le procès dénoté par le nom), alors que dans (2-b), le verbe *lancer* a une valeur aspectuelle inchoative : il exprime la phase initiale dans le déroulement de l'événement introduit par le nom, ce début étant temporellement localisé par le temps du verbe support. Assimilés dans la norme actuelle, ces deux cas de figure sont désormais distingués : d'une part, en conservant l'annotation standard (relation entre le verbe et le nom notée avec un `<TLINK>` de type `IDENTITY`), et d'autre part, en utilisant la balise `<ALINK>` dans le second cas pour marquer la relation entre le verbe aspectuel et son complément nominal.

Périphrases aspectuelles Beaucoup de langues réalisent des valeurs aspectuelles par le biais de périphrases, telles que *en train de* + V_{inf}, *en cours de* + N et *en voie de* + V_{inf} pour le français. Ces constructions sont elles aussi ignorées dans la norme actuelle ISO-TimeML. Nous traitons ces constructions en annotant une valeur aspectuelle sur la balise `<EVENT>` du complément événementiel ; les valeurs possibles sont celles déjà établies dans ISO-TimeML plus celles proposées dans la Section 3.2. Les valeurs de temps, d'aspect, de modalité et de polarité de la copule y sont également annotées.

Modalité ISO-TimeML permet de représenter la modalité attribuée à un événement dans une relation où un événement est subordonné par un verbe modal, par exemple. Le traitement actuel consiste à annoter la modalité sur l'événement subordonné par un attribut `modality` dans la balise `<EVENT>`, dont la valeur est de type XML CDATA (sans restriction donc). Ceci est vraisemblablement justifié pour l'anglais, où la modalité est exprimée essentiellement par des auxiliaires, non annotés en eux-mêmes. Mais dans une langue comme le français, la modalité est exprimée préférentiellement par des verbes pleins, qui peuvent être tensés, avoir des valeurs aspectuelles, et s'enchâsser les uns derrière les autres, et il nous semble préférable d'avoir une annotation spécifique pour ces verbes modaux. Suivant les catégories modales classiques (Palmer, 1986), nous proposons de limiter le jeu de valeurs à : `NECESSITY`, `POSSIBILITY` (pour le type épistémique), `OBLIGATION` et `PERMISSION` (pour le type déontique).

Cette dernière proposition doit être un peu discutée car elle conduit à s'éloigner un peu du point de vue surfacique qui prévaut en général dans la norme ISO-TimeML. En effet, au lieu de prévoir de marquer la modalité avec le lemme du verbe (ou plus généralement du prédicat) qui est responsable d'une subordination modale, ce qui nous

conduirait à conserver l'ambiguïté potentielle d'un verbe comme *devoir* (qui marque selon les cas la nécessité épistémique ou l'obligation déontique), nous proposons de lever l'ambiguïté pour indiquer au niveau de l'annotation la valeur modale elle-même. Notre choix est motivé par les raisons suivantes : d'une part, le lemme du verbe modal, puisqu'il est annoté, est facilement récupérable ; d'autre part, il a semblé, dans les premiers essais d'annotation, que la distinction entre les modalités ne posaient pas de problème important aux annotateurs, et il a donc semblé pertinent de profiter de l'occasion d'enrichir l'annotation.

3.2 Adaptations pour le français

Temps (verbaux) et aspect Le système français du temps et de l'aspect est différent de celui de l'anglais, et nous avons par conséquent proposé un jeu de valeurs appropriées pour l'annotation des temps verbaux (ainsi que les constructions en *en train de* + V_{inf}). Il s'agit d'une correspondance entre les valeurs ISO-TimeML et les valeurs aspectuo-temporelles. L'objectif n'est pas de fournir toutes les valeurs possibles des interprétations des temps verbaux, mais de fournir un ensemble de valeurs normalisées pour le français[3]. Voir le Tableau 1.

Groupe verbal	`tense`	`aspect`
mange	PRESENT	NONE
est en train de manger	PRESENT	PROGRESSIVE
a mangé	PAST	NONE
mangea	PAST	NONE
mangeait	IMPERFECT	NONE
était en train de manger	PAST	PROGRESSIVE
avait mangé	PAST	PERFECTIVE
avait été en train de manger	PAST	PERFECTIVE_PROGRESSIVE
mangera	FUTURE	NONE
sera en train de manger	FUTURE	PROGRESSIVE
aura mangé	FUTURE	PERFECTIVE
va manger	PRESENT	PROSPECTIVE
allait manger	IMPERFECT	PROSPECTIVE

TAB. 1 – Valeurs pour les temps verbaux et l'aspect en français.

Mode verbal Le mode verbal subjonctif est plus fréquemment employé en français qu'en anglais, et nous proposons d'ajouter la valeur `SUBJUNCTIVE` pour l'attribut `mood`, qui n'était pas prévu dans ISO-TimeML. De façon plus générale, il serait sans doute pertinent de spécifier systématiquement le mode (indicatif, subjonctif, conditionnel...) dans les annotations.

Verbes modaux Les verbes modaux en français (ex. *falloir, devoir, se pouvoir, etc.*) ne se comportent pas de la même façon que les auxiliaires modaux de l'anglais. Il s'agit plutôt de verbes lexicaux qui peuvent être conjugués à tous les temps, peuvent tomber sous la portée d'opérateurs de polarité et s'enchâsser les uns derrière les autres. Il est donc nécessaire de les annoter avec la balise `<EVENT>`, contrairement aux modaux de l'anglais. Nous proposons d'annoter les modaux du français avec la classe `MODAL`.

4 Méthodologie d'annotation

Nous présentons ci-dessous les points principaux de notre méthodologie d'annotation.

4.1 Échantillonnage de textes

Les textes source pour FTiB ont été sélectionnés à partir du corpus de *L'Est Républicain* du CNRTL. Le choix du domaine journalistique se justifie principalement par le nombre généralement important d'événements et d'ex-

[3]Nous avons fait le choix de conserver autant que possible le jeu de traits aspectuo-temporels de TimeML, ce qui conduit à certaines difficultés connues depuis longtemps (Kamp & Rohrer, 1983, p.ex.). Par exemple, pour le passé composé, nous avons privilégié l'interprétation la plus fréquente, analogue à un prétérit anglais, sur l'interprétation du type *present-perfect*. Une alternative qui mériterait sans doute d'être étudiée pourrait consister à choisir d'annoter avec les temps verbaux du français, en gardant toutes les ambiguïtés.

pressions temporelles dans ce type de textes. La distribution des textes choisis en fonction de leur sous-genre est résumée dans le Tableau 2. On notera que certains sous-genres sont plus fréquents que d'autres ; ce choix est motivé par deux raisons. D'abord pour favoriser une comparaison avec le TimeBank 1.2 de l'anglais, et deuxièmement, parce que ces sous-genres représentent une certaine diversité de style (p.ex., actualité politique) par rapport aux autres sous-genres, qui suivent plutôt un format particulier (p.ex., les nécrologies). Tous les textes du corpus contiennent des événements et des expressions temporelles. Nous reviendrons de manière plus détaillée sur les corrélations entre sous-genres textuels et contenu linguistique dans la section 5.

Sous-genre	**# documents**	**% documents**	**# tokens**	**% tokens**
Annonce	22	20.2%	1 679	10.4%
Bio	1	0.9%	186	1.1%
Actu. inter.	32	29.4%	5 171	31.9%
Actu. loc.	19	17.5%	4 370	27.0%
Actu. nat.	25	22.9%	3 347	20.7%
Nécrologie	2	1.8%	313	1.9%
Actu. sport	8	7.3%	1 142	7.0%
Total	109	100%	16 208	100%

TAB. 2 – Proportions de sous-genres de textes dans le French TimeBank.

4.2 Pré-annotation des marquables

Afin d'accélérer le processus d'annotation, nous avons opté pour une pré-annotation des marquables dans les textes, suivie d'une correction manuelle. L'annotation des relations a, quant à elle, été effectuée entièrement à la main. Le système d'annotation des marquables consiste en deux modules : le module *TempEx Tagger* et le module *Event Tagger*, que nous décrivons ci-dessous.

Le module ***TempEx Tagger*** balise les expressions temporelles (balise `<TIMEX3>`) et fixe ses attributs. Il repère également certains marqueurs de relation (balise `<SIGNAL>`), par exemple ceux qui apparaissent devant une expression temporelle. La technique choisie repose sur l'application de transducteurs Unitex (Paumier, 2008), qui s'appliquent directement sur du texte brut. Une des raisons de ce choix est que nous avons pu partir d'une batterie de transducteurs existants (Gross, 2002), que nous avons enrichie et adaptée. Les expressions sont classées selon leur type ISO-TimeML[4], et les valeurs de certains attributs sont calculées. La valeur de l'attribut `VALUE` n'est attribuée que dans un second temps, par un script qui calcule la valeur normalisée des expressions temporelles, y compris quand elles sont déictiques, comme *lundi dernier* ou *l'année prochaine* (la date de parution de l'article servant alors de point de repère).

Nous avons procédé à une évaluation comparative de ce module avec celui de (Parent *et al.*, 2008) appelé DEDO, et nous observons des performances très similaires sur un même corpus d'évaluation. La mesure de précision et de rappel pour la correspondance (*match*) correspond au balisage des mêmes empans textuels ; la mesure pour les valeurs correspond au calcul des valeurs d'attributs. Voir la table 3.

	Système	Précision	Rappel	F-score
Match	TempEx	**84.2**	**81.8**	**83.0**
	DEDO	83.0	79.0	81.0
Valeur	TempEx	55.0	44.9	49.4
	DEDO	**56.0**	**45.0**	**50.0**

TAB. 3 – Évaluation comparative du TempEx Tagger.

Le module ***Event Tagger*** s'occupe quant à lui des événements et des éventuels marqueurs qui réalisent une relation temporelle[5]. Il s'agit d'une suite d'applications de règles qui agissent sur les chunks et qui visent à éliminer ou à choisir les bons candidats pour l'annotation, sur la base de listes lexicales détaillées, et de divers

[4] `DATE` (*15/01/2001, le 15 janvier 1010, jeudi, demain*), `TIME` (*15h30, midi*), `DURATION` (ex. *trois jours, un an*) ou `SET` (ex. *tous les jours, chaque mardi*).

[5] Les prépositions comme *avant, après, etc.* qui introduisent un chunk événementiel.

critères contextuels. Ces règles supposent un texte déjà annoté en partie du discours, lemmatisé, et chunké. Nous avons choisi d'utiliser pour ce pré-traitement la chaîne de traitement Macaon (Nasr *et al.*, 2010).

Ce module repose sur certaines ressources lexicales, notamment un lexique de noms événementiels à large couverture. Le lexique est basé sur VerbAction (Hathout *et al.*, 2002) qui contient 9 393 paires (verbe, nom déverbal). Nous avons enrichi ce lexique par extraction de noms qui ne sont pas dans VerbAction. Ceci a été fait par recherche dans des moteurs de recherche de certains patrons, comme "*un * a eu lieu*", "*lors de la **" et "*le * se produit*" où * est susceptible d'être un nom d'événement. Cette méthode a rajouté 804 entrées au lexique des noms, notamment des noms d'événements non déverbaux, comme *anniversaire, apocalypse* et *grève* ainsi que des déverbaux n'apparaissant pas dans VerbAction.

Il n'a malheureusement pas été possible de comparer de façon fiable les performances du module Event Tagger avec le module similaire de (Parent *et al.*, 2008), le seul autre système développé pour cette tâche sur le français à notre connaissance. En effet, les évaluations ont été effectuées sur des corpus différents, quoique similaires, ce qui fait que les résultats ne sont qu'indicatifs. Pour le repérage des événements (les empans textuels des expressions), notre système a enregistré une précision de 62,2 (62,5 pour DEDO), un rappel de 89,4 (77,7), pour un F-score de 75,8 (69,3).

De conception assez simple, ces modules fournissent des résultats encore médiocres, mais suffisants pour permettre de considérablement accélérer les cycles d'annotation manuelle et ainsi de réduire le "coût" total de l'annotation.

4.3 Étapes d'annotation manuelles et validation

Après la pré-annotation des marquables, les textes ont été corrigés par trois annotateurs humains (à raison de deux annotateurs par texte). La correction a été faite avec les outils Callisto[6] et Tango[7], conçus pour les tâches en question. Le cycle auquel est soumis chaque document est décrit à la figure 1.

Notons que ce cycle se termine par la vérification de la cohérence des graphes temporels produits pour chaque document. Cette vérification a été faite par l'application de l'algorithme d'Allen (Allen, 1983) par saturation des graphes temporels (Tannier & Muller, 2008). À ce stade, le corpus contenait un total de 8 incohérences, qui ont été résolues à la main. le corpus final ne contient aucun graphe temporel incohérent. Pour comparer avec le TimeBank 1.2 de l'anglais, nous avons effectué la même vérification sur ce corpus et avons trouvé 18 graphes incohérents (sur le total de 183 fichiers). Enfin, les textes du corpus ont été validés selon une DTD ISO-TimeML pour le français, que nous fournissons avec le corpus.

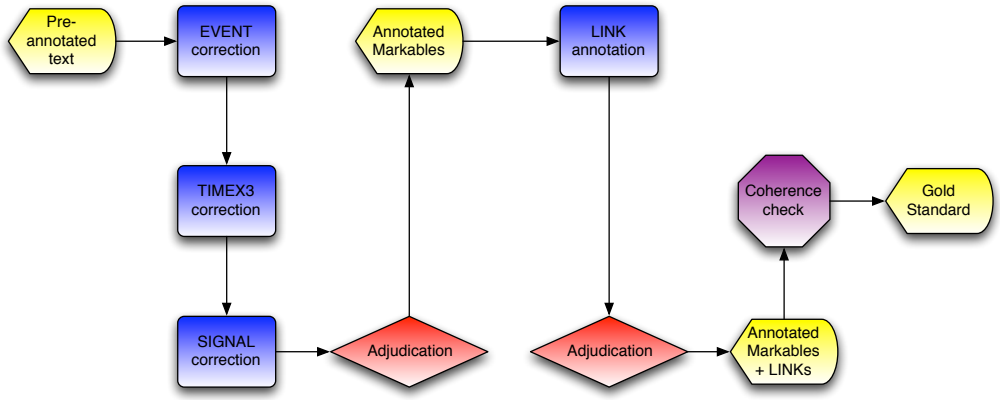


FIG. 1 – Schéma des étapes de la stratégie d'annotation adoptée.

[6] http://callisto.mitre.org/
[7] http://timeml.org/site/tango/tool.html

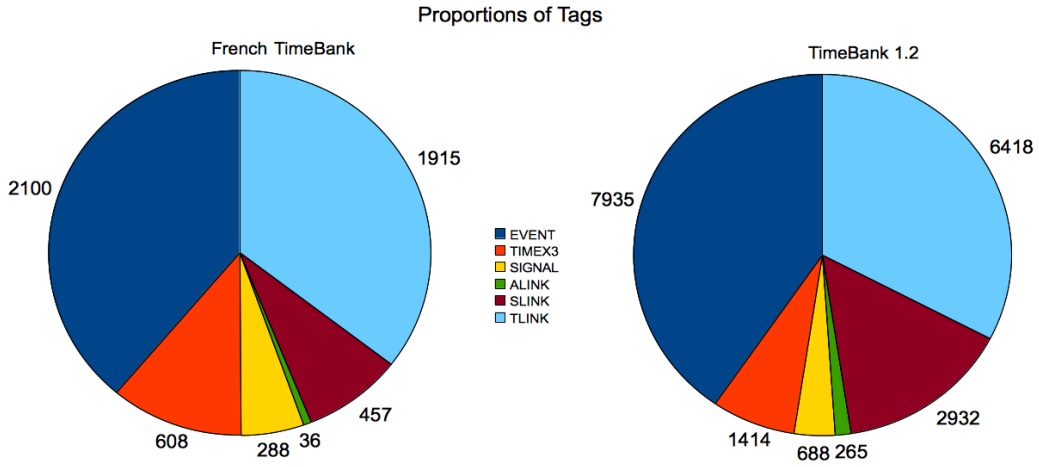


FIG. 2 – Contenu du French TimeBank comparé avec TimeBank 1.2.

5 French TimeBank

Notre projet pour le FTiB est de proposer un corpus de taille comparable à celle du TimeBank 1.2 de l'anglais (environ 61 000 tokens). La version 1.0 que nous présentons ici, et qui a été mise en ligne en janvier 2011, représente environ un quart de cette taille. Les quantités et proportions pour les divers éléments annotés sont donnés dans la Figure 2 avec, pour comparaison, les chiffres correspondants pour TimeBank 1.2.

On constate que les proportions des éléments annotés pour le français sont, pour la plupart, très proches de celles de TimeBank 1.2. La plus grande proportion de `<TLINK>` dans le French TimeBank est due au fait que nous avons annoté les relations temporelles entre les `<TIMEX3>` qui avaient une valeur pleinement spécifiée, alors que ce n'était pas le cas pour l'anglais. Lorsqu'on enlève ces relations, les proportions se rapprochent. Cette similarité dans les proportions nous semble indiquer que les consignes d'annotation ont été appliquées de façon comparable sur les deux corpus. Cela suggère également que, pour le genre journalistique, les distributions des différents types d'éléments sont similaires en anglais et en français.

Nous avons examiné l'effet de notre stratégie d'annotation sur le contenu du corpus, en particulier les effets de l'échantillonnage de textes et l'éventuel biais introduit par la pré-annotation automatique. En ce qui concerne l'échantillonnage, nous avons cherché à savoir si une corrélation existe entre le sous-genre de texte et le contenu linguistique servant à exprimer la temporalité. Nous présentons les résultats dans la Section 5.1. L'étude des effets de la pré-annotation se focalise sur les influences positive et négative du traitement préalable sur le résultat final. Nous présentons cette expérience dans la Section 5.2.

5.1 Sous-genre textuel et contenu linguistique

Nous avons observé un certain nombre de corrélations entre le sous-genre du texte et son contenu linguistique. Nous nous focaliserons ici sur les corrélations détectées entre le sous-genre et les types d'expressions temporelles employées, ainsi qu'entre le sous-genre et les classes des événements mentionnés.

La Figure 3 montre les pourcentages de chaque type d'expression temporelle, ainsi que les proportions des classes d'événements annotés pour chacun des sous-genres textuels.

La variation du nombre total d'expressions temporelles est due à la différence de proportions de chaque sous-genre dans le corpus. On constate que le sous-genre d'annonces contient une proportion importante (19/41, ou 46%) des expressions de type `TIME`, alors que les autres sous-genres en contiennent des proportions relativement basses (de 2% à 24%). Cela est d'autant plus significatif compte tenu du fait que les annonces représentent une proportion relativement faible du total des tokens du corpus. Il se trouve alors que ce sous-genre compense le manque de ce type d'expressions dans les autres sous-genres. Cela montre l'effet positif de l'inclusion de ce sous-genre, qui contribue visiblement à la diversité linguistique du corpus.

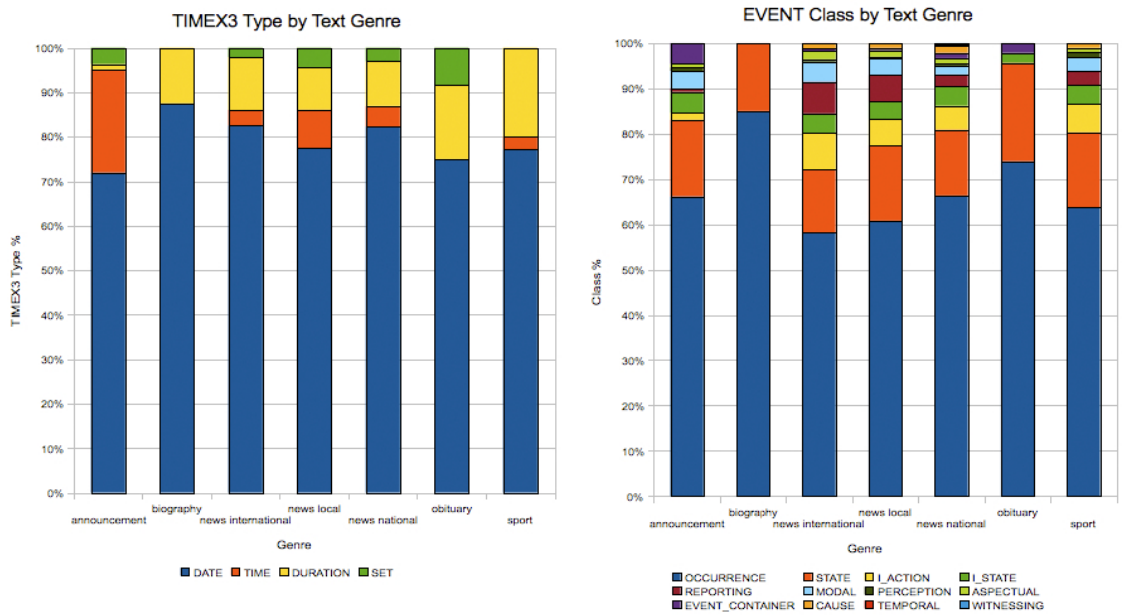


FIG. 3 – Distribution de marquables par sous-genre.

On constate également que les durées (DURATION) sont relativement peu fréquentes dans les annonces (2% du total), alors qu'elles sont plus utilisées dans les autres sous-genres, en particulier les actualités (de 21% à 32%) et les actualités sportives (13,5%). Là encore cela suggère un certain équilibre dans le corpus, une des motivations pour l'échantillonnage.

Les expressions de DATE sont de loin les plus fréquentes à travers les sous-genres, avec presque 80% des expressions annotées. Inversement, les expressions quantifiées ou de fréquence SET sont les moins fréquentes (3% du total). Les SET apparaissent dans des proportions relativement homogènes dans les annonces et les actualités, mais relativement peu ou pas du tout dans les autres genres. Les nécrologies et les biographies sont trop peu représentées dans le corpus pour pouvoir tirer des conclusions, mais il est intéressant de noter que les actualités sportives ne contiennent aucune expression de type SET. Ce n'est pas étonnant si l'on note que ces articles se concentrent sur la description d'un événement sportif particulier au lieu de décrire des événements récurrents.

Une étude préliminaire nous a permis de noter la présence de certaines corrélations entre les sous-genres textuels du corpus et les types d'expressions temporelles qu'ils contiennent. Des tendances assez claires apparaissent pour les annonces, qui contiennent une proportion relativement importante d'expressions d'heure et peu de durées, une tendance qui est compensée par les autres sous-genres du corpus. Les dates et les expressions quantifiées sont distribuées de façon plutôt homogène. On observe ainsi une certaine variété dans les expressions, ainsi qu'un équilibre, que l'on peut attribuer à la politique d'échantillonnage. Cela tend à valider le choix de sélectionner les textes en fonction du sous-genre de texte.

Maintenant, examinons les événements qui sont annotés dans les différents sous-genres de textes. Nous nous focaliserons sur le rapport entre les classes d'événements annotés et les sous-genres de textes. On remarque immédiatement la prépondérance de la classe OCCURRENCE, qui représente 62,1% de tous les événements annotés, 4 fois plus que la deuxième classe la plus fréquente, STATE à 15,4%. Cette tendance se trouve à travers tous les sous-genres, avec quelques exceptions. Premièrement, les articles d'actualité, en particulier les actualités internationales et locales contiennent des proportions significatives de la classe REPORTING, avec 7% et 5,4%, respectivement. Les autres sous-genres contiennent environ la moitié de cette proportion (3.1% pour le sport, 2.6% pour les actualités nationales, 0.8% pour les annonces et 0 pour les biographies et les nécrologies). 84 des 102 (82,5%) événements de classe REPORTING appartiennent aux sous-genres d'actualités internationales et locales. Cela montre l'importance d'avoir inclu ces sous-genres, qui contiennent des quantités significatives de discours rapporté, dans le corpus. La même tendance est apparente, même si elle est moins marquée, pour la classe MODAL, qui apparaît plus fréquemment dans les articles d'actualité, et dans des proportions légèrement plus faibles dans les annonces et les actualités sportives. Cela suggère que la modalité est une caractéristique générale de la langue

utilisée dans la description d'éventualités, malgré une légère corrélation avec le sous-genre du texte.

Les annonces se démarquent encore par une proportion relativement élevée de la classe EVENT_CONTAINER. Nous rappelons que cette classe est utilisée pour classifier les verbes comme *avoir lieu* et *se passer* qui prennent un sujet événementiel. Non seulement cette classe se trouve dans la plus grande proportion dans ce sous-genre de texte, mais elle y apparaît également le plus fréquemment (6 occurrences contre 5 dans les actualités nationales). Dans les annonces, toutes les occurrences de cette classe sont utilisées pour relier le sujet événementiel à une expression temporelle, comme dans (3-a). Cela reflète le fait que la fonction de tels documents est de préciser le moment précis auquel des événements auront lieu. Il est intéressant de constater que cette classe sert une tout autre fonction dans les textes de nouvelles internationales, où elle est annotée dans des cas de localisation spatiale d'un événement (3-b), ou la manière d'occurrence (3-c).

(3) a. *La distribution de lunettes spéciales pour l'éclipse aura lieu pour les administrés mardi 10 août...*
b. *Une violente explosion survenue dans la maison d'un membre des Brigades Ezzedine al-Qassam...*
c. *Le second remplace Edith Cresson, par qui le scandale est arrivé...*

Les classes I_ACTION et I_STATE apparaissent dans des proportions assez uniformes à travers les sous-genres (hormis dans les biographies et les nécrologies qui sont trop peu représentées). Les classes ASPECTUAL, REPORTING et PERCEPTION figurent toutes dans des proportions relativement très basses. La disparité entre la classe "par défaut", OCCURRENCE, et les autres suggère que la typologie des événements pourrait être affinée. Une possibilité serait de distinguer plus finement les différentes classes aspectuelles, par exemple pour annoter la différence entre les événements duratifs et ponctuels. Cette distinction est particulièrement pertinente lorsqu'on souhaite annoter les relations qui existent entre deux événements. Par exemple, un événement ponctuel peut être temporellement inclus dans un événement duratif, mais le contraire n'est pas possible. Ces annotations sont prévues dans les prochaines éditions de la norme ISO-TimeML.

5.2 Effets de la pré-annotation automatique

Nous avons choisi d'effectuer une pré-annotation automatique des marquables dans les textes, suivie d'une correction manuelle, une pratique courante dans les projets d'annotation linguistiques (Marcus *et al.*, 1993, p. ex.). Néanmoins, à l'heure actuelle, aucune évaluation des effets de la pré-annotation n'a été publiée pour la tâche de l'annotation temporelle en ISO-TimeML. Dans cette section, nous décrivons une expérience menée pour déterminer les effets de la pré-annotation dans le cadre de la création du French TimeBank. Nous avons examiné deux point principaux : l'effet sur le temps d'annotation et l'éventuelle introduction par l'annotation préalable d'un biais sur les choix des annotateurs. L'expérience a été effectuée sur un sous-ensemble de 8 documents (956 tokens, 121 <EVENT>, 27 <TIMEX3> et 18 <SIGNAL>) ayant déjà été annotés en marquables[8] suivant la stratégie décrite dans la Section 4.3. Nous avons mesuré le temps d'annotation manuelle à 85 minutes, alors que la correction d'une pré-annotation a été mesurée à 47 minutes – une réduction presque de moitié. Cette réduction est certainement due au fait que les balises contiennent des attributs fastidieux à annoter, avec des valeurs multiples ou qui doivent respecter un format très spécifique, notamment l'attribut value des <TIMEX3>. La pré-annotation accélère cette tâche en permettant à l'annotateur d'effectuer une simple vérification suivie d'une correction, si nécessaire. Une deuxième partie de cette expérience consistait à mesurer l'influence de la pré-annotation sur les choix des annotateurs, ce que nous appelons le *biais*[9]. Le biais positif représente les erreurs évitées par la pré-annotation et le biais négatif les erreurs attribuables à la pré-annotation. Afin de vérifier le biais, le document pré-annoté (D_p) et le document manuellement annoté (D_m) ont été comparés avec le document validé du corpus correspondant (D_r, le document de référence). Pour un marquable donné (empan de balise ou attribut), si les conditions suivantes étaient remplies, la différence dans les annotations était attribuée à un biais positif :

1. D_p et D_r ont la même annotation,
2. D_m est différent de D_p et D_r,
3. D_m est jugé incorrect.

Le biais négatif était mesuré de façon similaire, mais la condition 3 était :

3. D_m est jugé correct.

[8]Les relations ont été entièrement annotées à la main et donc cette expérience ne s'applique pas aux relations.
[9]Nous précisons qu'il ne s'agit pas ici de la notion de biais utilisée dans le domaine des statistiques.

Biais positif : le biais positif introduit par la pré-annotation se voit particulièrement dans l'annotation des expressions temporelles. Les erreurs manuelles pour l'annotation des événements étaient moins fréquentes. La première colonne du Tableau 4 donne les erreurs d'annotation manuelle repérées avec leur taux d'erreur[10].

Biais négatif : la plupart des erreurs attribuables à la pré-annotation étaient commises sur les événements. La deuxième colonne du Tableau 4 montre les erreurs introduites par la pré-annotation mais évitées dans l'annotation manuelle[11].

	Biais positif		**Biais négatif**	
Balise	**Type erreur**	**% erreur**	**Type erreur**	**% erreur**
<TIMEX3>	Mauvais empan	14.8	Mauvais empan	3.7
	`type` omis	25.9		
	`value` omis	18.5		
	Erreur de `value`	11		
	Erreur de format `value`	7.4		
<EVENT>	Fausse balise	4.1	Balise manquante	2.5
	`class` omis	3.3	Erreur de `class`	0.8
	Erreur de `class`	5.4	Erreur de `tense`	0.8
	Erreur de `tense`	2.5	Erreur de `aspect`	0.8
	Erreur de `aspect`	2.5		

TAB. 4 – Biais positif et négatif de la pré-annotation.

On voit que le nombre d'erreurs introduites par la pré-annotation et non repérées pendant la correction manuelle est relativement bas. Cela suggère que la pré-annotation a introduit peu d'erreurs, mais aussi que les annotateurs sont restés vigilants pour corriger celles qui restaient. Ces résultats montrent que la pré-annotation apporte plus d'avantages que d'inconvénients, notamment en réduisant de façon significative le temps d'annotation et le taux d'erreur humaine.

6 Conclusion

Dans cet article, nous avons présenté le French TimeBank (FTiB), un corpus de référence sur la temporalité pour le français. Cette ressource est librement disponible et adhère à la norme ISO-TimeML pour l'annotation temporelle. Bien qu'encore de taille modeste (un quart des tokens du TimeBank anglais), le FTiB devrait néanmoins grandement favoriser le développement et l'évaluation des systèmes pour le français. Bien évidemment, ce corpus va permettre l'usage de systèmes basés sur l'apprentissage automatique, mais il fournit également un matériau intéressant pour approfondir les études linguistiques sur la temporalité, ce que nous avons modestement entamé dans cet article. Par exemple, la détection du genre textuel pourrait tirer parti de caractéristiques distributionnelles des expressions temporelles et événementielles, comme suggère l'étude préalable que nous avons évoquée ici. Les analyses préliminaires des données que nous avons présentées ont fourni un premier aperçu du contenu du corpus.

La constitution de ce corpus nous a permis de tirer un certain nombre d'enseignements sur l'annotation temporelle, et l'annotation sémantique en général. Tout d'abord, nous avons pu constater que la norme ISO-TimeML semblait relativement stable et pouvait être appliquée au français, moyennant une série d'amendements et enrichissements (certains de ceux-ci dépassent d'ailleurs le cadre strict du français et ont une vocation multilingue). Ce travail nous a aussi conduit à discuter le principe d'annotation surfacique : il s'agit d'un principe souhaitable : il permet de rester neutre vis-à-vis des théories linguistiques du temps et de l'aspect (particulièrement nombreuses...), et aussi de conserver les ambiguïtés, pour permettre leur étude en tant que telles. Mais il doit être tout aussi clair qu'une annotation **intrinsèquement** surfacique reviendrait à ne marquer que ce qui est déjà visible, et par conséquent ne serait pas très utile. Nous avons donc fréquemment été conduits à introduire de l'interprétation dans l'annotation, en essayant de bien délimiter les cas concernés. C'est sur ce fil entre annotation redondante et surinterprétation que doit se tenir, nous semble-t-il, toute entreprise d'annotation sémantique.

Comme nous l'avons montré, la construction du FTiB a été le fruit d'une méthodologie réfléchie, basée sur un échantillonnage rigoureux et des cycles d'annotations combinant pré-étiquetage automatique et annotations/

[10] Le taux d'erreur est calculé comme le nombre d'erreurs divisé par le total de balises manuellement annotées $\times$ 100.
[11] Cette fois, le taux d'erreur est calculé comme le nombre d'erreurs divisé par le nombre de balises dans D_r $\times$ 100.

corrections humaines. Cette méthodologie a fourni des résultats positifs, notamment en termes de réduction du temps d'annotation et du taux d'erreur humaine. Notre démarche pourrait être suivie pour la création de corpus similaires au FTiB, ou dans d'autres tâches d'annotation. Notre expérience argumente en faveur d'une pré-annotation automatique suivie d'une correction par des annotateurs humains et on pourrait envisager de tirer profit des données existantes afin d'annoter le reste du corpus avec un système statistique.

Remerciements

Ce travail a été réalisé pendant le doctorat d'André Bittar, dans le laboratoire ALPAGE, sous la direction de Laurence Danlos, Pascal Amsili et Pascal Denis. Les auteurs souhaitent remercier, pour leur contribution à différents stades de ce travail, Philippe Muller, Michel Gagnon, et Gabriel Parent, sans oublier, bien sûr, Laurence Danlos.

Références

ALLEN J. F. (1983). Maintaining Knowledge About Temporal Intervals. In *Communications of the ACM*, volume 26, p. 832–843.

BACH E. (1986). The algebra of events. *Linguistics and Philosophy*, **9**(1).

BITTAR A. (2010a). *Building a TimeBank for French : a reference corpus annotated according to the ISO-TimeML standard*. PhD thesis, Université Paris Diderot, Paris, France.

BITTAR A. (2010b). *ISO-TimeML Annotation Guidelines for French*. Alpage-Université Paris Diderot, Paris, France.

CASELLI T. (2008). *It-TimeML : TimeML Annotation Guidelines for Italian, Version 1.0*. Rapport interne, Instituto di Linguistica Computazionale, C.N.R.

GROSS M. (2002). Les déterminants numéraux, un exemple : les dates horaires. *Langages*, (145), 21–38.

HATHOUT N., NAMER F. & DAL G. (2002). An Experimental Constructional Database : The MorTAL Project. In P. BOUCHER, Ed., *Many Morphologies*, p. 178–209. Somerville, Mass., USA : Cascadilla.

ISO (2008). *ISO DIS 24617-1 : 2008 Language Resource Management - Semantic Annotation Framework - Part 1 : Time and Events*. International Organization for Standardization, ISO Central Secretariat, Geneva, Switzerland.

KAMP H. & ROHRER C. (1983). Temporal reference in french. Manuscrit, Universität Stuttgart.

MARCUS M. P., SANTORINI B. & MARCINKIEWICZ M. A. (1993). Building a Large Annotated Corpus of English : The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

NASR A., BÉCHET F. & REY J.-F. (2010). MACAON : Une chaîne linguistique pour le traitement de graphes de mots. In *Actes de TALN 2010*, Montreal, Canada.

PALMER F. R. (1986). *Mood and Modality*. Cambridge, UK : Cambridge University Press.

PARENT G., GAGNON M. & MULLER P. (2008). Annotation d'expressions temporelles et d'événements en français. In *Actes de TALN 2008*, Avignon, France.

PAUMIER S. (2008). *Unitex 2.0 User Manual*. Université Paris-Est Marne-la-Vallée, Marne-la-Vallée, France.

PUSTEJOVSKY J., HANKS P., SAURÍ R., SEE A., GAIZAUSKAS R., SETZER A., RADEV D., SUNDHEIM B., DAY D., FERRO L. & LAZO M. (2003). The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics 2003*, p. 647–656.

PUSTEJOVSKY J., VERHAGEN M., SAURÍ R., LITTMAN J., GAIZAUSKAS R., KATZ G., MANI I., KNIPPEN R. & SETZER A. (2006). TimeBank 1.2. Linguistic Data Consortium.

TANNIER X. & MULLER P. (2008). Evaluation Metrics for Automatic Temporal Annotation of Texts. In E. L. R. A. (ELRA), Ed., *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

VENDLER Z. (1967). *Linguistics and Philosophy*. Ithaca, N.Y. : Cornel University Press.

Acquisition automatique de terminologie à partir de corpus de texte

Edmond Lassalle

(1) Orange Labs, 2 avenue Pierre Marzin
22 307 Lannion - France
edmond.lassalle@orange-ftgroup.com

Résumé :

Les applications de recherche d'informations chez Orange sont confrontées à des flux importants de données textuelles, recouvrant des domaines larges et évoluant très rapidement. Un des problèmes à résoudre est de pouvoir analyser très rapidement ces flux, à un niveau élevé de qualité. Le recours à un modèle d'analyse sémantique, comme solution, n'est viable qu'en s'appuyant sur l'apprentissage automatique pour construire des grandes bases de connaissances dédiées à chaque application. L'extraction terminologique décrite dans cet article est un composant amont de ce dispositif d'apprentissage. Des nouvelles méthodes d'acquisition, basée sur un modèle hybride (analyse par grammaires de chunking et analyse statistique à deux niveaux), ont été développées pour répondre aux contraintes de performance et de qualité.

Abstract :

Information retrieval applications by Orange must process tremendous textual dataflows which cover large domains and evolve rapidly. One problem to solve is to analyze these dataflows very quickly, with a high quality level. Having a semantic analysis model as a solution is reliable only if unsupervised learning is used to build large knowledge databases dedicated to each application. The terminology extraction described in this paper is a prior component of the learning architecture. New acquisition methods, based on hybrid model (chunking analysis coupled with two-level statistical analysis) have been developed to meet the constraints of both performance and quality.

Mots-clés : Apprentissage automatique, acquisition terminologique, entropie, grammaires de chunking
Keywords: Unsupervised learning, terminology acquisition, entropy, chunking analysis

1 Introduction

Une amélioration significative de la qualité des moteurs de recherche concerne l'identification des locutions en tant qu'unités de sens. C'est aussi une difficulté dans le cas de certaines applications d'Orange. Le problème est en effet de pouvoir prendre en compte une terminologie en constante évolution dans des domaines liés à l'actualité (presse, journaux télévisés...). Il s'agit en plus de traiter en continu des flux importants de données pour indexer les nouveaux documents entrants mais aussi pour acquérir une terminologie évanescente (*fuite de pétrole*, *nuage de cendres*, *Jean Paul II*, *Sidi Bouzid*, *Antoine de Léocour* ...). Les méthodes d'acquisition automatique de terminologie à partir de corpus trouvent ici leur entière justification.

Un examen de différents modèles d'apprentissage, de leur adéquation aux corpus dans nos applications va motiver une architecture hybride différente de celles connues et étudiées à ce jour. Ce choix oblige à innover dans les méthodes d'analyse linguistique et statistique pour répondre aux contraintes opérationnelles de qualité. L'objet de nos travaux est alors, d'avoir un système «homogène» pour limiter le biais statistique inhérent aux interactions dans tout modèle hybride. La loi binomiale régissant le comportement des mots constitue donc la seule hypothèse de départ. Des observations expérimentales, une modélisation formalisée permettent ensuite de dériver par calcul les autres lois. Les résultats obtenus vont confirmer la pertinence de cette démarche. Dans la suite de l'article, une description du modèle d'apprentissage, des méthodes d'analyse statistique va donner un éclairage sur le fonctionnement de notre composant linguistique.

2 Motivation d'un modèle hybride d'acquisition terminologique

Le choix d'une architecture est dicté par le type de corpus d'apprentissage. Le nôtre est constitué de textes décrivant des vidéos sur un mois d'actualités (http://www.2424actu.fr/actualite-du-jour/). A chaque instant, on dispose de 100 000 textes pour un total de 5 millions de mots. Chaque texte comprend un titre suivi d'un résumé court comme : «*Tunisie : affrontements à **Sidi Bouzid**. De nouveaux affrontements violents ont eu lieu dans la nuit dans la région de **Sidi Bouzid**, dans le **centre-ouest** de la Tunisie, faisant un blessé par balle et des **dégâts matériels** importants, a-t-on appris dimanche de **sources syndicales** Des centaines de Tunisiens ont participé à une manifestation.*»

Dans ce type de corpus, certaines locutions – étant communes (*dégâts matériels*, *sources syndicales*) – peuvent être obtenues hors méthodes d'apprentissage, mais d'autres (*Sidi Bouzid* ou *Camp Nou*) risquent de ne pas figurer dans un référentiel lexical qui serait établi *a priori*. Le problème à traiter est donc d'avoir un référentiel de mots simples exhaustif, incluant des mots inconnus. Une analyse visant à extraire des locutions devra ensuite identifier des constructions bien formées de groupes de mots, puis reconnaître la nature compositionnelle ou figée du sens porté par ces constructions, y compris celles comportant des mots inconnus. Les solutions à cette problématique peuvent être d'ordre statistique ou mixte, mais excluent une approche symbolique confrontée au problème d'exhaustivité.

2.1 Modèles statistiques

L'apport des méthodes statistiques concerne la quantification de la compositionalité. L'occurrence d'un mot m_i dans un corpus est modélisé par une loi de Bernoulli de paramètre p_i. Le comportement d'un mot dans le corpus est ensuite expliqué par sa fréquence d'occurrences et donc par une v.a.r de loi binomiale $B(n,p_i)$. Estimer le degré de compositionalité de deux mots contigus revient alors à déterminer le degré de dépendance des v.a.r associées à ces mots. Deux méthodes expérimentales permettent de réaliser ce calcul :

- La première nécessite une fenêtre d'observation (par exemple la phrase) pour estimer les probabilités d'occurrences et de cooccurrences à partir d'un comptage fréquentiel. Elle conduit au calcul de l'information mutuelle (Church et al., 1990) ou à la mesure de Dice (Smadja, 1993). Citons aussi pour cette méthode, le calcul de la log-perplexité (Kit, 2002) qui a l'avantage de prendre en compte des séquences de N mots mais nécessite en contre partie un modèle de langue pour viabiliser l'estimation de la probabilité de telles séquences.
- La seconde réalise un comptage fréquentiel direct de la cooccurrence, de la non-cooccurrence et des non-occurrences de deux mots contigus pour déterminer la log-vraisemblance des 2 v.a.r associées (Dunning, 1993) ou aussi leur corrélation via le calcul du χ^2.

Le résultat pour ces 2 méthodes est un classement suivant une «vraisemblance d'être une locution». La difficulté restant est de déterminer la valeur de seuillage, mais aussi de mesurer l'importance des termes par rapport au corpus applicatif.

Pour traiter ce dernier point, les modèles les plus avancés (Kit, 2002) (Vu et al., 2008) (Kageura et al., 1996) caractérisent les séquences extraites par le critère d'unithood, validant statistiquement la cohérence de la séquence, et par le critère de termhood, caractérisant la spécificité de la séquence par rapport au corpus applicatif. En l'absence d'analyse linguistique, le premier critère permet de valider la construction syntaxique de la séquence tandis que le second critère valide à la fois la non-compositionalité et l'importance de cette séquence. Cette approche est adaptée pour les domaines techniques où le vocabulaire est limité, où les expressions figées peuvent être longues comme *Altération des facteurs de coagulation sanguine*, où le critère de spécificité est assez proche du critère de non-compositionalité. Une variante intéressante (Frantzi et al., 1999) est d'introduire le filtrage de catégories grammaticales et de palier l'absence d'analyse syntaxique par des mesures statistiques (AC/NC-value).

2.2 Modèles hybrides

Le modèle le plus usité est basé sur un fonctionnement en tandem du composant linguistique et du composant statistique. L'avantage d'une telle architecture concerne la modularité. L'analyse linguistique est chargée d'annoter le corpus initial (étiquetage grammatical, parenthésage et étiquetage des syntagmes). L'analyse statistique reprend les informations annotées pour produire une liste de termes classés suivant un ordre de vraisemblance. Cette approche permet en plus de reprendre pour le deuxième composant (Daille, 1996) les mesures utilisées par les modèles statistiques. L'inconvénient du modèle en tandem concerne le biais statistique. Les évaluations que nous avons menées (Lassalle et al., 2011) sur Acabit ont indiqué un différentiel de 30% du taux de précision suivant que nous utilisons en amont, comme composant linguistique, l'analyseur de Brill (Brill, 1992) couplé au lemmatiseur Flem (Namer, 2000) ou l'analyseur Tilt (Heinecke et al., 2008).

Seul un couplage fin entre analyse linguistique et analyse statistique permettrait de minimiser ce biais. Ce qui exclut une réutilisation des mesures de classement des modèles statistiques car ces dernières nécessitent, dans le calcul, des données globales et non partielles comme c'est le cas dans un couplage fin. Cela nous conduit à spécialiser nos méthodes d'analyse statistique dans deux directions :

- la première pour détecter les éléments saillants (analyse de régularité)
- la seconde pour estimer la non-compositionalité des constructions syntaxiques.

Le rôle de l'analyse linguistique dans cette approche hybride est de proposer successivement des ensembles «statistiquement cohérents» de constructions syntaxiques. Ce que nous préciserons après avoir décrit dans un premier temps les analyses statistiques.

3 Analyse statistique de la régularité

Les finalités de l'analyse statistique décrite dans cette section sont triples. La même observation expérimentale permet en effet de déduire les caractéristiques des mots dans le corpus, suivant :

- une loi de distribution décrivant leur occurrence,
- des propriétés macroscopiques autorisant leur regroupement au sein de catégories grossières,
- et le degré de saillance permettant d'identifier les mots importants dans le corpus.

Seul, le calcul de saillance est prééminent dans l'acquisition de terminologie. La loi de distribution permet de déduire la loi conjuguée *a priori* et elle est plutôt utilisée dans nos modélisations bayésiennes, comme dans la catégorisation[1] ou dans l'indexation. Le regroupement des mots en catégories grossières, bien qu'utile dans le processus d'acquisition terminologique, nécessite une extension (restant à faire) du calcul de saillance.

3.1 Loi de distribution des mots

Si l'on accepte que l'occurrence d'un mot dans un corpus suit une loi de Bernoulli de paramètre p, alors sa fréquence d'apparition dans une fenêtre de n mots d'un corpus suit la loi binomiale B(n,p). La valeur de p est en général très faible, à l'exception des mots grammaticaux et des termes de domaine (dans le cas de corpus

[1]Blei D.M., (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research*

spécialisés comme ceux de la médecine, des finances,...). Il est donc possible pour les grandes valeurs de n d'approximer la loi binomiale B(n,p) par une loi de Poisson ou par une gaussienne discrétisée (Saporta 2006).

L'intérêt d'une loi de Poisson P(λ) par rapport à une gaussienne est d'avoir l'espérance et la variance égales à λ. Pour les grandes valeurs de λ (λ>18), P(λ) peut être confondue à une loi de Gauss (Saporta, 2006), avec l'avantage d'être caractérisée par un seul paramètre. L'estimation d'un seul paramètre (espérance = variance) plutôt que 2 présente un gain important en qualité dans l'apprentissage à condition que la loi de Poisson soit justifiée.

Le problème est donc de savoir, à partir d'observations expérimentales, quand représenter les fréquences d'occurrence par une loi de Poisson, c'est à dire, pour les grandes valeurs de fréquence quand représenter par une gaussienne à un seul paramètre ou par une gaussienne à 2 paramètres. Nous nous appuierons sur le théorème suivant (Saporta, 2006) pour affecter expérimentalement les mots observés dans l'une de ces 2 catégories.

Théorème :
Si X_n est une suite de variables binomiales B(n,p) telles que quand $n\to\infty$ et $p\to 0$, np tend vers une limite finie λ. Alors X_n converge en loi vers une variable de Poisson P(λ)

3.2 Méthode expérimentale

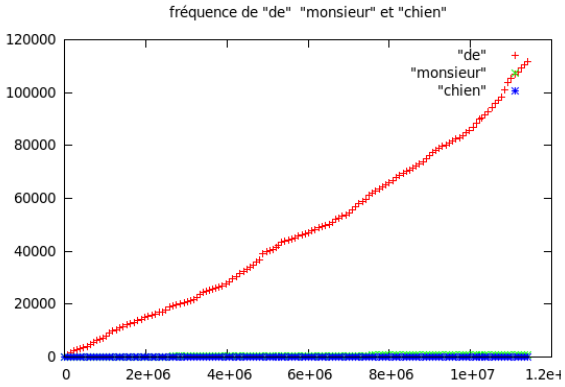
Une loi empirique comme celle de Zipf permet d'estimer si le contenu d'un texte est porteur de sens ou s'il relève d'une écriture aléatoire. Par contre, cette loi n'est pas adaptée à une analyse plus fine, car approximative et non discriminante pour les faibles valeurs de fréquence de mot (i.e classé en rang élevé dans la loi de Zipf). Nous proposons donc une nouvelle méthode d'analyse dynamique de corpus pour caractériser les probabilités d'occurrence des mots, et simultanément pour classer ces derniers en mots grammaticaux (mots " vides "), mots spécialisés de domaine ou mots courants :

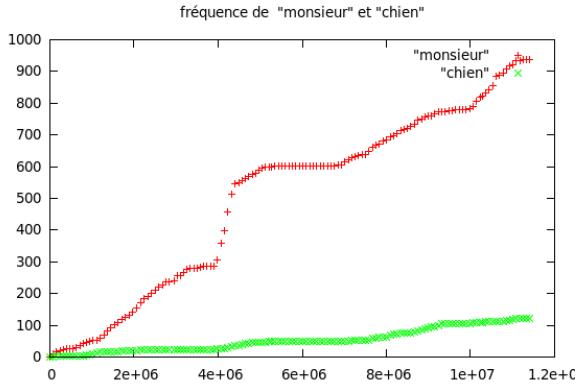
- Le corpus est analysé en flux continu. L'observation est réalisée périodiquement c'est-à-dire qu'on fige le comptage fréquentiel de tous les mots tous les k mots observés dans le corpus. Si, après avoir parcouru n premiers mots, on a décompté f_i occurrences d'un mot m_i alors $f_i \sim np_i$, où p_i est la probabilité d'apparition du mot m_i. D'après le précédent théorème, il suffit d'observer l'évolution de f_i en fonction de n quand n varie de 0 à taille maximale du corpus (que l'on considère comme très grand # ∞). En fonction de l'allure de la courbe $f_i(n)$ observée, on peut ensuite opter pour la loi décrivant le mieux la fréquence d'apparition du mot m_i.
- Si $f_i(x)$ tend vers une droite asymptote d'équation $y=c^{te}$ alors le théorème précédent s'applique. La distribution du mot m_i peut être alors modélisée par une loi de Poisson (et donc, pour les grandes valeurs de n, par une gaussienne à un seul paramètre λ). Dans une étude expérimentale, une courbe faiblement croissante, par exemple en log(x) peut aussi être acceptée comme une approximation acceptable de la droite asymptote $y=c^{te}$ (log-linéarité).

Expérimentalement, l'analyse de corpus «relativement» homogènes, comme le nôtre, montre que les fréquences des mots croissent plutôt linéairement. Nous retiendrons donc pour les grandes valeurs de n, un distribution gaussienne à 2 paramètres. De plus, l'analyse de la courbe d'évolution de chaque mot permet de classer ce dernier dans l'une des catégories précédemment évoquées. S'agissant d'un choix empirique des critères discriminants pour le classement, ce choix est justifié surtout par des observations dont l'exemple suivant est décrite en illustration.

3.3 Résultat expérimental et calcul de saillance

Les mots *de*, *monsieur*, *cheval* et *chien* ont été choisis pour représenter des classes de mots grammaticaux, de mots spécialisés et de mots d'emploi général. Leur courbe de fréquence cumulée est analysée sur notre corpus d'actualités. L'accroissement en fréquence du mot *de* est logiquement la plus rapide comme l'indique la figure ci-dessous. Comparativement, les courbes d'évolution des mots *chien* et *monsieur* paraissent plates. Ce n'est pas le cas comme l'indique la figure suivante lorsqu'on change le facteur d'échelle sur l'axe y. On constate aussi que la courbe de croissance du mot *de* est plus régulière autour de la droite qui la sous-tend tandis que les courbes de croissance des mots *monsieur* et *chien* sont plus dispersées.





On cherche donc à quantifier cette dispersion pour servir de critère de discrimination des mots à des fins de classement ou d'ordonnancement. La dispersion peut être traduite par la variance ou mieux, pour disposer d'une échelle de valeur uniformisée, par la forme normalisée qu'est le coefficient de variation.

Le calcul du coefficient de variation se fait comme suit : si f_1, f_2, ..., f_k désignent la suite de fréquences cumulées suivant le comptage décrit plus haut, et si $n_1,n_2,...,n_k$ désignent les nombres cumulés de mots parcourus pour décompter les f_i, alors la moyenne $\mu=\frac{f_k}{n_k}$ et la variance σ telle que $\sigma^2=(\sum_{i=1}^{k-1}\frac{f_{i+1}-f_i}{n_{i+1}-n_i}-\mu)^2$ permettent de calculer le coefficient de variation, égal à $\frac{\sigma}{\mu}$.

3.4 Utilisation du coefficient de variation

L'utilisation du coefficient de variation sur une échelle de valeur scalaire permet d'ordonner les mots (et les locutions une fois apprises) suivant un indice de notoriété. Intuitivement, ce ne sont pas les mots les plus fréquents qui présentent un intérêt mais plutôt ceux utilisés le plus régulièrement dans de nombreux contextes. En plus, en associant à chaque locution apprise sa catégorie grammaticale, et en se focalisant sur les catégories les plus porteuses d'information comme les groupes nominaux ou les patronymes, on arrive ainsi à extraire des éléments saillants mais évanescents comme *nuage de cendres*, *fuite de pétrole*...

3.5 Regroupement en catégories grossières

Le coefficient de variation permet d'estimer l'importance de chaque mot pris isolément par rapport au corpus. Expérimentalement, il permet une séparation effective des mots grammaticaux des autres mots. Mais pour regrouper les mots restants en catégories grossières, on a besoin de plus d'informations, et notamment de quantifier les interactions entre mots.

Pour pouvoir réutiliser les mêmes calculs expérimentaux que précédemment sur la fréquence des mots, et pour conserver une cohérence dans le formalisme de calcul, on remarquera qu'il existe un parallèle entre le coefficient de variation et la notion de tfxidf en recherche d'information (cette dernière correspond dans les modèles probabilistes à la probabilité d'avoir un document pertinent contenant un terme t). L'extension de cette mesure locale, liée à un document, vers une mesure globale sur le corpus se fait naturellement par la notion d'entropie $E(t)=\sum_{d\in D} -p_t\log(p_t)$. Plus un terme est uniformément distribué, plus sa valeur d'entropie est élevée. La notion d'entropie sur un terme isolé s'étend ensuite à celle sur des couples de termes t_1 et t_2 via la notion d'information mutuelle $I(t_1,t_2)=\sum_{d\in D} p(t_1,t_2)\log(\frac{p(t_1,t_2)}{p(t_1)p(t_2)})$. Plutôt que d'utiliser l'information mutuelle comme critère de regroupement des mots en catégories, on utilisera la notion de coefficient de corrélation linéaire entre couple de termes t_1 et t_2, qui est l'extension de la notion de coefficient de variation :

$\rho=\frac{\sigma_{t1,t2}}{\sigma_{t1}\sigma_{t2}}$ où $\sigma_{t1,t2}$ est la covariance de t_1 et t_2, et $\sigma_{t2}\sigma_{t2}$ leur variance respective .

Il s'agit *a posteriori* d'un calcul équivalent puisque $I(t_1,t_2)=-\frac{1}{2}\log(1-\rho^2)$.

La réalisation de cette partie est prévue pour la prochaine version du composant d'acquisition terminologique.

4 Analyse statistique de la compositionalité[2]

La compositionalité des mots est évaluée en linguistique par leur potentiel combinatoire. C'est un comptage fréquentiel, pour un mot donné, de l'appariement d'autres mots dans les constructions observées dans un corpus. Le potentiel combinatoire sert d'indicateur pour faciliter le travail d'analyse d'un lexicologue. Ce critère n'est cependant pas adapté à un apprentissage non supervisé, où l'analyse doit être réalisée automatiquement. Une notion plus appropriée concerne l'entropie, ce qui va être précisé ci-après.

Supposons que, dans un corpus, nous ayons observé 4 fois, le mot *bâton* dont deux avec le qualificatif *rouge*, une avec *bleu* et une avec *vert*. Si l'on souhaite ne garder qu'un seul indicateur qui résume la distribution, estimée à $(\frac{1}{2},\frac{1}{4},\frac{1}{4})$, la somme des probabilités présente peu d'intérêt comme indicateur. Par contre en étudiant la quantité d'information (Shannon, 1948) que chacun des précédents qualificatifs peut apporter au mot *bâton*, soit $(-\log(\frac{1}{2}),-\log(\frac{1}{4}),-\log(\frac{1}{4}))$ la moyenne attendue (espérance) est une bonne indication du degré de compositionalité du mot, soit $-\frac{1}{2}\log\left(\frac{1}{2}\right)-\frac{1}{4}\log\left(\frac{1}{4}\right)-\frac{1}{4}\log\left(\frac{1}{4}\right)=1.5$ dans le cas d'une échelle logarithmique en base 2.

Cette valeur d'entropie indique la quantité d'information que peut recevoir en moyenne chaque mot. Si un mot m_1 a été observé n fois dans un corpus, son entropie a une valeur entre 0 et log(n). Une valeur nulle traduit l'existence d'un mot m_2 dont la probabilité d'observer en cooccurrence avec m_1, vaut 1. Le mot m_1 est dans ce cas non compositionnel puisque fortement lié à m_2. C'est le cas des mots comme *cochère*, *aujourd* , *lurette* ou *escampette*. A l'opposé, une valeur maximale de l'entropie, log(n), correspond à la distribution équiprobable c'est-à-dire à un fort degré de compositionalité. Normalement, c'est vers cette valeur maximale que tendent les mots grammaticaux.

4.1 Champ de compositionalité

Intuitivement, si un mot est employé dans son sens compositionnel, il est fort possible de trouver, dans le corpus, ce mot associé à d'autres mots à des degré divers. Par exemple *bâton* peut être associé à *rouge*, *vert*, *jaune*... et peut-être moins à *joyeux*, *espiègle*, *content*. Le champ de compositionalité d'un mot m correspond à une distribution probabiliste sur l'ensemble des mots m_i et traduit la probabilité d'observer m_i sachant qu'on a observé le mot m. Cette distribution peut être résumée par sa moyenne (entropie), sa variance et sa loi de distribution. Le champ ainsi défini permet d'introduire la notion d'intervalle de confiance et de déterminer de quelle manière une construction est jugée compositionnelle. Dans le cas d'une locution comme *retour de bâton*, il n'existe pas de forme altérée ou modifiée ne comportant qu'une partie de mots de ce groupe. Si la locution est souvent employée dans le corpus, la fréquence d'association est plus élevée que le cas des constructions compositionnelles, ce qui doit permettre à une analyse statistique de conclure que l'un des mots *retour* ou *bâton* n'appartient pas au champ compositionnel de l'autre mot.

4.2 Méthode expérimentale

Pour chaque mot m_0, l'entropie et la variance sont déduites expérimentalement à partir d'un comptage fréquentiel de cooccurrences :

- Pour chaque mot m_0, on procède au comptage de cooccurrence $f_{0,i}$ (resp. $f_{i,0}$) des mots m_i contigus au mot m_0 à droite (resp. à gauche). Le positionnement gauche/droite reflète la nature séquentielle du corpus de texte.
- Pour évaluer le degré de compositionalité, le comptage ne devrait porter que sur les mots m_i ayant un sens compositionnel avec le mot m_0 et exclure les mots m_i lorsque m_0m_i constitue une locution. Au stade de l'apprentissage, on ne dispose pas d'une telle information. L'hypothèse est que les mots m_i constituant une locution sont en plus faible nombre que les mots m_i portant un sens compositionnel. Cela justifie l'approximation dans l'estimation de la moyenne et de la variance.
- Pour tenir compte de la masse absente (due au manque d'exhaustivité de tout corpus), on procède à un lissage de Laplace. La valeur de lissage est plus petite que 1, en raison des faibles fréquences de cooccurrence.

[2] Les notions de compositionalité, de champ de compositionalité... sont revues ici dans une logique calculatoire

- La probabilité $p_{0,i}$ d'observer le mot m_i est estimée par $p_{0,i}=\frac{f_{0,i}}{\sum_j f_{0,j}}$. L'entropie μ_0 et la variance σ_0 sont estimés par $\mu_0=-\sum_j p_{0,j}\log p_{0,j}$ et $\sigma_0^2=\sum_j p_{0,j}(\mu_0-\log p_{0,j})^2$.

4.3 Modélisation de la loi de compositionalité

Il reste à déterminer la loi de distribution des log $p_{0,i}$ pour pouvoir fixer l'intervalle de confiance à l'intérieur duquel une association de mots est considérée comme compositionnelle.

- Pour chaque mot m_0, la cooccurrence d'un mot m_i peut être considérée comme une épreuve de Bernoulli et la fréquence de cooccurrence comme une v.a.r X_i de loi binomiale de paramètre p_i.
- Pour les grandes valeurs de fréquence, la loi de X_i peut être approximée par une loi gaussienne. Nous nous intéressons pour la suite à la v.a.r $\frac{X_i}{n_0}$ où n_0 est le nombre total de cooccurrences observées pour le mot m_0. $\frac{X_i}{n_0}$ suit également une loi gaussienne que nous désignerons par X.

Pour la suite, μ_0 peut être considérée comme un résultat d'observation d'une v.a.r $Y=\sum_i \frac{X_i}{n_0}\log(\frac{X_i}{n_0})$. On est donc amené à étudier en premier la loi de X log(X) connaissant la loi de X.

4.4 Approximation de Y par une gaussienne.

X étant une distribution connue, on cherche, pour ce faire, à déterminer la fonction de distribution g de la v.a.r Y=X log(X) à partir de la fonction de distribution f de X.

La démarche classique consiste à évaluer à partir de F, fonction de répartition de X, la fonction de répartition G de Y alors : $G(y)=P(Y<y=\varphi(x))$ avec $\varphi(x)=-x\log(x)$.

La fonction φ n'est pas bijective. Elle est définie, s'agissant de valeurs de probabilité, sur l'intervalle $[0,1]$. Elle est croissante sur $[0,\frac{1}{e}]$ et décroissante sur $[\frac{1}{e},1]$.
La fonction inverse φ^{-1} est déterminée graphiquement à partir de φ par la symétrie axiale par rapport à la droite d'équation y=x. φ^{-1} est bivaluée et elle est composée d'une branche strictement croissante $\varphi_1^{-1}:[0,\frac{1}{e}]\rightarrow[0,\frac{1}{e}]$ et d'une branche strictement décroissante $\varphi_0^{-1}:[0,\frac{1}{e}]\rightarrow[\frac{1}{e},1]$.
Plus précisément, si W_0 et W_{-1} sont les branches définies sur $[-\frac{1}{e},1]$ de la fonction W de Lambert[3], partie réelle, alors $\varphi_0^{-1}(y)=\frac{-y}{W_0(y)}$ et $\varphi_1^{-1}(y)=\frac{-y}{W_1(y)}$. Par suite:
$P(Y<y)=P(X<\varphi_1^{-1}(x))+1-P(X>\varphi_0^{-1}(x))$ ce qui peut s'écrire : $G(y)=F(\frac{-y}{W_{-1}(y)})+1-F(\frac{-y}{W_0(y)})$. La dérivée de W étant $W'(y)=\frac{W(x)}{x(1+W(x))}$, les dérivées de $\frac{-y}{W_{-1}(y)}$ et de $\frac{-y}{W_0(y)}$ sont respectivement $\frac{-1}{1+W_{-1}(y)}$ et $\frac{-1}{1+W_0(y)}$ d'où la fonction de distribution : $g(y)=\frac{-f(\frac{-y}{W_{-1}(y)})}{1+W_{-1}(y)}+\frac{f(\frac{-y}{W_0(y)})}{1+W_0(y)}$ où g et f sont les dérivées respectives de G et F.

La fonction de Lambert est difficile à mettre en œuvre dans un calcul numérique du fait des phénomènes d'oscillation lorsqu'on doit utiliser son développement en série limitée. Nous nous contenterons donc de rechercher l'allure générale de la courbe g(y) afin de l'approximer par une fonction plus simple.

[3] la fonction de Lambert peut être visualisée ici : http://math.asu.edu/~kawski/MAPLE/274/images/Lambert8.gif

Domaine de variation de g

La fonction $g(y)=\frac{-f(\frac{-y}{W_{-1}(y)})}{1+W_{-1}(y)}+\frac{f(\frac{-y}{W_0(y)})}{1+W_0(y)}$ est définie sur [0,1/e] et de domaine de variation [0,1].

Pour $y \to 0$, $\frac{-1}{1+W_{-1}(y)} \to 0$ et $\frac{-1}{1+W_0(y)} \to 1$

Pour $y \to \frac{1}{e}$, $\frac{-1}{1+W_{-1}(y)} \to \infty$ et $\frac{-1}{1W_0(y)} \to \infty$

Si maintenant f est une partie gaussienne définie sur [0,1], f est associée à $W_{-1}(y)$ sur [0,1/e] et à $W_0(y)$ sur [1/e,1], 3 cas de figures se présentent suivant que l'espérance μ et la variance σ de la fonction f conduisent à un recouvrement important de la valeur critique 1/e par la gaussienne définie par f.

- Pour μ $\ll$ 1/e, c'est la composante $\frac{-f(\frac{-y}{W_{-1}(y)})}{1+W_{-1}(y)}$ dans g qui est prédominante. Par suite g peut être approximée par une gaussienne avec une asymétrie (skew négatif) d'autant moins marquée que μ est proche de 0.
- De manière similaire pour μ proche de 1, c'est la composantedans g $\frac{f(\frac{-y}{W_0(y)})}{1+W_0(y)}$ qui est prédominante. Et par suite g peut être approximée par une gaussienne avec une asymétrie (skew positif) d'autant moins marquée que μ est proche de 1.
- Dans le cas d'un recouvrement conséquent de la valeur critique 1/e par la gaussienne, l'allure de la distribution g nécessite une analyse approfondie, autour de 1/e, du comportement joint de $f(\frac{-y}{W_{-1}(y)})$ modulé par $1+W_{-1}(y)$ d'une part, et de $f(\frac{-y}{W_0(y)})$ modulé par $1+W_0(y)$ d'autre part. Ce cas ne sera pas traité ici.

En pratique, nous ne nous intéresserons qu'au premier cas, où μ << 1/e. En effet, la taille d'un vocabulaire type est de 50000 à 300000 mots (sans distinction des catégories grammaticales). Ce qui fait, dans nos estimations de μ à partir d'un comptage fréquentiel, et en effectuant un lissage de Laplace pour prendre en compte la masse absente, que la valeur de μ est très éloignée de 1/e et plutôt proche de 0. La représentation de la distribution g par une gaussienne est dans ce cas justifiée.

Si, maintenant, X_1 et X_2 sont 2 v.a.r de loi f_1 et f_2, alors la loi de X_1+X_2 est le produit de convolution f_1*f_2. Et dans le cas où X_1 et X_2 sont des gaussiennes, X_1+X_2 est aussi une gaussienne. En fonction des calculs estimatifs précédents et dans les conditions de nos expérimentations, nous admettrons que $Y=\sum_i \frac{X_i}{n_0}\log(\frac{X_i}{n_0})$ peut être approximée par une loi gaussienne.

4.5 Mise en œuvre de l'identification de non-compositionalité

Le comptage fréquentiel décrit dans §4.2 permet d'associer à chaque mot, pris individuellement, des caractéristiques de compositionalité à droite (resp. à gauche) via la moyenne et la variance. L'hypothèse d'une distribution gaussienne permet ensuite de définir un intervalle de confiance fixé expérimentalement à 95% (ce qui correspond à une valeur de 1.96 d'écart pour une gaussienne).

Pour tout mot m_1 de moyenne «à droite» $\mu_{d,1}$ et de variance «à droite» $\sigma_{d,1}$, si m_1 est suivi de m_2, de moyenne «à gauche» $\mu_{g,2}$ et de variance «à gauche» $\sigma_{g,2}$, m_1m_2 est non compositionnel:

- si $-\log(p_{dg,12}) < \mu_{d,1} - 1.96x\,\sigma_{d,1}$ où $p_{dg,12}$ est la probabilité d'avoir le mot m_2 qui suit le mot m_1
- ou si $-\log(p_{gd,21}) < \mu_{g,2} - 1.96x\sigma_{g,2}$ où $p_{gd,21}$ est la probabilité d'avoir le mot m_1 qui précède le mot m_2

5 Couplage du modèle linguistique

Le composant linguistique dispose au départ :

- d'un lexique du français comportant 300 000 formes fléchies, décrites par la partie du discours et des traits d'accord
- de règles de grammaires de chunking (Abney, 1994) de type hors contexte, décrites sous forme normale de Chomsky et regroupées par paquets homogènes
- de méta-règles régissant les paquets de règles afin de rendre, autant que possible, l'analyse déterministe.

De plus, la profondeur d'analyse est limitée pour couvrir des syntagmes de moins de 6 mots, ce qui est suffisant dans nos applications. Cette hypothèse permet de traduire les règles initiales en règles de grammaires régulières au sein de chaque paquet de règles.

Une première analyse lexicale du corpus permet de recenser le vocabulaire utilisé et de compléter le référentiel lexical initial par les nouveaux mots simples inconnus. L'ajout de ces mots inconnus dans le référentiel lexical est réalisé seulement après seuillage suivant leur fréquence d'occurrence et leur coefficient de variation.

5.1 Analyse lexicale et syntaxique du corpus

S'agissant de grammaires de chunking, l'absence du non-terminal initial S impose une analyse «bottom-up». Il s'agit donc d'une analyse LR classique (Aho et al., 1977) avec une utilisation particulière du chart parsing.

En effet, plutôt que de créer un espace de chart pour l'analyse de chaque phrase du corpus, on construit successivement des niveaux de chart couvrant tout le corpus et de la manière suivante :

- on dispose d'un référentiel lexical de mots simples et de locutions, et d'un référentiel des syntagmes en cours de construction
- le référentiel des syntagmes est vide au départ (éventuellement celui des locutions aussi)
- le référentiel lexical et le référentiel des syntagmes sont utilisés pour indexer tout le corpus
- le résultat de chaque indexation correspond alors à un niveau du chart
- une analyse du coefficient de variation des syntagmes du référentiel permet d'éliminer les éléments les moins pertinents
- une analyse de la compositionalité des syntagmes figurant dans le référentiel des syntagmes permet d'identifier les locutions et de les reverser dans le référentiel des locutions
- on applique ensuite un nouveau paquet de règles de grammaires pour identifier de nouveaux syntagmes et pour les reverser dans le référentiel des syntagmes

Le processus se termine après épuisement des paquets de règles.

5.2 Mise en œuvre du système

Les résultats qui suivent sont issus du corpus d'actualités décrit précédemment dans §2. Les données, en constante évolution, correspondent aux actualités de janvier 2011. Les listes ci-après correspondent à des extraits de patronymes et de groupes nominaux classés par ordre de pertinence décroissante. Un référentiel terminologique unique est dans un premier temps appris sur le corpus global d'actualités puis «projeté» sur des plus petits corpus thématisés, par analyse du coefficient de variation *intra* corpus.

Patronymes culturel	GN culturel	Patronymes sport	GN sport	Patronymes international	GN international
-nicolas sarkozy -johnny hallyday -frédéric mitterrand -conrad murray -dany boon -ben ali -john barry -brice taton -robert de niro - luc chatel	-golden globes -homicide involontaire -premier ministre -discours d un roi -grand palais -los angeles -première fois -poivre d arvor -sol majeur	-andy murray -caroline wozniacki -paris sg -jean pierre dick -claude onesta -kim clijsters -justine henin -wilfried tsonga -cyril despres -stanislas wawrinka	-autres sports -quarts de finale -championnats étrangers -tête de série -finale de la coupe -championnat du monde -fin de la saison -milieu de terrain	-ben ali -laurent gbagbo -nicolas sarkozy -sidi bouzid -zine ben -saad hariri -vincent delory -jean claude duvalier -silvio berlusconi	-premier ministre -affaires étrangères -service français -ancien président -départ du président -président déchu -président tunisien -conférence de presse -forces de l ordre

-xavier beauvois -beverly hills -marc olivier fogiel -sofia coppola -justin bieber -quentin tarantino -laurent gerra -caroline lachowsky -claude monet -françois fillon -ernest hemingway -alexandre jardin -jean dutourd	-biographie d hemingway -priorité santé -haute couture -bande dessinée -mise en scène -accusé de plagiat -tête de bois -meilleur film -jeu vidéo -nouvelles technologies -télé réalité -pluies diluviennes -premier album -bande dessinée d angoulême	-josé mourinho -michel desjoyeaux -stéphane sessegnon -paris fc -françois gabart -jean tigana -saint etienne -loïck peyron -carlos sainz -dimitri payet -brian joubert -tomas berdych -lionel messi	-ballon d or -journal du mercato -finale du tournoi -champion du monde -français jean -coupe de la ligue -ski alpin -match en retard -nuit des français -rumeurs du mercato -tenant du titre -quart de finale -nuit dernière -première fois -conférence de presse	-benoît xvi -françois fillon -mohamed elbaradei -alain juppé -nelson mandela -gilles trequesser -mohamed ghannouchi -antoine de léocour -eric zemmour -jean stéphane -johan vande -tarek amara -henri pierre -eric faye	-régime du président -ministre des affaires -nouveau gouvernement -journaliste de l afp -droits de l homme -jeunes français -président américain -ministère de l intérieur -démission du gouvernement -président sortant -union européenne -ministre de la défense -communauté internationale -français enlevés

6 Conclusion

L'approche que nous venons de décrire confirme qu'il est possible de concevoir un système d'acquisition de terminologie performant en temps d'exécution et aussi de très bonne qualité. Le taux de précision[4] obtenu est de l'ordre de 90% (Lassalle et al., 2011). Les principales raisons de ces performances sont liées à :

- une architecture de chart parsing couvrant tout le corpus, évitant ainsi des redondances d'analyse des mêmes syntagmes
- le regroupement des syntagmes analysées dans un même référentiel, permettant ainsi un couplage avec l'analyse statistique tout en minimisant le biais
- une spécialisation des analyses statistiques entre la détection des locutions et le classement de ces dernières en fonction du corpus applicatif

7 Annexe :

Extrait de la grammaire de chunking permettant d'identifier les patronymes :

#Cat prenoms.prenoms

- (CatLoc1 prenoms.prenoms) →(CatMot prenoms) (CatMot prenoms)
- (CatLoc1 prenoms.prenoms) →(CatMot particule.prefixe) (CatMot prenoms)

#Cat PRENOMS.particule

- (CatLoc1 PRENOMS.particule) →(CatMot prenoms) (CatMot particule)
- (CatLoc1 PRENOMS.particule) →(CatMot prenoms) (CatLoc1 particule.particule)
- (CatLoc1 PRENOMS.particule) →(CatLoc1 prenoms.prenoms) (CatMot particule)
- (CatLoc1 PRENOMS.particule) →(CatLoc1 prenoms.prenoms) (CatLoc1 particule.particule)

#syntagme PATRO
avec détection de non-compositionalité

- (CatLoc1 PATRO) →(CatLoc1 PRENOMS.particule) (CatMot patronyme) + (SeuilleOr $LOCBIN1)
- (CatLoc1 PATRO) →(CatLoc1 PRENOMS.particule) (CatMot prenoms) + (SeuilleOr $LOCBIN1)
- (CatLoc1 PATRO) →(CatLoc1 PRENOMS.particule) (CatMot v.stat) + (SeuilleOr $LOCBIN1)
- (CatLoc1 PATRO) →(CatLoc1 prenoms.prenoms) (CatMot patronyme) + (SeuilleOr $LOCBIN1)

[4] Le taux de rappel n'est pas pertinent pour un modèle d'apprentissage statistique. En effet, un nombre minimal d'occurrences (environ 4) d'une même locution est nécessaire pour que cette dernière puisse être identifiée, ce qui exclut des locutions dont la fréquence d'apparition est trop faible. Enfin, l'estimation de ce taux nécessite un recensement manuel des locutions dans le corpus de test, ce pour un coût en général prohibitif. Une solution (que nous n'avons pas mise en œuvre) consisterait à échantillonner le corpus pour estimer le nombre moyen de locutions observées tous les n mots analysés et de le comparer avec le nombre total des locutions extraites divisé par la taille (en nombre de mots) du corpus d'apprentissage.

- (CatLoc1 PATRO) →(CatLoc1 prenoms.prenoms) (CatMot v.stat) + (SeuilleOr $LOCBIN1)

Références

ABNEY S.T.,(1994). PARSING BY CHUNKS. *BELL COMMUNICATION RESEARCH*.

AHO A.,SETHI R., ULLMAN J.D.(1977). Compilers: Principles, Techniques, and Tools. *Dragon Book*.

BRILL E.,(1992). A Simple Rule Based Part of Speech Tagger. *ACL*.

CHURCH K., HANKS P.,(1996). WORD ASSOCIATION NORMS, MUTUAL INFORMATION, AND LEXICOGRAPHY. *COMPUTATIONAL LINGUISTICS*. 16, 22-29.

CORLESS ET AL.,(1996). On the Lambert W function. *Adv. Computational Maths*. 5, 329-359.

DAILLE B.,(1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *MIT Press.*, 49-66.

DUNNING T.D.,(1993). Accurate Methods for the Statistics. *Computational Linguistics. 19(1)*, 61-74.

FRANTZI K.T., ANANIADOU S., TSUJII J.,(1998). The C-value/NC-value Method of Automatic Recognition of Multi-word Terms. *ECDL'98*, 585-604.

HEINECKE J., SMITS G., CHARDENON C., GUIMIER DE NEEF E.,MAILLEBUAU E., BOUALEM M., (2008). TILT : plate-forme pour le traitement des langues naturelles. *TAL Vol. 49*.

KIT C.,(2002). Corpus Tools for Retrieving and Deriving Termhood Evidence. *The 5th East Asia Forum of Terminology*, 69-80.

LASSALLE E., CASIMIR P.K., GUIMIER DE NEEF E.,(2011). Evaluation des outils d'extraction terminologique Quezao et Acabit. *EGC 2011*, 131, 136.

NAMER F.,(2000). Flemm : Un analyseur flexionnel de français à base de règles. *Traitement Automatique des Langues pour la Recherche d'Information. Hermes*, 523-547.

NAZARENKO A., ZARGAYOUNA H., HAMON O., VAN PUYMBROUCK J.,(2009). Evaluation des outils terminologiques : enjeux, difficultés et propositions. *TA Vol. 50*, 257-281.

PAPOULIS A.,(2002). Probability, Random Variables ans Stochastic Processes. *Mac Graw Hill*.

SAPORTA G., (2006). Probabilité, analyse des données et statistique. *Ed. Technip*.

SHANNON C.E.,(1948). A Mathematical Theory of Communication. *The Bell System Technical Journal. 27*, 623-656.

SMADJA F.,(1993). XTRACT : An Overview. *Computer and the Humanities Kluwer Academic Publishers*.

TSURUOKA Y.,(2005). Chunk Parsing Revisited. *9th IWPT*.

VU T., AW A.T., ZHANG M.,(2008). Term Extraction Through Unithood And Termhood Unification. *IJNLP*.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Métarecherche pour l'extraction lexicale bilingue à partir de corpus comparables

Amir Hazem[1] Emmanuel Morin[1] Sebastián Peña Saldarriaga[2]
(1) Université de Nantes, LINA - UMR CNRS 6241
2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 03
(2) Synchromedia, École de technologie supérieure
1100 rue Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
amir.hazem@univ-nantes.fr, emmanuel.morin@univ-nantes.fr, spena@synchromedia.ca

Résumé. Nous présentons dans cet article une nouvelle manière d'aborder le problème de l'acquisition automatique de paires de mots en relation de traduction à partir de corpus comparables. Nous décrivons tout d'abord les approches standard et par similarité interlangue traditionnellement dédiées à cette tâche. Nous réinterprétons ensuite la méthode par similarité interlangue et motivons un nouveau modèle pour reformuler cette approche inspirée par les métamoteurs de recherche d'information. Les résultats empiriques que nous obtenons montrent que les performances de notre modèle sont toujours supérieures à celles obtenues avec l'approche par similarité interlangue, mais aussi comme étant compétitives par rapport à l'approche standard.

Abstract. In this article we present a novel way of looking at the problem of automatic acquisition of pairs of translationally equivalent words from comparable corpora. We first describe the standard and extended approaches traditionally dedicated to this task. We then re-interpret the extended method, and motivate a novel model to reformulate this approach inspired by the metasearch engines in information retrieval. The empirical results show that performances of our model are always better than the baseline obtained with the extended approach and also competitive with the standard approach.

Mots-clés : Corpus comparables, lexiques bilingues, métarecherche.

Keywords: Comparable corpora, bilingual lexicon, metasearch.

1 Introduction

L'extraction de lexiques bilingues à partir de corpus comparables est un domaine de recherche en pleine effervescence qui vise notamment à offrir une alternative crédible à l'exploitation de corpus parallèles. En effet, les corpus parallèles sont par nature des ressources rares notamment pour les domaines spécialisés et pour des couples de langues ne faisant pas intervenir l'anglais, là ou les corpus comparables sont par essence des ressources abondantes puisque composés de documents partageant différentes caractéristiques telles que le domaine, le genre, la période, etc. sans être en correspondance de traduction. Les lexiques bilingues extraits à partir de corpus comparables sont néanmoins d'une qualité bien inférieure à ce qui peut être obtenu à partir de corpus parallèles. Cette difficulté à extraire des lexiques bilingues peu bruités à partir de corpus comparables explique pourquoi ce champ de recherche n'a pas encore franchi le cap de l'industrialisation à la différence des corpus parallèles et reste encore majoritairement cantonné à une activité de recherche prometteuse. La principale difficulté des approches liées à l'exploitation de corpus comparables par rapport aux corpus parallèles pour l'extraction de lexiques bilingues, est l'absence d'éléments d'ancrage entre les documents des langues source et cible composant le corpus comparable. Face à cette difficulté les différentes approches liées à l'exploitation de corpus comparables reposent sur la simple observation qu'un mot et sa traduction ont tendance à apparaître dans les mêmes environnements lexicaux. La mise en œuvre de cette observation repose sur l'identification d'*affinités du premier ordre* (*i.e.* identifier les mots qui sont susceptibles d'être trouvés dans le voisinage immédiat d'un mot donné) ou d'*affinités du second ordre* (*i.e.* identifier les mots qui partagent les mêmes environnements lexicaux sans nécessairement apparaître ensemble) (Grefenstette, 1994a, p. 279). Les approches associées à l'identification de ces affinités sont, d'une

part, l'approche standard (Rapp, 1995; Fung & McKeown, 1997) qui est l'approche majoritairement exploitée, et d'autre part, l'approche par similarité interlangue (Déjean & Gaussier, 2002).

Dans cette article, nous reprenons à notre compte l'idée de (Fung, 1998) qui indique que l'extraction de lexiques bilingues à partir de corpus comparables peut être approchée comme un problème de recherche d'information. Dans cette représentation, la requête serait alors les mots à traduire et les documents retournés par le moteur de recherche les candidats à la traduction de ce mot. Et de la même manière que les documents retournés sont ordonnés suivant leur adéquation avec la requête, les traductions candidates sont classées en fonction de leur pertinence par rapport au mot à traduire. Nous souhaitons donc poursuivre plus en avant cette analogie et proposer une amélioration significative à l'approche par similarité interlangue en considérant l'extraction de lexiques bilingues comme un problème de fusion de résultats analogue à celui rencontré par les métamoteurs de recherche d'information. Nous faisons ainsi l'hypothèse que le fait de combiner différentes sources d'information permet de renforcer globalement la méthode par similarité interlangue.

Dans la suite de cet article, nous commençons par rappeler en section 2 les deux méthodes phares en extraction de lexiques bilingues à partir de corpus comparables, à savoir les méthodes dites standard et par similarité interlangue. La section 3 est quant à elle dédiée à la présentation de notre approche par métarecherche qui revisite l'approche par similarité interlangue. La section 4 se concentre sur l'évaluation des trois méthodes mises en œuvre et ouvre une discussion sur les limites de notre approche. Enfin la section 5 vient conclure ce travail.

2 Principales approches en extraction lexicale bilingue à partir de corpus comparables

Dans cette section, nous allons décrire les deux principales approches dédiées à l'extraction de lexiques bilingues à partir de corpus comparables, à savoir : l'*approche standard*, puis l'*approche par similarité interlangue*.

2.1 Approche standard

Les principaux travaux en extraction de lexiques bilingues à partir de corpus comparables sont basés sur une analyse du contexte lexical des mots et reposent sur la simple observation qu'un mot et sa traduction tendent à apparaître dans les mêmes contextes lexicaux. La mise en œuvre de cette observation repose sur l'identification d'*affinités du premier ordre* : « *Les affinités du premier ordre décrivent les mots qui sont susceptibles d'être trouvés dans le voisinage immédiat d'un mot donné.* [1] » (Grefenstette, 1994a, p. 279). Elles peuvent être représentées sous la forme d'un vecteur de contexte, où chaque élément du vecteur représente un mot qui apparaît dans différentes fenêtres contextuelles.

L'implémentation de l'approche standard peut être décrite par les quatre étapes suivantes (Rapp, 1995; Fung & McKeown, 1997) :

1. Identification des contextes lexicaux

Pour chaque partie du corpus comparable, le contexte de chaque mot plein i est extrait en repérant les mots qui apparaissent autour de lui dans une fenêtre contextuelle de n mots. Pour chaque mot i des langues source et cible, un vecteur de contexte $\mathbf{i}$ est ainsi obtenu. À chaque entrée $\mathbf{i}_j$ du vecteur est associée un score de cooccurrence des mots i et j. Habituellement, les mesures d'association comme l'information mutuelle (Fano, 1961), ou le taux de vraisemblance (Dunning, 1993) sont utilisées pour définir les entrées des vecteurs de contextes.

2. Transfert d'un mot à traduire

Les mots d'un vecteur de contexte i, pour lequel une traduction est recherchée, sont ensuite traduits en s'appuyant sur un dictionnaire bilingue. Si le dictionnaire propose plusieurs traductions pour un élément, nous ajoutons au vecteur de contexte de i l'ensemble des traductions proposées (lesquelles sont pondérées par la fréquence de la traduction en langue cible). Les entrées n'ayant pas de traductions dans le dictionnaire bilingue seront quant à elle tout simplement ignorées.

1. *First-order affinities describe what other words are likely to be found in the immediate vicinity of a given word*

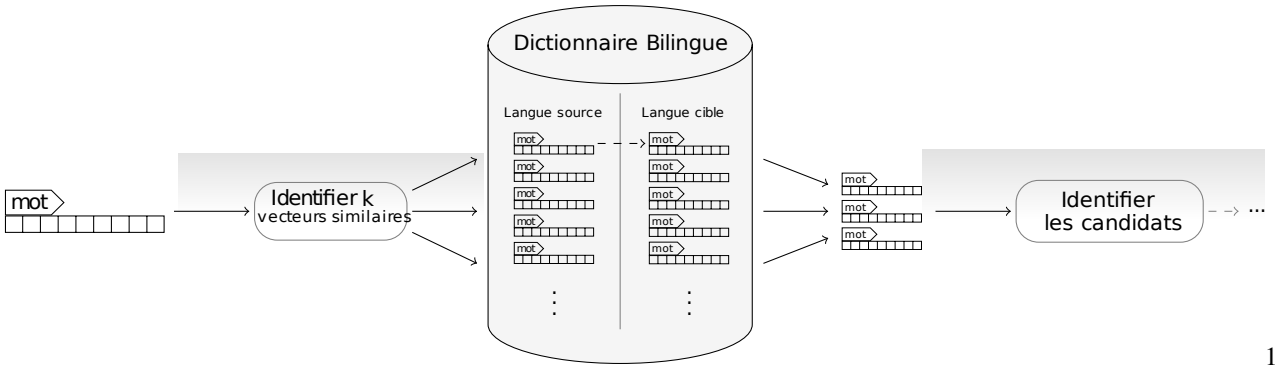
3. Identification des vecteurs proches du mot à traduire

Une mesure de similarité, $\mathrm{sim}(\bar{\mathbf{i}}, \mathbf{t})$, est utilisée pour calculer le score entre chaque mot, t, de la langue cible et le vecteur de contexte traduit du mot $\bar{\mathbf{i}}$. Parmi les mesures de similarité les plus souvent usitées pour cette tâche, nous retrouvons le cosinus (Salton & Lesk, 1968) ou le jaccard pondéré (Grefenstette, 1994b).

4. Obtention des traductions candidates

Les candidats à la traduction d'un mot i à traduire sont finalement ordonnés par ordre décroissant suivant leur score de similarité.

Deux remarques s'imposent ici en ce qui concerne cette approche standard. D'une part, cette approche met en œuvre différents paramètres (taille de la fenêtre contextuelle, choix des mesures d'association et de similarité...) dont il est parfois peu aisé d'identifier les valeurs adéquates pour une recherche optimum (voir par exemple le travail de (Laroche & Langlais, 2010) pour l'influence de ces paramètres sur la qualité des résultats). D'autre part, l'approche standard qui repose originellement sur des cooccurrences graphiques peut aussi être implémentée avec des cooccurrences syntaxiques (Yu & Tsujii, 2009; Otero, 2007).



1

FIGURE 1 – Illustration de la méthode par similarité interlangue.

2.2 Approche par similarité interlangue

Le principal inconvénient de l'approche standard est que ses performances dépendent grandement de la couverture du dictionnaire bilingue par rapport au corpus comparable. En effet, en traduisant un maximum d'entrées du vecteur de contexte du mot à traduire, on maximise les chances de retrouver sa traduction. Bien que la couverture du dictionnaire puisse être étendue en utilisant des dictionnaires spécialisés ou des thésaurus multilingues (Chiao & Zweigenbaum, 2003; Déjean *et al.*, 2002), la traduction des éléments du vecteur de contexte reste le cœur de cette approche.

Dans le but d'être moins dépendants de ce dictionnaire, (Déjean & Gaussier, 2002) ont proposé une extension de l'approche standard connue sous le nom d'approche par similarité interlangue. Cette approche se base sur l'idée que les mots ayant le même sens, partagent les mêmes environnements lexicaux. Elle repose sur l'identification d'*affinités du second ordre* : « *Les affinités du second ordre dévoilent quels mots partagent les mêmes environnements. Les mots partageant des affinités du second ordre n'ont pas besoin d'apparaître ensemble, mais leurs environnements sont semblables.* [2] »

Dans cette approche, le dictionnaire bilingue établit un pont entre les langues du corpus comparable. L'approche par similarité interlangue est basée sur ce principe et évite les traductions directes des éléments des vecteurs de contextes comme le montre la figure 1. L'implémentation de cette approche peut être réalisée en quatre étapes où la première et la dernière sont identiques à celles de l'approche standard (Déjean & Gaussier, 2002; Daille & Morin, 2005) :

2. Sélection des k plus proches voisins

Pour chaque mot à traduire, les k plus proches voisins sont identifiés parmi les entrées du dictionnaire selon

2. *Second-order affinities show which words share the same environments. Words sharing second-order affinities need never appear together themselves, but their environments are similar*

$\mathrm{sim}(\mathbf{i}, \mathbf{s})$. Chaque plus proche voisin est ensuite traduit à l'aide du dictionnaire bilingue, et le vecteur de contexte de langue cible $\overline{\mathbf{s}}$ correspondant à la traduction est sélectionné. Si pour un plus proche voisin il existe plusieurs traductions, $\overline{\mathbf{s}}$ est donné par l'union des vecteurs correspondant aux différentes traductions. Il est à noter que les vecteurs de contexte ne sont pas traduits directement, ce qui réduit l'influence du dictionnaire.

3. Identification des vecteurs proches du mot à traduire

La mesure de similarité, $\mathrm{sim}(\overline{\mathbf{s}}, \mathbf{t})$, est utilisée pour calculer le score entre chaque mot t de la langue cible en fonction des k plus proches voisins. Le score final attribué à chaque mot t est donné par :

$$\mathrm{sim}(\mathbf{i}, \mathbf{t}) = \sum_{s \in k\text{PPV}} \mathrm{sim}(\mathbf{i}, \mathbf{s}) \times \mathrm{sim}(\overline{\mathbf{s}}, \mathbf{t}) \quad (1)$$

Une autre manière de calculer le score de similarité a été proposée par (Daille & Morin, 2005). Les auteurs calculent alors le barycentre des vecteurs de contexte des k plus proches voisins.

3 Extraction lexicale bilingue par métarecherche

3.1 Motivations

L'approche proposée par (Déjean & Gaussier, 2002) introduit implicitement le problème du choix de la valeur adéquate de k dans les k plus proches voisins. D'une manière générale, la valeur optimale de k dépend des données mises en jeu. Cette valeur est souvent définie de façon empirique, bien qu'il soit possible de la déterminer par validation croisée. L'approche par similarité interlangue (ASI) appliquée à nos données s'est révélée très sensible vis-à-vis du paramètre k. En effet, pour des valeurs de k supérieures à 20, la précision chute de façon significative. De plus, il n'est pas possible de déterminer des intervalles de stabilité relative pour k. Le choix du paramètre k devient alors crucial.

En partant du principe que chaque mot contribue à la caractérisation du mot à traduire, notre proposition vise non seulement à améliorer la précision, mais aussi à être plus robuste vis-à-vis du nombre de plus proches voisins. En poussant l'analogie des approches inspirées de la RI (Fung & Lo, 1998) plus loin, nous proposons une nouvelle façon d'aborder le problème de l'extraction lexicale bilingue à partir de corpus comparables, en le considérant comme un problème de fusion de résultats analogue à celui rencontré par les métamoteurs de recherche.

L'objectif de la métarecherche est de fusionner les classements renvoyés par plusieurs systèmes de RI, en une liste unique, afin d'obtenir un système combiné qui soit plus performant que les systèmes individuels (Aslam & Montague, 2001). Puisque chacun des k plus proches voisins produit un classement différent, la métarecherche fourni un cadre adéquat pour exploiter l'information véhiculée par chacun des k classements. En outre, un intérêt particulier est donné aux mots candidats à la traduction d'un mot donné. En effet, partant du principe que les corpus contiennent beaucoup de bruit, il n'est pas rare de rencontrer des mots qui soient proches d'un nombre important de mots du dictionnaire, et ainsi, viennent parasiter le modèle et fausser les résultats. En effet, pour traduire un mot, le système choisit un nombre k de plus proches voisins en langue source, puis il cherche en langue cible les candidats les plus proches des traductions de ces k plus proches voisins sans tenir compte de la relation de ces candidats avec le reste des mots du dictionnaire. Pour pallier cela, nous construisons un modèle qui prend en compte cette information en accordant plus de confiance aux candidats qui sont plus proches des k plus proches voisins que du reste des entrées du dictionnaire.

3.2 Approche par métarecherche

Cette section décrit notre extension de l'approche par similarité interlangue. Les différents modes de fusion définis ici se basent sur les éléments décrits dans la table 1.

La première étape de notre méthode consiste à fixer le nombre de plus proches voisins d'un mot à traduire. La valeur de k est déterminée empiriquement. Cependant, intuitivement mais aussi à travers nos expériences, nous pouvons dire que la sélection d'un nombre faible de plus proches voisins est insuffisante dans la plus part des cas

Symbole	Définition
i	Le mot à traduire
t	Le mot candidat à la traduction de i
s	L'ensemble des plus proches voisins de i
$\overline{s}$	L'ensemble des traductions des plus proches voisins de i
k	Le nombre de plus proches voisins sélectionnés
n	L'ensemble de tous les voisins de t
u	Le nombre total de mots du dictionnaire
$occ_{\overline{s}(t)}$	L'effectif de t ie : avec combien de voisin t est-il en relation ?
$\mathrm{sim}(\overline{s_k}, t)$	Le score de similarité entre le k ième proche voisin de $\overline{s}$ et t
$\max_{\overline{s_k}}$	Le score maximum du candidat le plus proche de $\overline{s_k}$
$\max_{\overline{s}}$	Le score maximum du candidat le plus proche de l'ensemble $\overline{s}$
$\mathrm{sim}_k(s, t)$	Le score de similarité entre s et t par rapport au k ième plus proche voisin
$\mathrm{sim}(s, t)$	Le score de similarité entre s et t par rapport à l'ensemble des plus proches voisins
θ_t	Le paramètre de régulation ou facteur de confiance de t

TABLE 1 – Éléments de notation.

pour trouver la bonne traduction, et que la sélection d'un grand nombre de voisins, d'une part contredit la notion de plus proches voisins et d'autre part, induit la prise en compte de voisins éloignés qui peuvent fausser le modèle.

Une fois k fixé, nous considérons chaque liste de candidats renvoyée par un proche voisin indépendamment des autres. Ces candidats sont les mots dont les vecteurs de contexte sont les plus similaires au vecteur de contexte d'un voisin donnée. Dans nos expériences, la taille des listes a été fixée à 200. Partant du même principe que le choix du paramètre k. La taille de la liste joue un rôle important, en effet, une liste trop petite de candidats ne serait pas suffisante pour aider à trouver la bonne traduction, de la même façon, une liste trop importante de mots risque de rajouter du bruit car il faut garder en tête que les mots appartenant à une liste sont les traductions potentielles classées par ordre de score de similarité. Notre modèle privilégie les candidats qui apparaissent dans plusieurs listes, ainsi, plus l'effectif du mot candidat est important plus il a de chances d'être la bonne traduction, ceci dit, ceci reste valable si le candidat est bien classé, en revanche, s'il apparaît souvent mais en étant toujours mal classé par rapport aux différentes listes, ce mot a de fortes chances d'être une mauvaise traduction.

Dans l'approche par similarité interlangue, le calcul du score de similarité se fait sans prendre en considération les plus proches voisins d'une manière indépendante en amont, ainsi la fusion des scores est faite de telle sorte à ce que les candidats qui ont un score élevé par rapport à un proche voisin soit privilégiés par rapport à des candidats qui ont un score moins élevé mais qui apparaissent dans plusieurs listes. Notre approche vise à prendre en compte ce phénomène en normalisant les listes des plus proches voisins de la manière suivante :

$$\mathrm{sim}_k(i, t) = (\mathrm{sim}(i, s_k) \times \mathrm{sim}(\overline{s_k}, t)) \times \frac{\max_{\overline{s_k}}}{\max_{\overline{s}}} \tag{2}$$

Le raisonnement qui conduit à ce calcul est le suivant. Les scores des différents classements sont sur la même échelle car donnés par la même mesure de similarité. Ainsi, si $\max(l) \gg \max(m)$, cela veut dire que, selon le système, le classement l est plus « sûr » que le classement m (indépendamment de la réponse réelle).

Nous considérons ici que les classements, donnés par les k plus proches voisins du mot à traduire, sont le résultat de k moteurs de RI différents. À l'instar des métamoteurs de recherche, nous allons tenter de nous servir des scores des mots pour améliorer l'extraction bilingue.

Une des approches majeures en métarecherche est le modèle de fusion linéaire (LC) (Bartell *et al.*, 1994), où le score final d'un terme, i, est la somme pondérée de chacun des scores obtenus :

$$\mathrm{sim}(i, t) = \theta_t \times \frac{\sum_{j=1}^{k} \mathrm{sim}_j(i, t)}{\sum_{j=1}^{n} \mathrm{sim}(\overline{s_j}, t)} \tag{3}$$

Pour réduire l'influence des candidats à la traductions qui apparaissent dans différents contextes lexicaux et qui peuvent par leur forte fréquence d'apparition induire en erreur les systèmes d'extraction lexicale basés sur les

contextes, on se propose de prendre en compte ce phénomène en considérant en plus du score calculé à partir des k plus proches voisins, un score définit à partir de toutes les entrées du dictionnaire pour lequel le terme candidat est lié. L'équation 3 permet de calculer le score de similarité entre i et t en prenant en considération le score de similarité par rapport aux k plus proche voisins choisis, normalisé par la somme des scores de t par rapport à tout ses voisins pondéré par le paramètre de confiance θ. Le poids θ est donné par :

$$\theta_t = occ_{\overline{s}}(t) \times \frac{(u - (k - occ_{\overline{s}}(t)))}{(u - occ_n(t))} \quad (4)$$

L'équation 4 prend en compte l'effectif du candidat par rapport aux k plus proches voisins, c'est-à-dire, le nombre de voisins avec lesquels le mot t est en relation. Ceci est représenté par $occ_{\overline{s}}(t)$. Nous privilégions ainsi les mots avec un effectif élevé. Le numérateur $(u - (k - occ_{\overline{s}}(t)))$ permet de considérer l'effectif de t dans l'ensemble $\overline{s}$ par rapport à tous les mots du dictionnaire. On normalisera ensuite par la distribution de t par rapport à tous ses voisins à l'aide de $u - occ_n(t)$. Le paramètre θ permet donc d'accorder plus de confiance à un mot candidat à la traduction qui a un effectif élevé par rapports aux k plus proches voisins mais qui a aussi un effectif faible par rapport au reste de ses voisins.

4 Expériences et résultats

4.1 Ressources linguistiques

Dans le cadre de cette étude, nous avons construit un corpus comparable spécialisé français-anglais à partir de documents extraits du portail Elsevier [3]. L'ensemble des documents collectés relève du domaine médical restreint à la thématique du « cancer du sein ». Nous avons utilisé l'interface de recherche du portail pour sélectionner les publications scientifiques comportant dans le titre ou les mots clés le terme *cancer du sein* en français et *breast cancer* en anglais pour la période de 2001 à 2008. Les documents ont été nettoyés et normalisés à travers les traitements suivants : segmentation en occurrences de formes, étiquetage morpho-syntaxique et lemmatisation. Enfin, les mots agrammaticaux ont été supprimés et les mots apparaissant moins de deux fois dans la partie française et dans chaque partie anglaise écartés. Nous avons ainsi construit un corpus comparable spécialisé d'environ 1 million de mots qui est composé de 130 documents pour le français (7 376 mots distincts) et 103 documents pour l'anglais (8 457 mots distincts).

Le dictionnaire français-anglais nécessaire aux différentes approches comporte, après normalisation, 22 300 mots pour le français avec en moyenne 1,6 traductions par entrée. Il s'agit d'un dictionnaire de langue générale qui ne contient que peu de termes en rapport avec le domaine médical.

Pour évaluer les différentes approches utilisées dans cet article, nous avons sélectionné 400 couples de mots simples français-anglais à partir du meta-thesaurus UMLS [4] et du *Grand dictionnaire terminologique* [5]. Nous n'avons ensuite retenu que les couples pour lesquels le mot français apparaît au moins cinq fois dans la partie française et sa traduction au moins cinq fois dans la partie anglaise. Au terme de ce processus de sélection, nous disposons d'une liste de référence composée de 122 couples de termes simples français-anglais. Cette méthode de création d'une liste de référence est différente de celle proposée par (Déjean & Gaussier, 2002) qui construit sa liste à partir d'un sous ensemble du dictionnaire bilingue. Nous pensons que cette approche, plus fiable d'un point de vue statistique, ne correspond pas aux véritables difficultés rencontrées avec des corpus spécialisés. En effet, en domaine spécialisé les termes qui représentent une difficulté de traduction n'appartiennent par essence que rarement au dictionnaire de langue générale. En ce sens, nous préférons construire notre liste de référence en nous appuyant sur des nomenclatures attestées de termes du domaine non présent dans notre dictionnaire bilingue.

3. www.elsevier.com
4. www.nlm.nih.gov/research/umls
5. www.granddictionnaire.com

4.2 Paramètres expérimentaux

Trois paramètres communs à toutes les approches sont à fixer : i) la mesure d'association, ii) la mesure de similarité et iii) la taille de la fenêtre utilisée pour construire les vecteurs de contexte. Comme mesure de similarité nous avons choisit le jaccard pondéré (Grefenstette, 1994b) :

$$\mathrm{sim}(\mathbf{i},\mathbf{j}) = \frac{\sum_t \min(\mathbf{i}_t, \mathbf{j}_t)}{\sum_t \max(\mathbf{i}_t, \mathbf{j}_t)} \tag{5}$$

Les entrées du vecteur de contexte ont été déterminées par la mesure d'association du taux de vraisemblance (Dunning, 1993). La fenêtre contextuelle a été fixée à 7, partant de l'idée qu'elle approxime les dépendances syntaxiques. En plus de ces paramètres, notre approche ainsi que l'approche par similarité interlangue, ont besoin de définir le nombre de plus proche voisins.

Nous ne détaillons pas plus le choix de ces paramètres et renvoyons le lecteur vers (Morin, 2009) qui motive pour les mêmes ressources le choix de ces paramètres.

4.3 Résultats

Pour évaluer les performances de notre approche, nous utilisons comme référence l'approche par similarité interlangue (ASI) proposée par (Déjean & Gaussier, 2002). Nous comparons l'ASI avec les deux stratégies de l'approche par métarecherche définies dans la section 3 : i) le modèle qui se base sur les scores de similarité (AMS) sans tenir compte de la fiabilité des candidats ; ii) le modèle des sources multiples (AMF) qui prend en compte cette information. Nous allons étudier la stabilité des différentes stratégies de la méthode métarecherche en fonction de la variation des plus proches voisins.

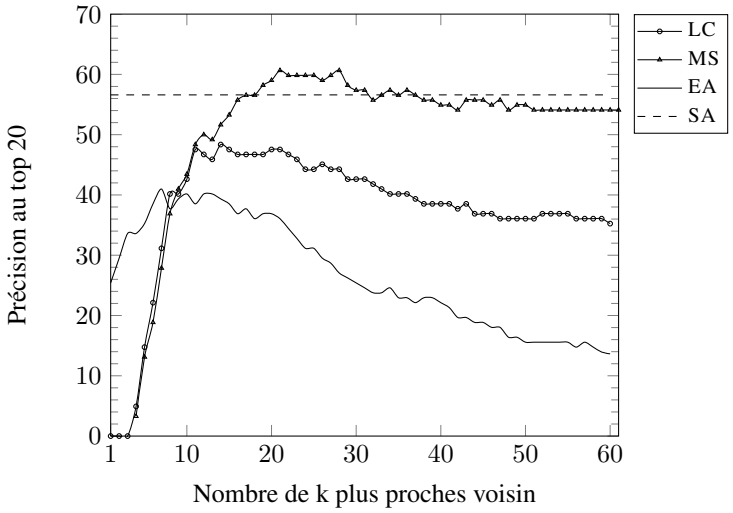


FIGURE 2 – Précision au top 20 en fonction du nombre de ppv.

La figure 2 montre la précision au top 20 en fonction de k. L'approche par similarité interlangue (ASI) atteint sa meilleure performance pour un $k = 7$ avec une précision de $40,98\%$, cette précision commence à décroitre d'une manière significative à partir de $k = 20$.

L'approche par métarecherche qui ne prend en considération que les scores de similarité (AMS) sans considérer la fiabilité des termes candidats à la traduction, montre de meilleurs résultats que la méthode de référence (ASI) à partir de $k = 10$ et obtient une précision maximale de $48,36\%$ pour un $k = 14$. On remarque aussi que la courbe correspondant au modèle AMS reste au-dessus de la méthode ASI malgré l'augmentation du paramètre k. La courbe correspondant au modèle de l'approche par métarecherche qui prend en compte la fiabilité des candidats

(AMF) est toujours au-dessus des autres à partir de $k = 10$. L'approche AMF améliore considérablement la précision et atteint sa meilleure performance avec $60,65\%$ pour un $k = 21$. Nous estimons que pour avoir une bonne exploitation des informations fournies par les différents k plus proches voisins en termes de score de similarité, notre système a besoin d'un minimum de k qui de par nos expériences semble être $k = 10$, ce qui explique les faibles résultats pour un $k < 10$. La raison des faibles résultats est simplement que notre système se base sur l'effectif des candidats à la traduction par rapport au paramètre k, en d'autres termes, de combien de proches voisins un candidat est-il proche ? il est évident qu'avec un k faible la notion d'effectif n'a pas assez de poids. Nous considérons aussi, que les candidats à la traduction proches d'un nombre très petit de voisins comme étant peu fiables. Ces candidats sont donc ignorés.

Nous pouvons noter à travers la figure 2 que les modèles AMF et AMS sont toujours meilleurs que la méthode de référence (ASI) (à partir de $k = 10$). De plus, ces modèles offrent une meilleure stabilité quant à la variation des k plus proches voisins. Quoique la précision décroisse en augmentant les valeurs de k, ceci se fait de manière moins rapide que l'approche de référence (ASI).

Nous comparons aussi nos résultats avec ceux obtenus par l'approche standard (AS). Celle ci est représentée dans la figure 2 par une droite car elle ne dépend pas du paramètre k. L'approche standard (AS) obtient une précision de $56,55\%$. Bien qu'elle soit au-dessus de l'approche par similarité interlangue (ASI) ainsi que du modèle AMS de l'approche par métarecherche, elle est en dessous de notre modèle AMF pour des valeurs de k entre 20 et 35. Nous pouvons ainsi considérer l'approche par métarecherche comme supérieure à l'approche de référence (ASI) mais aussi comme étant compétitive par rapport à l'approche standard (AS).

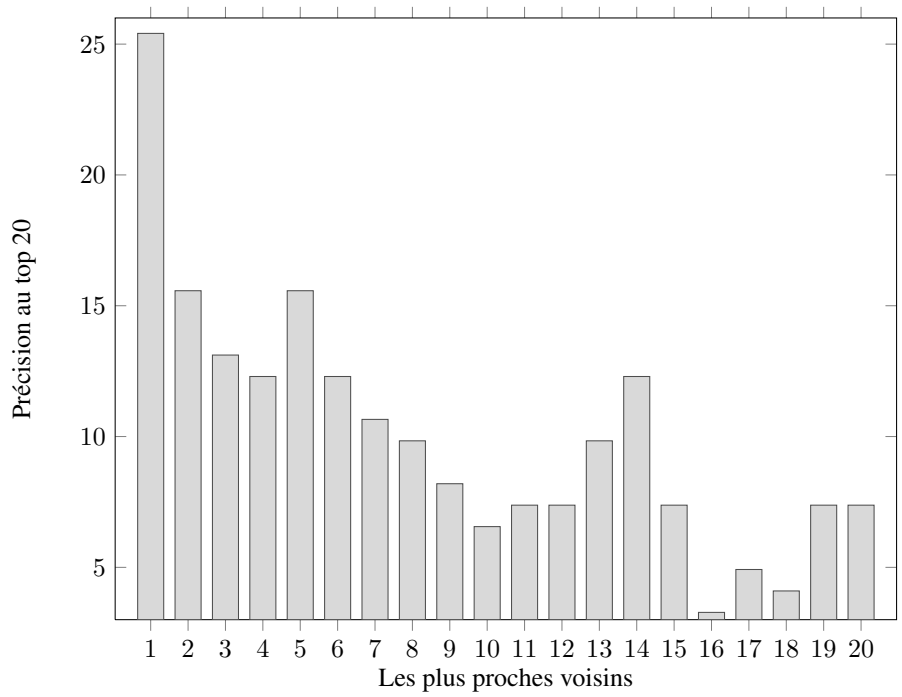


FIGURE 3 – Précision au top 20 pour chacun des 20-plus proches voisins . La précision est calculée en considérant les k plus proches voisins indépendamment les uns des autres.

La figure 3 montre la contribution de chaque plus proche voisin indépendamment des autres. Ceci confirme l'intuition que chaque proche voisin contribue à la caractérisation du mot à traduire, et confirme notre intuition sur le fait de les considérer indépendamment les uns des autres a priori et ceci en dressant pour chacun d'eux une liste de candidats, pour ensuite les combiner et ainsi améliorer les performances.

Il est à noter que les plus proches voisins sont ordonnés du plus proche voisin du mot à traduire au plus éloigné. Bien que chaque plus proche voisin ne puisse traduire qu'un nombre assez faible de mots, en utilisant l'idée de l'approche par métarecherche, nous pouvons améliorer les performances en termes de précision. Ainsi l'idée du

paradigme de notre méthode (AMF) est de prendre en compte l'information véhiculée par tous les plus proches voisins ainsi que le degré de fiabilité des candidats pour améliorer les performance du processus d'extraction lexicale.

Approches	Top 5	Top 10	Top 15	Top 20
AS	37,70%	45,08%	52,45%	56,55%
ASI	21,31%	31,14%	36,88%	40,98%
AMF	**40,98%**	**54,09%**	**56,55%**	**60,65%**

TABLE 2 – Précision aux tops 5, 10, 15, 20 des méthodes AS, ASI et AMF

Enfin comme dernier résultat, nous présentons dans le tableau 3 une comparaison des approches standard (AS), par similarité interlangue (ASI) et par méta-recherche (AMF), pour le top 5, 10, 15 et 20 et ceci en choisissant la meilleure configuration des paramètres de chaque approche. Nous constatons que notre approche AMF obtient une meilleure précision dans chaque situation. Partant du principe que les systèmes d'extraction lexicale tentent d'approcher le top 10 voir le top 5, nous considèreront nos résultats comme étant encourageants notamment pour le top 10 où AMF atteint $54,09\%$ ce qui n'est pas loin de l'approche standard au top 20.

4.4 Discussion

Les approches par similarité interlangue (ASI) et par métarecherche se basent sur les k plus proches voisins pour identifier les meilleurs candidats à la traduction. L'approche ASI effectue une fusion en amont, privilégiant ainsi une vue globale des k plus proches voisins. Ceci peut se révéler problématique, car une bonne traduction pourrait être noyée dans la masse, et ainsi être écartée de la liste des candidats, si des mots plus fréquents viennent à apparaître dans le contexte du mot à traduire. Plus précisément, si des mots obtiennent des scores de similarité très élevés par rapport à un seul plus proche voisin et que d'autres obtiennent des scores moins élevés mais proches de plusieurs plus proches voisins du mot à traduire, ces mots seront moins bien classés voir mal classés. Pour pallier ce problème, l'approche AMF considère dans un premier temps, chaque plus proche voisin comme étant une source d'information indépendante des autres, privilégiant ainsi sa liste de candidats en fixant une taille arbitraire (généralement autour de 200 dans nos expériences), pour ensuite effectuer une fusion en aval des plus proches voisins après avoir normalisé les scores de similarité comme décrit en section 3. En outre, l'approche AMF introduit une mesure de fiabilité, en considérant les plus proches voisins des mots candidats à la traduction comme étant proches des k plus proches voisins du mot à traduire, ainsi que tous les voisins de ces candidats, pour éloigner des mots qui apparaitraient trop fréquemment et dans trop de contextes. Car ne l'oublions pas, ces approches se basent uniquement sur une représentation graphique des données qui induit un certain volume de bruit, lequel serait sans doute mieux traité par une analyse plus fine du contexte. Il est évident que plus les mots sont fréquents dans le corpus plus on a une représentation riche de leur contexte. Cette remarque nous amène à nous interroger sur les fréquences des k plus proches voisins du mot candidat à la traduction. En effet, si un plus proche voisin apparaît fréquemment en langue source et que sa traduction en langue cible est faible ou inversement, quel serait l'impact de ce déséquilibre sur les résultats ? Aucune étude à notre connaissance n'a approfondi ce sujet. Quoique rien ne nous permette d'affirmer une quelconque relation entre ce déséquilibre et une éventuelle traduction erronée, nous pouvons néanmoins supposer que cela est nuisible à une représentation riche du contexte du mot, car un mot peu fréquent apporte moins d'information qu'un mot très fréquent et ceci toujours en se basant sur l'idée de la coloration graphique qui caractérise le contexte. Ainsi nous nous attellerons dans nos travaux futurs à étudier cette problématique. Enfin, les plus proches voisins ont été fixés d'une manière empirique dans les deux approches ASI et AMF, et dans toutes les évaluations. Nous avons fixé un même k pour tous les mots de la liste d'évaluation. L'état de l'art ne spécifie aucune manière efficace de choisir ce paramètre k. Néanmoins, nous sommes en droit de nous interroger pour savoir s'il existe un nombre k idéal de plus proches voisins qui puisse garantir une bonne traduction de tous les mots de la liste d'évaluation ? On serait plutôt tenté de dire qu'il existe un k pour chaque mot à traduire mais que celui-ci varie selon les mots. Là encore, nos travaux futurs devront répondre à cette question clé.

5 Conclusion

Nous avons présenté dans cet article une nouvelle manière d'aborder le problème de l'extraction lexicale bilingue à partir de corpus comparables en nous appuyant sur le principe des métamoteurs de recherche. Nous avons ainsi présenté une nouvelle approche simple et robuste qui revisite la méthode par similarité interlangue pour présenter un modèle inspiré par les métamoteurs de recherche d'information. Ce modèle qui prend en compte la distribution des candidats à la traduction non seulement par rapport au k plus proches voisins du mot à traduire mais aussi par rapport à tout leurs voisins, a permis un gain significatif en terme de précision. Les résultats empiriques que nous obtenons montrent que les performances de ce nouveau modèle sont toujours supérieures à celles obtenues avec l'approche par similarité interlangue pour $k > 10$, mais aussi comme étant compétitives par rapport à l'approche standard.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet METRICC (`www.metricc.com`) a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-CORD-009.

Références

ASLAM J. A. & MONTAGUE M. (2001). Models for Metasearch. In *SIGIR '01, proceedings of the 24th Annual SIGIR Conference*, p. 276–284.

BARTELL B. T., COTTRELL G. W. & BELEW R. K. (1994). Automatic combination of multiple ranked retrieval systems. In *SIGIR '94, proceedings of the 17th Annual SIGIR Conference*, p. 173–181.

CHIAO Y.-C. & ZWEIGENBAUM P. (2003). The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations. In R. BAUD, M. FIESCHI, P. LE BEUX & P. RUCH, Eds., *The New Navigators : from Professionals to Patients, Actes Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, p. 397–402, Amsterdam : IOS Press.

DAILLE B. & MORIN E. (2005). French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, p. 707–718, Jeju Island, Korea.

DÉJEAN H. & GAUSSIER E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, p. 1–22.

DÉJEAN H., SADAT F. & GAUSSIER E. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, p. 218–224, Tapei, Taiwan.

DUNNING T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, **19**(1), 61–74.

FANO R. M. (1961). *Transmission of Information : A Statistical Theory of Communications*. Cambridge, MA, USA : MIT Press.

FUNG P. (1998). A Statistical View on Bilingual Lexicon Extraction : From ParallelCorpora to Non-parallel Corpora. In D. FARWELL, L. GERBER & E. HOVY, Eds., *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, p. 1–16, Langhorne, PA, USA.

FUNG P. & LO Y. Y. (1998). An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics (COLING'98)*, p. 414–420.

FUNG P. & MCKEOWN K. (1997). Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97)*, p. 192–202, Hong Kong.

GREFENSTETTE G. (1994a). Corpus-Derived First, Second and Third-Order Word Affinities. In *Proceedings of the 6th Congress of the European Association for Lexicography (EURALEX'94)*, p. 279–290, Amsterdam, The Netherlands.

GREFENSTETTE G. (1994b). *Explorations in Automatic Thesaurus Discovery*. Boston, MA, USA : Kluwer Academic Publisher.

LAROCHE A. & LANGLAIS P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, p. 617–625, Stroudsburg, Pekin, Chine : Association for Computational Linguistics.

MORIN E. (2009). Apport d'un corpus comparable déséquilibré à l'extraction de lexiques bilingues. In *Actes de la 16ème Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN) Senlis France.*, p. 101–110.

OTERO P. G. (2007). Learning bilingual lexicons from comparable english and spanish corpora. In *Proceedings of Machine Translation Summit XI*, p. 191–198.

RAPP R. (1995). Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, p. 320–322, Boston, MA, USA.

SALTON G. & LESK M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, **15**(1), 8–36.

YU K. & TSUJII J. (2009). Bilingual dictionary extraction from wikipedia. In *Proceedings of Machine Translation Summit XII*.

Évaluation et consolidation d'un réseau lexical via un outil pour retrouver le mot sur le bout de la langue

Alain Joubert (1), Mathieu Lafourcade (1), Didier Schwab (2), Michael Zock (3)

(1) LIRMM, Université Montpellier II (2) LIG, Université Grenoble II (3) LIF-CNRS, Marseille
{alain.joubert, mathieu.lafourcade}@lirmm.fr, didier.schwab@imag.fr, michael.zock@lif.univ-mrs.fr

Résumé Depuis septembre 2007, un réseau lexical de grande taille pour le Français est en cours de construction à l'aide de méthodes fondées sur des formes de consensus populaire obtenu via des jeux (projet JeuxDeMots). L'intervention d'experts humains est marginale en ce qu'elle représente moins de 0,5% des relations du réseau et se limite à des corrections, à des ajustements ainsi qu'à la validation des sens de termes. Pour évaluer la qualité de cette ressource construite par des participants de jeu (utilisateurs non experts) nous adoptons une démarche similaire à celle de sa construction, à savoir, la ressource doit être validée sur un vocabulaire de classe ouverte, par des non-experts, de façon stable (persistante dans le temps). Pour ce faire, nous proposons de vérifier si notre ressource est capable de servir de support à la résolution du problème nommé 'Mot sur le Bout de la Langue' (MBL). A l'instar de JeuxdeMots, l'outil développé peut être vu comme un jeu en ligne. Tout comme ce dernier, il permet d'acquérir de nouvelles relations, constituant ainsi un enrichissement de notre réseau lexical.

Abstract Since September 2007, a large scale lexical network for French is under construction through methods based on some kind of popular consensus by means of games (JeuxDeMots project). Human intervention can be considered as marginal. It is limited to corrections, adjustments and validation of the senses of terms, which amounts to less than 0,5 % of the relations in the network. To appreciate the quality of this resource built by non-expert users (players of the game), we use a similar approach to its construction. The resource must be validated by laymen, persistent in time, on open class vocabulary. We suggest to check whether our tool is able to solve the *Tip of the Tongue* (TOT) problem. Just like JeuxDeMots, our tool can be considered as an on-line game. Like the former, it allows the acquisition of new relations, enriching thus the (existing) network.

Mots-clés Réseau lexical, JeuxDeMots, évaluation, outil de MBL, mot sur le bout de la langue
Keywords Lexical network, JeuxDeMots, evaluation, TOT software, tip of the tongue

Introduction

Grâce à un nombre important de participants à des jeux en ligne (notamment JeuxDeMots et PtiClic), nous avons obtenu un réseau lexical de grande taille pour la langue française (actuellement plus de 220000 termes[1], reliés par plus d'un million de relations sémantiques) représentant une connaissance générale commune. La communauté dispose donc d'une ressource lexicale dont nous souhaitons évaluer la qualité. Une évaluation manuelle pose au moins deux problèmes : d'une part, elle peut être biaisée par les compétences de l'évaluateur, et d'autre part, elle nécessite un temps prohibitif dès que l'on souhaite effectuer une évaluation quelque peu conséquente. Nous aurions pu envisager une évaluation automatique par comparaison avec une référence, mais à notre connaissance une telle référence n'existe pas, du moins

[1] Un terme peut être constitué de plusieurs mots (par exemple : *étoile de mer*)

pour la langue française, ayant une couverture et un nombre de types de relation suffisant. L'évaluation manuelle par échantillonnage ne nous semble pas satisfaisante car elle est nécessairement trop réduite, trop ponctuelle et d'une qualité difficile à apprécier. Nous avons donc décidé d'évaluer notre ressource, via un logiciel de détermination du « mot sur le bout de la langue » (MBL), évaluation qui pourrait être faite de façon permanente et avec un grand nombre d'évaluateurs en simple aveugle (ces derniers ne sachant pas qu'ils évaluent). Notre réseau lexical représentant des connaissances générales, un tel logiciel doit s'appliquer à un domaine ouvert, à savoir du vocabulaire tout venant, y compris des termes peu courants. Compte tenu du caractère sémantique du réseau lexical, l'outil de MBL opèrera exclusivement de manière quasi-sémantique, utilisant essentiellement des associations d'idées, des relations ontologiques ou celles de typicalité. Un mode d'accès par la phonétique, voire la notion de rébus, est donc exclu.

Nous commencerons cet article en présentant d'abord la problématique du MBL, pour rappeller ensuite brièvement le processus de constitution de notre réseau lexical, avant de présenter notre outil de MBL, dénommé AKI[2]. Enfin, nous commenterons les résultats obtenus grâce à AKI pour évaluer notre réseau lexical.

1 Problématique

Difficulté de l'évaluation - Nous sommes confrontés au problème d'évaluation de donnée lexicale, où aucun standard de référence n'est disponible et où une l'évaluation manuelle n'est pas envisageable. Dans un premier temps, on serait tenté de répondre aux questions de complétude et de précision (exactitude) :

- notre réseau lexical est-il « complet », à savoir comporte-t-il « tous » les termes et « toutes » les relations ?
- n'y a-t-il pas dans notre réseau des termes ou des relations erronés ?

Bien évidemment, la réponse stricte à ces deux questions est négative, ne serait-ce qu'en raison du caractère évolutif de la langue ; par exemple, le terme *révolution de jasmin* n'est apparu qu'en janvier 2011. Cependant, nous pouvons dégager une question plus réaliste :

- pour chaque terme de notre réseau lexical, l'ensemble des relations qu'il entretient avec d'autres termes suffit-il à le caractériser de façon unique ?

Dans l'affirmative, tout terme est susceptible d'être retrouvé via un ensemble réduit de termes indices. Ceci étant, nous avons créé un outil de "mot sur le bout de la langue" (MBL) pour réaliser cette évaluation.

2 Le problème du 'mot sur le bout de la langue' (MBL)

Le terme "manque de mot" désigne à la fois l'absence de terme dans le dictionnaire mental (Aitchison, 2003) d'un locuteur, ainsi que l'incapacité de pouvoir y accéder à temps. Nous nous intéressons ici uniquement à ce dernier cas. Le manque de mot, est une situation connue par tout producteur de langue notamment à l'oral (discours spontané). Cette défaillance sera qualifiée d'anomie, d'Alzheimer, ou de *mot sur le bout de la langue* (MBL), selon la durée et la fréquence du blocage et selon la nature d'information accessible (sémantique, phonologique) au moment crucial, la production écrite ou orale.

L'expression *« avoir le mot sur le bout de la langue »* (MBL) » ou, son analogue anglais, *« it's on the tip of my tongue »* (TOT), décrivent une forme de blocage très particulier. Un locuteur cherchant à exprimer une idée est conscient de connaître le terme, il sent sa production imminente (le plus gros du travail étant accompli), pourtant, il échoue tout près de la fin. La dernière partie est inaccessible. Tel un éternuement non consommé, la forme sonore reste bloquée, et la traduction du sens en forme linguistique n'aboutit pas. Ce qui caractérise le MBL et ce qui le distingue d'autres formes de manque de mot[3] c'est que le locuteur connaît

[2] http://www.lirmm.fr/jeuxdemots/AKI.php

[3] Comme indiqué, il y a d'autres cas de figure d'échec lexical. L'un, où le locuteur ignore tout simplement le mot recherché (cas fréquent en langue étrangère), et l'autre, où il connaît le terme, mais il n'arrive pas à l'évoquer à temps : c'est le blanc ou le vide total, situation typique pour des mots rares ou très techniques. D'ailleurs, beaucoup de gens utilisent le terme de MBL de manière générique, voulant lui faire endosser tout type de manque de mot. Ceci est impropre, car il peut y avoir différentes raisons pour

le terme, qu'il en est conscient et que le terme recherché est imminent, d'où l'expression, 'sur le bout de la langue' (Brown & McNeill, 1966). Que le locuteur connaisse le terme est démontrable. Soit il le produit spontanément peu de temps après (en général dans la journée), reconnaissant par ailleurs, et souvent avec soulagement, que c'est bien le terme recherché, soit il l'identifie dans une liste, tâche qu'il effectue avec une vitesse et certitude étonnante (taux d'erreurs extrêmement faible).

D'autres particularités du MBL sont le fait que le locuteur, sait énormément de choses concernant le mot-cible bien qu'il soit incapable de le produire : fragments de *sens* ou fonctions pratiques ('cela sert à s'orienter lors d'une navigation en mer'), *informations syntaxiques* (catégorie lexicale : nom/verbe ; genre grammatical : masculin/féminin), *informations morphologiques* (type et origine de l'affixe) ; *informations phonologiques* (contour intonatif, nombre de syllabes, première et dernière syllabe).

Les informations données par les personnes se trouvant dans cet état ont souvent été utilisées par des psychologues comme argument pour construire et justifier un modèle de production lexicale. La plupart des chercheurs s'accordent pour dire qu'il y a deux étapes se succédant avec, ou sans, chevauchement (Levelt et al. 1999; Ferrand, 1998 ; mais voir également Caramazza, 1997). L'une consiste à déterminer le *lemme* (pour un sens donné on choisit une forme lexicale, qui elle est abstraite), l'autre a pour fonction de déterminer la forme concrète (forme morphologique, graphémique ou phonologique), le *lexème*. Si l'enchaînement de ces deux étapes se fait généralement en continu, donc sans interruption, des problèmes peuvent survenir. Ainsi, il se peut que la première étape se déroule correctement mais pas la seconde, auquel cas on aboutit à l'état nommé MBL : l'information sémantique et grammaticale étant disponibles intégralement, mais pas l'information graphique ou phonologique. Bien entendu, on peut aussi imaginer que l'accès sémantique soit déficient, auquel cas il est logiquement impossible d'accéder à la forme phonologique, car, à moins de ne répéter un mot, il est impossible d'avoir accès à sa forme phonologique sans en avoir déterminé le sens.

Etant donné la faiblesse de la trace phonologique on pourrait être tenté à vouloir renforcer celle-ci, et c'est bien ce que certains psychologues ont suggéré (Abrams et al., 2007). Pourtant, ce n'est pas la voie que nous allons emprunter ici et il y a pour cela plusieurs raisons.

L'analyse d'erreurs (Rossi, 2001)[4] et l'étude du phénomène du MBL suggèrent que l'accès lexical se fait par deux voies : par le sens et par la forme (notamment les sons, phonèmes). Ceci dit, l'accès par le sens (boisson-vin) n'exclut nullement l'accès par des termes associés, par exemple, le terme 'vin' activant le terme 'fromage'. Pourtant, le terme 'vin' n'est pas un élément de sens du terme 'fromage'. C'est une co-occurrence, ou si l'on préfère, c'est une association sur l'axe syntagmatique. Outre cette co-occurrence, les deux termes entretiennent une relation sémantique (ou encyclopédique : 'en mangeant du fromage on boit du vin'). Deux autres points méritant être rappelés sont l'aspect relationnel des termes (ils sont du type associatifs : un terme *x* pouvant évoquer un terme *y* avec une probabilité *z*) et leur organisation sous forme de graphe. Ces deux caractéristiques sont capitales, et elles offrent plusieurs avantages :

- le fait que des termes soient liés élargit le champ de recherche : chaque terme source (mot donné en entrée) active un ensemble de termes associés (termes cibles potentiels), ensemble susceptible de contenir le terme recherché;
- le fait que les termes soient organisés sous forme de graphe permet leur accès par différents chemins. Si cette forme de représentation introduit une certaine redondance dans la représentation des données, elle a l'immense mérite de permettre de retrouver le bon chemin, au cas où l'on se serait trompé à un certain embranchement, situation guère possible, ou du moins beaucoup plus compliquée, dans le cas d'arborescences.

causer cette forme de blocage. Produire un mot suppose avoir effectué des traitements à différents niveaux (conceptuels, sémantiques, phonologiques). Or, l'échec (erreur, incomplétude) à n'importe lequel de ces niveaux peut bloquer la machine et produire ce qu'on appelle *manque de mot*. Le terme MBL ne décrit qu'une situation très particulière : le blocage se situe uniquement au niveau phonologique (informations erronées, informations manquantes), les informations venant des autres niveaux étant généralement disponibles dans leur intégralité.

[4] Des erreurs comme 'à ma gauche' au lieu de 'à ma droite' et 'élision' à la place de 'illusion' illustrent ces deux voies d'accès.

En somme, lorsqu'on est en état de MBL on peut essayer de retrouver un terme via d'autres termes phonologiquement proches (accès par la forme, Abrams, 2007; Zock; 2002), mais aussi via des termes ayant un lien sémantique. Nous nous sommes intéressés ici uniquement à cette dernière solution.

2.1 Hypothèses de travail

Il semble difficile de distinguer les deux usages suivants d'une application de MBL : 1) l'utilisation comme un utilitaire afin de retrouver un terme, 2) l'usage ludique de type devinette. Les motivations pour le second usage sont variables, mais en général visent à «mettre en difficulté le système ».

Il semble *a priori* difficile de savoir si les utilisateurs abordent notre outil de façon utilitaire ou ludique. Plutôt qu'effectuer une étude approfondie sur cette question (étude sans doute longue, difficile et coûteuse), nous allons donc considérer ces deux activités comme identiques. Plus précisément, nous partons des deux hypothèses suivantes :

- Hypothèse 1 : les termes recherchés par les utilisateurs de notre jeu MBL sont des termes de fréquence moyenne ou faible (termes de difficulté[5] moyenne ou importante). Les utilisateurs sont vraiment intéressés à trouver le terme recherché.
- Hypothèse 2 : le vocabulaire ciblé est de basse et de moyenne fréquence (termes de difficulté moyenne ou importante). Les joueurs cherchent à vérifier l'efficacité de l'outil.

Etant donné que le vocabulaire qui déclenche le MBL et celui avec lequel les joueurs de MBL jouent sont identiques, cela nous amène à postuler que « *l'évaluation d'un outil de MBL peut se faire grâce à des joueurs* ».

Par ailleurs, une seconde hypothèse de travail consiste à dire que l'éventail de comportements des joueurs est comparable à celui des personnes ayant réellement besoin de retrouver un mot. A savoir, leur motivation consiste à essayer de piéger l'outil soit avec un terme simple et des indices à la marge, soit avec un terme rare, improbable, ou récent, et des indices plus directs. On peut donc raisonnablement conclure qu'une telle évaluation est plus défavorable que celle portant sur des cas réels de MBL et qu'elle caractérise une ligne basse : *l'évaluation d'un outil de MBL via un jeu fournit une valeur plancher de ses performances.*

Ce sont ces deux hypothèses que nous tenterons de vérifier dans la suite de cet article.

3 Constitution du réseau lexical

3.1 JeuxDeMots[6] : construction du réseau

Le principe de base conduisant grâce à un jeu en ligne à la construction progressive du réseau lexical, à partir d'une base de termes préexistante, a déjà été décrit par (Lafourcade et Joubert, 2009). Une partie se déroule entre deux joueurs, en double aveugle et en asynchrone. Pour un même terme cible T et une même consigne C (synonymes, domaines, association libre…), les deux joueurs proposent des termes correspondant, selon eux, à cette consigne C appliquée à ce terme T. Ces propositions sont limitées en nombre, ce qui a pour effet d'augmenter leur pertinence, mais également dans le temps pour favoriser leur caractère spontané. Nous mémorisons alors les réponses communes à ces deux joueurs[7]. Les validations sont

[5] Nous faisons également l'hypothèse que ce que nous appelons la difficulté d'un terme est contra-variante à sa fréquence, la difficulté d'un terme exprimant à la fois la difficulté à trouver des indices s'y rapportant mais également la difficulté à faire émerger ce terme chez un interlocuteur.

[6] http://jeuxdemots.org

[7] La limitation dans le temps de la saisie des propositions des joueurs peut favoriser les fautes d'orthographe, mais, comme nous ne mémorisons que les réponses communes aux deux joueurs d'une même partie, l'expérience montre que ce risque est très limité et que seules subsistent les fautes d'orthographes qui de toutes façons auraient été faites par les joueurs (par exemple : *beau* pour *bot*, en parlant de *pied*).

donc faites par concordance des propositions entre paires de joueurs pour un même couple (C,T). Ce processus de validation rappelle celui utilisé par (von Ahn et Dabbish, 2004) pour l'indexation d'images ou plus récemment par (Lieberman et al., 2007) pour la collecte de « connaissances de bon sens ». À notre connaissance, il n'avait jamais été mis en œuvre dans le domaine de la construction des réseaux lexicaux.

La structure du réseau lexical ainsi obtenu s'appuie sur les notions de nœuds et de relations entre nœuds, selon un modèle initialement présenté par (Collins et Quillian, 1969) et davantage explicité par (Polguère, 2006). Chaque nœud du réseau est constitué d'une unité lexicale (terme, raffinement d'un terme ou segment textuel) liée aux autres termes via des relations des fonctions lexicales, telles que présentées par (Mel'čuk et al., 1995). Les relations obtenues grâce à l'activité des joueurs sont typées et pondérées[8] : elles sont typées par la consigne imposée aux joueurs, elles sont pondérées en fonction du nombre de paires de joueurs qui les ont proposées, comme indiqué dans (Lafourcade et Joubert, 2009). Plusieurs exemples de relations acquises ont été donnés dans (Lafourcade et Joubert, 2009). Au moment du lancement de JeuxDeMots en juillet 2007, le réseau comportait 152 000 termes (non reliés entre eux, c'est-à-dire aucune relation n'existait). Courant mars 2011, à l'issue d'environ 900 000 parties jouées par plus de 2500 joueurs, notre réseau compte 229 000 termes et plus de 1 100 000 relations.

3.2 PtiClic : consolidation du réseau

De manière analogue à JeuxDeMots (JDM), une partie de PtiClic (http://pticlic.org) se déroule, en double aveugle et asynchrone, entre deux joueurs. Un premier joueur se voit proposer un terme cible T, origine de relations, ainsi qu'un nuage de mots provenant de l'ensemble des termes reliés à T dans le réseau lexical produit par JDM. Plusieurs consignes correspondant à des types de relations sont également affichées. Le joueur associe, par cliquer-glisser, des mots du nuage aux consignes auxquelles il pense qu'ils correspondent. Ce même terme T, ainsi que le même nuage de mots et les mêmes consignes, sont également proposés à un deuxième joueur. Selon un principe analogue à celui mis en place pour JDM, seules les propositions communes aux deux joueurs sont prises en compte, renforçant ainsi les relations du réseau lexical.

Contrairement à JDM, PtiClic est un jeu fermé où les utilisateurs ne peuvent pas proposer de nouveaux termes, mais sont contraints de choisir parmi ceux affichés. Ce choix de conception a pour but de réduire le bruit dû aux termes mal orthographiés ou aux confusions de sens. PtiClic réalise donc une consolidation des relations produites par JDM et permet de densifier le réseau lexical. Notons également que PtiClic permet de créer de nouvelles relations entre termes précédemment reliés par au moins une relation d'un autre type (même si ce n'est pas l'objectif premier de ce logiciel). Afin de réduire le silence correspondant aux termes non proposés par les utilisateurs de JDM, (Zampa et Lafourcade, 2009) ont suggéré de générer le nuage de mots à l'aide de la LSA, en utilisant un corpus externe de grand volume (l'expérimentation réalisée utilise un corpus comportant une année du journal « Le Monde ».). Cette solution permet d'augmenter le réseau lexical par ajout de nouvelles relations, en proposant aux joueurs de nouveaux termes cibles sans liens à T dans le réseau.

3.3 Raffinement des termes : enrichissement du réseau

Le processus permettant d'aboutir aux raffinements de termes est décrit dans (Lafourcade et Joubert, 2010). Nous avons fait l'hypothèse que les sens d'usage, plus communément appelés usages, d'un terme correspondent dans le réseau aux différentes cliques auxquelles ce terme appartient. Cette approche est analogue à celle développée par (Ploux et Victorri, 1998) à partir de dictionnaires de synonymes. En calculant la similarité entre les différentes cliques d'un même terme, nous pouvons construire son arbre des usages nommés. La racine de l'arbre regroupe tous les sens de ce terme. Plus on s'éloigne de la racine, c'est-à-dire plus la profondeur des nœuds dans l'arbre est importante, plus on rencontre des distinctions fines d'usages. Les nœuds de profondeur 1 dans cet arbre correspondent généralement aux différents sens de ce terme répertoriés dans les dictionnaires traditionnels. Après un processus de validation par un expert

[8] Une relation peut donc être considérée comme un quadruplet : terme source, terme cible, type et poids de la relation. Entre deux mêmes termes, plusieurs relations de types (et de poids) différents peuvent exister.

lexicographe de ces différents sens, nous les intégrons dans le réseau en tant que nœuds de raffinement du terme considéré ; le réseau est ainsi enrichi de nouveaux nœuds à partir desquels ou vers lesquels les joueurs de JDM peuvent créer des relations. Actuellement, sur les 229 000 termes connus par le réseau, près de 5 000 ont été raffinés.

4 Un algorithme et un outil de MBL

AKI est un outil d'accès lexical accessible sur le Web à partir du portail JeuxDeMots ou directement à http://www.lirmm.fr/jeuxdemots/AKI.php. AKI peut être envisagé comme un jeu : l'utilisateur fait ″deviner″ un terme cible à l'ordinateur (espérant, éventuellement, de manière secrète, de mettre en défaut sa capacité à trouver un terme à partir d'indices). AKI peut également être considéré comme une assistance, pour retrouver un terme qu'on a sur le bout de la langue. L'utilisateur est invité à fournir, un par un, une succession de termes indices qui lui paraissent pertinents pour trouver le terme cible recherché. Ce mécanisme est comparable à celui de certains jeux télévisés. Après chacun de ces termes indices AKI fait une proposition. Si elle correspond au terme recherché, l'utilisateur valide la proposition, sinon il introduit un nouvel indice. Ce dialogue se poursuit jusqu'à ce que l'une des deux situations se produise : AKI trouve le terme cible ou il abandonne et demande à l'utilisateur de fournir la solution.

4.1 AKI : principe et réalisation

L'utilisateur saisit un premier terme indice i_1. En utilisant le réseau lexical, l'algorithme calcule la signature lexicale de i_1 : $S(i_1) = S_1 = t_1, t_2, \ldots$ où les t_i sont triés par activation décroissante. Autrement dit, t_1 est le terme pour lequel la somme des poids des relations le liant à i_1 est la plus élevée. La première proposition d'AKI, p_1, correspond à ce terme : $p_1 = t_1$. Si c'est le terme cible, l'utilisateur le valide et la partie est terminée, sinon, il est retiré de la signature S_1, ainsi que i_1 (qui ne peut pas être le terme cible). Donc, à ce stade, la signature courante est : $S'_1 = S_1 - \{p_1, i_1\}$. L'utilisateur est alors invité à saisir un deuxième terme indice i_2. L'algorithme calcule une deuxième signature lexicale par intersection entre la signature courante et celle de i_2 : $S_2 = (S'_1 \cap S(i_2)) - i_2$.

Le terme proposé p_2 est celui de S_2 dont l'activation est la plus forte. Autrement dit, parmi les termes reliés à la fois à i_1 et à i_2, AKI affiche celui pour lequel la somme des poids des relations le reliant à i_1 et à i_2 est la plus élevée, exception faite des termes déjà proposés par AKI pour cette partie (ainsi que des termes indices !). La signature courante est alors $S'_2 = S_2 - p_2$. D'une façon générale, à l'étape *n*, nous avons :

$$S_n = (S'_{n-1} \cap S(i_n)) - i_n \quad \text{et} \quad S'_n = S_n - p_n$$

où i_n est le n-ième indice fourni par l'utilisateur et p_n la n-ième proposition de AKI. Le nombre de termes constituant la signature diminue donc au fur et à mesure de l'insertion d'indices. Il est fréquent que la signature devienne vide, avant même que le terme cible n'ait été trouvé ; dans ce cas là, AKI ne peut plus proposer de terme. Le processus pourrait s'arrêter là, mais afin d'améliorer le taux de rappel, une ″procédure de rattrapage″ est mise en œuvre : au lieu d'effectuer des intersections de signatures, on utilise leur somme :

$$S_n = (S'_{n-1} + S(i_n)) - i_n \quad \text{et} \quad S'_n = S_n - p_n$$

Cette procédure favorise l'apprentissage en créant des relations entre des termes isolés. Aussi utile soit elle, cette astuce doit néanmoins être utilisée avec précaution. En effet, au-delà de quelques itérations, le nombre de termes constituant la nouvelle signature devient vite prohibitif ; notre expérience tend à montrer qu'il ne faut pas dépasser deux itérations. Au delà, l'algorithme donne des propositions trop éloignées des mots proposés. Le processus se termine lorsque AKI a trouvé le terme cible recherché, ou lorsque la signature lexicale courante devient vide (ce qui est relativement rare, compte tenu de la procédure de rattrapage). À partir de 5 indices (cette limite de 5 étant un paramètre modifiable) l'utilisateur a la possibilité d'abandonner en indiquant à AKI qu'il fait fausse route. En effet, nous avons estimé que si, au bout de 5 indices, AKI n'a pas trouvé le terme recherché, cela signifie probablement que ces indices ne sont pas pertinents. La figure 1 présente quelques exemples de parties.

Les utilisateurs d'AKI ont la possibilité de faire précéder leur indice de mot-clé faisant référence à des fonctions lexico-sémantiques. Actuellement il est possible d'utiliser dix fonctions : hyperonymie, hyponymie, synonymie, antonymie, domaine, matière, lieu (lieu typique où l'on peut trouver ce que l'on cherche), caractéristique, holonymie, méronymie. Elles correspondent toutes à un type de relation existant dans le réseau JeuxDeMots.

Figure 1 : Quelques exemples de parties. Dans les trois premiers cas, AKI a trouvé le terme cible. Dans le quatrième cas, ne pouvant plus faire de proposition, AKI a abandonné et l'utilisateur a saisi la bonne réponse (il s'agissait donc d'une utilisation ludique de AKI).

4.2 Consolidation du réseau

Dans l'hypothèse jeu, quand le terme cible n'a pas été trouvé par AKI, l'utilisateur est invité à le saisir. Il y a alors création dans le réseau lexical de relations typées « AKI » avec un poids très faible (+1 à chaque occurrence). Ces relations sont régulièrement vérifiées et validées (ou non) par un expert lexicographe. En effet, dans la mesure où c'est l'utilisateur lui-même qui choisit le terme cible, ainsi que les termes indices, la sécurité de la pertinence de telles relations peut difficilement être garantie (un joueur peut toujours commettre des erreurs, sans parler d'éventuels joueurs malveillants). Ceci est différent de JeuxDeMots où les relations sont créées par intersections de propositions de joueurs. Ici, il arrive fréquemment qu'un utilisateur joue plusieurs fois le même terme, avec des indices différents mais également avec des indices communs. Nous sommes en train de réfléchir comment sécuriser ces relations typées « AKI ».

5 Evaluation du réseau via AKI

Évaluation informelle - Nous avons mené une évaluation informelle des performances d'AKI à partir du jeu de société *Tabou* inversé. Le principe du jeu de société *Tabou* est de faire deviner un terme à des personnes à l'aide d'indices, en excluant certains termes dits tabous. La version commerciale de ce jeu fournit une collection de 500 fiches comprenant chacune un terme cible et 5 termes tabous. La version inversée de ce jeu consiste à faire deviner le mot cible en énumérant ces termes tabous, l'hypothèse étant que ces termes sont particulièrement évocateurs du terme cible lorsqu'ils apparaissent ensemble.

Ceci étant, nous avons soumis cette collection de 500 fiches à AKI ainsi qu'à trois personnes (à des fins de comparaisons). AKI a retrouvé le terme cible au plus tard au bout des 5 indices dans 494 cas (soit 98,8% de réussite). Les personnes prises comme références, ont globalement trouvé (dans les mêmes conditions) 402 fois (soit 80,4% de réussite). Ce dernier chiffre n'est qu'une indication, vue la faible taille de l'échantillon considéré, de trois participants.

5.1 Protocole

L'évaluation, tout comme l'apprentissage, ne se fait qu'en fonction de ce que les joueurs ont renseigné. Elle se fait donc sur du vocabulaire appartenant à la classe ouverte. Comme déjà mentionné, AKI peut être envisagé comme un jeu ou un outil de MBL. *A priori*, notre logiciel ne sait pas faire la distinction entre les deux usages. En effet, sur une seule partie et si AKI trouve la solution, nous ne pouvons pas savoir a priori si l'utilisateur connaissait ou s'il recherchait le terme cible. Par contre, si dans un laps de temps relativement court (de l'ordre de quelques minutes) un même terme est joué plusieurs fois, nous pouvons faire l'hypothèse qu'il s'agit d'une utilisation de type jeu (au moins à partir de la deuxième partie) où l'utilisateur essaie de faire trouver le terme cible par AKI, en proposant généralement des indices différents. Dans chacun des deux cas, jeu ou outil de MBL, les termes cibles sont majoritairement des termes de fréquence moyenne, voire faible. En effet, jouer pour trouver un mot fréquent ne présente pas un grand intérêt, et généralement on ne recherche pas grâce à un outil de MBL un terme courant. Les graphiques ci-après montrent l'évolution dans le temps du rapport entre le nombre de parties gagnées par AKI, parties où l'utilisateur a indiqué que le logiciel a trouvé le terme cible, et le nombre de parties jouées. Les graphiques de cette section reflètent 6522 parties réalisées entre le 30/12/2010 et le 24/01/2011.

5.2 Analyse quantitative et évolution dans le temps

Le premier graphique (figure 2) présente l'évolution du rapport entre le nombre de parties gagnées et le nombre de parties jouées par fenêtre glissante de 500. Par exemple, à l'abscisse 100 (correspondant à 2000 parties), la courbe correspond à la moyenne des valeurs entre les parties 1501 et 2000. Lorsqu'il y a moins de 500 valeurs, la courbe présente la moyenne des *n* premières valeurs. Ce graphique montre globalement une légère amélioration des résultats au cours du temps, avec un passage de 60% de réussite à 80%.

Nous avons analysé le type de mots joués par les utilisateurs. Nous avons considéré comme **courants** les mots issus de **l'échelle orthographique Dubois Buyse**[9], c'est-à-dire, ceux connus par un enfant de 12 ans. Nous considérions les autres comme normaux. Par exemple, *fourchette*, *pie*, *écureuil*, *restaurant* sont courants, tandis que *séquoia géant*, *Rabat* ou même *Akinator* sont considérés comme des termes normaux. On peut considérer que globalement, les termes normaux ont une fréquence d'utilisation allant de moyen à rare (les mots courants ayant une fréquence d'utilisation élevée). L'analyse des parties jouées, ainsi que le nombre de mots différents nous révèle que dans les deux cas, environ 25% concerne des mots courants : sur 1701 mots différents joués, 435 sont courants (soit 25,6%) et sur 6488 parties, 1565 concernent des mots courants (24,1%). Il est important de noter que les mots sont ceux qui étaient déjà bien complets dans le réseau lexical de JDM. On remarquera que l'addition pondérée par le nombre de parties des deux courbes (figures 3 et 4) donne la courbe de la figure 2.

Sur les parties jouées sur des mots courants, on observe globalement une stagnation des résultats, preuve que le réseau était déjà bien complet. Ce qui n'exclut pas que le réseau ait été enrichi de nouveaux résultats, bien que ceux-ci soient très peu visibles. En revanche, en ce qui concerne les mots normaux (ceux qui ne sont pas considérés comme courants), la progression est bien plus claire. Alors qu'au départ, la moyenne à 1000 était inférieure à 60%, elle atteint 80% à la fin. À quoi est due cette progression ? Une première explication possible serait que les joueurs découvrent AKI ; et ce n'est que petit à petit qu'ils réussissent à proposer des indices pertinents. Ceci est quelque peu contredit par l'expérience : en effet, il semble que les joueurs essaient de plus en plus de proposer des indices indirects afin de «mettre le système en défaut ». Il

[9] L'échelle orthographique Dubois Buyse permet d'indiquer les mots normalement acquis par 75% des enfants d'une classe d'âge. Nous considérons donc comme vocabulaire courant les mots bien orthographiés par 75% des enfants de 12 ans. On peut trouver cette liste, entres autres, à http://o.bacquet.free.fr/db2.htm

nous parait plus plausible que cette progression serait due à la capacité d'apprentissage du système. Cette hypothèse serait bien entendu à vérifier sur un plus long terme mais le nombre de relations acquises lors de cette expérience semble la corroborer.

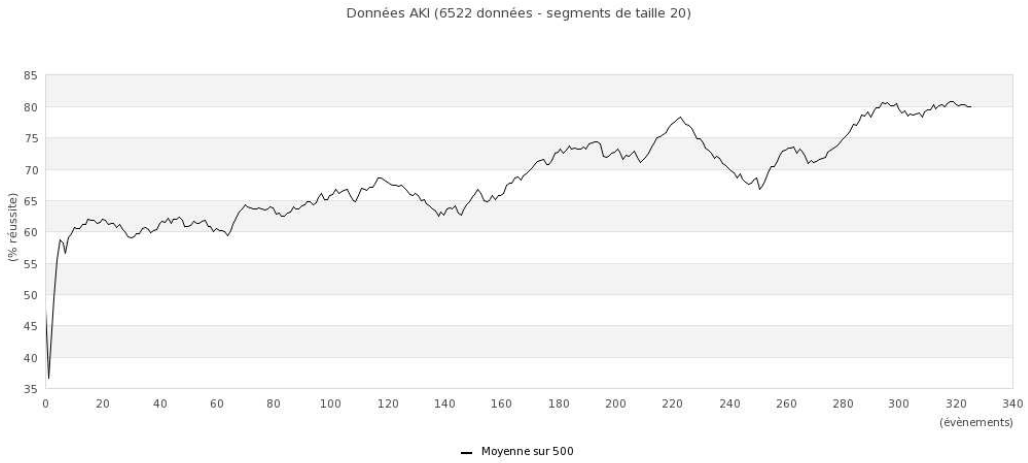


Figure 2 : Graphique montrant l'évolution dans le temps du rapport entre le nombre de parties AKI gagnées et le nombre de parties jouées (moyenne glissante sur les 500 dernières parties jouées)

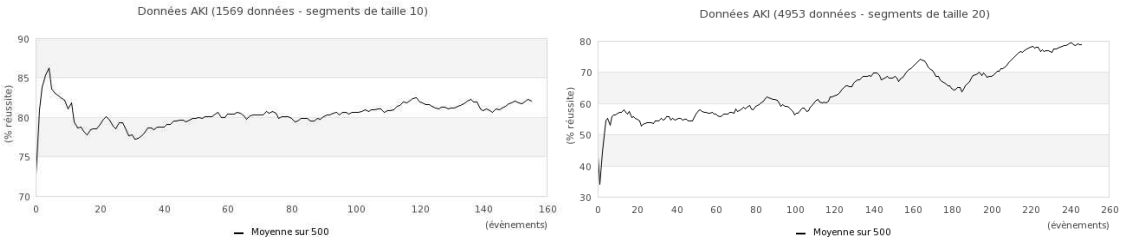


Figure 3 et 4 : A gauche, l'évolution dans le temps du rapport entre le nombre de parties gagnées et le nombre de parties jouées sur des **mots courants** (moyenne glissante sur les 500 dernières parties jouées) - A droite, l'évolution dans le temps du rapport entre le nombre de parties gagnées et le nombre de parties jouées sur des **mots de fréquence moyenne à faible** (moyenne glissante sur les 500 dernières parties jouées)

Acquisition de termes. Depuis le 1er janvier 2011, 208 nouveaux termes ont été insérés dans le réseau lexical via AKI. Ils résultent de 724 parties jouées. La quasi-totalité (90%) de ces termes sont des entités nommées (*DCRI, Révolution de jasmin, Noob, ...*), ou (10%) des termes composés et des néologismes divers (*sexe par surprise, gaz lacrymogène,…*), souvent liés à l'actualité.

Acquisition de relations. Depuis le 1er janvier 2011, 11434 relations on été acquises à l'aide de AKI (6546 parties). Si on ne compte que celles absentes du réseau, ce nombre tombe à 2105. Donc, en moyenne le réseau acquiert 1 nouvelle relation toutes les trois parties.

5.3 Analyse qualitative des parties

Sur le type de vocabulaire. Le vocabulaire (après une première analyse) se découpe en nombre de parties jouées en 24% de mots courants, le reste se divisant en 50% de mots de fréquence moyenne ou faible, et 26% de termes souvent nouveaux et liés à l'actualité. On peut considérer que ce dernier groupe est à

rapprocher des 50% si on partitionne les termes entre mots courants et les autres. Les termes liés à l'actualité conduisent souvent (69%) à un échec ce qui semble normal, étant donné qu'il s'agit de termes nouveaux (par exemple : *Révolution de jasmin, Jean-Luc Mélenchon, Médiator* - essentiellement des entités nommées) ou de termes déjà connus, mais recherchés via des indices nouveaux (*président, Tunisie, fuite => Ben Ali*). Certains mots déjà connus de AKI, c'est-à-dire présents dans le réseau lexical, sont réactualisés par l'actualité : *trafic d'organes, Lance Armstrong, aspartame*. Le compte sur l'ensemble des termes joués (indépendamment du nombre de parties) donne environ la même répartition de 1/4 de vocabulaire courant et de 3/4 de termes rares ou récents.

Sur les indices proposés. Le nombre moyen d'indices pour trouver un mot est de 2,8. Dans 40% des cas, un terme courant est trouvé dès le premier indice. Un peu moins de 3% des parties sont poursuivies au-delà de 5 indices, les 97% se divisant entre les trois cas suivants : a) le mot est trouvé avant, b) AKI échoue avant, ou c) l'utilisateur abandonne. Quand la recherche va au delà de 5 indices, la partie aboutit à une réussite dans 60% des cas. Il s'agit de termes de domaine fortement lexicalisés, et AKI est sur la bonne voie.

Une analyse des indices donnés lors des parties indique que moins de 0,3% des parties comporte au moins un indice apparemment non cohérent. Soit l'utilisateur a voulu volontairement mettre le système en défaut en donnant un indice sans rapport, soit il s'agit d'une erreur ou d'une confusion. On remarquera (avec satisfaction) que la quasi totalité des parties sont jouées "honnêtement" (ce qui peut s'expliquer par le manque d'intérêt à mettre en défaut le système avec des indices absurdes). On peut grouper les indices proposés en deux catégories :

- les indices **frontaux** (noté F) sont ceux qui amènent rapidement à la solution (trois indices au maximum). Dans le réseau, ils sont fortement connectés à la solution, en général de façon bidirectionnelle. Par exemple : *félin* pour *chat.*
- les indices **latéraux** (notés L) sont ceux qui sont très faiblement connectés à la solution et par ailleurs beaucoup plus fortement connectés à d'autres termes. Par exemple : *lisse, blanc, froid* pour *lavabo*.

Les parties concernant les mots courants correspondent à des parties dont la séquence type est : L+ (une succession d'indices latéraux). Plus le terme cible devient rare ou récent, plus la séquence type se rapproche de F+ (une succession d'indices frontaux). Il existe quelques autres schémas de parties, mais qui restent fort minoritaires. Les schémas les plus notables sont :

- L+ : uniquement des indices latéraux : *garçon, caillou, oiseau, miette* pour *Le Petit Poucet.*
- F+ : uniquement des indices frontaux : *sang* ou *couleur* pour *rouge* ou encore *mammifère, marin, défense* pour *morse*.
- L+, F : une série d'indices latéraux puis un dernier indice de type frontal permettant de trouver la solution *: blanc, lisse, dur, éléphant* pour *ivoire.*

On peut raisonnablement supposer que le schéma L+, F correspond à une activité ludique, plutôt qu'à une activité utilitaire. C'est sans doute également le cas pour le premier schéma, lorsqu'il s'agit de termes fréquents.

Sur le type d'activité. Pouvons nous déduire de l'activité enregistrée qu'il s'agit d'une activité ludique ou d'un usage MBL ? Sans doute seulement partiellement. En revanche, de nombreux indices nous portent à croire que la plupart des parties enregistrées lors de notre expérience relèvent du jeu. Nous savons que les liens partagés par nous ou des joueurs des réseaux sociaux Facebook et Twitter[10], ont généré 60% de l'activité de AKI. Ces liens proposaient de jouer tel ou tel mot. L'envoi du lien vers AKI à des listes de diffusion professionnelles (laboratoires, enseignements, sociétés savantes) générait une forte augmentation du trafic (pratiquement les autres 40%). À moins de ne penser que toutes ces personnes cherchaient le même mot à ce moment précis, on peut supposer que l'immense majorité de l'activité de AKI relève du jeu. Toujours pour aller dans ce sens, nous n'avons eu directement que deux témoignages attestant une utilisation non ludique ; dans ces deux cas, AKI s'est révélé fort utile puisqu'il a donné satisfaction aux utilisateurs.

[10] Pour voir en temps réel le compte des parties d'AKI, http://twitter.com/#!/Tot_aki

5.4 Conclusion de l'évaluation

À l'aune des résultats ci-dessus, il ne nous est pas permis de conclure avec certitude que les hypothèses 1 et 2 présentées au début de cet article sont globalement valides, mais elles ne sont pas invalidées pour autant. De nombreux indices nous laissent penser que quasiment toutes les parties analysées proviennent du jeu et non d'une utilisation réelle. Nous avons également montré que le vocabulaire utilisé durant ces jeux était le même vocabulaire que celui faisant l'objet du MBL. Ceci permet de valider notre première hypothèse de travail, à savoir, que *l'évaluation d'un outil de MBL peut se faire grâce à des joueurs.* En revanche, notre seconde hypothèse de travail —(*l'éventail des comportements des utilisateurs jouant avec un outil de MBL inclut le comportement de ceux ayant réellement besoin de retrouver un mot*)— demanderait une analyse plus fine.

Le réseau et sa consolidation via l'activité générée avec AKI permettent, dans le cas de vocabulaire complètement ouvert, de trouver le terme dans 78% des cas. Dans le cas de vocabulaire considéré comme courant, on se situe aux alentours de 82%. Enfin, dans le cas de vocabulaire filtré (issu du jeu Tabou inversé), on atteint 98,8%. On notera que, dans ce dernier cas, la performance des êtres humains se situe aux alentours de 80%.

Nous avons évalué les performances de cinq personnes sur du vocabulaire tout venant. A cette fin, nous avons choisi au hasard pour chacun d'eux 100 termes parmi ceux joués dans AKI et pour lesquels ce dernier avait soit majoritairement trouvé (50 termes) soit échoué (50 termes). Les indices donnés étaient les 5 termes les plus fortement associés dans le réseau. La performance globale a été de 46%, chiffre à comparer avec les 75-80% d'AKI.

Conclusion

Nous avons construit un réseau lexical évolutif de grande taille grâce à l'activité d'utilisateurs jouant en ligne (projet JeuxDeMots). Ces joueurs n'étant *a priori* pas des spécialistes, ce réseau représente un ensemble de connaissances générales communes. Avec des joueurs experts, il serait envisageable d'étendre ces connaissances, et donc le réseau, à des domaines spécialisés (n'est-ce pas déjà en partie le cas ?). L'intervention d'experts lexicographes, limitée à certaines corrections ainsi qu'à la validation des raffinements de termes, est ″négligeable″ compte tenu de la taille du réseau. Les questions concernant l'évaluation de la qualité d'une telle ressource, celles concernant son utilité, et la forme que peut prendre cette évaluation restent cependant ouvertes.

Le but poursuivi ici était d'évaluer la ressource lexicale ainsi produite à l'aide d'un logiciel (dénommé AKI). Celui-ci peut être considéré soit comme un jeu, soit comme un outil de MBL avec une approche exclusivement sémantique et lexicale. A l'heure actuelle, il n'y a aucune prise en compte de facteurs morphologiques ou phonologiques. AKI permet donc une évaluation à grande échelle du réseau par les utilisateurs eux-mêmes, qu'ils soient ou non des joueurs ayant contribué via JeuxDeMots. Quel que soit l'ensemble des termes considérés (termes courants ou termes de fréquence plus réduite) les performances d'AKI sont d'environ 80% ± 5%. Les résultats montrent par ailleurs que AKI est réellement utile, permettant de trouver des termes dans une ressource existante, tout en étant susceptible de l'enrichir grâce à sa capacité d'apprentissage.

On peut déduire des performance de AKI que 75% des termes pour lesquels il a été sollicité sont bien indexés, en tout cas suffisamment bien pour permettre le bon choix en cas de désambiguïsation lexicale (avocat: profession vs. fruit). L'évaluation se poursuit au long cours et les participants cherchant constamment à mettre en défaut AKI renforcent l'indexation, mais également l'évaluation avec une sévérité croissante – les deux se compensant. Il y a un auto-ajustement des joueurs en faveur d'indices faisant partie de la longue traîne. Une question restant cependant ouverte est de savoir à quel taux de réussite AKI va asymptotiquement plafonner. Cette valeur pourrait être un indice concernant une performance maximale en désambiguïsation lexicale en utilisant notre ressource.

Références

ABRAMS L., TRUNK D. L., & MARGOLIN S. J. (2007). Resolving tip-of-the-tongue states in young and older adults: The role of phonology. IN L. O. RANDAL (ED.), A*ging and the Elderly: Psychology, Sociology, and Health* (pp. 1-41). HAUPPAUGE, NY: NOVA SCIENCE PUBLISHERS, INC.

VON AHN L., DABBISH L. (2004). Labelling Images with a Computer Game. *ACM Conference on Human Factors in Computing Systems (CHI)*. pp. 319-326

AITCHISON J. (2003). W*ords in the Mind: an Introduction to the Mental Lexicon*. OXFORD, BLACKWELL.

BROWN R. & MCNEILL D. (1966). The "tip-of-the-tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior, 5*, pp. 325-337.

CARAMAZZA A. (1997). « How many levels of processing are there in lexical access ? » *Cognitive Neuropsychology*, 14, pp. 177-208.

COLLINS A, QUILLIAN M.R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behaviour*, 8(2), pp. 240-248.

FERRAND L. (1998). Encodage phonologique et production de la parole. *L'année psychologique*. vol. 98, n°3. pp. 475-509.

JI H., PLOUX S. AND WEHRLI E. (2003) Lexical knowledge representation with contexonyms. In *Proceedings of the 9th MT summit*, pp. 194-201

LAFOURCADE M., JOUBERT A. (2009). Similitude entre les sens d'usage d'un terme dans un réseau lexical. *Traitement Automatique des Langues*, vol.50/1, pp. 177-200

LAFOURCADE M., JOUBERT A. (2010). Détermination et pondération des raffinements d'un terme à partir de son arbre des usages nommés. *Traitement Automatique des Langues Naturelles (TALN'10)*. Montréal, 6 p.

LEVELT W., ROELOFS A. & A. MEYER. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1-75.

LIEBERMAN H., SMITH D.A., TEETERS A. (2007). Common Consensus: a web-based game for collecting commonsense goals. *International Conference on Intelligent User Interfaces (IUI'07)*. Hawaï, USA.

MEL'ČUK I.A., CLAS A., POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Editions Duculot AUPELF-UREF

MILLER G.A., BECKWITH R., FELLBAUM C., GROSS D. AND MILLER K.J. (1990). Introduction to WordNet: an on-line lexical database , *International Journal of Lexicography*, 3 (4), pp. 235-244.

PLOUX S., VICTORRI B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement Automatique des Langues*, vol.39/1, 161-182

POLGUÈRE A. (2006). Structural properties of Lexical Systems : Monolingual and Multilingual Perspectives. *Proceedings of the Workshop on Multilingual Language Resources and Interoperability (Coling/ACL)*, Sydney, pp. 50-59.

ROSSI M. (2001) : Les lapsus et la production de la parole. *Psychologie Française*, n° 46, pp. 27-41.

SITBON L. (2007). *Combinaisons de ressources linguistiques pour l'aide à l'accès lexical : études de faisabilité*, actes de RECITAL 2007, 5-8 juin 2007, Toulouse, France

SOWA J. (1992). *Semantic networks*, Encyclopedia of Artificial Intelligence, edited by S.C. Shapiro, Wiley, New York

SPENCE D.P. & OWENS K.C. (1990). Lexical co-occurrence and association strength, *Journal of Psycholinguistic Research*, 19 (5)

ZAMPA V., LAFOURCADE M. (2009). Evaluations comparées de deux méthodes d'acquisition lexicale et ontologique : JeuxDeMots vs Latent Semantic Analysis. *XVIèmes rencontre de Rochebrune : ontologie et dynamique des systèmes complexes, perspectives interdisciplinaires*

ZOCK M., FERRET O., SCHWAB D. (2010) Deliberate word access : an intuition, a roadmap and some preliminary empirical results, In *A. Neustein (éd.) 'International Journal of Speech Technology'*, 13(4):107-117, 2010. Springer Verlag.

ZOCK M. (2002). Sorry, what was your name again, or how to overcome the tip-of-the tongue problem with the help of a computer? *SemaNet workshop* (Building and Using Semantic Networks), Coling, Taipei, pp. 107-112.

Morphologie et Segmentation

TALN 2011, Montpellier, 27 *juin* –1er *juillet 2011*

Identifier la cible d'un passage d'opinion dans un corpus multithématique

Matthieu Vernier, Laura Monceaux, Béatrice Daille
Université de Nantes, LINA, 2, rue de la Houssinière 44322 Nantes
{Matthieu.Vernier, Laura.Monceaux, Beatrice.Daille}@univ-nantes.fr

Résumé. L'identification de la cible d'une d'opinion fait l'objet d'une attention récente en fouille d'opinion. Les méthodes existantes ont été testées sur des corpus monothématiques en anglais. Elles permettent principalement de traiter les cas où la cible se situe dans la même phrase que l'opinion. Dans cet article, nous abordons cette problématique pour le français dans un corpus multithématique et nous présentons une nouvelle méthode pour identifier la cible d'une opinion apparaissant hors du contexte phrastique. L'évaluation de la méthode montre une amélioration des résultats par rapport à l'existant.

Abstract. Recent works on opinion mining deal with the problem of finding the semantic relation between sentiment expressions and their target. Existing methods have been evaluated on monothematic english corpora. These methods are only able to solve intrasentential relationships. In this article, we focus on this task apply to french and we present a new method for solving intrasentential and intersentential relationships in a multithematic corpus. We show that our method is able to improve results on the intra- and intersential relationships.

Mots-clés : Fouille d'opinions, Identification des cibles, Méthode RankSVM.

Keywords: Opinion mining, Targeting sentiment expressions, RankSVM.

1 Introduction

Le début des années 2000 marque l'éclosion de la fouille d'opinions. Les travaux pionniers se sont principalement intéressés à la catégorisation globale de documents d'opinion, soit selon leur polarité (Turney, 2002; Torres-Moreno *et al.*, 2007), soit selon leur subjectivité (Wiebe & Riloff, 2005). Depuis, un très grand nombre de travaux traitent de données textuelles d'opinion dans des axes scientifiques et des domaines applicatifs très différents. Plus récemment, le recul sur dix ans de travaux permet selon nous de segmenter le domaine en cinq problématiques :

– **extraire les mots d'opinions** d'une langue pour construire des ressources et améliorer leur qualité (Baccianella *et al.*, 2010; Mathieu, 2006) ;
– **catégoriser globalement un document** selon l'opinion (Torres-Moreno *et al.*, 2007; Pang & Lee, 2008) ;
– **catégoriser des passages d'opinions** dans un document qui exprime des opinions hétérogènes (Wilson, 2008) ;
– **identifier la source** [1] d'une opinion (Choi *et al.*, 2005; Ruppenhofer *et al.*, 2008) ;
– **identifier la cible** [2] d'une opinion (Kessler & Nicolov, 2009; Jakob & Gurevych, 2010).

1. l'énonciateur d'une opinion.
2. le sujet sur lequel porte l'opinion.

L'identification de la cible d'un passage d'opinion fait partie des axes les plus récemment abordés dans la littérature du domaine. Aucun travail ne s'y est intéressé pour le traitement du français et, à notre connaissance, seuls deux travaux anglophones majeurs y consacrent une étude spécifique (Kessler & Nicolov, 2009; Jakob & Gurevych, 2010). Pourtant, afin d'analyser des textes dont le contenu est de plus en plus hétérogène (blogs, forums, etc.), cette problématique correspond à un besoin particulièrement criant. Elle nécessite de combiner des techniques connues en traitement automatique des langues (analyse syntaxique, résolution d'anaphore pronominale et nominale, segmentation thématique, etc.) pour proposer une solution souple et efficace (voir §. 2).

Dans cet article, nous nous intéressons spécifiquement à l'identification de la cible d'une opinion (voir §. 2). Nous renvoyons à (Vernier & Monceaux, 2010) pour la présentation d'une méthode et d'une ressource pour la détection automatique des passages d'opinions pour le français. Dans nos travaux précédents, nous avons développé le premier corpus francophone (*corpus Blogoscopie*) dans lequel les passages d'opinions et leur cible sont annotés (voir §. 3.1) (Dubreil *et al.*, 2008). Une étude statistique sur ce corpus permet de structurer le problème et de proposer différentes pistes de solutions (voir §. 3.2). Nous présentons quatre méthodes (dont deux méthodes *baseline*) pour identifier la cible d'un passage d'opinion (voir §. 4). Nous expérimentons ces quatre méthodes sur une sous-partie du corpus Blogoscopie et discutons des résultats obtenus (voir §. 5).

2 Une problématique récente peu explorée

L'identification de la cible d'une opinion a souvent été ignorée ou considérée comme un aspect de second plan. Par exemple, les travaux de catégorisation de l'opinion au niveau du document considèrent qu'un texte est monothématique et n'évalue qu'un seul objet donné (*un film, un livre, un appareil photo,* etc.). Ce n'est que récemment que quelques travaux anglophones ont directement axés leur étude sur cette problématique (Kessler & Nicolov, 2009; Jakob & Gurevych, 2010). Ils ont introduit le terme *target*, traduit ici par *cible*, pour désigner l'objet concerné par un passage d'opinion. Dans la pratique, ils ont limité cette cible à quelques catégories d'objets :
– un produit particulier ou une marque (*iPhone, EOS 5D de Canon, Apple, Matrix*, etc.) ;
– une caractéristique d'un produit (*la durée de vue de la batterie*, *le scénario d'un film*, etc.).
En fait, de part le large éventail d'objets qu'elle peut couvrir, la notion de cible est complexe à circonscrire. Elle est d'ailleurs souvent laissée floue dans les travaux du domaine. Dans notre cadre, nous considérons que tout objet peut être la cible d'une opinion. Nous tâchons de mieux appréhender cette notion en présentant quatre facteurs qui rendent complexe la problématique d'identification de la cible d'une évaluation (§. 2.1). Plusieurs approches ont été proposées mais celles-ci ne résolvent que partiellement la tâche et n'ont été appliquées qu'à l'anglais. Nous précisons leurs limites et motivons le travail présenté dans cet article (§. 2.2).

2.1 Une tâche complexe

L'identification de la cible d'une opinion consiste à relier un passage d'opinion avec l'objet du monde qu'il évalue. Dans (1), il y a ainsi trois passages d'opinions qui évaluent un objet cible unique. Néanmoins, cette tâche s'avère particulièrement complexe de part quatre facteurs que nous précisons ci-dessous.

(1) L'équipe de France va mal . Entre honte et déception .

Différentes formes textuelles pour un unique objet du monde Premier facteur, comme l'a introduit Benveniste (Benveniste, 1966), il importe de distinguer les signes linguistiques et les objets du monde auxquels ils font références. Un même objet peut être représenté dans le texte par différentes expressions nominales ou pronominales.

Dans (2) et (3), on parle de l'objet du monde *équipe de France de football* par une variante nominale métonymique (*les Bleus*) et par une anaphore pronominale (*elle*). Dans un but applicatif, il est nécessaire de regrouper les opinions qui portent sur le même objet et de nommer l'objet évalué le plus précisément possible. On ne pourra ainsi pas considérer le pronom *elle* comme la cible de l'évaluation *séduisante*. *Les Bleus* est une réponse intermédiaire plus acceptable pour identifier l'objet cible réellement évalué (*l'équipe de France de football*).

(2) [...] lors du fiasco des Bleus en juin.

(3) Elle a enfin été séduisante [...].

Les relations méronymiques entre objets Les objets sont potentiellement liés à d'autres objets par des relations méronymiques [3]. Dès lors, même si évaluer un méronyme A d'un mot B peut être une façon d'évaluer indirectement B, il importe de considérer le méronyme A comme la cible exacte de l'évaluation. Ainsi, dans (4) et (5), on évalue tout d'abord *la défense* et *le marquage sur les corners* de *l'équipe de France* et non *l'équipe de France* dans sa globalité.

(4) Disons-le clairement : la défense française n'a pas été très bonne, ce soir.

(5) Le marquage sur les corners reste encore approximatif.

Présence de plusieurs objets candidats autour de l'opinion Le troisième facteur est la présence de plusieurs objets distincts dans le contexte d'un même passage d'opinion. Dans (6), les deux passages d'opinions portent sur des cibles distinctes qu'il faut pouvoir déterminer parmi les quatres objets présents dans la phrase : *la dernière finale de la coupe du monde, l'équipe de France, le coursier, les pizzas*. L'objet le plus proche de l'opinion n'est pas nécessairement sa cible.

(6) [...] pizzas commandées lors de la dernière finale de coupe du monde avec l'équipe de France, coursier courageux, pizzas aussi banales que le match.

Proximité aléatoire entre l'opinion et sa cible Dernier facteur de complexité, la cible évaluée ne se situe pas toujours à proximité du passage d'opinion. Dans (7), le passage d'opinion *de la provocation* [4] porte sur *la ligne du sélectionneur* qui se trouve dans la phrase précédente. Le pronom *ce* en est une anaphore. Néanmoins, la présence de nombreux autres objets dans le contexte (*un iota, le jeu, les joueurs, la concrétisation de cette vue*) rend complexe l'identification de la cible réelle de l'opinion pour une approche automatique.

(7) force est de constater que la ligne du sélectionneur n'a jamais bougé d'un iota : le jeu appartient aux joueurs. Ce qui a souvent pu passer pour de la provocation n'est en fait [...]

Pour résoudre la problématique d'identification de la cible d'une opinion, il convient donc de considérer ces quatre facteurs. Le processus d'identification consiste alors à :

– repérer les différents objets impliqués dans un énoncé. Ceux-ci sont généralement représentés textuellement par des groupes nominaux ou pronominaux (tâche T_1) ;
– identifier l'objet cible qui est directement concerné par une opinion (tâche T_2) ;
– résoudre les relations d'anaphores nominales ou pronominales entre les objets (tâche T_3) ;
– résoudre les relations méronymiques entre les objets (tâche T_4).

3. Un méronyme A d'un mot B est un mot dont le signifié désigne une sous-partie du signifié de B.
4. Ce passage est considéré comme une opinion même s'il s'agit d'un discours rapporté.

2.2 Travaux liés et motivations

Pour résoudre l'identification de la cible d'une opinion, quatre types d'approches ont été proposées dans les travaux du domaine. Les solutions qu'elles proposent ne sont que partielles et nécessitent d'être améliorées. De plus, elles n'ont été évaluées que sur l'anglais et sur des corpus monothématiques.

Quatre approches proposées La **proximité** est la première hypothèse considérée pour relier une cible et une opinion. Ainsi, dans les travaux de Grefenstette *et al.* (2004) (textes journalistiques sur la politique) et Mishne & Glance (2006) (blogs sur le cinéma), les auteurs considèrent un mot ou une expression w comme objet cible. Toutes les opinions d'un document qui co-occurent avec w dans une certaine fenêtre de mots, sont considérées comme des opinions qui portent sur w. Cette tâche n'étant pas le coeur de leur travail, l'efficacité de cette méthode n'a pas été évaluée. Les **dépendances syntaxiques** entre une opinion et une cible constitue la deuxième méthode pour résoudre cette problématique. Hu & Liu (2004) se sont restreints aux chemins syntaxiques proches entre un adjectif et une cible dans un corpus de film. De la même façon, Bethard *et al.* (2004) se sont intéressés uniquement aux verbes d'opinion. Bloom *et al.* (2007) ont développé manuellement des listes de chemins de dépendances syntaxiques entre une opinion et sa cible dans des corpus de produits commerciaux. La précision obtenue varie autour de 0,70. Toutes ces approches ont été entraînées sur des corpus monothématiques ce qui tend à restreindre la richesse des cas observés. De plus, ces méthodes ne sont pas applicables dans les nombreux cas où la cible se situe dans une phrase différente de l'opinion. Kim & Hovy (2006) ont ainsi montré que dans une majorité des cas l'étude des dépendances syntaxiques ne suffit pas à déterminer la bonne cible et qu'elle est souvent confondue avec la source de l'opinion. En **combinant des informations** lexicales, grammaticales et syntaxiques par apprentissage, Kessler & Nicolov (2009) ont amélioré les résultats des approches syntaxiques sur un corpus monothématique de critiques de voiture. Leur méthode est la méthode état de l'art pour l'anglais, néanmoins elle se restreint toujours au grain phrase et ne traite pas les nombreux cas où la cible n'est pas présente dans la même phrase que l'opinion. Ils se sont également limités à l'étude des cas où l'opinion est un adjectif ou un verbe. Jakob & Gurevych (2010) sont à notre connaissance les seuls à sortir du grain phrase. Ils se sont intéressés spécifiquement à la **résolution d'anaphore pronominale** dans un corpus de critiques de film pour mieux identifier la cible exacte.

Motivations Dans ce travail, nous souhaitons d'une part adapter les travaux existants au français et vérifier ainsi les conclusions obtenues. D'autre part, nous souhaitons améliorer l'existant sur les trois aspects suivants :

– notre approche est **multithématique**. En effet, l'aspect monothématique des études existantes (souvent ciblées sur un genre particulier : les critiques d'objets commerciaux) tend à simplifier les cas linguistiques rencontrés ;
– notre approche **ne se limite pas à quelques catégories grammaticales** d'opinion. Les opinions peuvent ici être des groupes nominaux, adjectivaux, verbaux ou adverbiaux ;
– nous utilisons **le texte comme grain d'étude**. Les cas où la cible n'est pas présente dans la même phrase que l'opinion sont pris en compte.

Pour les besoins de notre travail, un corpus de référence où les passages d'opinions et les cibles sont annotées est nécessaire. Nous présentons dans la section suivante le corpus utilisé.

3 Corpus multithématique de référence

Actuellement, le seul corpus francophone disponible où les passages d'opinions et les cibles d'opinions sont annotés est le corpus de blogs *Blogoscopie* [5] (Dubreil *et al.*, 2008). Pour le problème présenté, les blogs sont intéressants car ils présentent une richesse linguistique plus grande que les textes de critiques d'objets commerciaux

5. Le corpus est disponible à l'adresse : http ://www.lina.univ-nantes.fr/Ressources.html

(beaucoup d'objets différents dans un même texte, syntaxes des phrases plus variées, etc.). Nous présentons brièvement le corpus Blogoscopie (§. 3.1) et analysons dans une deuxième partie des données statistiques sur le problème d'identification d'une cible d'opinion (§. 3.2).

Nature du corpus Le corpus Blogoscopie est constitué de 200 billets de blogs, et de 614 commentaires associés à ces billets, tous issus de la plateforme de blogs OverBlog [6]. Il est constitué d'environ 100 000 mots. La création du corpus Blogoscopie est basée sur un critère de représentativité thématique. Ainsi :
– 110 billets et 458 commentaires ont été sélectionnés à partir des 33 catégories de la plateforme OverBlog : *actualité, artiste, cinéma, consommation, console, croyance, détente, économie, gastronomie,* etc.
– 90 billets et 156 commentaires ont été sélectionnés à partir de dix mots-clés (9 billets par mot-clé) : *Beaujolais, développement durable, grève SNCF, Harry Potter, loi LRU, énergie nucléaire, Raymond Domenech,* etc.

L'exemple (8) présente une version allégée de l'annotation d'un passage du corpus Blogoscopie. Dans ce corpus, trois types d'objets principaux ont été annotés :
– 6 876 **objets**. Ces objets peuvent être dits :
  – « principaux » (OP) : si le sujet général d'un document concerne principalement cet objet. Un même document peut avoir plusieurs objets principaux ;
  – « associés » (OA) : si cet objet est une sous-partie d'un objet concerné.
– 4 909 **opinions** ;
– 4 129 **couples cible-opinion**.

(8) Le sommet a été atteint avec **<objet>** le nucléaire **</objet>** : **<objet>** M. Sarkozy **</objet>** **<opinion cible="M. Sarkozy">** a abusé **</opinion>** **<objet>** l'opinion publique **</objet>** en annonçant qu'il n'y aurait pas de **<objet>** "nouveaux sites" **</objet>**.

Précisons que dans ce corpus :
– les annotateurs n'ont pas pris en compte les anaphores pronominales ;
– la forme textuelle de l'objet cible la plus proche de l'opinion est considérée comme la cible. Lorsqu'une anaphore pronominale est présente, l'antécédent nominal le plus proche est considéré comme la cible.

Segmentation du corpus et statistiques Pour nos expérimentations, nous segmentons le corpus en deux :
– une partie « entraînement » constituée de 3 909 couples opinion-cible et de 5 584 objets (160 documents) ;
– une partie « test » constituée de 1 000 couples opinion-cible et de 1 292 objets (40 documents).

Nb OP / billet	1	2	3	4	5+
billets concernés	17 %	**35 %**	**19 %**	14 %	15 %

TABLE 1 – Nombre d'objets principaux par billet

Nb OA / billet	1-5	5-10	10-15	15+
billets concernés	**30 %**	23 %	15 %	**32 %**

TABLE 2 – Nombre d'objets associés par billet

Des statistiques sur la partie « entraînement » du corpus donnent un aperçu de la tâche. Le tableau 1 montre qu'une majorité de billets (et leurs commentaires) parle en général d'au moins deux ou trois objets principaux différents. Le tableau 2 montre que le nombre d'objets associés est très variable. Dit autrement, si dans une majorité de cas un blogueur articule le sujet de son billet autour de deux ou trois objets principaux, en revanche, on peut difficilement prédire le nombre d'objets associés qu'il va évoquer. Pour identifier la cible d'une opinion, il s'agit donc de la trouver parmi une liste d'objets pouvant aller de 2 à plus de 20. Le tableau 3 montre que la position de la cible est variable : soit très proche, soit très éloignée (voire dans une phrase différente dans 44 % des cas). Dans 47 % des cas, il y a au moins un objet intercallé entre l'opinion et sa cible.

6. http ://www.over-blog.com

Position relative de la cible	avant l'opinion		après l'opinion	
opinions concernées	**61 %**		39 %	
Distance - en nombre de mots -	0	1-4	5-10	11+
opinions concernées	**21 %**	**20 %**	11 %	**48 %**
Distance - en nombre de phrases -	0	1	2	3+
opinions concernées	**56 %**	9 %	6 %	19 %
Distance- en nombre de objets intercallés -	0	1	2	3+
opinions concernées	**53 %**	7 %	12 %	**28 %**

TABLE 3 – Étude de la position de la cible par rapport à son opinion

4 Méthodes utilisées

D'après les statistiques du corpus « entraînement » Blogoscopie, nous distinguons deux niveaux de complexité :
– la cible et l'opinion sont dans la même phrase (intraphrastique) : 56 % des cas du corpus d'entraînement ;
– la cible et l'opinion sont dans une phrase différente (interphrastique) : 44 % des cas du corpus d'entraînement.
Dès lors, nos objectifs sont les suivants :
– **évaluer les méthodes existantes** qui traitent uniquement les cas **intraphrastiques** afin de les valider sur le français, sur un corpus multithématique et en prenant davantage en compte les différentes catégories grammaticales des opinions ;
– **développer une méthode** qui traite les cas **intra- et interphrastiques** simultanément.
Pour le premier objectif, nous ré-implémentons la méthode état-de-l'art (**M-RankSVM**, voir §. 4.1.1) et nous la comparons à une méthode *baseline* (**M-Syntaxe**, voir §. 4.1.2). Pour le second objectif, nous proposons une méthode basée sur la saillance d'une cible (**M-Saillance**, voir §. 4.2.1). Celle-ci ré-utilise les résultats de **M-RankSVM** comme une amorce. Nous la comparons avec une méthode *baseline* (**M-Proximite**, voir §. 4.2.2). Pour ces quatre méthodes :
– les passages d'opinions sont annotés manuellement (annotations présentes dans le corpus Blogoscopie) ;
– le TreeTagger (Schmid, 1994) donne la catégorie grammaticale ;
– un algorithme de découpage en syntagmes non-récursifs, qui réimplémente la méthode de (Vergne & Giguet, 1998), permet d'identifier les groupes nominaux et pronominaux d'un texte. Ces expressions nominales et pronominales sont considérées comme l'ensemble des objets principaux et associés du texte.

4.1 Méthodes applicables à un niveau intraphastique

4.1.1 Approche par combinaison d'indices morpho-syntaxiques

La méthode état de l'art de Kessler & Nicolov (2009) a été évaluée sur un corpus monothématique anglais. Elle ne s'intéresse qu'aux opinions de type verbe et adjectif uniquement. Elle consiste à combiner des informations lexicales et sémantiques pour pallier les limites des approches purement syntaxiques. Nous adaptons leur approche pour traiter un corpus multithématique français sans restriction grammaticale (par la suite, **M-RankSVM**). La méthode prend en entrée : une phrase s à analyser, un passage d'opinion $opinion$ présent dans s, un ensemble d'objets o présents dans s et une fonction d'ordonnancement permettant de classer les objets de s. Cette fonction

d'ordonnancement est extraite par apprentissage sur le corpus « entraînement ».

Pour obtenir la fonction d'ordonnancement, nous utilisons l'algorithme RankSVM issue de SVMLight (Joachims, 2002). Le corpus « entraînement » Blogoscopie permet de fournir à la méthode :
– 3 909 exemples « positifs » correspondant aux couples « opinion-cible » présents dans le corpus. Ces exemples ont un rang 1 ;
– 4 000 exemples « négatifs » correspondant à des couples « objet non cible-opinion » présents dans le corpus. Ces exemples ont un rang 0 ;
Chacun de ces exemples est décrit par les caractéristiques présentées dans le tableau 4. À partir de ces exemples, RankSVM entraîne un modèle ayant pour objectif de déterminer une fonction de préférence qui maximise le *rang*. À l'image de Kessler & Nicolov (2009), nous utilisons les paramètres par défaut de RankSVM. Lors de la phase de test, cette méthode effectue un ordonnancement de tous les objets qui apparaissent dans la même phrase que le passage d'opinion en s'appuyant sur le modèle appris lors de la phase d'entraînement. L'objet classé premier est considéré comme la cible de l'opinion.

Caractéristique	**Exemple**
Chemin lexical lemmatisé entre l'opinion et sa cible	*pour*
Chemin grammatical entre l'opinion et sa cible	Prep.
Catégorie(s) grammatical(es) de l'opinion	DET NOM ADJ
Catégorie(s) grammatical(es) de la cible	DET NOM
Chemin de dépendance	↓ PP, ↓ DP
Nombre d'objets entre l'opinion et la cible	0
Nombre de mots entre l'opinion et la cible	1
Rang	**1**

TABLE 4 – Caractéristiques utilisées pour décrire chaque instance d'apprentissage. Exemple : « [...] est une défaite majeure pour l'écologie »

Chemin de dépendance(s)	**Fréquence**	**Exemple**
↑ AP	121	C'est une belle image (ex : $[_{TP}[_{DP}$ C'] est $[_{VP}$ $[_{DP}$ une $[_{NP}[_{AP}$ belle] image]]]])
↓ DP	91	Cette question me taraude
↓ PP, ↓ DP	65	J'ai une véritable fascination pour J.K Rowling
↑ TP, ↑ CP	59	Un événement qui contribue à alourdir le débât
↑ VP, ↓ DP	45	Ce jeu est accessible pour les enfants de 12 ans
↑ AdvP, ↓ DP	29	Cette histoire de travailler plus [...] ne tient [...] pas la route

TABLE 5 – Chemins de dépendances syntaxiques les plus fréquents entre une opinion et une cible dans le corpus d'entraînement Blogoscopie. AP = syntagme adjectival ; DP = syntagme déterminant ; PP = syntagme prépositionnel ; AdvP = syntagme adverbial ; CP = syntagme conjonctionnel ; NP = syntagme nominal ; TP = syntagme temporel ; VP = syntagme verbal

4.1.2 Approche syntaxique (*baseline*)

Nous comparons la méthode **M-RankSVM** avec une méthode qui repose uniquement sur une analyse des dépendances syntaxiques (par la suite, **M-Syntaxe**). Elle ne fonctionne également que dans un cadre intraphrastique. Elle prend en entrée : une phrase s à analyser, un passage d'opinion $opinion$ présent dans s, un ensemble d'objets O présents dans s, un ensemble de chemins de dépendances syntaxiques entre une opinion et une cible avec leur probabilité correspondante. Cet ensemble est déterminé à partir du corpus « entraînement » Blogoscopie.

Nous avons utilisé l'analyseur de dépendance syntaxique FIPS (Wehrli, 2007) sur 500 exemples [7] de couples opinion-cible qui apparaissent dans la même phrase dans le corpus d'entraînement Blogoscopie. Pour ces exemples, FIPS fournit un chemin de dépendance (visualisable sous forme d'arbre syntaxique (Fig. 1)). Nous classons par ordre de fréquence décroissante (voir tableau 5), les chemins de dépendances qui vont :
– de la fin d'une opinion vers le début d'une cible, si l'opinion est située avant la cible dans le texte ;
– du début de l'opinion vers la fin d'une cible, si l'opinion est située après la cible dans le texte.
Lors de la phase de test, cette méthode effectue une analyse syntaxique de la phrase contenant l'opinion. Le chemin de dépendance syntaxique le plus fréquent qui permet de relier $opinion$ avec un objet o ($\in O$) désigne l'objet o comme la meilleure cible potentielle. En utilisant un des chemins syntaxiques, il est possible que o soit un groupe pronominal (*il*, *aucun d'entre eux*, etc.).

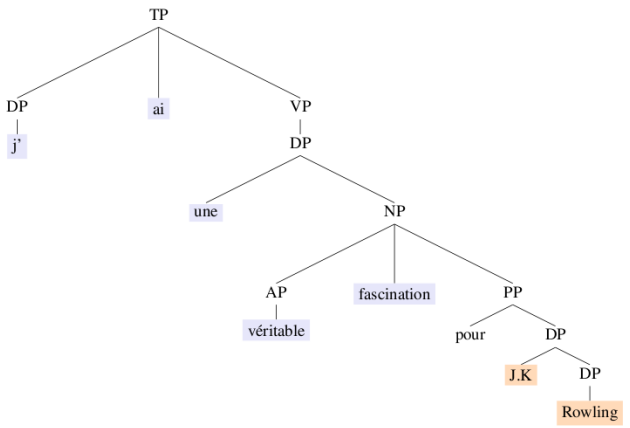


FIGURE 1 – Analyse syntaxique de « *J'ai une véritable fascination pour J.K Rowling* ».
L'opinion est séparée de sa cible par la dépendance ↓ **PP,** ↓ **DP**.

4.2 Méthodes applicables à un niveau intra- et interphrastique

Pour le second objectif, nous proposons une méthode (**M-Saillance**) pour traiter les cas intraphrastiques et interphrastiques. Nous comparons cette méthode à une méthode *baseline* (**M-Proximite**).

7. Le nombre de couples est réduit à 500 pour éviter les cas des phrases trop longues qui induisent en erreur l'analyseur FIPS.

4.2.1 Approche par mesure de saillance

Lors de la phase de test, en sortie de la méthode **M-RankSVM**, pour chaque document d, nous disposons d'un ensemble de couples d'opinion-cible déterminés automatiquement. Nous observons les quatre faits suivants :
– O_1 : certaines opinions n'ont **pas de cible affectée** (si la phrase contenant l'opinion ne possède aucun objet) ;
– O_2 : certaines opinions ont pour cible **un groupe pronominal** ;
– O_3 : certains couples cible-opinion ne sont **pas reliés par une dépendance syntaxique couramment observée** ;
– O_4 : certaines des cibles ne sont **pas évaluées ailleurs dans le document**.

Les couples qui vérifient l'une des observations O_1 ou O_2, ou qui vérifient simultanément O_3 et O_4 sont considérées comment des couples **faibles** (cas douteux). Les autres couples sont considérés comme des identifications **fortes** sur lesquelles nous pouvons nous appuyer. Notre méthode (**M-Saillance**) consiste à utiliser les résultats « forts » de **M-RankSVM** comme une amorce pour corriger les couples faibles. Cette méthode a deux objectifs :
– identifier une cible nominale pour les opinions qui n'en ont pas ou qui sont reliées à un groupe pronominal ;
– corriger l'affectaction des autres couples faibles.
Les hypothèses linguistiques de notre méthode sont les suivantes :
– **Hypothèse 1** : s'il n'existe pas de objet nominal pertinent dans la même phrase qu'une opinion, celui-ci existe probablement dans les phrases voisines ;
– **Hypothèse 2** : l'énonciateur d'un discours subjectif articule ses opinions au fil d'un document en évaluant les mêmes objets plusieurs fois. Plus un objet est la cible d'une opinion, plus il est saillant et plus il est probable qu'il soit de nouveau la cible d'une opinion dans le contexte.

Description Soit $opinion$, un passage d'opinion du document d, dont la relation avec sa cible actuelle est **faible**. Soit $segment$, l'ensemble des phrases voisines de l'opinion. $segment$ est limité à 4 phrases avant l'opinion et 4 phrases après. À partir des couples forts opinion-cible issus de **M-RankSVM**, nous classons par ordre décroissant les cibles les plus fréquemment liées à une opinion dans $segment$. Nous classons également les cibles les plus fréquentes dans d. Les cibles sont préalablement lemmatisées et les mots fonctionnels sont retirés (par exemple, *l'entraîneur de l'équipe de France* devient *entraîneur équipe France*). Pour tout objet o présent dans $segment$, nous mesurons une probabilité que o soit la cible de $opinion$ avec la formule suivante :

$$\text{P}(o, opinion) = \frac{\text{NB}(o, cible, segment)}{\text{NB}(o, segment).\text{NB}(cible, segment)} \cdot \frac{\text{NB}(o, cible, document)}{\text{NB}(o, document).\text{NB}(cible, document)} \quad (1)$$

$$\text{Cible}(opinion) = \text{MAX}(\text{P}(o, opinion)) \quad (2)$$

où :
– NB(o, $cible$, $segment$) est le nombre de fois que le objet o est une cible forte d'une opinion quelconque dans l'ensemble des phrases voisines d'$opinion$. De la même façon, NB(o, $cible$, $document$) est le nombre de fois que o est une cible forte d'opinion dans le document ;
– NB(o, $segment$) et NB(o, $document$) sont les nombres de fois où le objet o est présent dans le segment de phrases voisines et dans le document ;
– NB($cible$, $segment$) et NB($cible$, $document$) sont les nombres de cibles d'opinions dans le segment et dans le document.
P(o, $opinion$) représente la saillance de l'objet o dans le segment et le document. Cette mesure donne un score d'association entre le l'objet o et l'opinion émise dans le document et le segment en calculant la dépendance de ces trois variables. La cible de l'opinion choisie par **M-Saillance** est l'objet qui maximise cette probabilité.

4.2.2 Approche par proximité (*baseline*)

La méthode par proximité (par la suite, **M-Proximité**) est applicable au niveau intraphrastique et interphrastique. Elle prend en entrée : le document d à analyser, un passage d'opinion $opinion$ et sa position dans le document, un ensemble d'objets O présents dans d et leur position dans le document. La position dans le document est déterminée par une indexation de l'ensemble des mots du texte : $W = \{w_1, w_2, \ldots, w_n\}$. Le passage $opinion$ et l'ensemble des éléments de O sont délimités par un mot de début ($w_{\text{début}} \in W$) et un mot de fin ($w_{\text{fin}} \in W$).

Pour tout $o \in O$, cette méthode compte le nombre de mots qui séparent l'objet o de $opinion$ dans le document d. La ponctuation n'est pas prise en compte. Si $opinion$ est inclu dans o, la distance en mots est égale à 0. Le choix du meilleur candidat s'effectue sur la base de la plus petite distance en mots. En cas d'égalité, la cible qui précède l'opinion est préférée (nous justifions ce choix empirique par l'étude statistique réalisée précédemment).

$$\forall\, o \in O \text{ se situant avant } opinion, \text{ Proximite}(t) = w_{\text{fin}}(o) - w_{\text{début}}(opinion)$$

$$\forall\, o \in O \text{ se situant après } opinion, \text{ Proximite}(t) = w_{\text{fin}}(opinion) - w_{\text{début}}(o)$$

$$\text{Cible}(opinion) = \text{MIN}(\text{Proximite}(o))$$

5 Résultats

Type / **Score**	**Intraphrastique** Précision	**Interphrastique** Précision	**Total** Précision
M-Syntaxe	68.8 % (407/592)	non applicable	40.7 % (407/1 000)
M-RankSVM	71.5 % (423/592)	non applicable	42.3 % (423/1 000)
M-Proximite	53.5 % (317/592)	27.2 % (111/408)	42.8 % (428/1 000)
M-Saillance	**72.8 %** (431/592)	**60.0 %** (245/408)	**67.6 %** (676/1 000)

TABLE 6 – Résultats obtenus par les quatre méthodes sur le corpus de test Blogoscopie

Nous évaluons les 4 méthodes présentées précédemment sur les 40 documents du corpus test Blogoscopie. Ce corpus contient 1 000 couples cible-opinion : 592 de ces couples sont intraphrastiques, 408 sont interphrastiques.

Niveau intraphrastique Par rapport à une approche syntaxique, la combinaison d'indices morpho-syntaxiques améliore légèrement l'identification de la cible (71,5 % contre 68,8 %). Cette différence est néamoins beaucoup plus marquée dans les travaux de Kessler & Nicolov (2009). Par rapport à Kessler & Nicolov (2009) qui ont obtenu une précision de 74,8 % sur l'anglais, notre score de précision baisse d'environ 3 points sur le français (71,5 %). S'il est difficile de dire si cette baisse est significative, elle tend néanmoins à montrer que la méthode **M-RankSVM** est applicable pour traiter les cas intraphrastiques sur des corpus multithématiques en français. Elle est d'autant plus intéressante qu'elle fournit de bons indices à la méthode **M-Saillance** pour corriger certains couples intraphrastiques (72,8 % contre 71,5 %). L'hypothèse discursive consistant à reclasser les cibles d'opinions saillantes (celles qui sont fréquement liées à une opinion) semble être une hypothèse intéressante à développer.

Niveau interphrastique La méthode **M-Saillance** permet d'identifier correctement 60,0 % des cibles d'opinions en s'appuyant uniquement sur l'objet d'opinion le plus saillant dans une fenêtre de 4 phrases autour de l'opinion. Ce résultat est nettement supérieur à l'approche naïve consistant à prendre l'objet le plus proche (27,2 %). Ce score permet à la méthode **M-Saillance** d'être beaucoup plus performante que la méthode état de l'art **RankSVM** sur l'ensemble de la problématique d'identification de la cible d'une opinion (67.6 % contre 42,3 %). Ce score nous semble encore améliorable en introduisant d'autres types d'hypothèses sémantiques et discursives. Par exemple :

– si une cible A est évaluée fréquement positivement dans un document, si une cible B est évaluée fréquement négativement dans le même document, il est plus probable que A (respectivement B) soit reliée à une opinion positive (respectivement négative) dont la cible est inconnue ;
– si une ou plusieurs cibles appartenant au même graphe d'objets (modélisant les relations méronymiques entre objets) sont évaluées dans le même passage d'un document, il est probable qu'une opinion dont on ignore la cible porte sur un objet de ce graphe.

Ces hypothèses seraient intéressantes à comparer à des procédures de résolution d'anaphores nominales plus classiques pour le français. Néanmoins, ces procédures nous paraissent à ce stade encore complexes à mettre en place sur des corpus multithématiques et en particulier sur les blogs (dont le respect grammatical et orthographique est très variable). De ce point de vue, l'hypothèse de saillance d'un objet semble offrir une plus grande souplesse.

6 Conclusion

Nous avons abordé la problématique d'identification de la cible d'un passage d'opinion. Nos résultats montrent que la méthode état de l'art évaluée récemment sur des corpus anglophones monothématiques est adaptée pour un corpus français multithématique. Cette méthode, applicable au niveau intraphastrique uniquement, est utilisée comme amorce pour le développement de la nouvelle approche que nous proposons. Celle-ci est basée sur la saillance d'une cible d'opinion dans un segment textuel. Elle améliore l'état de l'art sur l'aspect intraphastique de la tâche. Elle permet également de couvrir davantage de problèmes en s'intéressant aux cas interphrastiques non traités dans le domaine jusqu'à présent. De ce point de vue nous améliorons significativement les résultats pour identifier la cible d'un passage d'opinion. Ces résultats nous semblent encore améliorables en prenant mieux en compte les relations sémantiques entre les objets d'un document.

Références

BACCIANELLA S., ESULI A. & SEBASTIANI F. (2010). Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

BENVENISTE E. (1966). *Problèmes de linguistique générale II*. Gallimard edition.

BETHARD S., YU H., THORNTON A., HATZIVASSILOGLOU V. & JURAFSKY D. (2004). Automatic extraction of opinion propositions and their holders. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

BLOOM K., GARG N. & ARGAMON S. (2007). Extracting appraisal expressions. In *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics*, p. 308–315, Rochester, New York : Association for Computational Linguistics.

CHOI Y., CARDIE C., RILOFF E. & PATWARDHAN S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*.

DUBREIL E., VERNIER M., MONCEAUX L. & DAILLE B. (2008). Annotating opinion - evaluation of blogs. In *Workshop on LREC 2008 Conference, Sentiment Analysis Metaphor, Ontology and Terminology (EMOT-08)*.

GREFENSTETTE G., QU Y., SHANAHAN J. G. & EVANS D. A. (2004). Coupling niche browsers and affect analysis for an opinion mining application. In *Proceedings of Recherche d'Information Assistée par Ordinateur*.

HU M. & LIU B. (2004). Mining opinion features in customer reviews. In *Proceedings of AAAI*, p. 755–760.

JAKOB N. & GUREVYCH I. (2010). Using anaphora resolution to improve opinion target identification in movie reviews. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, p. 263–268, Stroudsburg, PA, USA : Association for Computational Linguistics.

JOACHIMS T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD on Knowledge discovery and data mining*, KDD '02, p. 133–142, New York, NY, USA : ACM.

KESSLER J. S. & NICOLOV N. (2009). Targeting sentiment expressions through supervised ranking of linguistic configurations. In *3rd Int'l AAAI Conference on Weblogs and Social Media (ICWSM 2009)*.

KIM S.-M. & HOVY E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, p. 1–8, Stroudsburg, PA, USA : Association for Computational Linguistics.

MATHIEU Y. (2006). A computational semantic lexicon of french verbs of emotion. In W. B. CROFT, J. SHANAHAN, Y. QU & J. WIEBE, Eds., *Computing Attitude and Affect in Text : Theory and Applications*, volume 20 of *The Information Retrieval Series*, p. 109–124. Springer Netherlands.

MISHNE G. & GLANCE N. (2006). Predicting movie sales from blogger sentiment. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, p. 155–158.

PANG B. & LEE L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135.

RUPPENHOFER J., SOMASUNDARAN S. & WIEBE J. (2008). Finding the sources and targets of subjective expressions. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco : European Language Resources Association (ELRA). http//www.lrec-conf.org/proceedings/lrec2008/.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, p. 44–49.

TORRES-MORENO J. M., EL-BÈZE M., BÉCHET F. & CAMELIN N. (2007). Comment faire pour que l'opinion forgée à la sortie des urnes soit la bonne ? *Actes du 3ème Défi Fouille de Textes (AFIA 2007)*, p. 117–132.

TURNEY P. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. *Proceedings 40th Annual Meeting of the ACL (ACL'02)*, p. 417–424.

VERGNE J. & GIGUET E. (1998). Regards théoriques sur le «tagging». In *Actes de Traitement Automatique des Langues Naturelles (TALN'98)*, p. 22–31.

VERNIER M. & MONCEAUX L. (2010). Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques. *Traitement Automatique des Langues*, **51**(1), 125–149.

WEHRLI E. (2007). Fips, a deep linguistic multilingual parser. In *Proceedings of the Workshop on Deep Linguistic Processing*, DeepLP '07, p. 120–127, Stroudsburg, PA, USA : Association for Computational Linguistics.

WIEBE J. M. & RILOFF E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, number 3406 in Lecture Notes in Computer Science, p. 486–497.

WILSON T. A. (2008). *Fine-grained Subjectivity and Sentiment Analysis : Recognizing the Intensity, Polarity, and Attitudes of Private States*. PhD thesis, University of Pittsburgh.

TALN 2011, Montpellier, 27 *juin –*1*er juillet 2011*

Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français

Matthieu Constant[1] Isabelle Tellier[2] Denys Duchier[2]
Yoann Dupont[2] Anthony Sigogne[1] Sylvie Billot[2]
(1) Université Paris-Est, LIGM, CNRS, 5 bd Descartes, Champs-sur-Marne 77454 Marne-la-Vallée cedex 2
(2) LIFO, université d'Orléans, 6 rue Léonard de Vinci BP 6759, 45067 Orléans cedex 2
mconstan@univ-mlv.fr, isabelle.tellier@univ-orleans.fr, denys.duchier@univ-orleans.fr, yoann.dupont@etu.univ-orleans.fr, sigogne@univ-mlv.fr, sylvie.billot@univ-orleans.fr

Résumé. Dans cet article, nous synthétisons les résultats de plusieurs séries d'expériences réalisées à l'aide de CRF (Conditional Random Fields ou "champs markoviens conditionnels") linéaires pour apprendre à annoter des textes français à partir d'exemples, en exploitant diverses ressources linguistiques externes. Ces expériences ont porté sur l'étiquetage morphosyntaxique intégrant l'identification des unités polylexicales. Nous montrons que le modèle des CRF est capable d'intégrer des ressources lexicales riches en unités multi-mots de différentes manières et permet d'atteindre ainsi le meilleur taux de correction d'étiquetage actuel pour le français.

Abstract. In this paper, we synthesize different experiments using a linear CRF (Conditional Random Fields) to annotate French texts from examples, by exploiting external linguistic resources. These experiments especially dealt with part-of-speech tagging including multiword units identification. We show that CRF models allow to integrate, in different ways, large-coverage lexical resources including multiword units and reach state-of-the-art tagging results for French.

Mots-clés : Etiquetage morphosyntaxique, Modèle CRF, Ressources lexicales, Segmentation, Unités polylexicales.

Keywords: Part-of-speech tagging, CRF model, Lexical resources, Segmentation, Multiword units.

1 Introduction

Dans cet article, nous synthétisons les résultats de plusieurs séries d'expériences réalisées à l'aide de CRF (Conditional Random Fields ou "champs markoviens conditionnels" (Lafferty *et al.*, 2001; Tellier & Tommasi, 2011)) linéaires pour apprendre à annoter des textes français à partir d'exemples, en exploitant diverses ressources linguistiques externes. La tâche à laquelle nous nous sommes attachés est celle de la segmentation en unités lexicales des phrases d'un texte, couplée à celle de leur étiquetage en catégories morphosyntaxiques (ou "part of speech" en anglais).

Ces dernières années, l'étiquetage morphosyntaxique a atteint d'excellents niveaux de performance grâce à l'utilisation de modèles probabilistes discriminants comme les modèles de maximum d'entropie [MaxEnt] (Ratnaparkhi, 1996; Toutanova *et al.*, 2003), les séparateurs à vaste marge [SVM] (Giménez & Márquez., 2004) ou, déjà, les champs markoviens conditionnels [CRF] (Tsuruoka *et al.*, 2009). Il a par ailleurs été montré que le couplage de ces modèles avec des lexiques externes augmente encore la qualité de l'annotation, comme l'illustre (Denis & Sagot, 2009, 2010) pour MaxEnt. Néanmoins, les évaluations réalisées considèrent toujours en entrée un texte avec une segmentation lexicale parfaite, c'est-à-dire que les unités lexicales multi-mots, qui forment par définition des unités linguistiques, ont été parfaitement reconnues au préalable. Or cette tâche de segmentation est difficile car elle nécessite des ressources lexicales importantes. On notera que les systèmes tels que Macaon (Nasr *et al.*, 2010) et Unitex (Paumier, 2011) intègrent une analyse lexicale avec segmentation multi-mots ambiguë avant levée d'ambiguité par l'utilisation d'un modèle de Markov caché [HMM]. Dans cet article, nous proposons d'intégrer les deux tâches de segmentation et d'étiquetage dans un seul modèle CRF couplé à des ressources lexicales riches.

Le corpus d'apprentissage dont nous sommes partis provient du French Treebank (Abeillé *et al.*, 2003). Les ressources linguistiques externes utilisées sont de différentes natures. Nous avons ainsi exploité plusieurs dictionnaires : Lefff (Sagot, 2010) mais aussi DELA (Courtois, 2009; Courtois *et al.*, 1997), ainsi que des lexiques spécifiques comme Prolex (Piton *et al.*, 1999) et quelques autres incluant des noms d'organisation et des prénoms (Martineau *et al.*, 2009). Cet ensemble de dictionnaires est complété par une bibliothèque de grammaires locales qui reconnaissent différents types d'unités multi-mots (Constant & Watrin, 2008). Nous montrons que le modèle des CRF est capable d'intégrer de telles ressources de différentes manières et permet d'atteindre ainsi le meilleur taux actuel de correction pour la segmentation et l'étiquetage du français.

Dans la suite de cet article, nous commençons par présenter le modèle des CRF et le fonctionnement des bibliothèques logicielles que nous avons utilisées pour mener nos expériences. Nous décrivons ensuite le corpus d'apprentissage ainsi que la tâche que nous traitons, en détaillant les difficultés spécifiques que posent les unités multi-mots. Puis nous passons en revue les ressources à notre disposition et menons une réflexion méthodologique sur les différents moyens de les prendre en compte dans une chaîne de traitements qui fait appel à un CRF. La dernière partie est consacrée à la présentation des résultats de nos expériences. Ces travaux ont permis la mise au point de plusieurs segmenteurs-étiqueteurs qui sont librement disponibles.

2 Les CRF

2.1 Le modèle théorique

Les champs markoviens conditionnels ou CRF (Tellier & Tommasi, 2011) sont des modèles probabilistes discriminants introduits par (Lafferty *et al.*, 2001) pour l'annotation séquentielle. Ils ont été utilisés dans de nombreuses tâches de Traitement des Langues, où ils donnent d'excellents résultats (McCallum & Li, 2003; Sha & Pereira, 2003; Tsuruoka *et al.*, 2009; Tellier *et al.*, 2010).

Les CRF permettent d'associer à une observation x une annotation y en se basant sur un ensemble d'exemples étiquetés, c'est-à-dire un ensemble de couples (x, y). La plupart du temps (et ce sera le cas dans la suite de cet article), x est une *séquence d'unités* (ici, une suite d'unités lexicales) et y la *séquence des étiquettes correspondante* (ici, la suite de leurs catégories morphosyntaxiques, éventuellement enrichie pour coder la segmentation). Les CRF sont des modèles discriminants qui appartiennent à la famille des *modèles graphiques non dirigés*. Ils sont définis par X et Y, deux champs aléatoires décrivant respectivement chaque unité de l'observation x et son annotation y, et par un graphe $\mathcal{G} = (V, E)$ dont $V = X \cup Y$ est l'ensemble des nœuds (vertices) et $E \subseteq V \times V$ l'ensemble des

arcs (edges). Deux variables sont reliées dans le graphe si elles dépendent l'une de l'autre. Le graphe sur le champ Y des CRF linéaires, dessiné en Fig 1., traduit le fait que chaque étiquette est supposée dépendre de l'étiquette précédente et de la suivante et, implicitement, de la donnée x complète. Un dessin complet du graphe devrait ainsi également relier chaque variable Y_i à *chaque variable du champ X*, ce qu'on omet sur la figure pour la lisibilité.

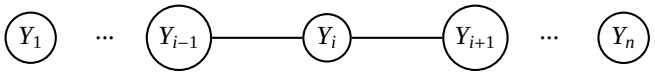


FIGURE 1 – graphe associé à un CRF linéaire

Dans un CRF, on a la relation suivante (Lafferty *et al.*, 2001) :

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in C} \exp\Big(\sum_k \lambda_k f_k(y_c, x, c) \Big) \quad avec$$

- C est l'ensemble des cliques (sous-graphes complètement connectés) de $\mathcal{G}$ sur Y : dans le cas du graphe de la Fig. 1, ces cliques sont constituées soit d'un nœud isolé, soit d'un couple de nœuds successifs.
- y_c l'ensemble des valeurs prises par les variables de Y sur la clique c pour un étiquetage y donné : ici, c'est donc soit la valeur d'une étiquette soit celles d'un couple d'étiquettes successives
- $Z(x)$ est un coefficient de normalisation, défini de telle sorte que la somme sur y de toutes les probabilités $p(y|x)$ pour une donnée x fixée soit égale à 1.
- Les fonctions f_k sont appelées *fonctions caractéristiques* (features) : elles sont définies à l'intérieur de chaque clique c et sont à valeurs réelles, mais souvent choisies pour donner un résultat binaire (0 ou 1). Elles doivent être founies au système par l'utilisateur. Par définition, la valeur de ces fonctions peut dépendre des étiquettes présentes dans une certaine clique c ainsi que de la valeur de *x n'importe où dans la donnée* (et pas uniquement aux indices correspondants à la clique c, ce qui donne beaucoup d'expressivité aux CRF).
- Les poids λ_k, à valeurs réelles, permettent d'accorder plus ou moins d'importance à chaque fonction f_k dont ils caractérisent le *pouvoir discriminant*. Ce sont les paramètres du modèle : l'enjeu de la phase d'apprentissage est de fixer leur valeur en cherchant à maximiser la log-vraisemblance sur un ensemble d'exemples déjà annotés (constituant le corpus d'apprentissage).

L'intérêt et l'efficacité des CRF proviennent de ce qu'ils prennent en compte des dépendances entre étiquettes reliées les unes aux autres dans le graphe. En cherchant le meilleur y, c'est-à-dire la meilleure *séquence d'étiquettes* associer à une donnée complète x, ils se comportent en général mieux qu'une série de classifications d'unités isolées. Mais cette prise en compte a un prix : la phase d'apprentissage d'un CRF peut être longue. Une fois cette phase réalisée, annoter une nouvelle séquence x de n mots en entrée revient alors à trouver le y qui maximise $p(y|x)$. L'espace théorique de recherche de ce meilleur étiquetage y est $|Y|^n$, où $|Y|$ est le nombre d'étiquettes distinctes possibles pour chaque nœud. Mais, grâce à des techniques de programmation dynamique, ce calcul peut être factorisé à l'intérieur des cliques et ramené à $K * n * |Y|^c$ où c est la taille de la plus grande clique ($c = 2$ pour les CRF linéaires) et K le nombre de fonctions caractéristiques. Une fois appris, l'étiqueteur est donc performant.

2.2 Les bibliothèques CRF++ et Wapiti

Notre objectif étant d'insérer des connaissances linguistiques dans un apprentissage réalisé à l'aide de CRF linéaires, il nous semble important de bien comprendre le fonctionnement concret des bibliothèques qui les implémentent. Plusieurs sont disponibles pour mettre en œuvre les CRF linéaires, notamment *crf.source.net* [1] de Sarawagi ou *Mallet* [2] de McCallum. Celles que nous avons utilisées sont *CRF++* [3] de Taku Kado et *Wapiti* [4] de Thomas Lavergne (Lavergne *et al.*, 2010), qui utilisent des moyens similaires pour instancier les fonctions caractéristiques qui entrent dans leur définition.

1. crf.sourceforge.net
2. http ://mallet.cs.umass.edu/
3. http ://crfpp.sourceforge.net/
4. http ://wapiti.limsi.fr

Corpus tabulaires. Les exemples d'apprentissage que requièrent ces bibliothèques sont des couples (x, y), où x est une séquence d'unités et y la séquence d'étiquettes correspondantes, de mêmes longueurs. Pour nous, une unité de x correspond à un "mot", mais elle peut être enrichie par d'autres propriétés, représentées par p attributs, du moment que ces derniers sont disponibles ou calculables aussi pour tout nouvel exemple x non étiqueté. Les attributs peuvent être des booléens (l'unité contient un chiffre, commence par une majuscule, est présente dans un lexique, etc.), des valeurs numériques (nombre de lettres, etc.) ou textuelles (valeur de l'unité ou de son préfixe ou suffixe de telle longueur, etc.). Une donnée étiquetée (x, y) de taille n se présente donc comme un tableau de n lignes et $p+1$ colonnes, où les p premières colonnes contiennent toutes les informations disponibles sur la donnée x et la dernière colonne les étiquettes y :

$$\begin{array}{ccccc} x_1^1 & x_1^2 & \cdots & x_1^p & y_1 \\ & & \vdots & & \\ x_i^1 & x_i^2 & \cdots & x_i^p & y_i \\ x_{i+1}^1 & x_{i+1}^2 & \cdots & x_{i+1}^p & y_{i+1} \\ & & \vdots & & \end{array}$$

Les exemples distincts sont séparés entre eux dans un même fichier par une ligne vide. Un corpus d'apprentissage est donc une suite de tels tableaux, tous de largeur $p + 1$, mais de hauteurs qui peuvent varier.

Patrons tabulaires. L'utilisateur des bibliothèques ne définit pas directement les fonctions caractéristiques du modèle ; il doit fournir des *patrons*. Il existe deux types de patrons correspondant aux deux tailles de clique possibles : les *unigrammes* pour les cliques de taille 1, et les *bigrammes* pour les cliques de taille 2.

Un patron unigramme est une sorte de carte perforée de même largeur $p + 1$ que nos tableaux, de hauteur quelconque sur les p premières colonnes mais ne pouvant capturer qu'une seule étiquette sur la colonne $p + 1$. Chaque position possible de cette carte sur un exemple définit une fonction caractéristique : celle qui renvoie la valeur 1 si la configuration de valeurs observée dans les perforations est satisfaite, 0 sinon. Les ronds dans le tableau précédent représentent les valeurs capturées par une telle carte, positionnée sur la ligne i d'une donnée. Chaque fonction caractéristique prend donc la forme d'une conjonction de critères booléens observée au moins une fois parmi les exemples et un patron en "génère" autant qu'il y a de positions où il peut s'appliquer dans le fichier d'exemples. Un patron permet de définir ainsi succinctement des milliers, voire des millions de fonctions caractéristiques. Un patron bigramme est similaire à un patron unigramme, mais on l'applique successivement à une position i, puis à la position suivante $i + 1$ et la fonction caractéristique obtenue est la conjonction de tous les critères rencontrés.

3 Corpus d'apprentissage pour la segmentation et l'étiquetage

3.1 Corpus FTB

Tout système d'annotation probabiliste supervisé requiert un corpus annoté de référence pour entraîner le modèle et ensuite l'évaluer. Pour notre tâche d'étiquetage morphosyntaxique intégrant la reconnaissance des unités multi-mots, il est donc nécessaire d'utiliser un corpus annoté en catégories grammaticales incluant l'annotation des unités polylexicales. Le corpus le plus complet en français est le corpus arboré de Paris 7 (Abeillé *et al.*, 2003), formé d'articles du journal *Le Monde* allant de 1989 à 1993. Il décrit la structure syntaxique des différentes phrases sous la forme d'arbres. Une unité de ce corpus peut être une ponctuation, un nombre, un mot simple ou une unité multi-mots. Au niveau morphosyntaxique, il existait initialement un jeu d'étiquettes de 14 catégories principales et de 34 sous-catégories. Pour notre tâche, nous utilisons un jeu d'étiquettes optimisé en 29 catégories pour l'analyse syntaxique (Crabbé & Candito, 2008) et réutilisé comme standard dans une expérience d'étiquetage morpho-syntaxique (Denis & Sagot, 2009). Les unités multi-mots codées sont de différents types : mots composés et entités nommées. Les mots composés comprennent des noms (*acquis sociaux*), des verbes (*faire face à*), des adverbes (*dans l'immédiat*), des prépositions (*en dehors de*). Il contient quelques types d'entités nommées : des noms d'organisation (*Société suisse de microélectronique et d'horlogerie*), des noms de famille (*Strauss-Kahn*), des noms de lieu (*Afrique du Sud*, *New York*).

Dans nos séries d'expériences, nous avons utilisé deux versions différentes du corpus : une version de 569 039 unités (au LIGM), une autre de 350 931 (au LIFO). Dans ces deux versions, nous n'avons repris que le niveau des feuilles, i.e. le niveau lexical. Nous en donnons un extrait ci-dessous :

```
Quant_à/P la/DET technique/NC ,/PONCT son/DET verdict/NC est/V implacable/ADJ ./PONCT
```

L'unité *Quant_à* est la fusion de deux mots simples (*Quant* et *à*), formant la préposition composée *quant_à*.

3.2 Unités lexicales multi-mots

Expressions multi-mots. Dans le consensus actuel du Traitement Automatique des Langues (TAL), les expressions multi-mots forment des unités linguistiques aux comportements lexicaux, syntaxiques et/ou sémantiques particuliers. Elles regroupent les expressions figées et semi-figées, les collocations, les entités nommées, les verbes à particule, les constructions à verbe support, les termes, etc. (Sag *et al.*, 2002). Leur identification est donc cruciale avant toute analyse sémantique. Elles apparaissent à différents niveaux de l'analyse linguistique : certaines forment des unités lexicales contigues à part entière (ex. *cordon bleu*, *San Francisco*, *par rapport à*), d'autres composent des constituants syntaxiques comme les phrases figées (*N0 prendre le taureau par les cornes* ; *N0 prendre N1 en compte*) ou les constructions à verbe support (*N0 donner un avertissement à N1* ; *N0 faire du bruit*).

Phénomènes traités. Dans cet article, nous ne traitons que les expressions multi-mots du niveau lexical, que nous appellerons dorénavant unités multi-mots ou polylexicales. Elles comportent les mots composés (noms, prépositions, adverbes, etc.), les entités nommées, les termes, les collocations nominales. Il existe une grande variété de phénomènes linguistiques rentrant dans cette catégorie et donc de nombreux critères d'identification. Les mots composés sont des séquences non compositionnelles de mots : ils présentent une opacité sémantique totale (*cordon bleu*, *tout à fait*) ou partielle (*vin blanc*), des contraintes syntaxiques et lexicales, etc. Il existe un continuum entre expressions figées et libres, ce qui rend leur identification encore plus difficile. Les collocations sont définies à partir de critères statistiques. Les entités nommées ont souvent une certaine compositionalité sémantique mais ont une syntaxe particulière : ex. *le 5 mars 2010* pour les dates, *Jacques Chirac* pour les noms de personnes.

Ressources. Les unités polylexicales peuvent être recensées dans des dictionnaires électroniques ou des grammaires locales. Les dictionnaires électroniques sont des listes qui associent des formes lexicales à des informations linguistiques comme les catégories grammaticales ou certains traits sémantiques (ex. *humain*, *concret*, etc.). Les grammaires locales (Gross, 1997; Silberztein, 2000) sont des réseaux récursifs de transitions décrits sous la forme de graphes d'automates finis. Chaque transition est étiquetée par un élément lexical (ex. *mange*), un masque lexical correspondant à un ensemble de formes lexicales encodées dans un dictionnaire (ex. <*manger*> symbolisant toutes les formes fléchies dont le lemme est *manger*) ou un élément non-terminal référant à un autre automate. Elles sont très utiles pour décrire de manière compacte des unités multi-mots acceptant des variations lexicales. Un système de transduction permet d'annoter les expressions décrites, comme la catégorie grammaticale ou l'analyse des composants internes pour les entités nommées par exemple (Martineau *et al.*, 2009).

Reconnaissance. La reconnaissance automatique des unités multi-mots est, la plupart du temps, réalisée à l'aide de ressources lexicales construites manuellement (ex. pour les expressions figées) ou apprises automatiquement (ex. collocations nominales). Par ailleurs, une grande partie des entités nommées, du fait de leur syntaxe particulière sont facilement décrites et reconnues à l'aide de grammaires locales (Friburger & Maurel, 2009; Martineau *et al.*, 2009), bien qu'il existe d'autres types d'approches telles que les systèmes statistiques (McCallum & Li, 2003) ou hybrides (Poibeau, 2009). L'identification de telles expressions est une tâche très difficile car les unités non décrites dans les ressources sont difficilement reconnaissables. Elle est d'autant plus difficile qu'elle dépend du contexte d'occurrence. En effet, une expression reconnue est souvent ambigue avec l'analyse en mots simples : par exemple, *il* ***en fait*** *une priorité* (mots simples) vs *j'ai* ***en fait*** *beaucoup travaillé* (mot composé). On observe parfois des chevauchements avec d'autres unités polylexicales comme dans la séquence *une pomme de terre cuite* où *pomme de terre* et *terre cuite* sont des mots composés. C'est pourquoi les outils existants de segmentation en unités multi-mots comme dans INTEX (Silberztein, 2000) ou SxPipe (Sagot & Boullier, 2008) produisent une segmentation ambiguë sous la forme d'automates finis acycliques pour éviter de prendre une décision définitive

trop hâtive. Cette analyse ambiguë peut alors être intégrée dans des traitements linguistiques tels que l'étiquetage morphosyntaxique (Nasr *et al.*, 2010; Paumier, 2011)) ou l'analyse syntaxique superficielle (Blanc *et al.*, 2007; Nasr *et al.*, 2010) et profonde (Sagot & Boullier, 2006).

3.3 Intégration d'un segmenteur et d'un étiqueteur

L'identification des unités multi-mots est similaire à une tâche de segmentation comme le chunking ou à la reconnaissance des entités nommées, qui identifient les limites de segments (chunks ou entités nommées) et les annotent. En effet, grâce à la représentation `IOB` [5] (Ramshaw & Marcus, 1995), segmenter un texte revient à annoter ses unités minimales. Pour combiner étiquetage morphosyntaxique et reconnaissance d'unités multi-mots, il suffit de concaténer les deux étiquetages en associant à chaque unité minimale une étiquette de la forme `X+B` ou `X+I`, où `X` est sa catégorie grammaticale et le suffixe indique si elle se trouve au début d'une unité multi-mots (`B`) ou dans une position "interne" (`I`). Le suffixe `O` est inutile car la fin d'un segment lexical correspond au début d'un autre (suffixe `B`) ou à une fin de phrase. Une telle procédure d'annotation détermine non seulement les limites des unités lexicales, mais aussi leur catégorie morphosyntaxique. Pour entraîner nos CRF, nous avons donc transformé le corpus d'apprentissage initial en isolant les unités composant les segments multi-mots et en les étiquetant conformément à cette nouvelle norme. L'exemple précédent est alors transformé en :

```
Quant/P+B à/P+I la/DET+B technique/NC+B ,/PONCT+B son/DET+B verdict/NC+B est/V+B
implacable/ADJ+B ./PONCT+B
```

Le jeu d'étiquettes initial est ainsi doublé, chaque étiquette se dédoublant en une variante `B` et une variante `I`. La reconnaissance des unités polylexicales dépendant fortement de la richesse de ressources lexicales utilisées, il s'agit maintenant de trouver les meilleures façons d'intégrer ce type d'informations dans nos CRF.

4 Exploitation d'une ressource externe

Dans cette section, nous commençons par présenter les différentes ressources que nous avons à notre disposition, et nous cherchons tous les moyens possibles de les prendre en compte dans un apprentissage avec des CRF.

4.1 Ressources

Même s'il existe de plus en plus d'études sur l'extraction automatique d'unités multi-mots, en particulier les collocations ou les termes (Daille, 1995; Dias, 2003; Seretan *et al.*, 2003), les ressources les plus riches et les plus précises ont été aquises manuellement. Pour notre étude, nous avons compilé diverses ressources lexicales sous la forme de dictionnaires morphosyntaxiques et de grammaires locales fortement lexicalisées. Nous avons utilisé notamment deux dictionnaires disponibles de mots simples et composés de la langue générale : DELA (Courtois, 2009; Courtois *et al.*, 1997) et Lefff (Sagot, 2010). Le DELA a été construit par une équipe de linguistes. Le Lefff a été automatiquement acquis et manuellement validé. Il résulte également de la fusion de différentes sources lexicales. En complément, nous disposons aussi de lexiques spécifiques comme Prolex (Piton *et al.*, 1999) composé de toponymes et d'autres incluant des noms d'organisation et des prénoms (Martineau *et al.*, 2009). Les nombres d'entrées de ces divers dictionnaires sont donnés dans le tableau 1.

Dictionnaire	#mots simples	#mots composés
DELA	690,619	272,226
Lefff	553,140	26,311
Prolex	25,190	97,925
Organisations	772	587
Prénoms	22,074	2,220

TABLE 1 – Dictionnaires morphosyntaxiques

5. I : Inside (intérieur du segment) ; O : Outside (hors du segment) ; B : Beginning (début du segment)

Cet ensemble de dictionnaires est complété par une bibliothèque de grammaires locales qui reconnaissent différents types d'unités multi-mots comme les entités nommées (dates, noms d'organisation, de personne et de lieu), prépositions locatives, déterminants numériques et nominaux. En pratique, nous avons utilisé une bibliothèque de 211 automates développée à partir de la bibliothèque en-ligne GraalWeb (Constant & Watrin, 2008).

4.2 Quelques statistiques préliminaires

Pour les expériences menées avec la variante du FTB la plus volumineuse, le corpus initial a été découpé en trois parties : 80% pour la phase d'entraînement (TRAIN), 10% pour le développement (DEV) et 10% pour le test. Cela nous a permis de faire quelques observations préalables.

Ainsi, dans le corpus FTB-DEV (avec étiquetage initial non transformé), nous avons observé qu'environ 97,4% des unités lexicales [6] sont présentes dans nos ressources lexicales (en particulier, 97% sont présentes dans les dictionnaires). Alors que 5% des unités sont inconnues (i.e. absentes du corpus d'apprentissage), 1,5% sont à fois inconnues et absentes des ressources lexicales, ce qui montre que 70% des unités inconnues sont couvertes par nos ressources. On observe également qu'environ 6% des unités sont multi-mots. En décomposant toutes les unités multi-mots du texte en unités minimales, on s'aperçoit qu'à peu près 15% d'entre elles sont incluses dans une unité multi-mots. Parmi les unités multi-mots codées dans le corpus FTB-DEV, 75,5% d'entre elles sont présentes dans nos ressources (87,5% en incluant le lexique du corpus d'entraînement). Ceci montre que 12,5% des unités multi-mots sont totalement inconnues et, par conséquent, seront sans doute très difficilement reconnaissables.

On observe, par ailleurs, que le corpus FTB ne couvre pas la reconnaissance de toutes les unités multi-mots. Tout d'abord, certains déterminants ou certaines entités nommées ne sont pas identifiés comme les déterminants nominaux, les dates, les noms de personne, les adresses postales. Par ailleurs, de nombreux noms composés sont manquants. Par exemple, après avoir appliqué nos ressources lexicales de manière non contextuelle (en excluant les grammaires locales reconnaissants des types d'entités nommées ou des déterminants nominaux non codés dans le FTB), nous avons manuellement observé sur le FTB-DEV qu'environ 30% des unités polylexicales de nos ressources "adaptées" ne sont pas prises en compte dans le corpus.

4.3 Méthodologie de prise en compte des ressources

Comment prendre en compte une ou plusieurs ressources lors d'une chaîne de traitements faisant appel à un apprentissage réalisé avec un CRF ? Dans le cadre de l'apprentissage de la ressource MElt$_{fr}$ (Denis & Sagot, 2009, 2010), les auteurs ont testé deux approches possibles :
– intégrer les propriétés des mots du lexique dans les fonctions caractéristiques du modèle d'apprentissage ;
– filtrer les étiquetages incompatibles avec les informations présentes dans la ressource.
Nous avons cherché toutes les façons possibles d'envisager cette intégration, ce qui nous a amené à en caractériser plus finement le mode opératoire, et à en trouver de nouvelles variantes. Nous les présentons ci-dessous, en discutant leurs intérêts et leurs limites. Elles peuvent s'organiser en deux familles principales, suivant que la ou les ressources disponibles sont mises à contribution comme des filtres *avant ou après* l'appel au CRF ou qu'elles sont utilisées *pendant la phase d'apprentissage*. L'approche "filtrage" requiert que les étiquettes qui figurent dans la ressource soient identiques à celles qui sont la cible de l'apprentissage, alors que ce n'est pas nécessairement le cas pour l'autre approche. Au cas où les conventions d'étiquetage ne sont pas les mêmes, une fonction de correspondance doit être préalablement appliquée.

Les ressources comme filtrage *a priori* ou *a posteriori* Les ressources peuvent être vues comme un moyen de contraindre, ou encore de *filtrer* les étiquetages possibles. Concrètement, ce filtrage peut opérer *avant* ou *après* l'appel au CRF. Le filtrage *a priori* consiste à définir l'espace de recherche des étiquetages possibles y d'une nouvelle chaîne x via un prétraitement fondé sur une ressource. Les analyseurs lexicaux actuels auxquels on soumet une phrase produisent en effet généralement un *dag* (graphe orienté acyclique) dont chaque chemin correspond à une séquence possible d'étiquettes. Les unités multi-mots peuvent être reconnues lors de cette étape, et figurer aussi dans le *dag*, comme cela a été évoqué section 3.2. Pour une phrase constituée de n unités minimales, il est évidemment plus facile et rapide de chercher le y qui maximise $p(y|x)$ (calculé suivant la formule des CRF) parmi

6. Les unités lexicales sont les unités autres que les nombres et les ponctuations.

l'ensemble des étiquetages du *dag* plutôt que sur l'espace de tous les $|Y|^n$ étiquetages possibles. Le filtrage est ainsi *a priori* mais l'apprentissage du CRF est néanmoins un pré-requis de la chaîne de traitements. Le filtrage *a posteriori*, lui, cherche non pas le meilleur étiquetage possible *y* d'une chaîne quelconque *x* mais les *m* meilleurs possibles (c'est une option généralement disponibles des bibliothèques CRF) et choisit le premier d'entre eux compatible avec la ressource. Les deux techniques donnent la même solution ; privilégier l'une ou l'autre dépend de la forme de la ressource. Leur intérêt est de garantir que dans la solution retenue, chaque mot reçoit une étiquette compatible avec ce que décrivent la ou les ressources consultées. Un filtrage peut d'ailleurs très bien se combiner avec une approche prenant en compte les ressources *pendant* la phase d'apprentissage.

Les ressources comme aide à l'apprentissage. D'après la section 2.2, quand nous faisons appel à une bibliothèque qui implémente les CRF linéaires, nous avons à notre disposition trois "leviers" d'action possibles :
– le choix des étiquettes et des propriétés des unités (les colonnes des données tabulaires)
– le choix des exemples (les lignes)
– le choix des fonctions caractéristiques (via les patrons), choix qui dépend fortement des précédents
Nous avons déjà vu qu'un choix pertinent d'étiquettes permettait de "coder" en quelque sorte les deux problèmes de la segmentation et de l'étiquetage simultanément. D'autres expériences ont montré l'intérêt de décomposer le jeu d'étiquettes en sous-étiquettes, notamment quand celles-ci sont trop nombreuses (Tellier *et al.*, 2010). Mais le problème auquel nous nous confrontons ici ne requiert pas un tel traitement, nous ne l'avons pas mis en œuvre.

Il est en revanche "naturel" d'insérer les informations des ressources en tant que propriétés des unités d'un exemple *x*, donc en jouant sur les colonnes $x_i^2, ..., x_i^p$. Plusieurs choix sont encore possibles pour cela, suivant qu'on se contente de concaténer les différentes étiquettes possibles d'une même unité pour en faire une seule colonne de nature textuelle, ou bien qu'on définisse autant de colonnes à valeur booléenne que d'étiquettes possibles dans l'ensemble de la ressource. Cela aura bien sûr des conséquences sur la définition des patrons qui génèrent les fonctions caractéristiques. Dans le cas des colonnes booléennes, la combinatoire des conjonctions possibles de plusieurs critères est explosive. Dans les deux cas, on peut soit garder les étiquettes des ressources telles quelles, soit les transformer pour qu'elles s'identifient à celles visées.

Enfin, il est aussi possible de considérer que chaque instance de couple (unité lexicale, étiquette) présent dans la ressource constitue à elle toute seule une "phrase" qu'on insère parmi les exemples étiquetés, en ajoutant de nouvelles lignes isolées dans le corpus d'apprentissage. Cela suppose bien sûr que les étiquettes qui figurent dans la ressource sont identiques à celles de l'étiquetage cible. L'idée sous-jacente de cette technique, très simple à appliquer, est que la présence dans une ressource équivaut à une occurrence attestée dans la langue, que l'on simule en l'insérant artificiellement dans le corpus d'apprentissage. Elle présente aussi toutefois quelques inconvénients :
– on introduit ainsi un biais sur les comptes d'occurrences puisque les différentes étiquettes possibles d'une unité donnent chacune lieu à un exemple, comme si elles étaient équiprobables. Il faut donc espérer que le reste de l'ensemble d'apprentissage soit suffisant pour compenser cette distorsion possible.
– en introduisant des "phrases" réduites à un mot, on va rendre inopérantes sur ces "phrases" particulières toutes les fonctions caractéristiques qui testent la valeur des unités ou des étiquettes voisins (et donc en particulier tous les bigrammes). Le poids de ces fonctions ne pourra être calculé que sur le reste des exemples.

5 Résultats des expériences

Les résultats présentés ici sont issus d'expériences menées en parallèle au LIFO (Orléans) et au LIGM (Paris-Est Marne-la-Vallée). Notons au préalable que les expériences ont été réalisées dans des environnements différents, sans coordination *a priori*, ce qui explique la difficulté à comparer précisément les résultats. Les expériences du LIFO ont été menées avec Wapiti [7] et évaluées par validation croisée en 10 parties : 9/10 pour l'apprentissage, 1/10 pour le test. Celles du LIGM ont utilisé CRF++ [8] et porté sur une variante du FTB plus volumineuse rendant plus coûteuse, mais aussi moins indispensable, une validation croisée : le corpus initial a alors été découpé en trois parties : 80% pour la phase d'entraînement (TRAIN), 10% pour le développement (DEV) et 10% pour le test.

Pour l'évaluation globale, nous avons précision = rappel = f-mesure. En effet, tous les mots ayant une unique étiquette, une erreur de précision sur une classe C1 correspond à une erreur de rappel sur une classe C2 et vice versa.

7. Ce programme a l'avantage d'opérer une sélection des fonctions caractéristiques *en cours d'apprentissage* grâce à une pénalisation L1.
8. L'algorithme de régularisation utilisé est L2 et le seuil de fréquence des traits a été fixé à 2.

Pour l'étiquetage avec segmentation, nous avons deux types d'évaluation : la f-mesure sur les unités minimales (LIFO) et sur les segments lexicaux (LIGM). Ceci explique les scores plus élevés pour LIFO sur cette tâche.

5.1 Evaluation de l'étiquetage avec segmentation parfaite

LIGM. Nous avons tout d'abord évalué l'étiquetage morphosyntaxique sur une segmentation multi-mots parfaite, au moyen d'un modèle CRF appris en utilisant des propriétés classiques des unités (forme lexicale, préfixes, suffixes, commence par une majuscule, etc.). Les expériences du LIGM ont porté sur deux méthodes d'intégration de la ressource lexicale externe décrite dans la section 4.1. La première méthode consiste à introduire, dans le fichier d'entraînement, une colonne supplémentaire (AC) représentant la concaténation des étiquettes trouvées dans la ressource pour l'unité courante. Nous obtenons alors un modèle LEX en utilisant tous les traits décrits dans la table 1(a). Nous notons STD le modèle incorporant les mêmes traits à l'exception de ceux issus de la ressource. La deuxième méthode consiste à procéder à un filtrage a priori de toutes les étiquettes absentes de la ressource pour chaque unité. Si l'unité est absente, toutes les étiquettes sont gardées. Les étiquettes des ressources ont été ajustées à celles du corpus pour le filtrage. Nous avons comparé les résultats avec d'autres outils d'étiquetage que nous avons tous entrainés sur le corpus FTB-TRAIN. Nous avons évalué TreeTagger (Schmid, 1994) basé sur des arbres de décision probabilistiques, SVMTool (Giménez & Márquez., 2004) basé sur les Séparateurs à Vastes Marges utilisant des traits indépendants de la langue, MElt (Denis & Sagot, 2009) basé sur un modèle MaxEnt incorporant en plus des traits dépendants de la langue issus de lexiques externes. Les lexiques utilisés pour entraîner et tester MElt intègrent toutes les ressources de la section 4.1 [9]. Les précisions obtenues sur le corpus FTB-TEST pour les différents systèmes sont données en pourcentage dans la table 1(b) avec un intervalle de confiance à 95% de +/-0,1.

(a) Types de traits

Traits internes unigrammes	
$w_0 = X$	$\&t_0 = T$
forme en minuscule de $w_0 = L$	$\&t_0 = T$
Préfixe de $w_0 = P$ avec $\|P\| < 5$	$\&t_0 = T$
Suffixe de $w_0 = S$ avec $\|S\| < 5$	$\&t_0 = T$
w_0 contient un tiret	$\&t_0 = T$
w_0 contient un chiffre	$\&t_0 = T$
w_0 commence par une majuscule	$\&t_0 = T$
w_0 est tout en majuscule	$\&t_0 = T$
w_0 commence par une majuscule et est en début de phrase	$\&t_0 = T$
classe d'ambiguité de w_0, $AC_0 = A$	$\&t_0 = T$
Traits contextuels unigrammes	
$w_i = X$, $i \in \{-2,-1,1,2\}$	$\&t_0 = T$
$w_i w_j = XY$, $(j,k) \in \{(-1,0),(0,1),(-1,1)\}$	$\&t_0 = T$
$AC_i = A$, $i \in \{-2,-1,1,2\}$	$\&t_0 = T$
Traits bigrammes	
$t_{-1} = T'$	$\&t_0 = T$

(b) Comparaison de systèmes d'étiquetage pour le français

	sans filtrage	avec filtrage
TreeTagger	96.4	-
SVMTool	97.2	-
MElt	97.6	-
CRF-STD	97.4	97.6
CRF-LEX	**97.7**	**97.7**

TABLE 2 – Résultats du LIGM avec segmentation parfaite

LIFO. Les expériences du LIFO ont porté sur une version du FTB moins volumineuse, en utilisant les traits unigrammes décrits dans la Table 3(a) sur une fenêtre $[-2..2]$ et les simples valeurs d'étiquettes pour les bigrammes. Les patrons bigrammes produisent en effet un très grand nombre de fonctions caractéristiques : cette restriction est destinée à limiter les calculs. La seule ressource à notre disposition était le Lefff. La première méthode utilisée pour le prendre en compte en cours d'apprentissage est de l'intégrer en tant que pourvoyeur de nouveaux exemples dans chaque fichier d'entraînement. Cette méthode augmente le temps d'apprentissage du simple au double voire triple selon les parties. La seconde méthode consiste à introduire des booléens en tant qu'attributs dans les colonnes des fichiers d'entraînement, chaque colonne représentant une étiquette possible dans le Lefff. Il a fallu alors générer par programme tous les patrons possibles qui combinent certains attributs entre eux. Cette méthode produit un grand nombre de fonctions caractéristiques mais Wapiti est capable de les gérer puisqu'il opère une sélection des fonctions caractéristiques les plus discriminantes *en cours d'apprentissage* (Lavergne *et al.*, 2010).

9. Nous avons regroupé ensemble tous les dictionnaires, ainsi que les unités reconnues lors de l'application des grammaires locales sur le corpus.

(a) Types de traits unigrammes

Valeur de l'unité
Commence par une majuscule
Est uniquement en majuscules
Est un chiffre
Est une ponctuation
3 dernières lettres

(b) Résultats

Sans lefff	96.5
Avec lefff (exemples)	96.6
Avec lefff (attributs booléens)	97.3

TABLE 3 – Résultats du LIFO avec segmentation parfaite

5.2 Evaluation de l'étiquetage avec identification des unités multi-mots

LIGM. Pour évaluer la tâche d'étiquetage intégrant la reconnaissance des unités multi-mots, nous avons entraîné trois modèles CRF sur le corpus FTB-TRAIN après avoir décomposé les unités multi-mots en séquences d'unités minimales étiquetées dans la représentation de type IOB (cf. section 3.3) : `STD`, `LEX` et `MWE`. Les deux premiers ont les mêmes types de traits que dans l'expérience précédente. Le modèle MWE est complété de traits issus de l'application non-contextuelle de nos ressources multi-mots sur le texte : une unité est associée à la catégorie grammaticale, la structure interne ou/et le trait sémantique de l'unité polylexicale reconnue à laquelle elle appartient, ainsi qu'à sa position relative dans l'unité (I, O ou B). Par exemple, le mot *de* dans le contexte du mot composé *eau de vie* présent dans nos ressources, sera associé à la catégorie grammaticale NC (nom), à la structure interne NPN (nom+préposition+nom) et à la position relative I (car il est en 2ème position). Ces trois systèmes ont été comparés avec SVMTool, entraîné sur le même corpus. Pour chaque segmenteur-étiqueteur appliqué sur le corpus TEST décomposé en unités minimales, nous avons calculé la f-mesure [10]. La précision et le rappel sont calculés par rapport aux segments lexicaux trouvés et non aux unités minimales simples. Les résultats sont synthétisés dans le tableau 3(a). La colonne `SEG` indique la f-mesure de la segmentation qui ne prend en compte que les limites des segments. La colonne `TAG` prend aussi en compte la catégorie grammaticale.

LIFO. Pour cette tâche, nous comparons les résultats obtenus (1) sans le Lefff, (2) avec le Lefff comme source d'exemples, (3) avec le Lefff comme source d'attributs booléens. Nos résultats évaluent la qualité de l'étiquetage des unités minimales avec les catégories intégrant B et I, et non celle de l'identification des unités multimots.

(a) LIGM (f-mesure sur les segments lexicaux)

	TAG	SEG
SVMTool	92.1	94.7
CRF-STD	93,7	95.8
CRF-LEX	93.9	95.9
CRF-MWE	**94.4**	**96.4**

(b) LIFO : méthodes d'intégration (f-mesure sur les unités minimales)

Sans lefff	94.5%
Avec lefff (exemples)	94.7%
Avec lefff (attributs)	95.2%

TABLE 4 – Apprentissage simultané étiquetage/segmentation

5.3 Description des segmenteurs-étiqueteurs proposés

Les diverses expériences décrites ci-dessus ont mené à la mise au point de segmenteurs-étiqueteurs qui sont librement disponibles. La chaîne de traitements de SEM [11] produite au LIFO a été écrite en Python. Le programme offre la possibilité d'exploiter ou non Lefff (sous forme d'attributs uniquement), en utilisant soit une segmentation rudimentaire écrite à la main (sans prise en compte de ressources externes), soit la segmentation acquise par le CRF. Le segmenteur-étiqueteur LGTagger [12] produit au LIGM est implanté en Java et comprend deux phases distinctes : (1) une analyse lexicale basée sur des ressources lexicales externes qui sert à filtrer les analyses (simples ou multi-mots) non décrites dans les ressources et qui produit un *dag* [13] ; (2) un décodeur qui détermine le

10. La formule de la f-mesure est la suivante : $f = \frac{2pr}{p+r}$ où p est la précision et r le rappel.
11. http ://www.univ-orleans.fr/lifo/Members/Isabelle.Tellier/SEM.html
12. http ://igm.univ-mlv.fr/~mconstan/research/software
13. Pour les mots simples inconnus de nos ressources, toutes les étiquettes possibles sont gardées comme candidates. Si l'analyseur n'a aucune ressource lexicale en entrée, il produit un *dag* représentant toutes les analyses possibles dans le jeu d'étiquettes.

chemin du *dag* le plus probable en fonction du modèle CRF appris. Il peut être exécuté avec ou sans segmentation multi-mots. Les ressources lexicales (pour le calcul des propriétés des fonctions caractéristiques et pour l'analyse lexicale) lui sont passées en paramètres. Les programmes d'Unitex (Paumier, 2011) sont utilisés pour l'application des ressources : consultation des dictionnaires et application des grammaires locales.

6 Conclusion

Dans cet article, nous avons montré que les tâches de segmentation et d'étiquetage sont intimement liées et qu'il est naturel de les traiter simultanément. L'écart entre la performance de l'étiquetage avec ou sans segmentation est de 2 à 4 points suivant la mesure utilisée : cela mesure le "coût" d'une bonne segmentation. Par ailleurs, nous avons montré l'intérêt certain d'intégrer des ressources lexicales dans un CRF, en particulier les ressources d'unités polylexicales utiles pour la segmentation. Nous voyons aussi qu'à ce niveau de performance, il est extrêmement difficile de gagner quelques dixièmes de points, même en mettant en jeu des ressources riches et variées.

Cet article a aussi été l'occasion d'une réflexion méthodologique poussée sur les différents moyens d'intégrer une ressource linguistique externe dans une chaîne de traitements faisant appel à un CRF. Une bonne partie de cette réflexion est d'ailleurs transposable à l'utilisation d'autres techniques d'apprentissage automatique. Les CRF, en intégrant fonctions caractéristiques locales et combinaison statistique globale, apparaissent comme un modèle particulièrement bien adapté à l'hybridation entre ressources symboliques et modèles statistiques. Grâce à cette intégration, il a été possible de produire en peu de temps des segmenteurs-étiqueteurs très performants.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In A. ABEILLÉ, Ed., *Treebanks*. Dordrecht : Kluwer.

BLANC O., CONSTANT M. & WATRIN P. (2007). Segmentation in super-chunks with a finite-state approach. In *Proceedings of the 6th Workshop on Finite-State Methods and Natural Language Processing (FSMNLP'07)*, p. 62 – 73.

CONSTANT M. & WATRIN P. (2008). Networking multiword units. In *Proceedings of the 6th International Conference on Natural Language Processing (GoTAL'08)*, number 5221 in Lecture Notes in Artificial Intelligence, p. 120 – 125 : Springer-Verlag.

COURTOIS B. (2009). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, **87**, 1941 – 1947.

COURTOIS B., GARRIGUES M., GROSS G., GROSS M., JUNG R., MATHIEU-COLAS M., MONCEAUX A., PONCET-MONTANGE A., SILBERZTEIN M. & VIVÉS R. (1997). *Dictionnaire électronique DELAC : les mots composés binaires*. Rapport interne 56, University Paris 7, LADL.

CRABBÉ B. & CANDITO M. H. (2008). Expériences d'analyse syntaxique statistique du français. In *Actes de TALN 2008 (Traitement automatique des langues naturelles)*, Avignon.

DAILLE B. (1995). Repérage et extraction de terminologie par une approche mixte statistique et linguistique. *traitement Automatique des Langues (TAL)*, **36**(1-2), 101–118.

DENIS P. & SAGOT B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 2009)*.

DENIS P. & SAGOT B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du francais. In *actes de TALN 2010*.

DIAS G. (2003). Multiword unit hybrid extraction. In *Proceedings of the Workshop on Multiword Expressions of the 41st Annual Meeting of the Association of Computational Linguistics (ACL 2003)*, p. 41–49.

FRIBURGER N. & MAUREL D. (2009). Finite-state transducer cascade to extract named entities in texts. *Theoretical Computer Science*, **313**, 94–104.

GIMÉNEZ J. & MÁRQUEZ. L. (2004). Svmtool : A general pos tagger generator based on support vector machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

Gross M. (1997). The construction of local grammars. In D. J. Lipcoll, D. H. Lawrie & A. H. Sameh, Eds., *Finite-State Language Processing*, p. 329–352. Cambridge, Mass. : The MIT Press.

Lafferty J., McCallum A. & Pereira F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, p. 282–289.

Lavergne T., Cappé O. & Yvon F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

Martineau C., Nakamura T., Varga L. & Voyatzi S. (2009). Annotation et normalisation des entités nommées. *Arena Romanistica*, **4**, 234–243.

McCallum A. & Li W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of CoNLL*.

Nasr A., Béchet F. & Rey J. F. (2010). Macaon : Une chaîne linguistique pour le traitement de graphes de mots. In *Traitement Automatique des Langues Naturelles - session de démonstrations*, Montréal.

Paumier S. (2011). Unitex 2.1 - user manual. http ://igm.univ-mlv.fr/˜unitex.

Piton O., Maurel D. & Belleil C. (1999). The prolex data base : Toponyms and gentiles for nlp. In *Proceedings of the Third International Workshop on Applications of Natural Language to Data Bases (NLDB'99)*, p. 233–237.

Poibeau T. (2009). Boosting Robustness of a Named Entity Recognizer. *International Journal of Semantic Computing*, **3**(1), 1–14.

Ramshaw L. A. & Marcus M. P. (1995). Text chunking using transformation-based learning. In *Proceedings of the 3rd Workshop on Very Large Corpora*, p. 88 – 94.

Ratnaparkhi A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 1996)*, p. 133 – 142.

Sag I. A., Baldwin T., Bond F., Copestake A. A. & Flickinger D. (2002). Multiword expressions : A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '02)*, p. 1–15, London, UK : Springer-Verlag.

Sagot B. (2010). The lefff, a freely available, accurate and large-coverage lexicon for french. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

Sagot B. & Boullier P. (2006). Deep non-probabilistic parsing of large corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*.

Sagot B. & Boullier P. (2008). Sxpipe 2 : architecture pour le traitement pré-syntaxique de corpus bruts. *Traitement Automatique des Langues*, **49**(2), 155–188.

Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, p. 252 – 259.

Seretan V., Nerima L. & Wehrli E. (2003). Extraction of multi-word collocations using syntactic bigram composition. In *Proceedings of the Fourth International Conference on Recent Advances in NLP (RANLP-2003)*, p. 424–431, Borovets, Bulgaria.

Sha F. & Pereira F. (2003). Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL*, p. 213 – 220.

Silberztein M. (2000). Intex : an fst toolbox. *Theoretical Computer Science*, **231**(1), 33–46.

Tellier I., Eshkol I., Taalab S. & Prost J. P. (2010). Pos-tagging for oral texts with crf and category decomposition. *Research in Computing Science*, **46**, 79–90. Special issue "Natural Language Processing and its Applications".

Tellier I. & Tommasi M. (2011). Champs Markoviens Conditionnels pour l'extraction d'information. In Eric Gaussier & François Yvon, Eds., *Modèles probabilistes pour l'accès à l'information textuelle*. Hermès.

Toutanova K., Klein D., Manning C. D. & Singer Y. Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, p. 252 – 259.

Tsuruoka Y., Tsujii J. & Ananiadou S. (2009). Fast full parsing by linear-chain conditional random fields. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, p. 790–798.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Segmentation et induction de lexique non-supervisées du mandarin

Pierre Magistry Benoît Sagot
Alpage, INRIA Paris–Rocquencourt & Université Paris 7,
Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
{pierre.magistry,benoit.sagot}@inria.fr

Résumé. Pour la plupart des langues utilisant l'alphabet latin, le découpage d'un texte selon les espaces et les symboles de ponctuation est une bonne approximation d'un découpage en unités lexicales. Bien que cette approximation cache de nombreuses difficultés, elles sont sans comparaison avec celles que l'on rencontre lorsque l'on veut traiter des langues qui, comme le chinois mandarin, n'utilisent pas l'espace. Un grand nombre de systèmes de segmentation ont été proposés parmi lesquels certains adoptent une approche non-supervisée motivée linguistiquement. Cependant les méthodes d'évaluation communément utilisées ne rendent pas compte de toutes les propriétés de tels systèmes. Dans cet article, nous montrons qu'un modèle simple qui repose sur une reformulation en termes d'entropie d'une hypothèse indépendante de la langue énoncée par Harris (1955), permet de segmenter un corpus et d'en extraire un lexique. Testé sur le corpus de l'Academia Sinica, notre système permet l'induction d'une segmentation et d'un lexique qui ont de bonnes propriétés intrinsèques et dont les caractéristiques sont similaires à celles du lexique sous-jacent au corpus segmenté manuellement. De plus, on constate une certaine corrélation entre les résultats du modèle de segmentation et les structures syntaxiques fournies par une sous-partie arborée corpus.

Abstract. For most languages using the Latin alphabet, tokenizing a text on spaces and punctuation marks is a good approximation of a segmentation into lexical units. Although this approximation hides many difficulties, they do not compare with those arising when dealing with languages that do not use spaces, such as Mandarin Chinese. Many segmentation systems have been proposed, some of them use linguistitically motivated unsupervized algorithms. However, standard evaluation practices fail to account for some properties of such systems. In this paper, we show that a simple model, based on an entropy-based reformulation of a language-independent hypothesis put forward by Harris (1955), allows for segmenting a corpus and extracting a lexicon from the results. Tested on the Academia Sinica Corpus, our system allows for inducing a segmentation and a lexicon with good intrinsic properties and whose characteristics are similar to those of the lexicon underlying the manually-segmented corpus. Moreover, the results of the segmentation model correlate with the syntactic structures provided by the syntactically annotated subpart of the corpus.

Mots-clés : Segmentation non-supervisée, entropie, induction de lexique, unité lexicale, chinois mandarin.

Keywords: Non-supervized segmentation, entropy, lexicon induction, Mandarin Chinese.

1 Introduction

La segmentation d'un texte en formes [1] est la première étape de presque tout traitement automatique de données textuelles. Pour la plupart des langues utilisant l'alphabet latin, dont le français ou l'anglais, un découpage selon les espaces et les symboles de ponctuation est une bonne approximation d'une segmentation en unités lexicales. À l'inverse, dans le cas des systèmes d'écriture utilisés par exemple pour écrire le chinois, le japonais, le thai, le khmer ou le vietnamien, la typographie n'est pas utilisée pour indiquer des frontières entre les mêmes unités linguistiques : en vietnamien, qui utilise une variante de l'alphabet latin, l'espace sépare des unités sous-lexicales. En chinois ou japonais, seuls les signes de ponctuation indiquent des frontières entre unités lexicales ; ailleurs, les caractères, qui représentent aussi des unités sous-lexicales, sont directement juxtaposés. L'étape de segmentation en unités lexicales est donc un problème délicat pour ces langues dites *non-segmentées*, et donne lieu à une littérature

1. Dans cet article, une *forme* est un segment continu de texte venant occuper de façon autonome une position syntaxique. Travaillant sur le mandarin, nous pouvons faire l'approximation qu'il y a identité entre la notion de forme et celle d'*unité lexicale*, Pour une discussion plus détaillées de l'unité lexicale en mandarin, se reporter à Packard (2000) ou en français à Nguyen (2006).

abondante (Zhao & Liu, 2010), y compris dans la communauté francophone (Seng *et al.*, 2009; Wu, 2010). Mais de tels travaux peuvent aussi être utiles pour les langues *segmentées*, en raison des cas de non-correspondance entre séparateurs et frontières d'unités lexicales, lesquelles restent difficiles à définir et à repérer quelle que soit la langue (Zhikov *et al.*, 2010).

Parmi les méthodes de segmentation, nous nous intéressons en particulier aux méthodes non supervisées qui cherchent une définition implicite du mot en faisant émerger la segmentation à partir des propriétés non-aléatoires de la distribution des formes en corpus. Ces méthodes sont difficiles à évaluer car elles ne s'adaptent pas à un standard donné. En contrepartie, elles présentent un plus grand potentiel d'adaptation à la dynamique d'une langue (changement de domaine, variantes géographiques, évolution diachronique, traitement des néologismes), et peuvent être utilisées pour la segmentation de langues peu ou pas dotées.

Nous décrivons ici une série d'expériences de segmentation en unités lexicales réalisées sur le chinois mandarin au moyen d'un système non-supervisé qui repose sur une hypothèse motivée linguistiquement formulée par (Harris, 1955), en adaptant sa modélisation présentée par (Tanaka-Ishii, 2005) dans le même but (Jin & Tanaka-Ishii, 2006). Nous insistons en particulier sur l'évaluation des résultats obtenus, tâche rendue délicate par la nature non-supervisée de l'approche et par la variété des conventions de segmentation qui existent pour le chinois mandarin.

Dans la section suivante nous présentons la tâche de segmentation et les problèmes que posent la méthode traditionnellement utilisée pour son évaluation. Les sections 3 et 4 présentent les systèmes dont nous nous inspirons et celui que nous avons développé. Nous utilisons ensuite notre système de segmentation pour extraire un lexique, dont nous proposons une évaluation (section 5). Enfin nous cherchons à corréler la sortie du système de segmentation étudié avec des informations syntaxiques extraites d'un corpus arboré.

2 La segmentation du chinois

La segmentation est la première étape de tout système d'analyse automatique du chinois écrit. En français et dans la majorité des langues utilisant l'alphabet latin, un découpage sur les espaces (et autour des signes de ponctuation), souvent appelé *tokenisation* et dont la sortie est un flux de *tokens*, constitue une première étape raisonnable, que l'on peut ensuite affiner pour identifier les cas de non-alignement entre *tokens* et *formes* (qui peuvent par exemple être des formes composées). À l'inverse, l'écriture chinoise ne comporte pas de séparateur typographique comme l'espace. Un découpage effectué uniquement autour des caractères de ponctuation produirait des *tokens* bien plus longs que des formes. À l'inverse, un découpage isolant chaque caractère chinois (ci-après *sinogramme*) ressemblerait plutôt à une segmentation en morph(ème)es qu'en formes. Il faut donc considérer un texte en chinois comme un flux de sinogrammes, la tâche de segmentation consistant à identifier entre quels sinogrammes il faut segmenter le texte afin de délimiter les formes, que l'on peut, en mandarin, assimiler à des *unités lexicales*.

2.1 État de l'art, enjeux actuels

Un grand nombre de méthodes ont été proposées pour effectuer une segmentation automatique. Certaines reposent sur des règles et des lexiques, d'autres utilisent des méthodes d'apprentissage automatique supervisé ou non-supervisé. Cinq campagnes du « *Chinese Word Segmentation Bakeoff* » ont été organisés par l'ACL, dont la dernière s'est tenue à l'été 2010. Zhao & Liu (2010) donnent un résumé des performances obtenues par les systèmes en compétition. Ils soulignent que si la précision peut sembler satisfaisante, la tolérance au changement de domaine et la reconnaissance des mots inconnus restent les limitations majeures.

Notons que lors de cette campagne, le système de base (*baseline*) et le meilleur système (*topline*) sont obtenus avec le même algorithme, un simple *maximum-matching* (*minimisation du nombre de mots*) reposant sur un inventaire d'unités lexicales, et ne se distinguent que par le lexique utilisé : la *baseline* utilise un lexique extrait à partir du corpus d'entraînement, tandis que la *topline* utilise un lexique extrait à partir de la totalité du corpus et connaît donc tous les formes attendues. Xue (2003), qui présentait un système d'apprentissage supervisé reposant sur une classification IOB (*Inside, Outside, Begin)* des sinogrammes, commente les résultats d'une autre heuristique simple qui repose sur un lexique, celle dite du *longest-match* gauche-droite (*plus longue chaîne d'abord*) : cette heuristique fournit de très bons résultats (f-mesure 0,952) si le lexique est exhaustif mais se dégrade très rapidement lorsque le corpus de test contient des mots inconnus (f-mesure de 0,898). Le *maximum-matching* utilisé lors du *bakeoff* obtient quant à lui des scores (f-mesure) supérieurs à 0,98 sur différents corpus (la *topline*) avec un lexique exhaustif,

et des scores de 0,72 à 0,88 selon les domaines dans la configuration *baseline*. Les 18 systèmes présentés lors du *segmentation bakeoff* ont tous obtenu des résultats intermédiaires entre ces deux niveaux. Il faut donc souligner l'importance pour cette tâche des ressources lexicales.

Parmi ces systèmes, et en général parmi les systèmes de segmentation du chinois mandarin, les deux principaux paradigmes d'apprentissage automatique ont été utilisés. Chacun présente des avantages et des inconvénients.

Les méthodes supervisées nécessitent un corpus d'entraînement constitué d'un ensemble de textes déjà segmentés (la réponse attendue, considérée comme « bonne »). À partir de ce jeu d'exemples, l'algorithme effectue une généralisation qui lui permet ensuite d'imiter la prise de décision effectuée par l'humain lors de la segmentation manuelle du corpus. De nombreuses méthodes d'apprentissage supervisé existent et les systèmes de segmentation actuels tendent à les combiner (cf. par exemple celui décrit par (Wu *et al.*, 2010), très bien classé au dernier *segmentation bakeoff*, qui repose sur un « *conditional support Markov model* »). Les méthodes supervisées sont celles qui obtiennent les meilleurs résultats, mais elles nécessitent l'utilisation d'un corpus d'entraînement dont la construction est longue et coûteuse. Ce corpus influe sur le comportement des systèmes qui dépendent de choix linguistiques particuliers, ainsi que de la nature du corpus (l'état de la langue à une époque donnée et pour un domaine donné). L'adaptation à d'autres domaines est un enjeu de recherche pour ce type de système.

Les méthodes non-supervisées n'utilisent pas de corpus pré-annoté mais se contentent d'une grande quantité de données brutes non-segmentées. L'hypothèse sous-jacente est que les données ne sont pas distribuées aléatoirement mais possèdent une certaine structure que l'on cherche à faire émerger par l'analyse de leur distribution. Parmi les méthodes utilisées pour la segmentation du chinois, on peut citer des approches utilisant l'information mutuelle, comme dans les travaux pionniers de Sproat *et al.* (1996), puis plus récemment des méthodes reposant sur l'algorithme *Expectation Maximization* (Peng & Schuurmans, 2001), ou sur la *Minimum Description Length* (Hua, 2000). La *complexité contextuelle*, inspirée des hypothèses de Harris (Harris, 1955) et utilisée dans les travaux de (Tanaka-Ishii & Jin, 2006; Jin & Tanaka-Ishii, 2006) est présentée plus en détail à la section 3. Les méthodes non supervisées présentent l'avantage de s'adapter à un corpus brut peu coûteux et plus facile à obtenir que des corpus segmentés manuellement. Mais elles sont difficiles à évaluer : il n'existe pas à priori de raison pour que la sortie d'un tel système corresponde à un guide de segmentation plutôt qu'à un autre.

2.2 Méthodes d'évaluation des systèmes de segmentation

Les différents systèmes de segmentation sont entraînés et évalués sur des parties de corpus dits « de référence » en utilisant les mesures classiques en apprentissage automatique (rappel, précision, f-mesure sur les formes ou sur les frontières), mais ce mode d'évaluation sous-estime une réalité linguistique complexe. Les différents corpus disponibles segmentés manuellement ne suivent pas les mêmes guides d'annotation. Ainsi le corpus de l'Université de Pékin suit le guide de Yu (1999) tandis que le corpus équilibré de l'*Academia Sinica* suit Huang *et al.* (1996) et que le Chinese Treebank respecte les conventions de Xia (2000).

Il a été plusieurs fois observé que le taux d'accord entre locuteurs natifs non linguistes à qui il était demandé de segmenter un texte est assez faible ((Sproat *et al.*, 1996) rapportent 76% de moyenne entre rappel et précision sur les mots, Jin (2007) rapporte une f-mesure de 0,839) . Ceci peut s'expliquer en partie par le fait que la tâche de segmentation recouvre différents problèmes qui ne sont spécifiques ni au mandarin ni à l'écriture chinoise : la définition des unités lexicales n'est triviale pour aucune langue et Packard (2000) propose 8 définitions différentes du « mot » : il fait remarquer que les critères phonologiques, syntaxiques, sémantiques, sociologiques, et autres ne coïncident pas toujours. La question de la segmentation soulève en effet des problèmes relatifs aux expressions multi-mots, au traitement des entités nommées et aux phénomènes de figement et de collocation (des exemples sont donnés à la section 5.3). Certains désaccords sur la segmentation relèvent d'une différence d'analyse morpho-syntaxique systématique et sont explicables, motivés et le plus souvent homogénéisables (Xia, 2000). C'est le cas par exemple du traitement de la marque du pluriel sur les nom humains (們 men) analysée en tant que suffixe ([N 們] = une unité) ou en tant que postposition (N + 們 = deux unités). Il en va de même pour les marques aspectuelles et résultatives sur les verbes. Les désaccords autour de figements lexicaux sont eux bien plus difficile à trancher (exemple : 全球暖化 quánqiúnuǎnhuà terre-entière-chaleur-devenir, *réchauffement planétaire*)

Enfin, dans un contexte applicatif, Wu (2003) note que différentes applications de TAL nécessitent différents critères de segmentation en amont. Dans notre cas, notre objectif premier est la construction de ressources lexicales à des fins de linguistique expérimentale sur corpus. Les contraintes et besoins que cela implique diffèrent donc légèrement des besoins posés par la conception d'applications TAL à visée plus industrielle.

2.3 Contexte et motivations

L'objectif de notre travail est notamment l'induction de lexiques et le pré-traitement de corpus à des fins d'études linguistiques. Ces corpus sont susceptibles de manifester une importante variation liée à trois facteurs au moins :
– l'espace : au travers des différentes variantes du mandarin pratiquées à Pékin, Hong Kong, Singapour ou Taïwan ;
– le temps : la publication récente des n-grammes de GoogleBooks ouvre de nouvelles possibilités pour une étude du lexique à la fois diachronique et quantitative, mais pose le problème de la segmentation sous un angle différent ; il est en effet exclu d'utiliser un système entraîné sur un corpus de la fin du XXe siècle pour segmenter des textes bien plus anciens [2] ;
– le domaine : des corpus de natures différentes, voire des corpus de spécialité, utiliseront des lexiques différents et en partie spécifiques, significativement différents de ce que l'on peut trouver dans les corpus d'apprentissage utilisés par les systèmes supervisés.

Les méthodes classiques de segmentation et d'évaluation exploitant des lexiques pré-existants ou des corpus segmentés manuellement semblent donc peu appropriées pour nos recherches où les mots inconnus et la tolérance à la variation nous intéressent particulièrement. D'un autre côté, les analyses proposées dans des travaux de linguistiques portant sur la définition de l'unité lexicale (Nguyen, 2006; Magistry, 2008) sont difficiles à automatiser.

Cette motivation à la fois linguistique et quantitative a nourri notre intérêt pour les méthodes non supervisées et particulièrement celles qui reposent sur l'hypothèse, motivée linguistiquement, de Harris (1955). Un exemple en est notamment les expérimentations menées sur corpus par Jin & Tanaka-Ishii (2006).

3 L'hypothèse harrissienne et sa reformulation entropique

Dans son article « *From phoneme to morpheme* », Harris (1955) formule l'hypothèse de l'existence d'un lien entre les frontières de morphèmes ou de mots et le nombre de successeurs possibles à une suite de phonèmes dans la chaîne parlée. Il effectue ensuite différentes expériences visant à confirmer cette hypothèse et à préciser la procédure de segmentation.

À l'époque, ces expériences ne pouvaient pas tirer parti de grands volumes de textes ou d'enregistrements et furent réalisées sous forme d'enquêtes. Plus récemment, cette idée a été déclinée pour réaliser différentes tâches telles que la détection de collocations (Frantzi & Ananiadou, 1996) ou la segmentation du chinois (Jin & Tanaka-Ishii, 2006). Les expériences sur le chinois reposent sur la reformulation de l'hypothèse de Harris dans le cadre de la théorie de l'information proposée par Tanaka-Ishii (2005), qui repose sur la notion d'*entropie* (notée H) : la distribution des successeurs possibles d'une suite de tokens, modélisée ici par un n-gramme x_n (de phonèmes ou de sinogrammes), permet de définir et de calculer une entropie $h(x_n)$, dite *entropie de branchement*, comme suit :

$$h(x_n) = H(\chi|x_n) = -\sum_{x \in \chi} P(x|x_n).\log P(x|x_n),$$

où χ est l'ensemble de tous les sinogrammes connus, mais aussi des lettres latines (en raison des expressions ou noms étrangers) et des chiffres arabes, et $P(x|x_n)$ est la probabilité conditionnelle de trouver le caractère x à la suite du n-gramme x_n. Cette reformulation a ainsi servi à tester l'hypothèse de Harris en corpus (Tanaka-Ishii & Jin, 2006).

Le modèle de Jin & Tanaka-Ishii (2006), qui est plus proche de l'article de Harris que notre système (décrit ci-dessous), est aussi beaucoup plus complexe. Il repose sur cinq modèles de langue (reposant sur des 1 à 5-grammes de sinogrammes) utilisés conjointement. Ceci permet de calculer une entropie de branchement après une séquence de longueur variable. Cependant à chaque intervalle entre deux sinogrammes, son système applique une série de critères qui lui permettent de décider de façon binaire si il faut segmenter ou non. La condition principale correspondant à l'hypothèse de Harris est qu'une frontière d'unité linguistique correspond à un point où l'entropie branchante atteint un maximum local. L'écriture chinoise produisant de nombreuses occurrences de mot d'un seul sinogramme, Jin et Tanaka-Ishii considèrent qu'il existe une frontière à chaque point où l'entropie est croissante. Leur système est ensuite évalué de façon classique par rappel/précision/f-mesure sur un extrait du corpus segmenté manuellement (celui de l'Université de Pékin suivant Yu (1999)).

2. Les n-grammes de Google ont toutefois été extraits après segmentation des textes, mais aucune information n'est donnée sur la méthode utilisée, ce qui pose problème pour l'exploitation de ces données.

Notre travail est motivé par l'idée que ce type de résultat binaire et ce mode d'évaluation ne révèle pas tout le potentiel de l'hypothèse et des modèles sous-jacents. Le système est évalué à chaque point du corpus alors que des généralisations pertinentes peuvent être obtenues à partir de l'ensemble des mesures de variation d'entropie effectuées, même bruités. Par ailleurs, les mesures de variations d'entropie ont des valeurs continues qui semblent à même de rendre compte plus finement du problème linguistique que les modes d'évaluation standard réduisent à une tâche de classification binaire.

Dans la section suivante nous présentons un système d'analyse qui conserve l'information sur les variations d'entropie afin de pouvoir utiliser celle-ci pour induire un lexique (section 5) ou corréler la variation d'entropie à la syntaxe sans la discrétiser (section 6).

4 Architecture de notre système de segmentation

Le modèle présenté ci-dessus peut être décliné en divers systèmes ayant en commun l'utilisation d'une mesure de « surprise » pour détecter les frontières. Contrairement aux travaux de Jin et Tanaka-Ishii, notre objectif n'est pas la segmentation en elle même mais l'induction de lexiques. Il nous est donc possible de ne pas prendre une décision binaire sur la segmentation à chaque intervalle entre deux sinogrammes mais de propager la mesure de « surprise » à un système qui prend une décision sur l'intégration ou non une suite de sinogrammes donnée à notre lexique.

Afin d'obtenir des résultats plus lisibles, nous avons choisi dans un premier temps d'utiliser un système simplifié ne reposant que sur un seul modèle de langue (4-grammes) qui calcule l'entropie branchante $h(x_3)$ à chaque inter-sinogramme et propage celle-ci pour en observer la variation.

Pour la séquence 台北市政府昨日開會決議… Táiběishìzhèngfǔ zuórì kāihuì juéyì… (*La municipalité de Taipei à décidé hier en réunion…*), notre chaîne de traitement produit la sortie suivante :

台 -4,98 北 1,11 市 1,53 政 -4,51 府 4,77 昨 -4,26 日 1,55 開 -0,06 會 -0,77 決 -0,92 議

Segmenter lorsque l'entropie est croissante produit donc le découpage : 台北 *(Taipei)* 市 *(ville)* 政府 *(gouvernement)* 昨日 *(hier)* 開會決議 (au lieu de 開會/決議 *tenir une réunion / décider*).

5 Induction de lexique : méthodologie et évaluation comparative

Dans cette section nous cherchons à induire un lexique à partir des informations sur la variations d'entropie et à définir une mesure de confiance dans les unités lexicales induites.

Une chaîne $w = c_1c_2...c_n$ est une unité lexicale candidate s'il en existe au moins une occurrence w_i dont les variations d'entropie inter-sinogrammes sont notées $e_{w_i,0}, \ldots, e_{w_in}$ avec $e_{w_i,k}$ la variation d'entropie après le k-ème sinogramme de la i-ème occurrence de w qui vérifie $e_{w_i,0} > 0,\ e_{w_i,n} > 0$ (l'entropie est croissante avant et après la chaîne) et $\forall k \in [1, n-1]\ ,\ e_{w_ik} \leq 0$ (l'entropie est décroissante ou constante à l'intérieur de la chaîne).

Nous avons appliqué notre segmenteur au corpus de l'Academia Sinica (Chen *et al.*, 1996), qui contient environ $7,7$ millions d'occurrences de 198 236 unités lexicales (segmentées manuellement) pour environ 12 millions de sinogrammes. Nous en avons ainsi extrait 193 714 unités lexicales candidates. Pour chaque unité lexicale candidate, nous disposons donc d'un ensemble d'occurrences auxquelles sont associées des variations d'entropie aux frontières et internes. Le corpus utilisé, qui compte 12 millions de sinogrammes, produit 193 714 unités lexicales candidates.

5.1 Métriques de confiance

Nous avons défini puis comparé différentes métriques pour filtrer le bruit parmi les unités lexicales candidates, qui combinent de façons différentes leur fréquence et une mesure de confiance, définie ci-dessous. La fréquence d'une unité lexicale candidate w est directement estimée à partir du nombre d'occurrence de celle-ci dans le corpus, noté $Nocc(w)$. Pour tirer parti de l'information sur la variation d'entropie, nous définissons pour chaque unité lexicale candidate une mesure de confiance définie à partir des variations d'entropie comme suit :

– pour chaque occurrence w_i de w, on définit une confiance locale $c_{w_i} = \min(e_{w_i,o},\ e_{w_i,n},\ -e_{w_i,1}, \ldots,\ -e_{w_i,n-1})$;

– on associe à la chaîne candidate w la confiance (globale) $c_w = \max_{i=1}^{Nocc(w)}(c_{w_i})$.

En d'autres termes, la confiance accordée à une occurrence est définie par valeur de variation d'entropie la moins fiable parmi celles ayant abouti à cette segmentation, et la confiance accordée à une unité lexicale est égale à la confiance de l'occurrence à laquelle on fait le plus confiance.

Différentes combinaisons de l'indice de confiance et du nombre d'occurrence sont alors possibles. Nous avons retenu les quatre combinaisons suivantes :

fréquence seule : $Nocc(w)$. Cette métrique simple présente l'avantage de cibler les unités lexicales couvrant le plus grand nombre d'occurrences en corpus, mais n'utilise pas l'information sur la variation d'entropie que nous conservons et se comporte ainsi comme les systèmes de segmentation binaires.

produit : $c_w \times Nocc(w)$. Cette métrique introduit la mesure de confiance, en lui conférant une importance identique à celle de la fréquence.

log : $c_w \times \log(Nocc(w))$. Cette métrique diminue l'impact des hautes fréquences.

confiance seule : c_w. Cette métrique ignore la mesure de fréquence et n'utilise que les informations extraites du modèle d'entropie.

Ces métriques sont choisies arbitrairement mais de manière à donner une importance croissante à la mesure de confiance afin d'évaluer sa pertinence.

5.2 Filtrage et évaluation du lexique de façon semi-supervisée.

Chacune des métriques présentées à la section précédente permet de définir un ordre de confiance sur les entrées lexicales, et donc de trier le lexique induit. On peut alors choisir par exemple de ne conserver que les n premières entrées. Pour un même n, chaque métrique conduit donc à un lexique filtré distinct. Il reste à choisir la meilleure métrique, puis un seuil sur cette métrique en dessous duquel filtrer le lexique, afin d'obtenir le meilleur compromis entre couverture et bruit.

Ces choix sont délicats à effectuer *a priori*, de même que l'évaluation de la qualité du lexique obtenu. Nous avons donc commencé par utiliser notre système dans une configuration semi-supervisée afin de pouvoir le comparer à une référence établie manuellement, et comprendre notamment quelle métrique semble se comporte le mieux.

Afin d'avoir une idée de la qualité d'un lexique L_i induit (filtré ou non), nous le comparons au lexique L_m extrait à partir de la segmentation effectuée manuellement à l'Academia Sinica. Rappelons qu'il ne s'agit pas d'un standard mais d'une analyse possible des données, motivée linguistiquement, mais qui n'est pas la seule. Pour comparer les deux lexiques nous avons utilisé l'*indice de Jaccard* et la *f-mesure* (qui donnent des résultats similaires). En l'absence de « bonne » ou de « mauvaise » réponse, ces indicateurs donnent tout de même une idée de la cohérence entre le résultat de notre système non-supervisé et une analyse effectuée manuellement. Ces deux mesures sont définies classiquement comme suit.

$$f(L_i, L_r) = \frac{2pr}{p+r}, \text{ avec } p(L_i, L_m) = \frac{|L_i \cap L_m|}{|L_i|} \text{ et } r(L_i, L_r) = \frac{|L_i \cap L_m|}{|L_m|} \; ; \quad jaccard(L_i, L_r) = \frac{|L_i \cap L_m|}{|L_i \cup L_m|}$$

Le lexique L_m obtenu à partir de la segmentation manuelle est trié par nombre d'occurrences (on considère que l'on a une confiance égale en toutes les entrées de ce lexique et l'on cherche à détecter en priorité les formes les plus fréquentes). Les différentes métriques décrites ci-dessus, qui combinent de diverses façons le nombre d'occurrences $Nocc$ et l'indice de confiance c permettent de moduler l'importance respective de ces deux quantités, depuis l'utilisation du nombre d'occurrence seul jusqu'à l'utilisation du seul indice de confiance.

Pour chacune des quatre métriques retenues, nous avons calculé le Jaccard et la f-mesure entre lexique induit filtré à différentes valeurs de la métrique et lexique manuel filtré à différents niveaux de fréquence. La figure 1 montre les résultats obtenus avec la f-mesure, en en indiquant les lignes de niveaux. Ceci nous permet de choisir un seuil en fonction d'un taux d'accord avec la référence. On observe que la qualité du tri par la fréquence se dégrade plus rapidement que les autres. À l'inverse, l'utilisation du seul indice de confiance ne semble pas accorder suffisamment d'importance aux formes fréquentes (le sommet est plus éloigné de l'origine du graphique).

Afin de nous faire une idée plus précise du contenu des lexiques induits par rapport à celui sous-jacent au corpus de l'Academia Sinica, nous avons choisi de les filtrer de la façon suivante : nous avons choisi le seuil de façon à maximiser la taille du lexique filtré tout en préservant une f-mesure d'au moins 0.6 par comparaison avec le lexique

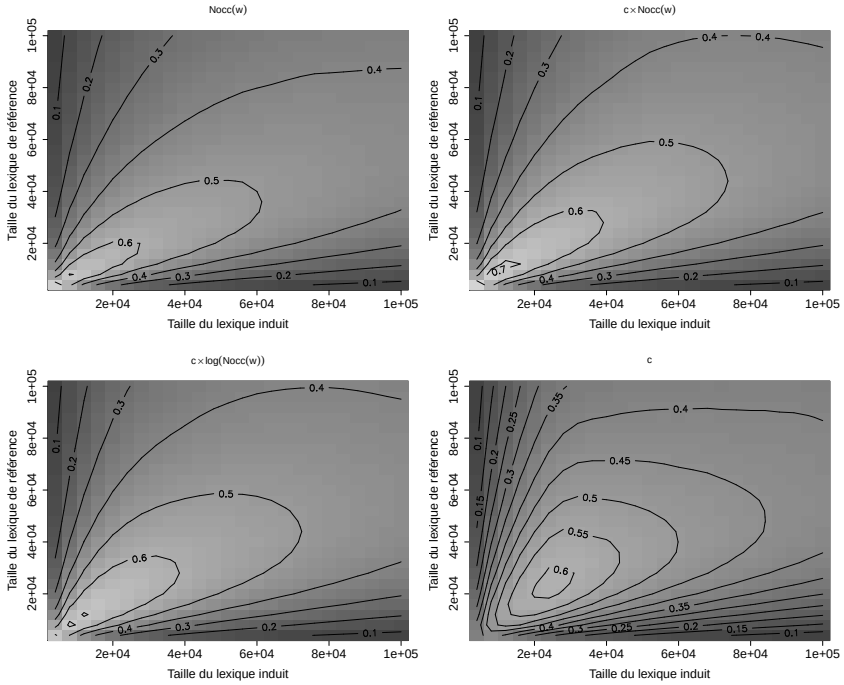


FIGURE 1 – comparaison des lexiques induits avec différentes mesures de confiance avec le lexique obtenu après segmentation manuelle (f-mesure)

mesure de confiance	taille $\|L\|$	formes validées $\|L \cap L_m\|$	occurrences couvertes	couverture	nombres de caractères
lexique manuel L_m	*116 844*	*116 844*	*7 584 040*	*100 %*	*5 861*
$Nocc(w)$	27 500	16 929	6 719 083	89 %	3 421
$c \times Nocc(w)$	38 000	24 045	6 901 992	91 %	4 016
$c \times \log(Nocc(w))$	38 000	24 571	6 892 594	91 %	4 070
c	31 000	23 620	6 695 416	88 %	4 097

TABLE 1 – Comparaison des lexiques induits L_i. Pour chaque mesure de confiance, le lexique est de taille maximale parmi ceux ayant une f-mesure de 0.6 par rapport au lexique L_m extractible du corpus de l'Academia Sinica.

manuel L_m (sur les graphiques de la figure 1, le seuil correspond donc à l'abscisse du point le plus à droite de la ligne à f-mesure de 0.6). Pour chacun des quatre lexiques ainsi extraits, nous avons effectué les mesures suivantes :
– nombre d'unités lexicales ($|L_i|$) ;
– nombre d'unités lexicales présentes dans la référence manuelle ($|L_i \cap L_m|$)
– nombre d'occurrences dans le corpus des unités lexicales communes (celles de $L_i \cap L_m$) ;
– couverture de $L_i \cap L_m$, c'est-à-dire proportion du corpus couverte par les unités lexicales communes ;
– nombre de sinogrammes distincts utilisés dans le lexique.
Les résultats sont données dans le tableau 1, où nous donnons également les valeurs correspondantes pour le lexique manuel L_m. On constate que l'utilisation de notre indice de confiance améliore bien la proportion de formes valides tandis que la fréquence reste une valeur intéressante pour optimiser la couverture du corpus.

Remarquons que les formes valides capturées par nos lexiques diffèrent sensiblement. Le tableau 2 donne les indices de Jaccard entre nos lexiques calculées 2 à 2 et confirme l'intérêt de combiner les deux informations de confiance et de fréquence.

	$Nocc(w)$		
$c \times Nocc(w)$	0,64	$c \times Nocc(w)$	
$c \times \log(Nocc(w))$	0,59	0,92	$c \times \log(Nocc(w))$
c	0,46	0,71	0,76

TABLE 2 – Indices de Jaccard entre les quatre lexiques induits filtrés, en se restreignant aux unités lexicales également présentes dans le lexique manuel.

type d'erreur	quantité	type d'erreur	quantité
suffixation	37	écriture non-chinoise	14
dates et nombres	35	conjonction	14
verbes	27	adverbes	10
expressions figée	21	entité nommée	4
translittération	2	autre	36

TABLE 3 – Répartition des faux négatifs

5.3 Analyse d'erreur

Dans cette section, nous présentons les résultats d'un analyse d'erreur, ou de divergence, entre le lexique de référence et le lexique construit à la section précédente au moyen de la mesure $c_w \times \log(Nocc(w))$. Nous avons concentré notre analyse sur deux axes : les unités lexicales de haute fréquence absentes de notre lexique induit mais présentes dans le lexique de référence (faux négatifs) et les chaînes considérées comme des unités lexicales avec un haut niveau de confiance mais absentes du lexique de référence (faux positifs). Pour chaque groupe, nous avons considérés les 200 premiers cas.

Le système mis en œuvre pour cet article est volontairement simpliste pour établir un système de base auquel se comparer. Les erreurs observées dans cette section suggèrent différentes pistes d'amélioration en amont (modification sur le modèle de langue utilisé) et en aval (ajout de règles linguistiques basées sur les catégories fermées).

5.3.1 Analyse des faux négatifs

Nous avons classé les faux négatifs suivant leur morphologie lorsque leur construction interne était transparente, dans le cas contraire nous avons les avons classé en parties du discours, mais le mandarin étant très ambigu sur ce point (le phénomène de conversion est fréquent), de nombreux cas sont restés non classés. Les résultats de cette analyse sont donnés dans le tableau 3 dont nous détaillons la moitié gauche ci-dessous.

Le type d'erreur le plus représenté concerne des unités lexicales, essentiellement nominales, construite sur le modèle *base+suffixe*. C'est là un phénomène de morphologie constructionnelle très productif en mandarin moderne dont on peut donner l'exemple 法務部長 fǎwùbùzhǎng *ministre de la justice* construit sur la base 法務部 fǎwùbù *ministère de la justice* à laquelle on ajoute le suffixe 長 zhǎng pour tête/chef (部 bù étant lui même un suffixe indiquant un ministère) . Dans (Magistry, 2008), des méthodes quantitatives ont permis d'estimer la productivité de ce procédé et ont ainsi montré que les règles très productives correspondent aux cas où le statut morphologique ou syntaxique de la composition est le plus discutable. C'est aussi un des points sur lesquels les guides de segmentation manuelle peuvent diverger. La grande proportion de ce type d'erreur n'est donc pas étonnante mais un soin particulier devra être apporté au traitement de ce phénomène dans des travaux futurs.

Les dates et nombres sont aussi une erreur attendue, la distribution des tokens qui les composent étant particulière.

Le cas des verbes est moins clair. cependant la présence de marques d'aspect (formant une petite classe fermée) directement à la suite du verbe peuvent induire notre système en erreur. Il coupera après le marqueur d'aspect. Il en va de même (mais dans une moindre mesure) des constructions *verbe+résultatif* (ex : 吃完 chīwán manger-finir, *avoir fini de manger*) qui bien que sécables (ex : 吃不完 chībuwán manger-négation-finir, *ne pas pouvoir finir de manger*) n'ont pas toujours un sens compositionnel. Certaines combinaisons sont lexicalisées et peuvent donner lieu à débat concernant leur bonne segmentation.

Parmi les 200 premiers nous avons compté 21 cas qui nous semblent être des figements à différents degrés, comme 高爾夫-球場 gāo'ěrfū-qiúchǎng golf-terrain, *terrain de golf*, l'unité pour *golf* étant autonome et celle pour *terrain* pouvant concerner tout type de terrain sport utilisant une balle. Mais ceci inclut aussi des « expressions en quatre

Type d'erreur	Qnt	Type d'erreur	Qnt
nom	17	verbe+DE	8
adverbe+verbe	15	adverbe+copule	8
suffixe+DE	14	adverbe+adverbe	8
nombre+classificateur	14	adverbe+avoir	7
verbe+aspect	13	adverbe+auxiliaire	7
démonstratif+classificateur	9	pronom+DE	6

TABLE 4 – Répartition des faux positifs

caractères » dont la concision et la structure interne relèvent d'un état antérieur de la langue (ex : 前所未有 avant/ce que/pas encore/avoir, *sans précédent*). Ce type d'erreurs regroupe ainsi des expressions figées dont certaines sont idiomatiques et d'autres compositionnelles (en quantités équivalentes).

5.3.2 Analyse des faux positifs

Nous avons ensuite analysé les faux positifs en observant leur composition interne. Une première remarque est que sur les 200 premiers, 198 sont des bigrammes et 2 sont des unigrammes (cette préférence disparaît à mesure que la confiance diminue). Nous avons donc classifié les erreurs en fonction des deux sinogrammes qui les composent. La dispersion est plus grande, nous ne donnons donc dans le tableau 4 que les types les plus importants.

Les 17 noms sont des unités lexicales dont le statut est discutable. 13 comportent en seconde position un élément appartenant à une classe très fermée de « mots » indiquant un lieu ou une direction (上,下,内,裡,中 shàng, xià, nèi, lǐ, zhōng *sur, sous, dans, à, au milieu*) et en première position un nom monosyllabique.

Les *adverbe+verbe* sont des combinaisons d'adverbe monosyllabique (*très, le plus, trop, aussi, relativement*) et de verbes d'état (*rapide, grand, petit, suffisant, nombreux, difficile).*

Suffixe+DE désigne un groupe composé en première position d'un suffixe nominal très productif (voir plus haut) et en seconde de 的 DE qui marque une relative ou la possession. Il s'agit ici d'une double erreur de segmentation dûe au fait que de nombreuses bases peuvent commuter devant ces suffixes et qu'il viennent terminer une large classe de nom et peuvent donc fréquemment se trouver devant le DE.

Les séquences *nombre+classificateur* et *démonstratif+classificateur* sont liées à la structure des groupes nominaux (non nus) en mandarin dans lesquels un classificateur est requis et doit obligatoirement être précédé d'un élément dénotant une quantité ou d'un démonstratif. Il est donc peu étonnant que notre système tende à les regrouper.

Remarquons aussi les 6 séquences pronom+DE qui se traduiraient par des possessifs *mon/le mien, ton/le tien, son/le sien…* auxquels s'ajoutent deux variantes avec un possesseur non-humain. On ne compte que 6 erreurs de ce type, mais c'est là une liste exhaustive des pronoms singuliers + DE.

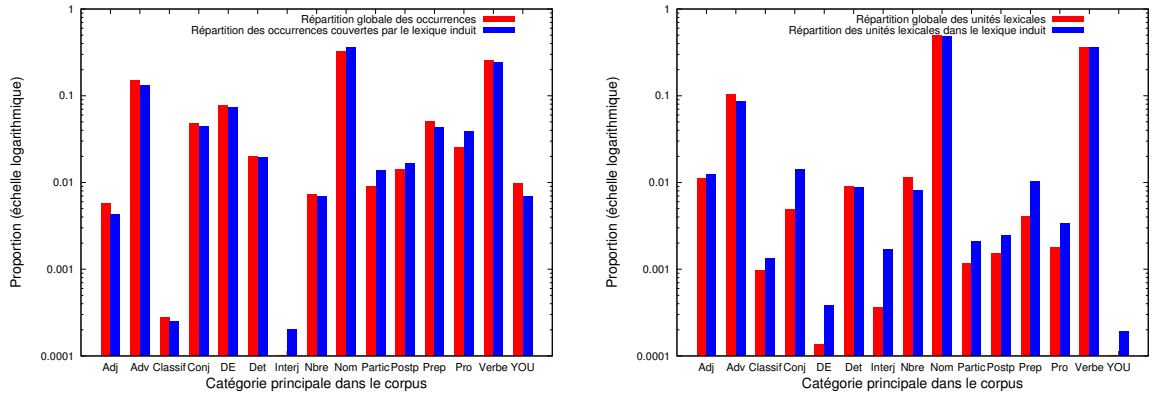
Une analyse de ces erreurs réalisée sur les caractères et non sur couples de caractères erronés montre que toutes ces erreurs incluent une unité lexicale qui est un unigramme très fréquent ou appartenant à une classe très fermée.

5.4 Caractéristiques générales du lexique

Chen *et al.* (1993) fournissent des informations sur la distribution globale d'un lexique du chinois mandarin et des occurrences en corpus. On peut ainsi vérifier si notre lexique induit possède bien les même propriétés que le lexique obtenu manuellement.

Tout d'abord, en observant le nombre d'occurrence des unités lexicale ordonnées par fréquence décroissante, on obtient une courbe zipfienne qui se superpose bien avec celle obtenue en utilisant le lexique manuel.

Les observations plus spécifiques au mandarin concernent la répartition des unités lexicales et de leurs occurrences en fonction de leur longueur en nombre de sinogrammes. Sur les 38 000 unités lexicales « de confiance » (ou les plus fréquentes pour le lexique manuel), le lexique induit est bien constitué principalement de bigrammes et de trigrammes (respectivement 65,8% et 27,3%) tandis que les unigrames constituent 6,5% du lexique. Le lexique manuel compte lui 7% d'unigrammes, 67,7% de bigrammes et 19% de trigrammes. Concernant le nombre d'occurrence observées dans le corpus, on obtient 37% d'unigramme, 54% de bigrammes et 7% de trigrammes (pour le corpus segmenté manuellement, on obtient respectivement 45%, 49% et 4%)



(a) (b)

FIGURE 2 – Comparaison des répartitions en catégories entre les occurrences de formes segmentées à l'identique par notre système et par le corpus lui-même (a), et entre les unités lexicales correspondantes (b).

6 Évaluation de la segmentation par comparaison avec les informations morphosyntaxiques et syntaxiques

Après avoir évalué notre modèle de segmentation au travers du lexique qu'elle permet d'extraire du corpus de l'Academia Sinica, nous avons poursuivi nos expériences d'évaluation en cherchant à tirer parti des annotations morphosyntaxiques et syntaxiques que fournit le Sinica Treebank (Chen *et al.*, 1996), qui en couvre une partie.

6.1 Répartition en catégories

L'étude de la distribution du lexique induit L_i en fonction de sa fréquence, bien qu'importante, ne suffit pas à montrer que ses propriétés sont cohérentes avec celles du lexique motivé linguistiquement L_m qui est sous-jacent au corpus de l'Academia Sinica. Nous avons donc cherché à comparer la répartition en catégories de ces deux lexiques. Nous avons donc identifié dans le treebank de l'Academia Sinica les occurrences de formes communes entre la segmentation du corpus et celle de notre système. Nous avons alors comparé la répartition de ces occurrences en parties du discours par rapport à celle de l'ensemble des occurrences de formes dans le corpus de l'Academia Sinica (figure 2a). Nous avons effectué la même chose avec unités lexicales correspondantes (figure 2b). Les catégories que nous avons utilisées sont les suivantes : Adj (adjectif), Adv (adverbe), Classif (classifieur), Conj (conjonction), DE (particules 的,之,得 et 地), Det (déterminant), Interj (interjection), Nbre, Nom, Partic (particule), Postp (postposition), Prep (préposition), Pro (pronom), Verbe et YOU (verbe 有 , *avoir*).

Nous discuterons de façon plus détaillée ces figures à la section suivante, mais on constate globalement une bonne corrélation entre les deux répartitions, tant pour les catégories les plus fréquentes que pour celles qui le sont moins.

6.2 Corrélation entre variation d'entropie et structure syntaxique

Notre modèle de segmentation reposant sur les variations d'entropie, il ne produit pas simplement pour chaque paire de sinogrammes adjacents une décision binaire (segmenter ou non), mais bien une mesure quantitative de la séparabilité des deux sinogrammes concernés. Nous avons cherché à confronter ce *degré de séparation linéaire* S_l avec les informations syntaxiques (arbres en constituants) fournies par le Sinica Treebank. L'intuition sous-jacente est que l'on pourrait constater une corrélation entre la séparabilité linéaire produite par notre modèle de segmentation et un *degré de séparation syntaxique*, mesure qui serait d'autant plus élevée que les deux sinogrammes étudiés appartiennent à des unités lexicales éloignées l'une de l'autre au sein de la structure en constituants.

Pour effectuer cette expérience, nous avons défini le degré de séparation syntaxique S_s entre deux unités lexicales adjacentes comme étant la longueur (en arcs) du plus court chemin permettant de les relier entre elles dans l'arbre de constituance. Il résulte de cette définition que S_s est nécessairement au moins égal à 2, ce qui est le cas lorsque

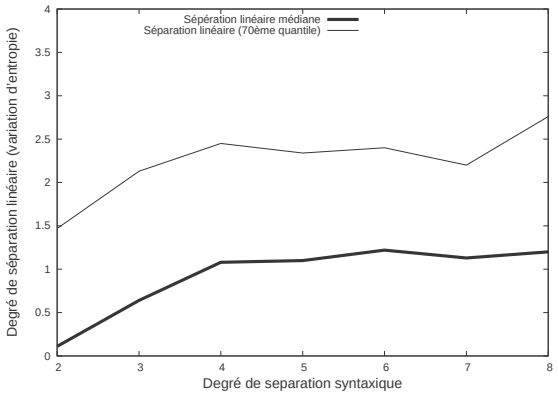


FIGURE 3 – Évolution de la séparation linéaire S_l (variation d'entropie) en fonction de la séparation syntaxique S_s sur les frontières communes au corpus de l'Academia Sinica et à la segmentation induite par notre système.

les unités lexicales ont un même nœud père.

La figure 3 montre pour chaque valeur de la séparation syntaxique S_s quelle est la valeur médiane de la séparation linéaire S_l, ainsi que son soixante-dixième quantile. On constate deux choses. Tout d'abord, lorsque le degré de séparation syntaxique est de 4 ou moins, il y a une nette corrélation entre S_l et S_s. Autrement dit, le modèle de segmentation utilisé réussit à capturer une partie des informations syntaxiques locales, de niveau terme voire *chunk*. En revanche, au-delà d'une séparation syntaxique de 4, la séparation linéaire médiane n'évolue quasiment plus. Deux hypothèses, non-exclusives l'une de l'autre, viennent à l'esprit. Tout d'abord, le modèle utilisé est 4-gramme, et il est difficile de capturer des frontières entre longs constituants avec un modèle local de ce type. Par ailleurs, le modèle simple que nous utilisons, inspiré de Harris, est un modèle très surfacique qui n'a aucune raison de pouvoir capturer des informations sur la macro-structure de l'arbre syntaxique d'une phrase.

Ce résultat, bien qu'obtenu à l'échelle de tout le corpus au moyen d'un calcul de médiane, est néanmoins prometteur : il ne semble pas exclu de pouvoir utiliser notre modèle de segmentation non-supervisé, tel quel ou sous une forme raffinée, non seulement pour induire une segmentation en unités lexicales et un lexique associé mais également pour identifier des collocations, termes, locutions et autres unités lexicales complexes, et de tenter de leur associer une structure interne. On peut ainsi espérer avoir accès à un moyen objectif, qui n'utilise pas de connaissance *a priori* et qui est donc indépendant de la langue, pour mettre en évidence le continuum qui relie les unités lexicales les plus classiques aux expressions semi-compositionnelles ou collocationnelles.

7 Conclusion et perspectives

Dans cet article, nous montrons sur le chinois mandarin qu'un modèle simple utilisant une hypothèse linguistiquement motivée mais indépendante de toute connaissance *a priori* sur une langue particulière donne des résultats prometteurs pour la segmentation non supervisée de textes et l'induction d'unités lexicales cohérentes avec des annotations de niveau syntaxique. De plus, certains résultats pouvant apparaître comme des erreurs de segmentation sont susceptible de questionner de façon constructive des analyses linguistiques traditionnelles parfois influencées par les états antérieurs de la langue.

Certaines erreurs résultent toutefois des limites de notre système dans son état actuel. En particulier, un traitement plus fin des catégories fermées (démonstratifs, DE,...) pourrait nettement améliorer les résultats tout en demandant une quantité d'analyse bornée par la taille de ces catégories. Mais d'autres améliorations destinées à rendre le modèle plus proche de considérations linguistiques pourront également être testées. Le modèle lui-même peut également faire l'objet de raffinements, par exemple pour comprendre si la prise en compte du mot situé à *droite* de l'intersinogramme considéré est de nature a améliorer les résultats.

Nous prévoyons de tester notre système sur des corpus relevant de différentes variétés du chinois mandarin, pour en étudier notamment les variations des distributions lexicales, mais également de le tester sur d'autres langues non-segmentées, pour valider l'approche sur un échantillon plus large de langues. Nous souhaitons également me-

ner des expérimentations sur diverses langues, y compris le français, pour segmenter non seulement des flux de sinogrammes mais ainsi, par exemple, des flux de phonèmes (en vue d'une segmentation en morphèmes) ou de *tokens* (en vue de l'identification d'unités lexicales multi-mots et de termes).

Références

CHEN C. Y., TSENG S. F., HUANG C. R. & CHEN K. J. (1993). Some distributional properties of Mandarin Chinese — A study based on the academia sinica corpus. In *Proceedings of Pacific Asia Conference on Formal and Computational Linguistics I*, p. 81–95.

CHEN K. J., HUANG C. R., CHANG L. P. & HSU H. L. (1996). Sinica corpus : Design methodology for balanced corpora. In *Proceedings of PACLIC 11th Conference*, p. 167--176.

FRANTZI K. T. & ANANIADOU S. (1996). Extracting nested collocations. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, p. 41–46.

HARRIS Z. S. (1955). From phoneme to morpheme. *Language*, **31**(2), 190–222.

HUA Y. (2000). Unsupervised word induction using MDL criterion. In *Proceedings of ISCSL*.

HUANG C. R., CHEN K. J. & CHANG L. L. (1996). Segmentation standard for chinese natural language processing. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, p. 1045–1048.

JIN Z. (2007). *A Study on Unsupervised Segmentation of Text Using Contextual Complexity*. PhD thesis, University of Tokyo, Graduate School of Information Science and Technology, Tokyo, Japon.

JIN Z. & TANAKA-ISHII K. (2006). Unsupervised segmentation of chinese text by use of branching entropy. In *Proceedings of the COLING/ACL on Main conference poster sessions*, p. 428–435.

MAGISTRY P. (2008). Productivité morphologique : Étude sur le chinois mandarin. Master's thesis, Université Paris Diderot, UFR de Linguistique, Paris, France.

NGUYEN . (2006). *Unité lexicale et morphologie en chinois mandarin*. PhD thesis, Université de Montréal, Montréal.

PACKARD J. L. (2000). *The morphology of Chinese : A linguistic and cognitive approach*. Cambridge Univ Pr.

PENG F. & SCHUURMANS D. (2001). Self-supervised chinese word segmentation. *Advances in Intelligent Data Analysis*, p. 238–247.

SENG S., BIGI B., BESACIER L. & CASTELLI E. (2009). Segmentation multiple d'un flux de données textuelles pour la modélisation statistique du langage. In *Actes de la conférence TALN 2009*, Senlis, France.

SPROAT R., GALE W., SHIH C. & CHANG N. (1996). A stochastic finite-state word-segmentation algorithm for chinese. *Computational linguistics*, **22**(3), 377–404.

TANAKA-ISHII K. (2005). Entropy as an indicator of context boundaries : An experiment using a web search engine. *Natural Language Processing–IJCNLP 2005*, p. 93–105.

TANAKA-ISHII K. & JIN Z. (2006). From phoneme to morpheme : Another verification using a corpus. *Computer Processing of Oriental Languages. Beyond the Orient : The Research Challenges Ahead*, p. 234–244.

WU A. (2003). Customizable segmentation of morphologically derived words in chinese. *International Journal of Computational Linguistics and Chinese Language Processing*, **8**(1), 1–27.

WU L.-C. (2010). Outils de segmentation du chinois et textométrie. In *Actes de la conférence TALN 2010*, Montréal, Canada.

WU Y. C., YANG J. C. & LEE Y. S. (2010). Chinese word segmentation with conditional support vector inspired markov models. In *Proceedings of the Joint Conference on Chinese Language Processing*.

XIA F. (2000). The segmentation guidelines for the penn chinese treebank (3.0). *IRCS Technical Reports Series*.

XUE N. (2003). Chinese word segmentation as character tagging. *International Jourrnal of Computational Linguistics and Chinese Language Processing*.

YU S. (1999). Guidelines for the annotation of contemporary chinese texts : word segmentation and POS-tagging. *Institute of Computational Linguistics, Beijing University, Beijing*.

H. ZHAO & Q. LIU, Eds. (2010). *The CIPS-SIGHAN CLP 2010 Chinese Word Segmentation Bakeoff*.

ZHIKOV V., TAKAMURA H. & OKUMURA M. (2010). An efficient algorithm for unsupervised word segmentation with branching entropy and MDL. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, p. 832–842.

TALN 2011, Montpellier, 27 *juin* –1er *juillet 2011*

Structure des trigrammes inconnus et lissage par analogie

Julien Gosme[1] Yves Lepage[2]
(1) GREYC, université de Caen Basse-Normandie, France
Julien.Gosme@unicaen.fr
(2) IPS, université Waseda, Japon
Yves.Lepage@aoni.waseda.jp

Résumé. Nous montrons dans une série d'expériences sur quatre langues, sur des échantillons du corpus Europarl, que, dans leur grande majorité, les trigrammes inconnus d'un jeu de test peuvent être reconstruits par analogie avec des trigrammes hapax du corpus d'entraînement. De ce résultat, nous dérivons une méthode de lissage simple pour les modèles de langue par trigrammes et obtenons de meilleurs résultats que les lissages de Witten-Bell, Good-Turing et Kneser-Ney dans des expériences menées en onze langues sur la partie commune d'Europarl, sauf pour le finnois et, dans une moindre mesure, le français.

Abstract. In a series of experiments in four languages on subparts of the Europarl corpus, we show that a large number of unseen trigrams can be reconstructed by proportional analogy using only hapax trigrams. We derive a simple smoothing scheme from this empirical result and show that it outperforms Witten-Bell, Good-Turing and Kneser-Ney smoothing schemes on trigram models built on the common part of the Europarl corpus, in all 11 languages except Finnish and French.

Mots-clés : analogie, trigrammes inconnus, trigrammes hapax, modèle de langue trigrammes, Europarl.

Keywords: proportional analogy, unseen trigrams, hapax trigrams, trigram language models, Europarl.

1 Introduction

Les techniques de lissage de modèles de langue reposent habituellement sur des hypothèses purement statistiques pour estimer la probabilité des évènements inconnus. Il y a dix ans, (Rosenfeld, 2000) constatait que :

> *Ironically, the most successful SLM techniques use very little knowledge of what language really is. The most popular language models (n-grams) take no advantage of the fact that what is being modeled is language.*

Nous présentons ici une technique de lissage pour les modèles de langue trigrammes qui repose sur la structure des évènements inconnus, c'est-à-dire la manière dont les trigrammes inconnus peuvent être construits à partir des trigrammes connus en utilisant une opération structurelle linguistiquement justifiée, l'analogie.

Le but du lissage des modèles de langue est d'attribuer des probabilités non-nulles aux évènements inconnus. Habituellement, les probabilités attribuées dépendent d'une caractérisation théorique des événements inconnus. L'hypothèse à l'origine de ce travail est que les trigrammes inconnus peuvent être caractérisés, dans une large mesure, par la similitude de leurs structures avec des trigrammes rares. Plus précisément nous montrons ci-dessous que, dans une large mesure, les trigrammes inconnus sont analogues aux trigrammes hapax.

En guise d'illustration, dans une de nos expériences préliminaires, le trigramme de mots *opportunité de servir* était un trigramme de notre jeu de test absent du corpus d'entraînement. Il se trouvait que ce trigramme pouvait être reconstruit par analogie à l'aide de trois trigrammes du corpus d'entraînement de la manière suivante :

opportunité de servir : *opportunité de* ~~*modifier*~~ :: ~~*qui pourrait*~~ *servir* : ~~*qui pourrait modifier*~~

La ligne précédente se lit ainsi : le trigramme inconnu *opportunité de servir* est au trigramme connu *opportunité de modifier* ce qu'un autre trigramme connu, *qui pourrait servir*, est à un dernier trigramme connu, *qui pourrait*

modifier. Les différents éléments du trigramme inconnu sont obtenus par similarité avec le second et le troisième trigrammes (*opportunité de* et *servir*) et peuvent être assemblés par différence avec le quatrième trigramme (mots barrés). En plus de permettre la reconstruction, les trois trigrammes ci-dessus étaient tous hapax dans le corpus d'entraînement.

La relation, telle celle donnée ci-dessus entre trigrammes de mots, qui énonce qu'A est à B ce que C est à D est appelée analogie. Un certain nombre de travaux en traitement automatique des langues exploitent l'analogie. Nous n'en citons que quelques-uns ici. Par exemple, sur des tâches de segmentation morphologique, (Lavallée & Langlais, 2010) ont récemment obtenu d'excellents résultats dans la découpe des mots par analogie. (Claveau & L'Homme, 2005), entre autres auteurs, avaient auparavant étudié, en faisant usage de l'analogie, dans quelle mesure la similarité liait la forme et le sens des mots : *connector* : *to connect* :: *editor* : *to edit*. En plus des analogies entre mots eux-mêmes, (Hathout, 2009) a récemment exploité les analogies entre définitions extraites du TLFi pour construire automatiquement des familles de mots liés par la forme et le sens. Dans le même ordre d'idée, (Langlais *et al.*, 2008) avaient proposé d'utiliser l'analogie pour forger de nouvelles équivalences terminologiques dans le domaine médical à cheval sur deux langues. Sur le seul plan sémantique, (Turney, 2008) a quant à lui présenté une approche générale au problème de l'association entre mots utilisant l'analogie entre vecteurs contextuels : *mason* : *stone* :: *carpenter* : *wood*, approche qu'il prétend généralisable aux relations de synonymie et d'antonymie. (Lepage & Denoual, 2005) quant à eux ont conçu un système de traduction automatique entièrement fondé sur l'analogie. Dans le cadre de la traduction automatique aussi, (Denoual, 2007) et (Langlais & Patry, 2007) ont montré la possibilité de traduire certains mots inconnus par analogie.

La définition de l'analogie que nous utilisons dans ce travail est détaillée et justifiée dans (Lepage, 2004). Nous l'appliquons aux trigrammes de mots. Un quadruplet de trigrammes de mots A, B, C et D est une analogie lorsque les contraintes suivantes sont vérifiées :

$$\begin{cases} d(A,B) &= d(C,D) \\ d(A,C) &= d(B,D) \\ |A|_m - |B|_m &= |C|_m - |D|_m, \forall m \end{cases}$$

Ici, d est la distance d'édition qui compte le nombre minimum d'insertions et de suppressions de mots nécessaires à la transformation d'un trigramme en un autre. [1] $|A|_m$ est le nombre d'occurrences du mot m dans le trigramme A. En reprenant l'exemple précédent :

$$\begin{array}{rcl} A &=& \textit{opportunité de servir} \\ B &=& \textit{opportunité de modifier} \\ C &=& \textit{qui pourrait servir} \\ D &=& \textit{qui pourrait modifier} \end{array}$$

on peut vérifier que $d(A,B) = d(C,D) = 2$ et $d(A,C) = d(B,D) = 4$. La relation entre nombres d'occurrences est vérifiée pour chaque mot :

mot m	$\lvert A\rvert_m - \lvert B\rvert_m = \lvert C\rvert_m - \lvert D\rvert_m$
opportunité	$1 - 1 = 0 - 0$
de	$1 - 1 = 0 - 0$
servir	$1 - 0 = 1 - 0$
modifier	$0 - 1 = 0 - 1$
qui	$0 - 0 = 1 - 1$
pourrait	$0 - 0 = 1 - 1$

Le bon sens veut que les trigrammes inconnus apparaissant dans un jeu de test qui peuvent être reconstruits par analogie avec des trigrammes d'un corpus d'entraînement soient considérés plus sûrs que ceux qui ne peuvent pas l'être. Une technique de lissage basée sur de simples décomptes devrait donc donner une plus forte ré-estimation aux trigrammes pouvant être reconstruits par analogie et une plus faible ré-estimation aux autres. Si, en plus, les trigrammes reconstruits peuvent l'être à l'aide de trigrammes hapax, la ré-estimation de leur effectif devrait être proche de 1 puisqu'ils sont alors proches structurellement des trigrammes hapax.

1. La distance d'édition de Levenshtein (Levenshtein, 1966) prend en compte la substitution comme opération d'édition supplémentaire.

La suite de l'article est divisée en deux parties : la section 2 est consacrée à vérifier l'hypothèse que les trigrammes inconnus sont structurellement analogues aux trigrammes hapax. Des expériences successives menées sur quatre langues européennes confirment, les unes après les autres, cette hypothèse. La section 3 présente alors une technique de lissage reposant sur cette propriété, et directement inspirée des techniques élémentaires de Lidstone et de Laplace. Des comparaisons effectuées avec quatre autres techniques de lissage classiques sur onze langues européennes montrent son efficacité, voire sa supériorité.

2 La structure des trigrammes inconnus d'un jeu de test

Nous menons des expériences sur les quatre langues suivantes : l'anglais, le français, l'allemand et le finnois. Ces langues ont été choisies pour leurs différentes richesses morphologiques. Sur une échelle croissante, on peut en effet placer successivement l'anglais, le français, l'allemand puis le finnois qui a la morphologie la plus riche.

Le corpus Europarl (Koehn, 2005) offre des textes alignés dans ces quatre langues,[2] ce qui permet de mener des expériences véritablement comparables. Pour les expériences de cette section, de l'ensemble de toutes les phrases correspondantes dans toutes les langues, nous avons extrait aléatoirement 100 000 phrases. Parmi elles, 90 000 phrases ont été sélectionnées aléatoirement, les mêmes dans toutes les langues, pour servir de corpus d'entraînement. Le jeu de test est constitué des 10 000 phrases restantes.

2.1 Proportion de trigrammes inconnus reconstruits

Pour vérifier dans quelle mesure les trigrammes inconnus d'un jeu de test peuvent être reconstruits par analogie, nous effectuons une première série d'expériences par validation croisée. Nous comptons simplement le nombre total de trigrammes inconnus du jeu de test reconstructibles par analogie à l'aide de trois trigrammes du corpus d'entraînement. Dès lors que la reconstruction est possible, le processus est interrompu pour ce trigramme.

Les résultats obtenus, reportés dans la table 1, montrent que la proportion de trigrammes inconnus dans le jeu de test reconstructibles par analogie avec trois autres trigrammes du corpus d'entraînement est supérieure à 80 % en anglais et en français et supérieure à 70 % en allemand. Cette proportion, appelée μ ici, est donc importante. Elle est calculée sur le nombre total de trigrammes inconnus différents (sans répétition) dont la proportion relativement au nombre total de trigrammes du jeu de test est appelée λ. Des expériences non rapportées ici montrent qu'en augmentant la taille des données, les valeurs de λ baissent tandis que les valeurs de μ augmentent. Les paramètres λ et μ seront exploités dans la section 3.1.

TABLE 1 – Nombre de trigrammes inconnus différents dans le jeu de test et proportion de trigrammes inconnus différents reconstructibles par analogie à l'aide de trois trigrammes du corpus d'entraînement.

	Trigrammes inconnus	
	du jeu de test (λ)	reconstruits (μ)
anglais	114,566 (60,04 %)	83,67 %
français	116,922 (57,81 %)	81,87 %
allemand	140,226 (68,97 %)	72,14 %
finnois	132,931 (83,33 %)	44,93 %

En finnois, la proportion de trigrammes inconnus différents reconstruits est faible avec seulement 45 %. Cette faible valeur est certainement explicable par la richesse morphologique de cette langue et donc l'absence relative de mots-fonctions permettant plus de commutations de mots dans les trigrammes. Nous pouvons dès à présent nous attendre à des résultats différents en finnois dans toute la suite de notre étude.

2. http://www.statmt.org/europarl

2.2 Patrons d'analogie les plus fréquents

Un trigramme de mots donné peut être obtenu par analogie à l'aide d'autres trigrammes de plusieurs façons. Par exemple, pour le trigramme *opportunité de servir*, on a, entre autres, les deux analogies suivantes qui utilisent des trigrammes différents du corpus d'entraînement et respectent bien la définition de l'analogie vue en introduction :

opportunité de servir : *opportunité de modifier* :: *qui pourrait servir* : *qui pourrait modifier*

opportunité de servir : *opportunité pour dire* :: *de servir le* : *pour dire le*

Ces deux analogies exemplifient deux patrons d'analogie différents donnés dans la table 2 et numérotés 1 et 2. Le patron 1 correspond au remplacement du bigramme correspondant à la partie gauche du premier trigramme par un autre bigramme et le remplacement de l'unigramme restant à droite par un autre unigramme : *opportunité de* est remplacé par *qui pourrait*, et *servir* est remplacé par *modifier*. Le patron 2 revient à trouver deux bigrammes dans les mêmes contextes droit et gauche : *de servir* et *pour dire* existent dans les mêmes contextes *opportunité* $\sim$ et $\sim$ *le*.

Dans le double but d'énumérer les patrons existant réellement en corpus et d'en déterminer les fréquences d'apparition respectives, nous énumérons simplement toutes les analogies existantes entre trigrammes à partir d'un échantillon aléatoire de 10 000 phrases dans chaque langue. Pour cela, nous avons contraint la méthode d'énumération de toutes les analogies d'un texte proposée dans (Gosme & Lepage, 2009) pour n'énumérer que les analogies entre trigrammes de mots. Ensuite, nous regroupons les instances d'analogies obtenues par patron et les comptons.

La table 2 donne les patrons d'analogie listés par ordre décroissant de fréquences pour l'anglais. Un résultat remarquable est que les cinq patrons d'analogie les plus fréquents dans les quatre langues apparaissent dans le même ordre avec des proportions semblables.

TABLE 2 – Patrons d'analogie entre trigrammes de mots dans un échantillon anglais de 10 000 phrases du corpus Europarl triés par proportions relatives sur l'ensemble des analogies entre trigrammes. Les symboles utilisés dans l'écriture des patrons d'analogie sont distincts deux à deux. Ces patrons respectent la définition de l'analogie donnée en introduction.

N°	A : B :: C : D	Proportion
1	*a b c* : *a b d* :: *e f c* : *e f d*	12,6 %
2	*a b c* : *a d e* :: *b c f* : *d e f*	9,1 %
3	*a b c* : *d b c* :: *e f a* : *e f d*	3,1 %
4	*a b c* : *a e c* :: *b c d* : *e c d*	2,7 %
5	*a b c* : *a b d* :: *b c e* : *b d e*	2,6 %
6	*a b c* : *a d e* :: *f b c* : *f d e*	2,4 %
7	*a b c* : *a d c* :: *b e f* : *d e f*	1,3 %
8	*a b c* : *a b d* :: *a e c* : *a e d*	0,9 %
⋮	⋮	⋮

2.3 Patrons d'analogie les plus rentables

Jusqu'à présent, nous avons montré, d'une part, qu'une grande majorité des trigrammes inconnus peuvent être reconstruits par analogie à l'aide de trigrammes issus du corpus d'entraînement (section 2.1) ; et nous avons identifié, d'autre part, les patrons d'analogie de trigrammes de mots les plus fréquents dans un même corpus (Section 2.2). L'étape suivante est d'identifier les patrons d'analogie qui permettent de reconstruire le plus de trigrammes inconnus d'un jeu de test à l'aide de trigrammes du corpus d'entraînement, autrement dit, les patrons les plus rentables. À cette fin, nous conduisons une nouvelle série d'expériences. En raison de la lourdeur en temps de calcul, nous limitons notre expérience aux cinq patrons d'analogie les plus fréquents listés dans la table 2, et nous procédons de la sorte : pour chaque trigramme inconnu du jeu de test, chaque patron d'analogie est essayé successivement dans l'ordre de la table 2. Dès lors qu'un patron d'analogie permet de reconstruire le trigramme inconnu en question, nous notons son rang et passons au trigramme inconnu suivant.

Les résultats sont présentés dans les tables 3((a))–(d)). Ils montrent les contributions cumulées des patrons d'analogie à la reconstruction des trigrammes inconnus. Le patron 1 contribue seul à la majorité de la reconstruction des trigrammes inconnus : plus de 70 % en anglais, français et allemand, mais seulement 61,5 % pour le finnois. Les patrons 1 et 2 suffisent à reconstruire environ 95 % des trigrammes inconnus en anglais, français et allemand, et presque 90 % en finnois.

TABLE 3 – Cumul des contributions des cinq patrons d'analogie les plus fréquents à la reconstruction des trigrammes inconnus dans les quatre langues étudiées. Les pourcentages présentés sont relatifs au nombre total de trigrammes inconnus.

(a) Anglais

N° de patron	Trigrammes reconstruits (μ)	Proportion cumulée
1	72 426 (63,22 %)	75,55 %
2	19 952 (17,42 %)	96,37 %
3	3 411 (2,98 %)	99,93 %
4	46 (0,04 %)	99,97 %
5	25 (0,02 %)	100,00 %
Total	95 860 (83,67 %)	100,00 %

(b) Français

N° de patron	Trigrammes reconstruits (μ)	Proportion cumulée
1	71,466 (61,98 %)	74,66 %
2	20,475 (17,51 %)	96,05 %
3	3 655 (3,13%)	99,87 %
4	92 (0,08 %)	99,97 %
5	35 (0,03 %)	100,00 %
Total	95 723 (81,87%)	100,00 %

(c) Allemand

N° de patron	Trigrammes reconstruits (μ)	Proportion cumulée
1	71 150 (50,74 %)	70,34 %
2	23 810 (16,98 %)	93,87 %
3	6 003 (4,28 %)	99,81 %
4	156 (0,11 %)	99,96 %
5	37 (0,03 %)	100,00 %
Total	101 156 (72,14 %)	100,00 %

(d) Finnois

N° de patron	Trigrammes reconstruits (μ)	Proportion cumulée
1	36 717 (27,62 %)	61,48 %
2	16 064 (12,08 %)	88,37 %
3	6 227 (4,68 %)	98,80 %
4	548 (0,41 %)	99,72 %
5	169 (0,13 %)	100,00 %
Total	59 725 (44,93 %)	100,00 %

Pour les quatre langues, les cinq patrons suffisent à reconstruire l'intégralité des trigrammes ; notre restriction se justifie donc a posteriori. Quelques exemples de reconstructions de trigrammes sont donnés dans les figures 1 et 2.

en justice et : *en est ,* :: *justice et de* : *est , de*
débat en tant : *débat de ce* :: *en tant qu'* : *de ce qu'*
coûts de la : *coûts et les* :: *de la plus* : *et les plus*

FIGURE 1 – Exemples de trigrammes de mots du corpus français d'Europarl respectant le patron 2, c'est-à-dire $\underline{a}\ b\ c : \underline{a}\ d\ e :: b\ c\ \underline{f} : d\ e\ \underline{f}$.

debate and we : *debate and far* :: *but as we* : *but as far*
Union have set : *Union have that* :: *a committee set* : *a committee that*
but they do : *but they must* :: *so we do* : *so we must*

FIGURE 2 – Exemples de trigrammes de mots du corpus anglais d'Europarl respectant le patron 1, c'est-à-dire $a\ b\ \underline{c} : a\ b\ \ \underline{d} :: e\ f\ \ \underline{c} : e\ f\ \ \underline{d}$.

2.4 Effectif suffisant pour la reconstruction d'un trigramme inconnu

Puisque l'hypothèse de la reconstruction massive des trigrammes inconnus par analogie est confirmée par les expériences précédentes, nous passons maintenant à l'étude des effectifs des trigrammes impliqués dans les reconstructions. Nous cherchons à savoir quels effectifs ont les trigrammes qui permettent la reconstruction des

trigrammes inconnus. Une supposition naturelle serait que les trigrammes d'effectifs semblables aient tendance à apparaître dans les mêmes analogies. Suivant cette supposition, on peut faire l'hypothèse que les trigrammes inconnus, c'est-à-dire apparaissant zéro fois dans le corpus d'entraînement, pourraient être reconstruits à l'aide de trigrammes apparaissant une fois dans le corpus, c'est-à-dire les trigrammes hapax. Nous confirmons ici cette hypothèse.

Nous effectuons une nouvelle série d'expériences afin d'obtenir les effectifs des trigrammes en relation d'analogie avec les trigrammes inconnus. En raison de la lourdeur des calculs, nous nous limitons à l'analyse du patron 1 : *a b c* : *a b d* :: *e f c* : *e f d*. Pour chaque instance de ce patron, nous définissons son *effectif maximum* comme l'effectif du trigramme le plus fréquent parmi les quatre trigrammes de l'analogie (comme le premier trigramme est inconnu, son effectif dans le corpus d'entraînement est évidemment zéro). Pour chaque trigramme inconnu reconstruit, nous mémorisons le minimum sur les effectifs maximums de toutes les analogies permettant de le reconstruire (effectif min-max). De cette mémorisation, et par inversion, pour chaque effectif min-max, nous pouvons compter le nombre de trigrammes inconnus reconstruits. Chaque effectif min-max est donc l'effectif suffisant à considérer pour trouver à coup sûr des trigrammes permettant la reconstruction de tant de trigrammes inconnus.

Ces décomptes sont donnés dans la table 4. Pour chaque *effectif min-max*, la table présente la quantité de trigrammes reconstruits et un pourcentage cumulé. Selon ces résultats, les instances d'analogie du patron 1 impliquant trois trigrammes hapax (effectif min-max = 1) permettent la reconstruction de plus de 95 % des trigrammes inconnus en anglais, 94 % en français ou en allemand et 91 % en finnois.

TABLE 4 – Pourcentages cumulés des trigrammes reconstruits, classés par effectif suffisant des trigrammes formant analogie pour leur reconstruction (colonne *effectif min-max*).

(a) Anglais

Effectif min-max	Trigrammes reconstruits	Pourcentage cumulé
1	54 227 (96,24 %)	96,24 %
2	1 288 (2,29 %)	98,53 %
3	345 (0,61 %)	99,14 %
4	127 (0,23 %)	99,36 %
5	99 (0,18 %)	99,54 %
⋮	⋮	⋮
523	1 (0,00 %)	100,00 %
TOTAL	56 345 (100,00 %)	—

(b) Français

Effectif min-max	Trigrammes reconstruits	Pourcentage cumulé
1	59 050 (94,07 %)	94,07 %
2	2 167 (3,45 %)	97,52 %
3	608 (0,97 %)	98,49 %
4	302 (0,48 %)	98,97 %
5	167 (0,26 %)	99,24 %
⋮	⋮	⋮
576	1 (0,00 %)	100,00 %
TOTAL	62 771 (100,00 %)	—

(c) Allemand

Effectif min-max	Trigrammes reconstruits	Pourcentage cumulé
1	41 272 (94,01 %)	94,01 %
2	1 475 (3,36 %)	97,36 %
3	465 (1,06 %)	98,42 %
4	219 (0,50 %)	98,92 %
5	124 (0,28 %)	99,21 %
⋮	⋮	⋮
412	1 (0,00 %)	100,00 %
TOTAL	43 904 (100,00 %)	—

(d) Finnois

Effectif min-max	Trigrammes reconstruits	Pourcentage cumulé
1	13 382 (91,02 %)	91,02 %
2	760 (5,217 %)	96,18 %
3	238 (1,62 %)	97,80 %
4	101 (0,68 %)	98,49 %
5	56 (0,38 %)	98,87 %
⋮	⋮	⋮
458	1 (0,01 %)	100,00 %
TOTAL	56 542 (100,00 %)	—

L'ensemble des résultats expérimentaux précédents conduit à la conclusion que non seulement les analogies entre trigrammes structurent les trigrammes inconnus, mais qu'en plus, la reconstruction des trigrammes inconnus est massivement possible avec des trigrammes d'effectif semblable, c'est-à-dire d'effectif 1.

Pour résumer l'étude empirique présentée ci-dessus, on peut donc dire que : *dans leur grande majorité les trigrammes inconnus sont analogues aux trigrammes hapax ; leurs structures et leurs effectifs sont semblables.*

3 Lissage de modèles trigrammes par analogie

Dans cette deuxième partie, nous allons exploiter les résultats de l'étude empirique précédente pour proposer une technique de lissage de modèles de langue. Notre proposition est volontairement simple et s'inspire de méthodes de lissage élémentaires : les lissages de Lidstone et de Laplace.

Habituellement, lorsqu'on utilise directement des outils tels que *SRILM* (Stolcke, 2002), on a l'habitude d'utiliser les techniques de lissage classiques connues pour donner des résultats acceptables. Des techniques de lissage plus élaborées ont été proposées afin de réduire la taille des modèles de langue, nous pensons en particulier au *clustering* (Brown *et al.*, 1992). Cependant, de telles techniques requièrent une phase de pré-traitement complexe, ce qui accroît le coût de calcul (Matsuzaki *et al.*, 2003). En comparaison, la méthode que nous proposons dans cet article n'extrait pas de connaissances supplémentaires des données d'entraînement. La structure des trigrammes inconnus est vérifiée au fil du calcul. Les principaux avantages de cette méthode sont sa simplicité et sa facilité d'utilisation.

3.1 Ré-estimation des effectifs

Redisons une vérité élémentaire : tout évènement inconnu apparaissant dans le jeu de test a un effectif nul dans le corpus d'entraînement. Immédiatement au-dessus de la classe des évènements d'effectif nul, vient la classe des évènements observés une seule fois dans le corpus d'entraînement : ce sont les hapax. Or, il est classique pour une technique de lissage d'essayer d'estimer la probabilité lissée des évènements inconnus en se basant sur les propriétés des événements classés selon leurs fréquences d'apparition : c'est la base du lissage de Good-Turing (Gale, 1994). Nous exploitons la même idée mais dans une mise en application plus simple.

Dans le lissage de Laplace, tout événement voit son effectif augmenté de 1. Dans notre technique de lissage, nous gardons cet incrément de 1 pour les événements connus. L'essence de notre technique de lissage tient dans la distinction faite entre évènements inconnus selon qu'ils peuvent être reconstruits par analogie ou non.

Nous donnons un fort avantage aux évènements inconnus qui peuvent être reconstruits par analogie au détriment de ceux qui ne peuvent pas l'être. Les résultats des expériences présentées en section 2.4 conduisent à proposer un effectif très proche de 1 pour les trigrammes reconstructibles puisqu'ils sont analogues aux trigrammes hapax Nous fixons leur effectif à $1 - \alpha$ avec α proche de 0. Ils deviennent donc de nouveaux quasi-hapax, alors que les anciens hapax sont ré-estimés avec un effectif de 1 + 1 = 2.

En désespoir de cause, nous affectons comme estimation des effectifs des trigrammes inconnus qui ne peuvent pas être reconstruits une valeur très proche de 0. Pour simplifier, nous utilisons la valeur α. On peut dire que cette partie du lissage est en fait un lissage de Lidstone.

Au total donc, la probabilité lissée d'un trigramme $h_i.m_i$ (h_i représente les deux mots précédant m_i) est ré-estimée selon chacun des trois cas suivants, avec N la longueur du texte, $|V|$ la taille du vocabulaire et δ restant à déterminer :

– trigrammes connus : $\dfrac{C(h_i.m_i)+1}{N+\delta\times|V|}$

– trigrammes inconnus pouvant être reconstruits par analogie : $\dfrac{1-\alpha}{N+\delta\times|V|}$

– trigrammes inconnus ne pouvant être reconstruits par analogie [3] : $\dfrac{\alpha}{N+\delta\times|V|}$

En reprenant les notations de la section 2.1 et de la table 1, nous notons λ la proportion de trigrammes inconnus différents et μ la proportion relative de trigrammes inconnus différents reconstruits par analogie. Les valeurs de λ et μ sont comprises entre 0 et 1. Avec ces notations :

– $(1-\lambda)$ est la proportion de trigrammes connus dans le jeu de test, λ étant la proportion de trigrammes inconnus dans le jeu de test ;
– $\mu\lambda$ est la proportion, sur l'ensemble du jeu de test, de trigrammes inconnus qui peuvent être reconstruits, μ étant la proportion de trigrammes inconnus reconstructibles ;

3. C'est en particulier le cas de tout trigramme contenant un mot inconnu. Un tel trigramme ne peut en effet être reconstruit par analogie de par la définition donnée en introduction (test sur le nombre d'occurrences des mots).

– et $(1-\mu)\lambda$ est le reste sur l'ensemble des trigrammes du jeu de test, c'est-à-dire la proportion de trigrammes inconnus ne pouvant être reconstruits par analogie.

La somme des probabilités de tous les trigrammes devant faire 1, la valeur de δ peut être déterminée :

$$\begin{aligned}\delta &= (1-\lambda)\times 1 \;+\; \mu\lambda\times(1-\alpha) \;+\; (1-\mu)\lambda\times\alpha \\ &= 1-(2\alpha\mu-\alpha-\mu+1)\lambda\end{aligned}$$

3.2 Estimation des paramètres

Dans la pratique, les paramètres λ et μ sont estimés dans une phase de pré-traitement. Le corpus d'entraînement est divisé en deux parties, l'une comprenant les neuf dixièmes du corpus, l'autre comprenant le dixième restant. La proportion de trigrammes inconnus dans la plus petite partie ainsi que la part de trigrammes inconnus reconstruits par analogie sont estimées par échantillonnage. Ces estimations deviennent les valeurs des paramètres λ et μ.

Concernant le paramètre α, des résultats d'expériences non présentés dans cet article nous ont conduits à le fixer à 10^{-6} pour toutes les langues.

3.3 Temps d'exécution

Afin de déterminer si un trigramme inconnu peut être reconstruit ou non par analogie, le corpus d'entraînement est mémorisé sous forme de deux tableaux de suffixes (sens de lecture normal et miroir). Lorsque la reconstruction d'un trigramme doit être testée, pour chaque patron d'analogie, une recherche appropriée à ce patron est effectuée dans ces tableaux de suffixes. Par exemple, pour le patron 1, le trigramme candidat *a b c* est décomposé en une partie gauche *a b* et une partie droite *c*. La recherche de ces séquences dans les tableaux de suffixes réduit aux trigrammes hapax est très rapide. Il suffit alors de prendre l'intersection en termes de positions de l'ensemble des unigrammes *d* qui suivent *a b* (sens de lecture normal) et de l'ensemble des bigrammes *e f* qui précèdent *c* (miroir). Dès qu'une position est trouvée dans l'intersection, nous en déduisons qu'il existe au moins un trigramme *e f d* dans le corpus d'entraînement et nous pouvons conclure en l'existence d'une analogie *a b c* : *a b d* :: *e f c* : *e f d*. Cela signifie que le trigramme *a b c* peut être reconstruit par analogie à l'aide de trigrammes du corpus d'entraînement. Une procédure similaire a été implantée pour le patron 2.

L'implantation des lissages classiques de *SRILM* permet de lisser environ 1 000 phrases par seconde en français quelle que soit la méthode de lissage et quelle que soit la taille du corpus d'entraînement sur une machine équipée d'un processeur 16 bits cadencé à 2 GHz et ayant 4 Go de mémoire.

Notre méthode de lissage effectue des recherches dans des tableaux de suffixes et nous nous attendons à ce que la vitesse de lissage dépende de la taille du corpus d'entraînement. Sur le même type de machine, nous mesurons la vitesse de notre méthode de lissage en fonction de la taille du corpus d'entraînement pour deux variantes : patron 1 seul et patrons 1 et 2. Nous utilisons des échantillons de la partie française d'Europarl avec des tailles variant de 900 à 320 000 phrases. Dans la seconde variante, le patron 2 est utilisé en deuxième instance dans la cas où le patron 1 n'a pas permis de reconstruire le trigramme.

Les courbes de la figure 3 donnent le nombre de phrases du jeu de test traitées par seconde en fonction de la taille du corpus d'entraînement. La vitesse de lissage de notre méthode dépend nettement de la taille du corpus d'entraînement. Pour de petits corpus, notre implantation traite 300 phrases par seconde. Cette vitesse chute à 100 phrases par seconde pour les corpus de plus grande taille et n'évolue plus vraiment à partir de 180 000 phrases. Les deux variantes sont similaires, ce qui signifie que la variante patrons 1 et 2 n'engendre qu'un faible surcoût de temps de traitement.

L'implantation actuelle de notre méthode de lissage par analogie, en Python, est donc dix fois plus lente que les implantations de *SRILM* en C++ des méthodes classiques de lissage. On peut raisonnablement espérer des temps comparables avec une implantation en C++ si l'on se fie aux règles très grossières donnant des accélérations par dix lors de réécritures de Python en C++. [4]

4. http://shootout.alioth.debian.org/u32q/benchmark.php?test=all&lang=gpp&lang2=python

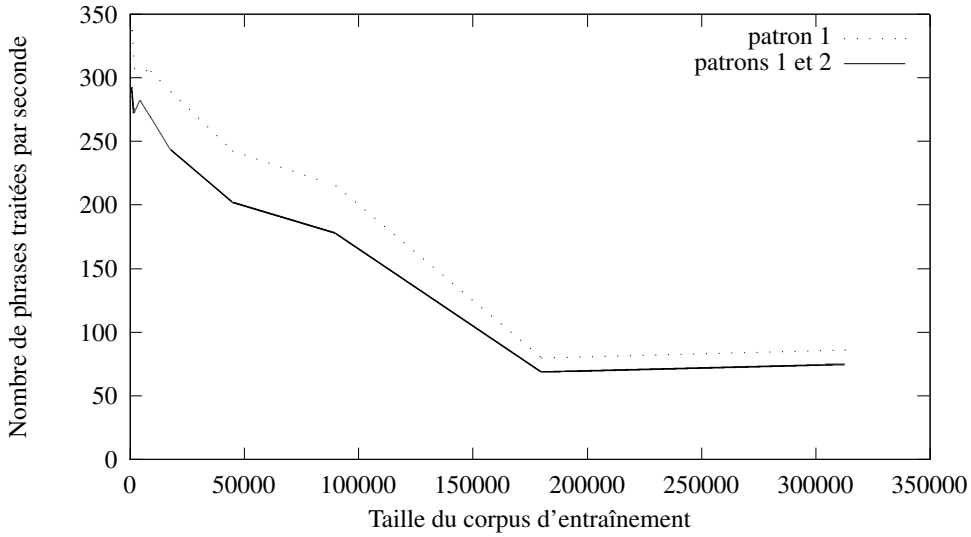


FIGURE 3 – Vitesse de la méthode de lissage par analogie pour différentes tailles du corpus d'entraînement, pour deux variantes de la méthode : patron 1 et patrons 1 et 2.

3.4 Évaluation des performances

Nous comparons notre méthode de lissage par analogie avec quatre méthodes de lissage classiques : Lidstone (Chen & Goodman, 1999), Witten & Bell (1991), Good-Turing (Gale, 1994) et Kneser & Ney (1995). Cette dernière méthode est souvent considérée comme la meilleure en pratique. [5] Pour ces quatre méthodes, nous utilisons les implantations de *SRILM* (Stolcke, 2002). Dans le cas du lissage de Lidstone, après optimisation, nous utilisons la valeur 10^{-3} pour le paramètre α pour chaque langue.

Les critères d'évaluation utilisés sont la divergence de Kullback-Leibler et la perplexité en mots. La divergence de Kullback-Leibler est définie pour chaque phrase par :

– entropie du jeu de test : $H(p) = -\sum_{i=1}^{l} p(m_i|h_i) \times \log_2 p(m_i|h_i)$;

– entropie d'un modèle de langue : $H(p,q) = -\sum_{i=1}^{l} p(m_i|h_i) \times \log_2 q(m_i|h_i)$;

– divergence de Kullback-Leibler : $D_{KL} = H(p,q) - H(p)$.

La perplexité en mots est définie comme la moyenne géométrique des inverses des probabilités réestimées. En notant n le nombre de mots du jeu de test : $PPL = 2^{\frac{-\sum_{i=1}^{n} \log_2 p(m_i|h_i)}{n}}$.

Dans ces formules, $p(m_i|h_i)$ est la probabilité conditionnelle obtenue sur le jeu de test, avec m_i le mot à la position i et h_i son histoire, c'est-à-dire les deux mots précédant m_i ; $q(m_i|h_i)$ est la probabilité conditionnelle lissée utilisant le corpus d'entraînement et l est la longueur de la phrase du jeu de test.

La comparaison est effectuée sur des données extraites d'Europarl en onze langues. Pour chaque langue, les phrases ayant une traduction en anglais sont retenues. Nous obtenons de cette manière onze corpus alignés de 383 237 phrases représentant 10 millions de mots ou plus dans chaque langue, sauf en finnois (seulement 8 millions). Chaque corpus est ensuite divisé en deux parties : 90 % du corpus pour l'entraînement, les 10 % restants servant de jeu de test. De cette manière, nos expériences sont véritablement comparables entre langues. Les statistiques concernant le corpus d'entraînement et le jeu de test pour chaque langue sont présentées dans la table 5.

Les estimations des paramètres λ et μ nécessaires à notre méthode de lissage sont détaillées dans la table 6.

5. « Kneser & Ney (1995) smoothing and its variants are generally recognized as having the best perplexity of any known method for estimating N-gram language models. » (Moore & Quirk, 2009). (Chen & Goodman, 1998) ont montré qu'une première version modifiée du lissage de Kneser-Ney « consistently had the best performance » sur l'ensemble de leurs tests et qu'une seconde version modifiée « [p]erform[ed] just slightly worse ».

TABLE 5 – Statistiques des corpus d'entraînement et des jeux de tests utilisés pour la comparaison.

	Corpus d'entraînement : 347 613 phrases			Jeux de test : 38 624 phrases		
Langue	Nbr total de mots ($\times 10^6$)	Taille du vocabulaire	Mots/phrase	Nbr total de mots ($\times 10^6$)	Taille du vocabulaire	Mots/phrase
da	9,46	153 425	27,21	1,06	46 117	27,36
de	9,51	167 942	27,36	1,06	51 398	27,48
el	10,00	149 247	28,76	1,12	52 671	28,89
en	9,94	67 819	28,60	1,11	25 854	28,76
es	10,47	100 410	30,12	1,17	37 128	30,27
fi	7,18	299 116	20,65	0,80	84 964	20,74
fr	10,95	86 567	31,51	1,22	33 403	31,65
it	9,88	99 252	28,42	1,10	36 624	28,54
nl	10,01	125 565	28,80	1,12	39 728	29,00
pt	10,29	102 800	29,59	1,15	38 041	29,73
sv	8,99	157 116	25,86	1,00	48 327	25,98

TABLE 6 – Proportion de trigrammes inconnus différents (λ) et proportion de trigrammes inconnus différents reconstructibles par analogie (μ) estimées à partir d'un échantillon de 10 % du corpus d'entraînement pour chaque langue, et valeurs correspondantes de δ ($\alpha = 10^{-6}$). Lors du calcul de δ, les valeurs de λ et μ sont ramenées entre 0 et 1.

	Trigrammes inconnus différents				
	(λ)	Reconstruits (μ)		(δ)	
		Patron 1	Patrons 1 et 2	Patron 1	Patrons 1 et 2
da	55,03 %	45,84 %	70,22 %	0,702	0,836
de	61,41 %	43,24 %	69,43 %	0,651	0,812
el	59,57 %	42,98 %	69,29 %	0,660	0,817
en	51,97 %	55,72 %	79,72 %	0,770	0,895
es	48,96 %	50,34 %	73,45 %	0,757	0,870
fi	**78,24 %**	**26,68 %**	**49,25 %**	**0,426**	**0,603**
fr	49,13 %	53,40 %	79,19 %	0,771	0,898
it	58,88 %	49,82 %	75,27 %	0,705	0,854
nl	54,56 %	52,00 %	75,94 %	0,738	0,869
pt	54,72 %	47,46 %	72,84 %	0,713	0,851
sv	60,18 %	47,25 %	71,28 %	0,683	0,827

TABLE 7 – Comparaison de la technique de lissage par analogie (patron 1 et patrons 1 et 2) avec quatre techniques de lissage classiques en onze langues.

	Perplexités en mots										
	da	de	el	en	es	fi	fr	it	nl	pt	sv
Patron 1	197,5	401,5	226,9	125,6	144,5	10099,8	106,0	149,0	181,0	141,6	334,6
Lidstone	171,0	247,1	179,3	107,4	107,6	1135,9	84,5	141,0	162,1	125,6	235,3
Witten-Bell	130,1	192,0	139,5	93,2	91,9	828,3	73,7	119,9	132,3	106,2	180,0
Good-Turing	128,9	189,2	138,1	92,6	91,0	**784,6**	**73,3**	119,1	131,0	105,3	177,7
Kneser-Ney	134,7	196,3	158,3	95,6	92,0	824,3	74,6	120,1	137,3	106,9	186,4
Patron 1 et 2	**107,8**	**182,4**	**116,4**	**90,9**	**85,8**	2876,6	73,7	**81,0**	**99,2**	**79,7**	**152,6**
	Divergences de Kullback-Leibler										
Patron 1	61,2	73,1	73,5	52,2	56,7	121,9	55,8	68,8	62,9	62,3	70,3
Lidstone	54,2	66,3	63,2	44,2	47,9	95,7	45,0	56,7	54,8	53,0	60,6
Witten-Bell	47,0	60,0	56,2	40,6	43,4	89,0	41,0	52,5	49,4	48,4	54,1
Good-Turing	46,5	59,3	55,7	40,0	42,9	87,6	40,5	52,0	48,8	47,8	53,5
Kneser-Ney	46,4	58,8	55,9	40,2	43,0	**87,2**	**40,4**	51,9	48,8	47,8	53,5
Patron 1 et 2	**43,5**	**50,2**	**51,8**	**38,7**	**41,5**	105,7	41,2	**48,5**	**44,8**	**44,3**	**48,4**

Nous rappelons que λ est la proportion de trigrammes inconnus différents et que μ est la proportion relative de trigrammes inconnus différents qui peuvent être reconstruits par analogie. Nous considérons deux variantes de notre méthode : la première n'utilise que le patron 1, la seconde utilise les patrons 1 et 2. Afin de rendre la technique de lissage indépendante du jeu de test, pour chaque langue les estimations des paramètres ont été obtenues automatiquement à partir d'un échantillon aléatoire formé d'un dixième du corpus d'entraînement comme décrit dans la section 3.2. Comme le montrent les chiffres de la table 6, l'utilisation du patron 2 en plus du patron 1 augmente sensiblement la valeur du paramètre μ : plus d'un quart en valeurs absolues. À l'exception du finnois, l'utilisation conjointe des patrons 1 et 2 permet la reconstruction de 70 % à 80 % des trigrammes inconnus. Les valeurs pour le finnois, en gras dans la table, sont nettement différentes des valeurs pour les autres langues.

Les résultats de l'évaluation des deux variantes de la méthode proposée sont présentés dans la table 7 :
– le patron 1 seul est insuffisant pour atteindre même le niveau du lissage de Lidstone. On obtient systématiquement les plus mauvais résultats dans les onze langues ;
– à l'exception du finnois, et dans une moindre mesure du français, l'ajout du patron 2 est suffisant pour obtenir des résultats bien meilleurs que ceux des quatre méthodes de lissage classiques.

La contre-performance sur le finnois n'est pas surprenante si l'on considère le nombre important de trigrammes inconnus et la faible proportion de ces trigrammes qui peuvent être reconstruits par analogie (voir table 6). Afin de remédier à ce problème, plutôt que d'accroître la quantité de données d'entraînement, il serait sans doute plus judicieux de segmenter les mots en morphèmes. Quant aux résultats en français, ils sont comparables à ceux des méthodes classiques.

4 Conclusion et perspectives

Dans cet article, à l'aide d'une série d'expériences sur quatre langues, nous avons montré qu'en majorité les trigrammes inconnus dans un jeu de test sont structurellement analogues aux trigrammes hapax d'un corpus d'entraînement.

De cette propriété, nous avons dérivé une méthode de lissage pour modèles de langue trigrammes. L'effectif des trigrammes connus est ré-estimé en appliquant un incrément de 1 comme dans le lissage de Laplace. Les trigrammes inconnus qui peuvent être reconstruits par analogie sont considérés comme quasi-hapax : leurs effectifs sont ré-estimés à une valeur proche de 1. Les trigrammes inconnus qui ne peuvent être reconstruits par analogie sont presque ignorés, leurs effectifs étant fixés à une valeur proche de 0 comme dans le lissage de Lidstone. En comparaison de techniques de lissage utilisant des techniques de *clustering*, cette méthode est simple ; elle ne construit que deux classes de trigrammes inconnus : ceux qui peuvent être reconstruits et les autres.

Des mesures sur onze langues ont montré que cette méthode de lissage donne de bons résultats en comparaison des techniques de lissage classiques, sauf dans le cas du finnois.

L'étude présentée ici laisse un certain nombre de points à examiner. Tout d'abord, cette étude a été consacrée aux trigrammes. Or, aujourd'hui, dans de nombreux domaines du traitement automatique des langues, comme par exemple la traduction automatique par approche statistique, on utilise des modèles de langue 5-grammes. Des expériences restent donc à effectuer avec des n-grammes d'ordres supérieurs pour savoir si de bons résultats peuvent aussi être obtenus. L'influence du nombre de patrons d'analogie sur l'entropie des modèles de langue obtenus reste elle aussi à étudier. Un autre point porte sur la taille des corpus utilisés. Les expériences rapportées ici visant à une comparaison sur plusieurs langues et les très grands corpus multilingues étant rares, la taille du corpus utilisé ici est relativement faible en regard de corpus monolingues dépassant le milliard de mots. Des expériences sur des corpus de tailles plus importantes restent donc à effectuer. Enfin, dans la perspective d'une intégration à la traduction automatique par approche statistique, les pouvoirs discriminants de la technique de lissage proposée ici restent à examiner.

5 Remerciements

Cet article décrit des résultats de recherche obtenus en partie grâce une subvention de l'université Waseda pour projets de recherche spécifiques (projet numéro : 2010A-906).

Références

BROWN P., PIETRA V., DESOUZA P., LAI J. & MERCER R. (1992). Class-based *n*-gram models of natural language. *Computational linguistics*, **18**(4), 467–479.

CHEN S. F. & GOODMAN J. (1998). *An empirical study of smoothing techniques for language modeling*. Rapport interne, Harvard university, Cambridge, Massachussetts.

CHEN S. F. & GOODMAN J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, **13**(4), 359–394.

CLAVEAU V. & L'HOMME M.-C. (2005). Terminology by analogy-based machine learning. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*, Copenhagen.

DENOUAL E. (2007). Analogical translation of unknown words in a statistical machine translation framework. In *Proceedings of Machine Translation Summit XI*, Copenhagen.

GALE W. (1994). Good-turing smoothing without tears. *Journal of Quantitative Linguistics*, **2**.

GOSME J. & LEPAGE Y. (2009). A first study of the complete enumeration of all analogies contained in a text. In *4th Language and Technology Conference (LTC 2009)*, p. 401–405, Poznań, Poland.

HATHOUT N. (2009). Acquisition morphologique à partir d'un dictionnaire informatisé. In T. NAZARENKO, D. ET POIBEAU, Ed., *Actes de la 16e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2009)*, p. 10 p. : ATALA.

KNESER R. & NEY H. (1995). Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1.

KOEHN P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the tenth Machine Translation Summit (MT Summit X)*, p. 79–86, Phuket.

LANGLAIS P. & PATRY A. (2007). Translating unknown words by analogical learning. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 877–886.

LANGLAIS P., YVON F. & ZWEIGENBAUM P. (2008). *Analogical Translation of Medical Words in Different Languages*, volume 5221/2008 of *Lecture Notes in Computer Science*, p. 284–295. Springer Berlin / Heidelberg : Springer Berlin / Heidelberg.

LAVALLÉE J. F. & LANGLAIS P. (2010). Analyse morphologique non supervisée par analogie formelle. In *TALN 2010*, p. 10 pages, Montréal, Québec, Canada.

LEPAGE Y. (2004). Analogy and formal languages. *Electronic notes in theoretical computer science*, **53**, 180–191.

LEPAGE Y. & DENOUAL E. (2005). Purest ever example-based machine translation : detailed presentation and assessment. *Machine Translation*, **19**, 251–282.

LEVENSHTEIN V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10**(8), 707–710.

MATSUZAKI T., MIYAO Y. & TSUJII J. (2003). An efficient clustering algorithm for class-based language models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, volume 4, p. 119–126 : Association for Computational Linguistics.

MOORE R. & QUIRK C. (2009). Improved smoothing for N-gram language models based on ordinary counts. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, p. 349–352 : Association for Computational Linguistics.

ROSENFELD R. (2000). Two decades of statistical language modelling : where do we go from here ? *Proceedings of the IEEE*, **88**(8), 1270–1278.

STOLCKE A. (2002). SRILM-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, volume 3, p. 901–904.

TURNEY P. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, p. 905–912, Manchester, UK : Coling 2008 Organizing Committee.

WITTEN I. & BELL T. (1991). The zero-frequency problem : Estimating the probabilities of novel events in adaptive text compression. *Information Theory, IEEE Transactions on*, **37**(4), 1085–1094.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Évaluer la pertinence de la morphologie constructionnelle dans les systèmes de Question-Réponse

Delphine Bernhard[1] Bruno Cartoni[2] Delphine Tribout[1]
(1) LIMSI-CNRS, 91403 Orsay, France
(2) Département de linguistique, Université de Genève, Suisse
bernhard@limsi.fr, bruno.cartoni@unige.ch, tribout@limsi.fr

Résumé. Les connaissances morphologiques sont fréquemment utilisées en Question-Réponse afin de faciliter l'appariement entre mots de la question et mots du passage contenant la réponse. Il n'existe toutefois pas d'étude qualitative et quantitative sur les phénomènes morphologiques les plus pertinents pour ce cadre applicatif. Dans cet article, nous présentons une analyse détaillée des phénomènes de morphologie constructionnelle permettant de faire le lien entre question et réponse. Pour ce faire, nous avons constitué et annoté un corpus de paires de questions-réponses, qui nous a permis de construire une ressource de référence, utile pour l'évaluation de la couverture de ressources et d'outils d'analyse morphologique. Nous détaillons en particulier les phénomènes de dérivation et de composition et montrons qu'il reste un nombre important de relations morphologiques dérivationnelles pour lesquelles il n'existe pas encore de ressource exploitable pour le français.

Abstract. Morphological knowledge is often used in Question Answering systems to facilitate the matching between question words and words in the passage containing the answer. However, there is no qualitative and quantitative study about morphological phenomena which are most relevant to this application. In this paper, we present a detailed analysis of the constructional morphology phenomena found in question and answer pairs. To this aim, we gathered and annotated a corpus of question and answer pairs. We relied on this corpus to build a gold standard for evaluating the coverage of morphological analysis tools and resources. We detail in particular the phenomena of derivation and composition and show that a significant number of derivational morphological relations are still not covered by any existing resource for the French language.

Mots-clés : Évaluation, Morphologie, Ressources, Système de Question-Réponse.

Keywords: Evaluation, Morphology, Resources, Question-answering system.

1 Introduction

Les systèmes de Question-Réponse (QR) ont pour objectif de fournir une réponse précise à une question. Pour ce faire, ils reposent généralement sur un composant de recherche d'information (RI) qui vise à apparier les mots de la question avec les mots des documents contenant la réponse potentielle. La principale difficulté pour les systèmes de RI réside dans le fait qu'une réponse peut se trouver dans un document qui ne reprend pas forcément les mots de la question. Les systèmes de RI et de QR doivent donc pouvoir récupérer les documents pertinents sans se baser uniquement sur l'identité formelle entre les mots de la question et les mots du document. À cette fin, la morphologie a souvent été préférée à une analyse sémantique plus complexe dans la mesure où deux mots reliés morphologiquement montrent généralement une similitude formelle qui permet de prendre en compte facilement leur relation sémantique. Les systèmes de RI et de QR intègrent donc généralement des connaissances morphologiques, que ce soit lors de l'indexation des documents ou lors de la recherche, en étendant les requêtes ou en les reformulant au moyen de mots morphologiquement reliés. Cette intégration est généralement effectuée de manière très générique, c'est-à-dire que toutes les relations morphologiques possibles, ou pour lesquelles on dispose d'une ressource, sont incluses. Par ailleurs, les évaluations sont effectuées de manière globale, en évaluant l'amélioration de la performance globale du système, et non l'impact de cet ajout.

La plupart des recherches menées dans ce domaine utilisent des techniques de désuffixation (*stemming*) basées sur des heuristiques simples qui suppriment la fin des mots (Lennon *et al.*, 1988; Harman, 1991; Fuller & Zobel,

1998). Ces méthodes s'avèrent efficaces pour les langues à morphologie moins riche comme l'anglais, mais ne sont pas disponibles pour toutes les langues (McNamee *et al.*, 2009). La plupart du temps l'utilisation de ces méthodes permet d'augmenter légèrement le rappel, mais ces techniques génèrent également du bruit. Bilotti *et al.* (2004) ont par exemple montré que des mots relativement éloignés comme *organisation* et *organ* sont réduits à la même racine par l'algorithme de désuffixation de Porter. Moreau & Claveau (2006) ont quant à eux utilisé une méthode d'acquisition automatique de connaissances morphologiques par apprentissage, et leur étude a montré que l'utilisation des connaissances morphologiques pour étendre les requêtes améliore les résultats pour la plupart des langues européennes qu'ils ont testées.

Dans chacune des études précédentes, les mots de la question sont étendus à l'ensemble des mots appartenant à la même famille morphologique, et les différents types de procédés (tels que la formation d'un nom déverbal ou la formation d'un nom désadjectival) ne sont pas distingués. Ainsi, tous les mots appartenant à la même famille morphologique sont considérés comme sémantiquement proches. Or, nous pensons que tous les mots morphologiquement reliés n'ont pas la même proximité sémantique et que certaines relations morphologiques sont plus pertinentes que d'autres dans le cadre de QR. Par exemple dans ce type de tâche, il nous semble plus pertinent d'étendre une requête contenant le verbe *diviser* au nom d'événement dérivé *division*, plutôt qu'à l'adjectif dérivé *divisible*. Cependant, à notre connaissance, aucune étude qualitative ni quantitative n'a été menée en ce sens, afin de déterminer quels types de relations morphologiques sont pertinents pour la recherche en QR.

Au delà de l'extension de requête, la morphologie trouve également sa place dans les méthodes de reformulation automatique des questions, qui visent à traiter des phénomènes de paraphrase entre question et réponse. Ainsi Ravichandran & Hovy (2002) proposent une méthode d'acquisition de patrons de reformulation de surface pour des types précis de questions. Ces patrons incluent entre autres des mots morphologiquement reliés tels que *discover*, *discovery* et *discoverer*, dans le cas d'une question portant sur la personne ayant fait une découverte donnée. Ces patrons sont ensuite utilisés pour extraire la réponse. Une approche similaire est proposée par Lin & Pantel (2001) et Hermjakob *et al.* (2002). Ces travaux ne se focalisent toutefois pas sur la morphologie et ne proposent donc pas d'évaluation spécifique. Seul Jacquemin (2010) évalue l'apport de la mophologie dans ce contexte en utilisant le lexique des verbes français de Dubois & Dubois-Charlier (1997) et les relations de dérivation qu'il contient pour automatiquement reformuler des énoncés, sur la base des relations de dépendance syntaxique.

Dans cet article, nous présentons les résultats d'une évaluation portant sur la *pertinence* des connaissances morphologiques dans un système de QR. Ces résultats permettent d'une part de déterminer quels types de ressources morphologiques sont nécessaires à l'amélioration des systèmes de QR, et d'autre part d'évaluer la couverture des ressources existantes pour une telle tâche.

Pour évaluer la pertinence des connaissances morphologiques dans un système de QR, nous avons tout d'abord constitué un corpus de paires question-passage contenant la réponse à partir de trois collections de données issues de campagnes d'évaluation en QR. Nous avons ensuite annoté ce corpus afin de déterminer quelles sont les relations morphologiques les plus fréquentes qui relient les mots de la question et les mots du passage. Enfin, nous avons analysé les résultats de cette annotation et évalué la couverture des ressources existantes en français pour les procédés morphologiques observés [1].

2 Constitution et annotation d'un corpus de paires question-passage réponse

2.1 Corpus de questions-passages

Nous avons constitué notre corpus de paires question-passage réponse à partir de trois collections de données utilisées pour l'évaluation de systèmes QR : Quæro, EQueR et Conique. Ces trois collections et le corpus qu'elles nous ont permis de constituer sont décrits ci-dessous. La table 1 présente les statistiques concernant ces trois collections.

1. Cet article présente les résultats de l'évaluation uniquement pour des *ressources morphologiques*. Les mêmes données de référence ont été utilisées pour évaluer des *outils d'analyse morphologique* (Bernhard *et al.*, à paraître)

	Quæro	EQueR-Medical	Conique
#Questions	350	200	201
#paires de question-passage	566	394	664
Longueur moyenne de la question	8,8	9,9	11,4
Longueur moyenne du passage réponse	38,5	29,0	92,4

TABLE 1 – Statistiques sur les sous-corpus de questions-réponses utilisés

2.1.1 Quæro

Le corpus français Quæro a été constitué dans le cadre du projet Quæro avec pour objectif d'évaluer des systèmes de QR (Quintard *et al.*, 2010). Le corpus de documents contient 2,5 millions de documents en français extraits de l'Internet et 757 questions, dont 250 pour la campagne de 2008 et 507 pour celle de 2009. La collection de documents a été constituée en prenant les 100 premières pages retournées par le moteur de recherche Exalead pour une série de requêtes trouvées dans les logs du moteur. Quant aux questions, elles ont été rédigées par des francophones, sur la base du contenu des documents pour la campagne 2008, et sur la base du log uniquement pour la campagne de 2009. Trois types de questions ont été formulées : des questions factuelles, des questions booléennes attendant une réponse de type oui/non, et des questions de type liste. Pour notre tâche d'annotation, nous avons constitué des paires question-passage, formées de l'ensemble des questions factuelles et des passages contenant la réponse qui ont été fournis par les systèmes et validés manuellement lors des deux campagnes d'évaluation 2008 et 2009. Nous avons ainsi obtenu 566 paires de questions-passages contenant la réponse, 338 pour la campagne 2008 et 228 pour la campagne 2009.

2.1.2 EQueR-Medical

Les données du corpus EQueR ont été constituées dans le cadre de la campagne d'évaluation EQueR-EVALDA pour les systèmes de question-réponse du français (Ayache *et al.*, 2006). La campagne comprenait deux tâches principales : (i) question-réponse générale sur une collection d'articles de journaux et de rapports sénatoriaux et (ii) question-réponse spécialisée sur une collection de textes médicaux. Pour ces deux tâches, les passages contenant les réponses retournées par les systèmes participants ont été validés manuellement par des spécialistes du domaine. Pour notre étude, seule la partie médicale a été retenue, constituant ainsi un ensemble de 394 paires de questions-passages, pour un total de 200 questions distinctes.

2.1.3 Conique

Le corpus Conique a été constitué dans le but d'étudier les justifications pertinentes pour les réponses des systèmes de QR (Grappy *et al.*, 2010). Les justifications des réponses fournissent un matériel supplémentaire pour l'utilisateur, afin qu'il ou elle puisse faire confiance à la réponse fournie par le système. Le corpus est basé sur un sous-ensemble de 291 questions de la campagne EQueR pour le français (Ayache *et al.*, 2006) et de plusieurs campagne CLEF. Les passages-réponses candidats ont été extraits de la version française de Wikipedia à l'aide d'un système de RI et ont ensuite été annotés par 7 annotateurs. Contrairement aux deux corpus décrits précédemment, les passages-réponses de Conique ne correspondent pas à une sortie d'un système de QR. Le corpus possède donc un taux de rappel extrêmement haut, et est exempt de tout biais inhérent aux systèmes de QR, comme les taux importants de mots identiques entre les questions et le passage. Nous avons pré-traité ce corpus, pour ne conserver que les justifications complètes ou partielles. De plus, nous avons réduit le passage à une longueur de trois phrases. Au total, le corpus constitué à partir de la collection Conique contient 664 paires de question-passage, pour 201 questions distinctes.

2.2 Annotation

Pour chaque paire question-passage réponse, nous avons manuellement annoté les mots de la question et les mots du passage afin de déterminer quels mots sont morphologiquement reliés et par quels types de relations.

Les annotations ont été effectuées par trois annotateurs indépendants [2] au moyen de l'outil d'alignement YAWAT (Germann, 2008). Cet outil a été initialement conçu pour aligner les mots de paires de phrases bilingues pour des campagnes d'évaluation de traduction automatique. Dans cette étude, nous l'avons utilisé pour aligner les mots (ou groupes de mots) dans des paires question-passage réponse, et pour assigner à ces paires de mots une étiquette, parmi les trois types de relations morphologiques suivantes : flexion, dérivation, composition. Les figures 1 à 3 présentent des exemples de paires question-passage impliquant respectivement des relations de flexion, dérivation et composition.

Q : Quand **est né** Philippe d'Orléans ?
R : Philippe d'Orléans **naquit** le 2 août 1674.

Q : Comment un *insuffisant* **rénal** doit-il être suivi ?
R : Du fait du risque de transmission nosocomiale du VHC chez les *insuffisants* **rénaux** hémodialysés et chez les transplantés rénaux, une surveillance annuelle de la sérologie doit être réalisée.

Q : À combien de milliards de dollars s'élève le **déficit** budgétaire américain ?
R : Politique budgétaire. Le PIB des États-Unis s'élève à environ 10 000 milliards de dollars et les **déficits** atteindraient au moins 300 ou 400 milliards de dollars en 2003

FIGURE 1 – Exemples de paires question-passage impliquant une relation de flexion

Q : En quelle année Martin Luther King **a**-t-il **été assassiné** ?
R : Il dit avoir été à côté du pasteur King à Memphis lors de son **assassinat**, le 4 avril 1968. Il est ordonné ministre baptiste à la fin de cette même année.

Q : Quels sont les quatre **réalisateurs** du film "Le jour le plus long" ?
R : Le Jour le plus long (The Longest Day) est un film américain **réalisé** par Ken Annakin, Andrew Marton, Bernhard Wicki et Gerd Oswald sorti en salle en 1962...

Q : La pose d'amalgame dentaire peut-elle provoquer des **allergies** ?
R : Il est certain que la pose d'amalgames peut entraîner des réactions **allergiques** plus ou moins graves et prononcées chez les patients.

FIGURE 2 – Exemples de paires question-passage impliquant une relation de dérivation

Q : Où Marcos fut-il **dictateur** ?
R : Imelda Marcos, le 22 février 2006. Imelda Romualdez Marcos (née le 2 juillet 1929) fut la femme de Ferdinand Marcos, **président-dictateur** des Philippines de 1965 à 1986.

Q : Le mercure est-il un métal **toxique** ?
R : En grande concentration ou lorsque l'exposition est prolongée, le mercure a des effets **neuro-toxiques** connus, principalement dans sa forme organique, soit le méthylmercure

Q : Qu'engendre la corticothérapie sur l'**os** ?
R : Les fractures de l'**ostéoporose** cortisonique surviennent au moins en partie en raison d'une perte osseuse induite par la corticothérapie

FIGURE 3 – Exemples de paires question-passage impliquant une relation de composition

Étant donné que deux mots morphologiquement liés peuvent être reliés par plus d'une relation, des instructions spécifiques ont également été définies. Ainsi, nous n'avons pas annoté les variantes flexionnelles des auxiliaires et des déterminants, dans la mesure où ils sont très fréquents et apportent donc peu d'information sémantique. Nous

2. Co-auteurs de cet article.

avons également décidé de donner la priorité aux relations de dérivation et de composition sur les relations flexionnelles. Par exemple, dans la paire de question-passage présentée à la Figure 4, il y a deux étapes morphologiques entre le nom *Australie* dans la question et l'adjectif féminin *australienne* dans la réponse : la première étape est la dérivation de l'adjectif *australien* à partir du nom propre ; la seconde est la flexion de l'adjectif dérivé au féminin. Dans ce cas, la relation morphologique la plus pertinente est la relation dérivationnelle entre le nom propre et l'adjectif, c'est pourquoi dans un tel cas seule la relation dérivationnelle a été annotée. Enfin, une étiquette spécifique "autre" a été utilisée pour annoter des mots qui ne sont par directement reliés morphologiquement, mais qui sont le résultat de deux constructions à partir de la même base. Par exemple dans la paire présentée à la Figure 5 le nom *utilité* et le verbe *utiliser* (sous la forme passive *est utilisée*) ne dérivent pas l'un de l'autre, mais sont tous deux dérivés de l'adjectif *utile*.

Q : Quelle est la capitale de l'**Australie** ?
R : le territoire sur lequel est située la capitale fédérale **australienne**, Canberra .

FIGURE 4 – Exemple de paires question-passage où deux types de relation sont présentes (flexion et dérivation)

Q : Quelle est l'**utilité** de la pierre d'alun ?
R : La pierre d'alun est un excellent déodorant corporel. Elle **est utilisée** pour neutraliser la transpiration, empêcher la fermentation et éliminer les mauvaises odeurs.

FIGURE 5 – Exemple de paires question-passage où la relation morphologique implique deux constructions différentes à partir de la même base

Nous avons mesuré la qualité des annotations à l'aide du coefficient kappa de Fleiss (Fleiss, 1971) [3]. Le coefficient κ varie en fonction du corpus et du type de relation considéré : il est fort (0,674) à presque parfait (0,83) pour la flexion, bon pour la dérivation (0,662 à 0,729) et faible (0,39) à bon (0,665) pour la composition.

Tous les désaccords ont été résolus et les données validées par l'ensemble des annotateurs afin de constituer un corpus de référence. Nous avons ensuite classé et caractérisé les paires de mots que nous avions considérés comme morphologiquement reliés lors de l'annotation. Les résultats de cette analyse sont présentés dans la section suivante et fournissent un panorama des relations morphologiques mises en jeu dans le cadre de Question-Réponse.

3 Analyse des résultats

Au terme de l'annotation, nous avons obtenu un ensemble de mots morphologiquement reliés jouant un rôle dans le lien opéré entre la question et le passage contenant la réponse. Plusieurs observations peuvent être faites suivant différents points de vue. Nous présentons tout d'abord la répartition des différentes relations morphologiques observées (flexion, dérivation et composition). Puis nous décrivons précisément les procédés de dérivation et de composition les plus fréquemment observés. Enfin, nous étudions la position du mot construit dans la paire et en particulier s'il se trouve dans la question ou dans le passage réponse.

3.1 Types de relations morphologiques

Les résultats de l'annotation de chaque sous-corpus en fonction des différents types de relations morphologiques sont présentés dans la table 2 [4]. Ces chiffres montrent que chaque sous-corpus semble favoriser un type particulier de relation morphologique : le sous-corpus Conique contient un majorité de relations de dérivation, le sous-corpus Quæro contient davantage de flexion, alors que pour le sous-corpus EQueR, c'est la composition qui semble la plus fréquente. De plus, si l'on étudie les relations morphologiques en fonction des sous-corpus, on constate que la composition est quasiment absente des sous-corpus Quæro et Conique.

3. Le kappa de Fleiss permet de mesurer l'accord inter-annotateurs lorsqu'il y a plus de deux annotateurs. Il a été calculé en fonction de l'accord des annotateurs sur la présence d'une paire de mots morphologiquement reliés pour une même paire de questions-réponse

4. Les paires de question-passage (paire qp) ne contiennent pas toujours des relations morphologiques, et certaines paires peuvent contenir plus d'une relation morphologique, impliquant parfois les mêmes mots.

Corpus (paire qp)	Flexion		Dérivation		Composition	
	nbr	%	nbr	%	nbr	%
Conique (664)	159	41,8	188	**49,5**	33	8,7
Quæro (566)	136	**61,8**	80	36,4	4	1,8
EQueR (394)	69	26,5	81	31,0	111	**42,5**

TABLE 2 – Flexion, dérivation et composition dans les trois sous-corpus

Il est notable que le sous-corpus Conique contient plus de relations de dérivation que le sous-corpus Quæro. Ceci est lié à la manière dont le corpus Conique a été construit. En effet, il n'a pas été constitué à partir des réponses fournies par un système de QR, mais sur la base de réponses identifiées et annotées manuellement. De plus, Conique contient en moyenne les passages les plus longs (c.f. table 1), ce qui peut expliquer la présence d'un nombre plus important de paires dérivationnelles dans ce sous-corpus. Quant à EQueR, la proportion importante de mots composés est liée au domaine de spécialité du sous-corpus qui contient un grand nombre de termes médicaux, ceux-ci étant souvent composés, comme le montre la figure 6.

Q : Quelle est la conséquence de la **corticothérapie** sur l'*os* ?
R : Le problème essentiel des **corticoïdes** réside dans leurs effets secondaires (... *ostéoporose*, *ostéonécrose* aseptique des têtes fémorales ou parfois humérales ...).

FIGURE 6 – Exemple de paire de question-passage de EQueR

Il est également intéressant de noter le rôle important de la dérivation dans les trois sous-corpus (entre 31 % et 49 %), et l'importance de la composition dans le domaine médical (42,5% dans EQueR). Ceci confirme l'intérêt d'inclure des connaissances morphologiques de ce type dans un système de QR.

Dans la suite, nous décrivons plus précisément ces deux types de construction, en analysant quels procédés morphologiques sont les plus présents. Nous ne nous attardons pas sur la morphologie flexionnelle dans le cadre de cet article car elle est considérée comme pertinente par les systèmes de QR existants. De plus, elle est généralement bien prise en compte par les systèmes de QR existant, notamment *via* l'utilisation de lemmatiseurs.

3.2 Dérivation

Comme nous venons de le montrer (table 2), la dérivation joue un rôle important dans les trois sous-corpus. Dans certain cas, la relation morphologique entre le mot de la question et le mot du passage-réponse implique plus d'une étape dérivationnelle : soit l'un des mots est plus complexe que l'autre, mais n'est pas un dérivé direct (par exemple, *lune* et *alunissage*, ce dernier dérivant du verbe *alunir*, lui-même dérivé de *lune*) ; soit les deux mots sont complexes et sont tous deux dérivés d'un même mot (par exemple *joueur* et *jouable* tous deux dérivés du verbe *jouer*). Le premier cas de figure représente 1,70% des relations observées et le second 8,30%, ce qui représente une proportion assez faible et indique que dans la grande majorité des cas les mots morphologiquement reliés entretiennent une relation de dérivation directe.

Les relations non directes impliquent une prédictibilité moindre, et influencent le choix des méthodes d'implémentation, comme nous l'expliquons dans la section 3.4. Dans un premier temps nous étudions donc uniquement les paires de mots morphologiquement reliés par une dérivation directe.

La table 3, qui présente la proportion des différents types de procédés morphologiques observés, montre que les sous-corpus diffèrent selon les procédés de dérivation majoritairement utilisés. Si Conique contient une majorité d'adjectifs dénominaux (environ 47% des procédés de dérivation), Quæro et EQueR montrent une préférence pour des procédés de nominalisation (avec respectivement 61% et 54% des procédés de dérivation).

	Exemple	Conique (174)		Quæro (70)		EQueR (70)	
		nbr	%	nbr	%	nbr	%
Nom > Adj	commerce > commercial	37	21	16	23	**28**	**40**
Nom propre> Adj	Afrique > africain	**45**	**26**	8	11,5	1	1
Nom > Nom	président > présidente	29	17	5	7	2	3
Nom propre > N	Arménie > Arméniens	6	3	8	11,5	2	3
Adj > Nom	national > nationalité	3	2	0	0	9	13
Verbe > Nom	traiter > traitement	41	24	**30**	**43**	25	36
Autres	complet > complètement	13	7	3	4	3	4

TABLE 3 – Procédés dérivationnels dans les paires de question-passage

3.2.1 Adjectifs dénominaux

Dans les trois sous-corpus, les adjectifs dérivés d'un nom propre sont toujours des adjectifs relationnels, qui peuvent être remplacés par un complément du nom équivalent, comme *chilien* dérivé de *Chili* ou *africain* dérivé de *Afrique*. Les adjectifs dérivés d'un nom commun sont la plupart du temps relationnels, comme le montrent les chiffres de la table 4. Par exemple, *commercial* dérivé de *commerce* ou *solaire* dérivé de *soleil*. Cependant, on trouve également quelques adjectifs qualificatifs comme *âgé* dérivé de *âge* ou *montagneux* dérivé de *montagne*. La table 4 présente les proportions d'adjectifs relationnels et qualificatifs dans notre corpus, et montre que ce sont les adjectifs relationnels qui sont les plus fréquents dans les trois sous-corpus.

	Adj. relationel		Adj. qualificatif	
	nbr	%	nbr	%
Conique (37)	23	62	14	38
Quæro (16)	10	62	6	38
EQueR (28)	24	86	4	14
Total (81)	57	70	24	30

TABLE 4 – Types d'adjectifs dénominaux

3.2.2 Procédés de formation des noms

En ce qui concerne les procédés de formation de noms, on trouve dans les trois sous-corpus un grand nombre de nominalisations déverbales, ainsi que quelques cas de nominalisations désadjectivales ou dénominales. La formation de noms sur des bases nominales est relativement rare, sauf dans Conique qui contient plusieurs noms de profession féminisés, comme *infirmier* et *infirmière*, *directeur* et *directrice*, *président* et *présidente*, que nous avons considérés comme dérivés l'un de l'autre et non comme deux formes fléchies du même mot. Nous avons également trouvé quelques noms diminutifs, comme *rame > ramette* et quelques préfixés comme *président > vice-président*. Nous avons aussi considéré les formations de noms à partir de noms propres comme des nominalisations. Ces noms dérivés sont principalement des gentilés comme *Colombien* dérivé de *Colombie*.

Les noms désadjectivaux sont rares dans les trois sous-corpus, voire inexistants dans le cas de Quæro. Ces noms désadjectivaux sont principalement des noms de propriété, comme *toxicité* construit sur *toxique*. La plupart des noms désadjectivaux se trouvent dans le corpus EQueR. Cela s'explique par le fait que le corpus médical contient beaucoup de noms de maladie ou de pathologie (comme *toxicité* ou *insuffisance*), or ces noms réfèrent la plupart du temps à la propriété de se trouver dans un état particulier (*toxicité* $\approx$ 'propriété d'être toxique', *insuffisance* $\approx$ 'propriété d'être insuffisant').

Quant aux noms déverbaux, qui sont les plus fréquents, ce sont essentiellement des noms d'événements, comme *débarquement* dérivé du verbe *débarquer*. Les noms d'événements représentent presque 85% des noms déverbaux, comme le montre la table 5. Cependant, on trouve également un petit nombre de noms d'agents dans les sous-corpus Conique et Quæro, comme *réalisateur* construit sur *réaliser*, et quelques cas de noms résultatifs comme *produit* dérivé de *produire*.

	Exemple	Conique (41)		Quæro (30)		EQueR (25)	
		nbr	%	nbr	%	nbr	%
Verbe > N événement	inaugurer > inauguration	**34**	**83**	**25**	**83**	**22**	**88**
Verbe > N agent	réaliser > réalisateur	4	10	4	13	0	0
Verbe > N autre	produire > produit	3	7	1	4	3	12

TABLE 5 – Types sémantiques des noms déverbaux dans les paires de questions-passage

3.2.3 Autres procédés de formation

Parmi les autres procédés de formation observés dans le corpus, on trouve des formations d'adverbes, comme *complètement* dérivé de *complet* ou *directement* construit sur *direct*, ainsi qu'un certain nombre de verbes préfixés, comme *déboucher*, ou d'adjectifs préfixés, comme *international*. Il est également intéressant de noter que nous n'avons observé aucun verbe désadjectival (comme *national* > *nationaliser*) et très peu de verbes dénominaux (quatre cas seulement dans le sous-corpus Conique, dont trois sont des verbes convertis : *border*, *fusionner* et *suicider*). La quasi-absence de verbes dénominaux peut s'expliquer par la faible prédictibilité du sens d'un verbe dénominal. Comme l'ont décrit Hopper & Thompson (1984), il existe une asymétrie entre les catégories lexicales, dans la mesure où un nom déverbal continue de référer à l'événement dénoté par le verbe base, alors qu'un verbe dénominal ne réfère pas à l'entité dénotée par le nom base, mais dénote un événement associé à cette entité. Or, les événements associés à une entité peuvent être nombreux et variés. Par exemple, le nom *destruction* dénote le même événement que sa base verbale *détruire*, alors qu'un verbe dénominal comme *hospitaliser* ne réfère pas à l'objet dénoté par la base nominale (*hôpital*), mais à l'un des événements liés à cet objet. Ainsi, dans le cadre d'un système de QR, la relation sémantique entre un nom et son verbe dérivé est moins informative que la relation entre un verbe et son nom dérivé.

3.3 Composition

En ce qui concerne la composition, comme nous l'avons vu dans la section 3.1 et la table 2, elle est surtout présente dans les sous-corpus EQueR et Conique, mais quasiment absente du sous-corpus Quæro. Dans notre analyse de la composition nous avons distingué les paires de question-passage contenant un composé et au moins l'un de ses constituants (comme dans *filmographie* composé de *film*), des paires contenant deux composés qui partagent un même constituant (comme *aéronautique* et *aéroport* partageant le constituant *aéro*). La table 6 présente les résultats de cette classification, et montre que le second cas (deux composés partageant un constituant) est avant tout présent dans le corpus spécialisé.

	composé-constituant(s)		2 composés	
	nbr	%	nbr	%
Conique (33)	**26**	**79**	7	21
Quæro (4)	**4**	**100**	0	0
EQueR (111)	**70**	**63**	41	37

TABLE 6 – Types de relations de composition

3.4 Conclusion sur les relations morphologiques observées dans les trois sous-corpus

Comme le montre l'analyse des résultats de notre annotation, la morphologie joue un rôle important pour établir le lien de similarité entre question et passage-réponse. D'après notre étude qualitative des relations morphologiques observées, la flexion est loin d'être la seule connaissance morphologique présente dans notre corpus, et la dérivation, tout comme la composition, jouent un rôle important. Notons également que les types de procédés morphologiques employés sont très similaires dans les corpus de langue générale, alors que la langue de spécialité – EQueR, corpus médical – montre des tendances nettement différentes.

Nous avons également étudié la position privilégiée du mot le plus complexe morphologiquement, afin se savoir s'il se trouve plutôt dans la question ou dans le passage réponse. La prédominance de l'une ou l'autre position

joue un rôle essentiel dans la manière de gérer la morphologie dans les systèmes de QR. Pour les paires impliquant une relation de dérivation, le mot complexe se trouve majoritairement dans le passage réponse (52% des cas dans EQueR, 59% dans Quæro et 65% dans Conique). Ce résultat confirme l'intérêt de l'expansion de requête aux mots dérivés des mots de la question.

De plus, le nombre important de relations de composition où les deux membres de la paire sont des composés partageant un même constituant (37% dans EQueR et 21% dans Conique) pointe les limites de l'apport de la morphologie dans de tels systèmes, étant donné qu'il est difficilement envisageable de vouloir générer tous les mots composés à partir d'un élément de la question.

Dans la suite de cet article, nous utilisons les résultats de l'annotation comme gold-standard (ensemble de référence) pour évaluer les ressources morphologiques existantes en français. Cette évaluation permet donc implicitement de connaître l'impact de telles ressources si elles étaient intégrées dans un système de QR. À noter également qu'en plus des ressources statiques rendant compte des relations morphologiques, il existe également des outils d'analyse, à base de règles ou d'heuristiques comme Dérif (Namer, 2009) [5].

4 Evaluation des ressources morphologiques existantes

En français, il n'existe pas de ressources contenant des relations morphologiques dérivationnelles à large échelle similaire à la base CELEX qui contient un nombre important d'informations morphologiques pour le néerlandais, l'anglais et l'allemand (Baayen *et al.*, 1995). Il existe pour le moment uniquement des ressources conçues pour traiter d'un phénomène morphologique particulier. Dans le cadre de notre étude, nous nous sommes intéressés à trois ressources qui couvrent des phénomènes morphologiques particulièrement fréquents dans notre corpus : les noms déverbaux et les adjectifs dénominaux. Ces ressources sont VerbAction, Dubois et Prolexbase. Nous laissons de côté dans le cadre de cette étude la couverture des procédés de composition, pour lesquels il n'existe pas de ressource.

4.1 Présentation des ressources existantes

Verbaction [6] est une ressource lexicale regroupant tous les noms d'actions dérivés d'un verbe (Hathout *et al.*, 2002; Hathout & Tanguy, 2002). Elle contient un total de 9 393 paires de nom-verbe.

Dubois [7] Cette ressource XML, constituée à partir du travail de (Dubois & Dubois-Charlier, 1997) est une description des verbes français regroupés en classes syntaxico-sémantiques, qui fournit également des informations sur les dérivés de ces verbes. Elle contient au total 25 609 entrées pour lesquels elle mentionne 33 955 dérivés.

Prolexbase [8] est un dictionnaire multilingue de noms propres (Tran & Maurel, 2006; Bouchou & Maurel, 2008). Bien qu'elle ne contienne pas explicitement de connaissances morphologiques, cette ressource fournit des informations sur les noms relationnels et les adjectifs associés aux noms propres. Par exemple, *Français* et *français* sont explicitement associés à l'entrée *France*. Au total, Prolexbase contient 76 118 lemmes et 20 614 relations dérivationnelles.

4.2 Résultats

L'évaluation de ces trois ressources dérivationnelles ne peut se faire sur l'ensemble du gold-standard dans la mesure où chacune d'elles a été conçue pour couvrir un phénomène morphologique spécifique. De ce fait nous avons évalué les ressources uniquement sur la partie du gold-standard concernée par le phénomène pour lequel

5. À ce sujet voir Bernhard *et al.* (à paraître).
6. http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=hathout&subURL=verbaction/main.html
7. http://rali.iro.umontreal.ca/Dubois/
8. http://www.cnrtl.fr/lexiques/prolex/

elles ont été conçues. La couverture de VerbAction et de Prolexbase a été évaluée en comptant le nombre de paires de mots morphologiquement reliés qui s'y trouvent. Dubois, en revanche, ne contient pas les dérivés mais mentionne uniquement leur existence et fournit des informations permettant de déduire la forme du dérivé. Pour évaluer Dubois nous avons donc pris en compte les cas où le dérivé pouvait être automatiquement calculé à partir des informations fournies.

La table 7 résume la couverture de VerbAction et de Dubois pour les noms d'événement observés dans notre corpus. La couverture de VerbAction est meilleure que celle de Dubois, en particulier pour les sous-corpus de langue générale Conique et Quæro. Quant aux noms déverbaux agentifs, Dubois couvre 100% des noms de Conique et 75% de ceux de Quæro (aucun nom agentif n'a été trouvé dans le sous-corpus EQueR). VerbAction est limité aux noms d'action et ne contient donc aucun nom d'agent.

	VerbAction		**Dubois**	
Corpus (nbr.)	nbr.	%	nbr.	%
Conique (34)	33	97	19	56
Quæro (25)	25	100	9	36
EQueR (22)	22	100	19	86
Total (81)	**80**	**99**	**47**	**58**

TABLE 7 – Couverture des ressources pour les noms d'événements déverbaux

Pour ce qui est des gentilés et des adjectifs relationnels dérivés de noms géographiques, les résultats de l'évaluation de Prolexbase sont présentés dans la table 8. Nous distinguons les gentilés (habitants d'un lieu), les adjectifs relationnels, et les noms de lieu ou d'entité institutionnelle que nous avons appelés "LocOrg". Les chiffres montrent que Prolexbase a une très bonne couverture pour les gentilés dérivés d'un nom de lieu, et pour les adjectifs relationnels dérivés d'un nom de lieu ou d'un gentilé [9].

Les ressources existantes VerbAction, Dubois et Prolexbase offrent donc une bonne couverture des noms déverbaux et des gentilés et adjectifs dérivés de noms propres. Cependant, si l'on évalue ces trois ressources sur l'ensemble des relations dérivationnelles observées dans le corpus, le taux de couverture global est relativement faible (environ 52%). Ce faible taux de couverture n'est pas étonnant dans la mesure où les ressources évaluées sont conçues pour des phénomènes particuliers. Cela démontre également qu'il reste un nombre important de relations morphologiques dérivationnelles pour lesquelles il n'existe pas encore de ressource exploitable. En premier lieu, il manque une ressource associant les adjectifs dénominaux aux noms dont ils dérivent, lorsqu'il ne s'agit pas de noms géographiques. Or, ce type de relation dérivationnelle est une des plus fréquentes dans notre corpus puisqu'elle concerne environ 21% des paires de mots reliés par un procédé dérivationnel dans le sous-corpus Conique, 23% dans le sous-corpus Quæro, et 40% dans le sous-corpus EQueR (cf. table 3). Une telle ressource spécifiant la relation entre un adjectif dénominal et son nom base permettrait donc d'augmenter de façon significative la couverture globale des ressources morphologiques du français pour les relations morphologiques observées dans notre corpus de question-passage.

5 Conclusion et perspectives

Nous avons présenté une étude détaillée des phénomènes morphologiques permettant de faire le lien entre question et réponse dans le cadre des systèmes de QR. Pour réaliser cette étude, nous avons constitué un corpus de paires de question-passage réponse à partir de divers corpus utilisés pour l'évaluation en QR. Nous avons réalisé une annotation détaillée du corpus, portant sur les liens morphologiques entre question et réponse. Cette annotation nous a permis d'obtenir des données de référence, que nous avons analysées de manière détaillée selon plusieurs axes : types de relations morphologiques, procédés dérivationnels utilisés, relations de composition. Cette analyse nous a permis de tirer les conclusions suivantes : (i) la morphologie dérivationnelle constitue une forte proportion des phénomènes morphologiques à l'œuvre dans le corpus, (ii) les phénomènes de dérivation observés concernent essentiellement les adjectifs dénominaux et les nominalisations verbales, (iii) le procédé de composition s'observe essentiellement dans le sous-corpus spécialisé de la langue médicale EQueR.

9. Dans le corpus Quæro, aucune paire gentilé>adjectif relationnel n'a été trouvée, et dans le corpus EQueR, seule une paire LocOrg>adjectif relationnel a été trouvée et est analysée correctement.

Corpus	**Relation morphologique (nbr.)**		**Trouvé dans Prolexbase**	
			nbr.	%
Conique	Gentilé - Adj. Rel	(1)	1	100
	LocOrg - gentilé	(6)	6	100
	LocOrg - Adj. rel	(45)	43	96
Quæro	LocOrg - Gentilé	(8)	5	62
	LocOrg - Adj. rel.	(8)	8	100
EQueR	LocOrg - Adj. rel.	(1)	1	100
Total		**69**	**64**	**93**

TABLE 8 – Couverture de Prolexbase pour les relations morphologiques de type "géographique"

Si ces résultats soulignent l'importance de la morphologie dans l'appariement en question-réponse, et montrent clairement quelles relations sont les plus pertinentes (car les plus fréquentes), ils ne permettent pas d'évaluer l'impact, notamment en termes de bruit, de la prise en compte de ces relations, fréquentes ou non, dans un système de QR. Une intégration modulaire de chacune des relations, et une évaluation précise de leur impact sur les résultats d'un système de QR permettraient sans doute d'avoir une meilleure idée sur la question.

Nous avons également évalué la couverture de ressources morphologiques existantes pour le français par rapport aux phénomènes observés. Si certains procédés bénéficient d'une très bonne couverture (noms d'événement dans VerbAction, gentilés et adjectifs relationnels dérivés de noms géographiques dans Prolexbase), d'autres souffrent d'un manque de ressource adaptée, comme par exemple les adjectifs dénominaux.

Les perspectives de ces travaux sont multiples. D'un point de vue linguistique, l'observation des relations morphologiques les plus fréquentes, et l'absence constatée de certaines autres, semblent indiquer que certains types de relations morphologiques sont plus informatifs et donc plus pertinents que d'autres. D'un point de vue applicatif cette hypothèse mériterait néanmoins d'être évaluée empiriquement. L'analyse a également permis de distinguer les procédés dérivationnels à intégrer de façon prioritaire dans les systèmes de QR. Nous envisageons d'intégrer ces observations dans un système de QR existant, en définissant notamment des patrons de reformulation de question basés sur la morphologie.

Remerciements Ces travaux ont été partiellement financés par OSEO dans le cadre du programme QUAERO.

Références

AYACHE C., GRAU B. & VILNAT A. (2006). EQueR : the French Evaluation campaign of Question-Answering Systems. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006)*, p. 1157–1160.

BAAYEN R. H., PIEPENBROCK R. & GULIKERS L. (1995). *The Celex Lexical Database (Release 2) [CD-ROM]*. Philadelphia, PA : Linguistic Data Consortium.

BERNHARD D., CARTONI B. & TRIBOUT D. (À paraître). A Task-Based Evaluation of French Morphological Resources and Tools : A Case Study for Question-Answer pairs. *Linguistic Issues in Language Technology - LiLT*.

BILOTTI M. W., KATZ B. & LIN J. (2004). What Works Better for Question Answering : Stemming or Morphological Query Expansion. In *Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*, Sheffield, England.

BOUCHOU B. & MAUREL D. (2008). Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres. *Traitement Automatique des Langues*, **49**(1), 61–88.

DUBOIS J. & DUBOIS-CHARLIER F. (1997). *Les verbes français*. Larousse-Bordas.

FLEISS J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**(5), 378–382.

FULLER M. & ZOBEL J. (1998). Conflation-based comparison of stemming algorithms. In *Proceedings of the Third Australian Document Computing Symposium*, p. 8–13, Sydney.

GERMANN U. (2008). Yawat : yet another word alignment tool. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (HLT '08)*, p. 20–23.

GRAPPY A., GRAU B., FERRET O., GROUIN C., MORICEAU V., ROBBA I., TANNIER X., VILNAT A. & BARBIER V. (2010). A Corpus for Studying Full Answer Justification. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

HARMAN D. (1991). How effective is suffixing ? *Journal of the American Society of Information Science*, **42**(1), 7–15.

HATHOUT N., NAMER F. & DAL G. (2002). Many Morphologies. chapter An Experimental Constructional Database : The MorTAL Project, p. 178–209. Cascadilla Press.

HATHOUT N. & TANGUY L. (2002). Webaffix : Discovering Morphological Links on the WWW. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, p. 1799–1804, Las Palmas de Gran Canaria, Espagne : ELRA.

HERMJAKOB U., ECHIHABI A. & MARCU D. (2002). Natural Language Based Reformulation Resource and Wide Exploitation for Question Answering. In *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002)*.

HOPPER P. & THOMPSON S. (1984). The discourse basis for lexical categories in universal grammar. *Language*, **60**, 703–752.

JACQUEMIN B. (2010). A Derivational Rephrasing Experiment for Question Answering. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

LENNON M., PIERCE D. S., TARRY B. D. & WILLETT P. (1988). An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science*, **3**(4), 177–183.

LIN D. & PANTEL P. (2001). Discovery of Inference Rules for Question Answering. *Natural Language Engineering*, **7**(4), 343–360.

MCNAMEE P., NICHOLAS C. & MAYFIELD J. (2009). Addressing morphological variation in alphabetic languages. In *SIGIR '09 : Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, p. 75–82, New York, NY, USA : ACM.

MOREAU F. & CLAVEAU V. (2006). Extension de requêtes par relations morphologiques acquises automatiquement. In *Actes de la Troisième Conférence en Recherche d'Informations et Applications CORIA 2006*, p. 181–192.

NAMER F. (2009). *Morphologie, Lexique et TAL : l'analyseur DériF*. TIC et Sciences cognitives. London : Hermes Sciences Publishing.

QUINTARD L., GALIBERT O., ADDA G., GRAU B., LAURENT D., MORICEAU V., ROSSET S., TANNIER X. & VILNAT A. (2010). Question Answering on Web Data : The QA Evaluation in Quæro. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

RAVICHANDRAN D. & HOVY E. (2002). Learning surface text patterns for a Question Answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, p. 41–47.

TRAN M. & MAUREL D. (2006). Prolexbase : un dictionnaire relationnel multilingue de noms propres. *Traitement Automatique des Langues*, **47**(1), 115–139.

Syntaxe

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Modèles génératif et discriminant en analyse syntaxique : expériences sur le corpus arboré de Paris 7

Joseph Le Roux Benoît Favre Seyed Abolghasem Mirroshandel Alexis Nasr
LIF - CNRS UMR 6166 - Université Aix Marseille
{joseph.le-roux, benoit.favre, alexis.nasr}@lif.univ-mrs.fr

Résumé. Nous présentons une architecture pour l'analyse syntaxique en deux étapes. Dans un premier temps un analyseur syntagmatique construit, pour chaque phrase, une liste d'analyses qui sont converties en arbres de dépendances. Ces arbres sont ensuite réévalués par un réordonnanceur discriminant. Cette méthode permet de prendre en compte des informations auxquelles l'analyseur n'a pas accès, en particulier des annotations fonctionnelles. Nous validons notre approche par une évaluation sur le corpus arboré de Paris 7. La seconde étape permet d'améliorer significativement la qualité des analyses retournées, quelle que soit la métrique utilisée.

Abstract. We present an architecture for parsing in two steps. First, a phrase-structure parser builds for each sentence an n-best list of analyses which are converted to dependency trees. Then these trees are rescored by a discriminative reranker. This method enables the incorporation of additional linguistic information, more precisely functional annotations. We test our approach on the French Treebank. The evaluation shows a significative improvement on different parse metrics.

Mots-clés : analyse syntaxique, corpus arboré, apprentissage automatique, réordonnancement discriminant.

Keywords: parsing, treebank, machine learning, discriminative reranking.

1 Introduction

On peut observer l'existence de deux approches en analyse syntaxique automatique. La première, dite générative, se fonde sur la tradition des langages formels et des systèmes de réécriture. L'analyse syntaxique est envisagée ici comme un processus permettant de passer d'une structure initiale (une chaîne d'entrée) à une structure finale (un arbre ou une forêt d'analyses). On utilise le plus couramment les grammaires algébriques qui peuvent s'analyser en temps polynomial. Malheureusement, l'hypothèse d'indépendance des réécritures qui sous-tend ce formalisme ne permet pas une analyse très fine de certains phénomènes, en particulier les dépendances à longue distance et les dépendances lexicales.

La seconde approche, dite discriminante, se fonde sur la « syntaxe comme théorie des modèles » (en anglais *model-theoretic syntax*, (Pullum & Scholz, 2001)) et a connu un regain d'intérêt grâce aux progrès réalisés dans le domaine de l'apprentissage automatique, plus précisément en classification automatique. Dans cette approche, la grammaire est vue comme un système de contraintes sur les structures syntaxiques correctes. Les mots de la phrase d'entrée sont eux-mêmes vus comme des contraintes sur les positions qu'ils occupent et l'analyse syntaxique revient à résoudre ces contraintes. Le problème majeur de cette seconde approche tient à sa complexité. Les contraintes pouvant en théorie porter sur divers aspects des structures finales, il n'est pas possible d'utiliser des techniques de programmation dynamique efficaces et il faut, dans le pire des cas, énumérer tous les arbres pour ensuite évaluer leur pertinence. Dans certains développements de cette approche, utilisés dans le présent travail, les contraintes sont uniquement d'ordre numérique. Une analyse y est représentée par un vecteur de traits et sa qualité se mesure par la distance entre ce dernier et l'analyse de référence.

Une manière de tirer profit des deux approches consiste, comme l'a proposé (Collins, 2000), à les combiner de façon séquentielle. Un analyseur de type génératif produit alors un ensemble de structures candidates pour un second module, discriminant, de façon à contraindre son espace de recherche. Cette approche par analyse puis réordonnancement (en anglais, *parsing/reranking*) est utilisée dans l'analyseur de Brown (Charniak & Johnson, 2005), adapté pour le français dans (Seddah *et al.*, 2009). Il est même intéressant de fournir au module de réordonnancement des analyses provenant de différents analyseurs, comme le montrent les résultats obtenus par (Johnson & Ural, 2010).

Notre architecture, représentée sur la figure 1, reprend ces deux étapes. Lors de la première étape, un analyseur syntagmatique traite chaque phrase d'entrée et produit la liste des n analyses les plus probables munies de leur probabilité. Elles sont ensuite annotées par un étiqueteur fonctionnel avec des fonctions syntaxiques classiques *sujet, objet, objet indirect...* Les analyses syntagmatiques enrichies d'annotations fonctionnelles sont alors converties en structures de dépendances. À l'issue de cette conversion, on dispose donc de candidats qui sont des structures de dépendances munies du score donné par le premier analyseur. Cette étape est réalisée par un analyseur syntagmatique PCFG-LA (Petrov *et al.*, 2006) couplé à l'étiqueteur et au convertisseur BONSAÏ [1].

Chaque structure de dépendances munie d'un score est ensuite traduite en un vecteur de traits. Ces traits représentent des configurations structurelles qui peuvent être absentes ou présentes dans l'arbre de dépendances et le vecteur indique leur nombre d'occurrences pour ce candidat. Le score attribué par l'analyseur syntagmatique est lui-même vu comme un trait. Les différents candidats sont finalement évalués par le réordonnanceur et le système retourne le meilleur candidat. Ce module est réalisé par notre implantation de l'algorithme MIRA (Crammer *et al.*, 2006).

Il nous paraît important de revenir ici sur deux aspects de notre architecture. Le premier est la réalisation de l'étape de réordonnancement sur des structures de dépendances et non pas sur des structures syntagmatiques, à l'image de (Charniak & Johnson, 2005). Notre analyseur produisant des structures syntagmatiques, il aurait en effet été plus naturel de réaliser le réordonnancement sur des structures syntagmatiques et de s'épargner ainsi leur conversion en structures de dépendances. Deux raisons sont à l'origine de ce choix. D'une part, il nous a semblé que de nombreuses contraintes se modélisent plus naturellement sous la forme de structures de dépendances que sous la forme de structures syntagmatiques. On pense en particulier à des contraintes sur les cadres de sous-catégorisation ou à des contraintes de sélection, qui font explicitement appel à la notion de fonction syntaxique. D'autre part, un certain nombre de travaux récents en analyse syntaxique (McDonald, 2006; Nivre *et al.*, 2007) reposent sur les structures de dépendances et il nous a semblé intéressant de proposer un système de réordonnancement pour ce type d'analyses.

1. disponibles sur le site `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`.

La raison pour laquelle nous n'avons pas utilisé directement un analyseur en dépendances, qui aurait été sans doute un choix plus naturel, est qu'il n'existe pas à notre connaissance d'analyseur en dépendances qui génère les n analyses les plus probables. Les analyseurs de (McDonald, 2006), ou de (Nivre *et al.*, 2007), par exemple, ne permettent pas de les produire, même si le premier peut les approximer. C'est donc une raison pragmatique qui nous a poussés à faire ce choix. Enfin, il était intéressant de vérifier si, comme pour l'anglais, les structures de dépendances obtenues par conversions de structures syntagmatiques sont de meilleure qualité que celles renvoyées par les analyseurs en dépendances directement.

Le réordonnanceur proposé dans ce travail, qui sera décrit en détail dans la section 3, partage plusieurs de ses caractéristiques avec l'analyseur de (McDonald, 2006), évoqué ci-dessus. Les deux reposent sur l'algorithme d'apprentissage MIRA (Crammer *et al.*, 2006) décrit dans la section 3. De plus, les contraintes considérées dans notre modèle sont inspirées de (McDonald, 2006). La différence fondamentale provient du fait que les seules analyses prises en comptes sont celles produites par le modèle génératif (il s'agit d'un réordonnanceur et non d'un analyseur). L'avantage de cette solution par rapport à (McDonald, 2006) est que nous ne sommes pas restreints à un ensemble de traits locaux. En effet la prise en compte de traits non prévus, de domaine de localité arbitraire, dans l'algorithme de (McDonald, 2006) suppose des modifications de l'algorithme d'analyse alors qu'ils peuvent être ajoutés facilement dans notre modèle de réordonnancement.

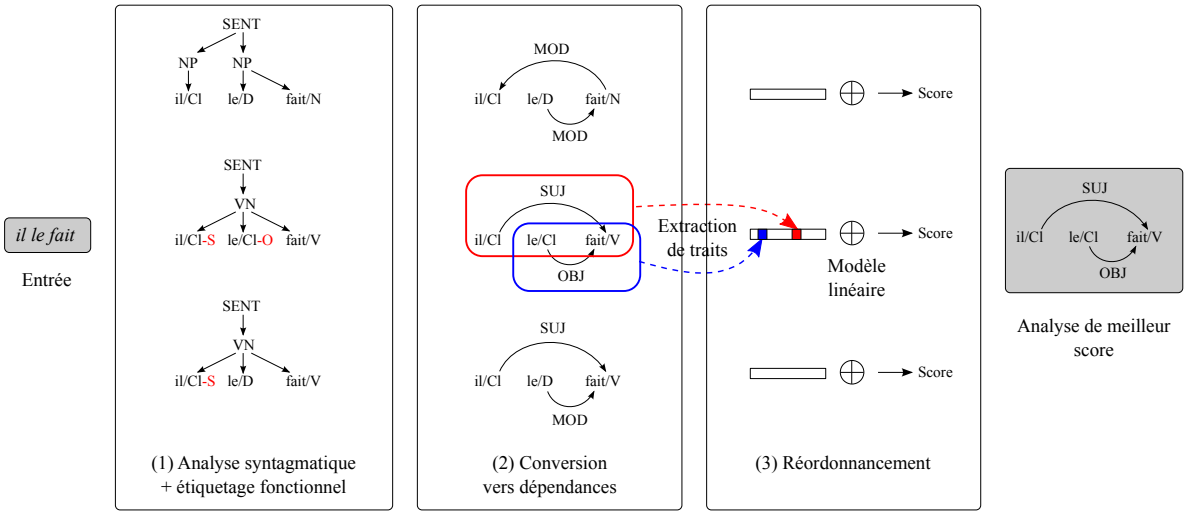


FIGURE 1 – Architecture de notre analyseur : (1) génération de n arbres syntagmatiques annotés en fonctions, (2) conversion vers une représentation en dépendances et extraction de vecteurs de traits, (3) calcul des scores à l'aide d'un modèle linéaire. L'analyse de meilleur score est considérée comme l'analyse finale.

La suite de l'article se présente de la façon suivante : nous décrivons en 2 les détails du modèle utilisé dans notre analyseur génératif puis en 3 le modèle de réordonnancement discriminant et les patrons de traits utilisés. La section 4 présente les résultats obtenus sur le corpus arboré développé à l'université Paris 7 (Abeillé *et al.*, 2003) et la section 5 présente les conclusions.

2 Modèle génératif

La première partie de notre système, l'analyse syntaxique classique, produit des structures en dépendances de surface grâce à un système séquentiel, à l'image de (Candito *et al.*, 2009, 2010b). Un analyseur, fondé sur les PCFG-LA, produit des structures syntagmatiques qui sont ensuite transformées en arbres de dépendances. Deux points nous distinguent des travaux précédents : (1) la liste d'analyses candidates est produite par un nouvel analyseur et (2) ces analyses ne sont pas considérées comme les structures finales mais seront traitées dans le réordonnanceur.

2.1 Les grammaires algébriques à annotations latentes

Les grammaires algébriques probabilistes à annotations latentes (PCFG-LA), introduites par (Matsuzaki *et al.*, 2005), peuvent être vues comme une façon de spécialiser automatiquement le jeu d'étiquettes d'une grammaire algébrique (PCFG) à partir d'un corpus de manière à en améliorer la précision.

Chaque symbole de la grammaire est enrichi d'annotations se comportant comme des sous-classes de ce symbole, et les probabilités des règles qui manipulent les symboles augmentés sont estimées par la méthode EM d'apprentissage non-supervisé, à partir des fréquences relatives observées sur le corpus. (Petrov *et al.*, 2006) proposent d'apprendre ces grammaires en plusieurs rondes : à chaque itération on divise une annotation d'un symbole en deux si l'apport des nouvelles annotations augmente la vraisemblance du corpus d'entraînement. Cette méthode permet d'obtenir une grammaire dans laquelle le nombre d'annotations est adapté au symbole et beaucoup plus compacte que celles obtenues par (Matsuzaki *et al.*, 2005).

On peut reprendre l'illustration des spécialisations d'étiquettes de parties du discours de (Petrov *et al.*, 2006) pour le français. Par exemple l'étiquette DET est divisée en quinze sous-étiquettes après cinq rondes d'apprentissage. Nous montrons dans la table 1 les trois mots les plus fréquents pour chaque annotation [2]. Même si la spécialisation est difficile à interpréter complètement, on note que les articles définis et indéfinis sont séparés des démonstratifs et possessifs d'une part et des cardinaux d'autre part. Il est important de noter que cette distinction est apprise par une méthode qui dépend largement des paramètres d'initialisation et qui ne garantit pas de trouver la spécialisation optimale.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	une	la	l'	un	les	les	ses	le	cette	son	NB	deux	NB	NB	NB
2	la	La	la	son	des	Les	ces	Le	Cette	ce	10	trois	7	40	1
3	La	L'	les	Un	Les	de	leurs	un	Ce	leur	3	quelques	trois	20	30

TABLE 1 – Division du symbole DET en annotations

Dans ce formalisme, le problèmes de l'analyse exacte devient NP-difficile mais (Petrov & Klein, 2007) décrivent comment l'approximer efficacement, en tirant également profit de la structure hiérarchique des annotations créées lors des rondes successives. Une forêt d'analyses est produite avec la PCFG d'origine et est ensuite élaguée avec les grammaires intermédiaires produites lors des rondes successives d'apprentissage [3], c'est-à-dire qu'on ne garde qu'un sous-ensemble *a priori* intéressant des analyses, de manière à restreindre au maximum l'espace de recherche lors du décodage avec la grammaire finale.

Nous utilisons notre propre analyseur de PCFG-LA [4]. La grammaire est apprise en cinq rondes sur le corpus d'entraînement : c'est le nombre de rondes optimal sur notre corpus. Pour pouvoir traiter les mots inconnus, les mots rares sont remplacés par des chaînes contenant des informations positionnelles et typographiques, en particulier leur suffixe. La liste des suffixes *intéressants* est collectée sur ce corpus en fonction du gain d'information pour l'étiquetage et permet de mieux traiter les mots inconnus (Attia *et al.*, 2010).

Ce type de grammaire a déjà été utilisé pour le français à partir du même corpus (Crabbé & Candito, 2008). Il a donné des résultats du niveau de l'état de l'art en matière d'analyse en constituants. (Candito & Crabbé, 2009) montrent que si l'on sait catégoriser les mots en classes lexicales (agrégats ou *clusters*), ces grammaires offrent les meilleures analyses en syntagmes pour le français. Nous reprendrons cette classification pour nos expériences (cf. section 4). Cependant, les structures en dépendances extraites des meilleures analyses en constituants ne sont pas aussi bonnes que celles obtenues directement par un analyseur en dépendances (Candito *et al.*, 2010b).

2.2 Structures de dépendances

Bien qu'une théorie syntaxique puisse se représenter à partir de syntagmes et de dépendances (Rambow, 2010), certains types d'informations sont plus ou moins faciles à décrire selon la représentation choisie. L'analyseur génératif étant appris sur des structures syntagmatiques, il est vraisemblable qu'une partie de l'information linguistique lui échappe parce qu'elle est implicite ou difficile à retrouver dans cette représentation.

2. Le symbole terminal NB est une chaîne générique qui remplace les nombres qui apparaissent peu de fois dans le corpus d'entraînement.
3. En réalité ces grammaires sont recalculées à la volée à partir de la grammaire finale et de la structure hiérarchique des annotations.
4. Cet analyseur est disponible pour les travaux académiques. Les lecteurs intéressés peuvent contacter le premier auteur.

Cette équivalence n'est de toute façon vraie que si les deux représentations offrent réellement les mêmes informations. Or, comme cela sera décrit dans la section 4, nous avons fait disparaître les fonctions du corpus d'apprentissage de l'analyseur syntagmatique pour aider lutter à contre l'effet de dispersion des données et améliorer l'apprentissage des relations de sous-constituance. Mais pour obtenir les relations de dépendances typées, il est nécessaire de disposer de l'information fonctionnelle.

Pour passer des structures en constituants aux structures en dépendances, nous utilisons le convertisseur développé par (Candito *et al.*, 2010a) et disponible dans la boîte à outils BONSAÏ. La conversion se décompose en deux étapes :

1. Les nœuds internes sont réannotés avec des étiquettes fonctionnelles, en utilisant un classifieur multiclasse par maximum d'entropie.
2. Les arbres ainsi décorés sont convertis en structures de dépendances par un ensemble de règles de propagations de têtes fonctionnelles et d'heuristiques.

Cette conversion effectuée, les analyses sont prêtes à être réordonnancées.

3 Modèle discriminant

Le modèle discriminant que nous proposons repose sur l'algorithme Margin-infused Relaxed Algorithm (MIRA) (Crammer *et al.*, 2006). Selon ce modèle, le score d'une analyse est calculé comme la combinaison linéaire des traits extraits à partir de cette analyse, pondérés par un vecteur de poids représentant les paramètres du modèle. MIRA, l'algorithme d'apprentissage des paramètres du modèle, est très similaire au Perceptron (Rosenblatt, 1958), et est donc rapide et peu gourmand en ressources, tout en offrant de meilleures performances.

Le réordonnancement discriminant des analyses se déroule selon deux étapes : apprentissage des paramètres du modèle et prédiction. L'étape de prédiction consiste à produire les n meilleures analyses du modèle génératif, extraire des traits pour caractériser ces analyses et attribuer un score à chaque analyse en fonction de ses traits et du modèle discriminant. L'analyse de meilleur score est alors sélectionnée comme sortie finale du système. L'étape d'entraînement consiste à utiliser des exemples de phrases avec leurs analyses de référence (ensemble d'apprentissage) pour déterminer les paramètres du modèle. Alors que la plupart des modèles discriminants tentent de minimiser le taux d'erreur global sur l'ensemble des exemples d'apprentissage, MIRA se contente de traiter les exemples un par un, ajustant son modèle pour que l'analyse sélectionnée pour la phrase courante soit celle qui est la plus proche de l'analyse de référence. Une telle approche, appelée « en ligne », limite les ressources nécessaires, processeur et mémoire, rendant possible l'apprentissage de modèles qui prennent en compte un très grand nombre de traits.

3.1 Définitions

On se place dans un espace vectoriel de dimension m où chaque dimension correspond à un trait. Certaines dimensions représentent la présence ou l'absence d'un trait (valeur booléenne), son nombre d'occurrence dans l'analyse (entier naturel), ou une valeur réelle quelconque (par exemple la probabilité de l'analyse, ou son logarithme, selon le modèle génératif). Une analyse p est alors représentée sous la forme d'un vecteur de réels $\phi(p)$. Dans le cas d'un indicateur de présence, la i^{e} coordonnée de $\phi(p)$ vaut 1 si p possède le i^{e} trait et 0 sinon. Un tel vecteur est généralement creux (une analyse ne possède en moyenne qu'une petite partie des différents traits possibles). Un modèle est un vecteur de poids w de dimension m dont la i^{e} coordonnée est le poids associé au i^{e} trait. Plus ce poids est important, plus le trait correspondant aura été jugé discriminant. Le score d'une analyse p n'est rien d'autre que le produit scalaire du vecteur $\phi(p)$ et du vecteur w :

$$score(p) = \sum_{i=1}^{m} w_i \times \phi_i(p) \tag{1}$$

Soit L la liste des n meilleures analyses produites par l'analyseur syntaxique génératif pour une phrase. On calcule le score de chaque analyse et l'analyse de score maximum $\hat{p}$ est choisie comme sortie finale du système :

$$\hat{p} = \underset{p \in L}{\text{argmax}}\, score(p) \tag{2}$$

La phase d'apprentissage consiste à utiliser les phrases d'entraînement et leurs analyses de référence pour déterminer le vecteur de poids w. Le classifieur MIRA commence avec un vecteur de poids nul (c'est-à-dire que toutes les analyses ont un score de zéro), et essaie de modifier ce vecteur de façon à ce que les bonnes analyses ait un score plus élevé que les mauvaises analyses. L'analyse la plus proche de la référence est nommée *oracle* (notée o). Il serait souhaîtable que l'oracle ait le meilleur score parmi les analyses proposée pour une phrase. Soit $error(p)$ le nombre de mauvaises dépendences (étiquette, position, direction) dans l'analyse p. L'oracle o est l'analyse pour laquelle $erreur()$ est minimale.

Les phrases de l'ensemble d'entraînement sont traitées séquentiellement. Pour chacune d'entre elles, on détermine la liste des n meilleures analyses candidates, puis l'analyse candidate de meilleur score, notée $\hat{p}$. Si cette analyse est différente de l'oracle ($\hat{p} \neq o$), cela signifie que le vecteur de poids w peut être amélioré. Dans ce cas, on recherche une modification de w qui assure que o ait un score plus élevé que l'analyse qui avait le meilleur score. Plus précisément, on souhaite que la différence entre leurs scores soit proportionnelle à la différence entre leur distance à la référence. Ainsi, une très mauvaise analyse aura un score bien plus faible qu'une analyse de qualité moyenne. Trouver une amélioration de w peut être formulé comme un problème d'optimisation sous la contrainte que la différence des scores de l'oracle et de l'hypothèse de plus haut score soit supérieure à la différence entre leurs distances à la référence. Comme il existe une infinité de vecteurs w satisfaisant cette contrainte, on recherche celui de plus petite norme. Ce problème s'écrit sous la forme suivante :

$$\text{minimiser : } ||w|| \text{ tel que : } score(o) - score(\hat{p}) \geq error(o) - error(\hat{p}) \tag{3}$$

Des méthodes classiques d'optimisation quadratique sous contrainte sont utilisées : tout d'abord la contrainte est introduite dans la fonction objective grâce à des multiplicateurs de Lagrange. De cette façon, les solutions qui violent le plus la contrainte sont pénalisées par rapport aux autres solutions. Enfin, la méthode de Hildreth donne la solution analytique suivante au problème quadratique non contraint :

$$w^\star = w + \alpha\,[\phi(o) - \phi(\hat{p})] \tag{4}$$

$$\alpha = max\left[0, \frac{error(o) - error(\hat{p}) - (score(o) - score(\hat{p}))}{||\phi(o) - \phi(\hat{p})||^2}\right] \tag{5}$$

Ici, $w^\star$ est le nouveau vecteur de poids, α est un taux de modification et $[\phi(o) - \phi(\hat{p})]$ est la différence entre le vecteur de traits de l'oracle et celui de l'analyse de plus haut score. Concrètement, cette mise à jour attire le vecteur de poids en direction de l'oracle et l'éloigne de $\hat{p}$. Cet algorithme d'apprentissage est très proche de l'algorithme du perceptron, et tout comme pour le perceptron, il est recommandé (1) de faire de multiples passes sur l'ensemble d'apprentissage et (2) de sauvegarder le vecteur de poids après chaque mise à jour et d'en faire la moyenne pour produire le vecteur de poids final [5]. L'algorithme 1 présente l'apprentissage MIRA.

3.2 Traits utilisés

La qualité d'un réordonnanceur dépend de la capacité de l'algorithme d'apprentissage à attribuer un bon poids aux traits discriminants, mais aussi, de manière cruciale, à la qualité des traits qui lui sont fournis. Ces traits peuvent porter sur n'importe quelle configuration lexico-syntaxique d'une analyse. L'ensemble des traits potentiels est par conséquent gigantesque. Pour être pertinent, un trait doit d'une part être assez général pour apparaître souvent et, d'autre part, permettre de discriminer les bonnes analyses des mauvaises. Il n'existe pas de méthode générale de sélection des traits pertinents, du fait de leur nombre et de leur caractère non monotone : un trait simple peut se révéler non pertinent mais l'extension de ce trait simple en un trait plus complexe peut l'être. L'espace des traits est par conséquent difficile à explorer systématiquement et la sélection des traits pertinents est une activité qui relève de l'art, pour reprendre un bon mot de (Charniak & Johnson, 2005).

Nous nous sommes inspirés de traits utilisés dans l'analyseur (McDonald, 2006) qui a montré de bonnes performances dans de nombreuses langues. Ils se divisent en cinq familles, chaque famille correspondant à un type de configuration. L'instanciation de ces patrons sur les sorties de l'analyseur a généré plus de 70 millions de traits. Nous décrivons ci-dessous les cinq familles de traits que nous illustrons sur la phrase *Les enfants mangent des glaces avec appétit* dans laquelle on s'intéressera en particulier à la dépendance objet *(mangent, glaces)*.

5. Dans la pratique, il n'est pas nécessaire de sauvegarder tous les vecteurs de poids, mais seulement deux vecteurs.

Algorithme 1 Entraînement MIRA

pour $i = 1$ à t **faire**
  pour chaque phrase de l'ensemble d'entraînement **faire**
    Générer les n meilleures hypothèses de l'analyseur en constituants.
    pour chaque hypothèse **faire**
      Extraire un vecteur de traits à partir de l'hypothèse.
      Calculer son score comme le produit scalaire entre ce vecteur et le vecteur de poids (éq. 1)
    fin pour
    Soit l'oracle, l'hypothèse la plus proche de la référence pour cette phrase.
    si l'hypothèse de meilleur score n'est pas l'oracle **alors**
      Calculer la différence entre le vecteur de traits de l'oracle et le vecteur de traits de l'hypothèse.
      Calculer le facteur α qui assure que l'oracle ait un meilleur score la prochaine fois (éq. 5)
      Ajouter au vecteur de poids ce facteur fois la différence entre les deux vecteurs (éq. 4)
    fin si
  fin pour
fin pour
Retourner un vecteur de poids moyen consituté à partir de l'état du vecteur de poids après chaque phrase d'entraînement.

Unigramme Les traits unigrammes sont les plus simples, ils ne mettent en jeu qu'une seule dépendance. Étant donné une dépendance entre les positions i et j de type l, gouvernée par x_i, notée $x_i \xrightarrow{l} x_j$, on crée deux traits, l'un pour le gouverneur x_i , l'autre pour le dépendant x_j qui prennent la forme de sextuplets (mot, lemme, partie du discours, statut dans la dépendance (gouverneur (G) ou dépendant (D)) , direction de la dépendance (droite (D) ou gauche (G)), type de la dépendance). On ajoute aussi tous les tuples avec une partie de l'information masquée pour lutter contre la dispersion des données lors de l'apprentissage.

Ainsi, la présence de la dépendance objet dans notre exemple donnera naissance aux deux traits :

`[mangent, manger, V, G, D, objet]` et `[glaces, glace, N, D, D, objet]`

mais aussi à tous les traits que l'on peut construire à partir des deux premiers par sous-spécification :

`[-, manger, V, G, D, objet], [mangent, -, V, G, D, objet] ...`
`[mangent, -,-,-,-,objet]`

Ce processus de sous-spécification des traits s'applique à toutes les familles de traits, on omettra de le répéter ci-dessous.

Bigramme Contrairement à la famille précédente qui ne prenait en compte qu'un des deux membres d'une dépendance, les traits bigrammes modélisent la cooccurrence des deux membres de la dépendance, à l'image des dépendances bi-lexicales de (Collins, 1997). Étant donné la dépendance $x_i \xrightarrow{l} x_j$, on crée un trait (mot x_i, lemme x_i, partie du discours x_i, mot x_j, lemme x_j, partie du discours x_j, distance [6] de i à j, direction de la dépendance, type de la dépendance).

L'exemple ci-dessus donnera donc naissance au trait suivant :

`[mangent, manger, V, glaces, glace, N, 2, D, objet]`

où 2 est la distance séparant *mangent* et *glaces* dans la chaîne linéaire.

Contexte linéaire Contrairement aux deux familles précédentes qui s'intéressaient à une dépendance indépendamment de sa réalisation dans la chaîne, on regarde ici les éléments qui séparent un gouverneur de son dépendant dans cette dernière. Étant donné la dépendance $x_i \xrightarrow{l} x_j$ On crée un trait avec les partie de discours de x_i, de x_j et de chaque mot entre les positions i et j. On crée également un trait comportant les parties de discours aux positions $i-1$, i, $i+1$, $j-1$, j, $j+1$. La dépendance objet de notre exemple donnera naissance aux deux traits :

`[V, D, N]` et `[N, V, D, N, P]`

Contexte syntaxique, nœuds frères Cette famille de traits ainsi que la suivante mettent en jeu deux dépendances dans deux configurations particulières. Étant donné deux dépendances $x_i \xrightarrow{l} x_j$ et $x_i \xrightarrow{m} x_k$, on crée un trait (mot x_i, lemme x_i, partie du discours x_i, mot x_j, lemme x_j, partie du discours x_j, mot x_k, lemme x_k, partie du discours x_k, distance de i à j, distance de i à k, direction de la première dépendance,

6. Cette distance est réduite à sept classes selon qu'elle est égale à 1, 2, 3, 4, 5, comprise entre 5 et 10, ou supérieure à 10.

type de la première dépendance, direction de la seconde dépendance, type de la seconde dépendance). Ce qui donne, dans notre exemple : `[mangent, manger, V, glaces, glace, avec, avec, P, 2, 3, D, objet, D, mod]`

Contexte syntaxique, chaînes Étant donné deux dépendances $x_i \xrightarrow{l} x_j \xrightarrow{m} x_k$, on crée un trait (mot x_i, lemme x_i, partie du discours x_i, mot x_j, lemme x_j, partie du discours x_j, mot x_k, lemme x_k, partie du discours x_k, distance de i à j, distance de i à k, direction de la première dépendance, type de la première dépendance, direction de la seconde dépendance, type de la seconde dépendance). Dans notre exemple : `[mangent, manger, V, avec, avec, P, appétit, appétit, N, 3, 4, D, mod, D, objet]`

On peut noter que les patrons de traits ne reposent que sur des connaissances présentes dans les données d'apprentissage. Nous n'avons pas ajouté de traits qui proviennent de connaissances linguistiques externes.

4 Expériences

Dans cette section, nous évaluons les performances de notre analyseur. Nous présentons d'abord le module génératif seul, puis les deux modules ensemble.

Nous utilisons dans nos expériences le corpus arboré de Paris 7 (Abeillé *et al.*, 2003) (dans la suite, FTB). Il contient 12 350 phrases annotées syntaxiquement en constituants, et en étiquettes fonctionnelles. Ce n'est pas directement ce corpus que nous manipulons mais deux transformations de celui-ci.

1. La première, FTB-UC, mise au point par (Crabbé & Candito, 2008), garde la structure en constituants mais simplifie le jeu d'étiquettes. En particulier, les fonctions disparaissent et les informations morphologiques sont réduites. C'est sur ce corpus que nous apprendrons la grammaire syntagmatique.
2. La seconde, FTB-UC-DEP, présentée dans (Candito *et al.*, 2009), est une conversion du FTB en dépendances. Cette conversion est réalisée à l'aide de règles de propagation de têtes et d'heuristiques. C'est sur ce second corpus que l'on apprendra l'analyseur discriminant.

Pour entraîner et évaluer notre système, nous divisons ces corpus en 3 : une partie pour l'entraînement (80%), une partie pour le développement (10%) et le reste pour l'évaluation finale.

Nous employons deux types de grammaires syntagmatiques, le premier appris directement sur FTB-UC (modèle simple), le second appris sur une version modifiée de FTB-UC dans laquelle les mots sont remplacés par leur classe d'équivalence (modèle agrégats), comme dans (Candito & Crabbé, 2009).

4.1 Évaluation du modèle génératif seul

Les performances de notre analyseur syntagmatique sur le corpus de développement sont résumées dans le tableau 2. Le F-score [7] est la moyenne harmonique du rappel (ici, les constituants de la référence retrouvés par l'analyseur) et de la précision (ici, les constituants prédits présents dans la référence). Nous donnons les scores oracles de notre analyseur quand il renvoie les 1, 10, 50 et 100 analyses les plus probables d'une phrase, pour donner une idée de la marge de progression possible.

Les résultats quantitatifs de la conversion en structures de dépendances sont également présentés dans la table 2. Le score d'attachement étiqueté (LAS) est le taux de dépendances typées correctement reconnues [8] par l'analyseur. Le score non-étiqueté (UAS) est ce même taux lorsque l'on ne tient pas compte du type des dépendances. Nous avons représenté sur la dernière colonne (GOLD) l'évaluation de la conversion en dépendances de l'analyse syntagmatique de référence. Ces résultats nous permettent d'évaluer la qualité de l'étiquetage fonctionnel, ils montrent que l'étiqueteur effectue à peu près 4% d'erreurs dans l'attribution de ces étiquettes. Comme précédemment, nous donnons aussi le score oracle quand notre analyseur renvoie plusieurs analyses.

7. Nous utilisons le logiciel `evalb` que nous avons modifié pour qu'il donne également le score oracle quand l'analyseur fournit une liste d'arbres.
8. La ponctuation n'est pas prise en compte.

	DEV-1	DEV-10	DEV-50	DEV-100	GOLD
F	84,41	89,35	91,71	92,56	100
LAS	86,01	89,25	90,86	91,48	96,04
UAS	89,60	92,70	94,23	94,80	100
F agrégats	85,02	89,98	92,33	93,27	100
LAS agrégats	86,99	89,93	91,61	92,25	96,04
UAS agrégats	90,72	93,48	95,05	95,68	100

TABLE 2 – Scores de l'analyseur génératif sur la partie de développement

Les résultats donnés ci-dessus nous permettent de tirer deux conclusions importantes. D'une part les résultats de l'analyseur sont du niveau de l'état de l'art pour l'analyse syntagmatique du français ($84,41\%$ de F-score). D'autre part, la marge de progression du réordonnanceur est importante puisque le score oracle LAS sur les 100 analyses les plus probables est de $91,48\%$ alors que le score de l'analyse la plus probable est de $86,01\%$ soit une marge possible de progression de $5,47\%$.

4.2 Ajout du réordonnanceur

4.2.1 Apprentissage

Le modèle discriminant, c'est-à-dire les instances des patrons de traits et leur poids, est appris sur le corpus d'entraînement. L'analyseur génératif produit 100 analyses par phrase [9] pour ce corpus qui servent d'exemples d'apprentissage au réordonnanceur. Le modèle donne 71 millions de traits pour la grammaire simple et 75 millions pour la grammaire d'agrégats. Notez qu'un tel nombre de traits n'est pas pénalisant car l'algorhtme discriminant ne donne un poids non nul qu'aux traits utiles.

4.2.2 Évaluation

	base	10	20	50	100
F	84,41	85,38	**85,58**	85,55	85,32
LAS	86,01	87,06	87,31	**87,39**	87,28
UAS	89,60	90,57	90,77	**90,83**	90,69

grammaire simple

	base	10	20	50	100
F	85,00	85,94	86,12	**86,20**	86,15
LAS	86,99	88,00	88,06	88,10	**88,17**
UAS	90,78	91,54	91,57	91,58	**91,62**

grammaire apprise sur agrégats

TABLE 3 – scores du réordonnanceur en fonction du nombre de candidats

Pour notre évaluation, nous avons testé plusieurs configurations sur le corpus de développement en faisant varier le nombre de candidats fourni au réordonnanceur lors de la phase de prédiction [10]. Les résultats sont présentés dans la table 3. Pour la métrique LAS, les meilleurs résultats sont obtenus en donnant 50 candidats au réordonnanceur dans le cas de la grammaire simple et 100 dans le cas de la grammaire d'agrégats. Mais la différence avec les autres configurations n'est pas significative.

Puisque la meilleure configuration pour les différentes grammaires n'est pas la même selon la métrique utilisée mais que la configuration à 50 candidats est toujours la meilleure selon l'une d'entre elles, c'est cette configuration qui est utilisée sur le corpus de test pour l'évaluation finale avec la grammaire simple et la grammaire d'agrégats. Les résultats, dans la table 4, sont du même ordre de grandeur [11]. Pour la grammaire simple, le réordonnanceur de notre système permet de passer d'un F-score de 85,09% à 86,02%, pour le LAS de 86,68% à 87,91% et pour le UAS de 90,22% à 91,31%. Sur les 3 métriques le réordonnanceur montre une amélioration significative [12]. Il est intéressant de noter que nos traits portent uniquement sur les structures en dépendances (mis à part le score

9. Pour certaines phrases, les phrases courtes en particulier, notre système renvoie moins de 100 analyses.
10. Les poids des traits sont toujours appris avec 100 candidats par phrase.
11. $F < 40$ est le F-score en constituants pour les phrases de moins de 40 mots.
12. Les différences de scores entre le système de base et les nouveaux systèmes incorporant le module discriminant sont statistiquement significatives avec une valeur $p < 0.01$. Les différences entre les nouveaux systèmes ne le sont pas.

attribué par l'analyseur PCFG-LA) et que le F-score qui mesure la qualité des arbres syntagmatiques est tout de même améliorée.

	base	réord.
F	85,09	86,02
F < 40	87,10	87,82
LAS	86,68	87,91
UAS	90,22	91,31

grammaire simple

	base	réord.
F	86,35	86,89
F < 40	88,45	88,74
LAS	87,37	88,45
UAS	91,08	91,91

grammaire apprise sur agrégats

TABLE 4 – Scores du système sur le corpus de test

4.2.3 Comparaisons

	F < 40	LAS	UAS
Ce travail	87,82	87,91	91,31
Ce travail + agrégats	**88,74**	**88,45**	**91,91**
MATE + MELT	–	88,17	90,15
BKY	88,2	86,8	91,0
MST	–	88,2	90,9

TABLE 5 – Comparaisons des différents résultats d'analyse syntaxique

Nous donnons un tableau récapitulatif de différents résultats d'analyseurs sur le FTB dans la table 5. Nous comparons notre système (nous avons choisi les configurations qui donnaient les meilleurs scores LAS sur le corpus de développement) avec l'analyseur en dépendances MATE (Bohnet, 2010), entraîné et évalué avec MELT (Denis & Sagot, 2010) pour l'étiquetage en parties du discours. Nous citons aussi les travaux de comparaisons de (Candito *et al.*, 2010b), avec un premier système (BKY dans la table) proche du nôtre avec un analyseur syntagmatique qui fournit une sortie convertie en arbre de dépendances. Le second système (MST dans la table) est l'analyseur MSTParser de (McDonald *et al.*, 2005). Ces deux systèmes utilisent aussi les agrégats et non directement les mots du texte.

4.2.4 Analyse

Une analyse du vecteur de poids produit par le modèle discriminant montre que seuls 27% des 75 millions de traits observés dans les données d'entraînement correspondent à des poids non nuls (modèle avec agrégats). Les autres traits ont donc été jugés non discriminants. Nous avons analysé les 1000 traits de plus grand poids positif, représentant les caractéristiques jugées les plus pertinentes pour discriminer les bonnes analyses et les 1000 traits de poids négatifs de plus grande valeur absolue, symptomatiques des mauvaises analyses.

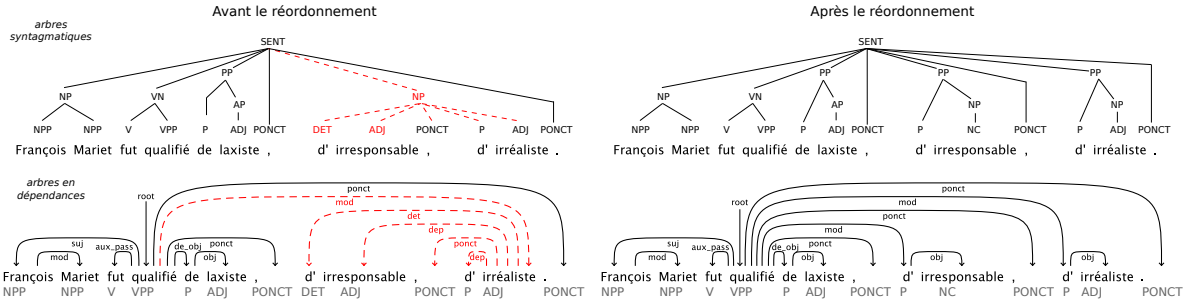


FIGURE 2 – Exemple de réordonnancement bénéfique (163e phrase de développement) : la meilleure analyse selon le modèle génératif est à gauche, la meilleure selon le modèle discriminant, la 35e hypothèse du modèle génératif, est à droite. Les erreurs, indiquées par des pointillés, sont propagées lors de la conversion en dépendances.

Cette analyse fait apparaître que les traits issus du repli [13] sont utiles pour caractériser de mauvaises analyses (63% des 1000 poids les plus négatifs). De plus, les traits négatifs font souvent référence à un indicateur d'échec de la lemmatisation. Comme cette lemmatisation est produite à partir d'un lexique et de la paire forme / partie du discours, et sous l'hypothèse que notre lexique est suffisamment riche, une erreur signifie que l'étiquetage en partie du discours est mauvais.

Le trait positif le plus discriminant est, sans surprise, le score donné par l'analyseur. Les autres traits positifs importants permettent d'avantager des étiquetages de parties du discours en fonction du contexte linéaire. On remarque aussi l'existence de traits qui incitent les conjoints à avoir la même catégorie dans les coordinations.

On voit donc que le réordonnancement peut remettre en question l'étiquetage de l'analyseur mais aussi qu'il peut influencer le traitement de phénomènes plus complexes comme la coordination. La figure 2 donne un exemple de phrase dont la meilleure analyse a été corrigée par le modèle discriminant.

Les patrons les plus représentés en positif sont **unigramme**, **bigramme**, **contexte linéaire**, **chaînes**, et leurs versions relachées, et en négatif **unigramme**, **bigramme**, **chaînes**, **nœuds frères**. Une analyse des résultats par type de dépendance montre que le réordonnanceur fait moins d'erreurs uniformément sur l'ensemble des types. La figure 2 donne un exemple de phrase dont l'analyse a été corrigée par le modèle discriminant.

5 Conclusion

Nous avons montré que l'ajout d'un module discriminant permet d'améliorer la qualité des analyses, comme nous avons pu le vérifier expérimentalement sur le corpus FTB à l'aide de trois métriques (F-score Parseval, LAS, UAS). Cependant, le gain est moins important que celui observé pour l'anglais (Charniak & Johnson, 2005). Sans procéder à une analyse d'erreurs exhaustive, on peut toutefois évoquer plusieurs points pouvant être améliorés.

En premier lieu, l'approche séquentielle est vulnérable aux erreurs en cascade. Bien que l'analyseur génératif fournisse plusieurs candidats, ce n'est pas le cas de l'étiqueteur fonctionnel. Les erreurs d'étiquetage ne sont donc pas récupérables. On peut envisager deux solutions ici : (1) permettre à l'étiqueteur de renvoyer une sortie ambiguë et laisser le réordonnanceur décider du meilleur étiquetage, et (2) utiliser des techniques plus sophistiquées dans la phase discriminante comme la prédiction structurée. On pourra ainsi se passer de l'étiqueteur.

Ensuite, cette approche à deux niveaux ne permet pas d'atteindre l'oracle, ce qui a déjà été observé sur l'anglais. Il est difficile de trouver un jeu de traits suffisamment général et qui soit utile pour une phrase particulière. Il y a encore des investigations à mener pour trouver un jeu de traits optimal, par exemple sur l'utilisation plus systématique de traits sur les structures de dépendances (par exemple en passant par des noyaux d'arbres), ou celle de connaissances externes, linguistiques ou collectées sur corpus (préférences lexicales, cadres de sous-catégorisation, verbes copulatifs, symétrie de la coordination...).

Enfin, notre système doit encore être amélioré pour une utilisation en situation réelle (par exemple, l'analyse de gros volumes de données provenant de la toile). L'extraction des traits pour le réordonnanceur, et vraisemblablement pour l'étiqueteur fonctionnel, est clairement le goulot d'étranglement. À titre d'exemple, pour analyser les 1235 phrases du corpus de test, les temps d'analyse en constituants, d'étiquetage fonctionnel et de conversion en structures de dépendances, d'extraction de traits et finalement de réordonnancement, sont respectivement : 9 min 55 s, 25 min 43 s, 51 min 12 s et 4 min 03 s.

Ce travail ouvre la voie à l'utilisation de données non étiquetées en plus d'un corpus arboré pour apprendre une meilleure grammaire générative (*self training* comme (McClosky *et al.*, 2008)) où le réordonnanceur évite au système d'apprendre sur les erreurs commises par l'analyseur génératif.

Remerciements

Ces travaux sont en partie financés par le projet ANR Sequoia ANR-08-EMER-013. Nous tenons à remercier Marie Candito qui nous a aidés à maîtriser BONSAÏ, Djamé Seddah qui nous a suggéré de tester notre architecture sur les grammaires d'agrégats, ainsi que les relecteurs anonymes.

13. C'est-à-dire qu'une partie de l'information est masquée (cf. section 3.2).

Références

ABEILLÉ A., CLÉMENT L. & FRANÇOIS T. (2003). *Treebanks*, chapter Building a treebank for French. Kluwer, Dordrecht.

ATTIA M., FOSTER J., HOGAN D., LE ROUX J., TOUNSI L. & VAN GENABITH J. (2010). Handling Unknown Words in Statistical Latent-Variable Parsing Models for Arabic, English and French. In *Proceedings of SPMRL*.

BOHNET B. (2010). Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of COLING*.

CANDITO M., CRABBÉ B. & DENIS P. (2010a). Statistical French Dependency Parsing : Treebank Conversion and First Results. In *Proceedings of LREC2010*.

CANDITO M.-H. & CRABBÉ B. (2009). Improving Generative Statistical Parsing with Semi-Supervised Word Clustering. In *Proceedings of IWPT 2009*.

CANDITO M.-H., CRABBÉ B., DENIS P. & GUÉRIN F. (2009). Analyse syntaxique du français : des constituants aux dépendances. In *Actes de TALN*.

CANDITO M.-H., NIVRE J., DENIS P. & HENESTROZA ANGUIANO E. (2010b). Benchmarking of Statistical Dependency Parsers for French. In *Proceedings of COLING'2010*.

CHARNIAK E. & JOHNSON M. (2005). Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In *Proceedings of ACL*.

COLLINS M. (1997). Three Generative, Lexicalised Models for Statistical Parsing. In *Proceedings of the 35th Annual Meeting of the ACL*.

COLLINS M. (2000). Discriminative Reranking for Natural Language Parsing. In *Proceedings of ICML*.

CRABBÉ B. & CANDITO M. (2008). Expériences d'analyse syntaxique du français. In *Actes de TALN*.

CRAMMER K., DEKEL O., KESHET J., SHALEVSHWARTZ S. & SINGER Y. (2006). Online Passive-Aggressive Algorithm. *Journal of Machine Learning Research*.

DENIS P. & SAGOT B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. In *Actes de TALN*.

JOHNSON M. & URAL A. E. (2010). Reranking the Berkeley and Brown Parsers. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 665–668, Los Angeles, California : Association for Computational Linguistics.

MATSUZAKI T., MIYAO Y. & ICHI TSUJII J. (2005). Probabilistic CFG with Latent Annotations. In *Proceedings of ACL*.

MCCLOSKY D., CHARNIAK E. & JOHNSON M. (2008). When is Self-Training Effective for Parsing ? In D. SCOTT & H. USZKOREIT, Eds., *COLING*, p. 561–568.

MCDONALD R. (2006). *Discriminative Training and Spanning Tree Algorithms for Dependency Parsing*. PhD thesis, University of Pennsylvania.

MCDONALD R., CRAMMER K. & PEREIRA F. (2005). Online Large-Margin Training of Dependency Parsers. In *Association for Computational Linguistics (ACL)*.

NIVRE J., HALL J., NILSSON J., CHANEV A., ERYIGIT G., KÜBLER S., MARINOV S. & MARSI E. (2007). Maltparser : A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, **13**(2), 95–135.

PETROV S., BARRETT L., THIBAUX R. & KLEIN D. (2006). Learning Accurate, Compact, and Interpretable Tree Annotation. In *ACL*.

PETROV S. & KLEIN D. (2007). Improved Inference for Unlexicalized Parsing. In *HLT-NAACL*, p. 404–411.

PULLUM G. K. & SCHOLZ B. C. (2001). On the distinction between model-theoretic and generative-enumerative syntactic frameworks. In *Logical Aspects of Computational Linguistics*.

RAMBOW O. (2010). The Simple Truth about Dependency and Phrase Structure Representations : An Opinion Piece. In *NAACL HLT*.

ROSENBLATT F. (1958). The Perceptron : A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*.

SEDDAH D., CANDITO M. & CRABBÉ B. (2009). Adaptation de parsers statistiques lexicalisés pour le français : Une évaluation complète sur corpus arborés. In *TALN*.

TALN 2011, Montpellier, 27 juin –1er juillet 2011

Apport de la syntaxe pour l'extraction de relations en domaine médical

Anne-Lyse Minard[1, 2] Anne-Laure Ligozat[1, 3] Brigitte Grau[1, 3]
(1) LIMSI-CNRS, BP 133, 91403 Orsay cedex
(2) Université Paris-Sud, 91400 Orsay
(3) ENSIIE, 1 square de la résistance, 91000 Évry
prenom.nom@limsi.fr

Résumé. Dans cet article, nous nous intéressons à l'identification de relations entre entités en domaine de spécialité, et étudions l'apport d'informations syntaxiques. Nous nous plaçons dans le domaine médical, et analysons des relations entre concepts dans des comptes-rendus médicaux, tâche évaluée dans la campagne i2b2 en 2010. Les relations étant exprimées par des formulations très variées en langue, nous avons procédé à l'analyse des phrases en extrayant des traits qui concourent à la reconnaissance de la présence d'une relation et nous avons considéré l'identification des relations comme une tâche de classification multi-classes, chaque catégorie de relation étant considérée comme une classe. Notre système de référence est celui qui a participé à la campagne i2b2, dont la F-mesure est d'environ 0,70. Nous avons évalué l'apport de la syntaxe pour cette tâche, tout d'abord en ajoutant des attributs syntaxiques à notre classifieur, puis en utilisant un apprentissage fondé sur la structure syntaxique des phrases (apprentissage à base de tree kernels) ; cette dernière méthode améliore les résultats de la classification de 3%.

Abstract. In this paper, we study relation identification between concepts in medical reports, a task that was evaluated in the i2b2 campaign in 2010, and evaluate the usefulness of syntactic information. As relations are expressed in natural language with a great variety of forms, we proceeded to sentence analysis by extracting features that enable to identify a relation and we modeled this task as a multiclass classification task based on SVM, each category of relation representing a class. This method obtained an F-measure of 0.70 at i2b2 evaluation. We then evaluated the introduction of syntactic information in the classification process, by adding syntactic features, and by using tree kernels. This last method improves the classification up to 3%.

Mots-clés : extraction de relation, domaine médical, apprentissage multi-classes, tree kernel.

Keywords: relation identification, medical domain, multiclass learning, tree kernel.

1 Introduction

L'extraction d'information permet d'obtenir des représentations structurées du contenu d'un corpus. Le domaine médical représente en cela un domaine d'application intéressant : en effet, les documents médicaux tels que les comptes-rendus cliniques contiennent de nombreuses informations sur le suivi médical des patients, et la structuration automatique de ces informations pourrait améliorer la prise en charge de ceux-ci.

L'extraction de ces informations amène à se poser différents problèmes, liés aux types d'information cherchés : la reconnaissance des termes du domaine dans les textes, des concepts qui leur sont liés, ainsi que l'identification des types de relations qui les lient dans les documents.

Nous nous sommes intéressées à l'identification de relations dans des comptes-rendus médicaux, tâche qui a fait l'objet d'une campagne d'évaluation i2b2 en 2010 [1]. Un premier travail a été réalisé modélisant l'identification des relations comme une tâche de classification multi-classes (Minard *et al.*, 2011b). Cette approche a été choisie car elle permet de caractériser les relations par des ensembles de traits lexicaux et de surface, et a conduit à l'obtention de bons résultats à i2b2. Néanmoins, les phrases étant parfois complexes, leur représentation par des traits de surface uniquement ne permet pas de capturer des relations entre termes distants. C'est pourquoi nous nous sommes posé la question de l'utilité des traits syntaxiques pour la reconnaissance de relations en domaine de spécialité : des attributs portant de l'information sur la structure syntaxique des phrases améliorent-ils l'extraction ? Et des approches par apprentissage sur des arbres syntaxiques sont-elles meilleures que des approches «sac de mots» ?

Après avoir présenté les travaux existant dans ce domaine, nous présenterons le contexte de notre étude, puis nous expliquerons nos méthodes et évaluerons notre approche.

2 L'extraction de relations

L'extraction de relations a donné lieu à de nombreux travaux, notamment dans le domaine biomédical. Les approches actuelles se fondent sur une classification automatique plus ou moins supervisée.

(Roberts *et al.*, 2008) proposent une approche classique fondée sur des SVM (machine à vecteurs de support) pour extraire des relations dans un corpus de spécialité : le corpus du projet CLEF (the Clinical E-Science Framework project). Ils extraient des relations entre des entités (ex : condition, médicament, résultat) et des modifieurs (ex : marqueur de négation) dans des dossiers de patients atteints d'un cancer. Les relations sont de sept types. Deux types d'entités (ou une entité et un modifieur) ne peuvent être reliées que par une relation (sauf entre une investigation et une condition où la relation peut être de deux types). La tâche est donc modélisée comme une classification binaire, c'est-à-dire que les SVM sont entraînés pour une décision entre une classe et toutes les autres. Les attributs qu'ils utilisent correspondent à des attributs de notre système de base :

– des attributs lexicaux : les mots (et les radicaux des mots) dans une fenêtre de 6 mots avant et après les entités en relation, les mots formant les entités, les mots situés entre les entités ;
– des attributs morpho-syntaxiques : les catégories morpho-syntaxiques et ces mêmes catégories généralisées (par exemple toutes les catégories verbales sont regroupées en *VB*) ;
– des attributs sémantiques : le type des entités en relation et des autres entités de la phrase.

D'autres travaux utilisent des attributs syntaxiques plus riches ((Zhou *et al.*, 2005), (Uzuner *et al.*, 2010)). En particulier, (Uzuner *et al.*, 2010) utilisent les dépendances syntaxiques entre les concepts sous forme d'attributs dans une approche vectorielle basée sur des SVM. Ils souhaitent typer des relations entre des problèmes, des tests et des traitements dans des comptes-rendus médicaux. Ils classent les relations en six catégories, comme par exemple les relations existantes entre une maladie qu'a le patient et un traitement, ou les relations entre une éventuelle maladie et un traitement. Ils utilisent des attributs classiques (l'ordre des concepts, la distance, des trigrammes lexicaux, les mots qui forment les concepts, les verbes, des bigrammes syntaxiques, etc.) ainsi que des attributs portant des informations sur les dépendances syntaxiques reliant les entités. Pour plusieurs relations ils obtiennent des F-mesures entre 0,60 et 0,85, mais pour les relations pour lesquelles il y a peu d'exemples dans le corpus d'apprentissage les F-mesures sont nulles. Ils évaluent leurs classes d'attributs, et montrent que les attributs qui apparaissent comme les plus utiles sont les trigrammes lexicaux et les mots qui forment les concepts. Les informations syntaxiques n'améliorent pas la classification notamment car les dépendances syntaxiques complètes ne sont trouvées que pour une faible proportion des phrases analysées.

Des travaux en domaine ouvert ont cependant montré que l'information structurelle utilisée sous forme d'arbres grâce à des tree kernels améliore la classification ((Culotta & Sorensen, 2004), (Zelenko *et al.*, 2003), (Zhang *et al.*, 2006)). En particulier, (Zhang *et al.*, 2006) ont étudié l'apport de la structure syntaxique des phrases pour l'extraction de relation en domaine général, en s'appuyant

1. https ://www.i2b2.org/NLP/Relations/

sur le corpus ACE 2003. Ils testent différentes sélections dans les arbres syntaxiques (arbre complet englobant les deux entités en relation, plus petit arbre commun, en conservant que les chunks, etc.), et ils montrent que les meilleurs résultats sont obtenus en utilisant le plus petit sous-arbre commun aux deux entités. (Culotta & Sorensen, 2004) utilisent des arbres de dépendance sur le même corpus, et montrent que les tree kernels sont meilleurs que l'information structurelle mise sous forme vectorielle. Ils testent deux types de tree kernels : *contiguous kernel* qui n'apparie pas les séquences qui sont interrompues par des nœuds non appariés, et *sparse tree* qui autorise les nœuds non appariés à l'intérieur de séquences appariées. Les meilleurs résultats sont obtenus avec des *contiguous kernel* associés à des kernels «sac de mots».

3 Objectif et méthodes

Nous nous sommes intéressées à l'extraction de relations en domaine de spécialité. Notre objectif était d'étudier comment intégrer des informations syntaxiques dans un système de classification automatique, et quel était l'apport de ce type d'informations. Nous avons considéré deux sous-tâches : la détection de la présence d'une relation entre deux entités, en l'occurrence des concepts médicaux, et la catégorisation de cette relation éventuelle selon des catégories prédéfinies. Ces sous-tâches ont été abordées comme des problèmes de classification supervisée bi-classes ou multi-classes.

Nous avons tout d'abord ajouté des attributs portant des informations sur la structure syntaxique des phrases aux vecteurs linéaires. Mais l'information syntaxique structurelle étant difficile à décrire par un vecteur d'attributs linéaires, nous avons ensuite utilisé les tree kernels qui permettent d'explorer les attributs contenus dans la structure des arbres en calculant la similarité des arbres deux à deux.

3.1 Classification

Nous avons utilisé des classifieurs à base de SVM car ils sont très présents dans l'état de l'art des systèmes d'extraction de relations. De plus ils donnent de bons résultats pour les tâches pour lesquelles il y a beaucoup d'attributs mais qui sont très épars, comme c'est souvent le cas en TAL. Pour tenir compte de l'information syntaxique, nous avons choisi d'utiliser une fonction kernel qui mesure la similarité entre deux arbres, en comptant le nombre de fragments en commun. L'arbre est découpé en fragments de différentes tailles. Deux options sont possibles, soit ST (*subtrees*) qui calcule tous les sous-arbres possibles avec tous leurs descendants (voir figure 1), soit SST (*subset tree*) qui autorise également les fragments d'arbres dont les feuilles ne sont pas des éléments terminaux, mais des chunks ou des étiquettes morpho-syntaxiques (voir figure 2). Pour la classification binaire, c'est-à-dire la détection de relation, nous avons utilisé le logiciel SVM-Light-TK version 1.5 de (Moschitti, 2006). Nous avons paramétré le classifieur de la façon suivante : combinaison d'arbres et de vecteurs comme type de fonction kernel, et somme des contributions des arbres et des vecteurs comme opérateur de combinaison. Nous avons choisi d'utiliser l'option SST qui est plus générale et donne de meilleurs résultats selon (Moschitti, 2006).

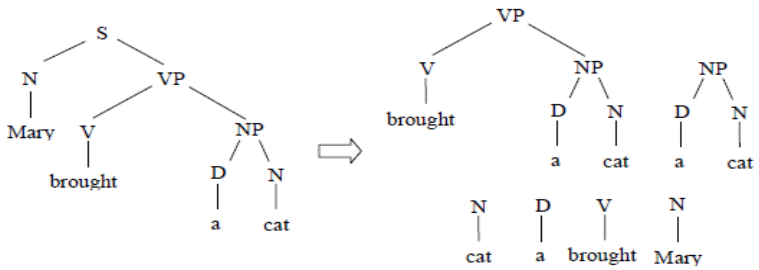


FIGURE 1 – Exemple de subtrees (ST) de (Moschitti, 2006)

Pour la classification multi-classes, c'est-à-dire la détection de relation et le typage des relations, nous avons utilisé le programme LIBSVM (Chang & Lin, 2001) avec une approche « un-contre-un ». Nous avons utilisé un noyau linéaire, ce choix est conseillé par (Hsu *et al.*, 2003) quand le nombre d'attributs est important par rapport au nombre d'instances. Nous avons fixé la valeur du paramètre *C* (*penalty parameter*) à *16* et celle du paramètre *gamma* (*kernel parameter*) à *0,03125*. Ces valeurs ont été choisies par l'outil *grid* fourni avec LIBSVM.

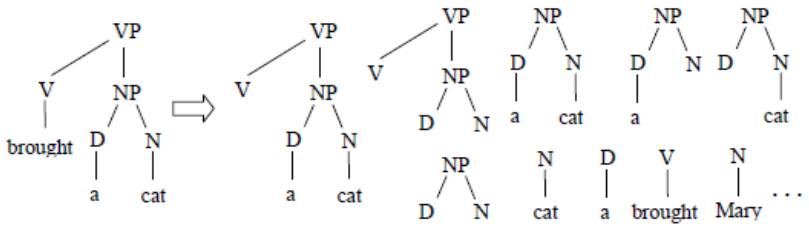


FIGURE 2 – Exemple de subset trees (SST) de (Moschitti, 2006)

3.2 Attributs de référence

Les attributs choisis pour la classification automatique prennent en compte des informations de surface, sur les distances entre mots par exemple, des informations lexicales, comme les mots formant les concepts, des informations morpho-syntaxiques, comme les catégories morpho-syntaxiques des mots, et des informations sémantiques grâce à un typage de concepts.

Tous les attributs présentés ici constitueront les attributs de notre système de référence, auxquels seront ajoutées des informations syntaxiques. La sélection des attributs est présentée de manière plus détaillée dans (Minard *et al.*, 2011b).

3.2.1 Attributs de surface

Les attributs de surface sont relatifs à la position des concepts dans la phrase ; ce sont les suivants :
– l'ordre des concepts : celui-ci influence en effet la façon dont la relation est exprimée ;
– la distance entre les deux concepts en termes de nombre de mots [2] : deux concepts ont d'autant moins de chance d'être en relation qu'il y a de nombreux mots entre les deux ;
– la présence d'autres concepts entre les deux concepts étudiés : la présence d'un troisième concept modifie les probabilités que les deux concepts soient bien en relation.

3.2.2 Attributs lexicaux

Plusieurs attributs lexicaux sont pris en compte :
– les mots, et leurs radicaux (stems) [3], qui composent les concepts, et le mot tête de chaque concept [4]. Nous avons considéré les radicaux de manière à regrouper les variantes flexionnelles et dérivationnelles ;
– les radicaux des trois mots à gauche et à droite des concepts. La taille de la fenêtre a été choisie après plusieurs tests : avec une fenêtre plus grande ou plus petite la précision augmente très légèrement mais le rappel diminue ;
– les radicaux des mots entre les concepts, ce qui permet de tenir compte de tous les mots entre les concepts ; c'est ici qu'est située l'information la plus utile à la classification ;
– les radicaux des verbes dans une fenêtre de trois mots avant et après chaque concept, et entre eux, le verbe marquant souvent la relation ;
– les prépositions entre les concepts.

3.2.3 Attributs morpho-syntaxiques

Le TreeTagger est appliqué sur les phrases à analyser, et son étiquetage est utilisé pour les attributs suivants :

– la catégorie morpho-syntaxique des mots dans une fenêtre de trois mots à gauche et à droite des concepts ;
– la présence d'une préposition entre les concepts, peu importe la préposition ;
– la présence d'une conjonction de coordination ou d'une virgule entre les concepts ;

2. Le découpage en mots est effectué par le TreeTagger (Schmid, 1994).
3. Pour obtenir les radicaux des mots nous utilisons le module PERL lingua : :stem.
4. La tête d'un concept correspond au mot précédant une préposition ou le dernier mot du concept, comme défini dans (Zhou *et al.*, 2005).

– si les concepts ne sont séparés que par un mot, un attribut marque la présence d'un signe de ponctuation. Cet attribut permet de considérer les énumérations différemment.

3.2.4 Attributs sémantiques

Enfin, des attributs sémantiques permettent de généraliser l'information portée par certains mots des phrases et concernent les concepts du domaine d'une part et les classes de verbes d'autre part :
– le type sémantique (issu de l'UMLS [5]) des mots dans une fenêtre de trois mots à gauche et à droite de chaque concept ;
– les types des concepts : c'est l'attribut le plus important car les relations possibles dépendent des types des deux concepts considérés ;
– les classes de VerbNet [6] (une extension des classes de Levin) des verbes dans un fenêtre de trois mots à gauche et à droite des concepts, et entre les concepts.

3.3 Informations syntaxiques

Les informations syntaxiques utilisées proviennent des arbres de constituants ; ces arbres syntaxiques ont été produits par l'analyseur de Charniak/McClosky (McClosky, 2010). Nous avons choisi d'utiliser cet analyseur car il est entraîné sur des textes biomédicaux et obtient de bons résultats dans ce domaine (f-score de 87,6% [7]). Les phrases ont été analysées après remplacement des abréviations, normalisation des dates, âges, noms propre et nombres, et annotation des concepts. La figure 3 est un exemple d'arbre fourni par l'analyseur à partir de la phrase suivante :

2 Low back strain requiring hospitalization for pain in 2002.

avec *hospitalization* étiqueté comme un concept de type traitement et *pain* étiqueté comme un concept de type problème (la balise <NUM> remplace un nombre et la balise <DATE> une date ou une année).

À partir de cet arbre nous avons produit le sous-arbre minimal reliant les deux entités possiblement en relation. Ce sous-arbre correspondant au chemin le plus court pour aller d'un concept à l'autre, nous l'appellerons «sous-arbre minimal complet». Nous avons également produit un sous-arbre plus restreint, qui est équivalent au sous-arbre minimal complet, sauf que nous n'avons pas gardé le contexte gauche de la première entité ni le contexte droit de la deuxième entité.

La figure 4 représente le sous-arbre minimal complet produit à partir de l'arbre présenté dans la figure 3 et la figure 5 le sous-arbre minimal. Ces trois arbres sont utilisés comme attributs pour la classification des relations.

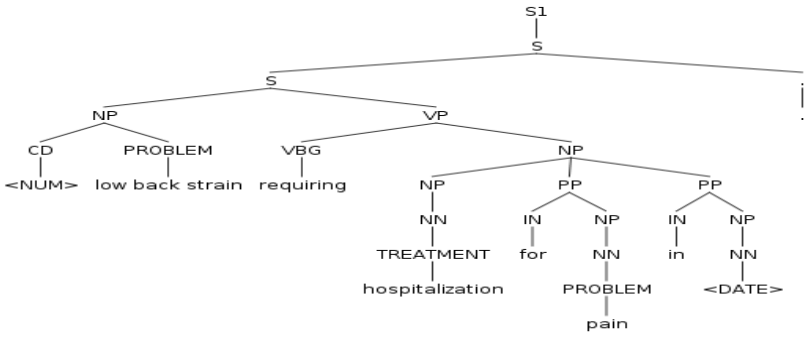


FIGURE 3 – Exemple de l'arbre complet

À partir du sous-arbre minimal nous avons calculé deux attributs : la taille du chemin reliant les deux entités (nous comptons le chemin entre les nœuds ayant pour valeur les types des concepts), et le constituant du nœud racine du sous-arbre. Pour le couple *hospitalization* et *pain*, la taille du plus petit chemin reliant les entités est sept et le constituant du nœud racine du sous-arbre minimal est *NP* (voir figure 5).

5. http ://www.nlm.nih.gov/research/umls/
6. http ://verbs.colorado.edu/ mpalmer/projects/verbnet.html
7. http ://www.cs.brown.edu/ dmcc/biomedical.html

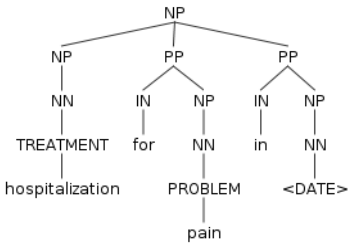


FIGURE 4 – Exemple du sous-arbre minimal complet entre les deux entités

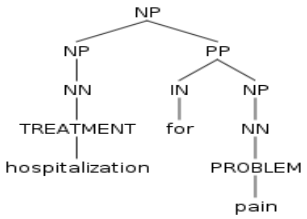


FIGURE 5 – Exemple du sous-arbre minimal entre les deux entités

4 Expérimentations

Les résultats présentés ici s'inscrivent dans le cadre de la campagne d'évaluation i2b2 en 2010. Trois tâches étaient proposées lors de cette campagne : la première consistait en l'annotation des problèmes médicaux, des traitements et des tests, la seconde en la détermination du statut d'assertion et la troisième en l'identification de relations. Nous avons travaillé sur la troisième tâche.

Les corpus, fournis par les organisateurs de la tâche i2b2, sont composés de comptes-rendus hospitaliers en anglais provenant de plusieurs centres médicaux aux États-Unis. Ces documents avaient été anonymisés et annotés manuellement. Un premier corpus a été fourni avant l'évaluation, composé de 350 documents. Puis, les organisateurs d'i2b2 ont fourni le corpus d'évaluation, qui comporte 477 documents.

Nous présentons dans une première partie les relations que nous devons identifier, dans une deuxième partie le corpus et nous terminons par les résultats des tests.

4.1 Relations considérées

La tâche consistait à annoter dans les comptes-rendus les relations existant entre deux concepts. Les concepts considérés étaient les suivants :

– les problèmes médicaux : les observations des patients ou cliniciens concernant ce qui n'allait pas ou semblait être causé par une maladie. Cette catégorie comprend notamment les maladies, syndromes, signes, et symptômes, les observations sur l'état mental du patient, les résultats anormaux de tests etc. ;
– les traitements : les procédures, interventions, substances et médicaments donnés à un patient pour tenter de résoudre un problème médical ;
– les tests : les procédures et examens effectués sur un patient ou un fluide corporel pour vérifier ou infirmer la présence d'un problème, ou pour avoir plus d'informations sur un problème.

Ces concepts peuvent être reliés par des relations, par exemple un examen a pu être prescrit pour analyser un problème médical, il peut également révéler un problème. Nous avons cherché à extraire ces relations dans les documents du corpus. Afin d'étudier spécifiquement l'extraction des relations, les documents sur lesquels nous avons travaillé comportaient déjà l'annotation de référence des concepts [8], et il s'agissait donc de déterminer si, étant donné deux concepts, ils étaient en relation, et si oui, quel était le type de la relation.

Trois ensembles de relations ont été définis :

– relations entre problème et traitement :
  – le traitement améliore le problème (TrIP)
    Exemple : «<pb>hypertension</pb> was controlled on <treat>hydrochlorothiazide</treat>»
  – le traitement aggrave le problème (TrWP), ce qui inclut les cas où le traitement est administré pour le problème mais ne l'améliore pas
    Exemple : «<pb>the tumor</pb> was growing despite the available <treat>chemotherapeutic regimen</treat>»

8. Dans les exemples présentés dans l'article nous annotons les problèmes médicaux avec la balise <*pb*>, les tests avec la balise <*test*> et les traitements avec la balise <*treat*>

  - le traitement cause le problème (TrCP)
    Exemple : «<treat>Bactrim</treat> could be a cause of <pb>these abnormalities</pb>.»
  - le traitement est administré en raison du problème (TrAP)
    Exemple : «<treat>antibiotic therapy</treat> for presumed <pb>right forearm phlebitis</pb>»
  - le traitement n'est pas administré en raison du problème (TrNAP)
    Exemple : «<treat>Relafen</treat> which is contra-indicated because of <pb>ulcers</pb>.»
- relations entre problème et test :
  - le test révèle le problème (TeRP), et plus généralement les tests menés accompagnés de leurs résultats
    Exemple : «<test>an echocardiogram</test> revealed <pb>a pericardial effusion</pb>»
  - le test est conduit en raison du problème (TeCP)
    Exemple : «<test>an VQ scan</test> was performed to investigate <pb>pulmonary embolus</pb>»
- relations entre problème et problème
  - un problème en indique un autre (PIP), incluant les cas où des problèmes révèlent différents aspects d'un même problème médical
    Exemple : «a history of <pb>noninsulin dependent diabetes mellitus</pb>, now presenting with <pb>acute blurry vision on the left side</pb>.»

Le tableau 1 indique le nombre de relations par catégorie dans le corpus d'entraînement et d'évaluation, ainsi que l'accord inter-annotateur [9]. Nous observons que l'accord inter-annotateur est faible pour certaines catégories comme TrIP et TrWP.

Relation	entraînement	évaluation	accord inter-annotateur
TrIP	**107**	198	**0,62**
TrWP	**56**	143	**0,58**
TrCP	296	444	0,82
TrAP	1423	2487	0,95
TrNAP	106	191	0,76
PIP	1239	1986	0,79
TeRP	1734	3033	0,96
TeCP	303	588	0,74
Toutes	5264	9070	

TABLE 1 – Nombre de relations par catégorie dans chaque corpus

4.2 Corpus

Sur le corpus d'entraînement (350 documents) la taille moyenne des phrases contenant au moins deux concepts est de 16,78 mots/phrase (les signes de ponctuation ne sont pas comptés comme des mots). La phrase la plus courte contient deux mots, la phrase la plus longue 212 et la médiane est située à 15 mots/phrase. (Coden *et al.*, 2005) compare la taille moyenne des phrases de trois corpus, le premier est un extrait du Penn TreeBank (composé d'articles de journaux), le second le corpus GENIA (composé de résumé de MedLine) et le troisième est le corpus MED composé de rapports clinique. Les tailles moyenne des phrases de ces trois corpus ainsi que du corpus pour la campagne i2b2 2010 que nous utilisons sont répertoriées dans le tableau 2. Ces données montrent que le corpus sur lequel nous travaillons est composé de phrases courtes, comparé au corpus GENIA. En effet les documents type rapport clinique sont composés de beaucoup de fragments de phrase (e.g. 1) et d'énumérations (e.g. 2) ce qui fait leur particularité.

	taille moyenne des phrases
Penn Treebank	24,16
MED	13,79
GENIA	27,18
i2b22010 corpus	16,78

TABLE 2 – Taille moyenne des phrases de différents corpus

(1) <pb>C5-6 disc herniation</pb> with <pb>cord compression</pb> and <pb>myelopathy</pb> .

9. L'accord inter-annotateur nous a été fourni par les organisateurs, il a été calculé à partir de Knowtator.

(2) Revealed <pb>icteric sclerae</pb> , <pb>the oropharynx with extensive thrush</pb> , and <pb>an ulcer under his tongue</pb> .

4.3 Prétraitement du corpus

Les fichiers du corpus ont été prétraités et normalisés. Tout d'abord, les abréviations connues ont été remplacées par leur forme complète, grâce à une liste. La liste a été constituée pour la campagne d'évaluation i2b2 2009 [10] à partir de la liste d'abréviations biomédicales formée par Berman [11] à laquelle ont été ajoutés les exemples du corpus du challenge i2b2 2009. Ainsi, *h.o.* a été converti en *history of* et *p.r.n.* en *as needed*. Puis, les données d'anonymisation (différentes pour chaque corpus) ont été remplacées par des balises *NAME*, *DATE* et *AGE*. *NUM* remplace toutes les valeurs numériques présentes dans les fichiers (principalement des dosages). Enfin, les textes ont été étiquetés par le TreeTagger et analysés par le parser de Charniak/McClosky.

4.4 Tests et résultats

4.4.1 Système : sans information syntaxique

Nous avons d'abord évalué l'approche vectorielle avec les attributs de base décrits dans la section 3.2. Les résultats sont donnés dans le tableau 3. Les résultats de la ligne *Toutes relations* sont la précision moyenne, le rappel moyen et la F-mesure moyenne pour la classification de toutes les relations. Nous avons utilisé ce système pour la campagne d'évaluation i2b2. Nous obtenons une F-mesure de 0,703, ce qui nous permet de nous placer troisième sur seize participants (Minard *et al.*, 2011a).

Relation	Précision	Rappel	F-mesure
TrIP	0,852	0,146	0,250
TrWP	0,000	0,000	0,000
TrCP	0,735	0,362	0,485
TrAP	0,743	0,689	0,715
TrNAP	0,461	0,062	0,110
PIP	0,781	0,530	0,632
TeRP	0,876	0,832	0,853
TeCP	0,870	0,251	0,390
Toutes relations	**0,807**	**0,622**	**0,703**
Médiane			0,664
1er système	0,720	0,753	0,736
2e système	0,773	0,693	0,731

TABLE 3 – Précision, rappel et F-mesure obtenus sur le corpus d'évaluation avec le système de base

4.4.2 Système : avec des informations syntaxiques sous forme vectorielle

Nous avons voulu voir si l'ajout d'informations syntaxiques améliorerait la classification des relations. Pour cela nous avons calculé des informations à partir de l'arbre de constituants produit par l'analyseur Charniak/McClosky, que nous avons ajoutées aux attributs de base. Il s'agit de la taille du chemin entre le premier concept et le deuxième concept, et du nom du constituant du nœud racine du sous-arbre minimal. Ce système obtient une F-mesure de 0,700. Les résultats détaillés sont présentés dans le tableau 4.

4.4.3 Sytème à base de tree kernels

L'ajout d'attributs n'améliorant pas la classification, nous avons testé l'ajout d'informations structurelles plus importantes que les deux attributs précédents. Pour cela nous avons utilisé le classifieur SVM-Light-TK basé sur des tree kernels. Comme certaines

10. https ://www.i2b2.org/NLP/Medication/
11. http ://www.julesberman.info/abbtwo.htm

Relation	Précision	Rappel	F-mesure
TrIP	0,875	0,141	0,243
TrWP	0,000	0,000	0,000
TrCP	0,733	0,360	0,483
TrAP	0,741	0,684	0,712
TrNAP	0,444	0,062	0,110
PIP	0,776	0,525	0,626
TeRP	0,877	0,833	0,854
TeCP	0,869	0,248	0,386
Toutes relations	**0,806**	**0,619**	**0,700**

TABLE 4 – Précision, rappel et F-mesure obtenus sur le corpus d'évaluation avec le système utilisant des informations syntaxiques sous forme vectorielle

relations ne sont pas suffisamment représentées dans le corpus, nous n'avons pas fait de classification multi-classes avec SVM-Light-TK. En effet pour 5 des 8 relations (TrIP, TrWP, TrNAP, TrCP et TeCP), le classifieur ne détectait pas de relation. Les évaluations que nous avons faites portent donc sur de la classification entre relation et non-relation, c'est-à-dire de la détection de relation.

Nous avons ajouté les arbres de constituants complets ainsi que les sous-arbres minimaux complets entre les deux entités possiblement en relation et les sous-arbres minimaux (c.f. 3.3), pour évaluer si d'autres informations contenues dans les arbres pouvaient être utilisées pour la détection des relations.

Les arbres complets contiennent des informations supplémentaires par rapport aux sous-arbres minimaux. Dans les informations supprimées, il y a du bruit qui peut gêner la classification, mais il y a également des déclencheurs des relations. Il semble donc pertinent d'utiliser les deux types d'arbres pour avoir le maximum d'informations. Les sous-arbres minimaux contiennent en moyenne la moitié du nombre de mots de l'arbre complet. Plus précisément les arbres complets ont un nombre moyen de mots par phrase de 21 [12] (sans compter la ponctuation, et en comptant les concepts comme un seul mot), et un nombre moyen de nœuds de 48. Alors que les sous-arbres minimaux ont un nombre moyen de mots de 8 et un nombre moyen de nœuds de 22.

Nous avons évalué plusieurs combinaisons à partir des arbres de constituants complets (AC), des sous-arbres minimaux complets (AMC), des sous-arbres minimaux (AM) et des attributs du système de base (ATT). Les résultats sont présentés dans le tableau 5 et dans la figure 6.

Combinaison	Précision	Rappel	F-mesure
AC	0,749	0,611	0,673
AMC	0,623	0,726	0,651
AM	0,819	0,625	0,709
ATT	0,835	0,709	0,767
AC + AM	0,790	0,708	0,747
AC + ATT	0,826	0,731	0,776
AMC + ATT	0,832	0,729	0,773
AM + ATT	0,828	0,724	0,772
AC + AM + ATT	0,776	0,804	**0,790**
AMC + AM + ATT	0,819	0,727	0,770
AC + AMC + ATT	0,818	0,726	0,769
AC + AMC + AM + ATT	0,816	0,730	0,771

TABLE 5 – Évaluation des combinaisons des différents apprentissages à base de tree kernels pour la détection de relations

Ces résultats nous montrent que l'information contenue dans les arbres seule n'est pas suffisante pour la classification des relations quel que soit l'arbre utilisé. De plus la combinaison des arbres minimaux et des arbres complets apportent des meilleurs résultats que les arbres complets seuls, mais la F-mesure n'atteint pas celle obtenue avec les attributs de base seuls. Les attributs apportent donc des informations supplémentaires par rapport aux arbres. Dans cette étude nous n'avons pas évalué l'apport de chaque attribut vectoriel mais il serait intéressant de savoir quelles données ne sont pas fournies par les arbres mais sont données par les attributs.

12. Le nombre de mots par phrase est différent de celui donné dans la section 4.2 car nous avons un arbre par couple de concepts, c'est-à-dire que si une phrase contient trois concepts, nous avons trois arbres dans le corpus.

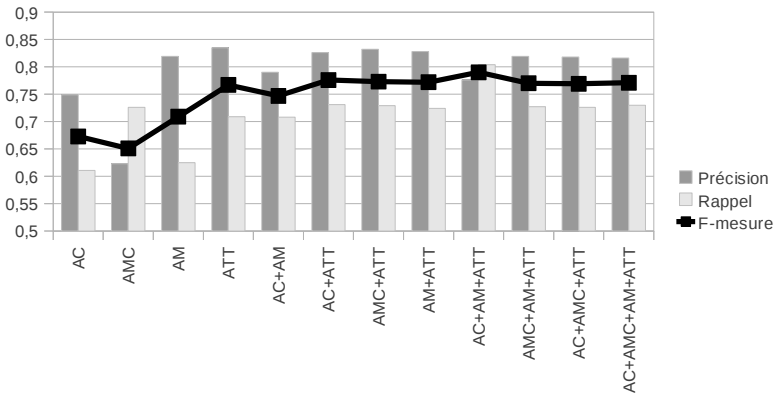


FIGURE 6 – Comparaison des différents apprentissages

Nous observons également que la meilleure combinaison est celle associant les arbres complets, les sous-arbres minimaux et les attributs (AC + AM + ATT). Les sous-arbres minimaux complets n'apportent pas d'informations supplémentaires. En effet ils apportent moins d'information que les arbres complets (la F-mesure pour AC + AM + ATT est de 0,790 et pour AMC + AM + ATT de 0,770), et plus d'informations bruitées que les arbres minimaux réduits (la F-mesure pour AC + AMC + ATT est de 0,771, contre 0,790 avec les arbres minimaux).

Nous avons effectué une première étude des relations détectées avec les attributs seuls et avec les attributs plus les arbres complets (AC + ATT). Nous avons observé que les arbres étaient utilisés pour détecter les relations entre des concepts éloignés dans la phrase. Les relations correctement détectées avec (AC + ATT) mais qui ne le sont pas avec (ATT) concernent deux concepts dont l'éloignement moyen est de 10,8 mots (ou ponctuations), alors que celles qui sont correctement classées par les deux systèmes concernent des concepts qui sont séparés en moyenne par 5,4 mots (ou ponctuations). Dans la figure 7 nous montrons une phrase contenant une relation de type PIP (un problème indique un autre problème) entre *increased tracer activity* et *active bleeding* ; les concepts sont séparés par 17 mots ou ponctuations. Cette relation a été correctement détectée lorsque les arbres étaient utilisés, mais elle n'est pas repérée avec l'utilisation des attributs seuls.

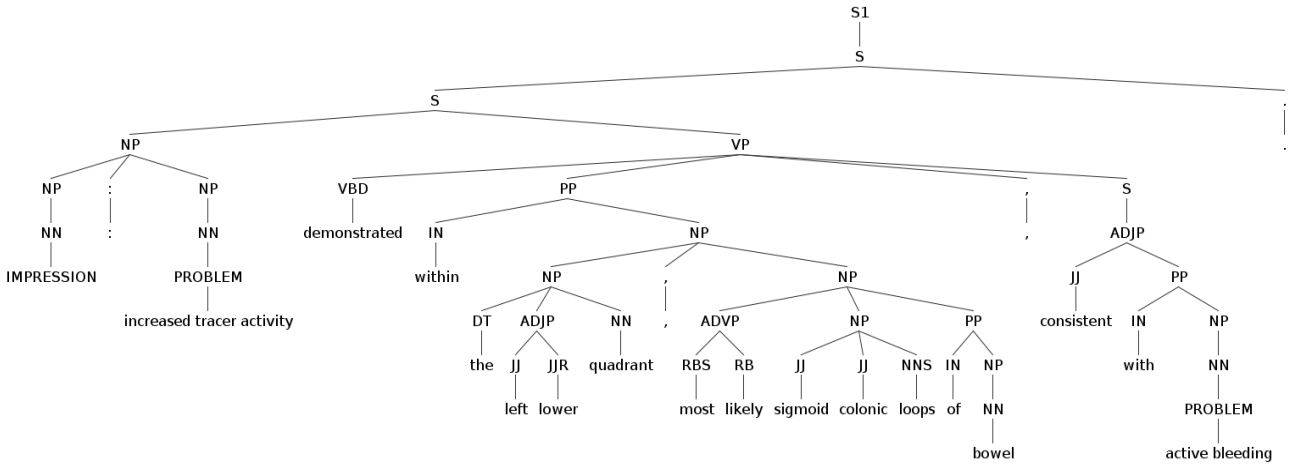


FIGURE 7 – Exemple de l'arbre complet d'une phrase contenant deux concepts reliés par une relation de type PIP

5 Conclusion

Dans cet article nous nous sommes intéressées à l'extraction de relations en domaine de spécialité. Nous avons développé un système de détection et typage de relations fondé sur une classification automatique, qui obtient une F-mesure d'environ 0,7. L'apport d'informations syntaxiques à ce système a ensuite été évalué. Les informations syntaxiques très simples ajoutées sous forme vectorielle n'ont pas amélioré la classification ; en revanche, l'utilisation des structures syntaxiques avec les tree kernels améliore la détection des relations, les meilleurs résultats étant obtenus avec une combinaison de l'arbre complet, le sous-arbre minimal et les attributs de base (augmentation de 3% de la F-mesure).

Il serait intéressant de poursuivre cette étude en faisant une classification multi-classes, pour évaluer l'apport des informations syntaxiques pour le typage des relations, ce qui nécessiterait d'augmenter le corpus d'entraînement, ou d'étendre l'étude à d'autres corpus de spécialité.

Remerciements

Ce travail a été partiellement financé par OSEO dans le cadre du programme Quæro.

Références

CHANG C.-C. & LIN C.-J. (2001). *LIBSVM : a library for support vector machines*. Software available at http ://www.csie.ntu.edu.tw/∼cjlin/libsvm.

CODEN A. R., PAKHOMOV S. V., ANDO R. K., DUFFY P. H. & CHUTE C. G. (2005). Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics*, **38**, 422–430.

CULOTTA A. & SORENSEN J. (2004). Dependency tree kernels for relation extraction. In *in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*.

HSU C.-W., CHANG C.-C. & LIN C.-J. (2003). *A Practical Guide to Support Vector Classification*. Rapport interne, Department of Computer Science, National Taiwan University.

MCCLOSKY D. (2010). Any domain parsing : Automatic domain adaptation for natural language parsing. *PHD Thesis, Department of Computer Science, Brown University*.

MINARD A.-L., LIGOZAT A.-L., BEN ABACHA A., BERNHARD D., CARTONI B., DELÉGER L., GRAU B., ROSSET S., ZWEIGENBAUM P. & GROUIN C. (2011a). Hybrid methods for improving information access in clinical documents : Concept, assertion, and relation identification. *Journal of the American Medical Informatics Association*. À paraître.

MINARD A.-L., LIGOZAT A.-L. & GRAU B. (2011b). Extraction de relations dans des comptes rendus hospitaliers. In *Actes des 22èmes Journées francophones d'Ingénierie des Connaissances (IC'2011)*.

MOSCHITTI A. (2006). Making tree kernels practical for natural language learning. In *In Proceedings of the Eleventh International Conference on European Association for Computational Linguistics (EACL), Trento, Italy, 2006*.

ROBERTS A., GAIZAUSKAS R. & HEPPLE M. (2008). Extracting clinical relationships from patient narratives. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, p. 10–18.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, p. 44–49.

UZUNER O., MAILOA J., RYAN R. & SIBANDA T. (2010). Semantic relations for problem-oriented medical records. *Artificial Intelligence in Medicine*, **50**, 63–73.

ZELENKO D., AONE C. & RICHARDELLA A. (2003). Kernel methods for relation extraction. *Journal of Machine Learning Research*, **3**, 1083–1106.

ZHANG M., ZHANG J. & SU J. (2006). Exploring syntactic features for relation extraction using a convolution tree kernel. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, p. 288–295.

ZHOU G., SU J., ZHANG J. & ZHANG M. (2005). Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the ACL*, p. 427–434.

TALN 2011, Montpellier, 27 *juin* –1er *juillet 2011*

Enrichissement de structures en dépendances par réécriture de graphes

Guillaume Bonfante, Bruno Guillaume, Mathieu Morey, Guy Perrier
INRIA Nancy-Grand Est - LORIA - Nancy-Université

Résumé. Nous montrons comment enrichir une annotation en dépendances syntaxiques au format du *French Treebank de Paris 7* en utilisant la réécriture de graphes, en vue du calcul de sa représentation sémantique. Le système de réécriture est composé de règles grammaticales et lexicales structurées en modules. Les règles lexicales utilisent une information de contrôle extraite du lexique des verbes français *Dicovalence*.

Abstract. We show how to enrich a syntactic dependency annotation of the *French Paris 7 Treebank* format, using graph rewriting, in order to compute its semantic representation. The rewriting system is composed of grammatical and lexical rules structured in modules. The lexical rules use a control information extracted from *Dicovalence*, a lexicon of French verbs.

Mots-clés : dépendance, French Treebank, réécriture de graphes, Dicovalence.

Keywords: dependency, French Treebank, graph rewriting, Dicovalence.

Introduction

Cet article propose une méthode d'enrichissement des structures en dépendances syntaxiques de surface et il applique cette méthode au *French Treebank de Paris 7* (par la suite noté FTB). Il entre dans la ligne de recherche ouverte par Bonfante *et al.* (2010) où nous montrions comment calculer — au moyen de la réécriture de graphes — la sémantique d'une phrase à partir de sa structure en dépendances syntaxiques. De manière plus générale, notre approche s'inscrit dans le contexte des méthodes exactes et symboliques de calcul en TAL.

Les représentations de la syntaxe en dépendances connaissent une popularité croissante pour l'évaluation et la comparaison d'analyses syntaxiques. Les raisons principales en sont données par Kahane (2001) : les dépendances syntaxiques sont lexicalisées et proches de la sémantique. Il existe très peu de corpus annotés en dépendances pour le français ; mais, récemment, Candito *et al.* (2009) ont montré comment produire une annotation en dépendances de surface du FTB à partir de son annotation en constituants (Abeillé *et al.*, 2003). Dans cet article, nous utilisons ce corpus pour tester notre système.

Dans Bonfante *et al.* (2010), nous avons proposé le principe de la réécriture de graphes pour calculer la sémantique à partir de la syntaxe. Nos entrées étaient des analyses syntaxiques profondes à la manière des structures tectogrammaticales du *Prague Dependency TreeBank* [1] (Hajič *et al.*, 2000). Dans notre cas, il s'agissait de structures enrichies du format PASSAGE [2]. Dans Bonfante *et al.* (2011), nous avons montré que l'on pouvait employer en fait le format FTB dès lors que certaines dépendances syntaxiques profondes étaient ajoutées : les arguments lexicalement ou grammaticalement déterminés des infinitifs et les antécédents des pronoms relatifs et réfléchis et

1. http://ufal.mff.cuni.cz/pdt2.0/
2. http://atoll.inria.fr/passage/

des sujets répétés. Dans les deux cas, les expérimentations demandaient une phase manuelle de préparation des données pour ajouter les annotations manquantes. Par ailleurs, nos systèmes surgénéraient car ils n'intégraient pas d'informations lexicales.

Nous pallions ici ces défauts en proposant un enrichissement automatique du FTB par des règles purement grammaticales et par d'autres qui utilisent de l'information lexicale extraite de Dicovalence (Van den Eynde & Mertens, 2003). Ces règles, en combinaison avec celles de l'article Bonfante *et al.* (2011), permettent donc d'*obtenir automatiquement une structure sémantique* [3] *à partir d'une annotation en dépendances provenant du* FTB.

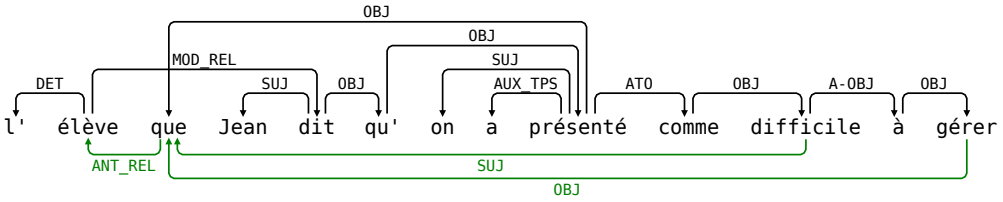
Les problèmes — échec de la réécriture, structures produites malformées — que nous avons rencontrés au cours de nos expérimentations, relèvent souvent d'erreurs d'annotations ou d'incohérences dans les structures du FTB. Notre système peut donc servir, par effet de bord, d'outil de fouilles d'erreurs et d'aide à la correction de corpus annotés.

Dans la section 1, nous posons le problème à travers un exemple introductif et nous présentons brièvement le modèle de réécriture de graphes choisi pour enrichir les annotations en dépendances du FTB. Puis dans la section 2, nous décrivons le système de règles de réécriture de graphes qui a été utilisé pour le faire et enfin dans la section 3 nous présentons la validation expérimentale qui a été effectuée sur le FTB.

1 Position du problème

De façon générale, pour calculer la sémantique, il est utile d'ajouter à l'annotation en syntaxe de surface deux types d'informations : les antécédents d'anaphores syntaxiques et certains actants syntaxiques. Dans la démarche proposée ici, la structure en dépendances de surface est celle du FTB. Nous nous imposons comme contrainte de ne pas modifier l'entrée mais seulement d'enrichir la structure à l'aide de nouvelles relations. Ce choix permet de rester complètement compatible avec d'autres travaux ou outils basés sur le format du FTB et de leur servir de guide. Il évite également de définir un nouveau format ad-hoc.

L'exemple donné ci-dessous est annoté par, en noir, les dépendances syntaxiques du guide d'annotation utilisé par (Candito *et al.*, 2009) [4] et, en vert, les dépendances à ajouter.



Pour les antécédents d'anaphores syntaxiques, nous nous en tenons à des phénomènes bien délimités : les antécédents des pronoms relatifs, des pronoms personnels réfléchis et des pronoms personnels répétitions de sujets. Retrouver l'antécédent d'un pronom relatif peut s'avérer complexe. Sur notre exemple, il faut remonter une chaîne de trois dépendances OBJ qui va du pronom relatif *« que »* jusqu'à la tête de la relative *« dit »* (en passant par *« présenté »* et *« qu' »*). De là, on retrouve l'antécédent *« élève »* en remontant la dépendance MOD_REL. Plus

3. une structure DMRS (Copestake, 2009)
4. http://www.linguist.univ-paris-diderot.fr/~mcandito/Rech/FTBDeps

généralement, la longueur des chaînes de dépendances n'est pas bornée. En outre, la chaîne peut contenir plusieurs sortes de relations et les structures de traits des nœuds rencontrés suivent certaines contraintes. Pour ces deux raisons, le calcul n'est pas immédiat.

Les actants syntaxiques à ajouter sont essentiellement les sujets des infinitifs s'ils sont présents dans la phrase. Nous mettons également dans cette catégorie les objets directs d'infinitifs comme dans la construction du *tough movement*. Ainsi dans l'expression *« un livre difficile à lire »*, l'objet de *« lire »* est *« livre »*. Manquent également dans le FTB les sujets des participes présents et des participes passés dans leur utilisation adjectivale. De façon plus générale, il paraît utile de traiter de façon uniforme les syntagmes adjectivaux et de leur associer systématiquement un sujet, que leur tête soit un participe ou un adjectif.

Comme les actants syntaxiques sont mutuellement dépendants, le calcul des relations SUJ et OBJ doit en tenir compte. Dans notre exemple, nous procédons dans l'ordre suivant. Nous commençons par marquer la dépendance SUJ entre l'adjectif *« difficile »* et son sujet *« que »*. En effet, *« que »* est l'objet direct de *« présenté »* et *« difficile »* en est l'attribut de l'objet. Partant de celle nouvelle relation, nous pouvons marquer la relation OBJ du verbe *« gérer »* vers le sujet de *« difficile »*. Cette annotation découle d'une propriété lexicale de l'adjectif *« difficile »* : son aptitude au *tough movement*.

Pour enrichir les structures de dépendances, nous employons le formalisme (β-calcul) que nous avons introduit dans Bonfante *et al.* (2010, 2011). Dans le cadre restreint de cet article, nous n'utilisons pas la réécriture de graphes de façon essentielle ; nos règle repèrent, dans des graphes, des motifs qui sont des arbres. Les seules transformations consistant à ajouter des arêtes, on pourrait certainement exprimer ces règles avec un formalisme moins puissant mais l'utilisation du β-calcul permet l'homogénéité de traitement avec les modules proposés dans les autres articles. Dans ce formalisme, un calcul procède par transformations successives des structures de dépendances jusqu'à leur normalisation. Le système est décrit par un ensemble fini de règles, chacune d'entre elles étant donnée par un motif et une liste de commandes élémentaires (ajout d'arête, suppression de nœud, etc). Une étape de calcul consiste à reconnaître le motif d'une règle dans le graphe courant et à modifier le graphe selon les commandes de la règle. La figure 1, en fin d'article, montre le déroulement de la réécriture sur la phrase *« Je trouve ce livre difficile à lire »*. L'encadré en haut de la figure présente les deux règles `ATTR-OBJ` et `TOUGH-MOVEMENT` utilisées dans cet exemple. Pour chaque règle, le schéma représente le *motif* reconnu, en rose et rouge. Au-dessous, est notée la liste de commandes effectuant la réécriture. Par exemple, pour la règle `ATTR-OBJ` , la liste se réduit à la commande *add_edge A –[SUJ]–> O* qui ajoute une relation SUJ de l'attribut de l'objet à l'objet du verbe.

Pour avoir un contrôle global sur le calcul, nous avons montré dans (Bonfante *et al.*, 2011) qu'une organisation modulaire des règles simplifie la tâche de développement d'un système de règles. De fait, des considérations linguistiques justifient à la fois la définition des modules et leur ordre d'application.

2 Règles utilisées pour l'enrichissement

L'enrichissement de l'annotation de surface se fait au moyen de quatre types de règles :
- des *règles grammaticales d'actants* déterminent de façon purement grammaticale certains sujets d'infinitifs, de participes et d'adjectifs,
- des *règles lexicales d'actants*, extraites d'un lexique, déterminent les sujets ou objets d'infinitifs compléments de verbes ou d'adjectifs à contrôle,
- des *règles d'antécédents* déterminent les antécédents des anaphores syntaxiques,
- des *règles de coordination* ajoutent les actants syntaxiques manquants aux conjoints.

2.1 Règles grammaticales d'actants

2.1.1 Sujets des syntagmes adjectivaux

Les participes têtes de syntagmes adjectivaux n'ont pas de sujets explicites dans le FTB car celui-ci ne contient que des arbres de dépendances. Le sujet d'un participe a déjà une autre fonction syntaxique ; le désigner reviendrait donc à lui attribuer un deuxième gouverneur. Voyons quelques exemples (les participes sont mis en gras) :

Jean arrive en **chantant**.
Abandonnée *de tous, elle ne sait plus que faire.*
Pierre sait Marie **délaissée** *par son mari.*
Un homme se **présentant** *comme son frère vient d'arriver.*

La configuration grammaticale dans laquelle se trouvent ces participes permet de déterminer leur sujet.

Nous étendons d'ailleurs la notion de sujet aux syntagmes adjectivaux qui ont un adjectif comme tête, pour deux raisons. Premièrement, au niveau sémantique, les adjectifs vont se traduire (comme les verbes) par des prédicats dont il faudra déterminer les arguments. Deuxièmement, les adjectifs peuvent se trouver dans des constructions variées : épithète, attribut du sujet, attribut de l'objet, dislocation. Toutes ces constructions sont compatibles avec le *tough movement*, comme le montrent les exemples suivants :

Je connais un livre **difficile** *à lire.*
Ce livre passe pour **difficile** *à lire.*
Je trouve ce livre **difficile** *à lire.*
Difficile *à lire, ce livre n'est pas à conseiller à tout le monde.*

Dans ces phrases, la relation entre *« livre »* et *« difficile »* implique que *« livre »* est objet de *« lire »*. Attribuer un sujet aux adjectifs permet de traiter les phénomènes comme le *tough movement* (Rezac, 2006) avec une seule règle, qui est indépendante de la configuration dans laquelle ces adjectifs apparaissent. Cela conduit d'ailleurs à une approche uniforme des adjectifs et des verbes, ce qui simplifie la définition des règles de calcul du sujet et des autres compléments.

Il y a autant de règles que de constructions intégrant des syntagmes adjectivaux. Ici, nous en considérons sept : épithète, attribut du sujet (avec ou sans préposition), attribut de l'objet (avec ou sans préposition), dislocation, gérondif. La règle `ATTR-OBJ` de la figure 1 (en fin d'article) s'applique aux attributs de l'objet sans préposition.

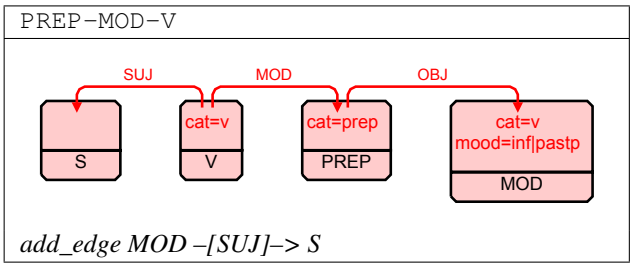
2.1.2 Sujets des infinitifs dans certaines constructions grammaticales

Comme les sujets des participes, les sujets des infinitifs ne sont pas exprimés dans le FTB ; les ajouter aboutirait également à violer la contrainte d'arbre des structures en dépendances. Certaines constructions syntaxiques associées à des mots grammaticaux contiennent pourtant des infinitifs dont le sujet n'est pas ambigu. Considérons les exemples suivants où les infinitifs concernés sont en gras :

Jean vient à Paris pour **travailler**.
Jean ne vient jamais à Paris sans **visiter** *la Tour Eiffel.*
Jean est trop poli pour ne pas **saluer** *Marie.*

Dans les deux premiers exemples, les prépositions *« pour »* et *« sans »* introduisent des compléments circonstanciels qui sont des infinitifs. Dans le dernier exemple, le couple *« trop . . . pour »* se construit avec un premier élément qui peut être un adjectif (ici *« poli »*), un adverbe ou un verbe et un second élément qui est toujours un infinitif (*« saluer »*).

Deux règles traitent ces différentes constructions. Dans le cas des infinitifs compléments circonstanciels, la règle `PREP-MOD-V` ci-contre fait apparaître que le sujet de l'infinitif est le sujet de la proposition principale modifiée [5]. Dans le cas de la construction *« trop + ADJ + pour + VINF »*, une autre règle fait apparaître le sujet de l'infinitif comme le sujet de l'adjectif. Là encore, attribuer un sujet aux adjectifs permet d'appliquer une même règle dans différentes configurations syntaxiques.



2.2 Règles lexicales d'actants

2.2.1 Le cas des verbes à contrôle ou à montée

Lorsqu'un infinitif est complément d'un verbe à contrôle ou à montée, son sujet, s'il est exprimé dans la phrase, est déterminé par ce verbe. Voyons quelques exemples où les infinitifs sont mis en gras et leur sujet profond est souligné.

Jean semble **changer** *d'avis.*
Jean permet à Marie de **venir**.
Jean promet à Marie de **venir**.
Jean propose à Marie de **venir**.

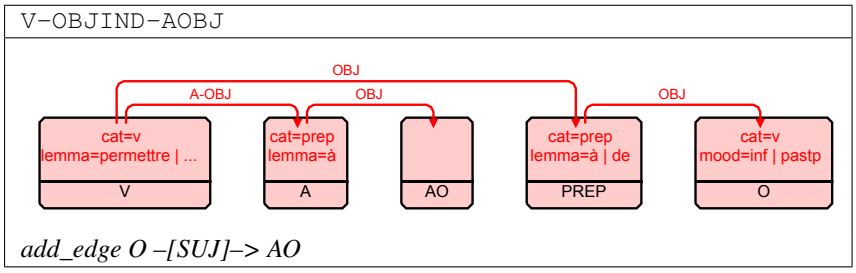
Dans la première phrase, *« sembler »* est un verbe à montée, l'infinitif qui le suit partage son sujet. Dans les deux suivantes, les verbes *« permettre »* et *« promettre »* déterminent le sujet de *« venir »* : *« Marie »* pour la première et *« Jean »* pour la deuxième. La dernière phrase est ambiguë, le sujet de *« venir »* peut être *« Jean »* ou *« Marie »*. Ce dernier exemple montre que le système que l'on a à décrire est nécessairement non confluent.

Les verbes à contrôle peuvent être groupés en différentes classes selon la fonction de l'infinitif contrôlé et selon la fonction du syntagme sujet de cet infinitif. Il y a trois possibilités pour chacun de ces deux facteurs : sujet, objet direct et objet indirect, donc six classes possibles (les deux fonctions syntaxiques sont nécessairement différentes). Pour chaque classe, il y a plusieurs règles, afin de gérer l'éventuelle présence d'un complémenteur *« à »* ou *« de »* introduisant l'infinitif.

Les verbes à contrôle possèdent, dans le lexique des verbes du français *Dicovalence* (Van den Eynde & Mertens, 2003), un ou plusieurs champs *PIVOT* qui contiennent leurs informations de contrôle. Ces informations sont extraites automatiquement de *Dicovalence* pour ancrer lexicalement les règles des verbes à contrôle, comme *« permettre »* dans l'exemple suivant :

```
VAL$    permettre: P0 P1 (P2)
NUM$    60700
FRAME$  subj:pron|n:[hum], obj:pron|n|compl|de_inf:[abs,mood:subj], ?objà:pron|n:[hum]
PIVOT$  P2/P0 [below de_inf in P1]
```

5. Il y a quelques exceptions à cette règle qui ne sont pas traitées : les constructions impersonnelles et certaines expression figées.



`P2/PO [below de_inf in P1]` se lit ainsi : l'objet indirect de *« permettre »* introduit par *« à »*, noté P2 dans *Dicovalence*, est le sujet, noté P0, d'une infinitive introduite par *« de »* (de_inf) qui est l'objet direct , noté P1, de *« permettre »*. Cela se traduit dans la règle `V-OBJIND-AOBJ` par la commande *add_edge O –[SUJ]–> AO* qui ajoute une relation SUJ de l'objet direct O vers le groupe nominal objet indirect AO.

2.2.2 Le cas des adjectifs à contrôle

Lorsqu'un infinitif est complément d'un adjectif à contrôle, son sujet ou son objet — quand il est présent dans la phrase — est déterminé lexicalement. Le *tough movement* est l'une de ces configurations. Voici quelques exemples où les infinitifs sont en gras :

Jean est lent à **comprendre**.
Le livre est difficile à **comprendre**.

Comme pour les verbes, c'est l'information lexicale sur les adjectifs *« lent »* et *« difficile »* qui détermine que le sujet du premier est le sujet de *« comprendre »*, alors que le sujet du second est l'objet direct de *« comprendre »*.

Les adjectifs à contrôle forment deux classes, selon que leur sujet est le sujet ou l'objet direct de l'infinitif contrôlé. Ces deux cas se traduisent chacun par une règle ; la règle `TOUGH-MOVEMENT` de la figure 1 correspond au cas où c'est l'objet de l'infinitif qui est contrôlé.

Il n'existe pas, pour les adjectifs, l'équivalent de Dicovalence pour les verbes. Nous avons donc relevé les adjectifs du corpus qui présentaient un argument infinitif et nous les avons classés manuellement pour ancrer les règles.

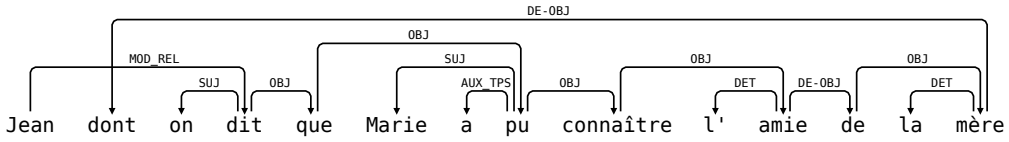
2.3 Les antécédents d'anaphores syntaxiques

Connaître les liens anaphoriques qui sont totalement déterminés par la syntaxe permet de calculer certaines co-références sémantiques. Dans ce travail, nous en distinguons trois types.

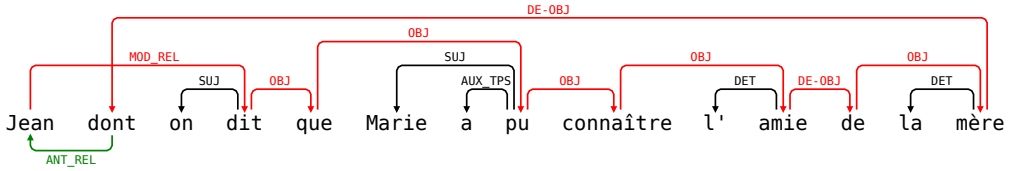
Le premier est le pronom réfléchi complément réel d'un verbe : son antécédent est le sujet de ce verbe. Par exemple, dans *« Jean ne veut pas* **se** *laver »*, l'objet direct de *« laver »* est *« se »*, qui co-réfère sémantiquement avec *« Jean »*. Une règle permet de détecter ce cas et une relation notée ANT_REFL entre le pronom et son antécédent est ajoutée.

Le second est le pronom personnel répétition d'un sujet : son antécédent est l'autre sujet du verbe. Dans *« Jean veut-***il** *manger ? »*, *« veut »* a deux sujets, dont *« il »* qui co-réfère sémantiquement avec *« Jean »*. Une relation ANT_REP est alors ajoutée.

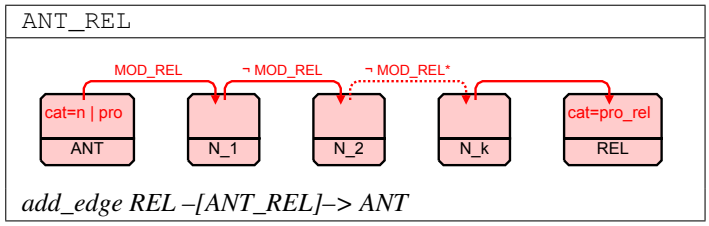
Le troisième type de lien anaphorique que nous considérons relie un pronom relatif à son antécédent. La création de ce lien, noté ANT_REL, est un problème plus complexe. Dans le FTB, il faut pour cela remonter les dépendances depuis le pronom relatif jusqu'à la tête de la relative, laquelle est reliée à l'antécédent par une dépendance MOD_REL. Considérons l'expression suivante annotée selon le guide du FTB :

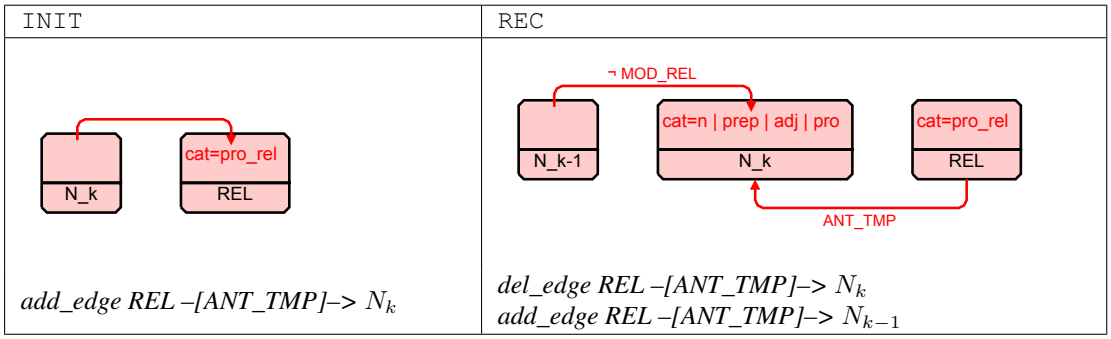


Dans cet exemple, on passe du pronom relatif à son antécédent en remontant successivement les dépendances DE-OBJ, OBJ, DE-OBJ, OBJ, OBJ, OBJ, OBJ, MOD_REL (en rouges dans la figure ci-dessous). Finalement, la nouvelle dépendance ANT_REL (en vert et en bas sur la figure), est produite.

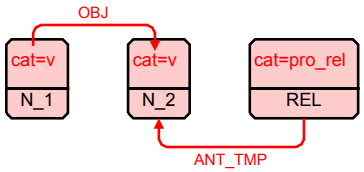
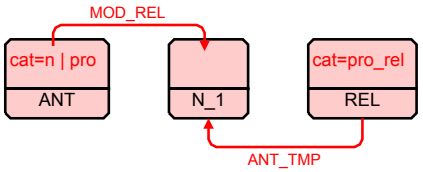


La transformation peut être décrite à l'aide du motif généralisé ci-contre. Toutefois, d'une part, ce type de motif ne permet pas d'exprimer les contraintes (comme les contraintes d'îlots) sur les structures de trait des noeuds. D'autre part, un tel motif n'est plus local, or cette propriété de localité du calcul permet une implantation efficace. La solution proposée ici emploie un lien temporaire ANT_TMP entre le pronom relatif et le point de la chaîne où on se situe dans sa remontée. Les quatre règles ci-dessous implémentent le motif généralisé en prenant en compte les contraintes d'îlots [6].





6. Des motifs négatifs (non représentés dans les figures) permettent de bloquer l'application d'une règle : par exemple, la règle INIT ne doit s'appliquer que sur un pronom relatif où il n'y a pas encore de lien ANT de créé.

RECV	STOP
	
del_edge REL –[ANT_TMP]–> N_2 *add_edge REL –[ANT_TMP]–> N_1*	*del_edge REL –[ANT_TMP]–> N_1* *add_edge REL –[ANT_REL]–> ANT*

Les relatives qui sont en apposition ou les constructions clivées peuvent être repérées par le même mécanisme de règle généralisée décomposée en règles atomiques (en changeant seulement la dernière règle). Les exemples suivants sont issus du corpus :

C'est une novice inspirée qui redressa [...]
Une mutation est en marche, qui ne s'arrêtera sans doute pas [...]

2.4 La coordination

La coordination est annotée dans le FTB de façon à respecter la contrainte d'arbre : un lien COORD va de la tête du premier conjoint jusqu'à la conjonction de coordination et un autre lien DEP_COORD va de la conjonction jusqu'à la tête du second conjoint. Si la tête du second conjoint est un verbe ou un adjectif, il est nécessaire de déterminer ses actants syntaxiques pour calculer sa représentation sémantique ; or les actants syntaxiques du second conjoint ne sont pas annotés dans le FTB quand ils sont partagés avec le premier conjoint. Considérons quelques exemples de coordination de syntagmes verbaux ou adjectivaux, où la tête du second conjoint est en gras et les actants partagés soulignés :

Jean déballe et **mange** son casse-croûte.
Abandonnée de son mari et **dégoûtée**, Marie ne sort plus.
Jean pense partir aujourd'hui et **pouvoir** rentrer dans un mois.

Dans le cadre de cet article, nous nous limitons à la recherche du sujet du second conjoint. Pour cela, on crée une dépendance SUJ entre le verbe tête du second conjoint et le sujet du premier conjoint, à condition que le second conjoint n'ait pas déjà un sujet. Deux règles différentes créent cette dépendance, selon que les syntagmes verbaux coordonnés sont introduits ou non par une préposition ou un complémenteur.

3 Mise en œuvre

3.1 Organisation en modules

Nous avons montré dans (Bonfante *et al.*, 2011) qu'une organisation modulaire permettait d'avoir un contrôle global sur le calcul et de simplifier le développement d'un système de règles. Les modules sont déterminés lin-

guistiquement : un module est un ensemble de règles contribuant à une transformation linguistique particulière qu'il est possible d'isoler relativement aux autres règles.

Cette organisation en modules permet également de gérer le comportement global du système de réécriture. Dans le cas présent, la terminaison est vérifiée pour chaque module, elle l'est donc globalement. Comme on l'a vu dans le cas des verbes à contrôle, notre enrichissement est nécessairement non-confluent mais cette non-confluence est très localisée (lexicalement) et donc contrôlable.

Par rapport à notre objectif général de construire une représentation sémantique à partir d'une représentation syntaxique en dépendances, les modules jouent un rôle essentiel. Par rapport à la tâche limitée qui est celle présentée dans cet article et qui consiste à enrichir les dépendances du FTB, ils ont une portée plus restreinte. Les interactions possibles entre les règles permettent difficilement de les isoler en modules.

Ainsi, les règles grammaticales d'actants et les règles lexicales d'actants peuvent se combiner de plusieurs façons. Prenons deux exemples. Dans la phrase *« Jean boit pour* **essayer** *d'***oublier** *»*, une règle grammaticale établit que *« Jean »* est sujet d'*« essayer »*, puis une règle lexicale ajoute que *« Jean »* est aussi sujet d'*« oublier »*. Dans la phrase *« Marie interdit à Jean de* **boire** *pour* **oublier** *»*, une règle lexicale établit *« Jean »* comme sujet de *« boire »*, puis une règle grammaticale établit que *« Jean »* est aussi sujet d'*« oublier »*.

Séparer les règles lexicales et grammaticales dans des modules différents ne permettrait pas de capturer toutes leurs interactions. C'est la raison pour laquelle nous répartissons les règles, au nombre de 47, présentées dans les sous-sections précédentes en trois modules ordonnés comme suit :
– le premier module ajoute les relations ANT des pronoms réfléchis, répétitions de sujets et relatifs vers leurs antécédents. Il contient 8 règles : 6 pour les relatives qui décomposent le motif généralisé, 1 pour les pronoms réfléchis et 1 pour les pronoms personnels répétitions de sujets ;
– le second module, SUJET, ajoute les relations SUJ des adjectifs, participes et infinitifs, que la règle soit grammaticale ou lexicale. Il contient 36 règles :11 règles grammaticales et 25 règles lexicales ;
– le troisième module, OBJET, ajoute les relations OBJ des infinitifs contrôlés dans des constructions de *tough movement*. Il contient 3 règles.

Le fait d'ajouter des sujets aux adjectifs permet de gérer le *tough movement* avec une seule règle. En revanche, il est alors nécessaire d'appliquer le module SUJET avant le module OBJET.

Enfin, la coordination intervient dans tous les modules et nécessite un traitement particulier. Reprenons un exemple déjà introduit pour la coordination :

Jean pense partir aujourd'hui et **pouvoir** rentrer dans un mois.

Dans cet exemple, il faut déterminer le sujet de *« partir »* avant d'en déduire que *« pouvoir »* partage ce sujet. Cela incite donc à appliquer les règles de coordination après le module SUJET. Cependant, pour déterminer le sujet de *« rentrer »*, il est nécessaire de connaître le sujet de *« pouvoir »*. De ce point de vue, il faudrait donc appliquer les règles relatives à la coordination avant le module SUJET.

Un moyen de résoudre cette contradiction est de particulariser chaque règle de coordination en autant de règles qu'il existe de modules concernés.

3.2 Expériences

Nous avons appliqué notre système aux 12 351 phrases annotées en dépendances que contient le FTB. Un extrait de 120 phrases est disponible en ligne [7].

7. http://wikilligramme.loria.fr/doku.php?id=taln2011

Les relations ANT, SUJ et OBJ que nous ajoutons sont très courantes dans le corpus : on en observe sur 87% des phrases. En outre, on ajoute, sur chaque phrase, en moyenne, trois nouvelles relations. On peut remarquer le faible nombre de relations OBJ ajoutées qui montrent que le *tough movement* est assez rare en corpus. Pour avoir une évaluation plus fine, notamment de la précision, il faudrait disposer d'une partie du corpus annotée manuellement qui n'existe malheureusement pas pour l'instant.

Au total, nous ajoutons :
– 3 691 relations ANT (3 152 pour les pronoms relatifs, 99 pour les pronoms réfléchis et 270 pour les pronoms personnels répétitions de sujets) ;
– 33 605 relations SUJ (23 940 pour les adjectifs et 9 665 pour les verbes) ;
– 19 relations OBJ.

Pour avoir une idée précise de la pertinence de ce système, il convient également d'essayer de détecter les cas problématiques, c'est-à-dire les cas où des relations sont susceptibles de manquer dans les structures produites par notre système de réécriture. Pour cela, nous avons observé les trois configurations suivantes :
– les verbes (sauf impératifs et auxiliaires) qui n'ont toujours pas de sujet (de surface ou profond) après réécriture, il y en a 3 548 ;
– les adjectifs sans sujet : 955
– les pronoms relatifs sans antécédent : 242

Pour chacune de ces configurations, nous avons étudié un échantillon de 100 phrases que nous avons classées manuellement. Le résultat de ce classement figure dans le tableau ci-dessous.

	Pas d'erreur	Annotation FTB	Problème de lexique	Problème de règles
verbes sans sujet	51	29	13	7
adjectifs sans sujet	32	45	1	22
pronoms relatifs sans antécédent	5	72	0	23

Les erreurs d'annotation du FTB que nous avons relevées sont systématiques et probablement liées à la conversion du FTB des constituants aux dépendances. L'annotation du FTB est donc la première source d'erreur dans notre processus. Les problèmes de lexique sont liés à des constructions nominales qui ne sont pas décrites dans Dicovalence. Nos règles ne couvrent pas certains phénomènes linguistiques : par exemple, un adjectif en emploi substantivé peut être antécédent d'une relative.

Discussion

(Gardent & Cerisara, 2010; Gardent, 2010) ont aussi utilisé la réécriture de graphes sur le FTB mais avec un objectif différent du nôtre, celui d'ajouter une annotation en rôles sémantiques pour les verbes. C'est pourquoi la plupart des règles qu'ils ont conçues vise à reformuler les constructions passives et causatives en constructions actives canoniques. Les quelques règles supplémentaires traitent de la coordination et des sujets des infinitifs. Les dernières, n'utilisant pas d'information lexicale, ont le défaut de systématiquement choisir pour sujet des infinitifs compléments celui du verbe dont ils dépendent. Enfin, l'étude n'aborde pas la question de la dépendance entre les règles et de leur ordre d'application.

Nous avons montré dans (Bonfante *et al.*, 2011) qu'il était possible de prendre compte les reformulations dans toute leur diversité (passif, moyen, causatif, impersonnel ...) à l'aide de la réécriture de graphes. Si nous avons écarté celles-ci du travail présenté dans cet article, c'est que nous nous étions fixé comme cadre pour ce travail d'enrichir les structures en dépendances du FTB sans modifier les relations existantes.

Intégrer les reformulations nécessiterait d'ajouter des règles correspondantes au système qui vient d'être présenté. Les interactions complexes entre ces règles nouvelles et les règles existantes font que cela ne peut pas se faire simplement par l'ajout d'un module avant ou après ceux qui ont été définis dans cet article. Par exemple, pour la phrase *« Jean est autorisé à* **partir** *»*, la reformulation du passif doit être effectuée avant la détermination du sujet profond de *« partir »*. C'est le contraire pour la phrase *« Jean demande à Marie d'être* **aidée** *de Pierre. »* En conséquence, l'intégration des reformulations devrait nous amener à restructurer le système de modules.

Nos expérimentations ont mis en évidence certaines limites du format d'annotation du FTB, qui sont dues essentiellement au choix des annotateurs de n'avoir que des arbres comme structures de dépendance. Premièrement, l'annotation des constructions causatives ne distingue pas le sujet et l'objet du verbe causé. Dans *« Jean fait* **manger** *un lapin »*, *« lapin »* est annoté OBJ de *« manger »*, que le lapin mange ou soit mangé. Deuxièmement, lorsque deux verbes sont coordonnés, l'annotation ne permet pas de retrouver les compléments partagés du second conjoint. Ainsi *« Jean* **déballe** *et* **mange** *son sandwich »* et *« Jean* **déballe** *son sandwich et* **mange** *»* ont la même annotation. Rien ne permet de distinguer que dans la première phrase, *« sandwich »* est également objet de *« déballe »* alors que dans la deuxième, *« sandwich »* n'est pas objet de *« mange »*.

Remerciements Nous remercions Sylvain Kahane pour une discussion qui a déclenché ce travail. Nous tenons également à remercier Anne Abeillé pour nous avoir autorisé à publier un extrait du FTB.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In *Treebanks. Building and Using Parsed Corpora*, chapter 10. Kluwer Academic Publishers.

BONFANTE G., GUILLAUME B., MOREY M. & PERRIER G. (2010). Réécriture de graphes de dépendances pour l'interface syntaxe-sémantique. In *TALN 2010*, Montréal, Canada.

BONFANTE G., GUILLAUME B., MOREY M. & PERRIER G. (2011). Modular graph rewriting to compute semantics. In *IWCS 2011*, p. 65–74, Oxford, UK.

CANDITO M.-H., CRABBÉ B., DENIS P. & GUÉRIN F. (2009). Analyse syntaxique statistique du français : des constituants aux dépendances. In *TALN 2009*, Senlis, France.

COPESTAKE A. (2009). *Invited Talk :* Slacker semantics : Why superficiality, dependency and avoidance of commitment can be the right way to go. In *EACL 2009*, Athens, Greece.

GARDENT C. (2010). Extraction des cadres syntaxiques à partir de P7dep. Notes transmises par l'auteur.

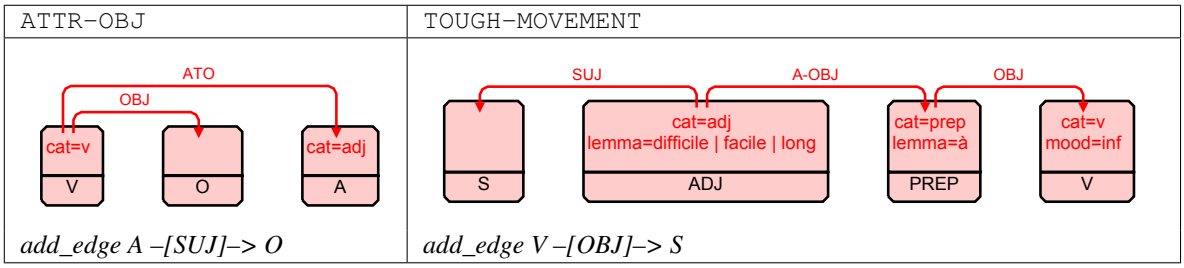
GARDENT C. & CERISARA C. (2010). Semi-Automatic Probanking for French. In *TLT9 – the ninth international workshop on Treebanks and Linguistic Theories, Tartu, Estonia*.

HAJIČ J., BÖHMOVÁ A., HAJIČOVÁ E. & HLADKÁ B. (2000). *The Prague Dependency Treebank : A Three-Level Annotation Scenario*, In *Treebanks : Building and Using Parsed Corpora*, p. 103–127. Amsterdam :Kluwer.

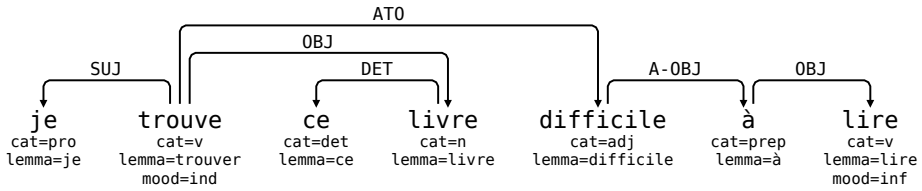
KAHANE S. (2001). Grammaires de dépendance formelles et théorie Sens-Texte. In *Tutoriel, TALN 2001*, volume 2, Tours.

REZAC M. (2006). *On tough-movement*, In *Minimalist Essays*, p. 288–325. Linguistik Aktuell/Linguistics Today 91. John Benjamins.

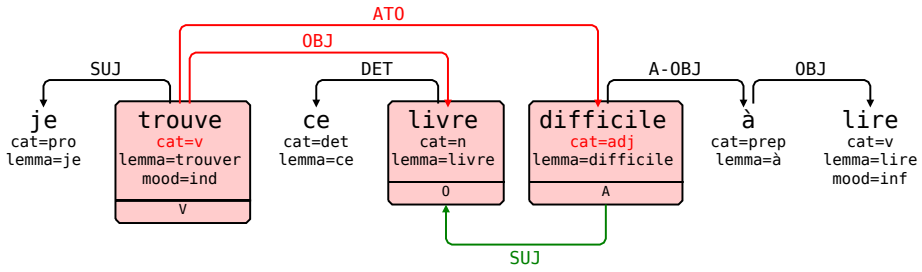
VAN DEN EYNDE K. & MERTENS P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *French Language Studies*, **13**, 63–104.



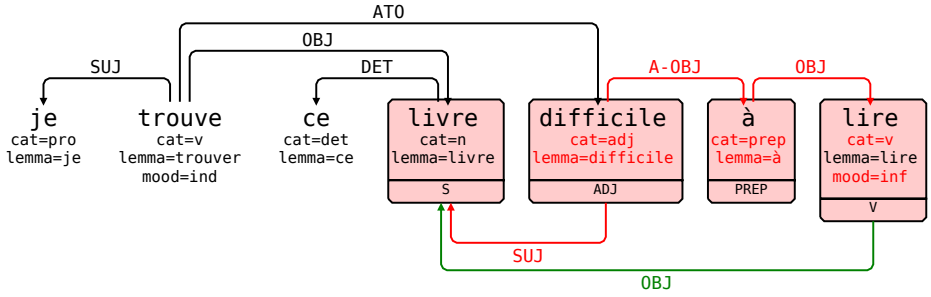
structure initiale



identification du motif de la règle ATTR-OBJ puis réécriture



identification du motif de la règle TOUGH-MOVEMENT puis réécriture



structure finale

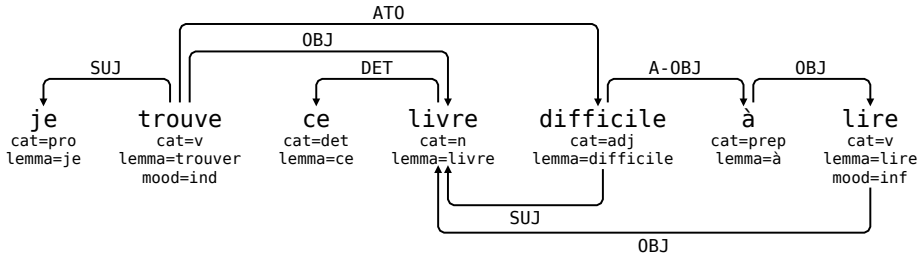


FIGURE 1 – *« Je trouve ce livre difficile à lire. »*

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Classification en polarité de sentiments avec une représentation textuelle à base de sous-graphes d'arbres de dépendances

Alexander Pak, Patrick Paroubek
alexpak@limsi.fr, pap@limsi.fr
Université de Paris-Sud, Laboratoire LIMSI-CNRS, Bâtiment 508,
F-91405 Orsay Cedex, France

Résumé. Les approches classiques à base de n-grammes en analyse supervisée de sentiments ne peuvent pas correctement identifier les expressions complexes de sentiments à cause de la perte d'information induite par l'approche « sac de mots » utilisée pour représenter les textes. Dans notre approche, nous avons recours à des sous-graphes extraits des graphes de dépendances syntaxiques comme traits pour la classification de sentiments. Nous représentons un texte par un vecteur composé de ces sous-graphes syntaxiques et nous employons un classifieurs SVM état-de-l'art pour identifier la polarité d'un texte. Nos évaluations expérimentales sur des critiques de jeux vidéo montrent que notre approche à base de sous-graphes est meilleure que les approches standard à modèles « sac de mots » et n-grammes. Dans cet article nous avons travaillé sur le français, mais notre approche peut facilement être adaptée à d'autres langues.

Abstract. A standard approach for supervised sentiment analysis with n-grams features cannot correctly identify complex sentiment expressions due to the loss of information incurred when representing texts with bag-of-words models. In our research, we propose to use subgraphs from sentence dependency parse trees as features for sentiment classification. We represent a text by a feature vector made from extracted subgraphs and use a state of the art SVM classifier to identify the polarity of a text. Our experimental evaluations on video game reviews show that using our dependency subgraph features outperforms standard bag-of-words and n-gram models. In this paper, we worked with French, however our approach can be easily adapted to other languages.

Mots-clés : analyse de sentiments, analyse syntaxique, arbre de dépendances, SVM.

Keywords: sentiment analysis, parsing, dependency tree, SVM.

1 Introduction

L'approche « sac de mots » est un des premiers modèles de représentation textuelle, qui est de nos jour encore souvent utilisé pour l'analyse de sentiments. Le texte y est représenté comme un ensemble de n-grammes sans prise en considération de leur ordre d'apparition dans le texte, ni des relations qui les relient au sein du texte. Des approches classiques en apprentissage automatique (Naive Bayes or SVM) utilisent ensuite cette représentation pour construire des systèmes de classification en sentiments des textes. L'exactitude [1] de ce genre d'approche peut être très élevée, tout particulièrement lorsque l'on utilise des techniques avancées de sélection de traits, en conjonction avec des lexiques additionnels extraits de textes identifiés au préalable comme porteur d'opinion. Cependant, nous sommes convaincus que des modèles capables d'identifier des expressions plus complexes de sentiments, allant au delà de la simple reconnaissance de construction comme « bon film » ou « jeu déplorable », doivent permettre d'obtenir de meilleurs systèmes de classification. Un des problèmes de l'approche sac de mots réside dans la perte d'information lors de la construction de la représentation des textes, vus comme des collections de termes dissociés. Or les relations qu'entretiennent les mots au sein du texte sont souvent très importantes dans la détermination précise du degré ou de la polarité d'un sentiment. Si nous considérons la phrase : « Ce film est **mauvais** », elle exprime de manière évidente un sentiment négatif et un système de classification standard à base d'unigrammes n'aura pas de mal à la classer comme négative, pourvu qu'il ait été suffisamment entraîné sur des données appropriées. Dans le cas d'un énoncé un peu plus complexe comme : « Ce film n'est **pas mauvais** », un modèle unigramme simple sera probablement mis en échec, mais un modèle utilisant des bigrammes sera lui capable de détecter l'occurrence de « pas mauvais » comme un terme à connotation positive. Considérons maintenant un exemple encore plus complexe comme : « Ce film est **étonnamment pas si mal** » et là les systèmes à base d'unigrammes et de bigrammes vont probablement se tromper. Dans cet exemple, il faudrait qu'ils soient associés à un traitement plus sophistiqué de la négation.

En plus d'être incapables de prendre en compte toutes les expressions de négation, les modèles n-grammes sont incapables de représenter les dépendances longue distance. Un modèle de bigrammes pourra identifier « J'ai apprécié » comme un motif à connotation positive dans l'énoncé « **J'ai apprécié** le film », mais pas dans « **J'ai** beaucoup **apprécié** le film ». Nous pensons qu'il faut recourir à d'autres modèles que le modèle sac de mots si nous voulons progresser dans l'identification automatique des sentiments en utilisant une classification plus fine, qui rende par exemple compte de l'intensité d'un sentiment en plus de sa polarité, car les modèles sac de mots ne nous fournissent pas assez d'information.

Pour aller au delà des modèles sac de mots, nous proposons d'utiliser les arbres de dépendances issus de l'analyse syntaxique des phrases pour générer des sous-graphes, qui serviront à représenter un texte. Un arbre de dépendances est une représentation graphique associée à une phrase, dans laquelle les nœuds correspondent aux mots de la phrase et les arcs représentent des relations syntaxiques entre les nœuds comme : objet, sujet, modifieur etc. La Figure 1 représente un arbre de dépendance syntaxique pour la phrase « Je n'aime pas beaucoup le poisson ». Une telle représentation des phrases est parfaitement adaptée à l'analyse de sentiment voire même à la fouille d'opinion car :

– À partir de l'arbre de dépendances, nous pouvons facilement identifier le sous-graphe contenant la négation « ne $\xrightarrow{NEGAT}$ aime ».
– Nous pouvons identifier les marqueurs d'intensité : « beaucoup $\xrightarrow{VMOD\_POSIT1}$ aime »
– De même pour la source d'une expression d'opinion : « Je $\xrightarrow{SUBJ}$ aime » et la cible d'une expression opinion : « aime $\xrightarrow{OBJ}$ poisson »

Comme pour les modèles à base de n-grammes, notre approche utilise un paramètre de taille pour calibrer les sous-graphes extraits des arbres de dépendance pour représenter un texte. Nous posons que la taille d'un sous-graphe est égale au nombre de ses arcs. Ainsi, un sous graphe de taille 1 contiendra un arc et deux nœuds, un sous graphe de taille 2 contiendra 2 arcs et 3 nœuds, etc. La Figure 2 contient la représentation de l'énoncé « J'aime bien le poisson » au moyen de sous-graphes de taille 2.

Dans la section suivante, nous expliquons en détails comment obtenir la représentation à base de sous-graphes d'un texte à partir des arbres de dépendances syntaxiques qui lui sont associés. Ensuite, nous montrons comment utiliser cette représentation pour indexer des critiques de jeux vidéo et pour entraîner un classifieur en polarité de sentiments à base de SVM. Nous présentons notre protocole d'évaluation et les résultats obtenus par notre modèle

1. accuracy

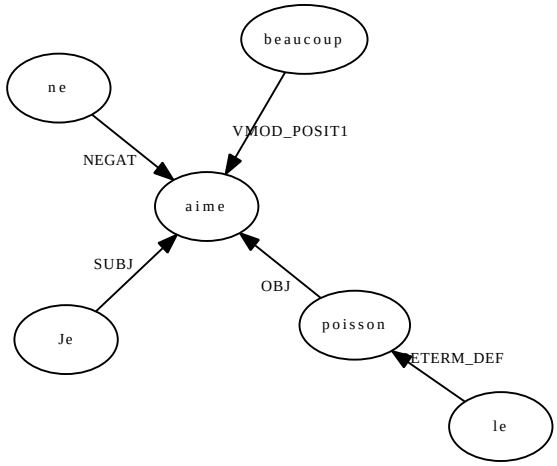


FIGURE 1 – Un arbre de dépendance syntaxique pour la phrase « Je n'aime pas beaucoup le poisson ». Les nœuds représentent des mots , les arcs des relations entre les mots. Le mot «pas» ne figure pas explicitement dans le diagramme, car il est encodé par la relation de négation.

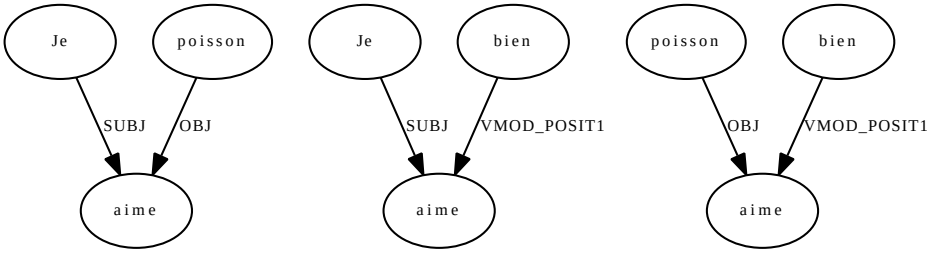


FIGURE 2 – Une représentation de la phrase « J'aime bien le poisson » avec des sous-graphes de taille 2. La relation « déterminant » contenant le nœud « le » a été écartée.

dans la Section 3, une présentation des recherches antérieures dans le domaine dans la Section 4 et la conclusion sur nos travaux en section 5.

2 Notre approche

2.1 Représentation à base de sous-graphes de dépendances

Nous utilisons la sortie en dépendances typées de l'analyseur syntaxique *Xerox Incremental Parser* (XIP) (Aït-Mokhtar *et al.*, 2002) pour construire l'arbre de dépendances de la phrase. La Table 1 contient un exemple de dépendances produites par XIP.

SUBJ(VERB :aime, PRON :Je)
OBJ(VERB :aime, NOUN :poisson)
VMOD_POSIT1(VERB :aime, ADV :beaucoup)
DETERM_DEF(NOUN :poisson, DET :le)
NEGAT(VERB :aime)

TABLE 1 – Les dépendances produites par XIP pour la phrase : « Je n'aime pas beaucoup le poisson »

Chaque ligne de la sortie de XIP contient une unique dépendance qui correspond à une description des relations grammaticales entre les mots de la phrase (de Marnee & Manning, 2008). Chaque dépendance peut être vue

comme un triplet <Type, Source, Cible>, où *Type* détermine la relation grammaticale (ex. sujet, objet, etc.) entre la *Source* et la *Cible*. La source et la cible sont représentés comme des mots associés à leur étiquette grammaticale. XIP produit aussi des relations unaires, que nous catégorisons en deux types distincts, avec pour chacun un traitement spécifique :

1. **Négations** (ex. `NEGAT(VERB:aime)`) Nous transformons une relation unaire en relation ternaire par l'ajout de la particule 'ne' comme cible. Ainsi, nous obtenons : `NEGAT(VERB:aime, NEG:ne)`

2. **Entités** XIP reconnaît et étiquette les entités telles que les noms de personnes, dates, temps, noms de lieux etc. Ces informations n'étant pas utiles pour la détection des sentiments, elles sont ignorées.

Nous écartons aussi la relation `SEQNP`, qui indique les énumérations dans les phrases ; ceci afin de réduire la taille de l'index, la suppression de cette relation n'ayant pas d'impact notoire sur nos résultats.

De l'ensemble de relations produit par XIP pour chaque énoncé, nous voulons obtenir un arbre dans lequel chaque nœud possède un sens complètement déterminé. Dans notre exemple, un nœuds comme « ne » n'a pas de sens intrinsèque et le nœud « aime » possède un sens partiel (il lui manque la prise en compte de la négation dans son interprétation). Par conséquent, nous avons besoin de fusionner certains nœud et de retirer certaines relations. Nous avons décidé de réduire le nombre de relations avec lesquelles travailler, car XIP produit plus de 90 types de relations (une liste complète est présentée en 5.).

2.1.1 Réduction du jeu de relations

Nous avons simplifié le jeu de relations de dépendances en ne considérant que les classes génériques en appliquant les règles d'assimilation de la Table 2.

NMOD_* -> NMOD	les modifieurs de nom (ante et post posés)
VMOD_* -> VMOD	les différents modifieurs de verbe
SUBJ_* -> SUBJ	les différents sujets
OBJ_* -> OBJ	les différents compléments d'objet directs
DEEPSUBJ* -> SUBJ	le sujet profond est assimilé au sujet de surface

TABLE 2 – Règles de simplification des relations de dépendances.

En outre, lors de la construction de l'arbre de dépendance, nous excluons certains arcs qui ne sont pas indispensables à notre analyse :
– Les déterminants, ainsi « le $\xrightarrow{DETERM\_DEF}$ film » devient « film », mais nous conservons les quantificateurs (`DETERM_NUM`, `DETERM_QUANT`, `DETERM_QUANT_DEF`, `DETERM_QUANT_DEM`).
– Les pronoms possessifs, ainsi « mon $\xrightarrow{DETERM\_POSS}$ livre » devient « livre ».
– Les relations modifieur de nom `NMOD`, lorsque la source et la cible sont tous deux des noms, ainsi « livre $\xrightarrow{NMOD\_POSIT1}$ cuisine » devient « livre ».
Au final, le jeu de relation de dépendances que nous considérons pour notre analyse est donné en Table 3.

ADJMOD	modifieur d'adjectif
ADVMOD	modifieur d'adverbe
DETERM_NUM	déterminant numérique
DETERM_QUANT	quantificateur
NMOD	modifieur de nom
OBJ	complément d'objet direct
SUBJ	sujet (de surface ou profond)
VMOD	modifieur de verbe

TABLE 3 – Jeu de relations de dépendances final.

2.1.2 Combinaison de nœuds

Nous utilisons les règles suivantes pour combiner les nœuds :
– nœuds liés par la relation de négation (NEGAT), ainsi l'arc « ne $\xrightarrow{NEGAT}$ aime » devient un nœuds simple « ne aime »
– verbes auxiliaires et principaux (AUXIL), ainsi l'arc « a $\xrightarrow{AUXIL}$ aimé » devient un nœuds simple « a aimé »
– verbes passifs, réfléchis et composés (AUXIL, AUXIL_PASSIVE, REFLEX, OBJ_SPRED, COORDITEMS_SC)

Un exemple d'arbre issu de l'application des règles précédentes est donné dans la figure Figure 3.

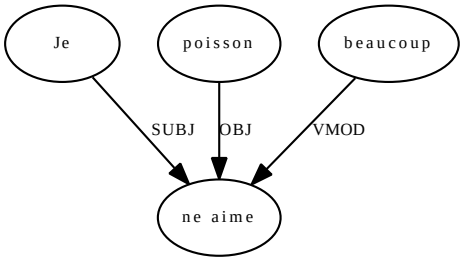


FIGURE 3 – Arbre de dépendance obtenu pour la phrase « Je n'aime pas beaucoup le poisson », après combinaison des nœuds et réduction des arcs.

Finalement, la phrase est représentée par un le jeu de tous les sous-graphes possibles pour une taille S, où S est égal au nombre d'arc des sous-graphes. Dans nos expériences, nous avons utilisé $S = 1, 2, 3$.

2.1.3 Nœud universel

La majorité des expressions de sentiment ont la même structure grammaticale. Par exemple, dans les expressions suivantes : « J'aime le poisson » et « J'aime le film » seul l'objet diffère tandis que le reste de la construction reste le même. Nous aimerions entraîner notre système à reconnaître ces expressions. Pour cela, nous avons ajouté un nœud universel, représentant la classe de tous les mots, dans les sous-graphes (Arora *et al.*, 2010).

Pour chaque sous-graphe obtenu à l'étape précédente, nous générons une permutation des sous-graphes contenants plusieurs nombres (de 0 à $S - 1$) de nœuds universels. Pour ce faire, nous remplaçons tout à tour chaque nœud d'un sous-graphe par un nœud universel, sauf pour les verbes, les adjectifs et les adverbes car ils peuvent exprimer des sentiments. Par ailleurs, nous interdisons d'avoir deux nœuds universels adjacents. Un exemple de l'emploi des nœuds universels avec la phrase « Je $\xrightarrow{SUBJ}$ aime $\xrightarrow{OBJ}$ poisson » est décrit dans la Figure 4.

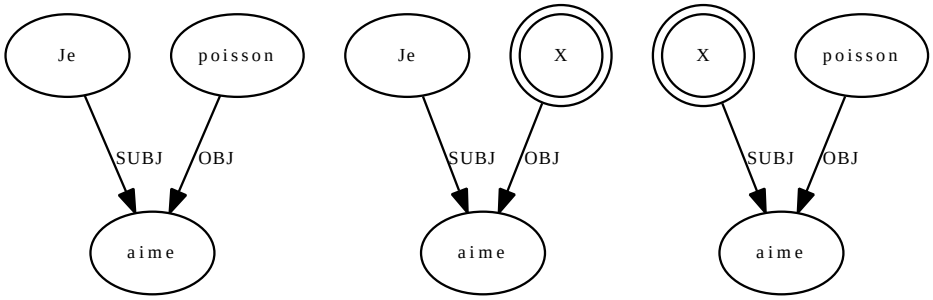


FIGURE 4 – Sous-graphes avec nœud universel (X) obtenus pour « J'aime le poisson »

2.2 Construction des vecteurs de traits

Nous représentons un texte donné T comme un vecteur de traits $T = \{w_1, w_2, \cdots, w_K\}$, où w_i est le poid d'un sous-graphe i dans le texte T et K est le nombre de sous-graphes dans T. Nous utilisons le schéma de pondération delta TFIDF lissé, car il a permis d'obtenir la meilleure performance dans des recherches antérieures (Paltoglou & Thelwall, 2010).

$$w_i = tf_i \cdot \Delta idf_i \tag{1}$$

$$\Delta idf_i = log \frac{N_1 \cdot df_2 + 0.5}{N_2 \cdot df_1 + 0.5} \tag{2}$$

où N_1 et N_2 représentent le nombre total de documents de classe 1 et 2, df_1 et df_2 sont des classes de fréquences du graphe i (c.a.d. le nombre de documents de classes 1 et 2 dans lesquelles le graphe apparaît). Dans notre cas, les classes 1 et 2 sont des documents positifs et négatifs.

3 Expériences et résultats

3.1 Les données

Nous utilisons des critiques de jeux vidéo issues du projet DOXA [2], dont le but est la construction d'une plateforme industrielle de fouille d'opinion.

Le corpus est constitué de critiques de jeux vidéo provenant de 8 sites dédiés [3]. Le corpus et ses annotations sont décrites dans (Paroubek *et al.*, 2010). Les annotations synthétisent les sentiments exprimés par les auteurs des critiques au niveau du document et du paragraphe (définis arbitrairement comme un empan de texte d'environ 100 mots). Un exemple d'annotation est fourni dans la table 4.

Attribut	Valeur
catégorie sémantique	une liste de 1 à 5 catégories d'opinion DOXA, ex. *« recommandation_suggestion »*
polarité	$-, \pm, +$, neutre
intensité	faible-moyen, fort
thème	la cible de l'expression d'opinion sélectionnée dans une taxonomie du domaine considéré (une liste de 1 à 5 concepts)
lien	lorsque plusieurs catégories sémantiques et plusieurs thèmes sont présents, le lien peut être fait entre certaines opinions s'ils sont plus particulièrement associés.
justification	référence au paragraphe/segment de texte qui représente au mieux l'opinion annotée

TABLE 4 – Annotation d'opinon DOXA au niveau document et paragraphe.

Dans les annotations DOXA, la polarité d'un sentiment est exprimée au moyen d'une échelle de six valeurs : *neutre, très-négatif, faible-moyen-négatif, mixte, faible-moyen-positif, fort-positif*. Nous avons sélectionné tous les documents ayant une polarité positive (*forst-positif* et *faible-moyen-positif*), ainsi que tous les documents avec une polarité négative (*fort-négatif* et *faible-moyen-négatif*) que nous avons répartis dans deux classes distinctes. Nous n'avons pas utilisé les documents annotés comme *neutre* (pas d'expression de sentiment) ni ceux annoté *mixte* (qui contiennent à la fois des expressions positives et négatives, résultant en une interprétation mitigée). Notre corpus contient donc 387 documents considérés à teneur positive et 250 à teneur négative. Nous avons ensuite divisé le sous corpus des documents positifs en deux parties : un corpus d'entraînement et un corpus de tests, en sélectionnant pour ce dernier, tous les documents qui ont été annotés par deux annotateurs. Le sous-corpus négatif a subit le même découpage. La Table 5 résume les caractéristiques de notre corpus.

2. https ://www.projet-doxa.fr/index.php

3. www.ecrans.fr, www.gamehope.com, www.gamepro.fr, www.jeuxactu.com, www.jeuxvideo.com, www.jeuxvideo.fr, www.play3-live.com

Classe	Entraînement	Tests
Positif	334	53
Negatif	197	35
Total	531	88

TABLE 5 – Nombre de documents par classe

3.2 Évaluation

Nous avons entraîné un classifieur SVM à base de n-grammes (Pang *et al.*, 2002) avec un schéma de pondération delta TFIDF (Paltoglou & Thelwall, 2010), que nous utilisons pour obtenir une mesure de performance de base. Les négations ont été traitées en attachant la particule de négation successivement au mot qui la précède et au mot qui la suit lors de la génération des n-grammes (Pak & Paroubek, 2010). Nous avons généré trois types de classieurs, respectivement à base de n-grammes, de bigrammes et de trigrammes.

De manières similaire, pour notre modèle à base de sous-graphes de dépendances, nous avons utilisé trois types de modèles, utilisant respectivement des sous-graphes de taille 1, 2 et 3.

Aussi bien pour le modèle à n-gramme que pour notre modèle à sous-graphes de dépendances, nous avons utilisé une implémentation libre de classifieur SVM issue de la librairie LIBLINEAR (Fan *et al.*, 2008), avec des valeurs de paramètre par défaut et un noyau linéaire. Le classifieur a d'abord été entraîné sur un jeu de 531 documents puis évalué sur un ensemble de 88 documents. L'exactitude moyenne et la précision moyenne (Manning & Schütze, 1999) ont été choisies comme mesures d'évaluation.

$$exactitude = \frac{vp + vn}{vp + vn + fp + fn} \tag{3}$$

$$precision = \frac{vp}{vp + fp} \tag{4}$$

où `vp` est le nombre de documents classés correctement comme positifs (*vrais positifs*), `vn` est le nombre de document classés correctement comme étant négatifs (*vrais négatifs*), `fp` est le nombre de document incorrectement identifiés comme positifs (*faux positifs*) et `fn` est le nombre de document incorrectement identifiés comme négatifs (*faux négatifs*).

3.3 Résultats

Les résultats de l'évaluation sont donnés dans la Table 6. Les mentions unigramme, bigramme et trigramme correspondent respectivement aux trois modèles de base n-gramme, tandis que les mentions subgraph-1, subgraph-2 et subgraph-3 correspondent à nos modèles à sous-graphes de dépendances, respectivement de taille 1, 2 et 3.

Modèle	Exactitude moy. (%)	Précision moy. (%)	Préc$_{pos}$ (%)	Préc$_{neg}$ (%)
unigramme	73.86	69.57	90.57	48.57
bigramme	72.73	69.11	86.79	51.43
trigramme	64.77	60.08	83.02	37.14
subgraph-1	**78.41**	**74.80**	**92.45**	**57.14**
subgraph-2	64.77	61.05	79.25	42.86
subgraph-3	60.23	59.22	64.15	54.29

TABLE 6 – Comparaison des mesures d'exactitude et de précision pour des modèles unigramme, bigramme et trigramme par rapport à nos modèles à sous-graphes de dépendances de tailles 1, 2 et 3.

Comme le montre la table de mesures, la meilleure valeur d'exactitude est obtenue avec un modèle à sous-graphes de dépendances de taille 1 (78.41%). Quant à elle, la meilleur valeur d'exactitude pour les modèles à n-grammes est obtenue avec un modèle unigramme (73.86%). Les performances des modèles n-gramme se dégradent au fur et à mesure que l'ordre du modèle augmente. Le même phénomène se produit avec les modèles à base de sous-graphes de dépendances : l'exactitude diminue avec l'accroissement de la taille des sous-graphes. D'après nous,

ce phénomène provient de la taille des données, qui n'est pas suffisante pour que les modèles d'ordres supérieurs soient confrontés à suffisamment d'exemples d'apprentissage.

Puisque dans nos données, nous avons plus de documents d'opinion positive, la précision moyenne de classification est meilleure pour ces derniers. La combinaison des modèles unigramme, bigramme and trigramme en vue d'obtenir de meilleurs résultats de classification n'a pas répondu à nos attentes de manière significative. De la même manière, la combinaison des modèles à base de sous-graphes de dépendances de différentes tailles n'a pas produit d'amélioration significative non plus.

Dans les Figure 5 et 6 nous présentons les 10 sous-graphes les plus fréquents de taille 1 et les 5 sous-graphes les plus frèquents de taille 2 (selectionnés avec le score Δidf) respectivement pour les classes de documents positifs et négatifs.

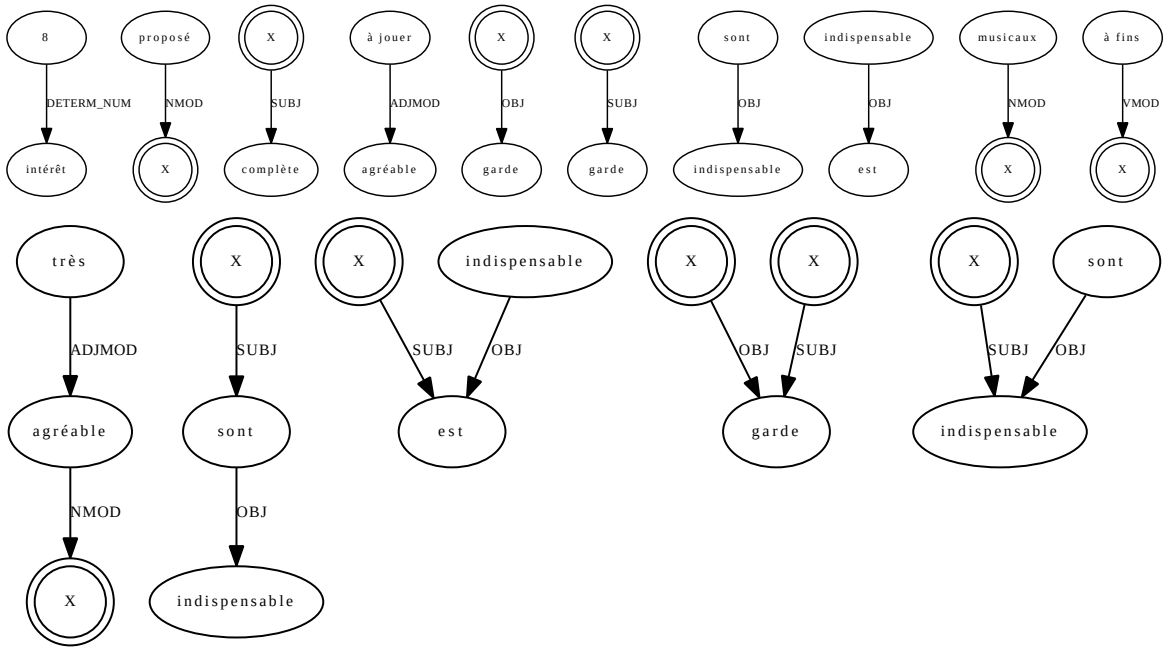


FIGURE 5 – Sous-graphes extraits des critiques de jeux vidéo positives

4 Travaux apparentés

Une première expérience par (Pang *et al.*, 2002), utilisant la représentation « sac de mots » avec des traits binaires et des classfieurs SVM, est devenue une base pour de nombreux travaux dans le domaine de la classification des sentiments. Les auteurs ont amélioré leur système dans (Pang & Lee, 2004) en utilisant un détecteur de subjectivité basé sur la notion de coupe minimale dans un graphe. L'utilisation d'un détecteur de subjectivité a permi de diminuer le bruit et se concentrer uniquement sur les phrases exprimant des sentiments. Cette méthode a amélioré la précision de 82.7% à 86.4%. Par la suite, de nombreux travaux ont utilisé des techniques avancées et des lexiques additionels pour augmenter l'espace de trait ou bien pour affiner la sélection des traits pertinents, améliorant ainsi la précision de la classification. (Whitelaw *et al.*, 2005) utilise des groupes d'appréciation, comme *« very good »* (très bon) ou « not terribly funny » (pas vraiment drôle) dans le cadre de la théorie de l'appréciation (*Appraisal theory*) en combinaison avec le modèle « sac de mots » et a obtenu une précision de 90.2% sur le jeu de données de critiques de films. (Aue & Gamon, 2005) a utilisé les SVM avec une sélection de traits par registre de probabilité et a obtenu une précision de 90.2% sur le même jeu de données.

L'arbre de dépendances des phrases a été largement utilisé dans le domaine de l'analyse de sentiments. Une recherche récente par (Arora *et al.*, 2010) a noté les problèmes de la représentation habituelle des textes par une approche « sac de mots ». Les auteurs suggéraient d'utiliser leur algorithme pour extraire les traits de sous-graphe

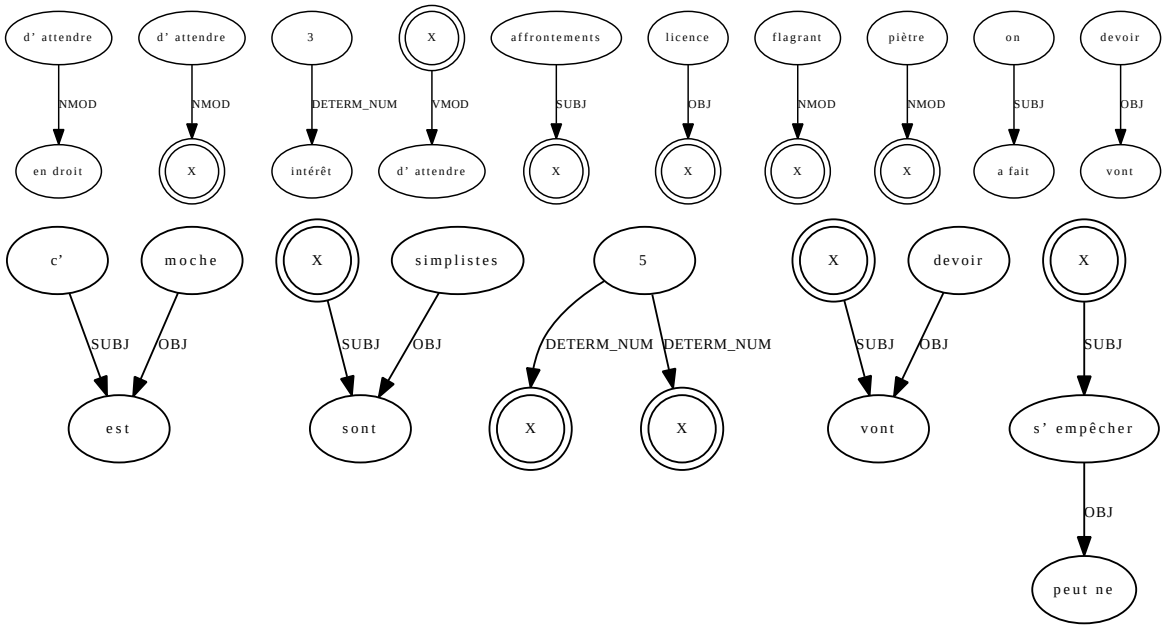


FIGURE 6 – Sous-graphes extraits des critiques de jeux vidéo négatives

par la programmation génétique. Cependant, les traits obtenues n'étaient pas utilisées pour remplacer le modèle n-gramme classique, mais plutôt comme un jeu de traits complémentaire. Un travail récent par (Nakagawa *et al.*, 2010) utilise un arbre de dépendances pour obtenir des traits qui sont utilisées pour entraîner un classifieur CRF pour la détection de la polarité des sentiments. Dans (Zhuang *et al.*, 2006), les auteurs utilisent des arbres de dépendances pour extraire les paires trait-opinion, ou le premier membre de la paire est un terme trait (ex. *« movie »*/film) et le second est un porteur d'opinion (ex. *« masterpiece »*/chef d'œuvre). Les arbres de dépendances sont utilisés afin d'établir les relations entre les mots traits et les mots-clés d'opinion. Dans (Chaumartin, 2007), l'arbre de dépendance est utilisé pour normaliser des titres vers des formes grammaticalement correctes, avant analyse des sentiments. Dans (Meena & Prabhakar, 2007), les auteurs utilisent l'arbre de dépendances et WordNet pour effectuer une analyse en sentiments.

5 Conclusion

Avec l'explosion du nombre de blogs et le développement des réseaux sociaux, la fouille d'opinion et l'analyse de sentiments sont devenus des domaines d'intérêt pour la recherche. Un travail pionnier sur la classification supervisée en sentiments à base de n-grammes ayant produit des résultats prometteurs, de nombreux chercheurs ont développé ce type de modèle. Cependant, l'approche « sac de mots » pour représenter un texte ne permet pas de prendre en compte des expressions complexes de sentiments et ne se prête que difficilement à l'utilisation de modèles sophistiqués de sentiments, qui nécessitent d'identifier entre autres, l'intensité d'une opinion ou la source/cible d'une expression d'opinion. Clairement, un nouveau type de modèle est nécessaire afin d'obtenir de meilleures performances en classification automatique de sentiments et en fouille d'opinion. Dans nos travaux, nous avons développé une nouvelle représentation à base de sous-graphes extraits des arbres de dépendances syntaxiques. Nous représentons un texte comme une collection de sous-graphes, où les nœuds sont des mots (ou des classes de mots) et les arcs des dépendances syntaxiques entre ceux-ci. Une telle représentation évite la perte d'information associée à l'emploi de modèles « sac de mots » pour représenter un texte, ces derniers étant basés uniquement sur des collections de n-grammes de mots. Nous avons testé notre modèle sur un ensemble de critiques de jeux vidéo, développé dans le cadre du projet DOXA sur la fouille d'opinion. Ainsi nous avons pu montrer qu'un classifieur SVM utilisant des traits construits à partir des sous-graphes extraits des arbres de dépendances, donne de meilleurs résultats que les systèmes traditionnels à base d'unigrammes. L'exactitude la plus élevée que nous ayons mesurée sur des textes en français est de 75%. Nous pensons que cette mesure peut

encore être améliorée par l'utilisation de techniques avancées de sélection de traits ou l'utilisation de lexiques dédiés à l'analyse de sentiments et d'opinion.

Remerciements

Ces travaux ont reçu le soutien financier du projet DOXA du pôle de compétitivité CAP-DIGITAL.

Références

AÏT-MOKHTAR S., CHANOD J.-P. & ROUX C. (2002). Robustness beyond shallowness : incremental deep parsing. *Nat. Lang. Eng.*, **8**, 121–144.

ARORA S., MAYFIELD E., PENSTEIN-ROSÉ C. & NYBERG E. (2010). Sentiment classification using automatically extracted subgraph features. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, p. 131–139, Morristown, NJ, USA : Association for Computational Linguistics.

AUE A. & GAMON M. (2005). Customizing Sentiment Classifiers to New Domains : a Case Study. In *Proc. International Conference on Recent Advances in NLP*.

CHAUMARTIN F.-R. (2007). Upar7 : a knowledge-based system for headline sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, p. 422–425, Morristown, NJ, USA : Association for Computational Linguistics.

DE MARNEE M.-C. & MANNING C. D. (2008). Stanford typed dependencies manual. *http ://nlp.stanford.edu/software/dependencies_manual.pdf*.

FAN R.-E., CHANG K.-W., HSIEH C.-J., WANG X.-R. & LIN C.-J. (2008). Liblinear : A library for large linear classification. *J. Mach. Learn. Res.*, **9**, 1871–1874.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA, USA : MIT Press.

MEENA A. & PRABHAKAR T. V. (2007). Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In *Proceedings of the 29th European conference on IR research*, ECIR'07, p. 573–580, Berlin, Heidelberg : Springer-Verlag.

NAKAGAWA T., INUI K. & KUROHASHI S. (2010). Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, p. 786–794, Morristown, NJ, USA : Association for Computational Linguistics.

PAK A. & PAROUBEK P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC10)*, Valletta, Malta : European Language Resources Association (ELRA).

PALTOGLOU G. & THELWALL M. (2010). A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, p. 1386–1395, Morristown, NJ, USA : Association for Computational Linguistics.

PANG B. & LEE L. (2004). A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, p. 271–278.

PANG B., LEE L. & VAITHYANATHAN S. (2002). Thumbs up ? : sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, p. 79–86, Morristown, NJ, USA : Association for Computational Linguistics.

PAROUBEK P., PAK A. & MOSTEFA D. (2010). Annotations for opinion mining evaluation in the industrial context of the doxa project. In N. C. C. CHAIR), K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER & D. TAPIAS, Eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

WHITELAW C., GARG N. & ARGAMON S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, p. 625–631, New York, NY, USA : ACM.

ZHUANG L., JING F. & ZHU X.-Y. (2006). Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, p. 43–50, New York, NY, USA : ACM.

Annexe A. Liste des dépendences produites par XIP

```
ADJMOD
    ADJMOD_POSIT1
    ADJMOD_PROPQUE

ADVMOD
AUXIL_PASSIVE

CONNECT
    CONNECT_REL
    CONNECT_SUBJ

COORD
    COORDITEMS
    COORDITEMS_SC

COREF_POSIT1_REL
    COREF_REL

DATE*
    DATE_PERIODE*

DEEPOBJ
DEEPSUBJ
    DEEPSUBJ_PASSIVE
    DEEPSUBJ_PROPQUE

DETERM
    DETERM_DEF
    DETERM_DEM
    DETERM_INT
    DETERM_NUM
    DETERM_POSS
    DETERM_QUANT
    DETERM_QUANT_DEF
    DETERM_QUANT_DEM

LIEU*
    LIEU_BATIMENT*
    LIEU_CONTINENT*
    LIEU_IMPERSO*
    LIEU_PAYS*
    LIEU_PAYS_REGION*
    LIEU_QUARTIER*
    LIEU_REGION*
    LIEU_REGION_VILLE*
    LIEU_VILLE*

NEGAT*
    NEGAT_SUBJ

NMOD
    NMOD_NUM
    NMOD_POSIT1
    NMOD_POSIT2
    NMOD_POSIT3
    NMOD_PROPQUE
    NMOD_REL

OBJ
    OBJ_COORD
    OBJ_COORD_SPRED
    OBJ_PROPQUE
    OBJ_PROPQUE_COORD
    OBJ_PROPQUE_SPRED
    OBJ_REL
    OBJ_SPRED
    OBJ_SUBJ

ORG*
    ORG_BATIMENT_LIEU*
    ORG_ENTREPRISE*

PERSONNE*
PREPMOD
PREPOBJ_REL
SEQNP**
SUBJ
    SUBJ_COORD
    SUBJ_IMPERSO
    SUBJ_IMPERSO_COORD
    SUBJ_IMPERSO_PASSIVE
    SUBJ_PASSIVE
    SUBJ_PASSIVE_COORD
    SUBJ_PASSIVE_PROPQUE
    SUBJ_PASSIVE_REL
    SUBJ_PROPQUE
    SUBJ_REFLEXIVE
    SUBJ_REL
    SUBJ_REL_COORD
    SUBJ_SUBJ

SUBJCLIT
    SUBJCLIT_PASSIVE

URL*
VMOD
    VMOD_COORD
    VMOD_COORD_SPRED
    VMOD_IMPERSO
    VMOD_POSIT1
    VMOD_POSIT1_SUBJ
    VMOD_POSIT1_SUBORD
    VMOD_POSIT2
    VMOD_PROPQUE
    VMOD_REL
    VMOD_SPRED
    VMOD_SUBJ
    VMOD_SUBORD
```

Les relations sont marquées d'un astérisque (*), une séquence de relation SEQNP est marquée par (**).

Une modélisation des dites alternances de portée des quantifieurs par des opérations de combinaison des groupes nominaux

Sylvain Kahane

Modyco, Université Paris Ouest Nanterre & CNRS / Alpage, INRIA
sylvain@kahane.fr

Résumé.

Nous montrons que les différentes interprétations d'une combinaison de plusieurs GN peuvent être modélisées par deux opérations de combinaison sur les référents de ces GN, appelées combinaison cumulative et combinaison distributive. Nous étudions aussi bien les GN définis et indéfinis que les GN quantifiés ou pluriels et nous montrons comment la combinaison d'un GN avec d'autres éléments peut induire des interprétations collective ou individualisante. Selon la façon dont un GN se combine avec d'autres GN, le calcul de son référent peut être fonction de ces derniers ; ceci définit une relation d'ancrage de chaque GN, qui induit un ordre partiel sur les GN. Considérer cette relation plutôt que la relation converse de portée simplifie le calcul de l'interprétation des GN et des énoncés. Des représentations sémantiques graphiques et algébriques sans considération de la portée sont proposées pour les dites alternances de portée.

Abstract.

We show that the various interpretations of a combination of several Noun Phrases can be modeled by two operations of combination on the referent of these NPs, called cumulative and distributive combinations. We study definite and indefinite NPs as well as quantified and plural NPs and we show how the combination of an NP with other NPs can induce collective or individualizing interpretations. According to the way a NP combine with another NP, the calculation of its referent can be a function of the latter; this defines an anchoring relation for each NP, which induces a partial order on NPs. Considering this relation rather than the converse scope relation simplifies the calculation of the interpretation of NPs and utterances. Graphic and algebraic semantic representations without considering scope are proposed for the so-called scope alternations.

Mots-clés : portée des quantifieurs, cumulatif, collectif, distributif, référent de discours, ancrage.

Keywords: quantifier scope alternation, cumulative, collective, distributive, discourse referent, anchoring.

1 Introduction

Nous nous intéressons dans cet article à l'interprétation des GN définis, indéfinis et quantifiés notamment lorsqu'ils se trouvent dans la portée d'un autre GN, comme dans les exemples suivants :

(1) **a.** *Les invités ont trouvé* ***un chat****.*
b. *Au moins la moitié des profs pense qu'****aucun étudiant*** *ne parle* ***deux des langues étudiées cette année****.*
c. *Les trois filles ont donné* ***deux pommes*** *à* ***deux garçons****.*

Il est bien connu que de tels énoncés sont ambigus et une abondante littérature concerne le calcul de leurs interprétations possibles (cf. Link 1997, Corblin 2002, Dobrovie-Sorin & Beyssade 2004, Nickel à par., pour des analyses détaillées de la littérature).

L'objectif de cet article est de faire une présentation unifiée du calcul des interprétations possibles de ces GN, alors que ces questions sont souvent abordées séparément selon qu'on s'intéresse aux GN indéfinis plutôt qu'aux GN définis, aux GN pluriels ou aux GN coordonnées ou encore aux GN quantifiés. La résolutions de chacun de ces cas tend à faire appel à des concepts différents : la résolution d'anaphore et la recherche d'un antécédent pour les GN définis, les interprétations référentielles ou non pour les GN indéfinis, les possibilités d'interprétations collectives ou individualisantes[1] pour les GN pluriels, des parallélismes pour les GN coordonnés (déclenchés par un adverbe comme *respectivement*), ou encore les questions de portée pour les GN quantifiés.

L'un des contributions de notre article est de montrer que le problème de l'interprétation des GN n'est pas tant de calculer la portée des GN que de décider si un GN doit ou non être interprété comme étant dans la portée d'un autre GN. Autrement dit, nous proposons de renverser le calcul de la portée : il ne s'agit plus d'associer une portée à un GN, mais d'associer un ancrage à un GN. Cette relation d'ancrage est la relation converse de la relation de portée (X s'ancre sur Y si et seulement si X est interprété comme étant dans la portée de Y). On peut penser que ceci ne change rien au problème et pourtant cela change beaucoup de choses : cela change la représentation sémantique des énoncés, puisque les GN quantifiés et les quantifieurs[2] n'ont plus à proprement parler une portée, alors que tout GN possède un ancrage.

L'autre contribution de notre article est de montrer que l'on peut modéliser les différentes interprétation d'une prédication sur plusieurs GN par deux opérations de combinaisons sur les référents de ces GN : la combinaison cumulative et la combinaison distributive. Nous verrons que le choix d'une opération plutôt qu'une autre est directement déductible de la relation d'ancrage entre les référents des GN, relation qui définit un ordre partiel sur les GN, et nous en déduirons diverses représentations sémantiques d'une prédication sur plusieurs GN. Sans relever directement du cadre de la DRT (Kamp & Reyle 1993), notre travail s'inscrit dans les horizons ouverts par la sémantique dynamique, consistant à introduire pour chaque GN un référent de discours et à calculer dynamiquement la valeur de ce référent.

Dans cet article, nous commencerons par présenter les différentes interprétations que peuvent avoir des GN définis lorsqu'ils sont combinés et nous définirons les notions d'interprétations collective, individualisante et synchrone (Section 2). Nous étendrons ensuite ces notions à l'interprétation des GN indéfinis et nous introduirons les opérations de combinaison cumulative et distributive, ainsi que la notion d'ancrage (Section 3). Nous verrons comment ces opérations se combinent lorsqu'on considère trois GN et nous montrerons l'équivalence entre les différentes configurations d'opérations et les différentes relations d'ancrage entre ces GN (Section 4). Nous étudierons ensuite la modélisation des GN quantifiés sans introduire de portée et nous reviendrons sur les indéfinis pluriels et singuliers (Section 5).

---

[1] On oppose généralement *collectif* à *distributif*, mais le même *distributif* est aussi utilisé pour un autre concept où il s'oppose à *cumulatif*. Nous utiliserons le terme *distributif* uniquement en opposition à *cumulatif* et *individualisant* en opposition à *collectif*.

[2] Nous appellerons *quantifieurs* les objets linguistiques comme *chaque* et *quantificateurs* les objets mathématiques de la logique frégéenne.

2 Interprétations collectives, individualisantes et synchrones des GN

Nous allons commencer notre étude par les combinaisons de GN définis comme en (2) :

(2) *Les trois filles ont déplacé les quatre pianos*.

L'énoncé (2) indique que les quatre pianos ont été déplacés et que les agents de ce déplacement sont les trois filles. Cet énoncé est fondamentalement vague : il n'indique pas si les filles ont agi individuellement ou collectivement, ni quelle fille a participé à quel déplacement de piano. Une formalisation des différentes interprétations d'un énoncé tel que (2) a été proposée par Gillon (1987, 1996) : soit A l'ensemble des trois filles et B l'ensemble des quatre pianos. Alors l'énoncé (2) implique qu'il existe un recouvrement[3] $(A_i)_{1\leq i\leq n}$ de A et un recouvrement $(B_i)_{1\leq i\leq n}$ de B tel que pour chaque i de 1 à n, l'ensemble des filles de A_i aient collectivement déplacé l'ensemble des pianos de B_i (l'énoncé décrivant ainsi un ensemble de n évènements). Pour des pianos, notre connaissance du monde nous incite à considérer que chaque piano a été déplacé séparément (et que donc les B_i sont des singletons), mais pour un énoncé tel que (3), on imagine facilement une situation où chaque fille doit soulever à son tour l'ensemble des quatre chaises.

(3) *Les trois filles ont soulevé les quatre chaises*.

Parler ici des différentes interprétations est assez commode, mais en fait abusif, pour ne pas dire erroné. En effet, l'interprétation que l'on fait d'un énoncé tel que (2) peut fort bien rester vague. Plus précisément, ce que nous décrivons lorsque nous parlons des différentes interprétations sont les différentes situations que peut recouvrir l'interprétation proprement dite de l'énoncé. Cela reste vrai lorsqu'on s'intéresse à la synthèse : le locuteur peut avoir une idée plus ou moins précise du procès qu'il décrit et des situations exactes qu'il recouvre.

Parmi les interprétations de (2)-(3), on distingue certaines interprétations « extrêmes » : l'interprétation du GN sujet est dite *collective* si tous les A_i sont égaux à A et *individualisantes* si tous les A_i sont des singletons (contrairement à certains usages, nous réservons le terme *distributif* à un autre type d'interprétation dont il va être question ensuite). Trois remarques s'imposent :

1) Les interprétations « extrêmes » semblent de loin les plus accessibles, mais des contraintes lexicales ou pragmatiques peuvent permettre des interprétations intermédiaires (Gillon 1987). On peut forcer les interprétations collective, individualisante ou intermédiaire d'un des actants (sujet ou objet) par l'ajout de modifieurs :

(4) *Les trois filles ont déplacé les quatre pianos ensemble/l'un(e) après l'autre/deux par deux*.

Lorsque A et B ont le même nombre d'éléments, les interprétations doublement individualisantes minimales[4] reviennent à aligner les éléments de A et de B. Une telle interprétation sera dite *synchrone*. Elle est forcée par l'ajout de l'adverbe *respectivement* et favorisée par le parallélisme syntaxique des deux groupes (cf. (5)b qui déclenchera moins immédiatement cette interprétation que (5)a) :

(5) **a.** *Adi et Béa ont (respectivement) mangé la tartelette aux fraises et la tartelette aux poires*.
b. *Les jumelles ont (?? respectivement) mangé les deux tartelettes*.

2) Comme il a été fort bien montré dans la littérature (Landman 1989, Link 1997), ce n'est pas le GN qui possède une interprétation collective ou individualisante, mais c'est le procès qui induit une interprétation collective ou individualisante du GN *à l'intérieur du procès*. Autrement dit, un même GN impliqué dans deux procès peut très bien avoir une interprétation collective à l'intérieur d'un procès et individualisante à l'intérieur de l'autre :

[3] $(A_i)_{1\leq i\leq n}$ est un *recouvrement* de A si la réunion des A_i est égale à A ($A = \cup_{1\leq i\leq n} A_i$). Cette notion ne suppose pas que les A_i sont deux à deux disjoints.

[4] Il existe aussi une interprétation doublement individualisante *maximale* où chaque élément de A agit sur chaque élément de B.

(6) **a.** *Adi et Béa se sont rencontrées dans un bar et ont bu une bière* (Dowty 1986, Lasersohn 1989)
b. *Les trois filles ont amené une chaise chacune et une table (ensemble).*

Ceci amène un certain nombre d'auteurs à dire que l'interprétation collective ou individualisante du GN n'est pas une propriété du GN, mais est déclenché par une propriété du GV. Bien qu'il soit préférable d'associer cette propriété au GV plutôt qu'au GN, cela n'est pas non plus satisfaisant. En effet, le problème ne concerne pas seulement le GN sujet (celui qui se combine avec le GV), mais tous les GN quelle que soit leur position, et les données montrent une certaine symétrie entre GN sujet et objet. Nous considérons pour notre part que l'interprétation collective ou individualisante d'un GN est la *projection* sur le GN d'une propriété de la *combinaison* de ce GN avec le verbe et d'autres GN. Autrement dit, le GN n'a pas, pris isolément, d'interprétation collective ou individualisante ; c'est au sein d'une combinaison avec un autre élément (un GN, mais aussi un prédicat verbal) que le GN est considéré collectivement ou individu par individu et cette interprétation particulière du GN reste localisée au sein de cette combinaison.

3) Nous avons illustré l'interprétation des GN définis avec un procès à deux participants, mais les mêmes remarques valent quel que soit le nombre de participants : cf. (7)a et b qui mettent respectivement en jeu 1 et 3 participants, pour lesquels il y a le même flou sur la participation collective ou individualisante de chaque participant :

(7) **a.** ***Les étudiants*** *sont arrivés en retard.*
b. ***Les enfants*** *ont donné* ***des jouets de bébé*** *à* ***leurs petits cousins****.*

La définition donnée pour deux GN peut être étendue à un nombre quelconque de dimensions.

3 Combinaisons cumulative et distributive des GN

Considérons maintenant des variantes de (2) où l'objet et/ou le sujet sont indéfinis :

(8) **a.** *Les trois filles ont déplacé quatre pianos.*
b. *Trois filles ont déplacé les quatre pianos.*
c. *Trois filles ont déplacé quatre pianos.*

Chacun de ces trois énoncés possèdent, parmi d'autres, une série d'interprétations où il y a seulement trois filles et quatre pianos en jeu comme en (2), à la différence que les GN en question sont indéfinis, c'est-à-dire qu'ils ne sont pas présentés comme référents à un groupe identifiable par l'interlocuteur. Les interprétations de ce type, où aucun GN n'est dans la portée de l'autre, sont appelées *cumulatives* (Scha 1981, Schwarzschild 1996, Nickel à par.). Nous allons donner une définition formelle des interprétations cumulatives. Si GN est un groupe nominal, on note $\underline{GN}$ son référent. Pour l'instant nous considérons des cas simples où GN possède un référent fixe $\underline{GN}$, qui est donc un ensemble d'objets du monde[5]. Soit P un prédicat binaire modélisant un verbe bi-actanciel ; P(A , B) signifie qu'il existe une P-événement impliquant l'ensemble des individus de A et l'ensemble des individus de B. Nos énoncés dénotent en général une série d'événements (cf. Krifka 1990 pour la relation entre quantification nominale et événements).

> **Définition 1.** GN1 et GN2 *se combinent cumulativement* (par le prédicat P) s'il existe des recouvrements $(A_i)_{1\leq i\leq n}$ et $(B_i)_{1\leq i\leq n}$ de $\underline{GN1}$ et $\underline{GN2}$ tel que pour chaque i de 1 à n, on ait $P(A_i , B_i)$. On note $\underline{GN1} \otimes \underline{GN2}$ une combinaison cumulative de $\underline{GN1}$ et $\underline{GN2}$ et $P(\underline{GN1} \otimes \underline{GN2})$ une interprétation cumulative de P.

La combinaison cumulative est commutative et chaque GN possède un référent indépendant du référent de l'autre GN. Comme on l'a vu, il y a parmi les interprétations cumulatives des interprétations collectives ou individualisantes, pour le sujet et/ou l'objet, et d'autres qui ne le sont pas, néanmoins la combinaison n'affecte pas le calcul des référents des GN : les interprétations collectives ou idividualisantes ne valent qu'au sein de la combinaison et pour l'application du prédicat P. Notons encore que P s'applique à un objet plus riche que le produit cartésien $\underline{GN1} \times \underline{GN2}$ ou le couple $(\underline{GN1} , \underline{GN2})$.

[5] Ce qu'on appelle le monde est en fait une représentation du monde dans le savoir partagé des interlocuteurs.

Du fait du caractère indéfini du GN, il existe un autre type d'interprétation. En (8)a, le GN *quatre pianos* peut ne pas avoir un référent absolu, mais relatif au GN sujet *les trois filles*. Autrement dit, il est possible d'interpréter (8)a comme décrivant une situation où chacune des trois filles a déplacé quatre pianos a priori différents à chaque fois (la question de savoir si ce sont les quatre mêmes pianos n'est tout simplement pas exprimée et est indécidable et non pertinente dans l'interprétation). Cette interprétation est dite *distributive* et elle met en jeu potentiellement 3×4 pianos[6].

> **Définition 2.** GN2 se *combine distributivement* avec GN1 (par le prédicat P) si pour tout individu a_i dans GN1, il existe un référent B_i de GN2 tel que $P(\{a_i\} \otimes B_i)$[7]. On note GN1 $\odot$ GN2 une combinaison distributive de GN2 par rapport à GN1 et P(GN1 $\odot$ GN2) une interprétation distributive de P.

Comme on le voit, dans une interprétation distributive, le référent de GN2 est une fonction sur les individus du référent de GN1.

> **Définition 3.** GN2 *s'ancre* sur GN1 si le référent de GN2 est fonction (des individus) du référent de GN1. Le référent de GN2 est dit *relatif* (à celui de GN1). Lorsqu'un GN ne s'ancre sur aucun autre GN, nous dirons que le référent du GN est *absolu* et nous considérons que le GN s'ancre dans le contexte.

On peut, à l'instar de Steedman 2009, modéliser de tels référents par des fonctions de Skolem généralisées. Nous verrons plus loin une représentation en termes d'ancrage plus ou moins équivalente.

Revenons à nos exemples et à leurs différentes interprétations. L'interprétation distributive (du sujet par rapport à l'objet) en (8)b est beaucoup plus difficile qu'en (8)a (où il s'agissait de la distribution de l'objet par rapport au sujet), mais pas totalement exclue. Par exemple dans le cadre d'une enquête, on pourrait dire : *Trois filles ont déplacé ces deux pianos. Voici les traces des trois qui ont déplacé le piano de gauche et ici les traces des trois qui ont déplacé le piano de droite*. Plus généralement il a déjà été noté qu'un GN indéfini sujet pouvait être dans la portée du GN objet (on parle alors de portée inversée, Beyssade 2006) :

(9) **a.** *Un guide accompagne chaque visiteur dans la crypte.*
b. *Un guide accompagne la plupart des visiteurs dans la crypte.*
c. *Un guide n'accompagne aucun visiteur dans la crypte.*

L'énoncé (8)c, avec deux GN indéfinis, possède a priori les deux interprétations distributives : distributivité de l'objet par rapport au sujet et distributivité (beaucoup moins probable) du sujet par rapport à l'objet. Pour chacune de ces interprétations distributives, l'un des GN s'ancre sur l'autre et le calcul de ses référents possibles se fait par rapport à l'autre GN.

Nous estimons que les interprétations cumulatives et distributives sont les seules possibles pour les énoncés en (8) et qu'il n'existe pas d'interprétations intermédiaires (voir juste après pour des cas où elles sont possibles). Par exemple, l'énoncé (8)a ne peut couvrir une situation où deux filles ont déplacé ensemble quatre pianos et la troisième a déplacé séparément quatre autres pianos ; soit certaines filles ont agi collectivement et alors l'interprétation est obligatoirement cumulative et il n'y a qu'un jeu de quatre piano, soit il y a plusieurs jeux de quatre pianos et alors toutes les filles ont agi individuellement.

Notons encore une difficulté supplémentaire. Nous avons fait comme si l'interprétation distributive était nécessairement individualisante, c'est-à-dire comme si le GN devait se distribuer par rapport aux individus du GN sur lequel il s'ancre. Plus qu'individualisante, l'interprétation distributive est *séparative* : elle sépare le GN ancre en un certain nombre de paquets par rapport auquel le deuxième GN se distribue[8]. Une telle

---

[6] Les situations où chacune des trois filles a déplacé les quatre mêmes pianos sont à la fois distributive et cumulative. Autrement, les deux notions recouvrent des situations différentes.

[7] On notera que, si les interprétations distributives forcent une interprétation individualisante de GN1, elles restent par contre floues sur l'interprétation collective ou non de GN2 et c'est pourquoi on a a priori $P(\{a_i\} \otimes B_i)$ et non $P(\{a_i\}, B_i)$.

[8] C'est l'occasion de revenir sur notre première note concernant l'emploi du terme *distributif* dans la littérature : les interprétations distributives étant la plupart du temps individualisantes, notamment dans

distribution dite *partielle* (c'est-à-dire où la séparation ne se fait pas au niveau individuel) n'est possible qu'avec une structure syntaxique particulière du GN indiquant la partition (et est donc impossible dans des exemples comme (8)). Ainsi en (10)a, une interprétation possible est que les filles d'une part et les garçons d'autre part aient déplacé chacun un piano :

(10) **a.** *Les filles et les garçons ont déplacé un piano.*
b. *Les filles et les garçons ont déplacé quatre pianos.*
c. *Les filles et les garçons ont (respectivement) déplacé le piano et l'armoire.*

Comme on peut s'en convaincre avec (10)b, dans une interprétation à distribution partielle, pour chaque groupe ayant déplacé quatre pianos toutes les situations d'une interprétation cumulative sont possibles. Ce qu'il est important de dire sur l'interprétation exacte de tels énoncés, c'est que le flou sur la situation exacte ne vise pas être levé et que ce qui est dit alors c'est que chacun des deux groupes a déplacé quatre pianos et que la façon dont chacun des deux groupes s'y est pris n'importe pas dans ce qui est dit et ce qu'on doit en déduire. Notons également que l'interprétation synchrone d'un énoncé comme en (10)c est possible sans qu'il y ait une interprétation individualisante des GN : il suffit d'une interprétation séparative du GN, suggérée par la syntaxe du GN, pour permettre l'alignement filles-piano et garçons-armoire.

4 Ancrage, combinaisons et représentation sémantique

Nous allons maintenant nous intéresser à des énoncés combinant trois GN. Ceci nous permettra de faire le lien entre les différentes configurations d'ancrage et la composition de nos deux opérations de combinaison, ⊗ et ⊙.

(11) **a.** ***Les trois filles*** *ont donné* ***deux pommes*** *aux* ***deux garçons****.*
b. ***Les trois filles*** *ont donné* ***deux pommes*** *à* ***deux garçons****.*

Nous allons décrire les différentes interprétations de ces énoncés. Les référents des trois GN sont notés X = *les trois filles*, Y = *deux pommes*, Z = (*les*) *deux garçons*, tandis que Ω désigne le contexte. Pour chaque interprétation de l'énoncé, nous proposons trois représentations sémantiques.

- La première indique uniquement les ancrages des GN : chaque ancrage est représenté par une flèche ⋯> pointant sur l'ancre du GN. Cette représentation, similaire à celle proposée par Lesmo & Robaldo (2004) et Robaldo (2007), n'indique pas les combinaisons cumulatives. Elle est équivalente aux suivantes à la condition qu'on sache s'il y a combinaison ou non entre les GN.

- Nous avons vu que, si un GN s'ancre sur un autre, c'est qu'il se combine distributivement avec ce dernier. A l'opposé, s'il n'y a pas de relations d'ancrage entre deux GN et que ces GN sont pris sous le même prédicat et doivent donc se combiner, la combinaison est nécessairement cumulative. Notre deuxième représentation indique pour chaque couple de GN s'il se combine cumulativement ou distributivement. Nous gardons la flèche ⋯> pour la combinaison distributive et nous notons la combinaison cumulative par un rectangle ▭ incluant les éléments combinés.

- Notre troisième représentation est une formule algébrique indiquant comment les référents des GN se combinent par les opérations ⊗ et ⊙.

L'énoncé (11)a possèdent deux GN définis (de référents X et Z) qui se combinent nécessairement cumulativement ; il possède alors quatre interprétations selon l'ancrage de Y. La Fig. 1 donne les trois représentations sémantiques de chacune des interprétations. Une quatrième ligne indique le nombre de filles (X), de pommes (Y) et de garçons (Z) mis en jeu.

Dans l'interprétation 1, Y s'ancre dans le contexte et les trois GN se combinent cumulativement : il y a 3 filles, 2 pommes et 2 garçons et on ne sait pas exactement qui a donné quoi à qui. La définition de la combinaison cumulative peut facilement être généralisée à un nombre quelconque de GN et nous notons X ⊗ Y ⊗ Z la combinaison cumulative de trois GN.

---

des exemples comme en (8) qui servent de base aux études, on peut comprendre les raisons de la confusion entre les deux notions.

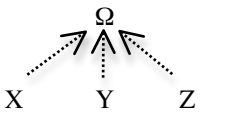
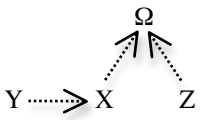
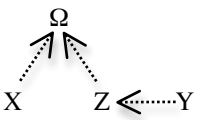
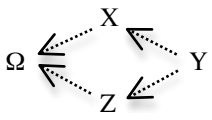
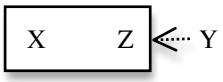
			
			
X ⊗ Y ⊗ Z	Z ⊗ (X ⊙ Y)	X ⊗ (Z ⊙ Y)	(X ⊗ Z) ⊙ Y
(3, 2, 2)	(3, 6, 2)	(3, 4, 2)	(3, 12, 2)
Interprétation 1	**Interprétation 2**	**Interprétation 3**	**Interprétation 4**

Figure 1. Les quatre interprétations de (11)a

Dans l'interprétation 2, Y s'ancre sur X : il s'ensuit que chacune des 3 filles a donné 2 pommes et qu'il y a potentiellement 6 pommes (on ne peut exclure que la même pomme ait été donnée plusieurs fois). Y se combine donc distributivement avec X (qui reçoit ainsi une interprétation individualisante) tout en se combinant de manière cumulative avec Z. Nous notons cela Z ⊗ (X ⊙ Y). La situation est analogue dans l'interprétation 3 : Y s'ancre sur Z et chaque garçon reçoit 2 pommes.

L'interprétation 4 est plus complexe : chacune des 3 filles donne 2 pommes à chacun des 2 garçons et il y a potentiellement 12 pommes. Il y a dans ce cas un double ancrage de Y simultanément sur X et Z. Le double ancrage n'est en fait possible que parce que X et Z se combinent cumulativement, et on peut considérer de manière équivalente que Y s'ancre sur la combinaison cumulative X ⊗ Z. Il est à noter qu'à chaque fois que Z s'ancre sur une GN ou une combinaison de GN, il induit une interprétation individualisante de ces GN. Dans le cas du double ancrage, il y a donc une interprétation doublement individualisante de X et Z (maximale ou minimale).

L'énoncé (11)b possède a priori six interprétations supplémentaires. Trois de ces interprétations sont les symétriques des interprétations 2, 3, 4, où X et Y sont combinés cumulativement et où les valeurs de (X, Y, Z) sont (3, 2, 6), (3, 2, 4) et (3, 2, 12). Ces interprétations sont pragmatiquement peu probables puisqu'elles supposent que les 2 pommes aient été données plusieurs fois. La Fig. 2 donne les représentations sémantiques des trois dernières interprétations :

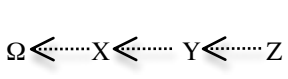
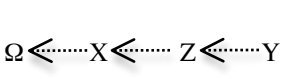
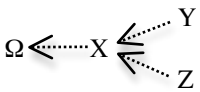
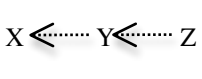
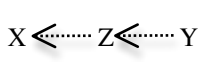
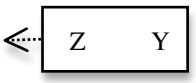
		
		
X ⊙ Y ⊙ Z	X ⊙ Z ⊙ Y	X ⊙ (Y ⊗ Z)
(3, 6, 12)	(3, 12, 6)	(3, 6, 6)
Interprétation 8	**Interprétation 9**	**Interprétation 10**

Figure 2. Trois interprétations de (18)b.

Comme on le voit, il est possible de s'ancrer sur un GN qui est lui-même ancré sur un autre GN (interprétations 8 et 9). Bien que non commutative, l'opération ⊙ est associative et nous notons X ⊙ Y ⊙ Z la double combinaison distributive de X avec Y et de Y avec Z. Dans l'interprétation 10, Y et Z s'ancrent tous les deux sur X ; ils se combinent donc cumulativement et on peut considérer que c'est la combinaison Y ⊗ Z qui s'ancre sur X, d'où la formule algébrique X ⊙ (Y ⊗ Z). En effet, dans ce cas, chacune des 3 filles donne 2 pommes à 2 garçons sans qu'on sache précisément si le don est collectif ou individuel.

Pour conclure cette section, nous allons dire quelques mots sur la relation d'ancrage. La relation d'ancrage définit un ordre partiel sur l'ensemble {X, Y, Z}. Dans la mesure où X est un GN défini (et en plus sujet syntaxique), il ne peut s'ancrer sur un autre GN et est donc nécessairement un plus petit élément pour cet ordre partiel. Comme on peut le voir, il y a autant d'interprétations que d'ordres partiels sur cet ensemble à 3 éléments avec X comme plus petit élément. A chaque configuration de la relation d'ancrage, nous associons de manière biunivoque une combinaison de X, Y et Z. Le fait que les relations de portée entre GN forme un ordre partiel a été souligné par les travaux d'Hinkitta (1976), ce qui a entrainé un courant de recherche sur les formules logiques avec un ordre partiel sur les quantificateurs (Barwise 1979, Sher 1990, Robaldo 2007). Notre opération $\otimes$ définit en quelque sorte l'opération de combinaison de deux quantificateurs qui ne sont pas dans la portée l'un de l'autre et fait la connexion entre ces travaux et les travaux de Gillon (1987, 1996).

Il est généralement considéré qu'on ne peut indiquer la portée d'un élément X dans une représentation uniquement sous forme de graphe (cf. Sowa (1987) pour un inventaire précis de ce qui peut et ne peut pas être représenté par un graphe en sémantique). En un sens, on peut le faire en ajoutant un lien vers chaque élément qui est dans la portée de X. Dire qu'on ne peut pas, c'est en fait dire qu'on ne peut pas représenter la portée par des moyens finis (puisqu'un nombre potentiellement illimité d'élément peut être dans la portée de X, chacun nécessitant un lien) et qu'on ne saisit pas le fait que LA portée de X est UNE caractéristique de X que l'on souhaite modéliser par UN objet. Dans notre cas, nous représentons la relation converse qu'est la relation d'ancrage. Cette relation d'ancrage peut être représentée par des liens dans un graphe. Comme, en plus, chaque élément possède généralement une ancre, deux au plus, on a de ce point de vue une condition de finitude que ne respectent pas les relations de portée. On peut également représenter les relations prédicatives entre les signifiés sous forme de graphe et obtenir ainsi une représentation sémantique entièrement basée sur des graphes (Mel'čuk 1988, Kahane 2005, Robaldo 2007, Copestake 2009).

5 Quantifieurs, ancrage et portée

Nous n'avons pas jusque-là abordé frontalement la question des quantifieurs. Rappelons que Frege proposait un traitement distinct des définis d'un part, modélisés par des constantes, et des indéfinis d'autre part, modélisés par des variables introduites par des quantifieurs. La sémantique dynamique a défendu une distinction entre les indéfinis et les GN quantifiés comme *chaque N* et opposé les GN référentiels comportant les définis et certains emplois des indéfinis aux GN quantifiés (Kamp & Reyle 1993, Corblin 2002).

Nous proposons de considérer que tous les GN introduisent un référent de discours. Certains GN introduisent par défaut un référent fixe comme les définis et les indéfinis. Les calculs de ce référent diffèrent ensuite, puisque l'indéfini va pointer sur un nouveau référent de discours, tandis que le défini va pointer vers un référent de discours connu ou identifiable[9]. Les GN quantifiés introduisent au contraire un référent mobile, que l'on peut modéliser par une variable ou un parcours d'un certain ensemble contextuellement inférable. A part leur caractère intrinsèquement mobile, ces référents de discours restent de nature comparable aux référents fixes et peuvent être repris anaphoriquement par des pronoms, même si leur accessibilité est plus restreinte :

(12) ***Chaque candidat*** *remplit un questionnaire*. ***Il*** *le rend avant de partir*. (Corblin 2002 : 238)

On peut comparer le fonctionnement du GN quantifié *chaque candidat* à celui du GN pluriel *les candidats*.

(13) ***Les candidats*** *remplissent un questionnaire*. ***Ils*** *le rendent avant de partir*.

Les phrases (12) et (13) sont synonymes et dans les deux cas, le GN *un questionnaire* s'ancre sur le GN sujet. Il y a néanmoins une différence importante : *les candidats* possède un référent pluriel (et est repris par un pronom pluriel), tandis que *chaque candidat* possède un référent singulier. Lorsque *un questionnaire* se combine distributivement avec *les candidats*, son référent est une fonction sur les individus du référent de

[9] Une des raisons qui justifie, à notre avis, un traitement unifié des définis et des indéfinis est que dans de nombreuses langues, comme le russe ou le mandarin, cette distinction n'est pas grammaticalisée. Autrement dit, il est possible de construire un GN sans que ni le locuteur, ni l'interlocuteur n'aient à décider si le référent de ce GN est identifiable ou non.

les candidats. Lors de l'ancrage sur *chaque candidat*, c'est le caractère intrinsèquement mobile du référent de ce GN qui donne à *un questionnaire* un référent de nature fonctionnelle. Nous utiliserons le même opérateur ⊙ pour désigner la combinaison entre un GN quantifié et un GN s'ancrant sur lui, mais la définition est un peu différente.

> **Définition 3.** GN2 se *combine distributivement* avec le GN quantifié à GN1 (par le prédicat P) si pour tout toute valeur a_i de GN1, il existe un référent B_i de GN2 tel que $P(\{a_i\} \otimes B_i)$ et la combinaison sera notée GN1 ⊙ GN2. La combinaison sera dite *cumulative* si GN2 ne dépend pas des valeurs de GN1 et la combinaison sera notée GN1 ⊗ GN2.

Une conséquence de la similarité de fonctionnement des GN pluriels et quantifiés est qu'il n'y a pas plus de raison de parler de portée pour *chaque candidat* que pour *les candidats*. Les deux GN permettent à un autre GN de s'ancrer sur eux et de déclencher un parcours de l'ensemble des candidats. Nous considérons, en conséquence, qu'un quantifieur n'introduit pas de portée ; il possède un unique argument sémantique qui est sa restriction, c'est-à-dire l'ensemble que parcourt son référent mobile, et qui est déterminé par la dénotation du GN. Rappelons que dans la Théorie des Quantifieurs Généralisés (Mostowski 1957, Barwise & Cooper 1981), un quantifieur est associé à un opérateur à deux arguments, qui sont respectivement sa restriction (P) et sa portée (Q) :

(14) **a.** *Tout homme est mortel* : ∀x [homme(x) → mortel(x)]
b. *tout* : λPλQ.(∀x [P(x) → Q(x)])

Nous pensons que le calcul du sens d'un énoncé comme (14)a est plus simple et plus proche de la syntaxe de la langue que ne le suggère la modélisation en (14)b : le lexème *tout* se combine syntaxiquement avec *homme*, et donc son signifié 'tout' est un prédicat prenant 'homme' comme argument. Le prédicat 'tout' appliqué à 'homme' a pour effet de construire un référent mobile qui parcourt l'ensemble des hommes. Le tout est ensuite argument de 'mortel', c'est-à-dire que chacun des référents potentiels de *tout homme* possède la propriété 'mortel'. Le quantifieur est un sémantème comme les autres et il peut lui-même être argument d'un autre sémantème (***presque*** *tout homme* ou *tout homme* ***sauf Jés***).

Terminons notre étude en revenant aux GN indéfinis. Considérons les énoncés suivants :

(15) **a.** *Les invités ont trouvé un chat*.
b. *Les filles ont déplacé un piano*.

L'énoncé (15)a possède deux interprétations : un interprétation distributive où chaque invité a trouvé un chat et une interprétation où il y a un chat que les invités ont trouvé. Il existe une analyse traditionnelle de cette ambiguïté en terme d'inversion de portée, où X = *les invités* et Y = *un chat* sont respectivement traités comme des quantifications universelle et existentielle :

(16) **a.** Pour tout invité *x*, il existe un chat *y* tel que *x* a trouvé *y*.
b. Il existe un chat *y* tel que pour tout invité *x*, *x* a trouvé *y*.

Cette modélisation n'est pas satisfaisante : si (16)a correspond bien à la première interprétation, (16)b, par contre, ne prend pas en compte que les invités ont collectivement trouvé le chat. Cette inadéquation devient évidente avec des exemples comme (15)b qui suppose un effort collectif.

Une autre modélisation consiste à considérer que, pour la deuxième interprétation, il n'y a pas d'inversion de portée, mais que l'indéfini Y échappe à la portée de X sans que pour autant X se retrouve dans la portée de Y (Sag & Fodor 1982, Beyssade 2006). C'est ce que décrit la combinaison cumulative. Notre modélisation ne traitent plus les indéfinis ou les pluriels comme des quantificateurs et nous ne gérons pas les différences d'interprétation de (15)a par des différences de portée, mais par des différences d'opérations de combinaison. Notons que pour les GN singuliers, on ne peut pas distinguer combinaisons distributive et cumulative : si GN1 = *un N*, GN1 ⊙ GN2 = GN1 ⊗ GN2 quel que soit GN2. Nous considérons néanmoins que pour la deuxième interprétation il s'agit d'une interprétation cumulative, car l'ancrage du sujet sur l'objet est un phénomène marginal (cf. la discussion sur les exemples en (8) et (9)).

D'autres arguments encore vont dans le sens du rejet d'une modélisation de ces GN par des quantifieurs avec une portée.

1) Le premier est la base de théories comme la DRT (Kamp & Reyle 1993, Corblin 2002), qui ont développées la notion de référent de discours sur laquelle se base notre analyse en termes d'ancrage. Considérons les fameuses « donkey sentences » :

(17) **a.** *Jil possède un âne. Elle le bat.*
b. *Si une fermière possède un âne, elle le bat.*

En (17)a, si on formalise *un âne* par une variable introduite dans la portée d'un quantificateur existentiel, on est confronté au problème de pouvoir réutiliser cette variable et donc de devoir prolonger la portée de ce quantifieur au delà des bornes de la première proposition (il existe un âne *y* tel que Jil possède *y* et tel que Jil bat *y*). En (17)b, la seule solution en terme de portée est de sortir les quantificateurs existentiels introduits par les indéfinis de la portée de *si* et d'en faire des quantificateurs universels (pour tout fermière *x* et tout âne *y*, si *x* possède *y*, alors *x* bat *y*).

En termes d'ancrage et de construction de référents de discours, les problèmes s'effacent : un GN indéfini crée un nouveau référent de discours, ce référent de discours étant ensuite accessible pour un pronom ou un GN défini qui voudrait pointer sur lui. Lorsque le GN indéfini Y ne s'ancre pas directement dans l'univers de discours, il crée quand même un référent de discours, lequel est assujetti à l'univers de l'élément sur lequel il s'est ancré. En (17)b, les deux GN indéfinis sont dans la portée de *si* : il vont donc s'ancrer dans l'univers de discours hypothétique ouvert par cet élément. En dehors de cela, ils fonctionnent comme n'importe quel référent de discours et sont accessibles à la reprise par un pronom, tant qu'on reste dans le sous-univers hypothétique ouvert par *si*. Plus généralement, un pronom peut très bien reprendre (par le biais d'une relation anaphorique), un élément lié à un autre référent de discours (voir aussi les exemples (12) et (13)) :

(18) *Tous mes amis avaient **une voiture**, mais plusieurs **l'**ont vendue.*

2) Il existe d'autres situations où un GN indéfini échappe à la portée d'un élément. Considérons l'exemple suivant proposé par Fodor & Sag (1982) :

(19) *If **a friend of mine from Texas** had died in the fire, I would have inherited a fortune.*

Il y a deux interprétations possibles de (19) : une où *a friend of mine from Texas* possède une référence absolue (favorisée ici par le contexte) et une autre où ce GN est dans la portée de *if*. Pour que le GN indéfini échappe à la portée de *if*, il faut que le locuteur ait en tête quelqu'un de bien précis, mais qu'il présente par un indéfini, car il le considère comme non identifiable par l'interlocuteur. Ceci argumente, à notre avis, davantage pour un ancrage direct dans l'univers de discours que pour une inversion de portée.

3) Tous les indéfinis ne fonctionnent pas de la même façon. Comparons :

(20) **a.** *Au moins la moitié des étudiants parlent **deux des langues** étudiées cette année.*
b. *Au moins la moitié des étudiants parlent **deux langues** étudiées cette année.*

L'exemple (20)a peut avoir deux interprétations dont l'une dite à portée large, c'est-à-dire où, selon notre point de vue, l'indéfini *deux des langues* s'ancre directement dans l'univers de discours. Par contraste, (20)b peut beaucoup plus difficilement avoir cette deuxième interprétation. Comment modéliser cette différence ? Si nous raisonnons en terme de portée, alors il faut considérer qu'un indéfini pur comme *deux langues* peut plus difficilement prendre dans sa portée un autre quantifieur qu'un partitif comme *deux des langues* qui combine un indéfini et un défini (***deux** de* [***les** langues*]). On ne voit pas très bien en quoi le défini renforcerait la portée du GN. Par contre, si on raisonne en terme d'ancrage, cela devient nettement plus motivé : un GN comportant une part de défini s'ancrera plus facilement directement dans l'univers de discours, comme le fait généralement un défini.

6 Conclusion

La principale contribution de cet article réside dans l'introduction des opérations de combinaison cumulative et distributive et dans le lien fait entre la composition de ces opérations et les relations d'ancrage.

Le problème de l'interprétation des GN est tentaculaire. Nous sommes loin d'avoir abordé toutes les questions qu'il soulève et nous avons probablement voulu en couvrir trop dans cet article en considérant aussi bien les GN définis et indéfinis que les GN pluriels ou quantifiés. Nous avons étudié la combinaison de GN pris dans une même prédication, mais nous n'avons pas étudié des combinaisons à l'intérieur d'un GN comme pour *un représentant de chaque village* où l'article indéfini sert à saisir un individu d'une classe ('représentant de chaque village') dont la définition contient un quantifieur et induit ainsi un référent mobile[10]. Ceci nous amènerait à définir le fonctionnement de la modification (c'est-à-dire le cas où l'un des arguments d'un prédicat est son gouverneur syntaxique) et son interaction avec le calcul des référents.

Considérer l'ancrage plutôt que la portée simplifie, à notre avis, aussi bien la représentation sémantique que son calcul. D'autres auteurs ont considéré l'ancrage plutôt que la portée. Dobrovie-Sorin & Beyssade (2004 : 141) parle de *dépendance référentielle*, mais elles ne vont pas jusqu'à substituer cette relation à la portée dans leurs représentations sémantiques. Nous n'avons pas retenu le terme de *dépendance* qui est déjà souvent utilisé pour désigner les relations prédicat-argument (Mel'čuk 1988), lesquelles correspondent assez naturellement aux dépendances syntaxiques. Robaldo 2007 propose des représentations en termes d'ancrage similaire aux nôtres, mais les GN sont représentés par des quantifieurs et les représentations sont interprétées par des formules logiques. Une autre représentation de l'ancrage est proposée par Steedman 2009 grâce à des fonctions de Skolem généralisées, mais la combinaison cumulative n'est pas clairement dégagée, me semble-t-il.

Nous avons proposés trois modes de représentation de la composition des GN, mais il reste à intégrer cela dans une représentation globale des énoncés et à formaliser pleinement la sémantique de nos opérations de combinaison. On aura noté que nous ne faisons pas usage de variables dans nos représentations, qu'elles soient sous forme graphique ou algébrique. L'introduction de variables dans les expressions algébriques est un moyen pratique de noter le point d'articulation entre deux objets sémantiques que l'on compose (par exemple entre un quantificateur et un prédicat), mais ce moyen peut être contourné par l'introduction d'opérateurs de composition (cf. Desclés & Kye-Seop 2006 pour des notations algébriques de formules du premier ordre à l'aide d'opérateurs plutôt que de variables). L'autre conséquence de la modélisation par des opérations de combinaison est de ne plus avoir à raisonner en termes de portée. Les conséquences sont importantes : les quantifieurs et les GN en général n'ont plus de portée, mais à l'inverse chaque GN possède une ancre. La représentation de ces éléments s'en trouve grandement simplifiée et le calcul de l'ancrage/portée peut être rapproché du calcul des relations anaphoriques.

Aborder le problème de la combinaison des GN en terme d'ancrage et d'opérations de combinaison ouvre à notre avis des perspectives nouvelles pour le calcul de l'interprétation des énoncés. Nous aimerions en particulier souligner que les combinaisons cumulatives sont courantes et mérite un mode de représentation élémentaire que ne permet pas la logique classique, même avec des représentations sous-spécifiées pour la portée. De plus, la combinaison cumulative subsume un grand nombre d'interprétations particulières et il serait nécessaire d'étudier comment différents facteurs lexicaux, prosodiques[11] et pragmatiques réduisent l'espace des interprétations possibles.

Remerciements

L'article a subi de profondes modifications entre sa soumission et sa version finale. Il s'agit d'un travail de longue haleine commencé il y a au moins trois ans et dont la version actuelle n'est certainement pas aboutie. Je remercie pour leurs nombreux commentaires Claire Beyssade, Laurence Danlos, Paola Pietrandrea et Alain Polguère, et tout particulièrement les quatre relecteurs de la conférence.

[10] Il semble aussi que les phrases génériques attribuent à *un N* un référent mobile :

(i) ***Une baleine*** *mange plusieurs tonnes de planctons en* ***une journée***.

[11] La prosodie dépend en partie de la structure communicative (ou *information packaging*) de l'énoncé, qui joue elle-même un rôle important dans les relations d'ancrage/portée, mais aussi dans les interprétations cumulatives. L'interprétation synchrone est par exemple favorisée par un parallélisme prosodique.

Références

BARWISE J., COOPER R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy* 4, 159-219.

BARWISE J. (1979) On Branching Quantifiers in English, *Journal of Philosophical Logic* 8, 47-80.

BEYSSADE C. (2006). Portée. In D. Godard, L. Roussarie, F. Corblin (éd.), *Sémanticlopédie: dictionnaire de sémantique*, GDR Sémantique & Modélisation, CNRS, http://www.semantique-gdr.net/dico/.

CORBLIN F. (2002). *Représentation du discours et sémantique formelle.* Paris : PUF.

DESCLES J.-P., KYE-SEOP CH. (2006), Analyse critique de la notion de variable : points de vue sémiotique et formel, *Mathématiques et sciences humaines*, 43-102.

DOBROVIE-SORIN C., BEYSSADE C. (2004). *Définir les indéfinis.* Paris : CNRS Editions.

FODOR J., SAG I. (1982). Referential and Quantificational Indefinites. *Linguistics and Philosophy* 5, 355-398.

GILLON B. (1987). The readings of plural noun phrases. *Linguistics and Philosophy* 10:2, 199-219.

GILLON B. (1996). Collectivity and distributivity internal to English noun phrases. *Language Sciences*, 18:1-2, 443-468.

HINTIKKA J. (1976) Partially Ordered Quantifiers vs. Partially Ordered Ideas, *Dialectica* 30, 89-99.

KAHANE S. (2005). Structure des représentations logiques, polarisation et sous-spécification. Actes de *TALN*, Dourdan, 153-162.

KAMP H., REYLE U. (1993). *From Discourse to Logic*. Dordrecht : Kluwer.

KRIFKA M. (1990). Four thousand ships passed through the lock: object-induced measure functions on évents. *Linguistics and Philosophy* 13: 487-520, 1990.

LANDMAN F. (1989). Groups. *Linguistics and Philosophy* 12:5-6, 559-605 (Partie I), 723-744 (Partie II).

LASERSOHN P. (1989). On the Readings of Plural Noun Phrases. *Linguistic Inquiry* 20:1, 130-134.

LESMO L., ROBALDO L. (2004) Dependency Tree Semantics and Underspecification. Proc. of *International Conference On Natural language processing* (*ICON2004*), Hyderabad, India.

LINK G. (1997). Ten Years of Research on Plurals – Where Do We Stand? In F. Hamm, E. Hinrichs (éd.) *Plurality and Quantification*, Dordrecht: Kluwer, 19-54.

MEL'ČUK I. (1988). *Dependency Syntax: Theory and Practice*. Albany : SUNY Press.

MOSTOWSKI A. (1957). On a generalization of quantifiers. *Fundamenta Mathematica* 44, 12-36.

NICKEL B. (à par.). Plurals. In Graff-Fara D., Russell G. (éd.), *The Routledge Handbook for Philosophy of Language*. Manuscrit en ligne.

ROBALDO L. (2007) *Dependency Tree Semantics*. Ph. D Thesis, Université de Turin.

SCHA R. (1981). Distributive, Collective, and Cumulative Quantification. In J. Groenendijk, T. Janssen, M. Stokhof (éds), *Formal Methods in the Study of Language,* Amsterdam : University of Amsterdam, 483-512.

SCHWARZSCHILD R. (1996). *Pluralities*. Kluwer.

SHER G. (1990) Ways of Branching Quantifiers. *Linguistics and Philosophy* 13, 393-422.

STEEDMAN M. (2009). Surface-Compositional Scope-Alternation Without Existential Quantifiers. Manuscrit en ligne (draft 5.2).

SOWA J. (1987). Semantics networks. In S. C. Shapiro (éd.), *Encyclopedia of Artificial Intelligence*, New York : Wiley ; version mise à jour en ligne.

Discours

Analyse automatique de la modalité et du niveau de certitude : application au domaine médical

Delphine Bernhard[1] Anne-Laure Ligozat[1, 2]
(1) LIMSI-CNRS, B.P. 133, 91403 Orsay Cedex
(2) ENSIIE, 1 square de la résistance, 91000 Évry
bernhard@limsi.fr, annlor@limsi.fr

Résumé. De nombreux phénomènes linguistiques visent à exprimer le doute ou l'incertitude de l'énonciateur, ainsi que la subjectivité potentielle du point de vue. La prise en compte de ces informations sur le niveau de certitude est primordiale pour de nombreuses applications du traitement automatique des langues, en particulier l'extraction d'information dans le domaine médical. Dans cet article, nous présentons deux systèmes qui analysent automatiquement les niveaux de certitude associés à des problèmes médicaux mentionnés dans des compte-rendus cliniques en anglais. Le premier système procède par apprentissage supervisé et obtient une f-mesure de 0,93. Le second système utilise des règles décrivant des déclencheurs linguistiques spécifiques et obtient une f-mesure de 0,90.

Abstract. Many linguistic phenomena aim at expressing the speaker's doubt or incertainty, as well as the potential subjectivity of the point of view. Most natural language processing applications, and in particular knowledge extraction in the medical domain, need to take this type of information into account. In this article, we describe two systems which automatically analyse the levels of certainty associated with medical problems mentioned in English clinical reports. The first system uses supervised machine learning and obtains an f-measure of 0.93. The second system relies on a set of rules decribing specific linguistic triggers and reaches an f-measure of 0.90.

Mots-clés : Modalité épistémique, Niveau de certitude, Domaine médical.

Keywords: Epistemic modality, Certainty level, Medical domain.

1 Introduction

En traitement automatique des langues, les informations contenues dans les textes sont souvent considérées comme affirmées et vérifiées. Or, de nombreux phénomènes linguistiques visent à exprimer le doute ou l'incertitude de l'énonciateur, ainsi que la subjectivité potentielle du point de vue. Il y a ainsi une gradation dans les niveaux de certitude associés à une information : elle peut être vraie, possible ou fausse. Une information peut également n'être vraie que dans certaines conditions, ou être hypothétique.

Dans certains domaines, il est particulièrement important de savoir si l'information donnée dans un document est certaine ou pas. Par exemple dans le domaine médical, si l'on tente d'analyser des relations entre un médicament et des symptômes décrites dans un texte, il est nécessaire de savoir si le symptôme est présent ou pas, ou encore s'il est susceptible d'être développé par le patient. En questions-réponses, notamment sur des corpus de documents issus du web, le niveau de certitude d'une information peut également être utile : une réponse comme «La tour Eiffel est une tour de 327 mètres de hauteur» à la question «Quelle est la taille de la tour Eiffel» est plus précise que la réponse «Je pense que la tour Eiffel fait environ 300 mètres» et devra être considérée comme plus fiable.

Ces divers aspects ne sont encore que très rarement pris en compte dans les applications développées actuellement en traitement automatique des langues, même s'il existe des travaux récents visant à détecter l'incertitude, la modalité épistémique, la spéculation ou encore les opinions.

Dans cet article, nous nous intéressons à l'analyse automatique de la modalité et du niveau de certitude dans le domaine médical. Plus particulièrement, nous nous attachons à étudier ces phénomènes lorsqu'ils portent sur des problèmes médicaux (maladies, syndromes, virus, bactéries, symptômes) mentionnés dans des compte-rendus

cliniques en anglais afin de savoir si le problème est présent, absent, hypothétique, soumis à certaines conditions, ou encore associé à une personne différente du patient dont il est question. Nous utilisons pour ce faire des données produites et annotées dans le cadre du challenge international i2b2/VA 2010 [1].

L'article est organisé comme suit : nous détaillons dans un premier temps l'état de l'art (section 2). Puis nous décrivons les données utilisées (section 3). Dans la section 4, nous présentons nos deux systèmes et exposons les résultats de leur évaluation dans la section 5.

2 État de l'art

Les notions de modalité et de niveau de certitude couvrent des phénomènes linguistiques variés qu'il est nécessaire de prendre en compte pour effectuer une analyse sémantique profonde des textes. Après une mise au point terminologique, nous détaillons le traitement de ces problématiques pour la fouille d'opinion et l'extraction d'informations dans le domaine médical.

2.1 Mise au point terminologique

La modalité est définie par Lapaire & Rotgé (1998) comme « une prise de position concernant la valeur de vérité d'une proposition » et « la manière dont un sujet pensant et parlant se prononce sur un contenu propositionnel ». Dans cet article, nous nous intéressons plus particulièrement à la modalité épistémique qui traduit le niveau de certitude de l'énonciateur par rapport à l'énoncé. Cette notion est à rapprocher des procédés euphémistiques (en anglais, *hedging*) qui visent à atténuer ou à moduler la force d'une assertion.

De nombreux marqueurs textuels servent à exprimer le degré de certitude de l'énonciateur : adjectifs épistémiques (« possible », « impossible »), adverbes épistémiques (« probablement »), verbes (« savoir », « croire ») et locutions verbales (« il est possible que »), propositions conditionnelles, etc (Rubin *et al.*, 2006; Rubin, 2007; Saurí & Pustejovsky, 2009).

Rubin (2010) distingue cinq niveaux de certitude : absolue, élevée, modérée, faible et incertaine. À ces niveaux de certitude sont associées diverses dimensions contextuelles : *perspective* (point de vue de l'auteur / discours rapporté), *focus* (opinion / faits) et *temporalité* (passé, présent, futur, hors de propos ou ambigu). L'annotation manuelle de 2 243 phrases issues d'articles de journaux en anglais selon cette typologie a montré que plus de la moitié des phrases contiennent des marqueurs de modalité épistémique, avec 1 727 occurrences de marqueurs, classés dans 47 classes syntaxico-sémantiques (Rubin, 2010). Les marqueurs les plus fréquents sont les auxiliaires modaux (« must », « could »), les adjectifs superlatifs et les adverbes intensifieurs (« much », « so »).

Saurí & Pustejovsky (2009) définissent quant à eux les notions de fait avéré (*fact*) et fait non avéré (*counterfact*), modulés par un niveau de certitude (certain, probable, possible). Ces catégories ont été utilisées pour l'annotation du corpus anglais FactBank.

2.2 Fouille d'opinion

La fouille d'opinion vise à détecter de manière automatique les unités textuelles (mots, syntagmes, phrases ou textes) qui portent une marque de subjectivité et à déterminer leur polarité ou orientation (positive, négative ou mixte) (Pang & Lee, 2008). Les systèmes de fouille d'opinion utilisent souvent des lexiques affectifs regroupant des mots et expressions subjectifs, associés à leur polarité et éventuellement l'intensité de leur orientation. L'utilisation en contexte de tels lexiques nécessite la prise en compte des phénomènes linguistiques pouvant modifier l'orientation *a priori* des expressions subjectives. Ces phénomènes incluent notamment la modification adverbiale, conduisant à une modulation de l'intensité d'une opinion (« bon », « très bon »), voire son inversion par des adverbes de négation (« bien », « pas bien »). En pratique, les phénomènes de négation sont identifiés par diverses heuristiques. Ainsi, Pang *et al.* (2002) marquent les mots modifiés par une négation s'ils apparaissent entre un négateur et une marque de ponctuation. Na *et al.* (2005) utilisent quant à eux des patrons syntaxiques pour identifier de manière plus précise les expressions incluant un négateur.

1. https://www.i2b2.org/NLP/Relations/Main.php

L'analyse automatique de la subjectivité est également influencée par d'autres phénomènes, tels que la présence d'une proposition conditionnelle (Narayanan *et al.*, 2009). Une phrase conditionnelle peut contenir des expressions subjectives sans pour autant formuler une opinion (ex : « Si cet auteur écrit un bon livre, je l'achèterai »). D'autre part, l'opinion exprimée est souvent déterminée conjointement par la proposition conditionnelle et le conséquent (ex : « Si tu veux lire un bon livre, je te conseille celui-ci »). L'analyse de phrases conditionnelles en fouille d'opinion repose en général sur des critères spécifiques, tels que le temps des verbes ou la présence de connecteurs conditionnels.

2.3 Extraction d'information en domaine biomédical

En domaine biomédical, il est particulièrement important d'analyser la modalité et le niveau de certitude car les formes modalisatrices sont très largement utilisées dans les documents biomédicaux pour indiquer des impressions, des explications possibles ou des résultats négatifs (Vincze *et al.*, 2008), comme le montrent les exemples suivants :
– «These findings that **may be** from an acute pneumonia include minimal bronchiectasis as well.»
– «Stable appearance the right kidney **without** hydronephrosis.»
– «The treatment **seems to be** successful.»
– «Right upper lobe volume loss and **probably** pneumonia.»
Différents niveaux de certitude peuvent être distingués. Le niveau le plus traité est celui de la négation, mais une gradation plus fine peut également être utilisée avec, par exemple, des niveaux signalant une possibilité ou une condition. La détection du niveau de certitude peut se faire soit au niveau global de la phrase, soit à l'intérieur d'une phrase.

Concernant la négation, sa détection dans les documents biomédicaux a été largement étudiée. Chapman *et al.* (2001) ont proposé un algorithme appelé NegEx détectant les négations et identifiant les termes médicaux dans la portée d'une négation [2]. Le principe est le suivant : les déclencheurs de négation sont annotés dans la phrase (comme par exemple «no» ou «denies») et répartis en deux classes selon qu'ils apparaissent avant ou après les termes dans leur portée (les deux déclencheurs précédents apparaissent par exemple avant les termes qu'ils qualifient) ; des pseudo-déclencheurs sont également annotés, c'est-à-dire des termes contenant des déclencheurs de négation, mais qui n'en sont pas, comme «no increase». Puis les portées de ces déclencheurs sont définies : par défaut, la portée d'un déclencheur va de ce déclencheur à la fin (ou au début) de la phrase, mais peut être interrompue par la présence d'un autre déclencheur ou d'un déclencheur de fin de clause (comme par exemple «presenting», qui marque la fin de portée de «History» dans la phrase «History of COPD, presenting with shortness of breath»).

Mutalik *et al.* (2001) ont également proposé un outil de détection des négations dans des documents médicaux. Des marqueurs de négation sont repérés puis une grammaire associe ces marqueurs avec des concepts. Enfin, Huang & Lowe (2007) ont utilisé une approche mixte combinant des expressions régulières et une analyse syntaxique. Le traitement de la négation est également inclus dans certains systèmes plus généraux comme le système Medical Language Extraction and Encoding (MedLEE) (Friedman *et al.*, 1994).

D'autres travaux concernent de façon plus générale le traitement de la modalité. La première étude analysant ce problème d'un point de vue linguistique et informatique est celle de Light *et al.* (2004) ; les auteurs ont analysé les marqueurs de modalité dans des résumés d'articles scientifiques médicaux issus de PubMed [3], annoté manuellement des phrases selon la présence de modalisation et entraîné un classifieur à détecter de telles phrases dans des résumés d'articles. Plusieurs travaux se sont également intéressés à la classification de phrases en fonction de la présence de modalité. Medlock & Briscoe (2007) se fondent sur une approche faiblement supervisée pour cette classification, et utilisent uniquement les mots des phrases comme attributs. Leur classification obtient un point d'équilibre (*break-even point* ou BEP) entre rappel et précision de 0,76. Ils ont par ailleurs rendu public le corpus d'articles scientifiques biologiques qu'ils utilisent dans leurs travaux. Szarvas (2008) a étendu cette approche en ajoutant les bigrammes et trigrammes aux attributs du classifieur, en sélectionnant les attributs pour réduire le nombre de mots-clés candidats et en ajoutant des listes de mots-clés externes, pour obtenir un BEP de 0,85. Kilicoglu & Bergler (2008) soulignent le rôle de la syntaxe dans la détection des marqueurs de modalité, et ajoutent aux attributs des informations syntaxiques sur les relations de dépendance afin de tenir compte du contexte syntaxique des marqueurs.

2. Cet outil nous a servi de système de référence lors de notre évaluation.
3. `http://www.ncbi.nlm.nih.gov/pubmed`

Une limitation des ces travaux est qu'ils ne considèrent la modalité qu'au niveau des phrases, alors que plusieurs modalités peuvent en réalité être présentes dans une même phrase comme le notent par exemple Wilbur *et al.* (2006). Un niveau de détection plus précis consiste à identifier des marqueurs de modalité et leur portée. Ainsi, dans l'exemple «Right upper lobe volume loss and probably pneumonia.», «probably» est un marqueur de modalité dont la portée est le seul terme «pneumonia». Le corpus BioScope (Vincze *et al.*, 2008) a été utilisé pour évaluer la reconnaissance de telles informations, notamment dans le cadre de la tâche d'évaluation «Learning to detect hedges and their scope in natural language text» à CoNLL 2010 [4]. Morante & Daelemans (2009) par exemple ont considéré la détection des marqueurs et de leur portée comme un problème de classification supervisée, en entraînant et combinant plusieurs classifieurs. Ils obtiennent une f-mesure de 0,85 pour la détection des marqueurs, et de 0,66 pour la détection des portées de ces marqueurs.

Comme nous venons de le montrer, l'analyse de la modalité et du niveau de certitude a fait l'objet de divers travaux dans le domaine biomédical. Les méthodes utilisées sont soit à base de règles et de lexiques, soit à base d'apprentissage, lorsqu'il existe des données annotées en quantité suffisante. Dans cet article, nous décrivons et comparons de manière détaillée deux méthodes différentes pour une tâche précise, à savoir la classification d'assertions portant sur des problèmes médicaux en anglais.

3 Description des données

Nous nous sommes intéressées à la détection de modalité et du niveau de certitude intra-phrases dans le cadre de la campagne d'évaluation i2b2 2010 [5]. L'objectif d'une des tâches de cette campagne était la classification d'assertions relatives à des problèmes médicaux. Il s'agissait donc de donner une valeur de certitude ou de modalité à des concepts médicaux qui avaient été préalablement annotés dans des compte-rendus de patients.

Plutôt que de calculer la portée de chaque marqueur possible de modalité, l'accent est mis ici sur les concepts étudiés, pour lequels on cherche à connaître une modalité. Nous détaillerons dans cette section les données que nous avons utilisées pour nos expériences, puis présenterons dans la section suivante les deux systèmes d'analyse développés.

3.1 Corpus

Les corpus sont composés de 826 compte-rendus médicaux en langue anglaise provenant de trois hôpitaux américains. Ils ont préalablement été annotés en concepts [6], qui sont :

– les **problèmes médicaux**, définis comme les observations concernant le corps ou l'esprit du patient et considérées comme anormales ou provoquées par une maladie, ce qui englobe les maladies, syndromes, virus, bactéries, symptômes...
– les **traitements**, regroupant les procédures, interventions ou médicaments donnés au patient pour résoudre un problème médical ;
– les **tests**, c'est-à-dire les procédures ou mesures faites sur le patient pour trouver des informations sur un problème médical.

Ces compte-rendus ont également été anonymisés, afin de supprimer les noms de personne, de lieu, les adresses, et autres informations géographiques précises, les dates relatives à un individu comme une date de naissance, les numéros de téléphone et fax, les courriels, les numéros de sécurité sociale etc. Dans les documents, toutes ces informations sont remplacées par des balises génériques comme «NAME», «ADDRESS», ou encore «AGE».

3.2 Catégories d'assertions

Une classe d'assertion a été attribuée à chaque problème médical du corpus. Six catégories d'assertions ont été définies :

4. http://www.inf.u-szeged.hu/rgai/conll2010st/
5. https://www.i2b2.org/NLP/Relations/Main.php
6. Pour une méthode d'annotation automatique des concepts, voir (Minard *et al.*, 2011).

– **présent** : le problème associé au patient est présent. C'est également la catégorie par défaut pour les problèmes médicaux qui n'ont pu être classés dans une autre catégorie.
  Exemples : «*the wound* was noted to be clean with *mild serious drainage*», «history of *chest pain*»
– **absent** : le patient ne présente pas ce problème. Cette catégorie inclut les problèmes résolus.
  Exemples : «patient denies *pain*», «history inconsistent with *stroke*»
– **possible** : le patient peut avoir le problème, mais ce n'est pas certain.
  Exemples : «this is very likely to be *an asthma exacerbation*», «We suspect this is not *pneumonia*»
– **conditionnel** : le patient ne rencontre le problème que dans certaines conditions, comprenant en particulier les allergies.
  Exemples : «Penicillin causes *a rash*», «patient reports *shortness of breath* upon climbing stairs»
– **hypothétique** : le patient est susceptible de développer le problème dans le futur.
  Exemples : «if you experience *wheezing* or *shortness of breath*», «ativan 0.25 to 0.5 mg IV q 4 to 6 hours prn *anxiety*»
– **associé à quelqu'un d'autre** : le problème concerne quelqu'un d'autre que le patient.
  Exemples : «Family history of *prostate cancer*», «brother had *asthma*»

Catégorie d'assertion	Corpus d'entraînement	Corpus d'évaluation
Présent	8 052	13 025
Absent	2 536	3 609
Possible	535	883
Conditionnel	103	171
Hypothétique	651	717
Associé à qqun d'autre	92	145
Total	11 969	18 550

TABLE 1 – Nombre d'assertions par catégorie dans les corpus d'entraînement et d'évaluation

La table 1 détaille le nombre d'assertions annotées pour chaque catégorie dans les corpus d'entraînement et d'évaluation. La majorité des problèmes médicaux évoqués dans les documents sont *présents* : 67% des assertions appartiennent à cette catégorie. La répartition des diverses catégories est donc très déséquilibrée.

Les organisateurs d'i2b2 ont calculé l'accord inter-annotateur sur le corpus d'entraînement, annoté par 12 annotateurs. Cette information, présentée dans la table 2, est intéressante car elle montre que pour certaines catégories, les annotateurs ont fait des choix d'annotation relativement différents. L'accord inter-annotateurs sur l'ensemble des catégories est élevé, mais pour certaines catégories notamment «conditionnel», il est très faible. On peut donc supposer que les résultats obtenus par des systèmes automatiques seront également plus faibles pour ces catégories.

Catégorie d'assertion	Accord
Présent	0,89
Absent	0,94
Possible	0,60
Conditionnel	0,44
Hypothétique	0,79
Associé à qqun d'autre	0,89
Total	0,91

TABLE 2 – Accords inter-annotateurs pour chaque catégorie d'assertion

4 Description des systèmes

Dans cette section, nous décrivons tout d'abord les systèmes de référence (ou *baseline*) et les résultats qu'ils obtiennent. Puis nous détaillons les systèmes que nous avons développés afin d'identifier de manière automatique la catégorie d'assertion associée à un problème médical.

4.1 Références

Nous avons considéré plusieurs références pour évaluer l'apport de nos systèmes :
– la première méthode de référence consiste à tout étiqueter comme « présent », ce qui donne une f-mesure totale de 0,67 (proportion de présents dans le corpus) ;
– la deuxième référence consiste à utiliser le système NegEx [7] sans aucune modification, qui n'annote donc que les catégories «présent» et «absent», ce qui donne une f-mesure de 0,83 ;
– la troisième référence consiste à utiliser le système ConText (extension de NegEx pour tenir compte de la catégorie associé à quelqu'un d'autre) sans aucune modification, donc sans les catégories «possible», «conditionnel» et «hypothétique», ce qui donne une f-mesure de 0,86.

4.2 Prétraitements

4.2.1 Annotation des concepts

Nous avions à notre disposition l'annotation en concepts faite manuellement par les organisateurs d'i2b2. Nous avons donc annoté les documents en balisant les termes qui correspondaient à un concept médical reconnu, c'est-à-dire à un problème, un traitement ou un test.

4.2.2 Traitement des coordinations

L'étude des données de développement a montré que de nombreux concepts sont coordonnés à l'aide de virgules ou de conjonctions de coordination, comme par exemple « pleural effusion or pneumothorax ». Nos deux systèmes utilisent les contextes directs des concepts, sous forme de fenêtre de mots, afin d'identifier la catégorie de l'assertion. La présence de séquences de concepts coordonnés peut donc conduire à la prise en compte d'un contexte gauche ou droit réduit, comportant principalement d'autres problèmes coordonnés. Dans ce cas, des indices essentiels pour l'identification de la catégorie de l'assertion peuvent se trouver en dehors de la fenêtre contextuelle considérée. Le rôle important de la coordination a également été démontré pour une autre tâche d'extraction d'information, l'extraction d'événements (Kilicoglu & Bergler, 2009). Nous avons donc pré-traîté les données pour identifier les séquences de concepts coordonnés et ainsi redéfinir les fenêtres contextuelles utilisées : les fenêtres gauches se terminent au début d'une séquence de concepts coordonnés tandis que les fenêtres droites débutent à la fin d'une séquence de concepts. Ces fenêtres contextuelles sont partagées par tous les concepts qui apparaissent dans la même séquence. Plus de la moitié des concepts étaient inclus dans une coordination de concepts dans le corpus d'entraînement.

4.3 Système par apprentissage supervisé

Le système par apprentissage supervisé considère l'identification d'assertions comme une tâche de classification. Nous avons entraîné une machine à vecteurs de support (SVM) avec libsvm (Chang & Lin, 2001) sur la base d'attributs binaires et d'un noyau RBF [8]. Les paramètres optimaux ont été sélectionnés automatiquement par validation croisée en 5 sous-ensembles [9]. Pour le développement du système, nous avons utilisé un corpus d'entraînement de 241 fichiers et un corpus de test de 54 fichiers.

Nous avons utilisé quatre types d'attributs :
– **attributs contextuels lexicalisés** : mots et mots désuffixés dans une fenêtre de cinq mots à gauche et à droite du concept cible. Nous avons également effectué des expériences complémentaires avec les étiquettes morpho-syntaxiques mais celles-ci n'apportent pas d'améliorations significatives ;

7. `http://code.google.com/p/negex/`
8. Il aurait également été possible d'utiliser les CRF (Conditional Random Fields) afin d'étiqueter de manière plus précise les déclencheurs et leur portée. Toutefois, ces annotations n'étaients pas fournies dans le corpus d'apprentissage et il aurait donc été nécessaire d'utiliser un corpus externe, comme Bioscope par exemple (Vincze *et al.*, 2008). Or, seules les catégories de négation et de spéculation sont annotées dans Bioscope. Nous avons donc renoncé à cette possibilité.
9. Cette étape a été réalisée de manière automatique en utilisant le script `easy.py` fourni avec libsvm.

– **déclencheurs** : nous avons utilisé les déclencheurs définis pour le système à base de règles (voir section suivante), avec quelques déclencheurs supplémentaires. Ces déclencheurs sont identifiés dans une fenêtre de cinq mots à gauche et à droite du concept cible. Nous avons également identifié quelques déclencheurs internes au concept, tels que « on exertion » (« à l'effort ») qui indique la catégorie conditionnel ;
– **attributs internes au concept cible** : mots et mots désuffixés formant le concept et présence du préfixe « non » dans un des mots ;
– **attributs spécifiques à une séquence de concepts coordonnés.** Dans le cas où le concept cible apparaît dans une séquence de concepts coordonnés, nous utilisons des attributs spécifiques qui correpondent aux mots et aux mots désuffixés de la séquence. Par exemple, pour la séquence « pleural effusion **or** pneumothorax », les mots « pneumothorax », « pleural » et « effusion » ainsi que la racine « effus » sont utilisés comme attributs.

4.4 Système à base de règles

L'algorithme utilisé est très proche de celui de NegEx (voir section 2.3) mais est étendu aux six catégories d'assertions à détecter. Quatre types de déclencheurs ont été définis (les exemples suivants sont donnés pour la catégorie «absent») :
– ceux qui précèdent le problème comme «denies», «never had» ou «negative for» ;
– ceux qui suivent le problème comme «was ruled out», «is absent» ou «was stopped» ;
– ceux qui sont inclus dans le problème comme «afebrile» (dans le cas, le déclencheur est le problème lui-même) ou «allergy» ;
– ceux qui limitent la portée d'un des déclencheurs précédents, comme «but».
Un même terme peut être de plusieurs types : ainsi, «ruled out» peut précéder ou suivre le problème associé. Les termes de «pseudo-négation» de NegEx (qui ressemblent à des termes de négation mais n'en sont pas comme la double négation «not ruled out») ne constituent pas un type distinct dans notre système, mais sont éliminés en vérifiant le contexte des termes de négation (par exemple on vérifiera que le mot «not» ne précède pas «ruled out»).

Les déclencheurs ont d'abord été définis manuellement par une étude de corpus, puis les listes ont été complétées grâce aux résultats du système par apprentissage : les attributs les plus utiles à la classification ont été étudiés, et éventuellement ajoutés aux listes. Le nombre moyen de déclencheurs par catégorie d'assertion varie d'une cinquantaine pour les déclencheurs précédant le problème à une quizaine pour les déclencheurs inclus dans le problème.

Trois expressions régulières sont utilisées pour rechercher des déclencheurs avant, après ou à l'intérieur des problèmes, sachant que des mots non déclencheurs peuvent être autorisés entre le déclencheur et le problème (le nombre de ces mots a été fixé par une étude du corpus d'entraînement pour chaque catégorie d'assertion) :
– <déclencheur> <mots non déclencheurs> {0,n} <problème> ;
– <problème> <mots non déclencheurs> {0,n} <déclencheur> ;
– <problème> <déclencheur> >.
Ainsi, dans la phrase «Also with *multiple masses consistent with* metastasis», «consistent with» est un déclencheur de la catégorie «possible» suivi directement du problème «metastasis». La première expression régulière sera donc déclenchée, et donnera la catégorie «possible» à ce problème.

Ces listes de déclencheurs et expressions régulières ont été implémentés avec l'outil WMatch (Rosset *et al.*, 2008; Galibert, 2009) [10] qui présente une vitesse d'exécution bien supérieure à celle de NegEx.

Une priorité a été attribuée à chaque catégorie d'assertion, en fonction du guide d'annotation i2b2 : associé à quelqu'un d'autre > absent > possible > conditionnel > hypothétique.

5 Évaluation et discussion

Le corpus d'évaluation comporte 18 550 assertions (voir table 1). Les résultats ont été évalués en termes de rappel, précision et f-mesure pour chaque catégorie d'assertion, ainsi que pour toutes les catégories :

10. Moteur d'expressions régulières développé au LIMSI, disponible sur demande.

$$\text{rappel} = \frac{\#\text{problèmes correctement attribués à la catégorie d'assertion i}}{\#\text{problèmes de la catégorie d'assertion i}}$$

$$\text{précision} = \frac{\#\text{problèmes correctement attribués à la catégorie d'assertion i}}{\#\text{problèmes attribués à la catégorie d'assertion i}}$$

$$\text{f-mesure} = \frac{2*\text{rappel}*\text{précision}}{\text{rappel}+\text{précision}}$$

5.1 Résultats de l'évaluation

Les résultats obtenus par les deux systèmes sur le corpus d'évaluation [11] sont détaillés dans la table 3.

	Apprentissage			Règles		
Catégorie	**Rappel**	**Précision**	**F-mesure**	**Rappel**	**Précision**	**F-mesure**
Toutes	0,93	0,93	0,93	0,90	0,90	0,90
Présent	0,97	0,94	0,96	0,95	0,92	0,93
Absent	0,95	0,93	0,94	0,85	0,93	0,89
Possible	0,54	0,74	0,62	0,57	0,61	0,59
Hypothétique	0,83	0,93	0,88	0,74	0,86	0,80
Conditionnel	0,24	0,74	0,36	0,27	0,29	0,28
Associé à quelqu'un d'autre	0,78	0,86	0,82	0,95	0,76	0,84

TABLE 3 – Résultats pour le corpus d'évaluation

Les deux systèmes développés obtiennent des résultats significativement [12] meilleurs que les trois systèmes de référence. Le système par apprentissage supervisé obtient de très bons résultats, et s'est classé 5e sur 21 participants à i2b2 2010.

Si l'on compare plus en détails les deux systèmes, on constate que le système par apprentissage a tendance à privilégier la précision sur le rappel. La précision qu'il obtient est généralement supérieure ou égale à celle du système à base de règles. En outre, le système par apprentissage se caractérise par une f-mesure supérieure à celle du système à base de règles, sauf pour la catégorie «associé à quelqu'un d'autre». Dans ce dernier cas, les déclencheurs utilisés par le système à base de règles permettent d'obtenir un très bon rappel de 0,95. Cette comparaison démontre la complémentarité des deux systèmes, qui gagneraient à être combinés.

Lors du challenge i2b2, le meilleur score à cette tâche a été obtenu par l'équipe du National Research Council Canada, avec une f-mesure de 0,9362 (contre 0,9311 pour notre système par apprentissage). Les méthodes des meilleurs systèmes sont assez proches de celle utilisée par notre système par apprentissage. Les SVM sont généralement utilisés pour la classification, avec des attributs portant sur les mots, l'annotation sémantique des mots, l'étiquetage morpho-syntaxique ou encore des déclencheurs spécifiques. On trouve également des attributs différents, qu'ils serait intéressant d'intégrer aux futures versions de notre système : n-grammes de caractères, attributs spécifiques à la phrase tels que le temps du verbe principal ou encore attributs spécifiques au document considéré (longueur).

5.2 Analyse des erreurs

La table 4 détaille la matrice de confusion pour le système par apprentissage. Le système présente une tendance à sur-annoter les catégories «présent» et «absent». Ce résultat était prévisible dans la mesure où il s'agit également des catégories les plus représentées dans le corpus d'apprentissage. La catégorie «présent» est également celle qui compte le plus grand nombre de faux positifs. Pour les classes plus rares dans le corpus d'apprentissage, et notamment la classe «conditionnel», le système se caractérise par un rappel réduit, corrélé à une bonne précision, de 0,74 à 0,93 en fonction de la classe considérée.

11. Ces résultats sont très légèrement différents de ceux obtenus à l'évaluation i2b2 pour le système par règles, du fait d'un changement de système d'annotation : lors de l'évaluation, nous avions utilisé GenConText, alors que les résultats présentés dans cet article sont obtenus avec WMatch.

12. p<0,05 selon le test de Student

		Apprentissage						
		Présent	**Absent**	**Possible**	**Cond.**	**Hypo.**	**Assoc. autre**	**Total**
Référence	**Présent**	12629	199	134	13	34	16	13025
	Absent	168	3418	20	0	2	1	3609
	Possible	381	16	475	1	10	0	883
	Cond.	117	12	1	41	0	0	171
	Hypo.	94	12	14	0	595	2	717
	Assoc. autre	16	16	0	0	0	113	145
	Total	13405	3673	644	55	641	132	18550

TABLE 4 – Matrice de confusion pour le système à base d'apprentissage

La matrice de confusion de la table 5 présente les erreurs de catégorisation détaillées du système à base de règles. Les confusions les plus fréquentes concernent principalement les catégories «présent» et «absent» ; le système a également tendance à surannoter en «possible». Enfin, le très bon rappel sur la catégorie «associé à quelqu'un d'autre» se traduit par très peu de confusion dans cette catégorie.

		Règles						
		Présent	**Absent**	**Possible**	**Cond.**	**Hypo.**	**Assoc. autre**	**Total**
Référence	**Présent**	12352	167	285	110	73	38	13025
	Absent	514	3077	9	2	1	6	3609
	Possible	331	34	505	3	10	0	883
	Cond.	117	6	1	47	0	0	171
	Hypo.	152	9	23	2	531	0	717
	Assoc. autre	5	2	0	0	0	138	145
	Total	13471	3295	823	164	615	182	18550

TABLE 5 – Matrice de confusion pour le système à base de règles

Les erreurs du système à base de règles ont de multiples causes (dans les exemples suivants, les problèmes médicaux à analyser sont remplacés par le mot PROBLEM) :
– les listes de déclencheurs sont incomplètes : ainsi, pour la catégorie «associé à quelqu'un d'autre», les cas de dons d'organes anonymes n'avaient pas été pris en compte («The liver was from a gentleman who had died from PROBLEM» [13]) et donc les déclencheurs associés («gentleman» ici) n'étaient pas dans les listes ;
– le déclencheur peut être trop loin du problème pour que les deux soient associés, notamment dans le cas d'anaphores ; ainsi pour la phrase «The patient declined the procedure stating that her mother had mitral valve regurgitation and she lived for many years without PROBLEM», le déclencheur «her mother» de la catégorie «associé à quelqu'un d'autre» est trop loin du problème à analyser ; l'utilisation d'un outil de résolution d'anaphore pourrait ici être utile ;
– un déclencheur est détecté, mais il n'est en réalité pas relié au problème, notamment lorsque l'avis de la famille est évoqué («According to the family») ; un contexte gauche au déclencheur «family» pourrait ici être précisé pour éliminer ces cas, mais les formulations plus complexes comme «The family is confident in any decision the doctor would make concerning her PROBLEM» sont plus difficiles à traiter ;
– la portée du marqueur est trop importante ou trop faible : dans la phrase «The patient is on no specific medications for his PROBLEM», le marqueur «no» est repéré, mais ne s'applique en réalité pas au problème ;
– deux marqueurs sont détectés, et l'ambigüité est mal résolue, comme par exemple dans la phrase «PROBLEM was less likely given her negative angiogram.» où «likely» est bien déclencheur de la catégorie «possible», mais le déclencheur «negative» est favorisé du fait des priorités entre catégories ;
– quelques erreurs sont liées à des problèmes d'anonymisation des données, qui perturbent l'application des règles ;
– quelques incohérences d'annotation ont également été notées, ce que prévoyait l'accord inter-annotateur.

Enfin, la matrice de confusion de la table 6 permet de comparer les annotations réalisées par les deux systèmes.

13. Les exemples présentés sont issus d'exemples réels, mais sont généralement simplifiés pour être plus lisibles.

		Règles						
		Présent	**Absent**	**Possible**	**Cond.**	**Hypo.**	**Assoc. autre**	**Total**
Apprentissage	**Présent**	12711	125	312	129	87	41	13405
	Absent	497	3142	11	2	2	19	3673
	Possible	129	22	491	0	2	0	644
	Cond.	22	1	1	31	0	0	55
	Hypo.	99	5	8	2	524	3	641
	Assoc. autre	13	0	0	0	0	119	132
	Total	13471	3295	823	164	615	182	18550

TABLE 6 – Matrice de confusion pour les deux systèmes

Les systèmes sont fortement en désaccord pour la catégorie «conditionnel» : seules 31 annotations sont partagées. C'est également la catégorie qui obtient les moins bons résultats globalement. La deuxième catégorie la plus problématique correspond aux assertions du type «possible». Le système par apprentissage tend à sous-annoter cette catégorie et par conséquent de nombreuses assertions annotées comme «possible» par le système à base de règles sont associées à la catégorie «présent» par le système par apprentissage.

6 Conclusion et perspectives

Nous nous sommes intéressées à l'analyse de la modalité et du niveau de certitude dans des textes médicaux, et avons développé deux systèmes dans le cadre de la campagne d'évaluation i2b2 2010. Le premier procède par apprentissage supervisé et s'appuie sur des attributs lexicaux, morphologiques et sémantiques. Le second est un système à base de listes et de règles. Ces deux systèmes obtiennent des résultats significativement meilleurs que les trois références considérées, et au niveau de l'état de l'art : environ 0,93 et 0,90 de f-mesure.

Les perspectives de poursuite de ces travaux sont nombreuses. Ces systèmes pourraient notamment être étendus afin de travailler éventuellement sur plusieurs phrases, et de tenir compte des phénomènes discursifs. Il serait également souhaitable de combiner les deux systèmes, en particulier pour profiter du très bon niveau de rappel du système à base de règles pour la catégorie «associé à quelqu'un d'autre».

Nous envisageons également d'adapter les systèmes présentés au français. Pour le système à base de règles, il sera nécessaire d'identifier les déclencheurs correspondant ou aditionnels pour le français. Le système à base d'apprentissage requiert quant à lui de larges quantités de données annotées. Dans la mesure où ceci constitue une tâche nécessitant des annotateurs experts humains, nous nous appuierons sur la méthode de l'apprentissage actif (*active learning*).

Nous souhaitons également généraliser ces travaux à d'autres domaines, afin d'analyser la modalité et les niveaux de certitude dans des contextes plus larges. En particulier, nous envisageons l'étude des phénomènes similaires dans les textes journalistiques et, plus globalement, les textes subjectifs visant à exprimer des opinions.

Remerciements

Ce travail a été partiellement financé par le projet Quæro (financement Oseo, agence française pour l'innovation et la recherche). Les données médicales utilisées proviennent du consortium Informatics for Integrating Biology to the Bedside (i2b2) grâce aux financements numéros U54LM008748 de la National Library of Medicine, VA HSR HIR 08-374 du Consortium for Healthcare Informatics Research (CHIR), et VA HSR HIR 08-204 du VA Informatics and Computing Infrastructure (VINCI).

Références

CHANG C.-C. & LIN C.-J. (2001). *LIBSVM : a library for support vector machines*. Outil disponible à l'adresse

suivante : http://www.csie.ntu.edu.tw/~cjlin/libsvm.

CHAPMAN W., BRIDEWELL W., HANBURY P., COOPER G. & BUCHANAN B. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, **34**(5), 301–310.

FRIEDMAN C., ALDERSON P. O., AUSTIN J. H., CIMINO J. J. & JOHNSON S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, **1**(2), 161.

GALIBERT O. (2009). *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. PhD thesis, Thèse de doctorat en informatique, Université Paris-Sud 11, Orsay, France.

HUANG Y. & LOWE H. (2007). A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. *Journal of the American Medical Informatics Association*, **14**, 304–311.

KILICOGLU H. & BERGLER S. (2008). Recognizing speculative language in biomedical research articles : a linguistically motivated perspective. *BMC Bioinformatics*, **9**(Suppl 11).

KILICOGLU H. & BERGLER S. (2009). Syntactic dependency based heuristics for biological event extraction. In *BioNLP '09 : Proceedings of the Workshop on BioNLP*, p. 119–127.

LAPAIRE J.-R. & ROTGÉ W. (1998). *Linguistique et grammaire de l'anglais, 3e édition*. Amphi 7. Presses Universitaires du Mirail.

LIGHT M., QIU X. & SRINIVASAN P. (2004). The language of bioscience : Facts, speculations, and statements in between. In *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases : tools for users*, p. 17–24.

MEDLOCK B. & BRISCOE T. (2007). Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 992–999, Prague, Czech Republic.

MINARD A.-L., LIGOZAT A.-L., BEN ABACHA A., BERNHARD D., CARTONI B., DELÉGER L., GRAU B., ROSSET S., ZWEIGENBAUM P. & GROUIN C. (2011). Hybrid methods for improving information access in clinical documents : Concept, assertion, and relation identification. *Journal of the American Medical Informatics Association*. À paraître.

MORANTE R. & DAELEMANS W. (2009). Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on BioNLP*, p. 28–36.

MUTALIK P., DESHPANDE A. & NADKARNI P. (2001). Use of general-purpose negation detection to augment concept indexing of medical documents. *Journal of the American Medical Informatics Association*, **8**(6), 598.

NA J.-C., KHOO C. & WU P. H. J. (2005). Use of negation phrases in automatic sentiment classification of product reviews. *Library Collections, Acquisitions, and Technical Services*, **29**(2), 180 – 191.

NARAYANAN R., LIU B. & CHOUDHARY A. (2009). Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 1 of *EMNLP '09*, p. 180–189, Stroudsburg, PA, USA.

PANG B. & LEE L. (2008). *Opinion mining and sentiment analysis*, volume 2 of *Foundations and Trends in Information Retrieval*. Now Publishers Inc.

PANG B., LEE L. & VAITHYANATHAN S. (2002). Thumbs up ? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 79–86.

ROSSET S., GALIBERT O., BERNARD G., BILINSKI E. & ADDA G. (2008). The LIMSI participation to the QAst track. In *Working Notes of CLEF 2008 Workshop*, Aarhus, Danemark.

RUBIN V., LIDDY E. & KANDO N. (2006). Certainty Identification in Texts : Categorization Model and Manual Tagging Results. In W. B. CROFT, J. SHANAHAN, Y. QU & J. WIEBE, Eds., *Computing Attitude and Affect in Text : Theory and Applications*, volume 20 of *The Information Retrieval Series*, p. 61–76. Springer Netherlands.

RUBIN V. L. (2007). Stating with Certainty or Stating with Doubt : Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. In *Proceedings of the Human Language Technologies Conference : The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007) ; Companion Volume, Short Papers*, p. 141–144.

RUBIN V. L. (2010). Epistemic modality : From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management*, **46**(5), 533 – 540.

SAURÍ R. & PUSTEJOVSKY J. (2009). FactBank : a corpus annotated with event factuality. *Language Resources and Evaluation*, **43**, 227–268.

SZARVAS G. (2008). Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of ACL-08 : HLT*, p. 281 – 289, Colombus, Ohio, USA.

VINCZE V., SZARVAS G., FARKAS R., MORA G. & CSIRIK J. (2008). The bioscope corpus : biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, **9**(Suppl 11), S9.

WILBUR W., RZHETSKY A. & SHATKAY H. (2006). New directions in biomedical text annotation : definitions, guidelines and corpus construction. *BMC bioinformatics*, **7**(1), 356.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Analyse discursive et informations de factivité

Laurence Danlos
ALPAGE, Université Paris Diderot (Paris 7), 175 rue du Chevaleret, 750013 Paris
Laurence.Danlos@linguist.jussieu.fr

Résumé. Les annotations discursives proposées dans le cadre de théories discursives comme RST (Rhetorical Structure Theory) ou SDRT (Segmented Dicourse Representation Theory) ont comme point fort de construire une structure discursive globale liant toutes les informations données dans un texte. Les annotations discursives proposées dans le PDTB (Penn Discourse Tree Bank) ont comme point fort d'identifier la "source" de chaque information du texte — répondant ainsi à la question qui a dit ou pense quoi ? Nous proposons une approche unifiée pour les annotations discursives alliant les points forts de ces deux courants de recherche. Cette approche unifiée repose crucialement sur des information de factivité, telles que celles qui sont annotées dans le corpus (anglais) FactBank.

Abstract. Discursive annotations proposed in theories of discourse such as RST (Rhetorical Structure Theory) or SDRT (Segmented Representation Theory Dicourse) have the advatange of building a global discourse structure linking all the information in a text. Discursive annotations proposed in PDTB (Penn Discourse Tree Bank) have the advatange of identifying the "source" of each information — thereby answering to questions such as who says or thinks what ? We propose a unified approach for discursive annotations combining the strengths of these two streams of research. This unified approach relies crucially on factivity information, as encoded in the English corpus FactBank.

Mots-clés : Discours, Analyse discursive, Factivité (véracité), Interface syntaxe-sémantique, RST, SDRT, PDTB, FactBank.

Keywords: Discourse, Discursive analysis, Factuality (vericity), Syntax-semantic interface, RST, SDRT, PDTB, FactBank.

1 Introduction

L'analyse discursive d'un texte s'effectue généralement en deux étapes : la première consiste à segmenter le texte en "unités de discours élémentaires" (EDU, Elementary Discourse Unit), la seconde consiste à construire la "structure du discours ", cette structure reposant sur les "relations de discours" ("relations rhétoriques") qui relient deux segments de discours en spécifiant le rôle d'un segment par rapport à l'autre, spécifiant par là-même l'intention communicative de l'auteur du texte. Un segment de discours est soit une EDU soit un segment complexe groupant plusieurs EDU avec récursivement leur structure discursive. C'est cette approche de l'analyse discursive qui est adoptée dans les deux principales théories du discours, RST (Rhetorical Structure Theory, (Mann & Thompson, 1988; Taboada & Mann, 2006)) et SDRT (Segmented Dicourse Representation Theory, (Asher, 1993; Asher & Lascarides, 2003)) pour lesquelles des corpus ont été annotés manuellement, RST-corpus pour RST en anglais (Carlson *et al.*, 2003), et ANNODIS pour SDRT en français (Péry Woodley *et al.*, 2009).

Parallèlement à ces travaux, des applications récentes du TAL comme la détection d'opinion ou les systèmes de question/réponse ont fait surgir le besoin de savoir qui pense quoi ou qui a dit quoi. Ceci nécessite en premier lieu de pouvoir identifier la "source" d'une information se trouvant dans un texte : est-elle attribuée à l'auteur du texte (le "locuteur") ou à une autre personne mentionnée dans le texte ? De plus, il faut pouvoir déterminer si une information concernant un événement (une éventualité) présente cet événement comme correspondant à une situation du monde ou comme une simple possibilité ou hypothèse ; en termes techniques, il faut pouvoir déterminer la "factivité des événements". Ces deux aspects sont intrinsèquement liés dans la mesure où la factivité d'un événement peut être évaluée différemment, par exemple, par le locuteur et par une source autre que le locuteur. Ainsi, dans *Fred a dit que Jane était la plus belle*, la source de l'information "Jane est la plus belle"

est Fred, la source de l'information que Fred a émis des propos est le locuteur. L'information subjective "Jane est la plus belle" est prise en charge par Fred qui est prêt à défendre son point de vue tandis qu'elle n'est pas prise en charge par le locuteur (voir la notion de "commitment" (Hamblin, 1970)). Ce courant de recherche a donné lieu à des corpus annotés pour la factivité, principalement FactBank pour l'anglais (Saurí, 2008; Saurí & Pustejovsky, 2009) (Section 3).

Le PDTB (Penn Discourse Tree Bank, (PDTB Group, 2008)) est un corpus annoté pour l'anglais qui allie tant des informations sur la structure discursive du texte que des informations sur la source et le degré de factivité des éventualités présentées dans le texte. C'est donc un effort notable pour aller vers une compréhension profonde d'un texte. Le travail présenté dans cet article s'inscrit dans la lignée du PDTB : il vise à jeter les bases théoriques d'un manuel pour annoter tant la structure discursive que les informations de factivité (relativement à une source) d'un texte français. Il se départit néanmoins du PDTB sur les points suivants :

– Le PDTB ne repose pas sur une théorie du discours mais sur un formalisme D-LTAG (Discourse Lexicalized Tree Adjoining Grammar, (Webber, 2004; Forbes-Riley *et al.*, 2006) qui étend un formalisme d'analyse syntaxique (LTAG) au discours. L'objectif du PDTB n'est pas d'annoter la structure discursive globale d'un texte. L'objectif principal est d'annoter pour chaque connecteur de discours explicite (lexicalement marqué) et pour certains (mais pas tous) connecteurs implicites (phonologiquement vides) leurs arguments (Section 2). A l'inverse, notre travail s'inscrit dans la perspective d'annoter une analyse discursive complète d'un texte comme en RST ou SDRT.
– Dans le PDTB, les informations de factivité sont annotées après que les arguments des connecteurs de discours ont été identifiés. Nous montrerons à l'inverse que les informations de factivité doivent être annotées en premier car elles sont primordiales pour déterminer les arguments des connecteurs (Section 4.1).
– Dans le PDTB, les informations sur les sources d'un connecteur de discours et de ses arguments sont aussi annotées après que les arguments de la relation de discours ont été identifiés et avant les informations de factivité. Nous montrerons à nouveau que les informations de factivité doivent être annotées en premier car elles sont primordiales pour déterminer la source des relations de discours et par là-même la source de leurs arguments (Section 4.2).

Pour présenter les différentes positions théoriques que nous défendons, nous nous concentrons dans cet article sur des discours dont au moins une phrase (la première ou la seconde) est de forme *N0$_{hum}$ V W que P = : Fred a dit à Marie que Zoé allait venir pour Noël*, c'est-à-dire une phrase complexe contenant un segment *N0$_{hum}$ V W* dont la tête est un verbe *V* qui sous-catégorise une complétive *P*. Le symbole *N0* désigne le sujet de *V* qui est considéré ici comme humain, *W* désigne un ensemble éventuellement vide de compléments (en incluant les clitiques) ou d'ajouts (en incluant les particules négatives).

La Section 2 discute de la relation de discours $Attribution$ qui est utilisée en RST et SDRT et que nous adoptons pour relier *N0$_{hum}$ V W* à la complétive *P* dans une phrase de forme *N0$_{hum}$ V W que P*. La relation $Attribution$ n'étant pas utilisée dans le PDTB, nous présentons dans cette section les choix alternatifs effectués dans le PDTB. La Section 3 présente les informations de factivité telles qu'elle son annotées dans le corpus FactBank et que nous adoptons. La Section 4 met en avant les positions théoriques que nous défendons en examinant les analyses discursives des discours dont la première phrase est de forme *N0$_{hum}$ V W que P* (Section 4.1) et de ceux dont la seconde phrase est de cette forme (Section 4.2). La Section 5 montre qu'il n'est pas nécessaire d'annoter la source des arguments des relations de discours : ces informations peuvent être déduites des annotations sur la source des relations de discours (lorsqu'on a recours à la relation $Attribution$). La Section 6 conclut en présentant des perspectives pour une annotation de corpus (français) suivant les principes théoriques défendus dans cet article.

2 Relation de discours $Attribution$

Considérons le discours (1) qui enchaîne deux phrases à complétive. Les deux complétives (soulignées) sont liées par la relation $Résultat$ marquée par le connecteur adverbial *du coup*. Par conséquent, il faut que ces deux complétives soient considérées comme des EDU.

(1) Fred a dit à Marie que Zoé allait venir pour Noël. Ensuite, Marc a ajouté que du coup Sue ne viendrait pas.

Il y a donc consensus pour considérer que, dans un discours indirect de forme *N0$_{hum}$ V W que P*, la complétive *P* forme un EDU. Il en est de même pour les discours directs (*"Zoé va venir pour Noël", a dit Fred à Marie*) où

la citation forme un EDU. Par contre, les positions divergent sur la question de savoir si le segment attributif — la séquence $N0_{hum}$ *V W* dans un discours indirect ou l'incise de citation dans un discours direct — forme un EDU ou non. Examinons les différentes positions.

– En RST, il est posé que le segment attributif forme un EDU lié au segment attribué par la relation $Attribution$ (Wolf & Gibson, 2006; Redeker & Egg, 2006). Cette relation de discours n'est pas standard puisque le segment attributif ne forme pas un segment autonome (syntaxiquement et sémantiquement) et que $Attribution$ ne joue aucun rôle rhétorique.
– La SDRT adopte la relation $Attribution$ comme la RST mais avec une différence de taille : le segment attributif est considéré comme complet (Hunter *et al.*, 2006). En effet, il est posé qu'il contient une variable existentiellement quantifiée correspondant à l'argument phrastique du verbe du segment attributif, cette variable étant une anaphore liée par (le contenu sémantique de) le segment attribué. Sans entrer dans des détails trop techniques, retenons que cette position revient à analyser un discours indirect tel que *Fred a dit à Marie que Zoé allait venir* comme *Fred l'a dit à Marie, que Zoé allait venir* et un discours direct tel que *"Zoé va venir pour Noël", a dit Fred* comme *"Zoé va venir pour Noël", Fred l'a dit*.
La différence d'utilisation de $Attribution$ en RST et SDRT est illustrée sur l'exemple (2a) avec segmentation en EDU : (2b) présente l'analyse discursive en RST, (2c) celle en SDRT. Ces deux analyses s'accordent sur le point suivant : le premier argument de $Résultat$ n'est pas le segment 2 ; ceci est exclu car la cause du segment 3 *il l'a énervée* construit autour du verbe causatif *énerver* ne peut être qu'un acte de Fred (Pustejovsky, 1995; Danlos, 2000). Les analyses (2b) et (2c) ne diffèrent que par le contenu sémantique qui est éventuellement donné au premier argument de $Attribution$: non existant en RST qui pose que le premier argument de $Résultat$ est le segment complexe formé par l'ensemble segment attributif et segment attribué, noté $[1, 2]$; en revanche, SDRT considère que le premier argument de $Résultat$ est le segment 1 dont le contenu sémantique est considéré comme équivalent à *Fred l'a dit à Marie*.

(2) a. (Fred a dit à Marie)$_1$ que (Zoé allait venir pour Noël)$_2$. (Il l'a énervée)$_3$.
b. $Attribution(1, 2) \wedge Résultat([1, 2], 3)$
c. $Attribution(1, 2) \wedge Résultat(1, 3)$

Par contre, les deux théories proposent la même analyse pour (3a), voir (3b).

(3) a. (Fred est très énervé en ce moment)$_1$. (Jane dit)$_2$ qu'(il est gravement malade)$_3$.
b. $Attribution(2, 3) \wedge Explication(1, 3)$

– Dans le PDTB, la relation $Attribution$ n'est pas utilisée car ce n'est pas une relation de discours standard (Prasad *et al.*, 2006). Les annotations effectuées dans le PDTB sont détaillées ci-dessous, mais, en faisant abstraction du format particulier utilisé, elles sont présentées en (4) et (5) pour (2a) et (3a) respectivement. En (4), les arguments de $Résultat$ correspondent aux deux phrases. En (5), le second argument de $Explication$ est le segment 2 (qui comporte un trait spécifiant son segment attributif *Jane dit que*, cf ci-dessous).

(4) a. (Fred a dit à Marie que Zoé allait venir pour Noël)$_1$. (Il l'a énervée)$_2$.
b. $Résultat(1, 2)$

(5) a. (Fred est très énervé en ce moment)$_1$. Jane dit qu'(il est gravement malade)$_2$.
b. $Explication(1, 2)$

Les analyses proposées en RST et dans le PDTB peuvent être considérées comme équivalentes : nous laissons au lecteur le soin de se convaincre que les analyses (2b) et (4b) ou (3b) et (5b) reviennent fondamentalement au même, ne différant que par l'utilisation ou non de $Attribution$. Par contre, la position prise en SDRT est radicalement différente : les analyses (2b) à la RST et (2c) à la SDRT divergent sur le premier argument de $Résultat$. Nous écartons la position de SDRT pour deux raisons : la première est que l'analyse proposée en (2c) est contre-intuitive et peut donc amener à des erreurs d'annotation ; la seconde est qu'elle n'est pas linguistiquement justifiée pour des discours directs [1] dont l'incise de citation a pour tête un verbe comme *mentir* qui ne permet pas de discours indirect, (6a). Il est donc linguistiquement injustifié d'analyser (6b) comme **"Zoé va venir pour Noël", Fred le mentit* car la citation en (6b) ne correspond pas à un argument du verbe *mentir*. Celui-ci — comme plus de 500 autres verbes — demande une entrée lexicale spéciale pour prendre en compte son emploi comme tête d'une incise de citation (Danlos *et al.*, 2010).

(6) a. *Fred a menti que Zoé allait venir pour Noël.
b. "Zoé va venir pour Noël", mentit Fred.

1. Rappelons que les positions prises en RST, SDRT et dans le PDTB s'appliquent tant aux discours indirects qu'aux discours directs.

Entre les positions quasiment équivalentes prises en RST et dans le PDTB, nous préférons celle prise en RST qui a recours à $Attribution$. En effet, le recours à $Attribution$ permet de construire la structure discursive globale d'un texte en intégrant tout fragment de texte, y compris tout segment attributif [2]. En résumé, nous adoptons la position prise en RST et donc des analyses comme (2b) et (3b) [3].

La relation $Attribution$ est utilisée en RST (ou SDRT) pour relier le segment attributif *N0*$_{hum}$ *V W* et la complétive lorsque *V* est un verbe de discours rapporté comme *dire* mais aussi lorsque *V* est un verbe d'attitude propositionnelle comme *craindre, croire, douter* ou *réaliser*. De ce fait, $Attribution$ ne prend pas en compte les différences sémantiques entre tous ces verbes (leur factivité, par exemple) ni d'ailleurs le fait que le verbe soit sous la portée d'une négation. Il faut donc avoir recours à d'autres mécanismes pour prendre en compte ces différentes propriétés du segment attributif qui influent sur l'analyse discursive. De tels mécanismes seront présentés à la Section 3.

Auparavant, examinons la solution adoptée dans le PDTB où la relation $Attribution$ n'est pas utilisée. Pour chaque connecteur de discours explicite, sont annotés non seulement ses arguments mais aussi la source de la relation de discours marquée par le connecteur et la source de chacun de ses arguments. Ainsi, pour l'exemple (7a) de (Prasad *et al.*, 2006) où le connecteur *while* est souligné, ses argument Arg1 et Arg2 sont repérés par les segments de texte respectivement en italiques et en gras. Le segment attributif, *purchasing agent said*, placé dans une boîte, ne fait partie d'aucun de ces arguments. Ces annotations sont complétées par le tableau en (7b) qui indique la valeur du trait [Source] pour les sources de la relation REL (marquée par *while* et identifiée comme étant $Contraste$), de Arg1 et de Arg2. La valeur "Wr" est utilisée pour l'auteur ("writer") du texte, "Inh" indique que la valeur de [Source] est héritée de celle de REL, "Ot" ("other") est utilisée pour un (ou des) individu(s) autre(s) que l'auteur (il s'agit des purchasing agents pour Arg2). Comme la valeur du trait [Source] de Arg2 est "Ot", cet argument possède un trait qui spécifie son segment attributif, ici *purchasing agent said*. Le tableau comporte d'autres informations relatives à la factivité et la polarité, les traits [Type], [Polarity] et [Determinacy], mais nous ne les détaillerons pas ici, non pas parce que nous les considérons comme non pertinentes mais parce que, d'une part, nous préférons celles mises au point dans FactBank qui sont plus sophistiquées (Section 3) et que, d'autre part, nous pensons que les informations de factivité doivent être déterminées **avant** d'établir quels sont les arguments d'un connecteur de discours (Section 4.1) et quelle est sa source (Section 4.2).

(7) a. *Factory orders and construction outlays were largely flat in December* while purchasing agents said **manufacturing shrank further in October**.

b.

	REL	Arg1	Arg2
[Source]	Wr	Inh	Ot
[Type]	Comm	Null	Comm
[Polarity]	Null	Null	Null
[Determinacy]	Null	Null	Null

Le PDTB apporte donc un grand soin à des annotations permettant de déterminer qui a dit quoi, ce qui est effectivement une question importante pour de nombreuses applications du TAL. Nous nous inspirons de leur approche pour cet aspect de l'annotation discursive. Par contre, dans le PDTB, seuls les arguments de certains connecteurs implicites (phonologiquement vides) sont annotés. Ainsi sont prises en compte les relations de discours qui doivent être inférées entre deux phrases adjacentes juxtaposées à l'intérieur d'un même paragraphe, mais pas celles entre deux phrases adjacentes séparées par une marque de paragraphe (PDTB Group, 2008). Il n'y a donc pas annotation de la structure discursive globale du texte, alors que c'est l'ambition affichée dans le RST-corpus (Carlson *et al.*, 2003) et ANNODIS (Péry Woodley *et al.*, 2009) [4]. De plus, rien ne garantit qu'on puisse déduire la structure discursive globale d'un paragraphe à partir des annotations faites sur les arguments des connecteurs (explicites ou implicites) figurant à l'intérieur de ce paragraphe. Le problème ne vient pas des segments attributifs qui peuvent

2. Signalons toutefois que le recours à $Attribution$ pose un léger problème technique lorsque des compléments ou ajouts du segment attributif apparaissent après la complétive, voir *Fred a prévenu Marie que Zoé allait venir pour Noël par fax et par e-mail*. De tels exemples seront ignorés dans cet article.

3. La relation $Attribution$ est considérée par tous les auteurs qui l'utilisent comme satellite-nucleus (dans les termes de la RST) ou subordonnante (dans les termes de la SDRT). Néanmoins, en RST, les auteurs ne sont pas d'accord sur le nucleus de cette relation subordonnante : (Wolf & Gibson, 2006) considèrent que c'est le segment attributif tandis que (Redeker & Egg, 2006) considèrent que c'est le segment attribué. En SDRT, il est posé que c'est parfois l'un parfois l'autre, selon par exemple l'emploi intensionnel versus évidentiel de *dire*, voir (2) versus (3). Cette position de la SDRT nous paraît tout à fait justifiée, mais, faute de place, nous ne discuterons pas de la question du nucleus de $Attribution$ dans cet article.

4. L'annotation effectuée dans ANNODIS pour le français est donc nettement plus ambitieuse que celle effectuée dans le PDTB. Mais elle ne concerne que 4000 relations de discours contre plus de 40 000 dans le PDTB.

être facilement intégrés à l'analyse discursive en faisant appel à la relation $Attribution$ telle qu'utilisée en RST. Il vient de ce qu'un fragment de texte qui n'est intégré dans aucun argument d'aucun connecteur du paragraphe est tout bonnement et simplement ignoré. En d'autres termes, l'approche du PDTB — identifier les connecteurs explicites et certains connecteurs implicites puis annoter leurs arguments — est orthogonale à l'approche incrémentale préconisée en RST ou SDRT — identifier un nouveau segment de discours et l'intégrer à l'analyse discursive déjà construite. Nous nous inspirons de cette approche incrémentale, qui s'inscrit dans le courant de la sémantique dynamique, pour cet aspect de l'analyse discursive visant à construire la structure discursive globale d'un texte.

En conclusion, comme préconisé en RST (ou SDRT), il semble justifier d'utiliser la relation $Attribution$ tout en ayant recours à d'autres mécanismes pour indiquer les propriétés sémantiques de son premier argument, un segment attributif. Néanmoins, en RST ou SDRT, la question de savoir quelles sont les sources d'une relation de discours et de ses arguments est à tort passée sous silence. Nous considérons qu'il doit être inscrit dans les objectifs de l'analyse discursive de déterminer quelles sont ces sources. Plus précisément, nous considérons que l'analyse discursive doit indiquer la source de chaque relation de discours ; pour indiquer qu'une relation de discours R est attribuée à la source s_j, nous utilisons la notation R_{s_j} dans l'analyse discursive. Pour les arguments d'une relation de discours, nous montrerons à la Section 5 qu'il n'est pas nécessaire d'annoter leur source car celle-ci peut être déduite des informations contenues dans la structure discursive, à savoir les sources des relations de discours et la présence de relation(s) $Attribution$.

Une des difficultés rencontrées dans l'analyse discursive pour des exemples mettant en jeu $Attribution$ est de savoir si c'est le segment attribué ou l'ensemble segment attributif et segment attribué qui est argument d'une relation de discours donnée, voir le contraste entre (2b) et (3b). C'est à cette difficulté que nous allons nous attaquer, entre autres, dans la Section 4. Auparavant, présentions les informations de factivité annotées dans le corpus FactBank.

3 Présentation de FactBank

FactBank est un corpus (anglais) annoté pour la factivité événementielle. Pour chaque événement (éventualité) e_i d'une phrase, les informations de factivité sont données relativement à une source donnée, soit l'auteur soit un (ou des) individu(s) à qui est attribuée les informations concernant e_i. Par exemple, dans une construction à complétive de forme *N0*$_{hum}$ *V W que P*, les informations de factivité concernant l'événement décrit dans la complétive *P* sont données par rapport à l'auteur et au référent de *N0* (si différent de l'auteur). D'une manière générale, une information de factivité est de la forme $f(e_i, s_j) = x$, où s_j désigne l'auteur ou la source des informations concernant e_i [5]. La valeur x d'une information de factivité est une paire $Mod(x)Pol(x)$ contenant une valeur de modalité et une valeur de polarité. Les valeurs de modalité sont au nombre de quatre : certain (CT), probable (PR), possible (PS), non-spécifié (U), avec une relation d'ordre entre ces valeurs. Les valeurs de polarité sont au nombre de trois : positive (+), négative (-) et non-spécifié (u) [6]. A titre d'illustration, pour la phrase en (8) avec sa segmentation en deux EDU, les informations de factivité à la FactBank sont données en (i). L'événement noté e_i est l'événement décrit dans le segment i sans les éventuelles informations de modalité ou de polarité négative ; par exemple, pour le segment 2, e_2 correspond à *Fred ira à Dax*. Pour e_1, la seule source pertinente est l'auteur qui affirme que cet événement est vrai (l'auteur a asserté la phrase donc le segment 1), d'où $f(e_1, auteur) = CT+$. Pour e_2, les informations de factivité sont évaluées relativement à l'auteur et à Jane. Concernant l'auteur, la sémantique du verbe *penser* implique que l'auteur ne s'engage pas [7], ce qui veut dire qu'il ne se prononce pas ou ne veut pas se prononcer sur le contenu propositionnel de *P*, d'où $f(e_2, auteur) = Uu$. Concernant Jane, *penser* implique que Jane juge le contenu propositionnel de *P* comme possible (PS) ; comme *P* en (8) est sous une polarité négative, $f(e_2, Jane) = PS-$.

(8) (Jane pense)$_1$ que (Fred n'ira pas à Dax)$_2$.
(i) $f(e_1, auteur) = CT+ \wedge f(e_2, auteur) = Uu \wedge f(e_2, Jane) = PS-$

5. En fait, il peut y avoir plusieurs sources quand, par exemple, il y a enchâssement d'une complétive dans une complétive, voir *Luc pense que Zoé a dit que P*. Néanmoins, nous laissons de côté pour l'instant ces cas complexes.

6. Les paires $Mod(x)Pol(x)$ sont cependant au nombre de 8 et non de 12 car certaines combinaisons de modalité et polarité ne font pas sens, par exemple $U+$ ou $U-$

7. Nous utilisons *s'engager* comme traduction de *commit* communément utilisé dans la littérature anglophone.

FactBank n'est pas concerné par les relations de discours et par l'analyse discursive : les annotations de factivité sont effectuées phrase par phrase, sans tenir compte du contexte discursif ni des connaissances du monde, même celles concernant "l'autorité morale" d'une source d'information. Ces annotations proviennent uniquement de connaissances en sémantique lexicale, en particulier des propriétés sémantiques des verbes à complétive, et de la présence de marqueurs de modalité (épistémique) ou de polarité, l'interaction entre ces différents facteurs ayant été modélisée. Nous allons montrer que dans une approche de sémantique dynamique, la mise à jour de l'analyse discursive par un nouveau segment de discours doit se dérouler en trois étapes :

(i) détermination des informations de factivité événementielle dans le nouveau segment sur la seule base de connaissances linguistiques,

(ii) mise à jour de la structure discursive en s'appuyant d'une part sur ces informations de factivité événementielle d'autre part sur d'autres connaissances linguistiques, des connaissances pragmatiques et des connaissances du monde ; cette mise à jour doit intégrer l'identification de la source des relations de discours introduites dans la structure discursive,

(iii) révision et/ou complétion des informations de factivité événementielle en fonction de l'analyse discursive.

FactBank n'étant pas concerné par les relations discursive, celles-ci ne sont pas annotées d'informations de factivité. Néanmoins, il est nécessaire de disposer de telles annotations, qui existent d'ailleurs dans le PDTB (Section 2). Nous ajoutons donc dans les tâches de l'annotation de l'analyse discursive l'annotation des informations de factivité concernant les relations de discours (tâche qui doit être effectuée à l'étape ii).

4 Analyse discursive de discours comportant $Attribution$

Nous allons examiner l'analyse discursive de discours comprenant une phrase à complétive de forme *N0*$_{hum}$ *V W que P*, mettant donc en jeu une relation $Attribution$. Dans un premier temps, nous discutons du cas où la phrase à complétive apparaît à l'initiale du discours et est suivie d'une phrase simple (i.e. sans relation $Attribution$) introduite par un connecteur de discours *Conn* éventuellement précédé d'un signe de ponctuation (*Ponct*), soit des discours de forme *N0*$_{hum}$ *V W que P (Ponct) Conn P'*. Nous montrerons que les informations de factivité sont primordiales pour déterminer l'argument gauche de *Conn*. Dans un second temps, nous discutons du cas où la phrase à complétive apparaît après une autre phrase (avec ou sans relation $Attribution$), soit des discours de forme *P. (Conn) N0*$_{hum}$ *V W que (Conn) P'*. Nous montrerons que les informations de factivité sont primordiales pour identifier la source de la relation de discours marquée par *Conn*.

Avant d'entrer dans le vif du sujet, présentons nos conventions de segmentation en EDU, qui s'harmonisent avec celles du PDTB. Un connecteur de discours, qu'il soit de type adverbial ou conjonction, n'est pas considéré comme faisant partie d'une EDU. Cette convention n'est pas celle adoptée dans le projet ANNODIS (Péry Woodley *et al.*, 2009) où un connecteur est intégré à l'EDU correspondant à sa phrase hôte. Ce choix effectué dans ANNODIS préjuge de la portée sémantique des connecteurs. Ainsi, pour le discours en (9a) — dont il sera longuement question à la Section 4.2 —, la segmentation d'ANNODIS donnée en (9b) préjuge à tort que *ensuite* a portée sémantique sur *Jane croit*, contrairement à la nôtre donnée en (9c).

(9) a. Fred ira à Dax pour Noël. Ensuite, Jane croit qu'il ira à Pau.
b. (Fred ira à Dax pour Noël)$_1$. (Ensuite, Jane croit)$_2$ (qu'il ira à Pau)$_3$.
c. (Fred ira à Dax pour Noël)$_1$. Ensuite, (Jane croit)$_2$ qu'(il ira à Pau)$_3$.

Lorsqu'un connecteur adverbial ne se trouve pas en tête de l'EDU qui correspond à sa phrase hôte, comme dans l'exemple (10a) où *ensuite* se trouve au milieu du noyau verbal de sa phrase hôte, il est possible d'avoir recours à une "forme normalisée de discours" (Danlos, 2009) qui fait abstraction de la position de l'adverbial — tout en gardant une trace de cette position car il existe de cas où elle induit une différence de sens (Bras, 2008). La segmentation de (10a) est alors celle donnée en (10b).

(10) a. Fred ira à Dax pour Noël. Il ira ensuite à Pau.
b. (Fred ira à Dax pour Noël)$_1$. ensuite^{interne} (Il ira à Pau)$_2$.

4.1 *Attribution* dans la première phrase (*N0$_{hum}$ V W que P (Ponct) Conn P'*)

Pour ne pas introduire de bruit dans l'analyse des données, nous allons nous concentrer principalement sur des discours *Fred V W qu'il détestait les grévistes (Ponct) Conn il est syndicaliste*, en nous contentant de faire varier *V W*, soit les propriétés du segment attributif, et de faire varier *Conn* entre la conjonction de subordination *alors que* et l'adverbial *pourtant*, ces deux connecteurs marquant la relation $Contraste$. Nous allons montrer qu'il existe trois interprétations pour ces discours de forme *N0$_{hum}$ V W que P (Ponct) Conn P'*.

Dans la première interprétation, observée uniquement avec *Conn = alors que*, soit une conjonction de subordination et non un connecteur adverbial, le segment attribué est *P alors que P'*, ce qui se traduit en discours direct par *"P alors que P'", V N0 W*, voir (11b). La citation de (11b) ou le discours rapporté de (11a) met en jeu une "violation d'attente", à savoir l'attente $syndicaliste(x) > \neg détesterLesGrévistes(x)$. C'est cette violation d'attente qui légitime la relation $Contraste$ marquée lexicalement par *alors que* et dont la source est Fred. Avec cette interprétation, l'analyse discursive de (11a) est celle donnée en (11c). Les informations de factivité sont données en (11d). La relation $Contraste$ posée par Fred doit être évaluée par rapport à l'auteur : la sémantique de *dire* implique que l'auteur ne s'engage pas sur les propos de Fred, soit $f(Contraste_{Fred}(2,3), auteur) = Uu$, ce qui implique *a priori* $f(e_2, auteur) = Uu \wedge f(e_3, auteur) = Uu$. Par contre, $f(Contraste_{Fred}(2,3), Fred) = CT+$ implique *a priori* $f(e_2, Fred) = CT+ \wedge f(e_3, Fred) = CT+$.

(11) a. (Fred a dit)$_1$ qu'(il détestait les grévistes)$_2$ alors qu'(il est syndicaliste)$_3$.
b. "Je déteste les grévistes alors que je suis syndicaliste", a dit Fred.
c. $Attribution_{auteur}(1, [2,3]) \wedge Contraste_{Fred}(2,3)$
d. $f(Contraste_{Fred}(2,3), auteur) = Uu \wedge f(e_2, auteur) = Uu \wedge f(e_3, auteur) = Uu$
$f(Contraste_{Fred}(2,3), Fred) = CT+ \wedge f(e_2, Fred) = CT+ \wedge f(e_3, Fred) = CT+$.

Nous écartons désormais cette interprétation, en ne retenant pour les exemples suivants que celles mettant en jeu $Attribution_{auteur}(1,2)$, interprétations privilégiées avec *Conn = pourtant*. Il reste alors deux possibilités pour l'argument gauche de $Contraste$: $Contraste_{auteur}(2,3)$ ou $Contraste_{auteur}([1,2],3)$ avec $Attribution(1,2)$. Nous allons montrer que le choix entre l'une ou l'autre de ces possibilités dépend des informations de factivité venant du segment attributif *N0 V W*.

Commençons par examiner des exemples où e_2 et e_3 mettent en jeu une violation d'attente, e.g. (Fred déteste les grévistes)$_2$ et (Fred est syndicaliste)$_3$. Considérons (12a) construit autour du verbe *réaliser* qualifié de "factif" dans la littérature. Les informations à la FactBank sont données en (12b). L'auteur croit en la véracité de $e2$ et $e3$ qui mettent en jeu la violation d'attente $syndicaliste(x) > \neg détesterLesGrévistes(x)$. Il peut donc poser la relation $Contraste$ entre les segments 2 et 3, ce qui débouche sur l'analyse donnée en (12c). L'intention communicative de l'auteur est de montrer que Fred est incohérent avec lui-même.

(12) a. (Fred a réalisé)$_1$ qu'(il détestait les grévistes)$_2$. Pourtant (il est syndicaliste)$_3$.
b. $f(e_1, auteur) = CT+$
$f(e_2, Fred) = CT+ \wedge f(e_2, auteur) = CT+$
$f(e_3, auteur) = CT+$
c. $Attribution_{auteur}(1,2) \wedge Contraste_{auteur}(2,3)$

Passons maintenant à (13a) construit autour de *prétendre* qui est un "factif négatif", voir les informations de factivité concernant e_2 en (13b) où $f(e_2, auteur) = CT-$ indique que d'après l'auteur ce n'est certainement pas le cas que Fred déteste les grévistes. L'auteur ne peut donc pas poser qu'il y a violation de l'attente $syndicaliste(x) > \neg détesterLesGrévistes(x)$. Il ne peut donc poser qu'un contraste entre le fait que Fred a prétendu qu'il détestait les grévistes et e3, ce qui débouche sur l'analyse donnée en (13c). L'intention communicative de l'auteur est de montrer que Fred a menti.

(13) a. (Fred a prétendu)$_1$ qu'(il détestait les grévistes)$_2$. Pourtant (il est syndicaliste)$_3$.
b. $f(e_2, Fred) = CT+ \wedge f(e_2, auteur) = CT-$
c. $Attribution_{auteur}(1,2) \wedge Contraste_{auteur}([1,2],3)$

Finalement, considérons (14a) construit autour de *dire* avec les informations de factivité pour $e2$ données en (14b) : l'auteur ne se prononce pas sur la véracité de $e2$. Le discours (14a) ne suffit pas à lui-seul pour déterminer l'opinion

de l'auteur, qui peut d'ailleurs être hésitant comme en témoigne le fait que (14a) peut être prolongé par une phrase qui indique explicitement son hésitation, (14c). Nous considérons donc (14a) comme ambigu, ambiguïté traduite dans l'analyse discursive en (14d) comportant une disjonction. Cette disjonction se traduit par les informations de factivité données en (14e).

(14) a. (Fred a dit)$_1$ qu'(il détestait les grévistes)$_2$. Pourtant (il est syndicaliste)$_3$.
b. $f(e_2, Fred) = CT+ \wedge f(e_2, auteur) = Uu$
c. Fred a dit qu'il détestait les grévistes. Pourtant il est syndicaliste. Soit il est incohérent avec lui-même soit il a menti.
d. $Attribution_{auteur}(1,2) \wedge (Contraste_{auteur}([1,2],3) \vee Contraste_{auteur}(2,3))$
e. $f(Contraste_{auteur}([1,2],3), auteur) = PS+ \wedge f(Contraste_{auteur}(2,3)), auteur) = PS+$

Tournons-nous rapidement vers des exemples où le segment 2 est le contraire de 3, e.g. (il pleuvait)$_2$ et (il ne pleuvait pas)$_3$. Comme un individu ne peut pas croire simultanément en une chose et son contraire[8], on a $POL(f(e_2, s_j)) = -POL(f(e_3, s_j))$ avec $-u = u$. Cette règle explique l'incohérence de *#Fred a réalisé qu'il pleuvait. Pourtant, il ne pleuvait pas.* construit avec le verbe factif *réaliser*. Cette règle explique aussi que le discours en (15a) construit avec *dire* est non ambigu, contrairement au discours en (14a) construit aussi avec *dire*. En effet, l'auteur ne pouvant poser $Contraste(2,3)$, il pose $Contraste([1,2],3)$, voir l'analyse en (15b) qui indique que l'intention communicative de l'auteur est de montrer que Fred a menti.

(15) a. (Fred a dit)$_1$ qu'(il pleuvait)$_2$. Pourtant (il ne pleuvait pas)$_3$.
b. $Attribution_{auteur}(1,2) \wedge Contraste_{auteur}([1,2],3)$

Cette discussion sur les analyses discursives des discours de forme *(N0 V W)*$_1$ *que (P)*$_2$ *(Ponct) Conn (P')*$_3$ amène les remarques suivantes respectivement sur le PDTB et SDRT :

1) PDTB : Tout annotateur "naïf" se reposant sur son intuition immédiate se jetterait probablement sur la violation d'attente $syndicaliste(x) > \neg détesterLesGrévistes(x)$ pour poser systématiquement $Contraste(2,3)$ dans les exemples (12)-(14). Or cette analyse est fausse pour (13) et (14). Ce point est à relier au suivant : dans le PDTB, les arguments Arg1 et Arg2 d'une relation de discours REL sont annotés avant les informations de factivité (qui sont ajoutées sur REL, Arg1 et Arg2, voir le tableau en (7b)). Les exemples (12)-(14) montrent qu'à l'inverse il faut annoter les informations de factivité **avant** d'annoter les arguments d'une relation de discours donnée.

2) SDRT : En SDRT, la relation $Contraste$ est considérée comme "véridique", ce qui est défini dans (Asher & Lascarides, 2003) comme une relation impliquant la véracité (du contenu propositionnel) de ces deux arguments. Néanmoins, les discours mis en avant par ces auteurs ne sont que des discours simples dans la mesure où ils ne mettent jamais en jeu la relation $Attribution$. En d'autres termes, dans ces discours, tout est attribué à l'auteur, les relations de discours et leurs arguments, ce qui permet d'ignorer le fait que les informations de véracité (factivité) doivent être évaluées relativement à plusieurs sources. Il faut donc réviser la notion de relation de discours véridique en prenant en compte l'évaluation des informations de factivité relativement à plusieurs sources (Danlos, 2011).

4.2 $Attribution$ dans la seconde phrase (*P. (Conn) N0*$_{hum}$ *V W que (Conn) P'.*)

Pour ne pas introduire de bruit dans l'analyse des données, nous allons nous concentrer uniquement sur des discours avec $Conn = ensuite$, ce connecteur marquant la relation $Narration$ de succession temporelle entre deux événements. Considérons d'abord des phrases *Ensuite, N0 V W que P* (avec une relation $Attribution$) qui sont énoncées dans un contexte gauche comprenant aussi une relation $Attribution$, par exemple (16a) (le contexte gauche, i.e. *P*, est mis en italiques). Dans cet exemple, *ensuite* a portée sur *elle a cru* : c'est un ajout syntaxique et sémantique sur le noyau verbal, qui peut être déplacé à l'intérieur de celui-ci sans changement de sens, (16b), mais qui ne peut pas être déplacé à l'intérieur la complétive sans induire un changement de sens important, (16c).

(16) a. *Jane a (d'abord) cru que Fred irait à Dax pour Noël.* Ensuite, elle a cru qu'il irait à Pau.
b. = *Jane a (d'abord) cru que Fred irait à Dax pour Noël*. Elle a ensuite cru qu'il irait à Pau.

8. Citons (Prabhakaran *et al.*, 2010) : "We cannot both believe something and not believe it : *#John won't be here, but nevertheless I think he may be here.*"

c. $\neq$ *Jane a (d'abord) cru que Fred irait à Dax pour Noël.* Elle a cru qu'ensuite il irait à Pau.

Le discours (16a) décrit la succession temporelle de croyances de Jane. En suivant sa segmentation donnée en (17a), il doit donc être analysé avec $Narration([1,2],[3,4])$. La relation $Narration$ est posée par l'auteur, tout comme les deux relations $Attribution$. Au total, l'analyse discursive de (16a) est celle présentée en (17b).

(17) a. *(Jane a (d'abord) cru)*$_1$ *que (Fred irait à Dax pour Noël)*$_2$. Ensuite, (elle a cru)$_3$ qu'(il irait à Pau)$_4$.
b. $Attribution_{auteur}(1,2) \wedge Attribution_{auteur}(3,4) \wedge Narration([1,2],[3,4])$

Considérons maintenant des phrases ***Ensuite, N0 V W que P*** qui sont énoncées dans un contexte gauche ne comprenant pas de relation $Attribution$, par exemple (18a). Dans cet exemple, le déplacement de *ensuite* à droite de *croit* débouche sur une incohérence, (18b), mais le déplacement de *ensuite* à l'intérieur de la complétive n'induit pas de différence de sens majeur, (18c).

(18) a. *Fred ira à Dax pour Noël.* Ensuite, Jane croit qu'il ira à Pau.
b. # *Fred ira à Dax pour Noël.* Jane croit ensuite qu'il ira à Pau.
c. $=$ *Fred ira à Dax pour Noël.* Jane croit qu'ensuite il ira à Pau.

L'analyse syntaxique de (18a) pose problème pour l'interface syntaxe-sémantique mais nous n'aborderons pas ce sujet ici (Dinesh *et al.*, 2005). En (18a) et (18c), l'événement du segment 1 (le voyage de Fred à Dax) est présenté comme précédant dans le temps l'événement du segment 3 (le voyage de Fred à Pau), on a donc $Narration(1,3)$. Il reste à examiner quelle est la source de $Narration$. Commençons par un exemple comme (18c) où *ensuite* se situe dans la complétive. Une source, que ce soit l'auteur ou non, ne peut poser de relation de succession temporelle entre deux événements que si cette personne est au courant des deux événements en jeu, ou tout du moins si elle pense qu'ils se sont passés (ou vont se passer) certainement, probablement ou peut-être [9]. Examinons les conséquences de cette affirmation pour (18c), répété en (19a) avec segmentation en EDU et dont les informations de factivité concernant e_3 sont données en (19b). L'auteur ne peut pas poser une relation de succession temporelle entre e_1 et e_3 car il ne s'engage pas sur le statut factuel de e_3. En revanche, Jane peut poser une telle relation car elle juge e_3 probable. L'analyse discursive est donc celle donnée en (19c) où la source de la relation $Narration$ est Jane. L'auteur ne prend pas en charge cette relation, (19d). Il peut d'ailleurs la mettre en question dans une troisième phrase, voir (19e), qui amène à une révision de la valeur de factivité donnée en (19d). Par ailleurs, les informations de factivité événementielle ne disent rien sur $f(e_1, Jane)$ puisque la source de 1 est l'auteur, mais l'analyse discursive avec $Narration_{Jane}(1,3)$ amène à poser $f(e_1, Jane) = CT+$ (ou à la rigueur $PR+$ou $PS+$). Ces données montrent que les informations de factivité événementielle doivent être révisées ou complétées après l'analyse discursive.

(19) a. *(Fred ira à Dax pour Noël)*$_1$. (Jane croit)$_2$ qu'ensuite (il ira à Pau)$_3$.
b. $f(e_3, Jane) = PR+ \wedge f(e_3, auteur) = Uu$
c. $Attribution_{auteur}(2,3) \wedge Narration_{Jane}(1,3)$
d. $f(Narration_{Jane}(1,3), auteur) = Uu$
e. *Fred ira à Dax voir sa mère.* Jane croit qu'ensuite il ira à Pau voir son père. *Mais je pense qu'il ira voir son père avant d'aller voir sa mère.*

Passons à (20a) dont les informations factuelles pour e_3 sont données en (20b). La mère de Fred n'étant pas au courant de l'événement e_3, elle ne peut pas poser de succession temporelle entre $e1$ et $e3$. La relation de succession temporelle ne peut donc être posée que par l'auteur, voir l'analyse en (20c).

(20) a. (Fred ira à Dax pour Noël voir sa mère)$_1$. (Celle-ci ne sait pas)$_2$ qu'ensuite (il ira à Pau voir son père)$_3$.
b. $f(e_3, MèredeFred) = Uu \wedge f(e_3, auteur) = CT+$
c. $Attribution_{auteur}(2,3) \wedge Narration_{auteur}(1,3)$

L'exemple (20a) contredit une idée qu'on aurait pu avoir *a priori*, à savoir qu'un adverbial qui se situe dans une complétive *P* introduite par *N0 V W* a pour source le référent de *N0* [10]. Le contraste entre (19c) et (20c) montre que

9. Par souci de simplification, nous ignorons dans cette discussion la succession temporelle d'événements dont un au moins est sous la portée d'une polarité négative.

10. (Bonami & Godard, 2008) mettent aussi en avant des exemples où un adverbe évaluatif tel que *bizarrement* se situe dans une complétive sans avoir pour source le référent de *N0*.

les informations de factivité sont nécessaires pour déterminer la source des relations de discours (en plus d'être nécessaires pour déterminer leurs arguments, comme montré à la section précédente).

Examinons (21a) où la seconde phrase comporte une coordination. Son analyse discursive, construite par analogie avec celle de (19a), est donnée en (21a). Celle-ci est un contre-exemple à la contrainte de la frontière droite posée en SDRT (Asher & Lascarides, 2003). Rappelons que ces auteurs étudient seulement des discours ne comportant que des assertions de l'auteur et que donc toutes les relations de discours sont attribuées à l'auteur. La contrainte de la frontière droite a pour effet d'interdire $R(\alpha, \beta) \wedge R'(\alpha, \gamma)$ lorsque R est coordonnante. Si cette interdiction est probablement valable lorsque seules sont considérées des assertions de l'auteur, soit R_{auteur} et R'_{auteur}, elle ne l'est plus lorsqu'on considère des sources autres que l'auteur : voir l'analyse en (21b) avec $Narration_{Jane}(1,3) \wedge Narration_{Zoé}(1,5)$ où $Narration$ est sans conteste coordonnante. Il est donc nécessaire d'étudier la validité de la contrainte de la frontière droite dans des discours mettant en jeu la relation $Attribution$.

(21) a. *(Fred ira à Dax pour Noël)*$_1$. (Jane croit)$_2$ qu'ensuite (il ira à Pau)$_3$ et (Zoé croit)$_4$ qu'(ensuite) (il ira à Bayonne)$_5$.
b. $Attribution_{auteur}(2,3) \wedge Narration_{Jane}(1,3) \wedge Attribution_{auteur}(4,5) \wedge Narration_{Zoé}(1,5)$

Tournons-nous maintenant vers la source de *ensuite* dans des exemples de forme *Ensuite, N0 V W que P* en considérant les trois discours obtenus à partir de (19a)-(21a) en déplaçant *ensuite* en tête de la phrase, voir (22). En (22a), qui répète (18a), la source de *ensuite* n'est pas claire : on peut hésiter entre l'auteur ou Jane [11]. En (22b), la source de *ensuite* est l'auteur comme en (20a). Enfin, en (22c), la source de *ensuite* est l'auteur alors que c'est Jane ou Zoé en (21a).

(22) a. *Fred ira à Dax pour Noël.* Ensuite, Jane croit qu'il ira à Pau.
b. *Fred ira à Dax pour Noël.* Ensuite, Jane ne sait pas qu'il ira à Pau.
c. *Fred ira à Dax pour Noël.* Ensuite, Jane croit qu'il ira à Pau et Zoé croit qu'il ira à Bayonne.

En conclusion, la position du connecteur, en tête de la phrase complexe ou en tête de la complétive, n'induit pas de différence de sens majeur à l'exception du fait qu'elle peut éventuellement influer sur la source du connecteur (et donc sur la source de la relation de discours qu'il exprime).

5 Source des arguments d'une relation de discours

Montrons brièvement qu'une structure discursive faisant appel à $Attribution$ et composée d'une conjonction de formules $R_{s_j}(\alpha, \beta)$ où les sources s_j des relations R sont annotées suffit à déterminer la source des arguments α et β de R_{s_j}, quelle que soit R, et par là-même récursivement la source de chaque EDU du texte.

1) $R = Attribution$. Le premier argument α est alors forcément une EDU. En effet, $Attribution$ amène à décomposer un enchâssement de plusieurs complétives : (23a) reçoit la segmentation en (23b) et l'analyse discursive en (23c).

(23) a. Fred pense que Zoé a dit qu'il neigeait.
b. (Fred pense)$_1$ que (Zoé a dit)$_2$ qu'(il neigeait)$_3$.
c. $Attribution_{auteur}(1, [2,3]) \wedge Attribution_{Fred}(2,3)$

$Attribution_{s_j}(\alpha, \beta)$ indique que la source de α est s_j. Le second argument β peut être une EDU ou un segment complexe. Lorsque β est une EDU, sa source est déterminée par le segment attributif α. Si celui-ci est de forme *N0 V W* alors la source de β est le référent de *N0*. Lorsque β est un segment complexe, la structure discursive de ce segment permet récursivement de déterminer la source de ses EDU. Ainsi, en (23a), la source de 1 est l'auteur, celle de 2 Fred et celle de 3 Zoé.

2) $R \neq Attribution$. Si les arguments α et β de R sont des EDU, alors soit ils apparaissent l'un et/ou l'autre dans la structure discursive comme argument d'une relation $Attribution$, par exemple $Attribution_{s_k}(\gamma, \beta)$, et l'étape précédente s'applique alors pour déterminer leur source, soit ce n'est pas le cas et leur source est alors l'auteur. Si les arguments α et β ne sont pas des EDU, le mécanisme s'applique récursivement.

11. Ce type d'hésitation sur la source d'un connecteur a amené les auteurs du PDTB à faire un choix par défaut, à savoir l'auteur.

6 Conclusion et perspectives

En nous inspirant de théories sur le discours (RST et SDRT) d'une part, et du corpus annoté PDTB d'autre part, nous avons mis en évidence que le calcul ou l'annotation de l'analyse discursive d'un texte devait produire une structure hiérarchique connexe où tous les éléments d'information devaient être reliés — ce qui n'est pas le cas dans le PDTB — et où les sources des différentes relations de la structure discursive devaient être calculées ou annotées — ce qui n'est pas le cas en RST ou SDRT. En corollaire, nous avons montré qu'une structure discursive qui utilise la relation $Attribution$ et qui définit la source de chaque relation de discours est suffisante pour déterminer la source de chaque EDU du texte (Section 5).

Nous avons montré de plus que les informations de factivité à la FactBank étaient primordiales non seulement pour identifier les arguments d'une relation de discours donnée (Section 4.1) mais aussi pour identifier sa source (Section 4.2), et donc, d'après notre corollaire, identifier la source de chaque EDU du texte. A rebours, nous avons montré que les informations de factivité à la FactBank devaient être révisées ou complétées après le calcul ou l'annotation de la structure discursive.

Dans nos perspectives de recherche, nous incluons :

1) L'extension du travail présenté ici à d'autres constructions que les discours indirects de forme *N0*$_{hum}$ *V W que P* discutées dans cet article. Nous avons déjà amorcé l'étude des discours directs dans (Danlos *et al.*, 2010). Il reste les constructions à infinitive (avec des "verbes à montée" ou "des verbes à contrôle"), les constructions impersonnelles et les constructions mettant en jeu un "verbe de discours" dont le sujet réfère à un événement (*Ceci a précédé/expliqué/prouvé que P*) (Danlos, 2006). Pour chacune de ces constructions, qui présentent un événement enchâssé sous un autre, il faut en premier lieu déterminer quelle est la segmentation en EDU — en se posant la question de savoir si cette segmentation peut être obtenue à partir d'une analyse syntaxique profonde. Il faut ensuite déterminer les différentes analyses discursives possibles, en s'appuyant sur les informations de factivité comme nous l'avons fait ici pour les discours indirects.

2) La réalisation d'un corpus annoté discursivement pour le français qui combine les avantages de RST, de SDRT et du PDTB en suivant les conclusions de cet article. Pour cette tâche, nous pouvons bénéficier du corpus French TimeBank (Bittar, 2010) qui peut servir de première étape pour fournir un corpus French FactBank — le corpus anglais FactBank a été réalisé comme une seconde couche d'annotations sur TimeBank (Saurí & Pustejovsky, 2009). Les informations de factivité du French FactBank recevant ultimement une autre couche d'annotations concernant la structure discursive. Cette tâche est de longue haleine et ne saurait être confiée à des annotateurs "naifs", mais est-il besoin de rappeler que seuls des corpus annotés proprement permettront aux techniques d'apprentissage (supervisées) de fournir des résultats dignes d'intérêt ?

Remerciements

Je remercie chaleureusement Owen Rambow avec qui j'ai eu des discussions tout à fait éclairantes sur différents thèmes abordés dans cet article, Sylvain Kahane et Philippe Muller pour leur relecture attentive, et les reviewers anonymes de TALN.

Références

ASHER N. (1993). *Reference to Abstract Objects in Discourse*. Dordrecht : Kluwer.

ASHER N. & LASCARIDES A. (2003). *Logics of Conversation*. Cambridge : Cambridge University Press.

BITTAR A. (2010). *Building a TimeBank for French : A Reference Corpus Annotated According to the ISO-TimeML Standard*. PhD thesis, Université Paris Diderot (Paris 7).

BONAMI O. & GODARD D. (2008). Lexical semantics and pragmatics of evaluative adverbs. In L. MCNALLY & C. KENNEDY, Eds., *Adverbs and Adjectives : Syntax, Semantics, and Discourse*, p. 274–304. Oxford University Press.

BRAS M. (2008). *Entre relations temporelles et relations de discours*. Université de Toulouse le Mirail : Dossier d'HDR.

CARLSON L., MARCU D. & OKUROWSKI M. E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. VAN KUPPEVELT & R. SMITH, Eds., *Current Directions in Discourse and Dialogue*, p. 85–112. Kluwer Academic Publishers.

DANLOS L. (2000). Event coreference in causal discourses. In P. BOUILLON & F. BUSA, Eds., *The Language of Word Meaning*, p. 216–241. Cambridge University Press.

DANLOS L. (2006). Discourse verbs and discourse periphrastic links. In *Proceedings of the second workshop on Constraints in Discourse (CID)*, Maynooth, Ireland.

DANLOS L. (2009). D-STAG : un formalisme d'analyse automatique de discours basé sur les TAG synchrones. *Revue TAL*, **50**(1), 111–143.

DANLOS L. (2011). Factivity information and veridicality of discourse relations. In *Proceedings of the Constraints in Discourse workshop CID 2011*, Agay-Roches rouges, France.

DANLOS L., SAGOT B. & STERN R. (2010). Analyse discursive des incises de citations. In *Actes du Second Colloque Mondial de Linguistique Française"*, New-Orleans, USA.

DINESH N., LEE A., MILTSAKAKI E., PRASAD R., JOSHI A. & WEBBER B. (2005). Attribution and the (non)alignment of syntactic and discourse arguments of connectives. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II : Pie in the Sky*, p. 29–36, Ann Arbor, Michigan : Association for Computational Linguistics.

FORBES-RILEY K., WEBBER B. & JOSHI A. (2006). Computing discourse semantics : The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, **23**(1).

HAMBLIN C. L. (1970). *Fallacies*. London : Methuen.

HUNTER J., ASHER N., REESE B. & DENIS P. (2006). Evidentiality and intensionality : Two uses of reportative constructions in discourse. In *Proceedings of the Constraints in Discourse Workshop (CID'06)*, Maynoth, Ireland.

MANN W. C. & THOMPSON S. A. (1988). Rhetorical Structure Theory : Toward a functional theory of text organization. *Text*, **8**(3), 243–281.

PDTB GROUP (2008). *The Penn Discourse Treebank 2.0 Annotation Manual*. Rapport interne, Institute for Research in Cognitive Science, University of Philadelphia.

PÉRY WOODLEY M.-P., ASHER N., ENJALBERT P., BENAMARA F., BRAS M., FABRE C., FERRARI S., HO DAC L.-M., LE DRAOULEC A., MATHET Y., MULLER P., PRÉVOT L., REBEYROLLE J., TANGUY L., VERGEZ COURET M., VIEU L. & WIDLÖCHER A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. In *Proceedings of TALN 2009*, p. 190–196, Senlis, France.

PRABHAKARAN V., RAMBOW O. & DIAB M. (2010). Automatic committed belief tagging. In *Proceedings of COLING 2010*, Pékin, Chine.

PRASAD R., DINESH N., LEE A., JOSHI A. & WEBBER B. (2006). Attribution and its annotation in the Penn Discourse Treebank. *Revue TAL*, **47**(2).

PUSTEJOVSKY J. (1995). *The generative Lexicon*. Cambridge : The MIT Press.

REDEKER G. & EGG M. (2006). Says who ? on the treatment of speech attributions in discourse structure. In *Proceedings of the Constraints in Discourse Workshop (CID'06)*, Maynoth, Ireland.

SAURÍ R. (2008). *A Factuality Profiler for Eventualities in Text*. PhD thesis, Brandeis University.

SAURÍ R. & PUSTEJOVSKY J. (2009). FactBank : A corpus annotated with event factuality. *Language Resources and Evaluation*, **43**, 227–268.

TABOADA M. & MANN W. (2006). Rhetorical Structure Theory : Looking back and moving ahead. *Discourse Studies*, **8**(3), 423–459.

WEBBER B. (2004). D-LTAG : extending lexicalized TAG to discourse. *Cognitive Science*, **28**(5), 751–779.

WOLF F. & GIBSON E. (2006). *Coherence in Natural Language : Data Structures and Applications*. London : The MIT Press.

TALN 2011, Montpellier, 27 *juin* –1er *juillet 2011*

Paraphrases et modifications locales dans l'historique des révisions de Wikipédia

Camille Dutrey[1] Houda Bouamor[2,3] Delphine Bernhard[2] Aurélien Max[2,3]
(1) INALCO, Paris, France
(2) LIMSI-CNRS, Orsay, France
(3) Univ. Paris-Sud, Orsay, France
camille@dutrey.fr {prénom.nom}@limsi.fr

Résumé. Dans cet article, nous analysons les modifications locales disponibles dans l'historique des révisions de la version française de Wikipédia. Nous définissons tout d'abord une typologie des modifications fondée sur une étude détaillée d'un large corpus de modifications. Puis, nous détaillons l'annotation manuelle d'une partie de ce corpus afin d'évaluer le degré de complexité de la tâche d'identification automatique de paraphrases dans ce genre de corpus. Enfin, nous évaluons un outil d'identification de paraphrases à base de règles sur un sous-ensemble de notre corpus.

Abstract. In this article, we analyse the modifications available in the French Wikipédia revision history. We first define a typology of modifications based on a detailed study of a large corpus of modifications. Moreover, we detail a manual annotation study of a subpart of the corpus aimed at assessing the difficulty of automatic paraphrase identification in such a corpus. Finally, we assess a rule-based paraphrase identification tool on a subset of our corpus.

Mots-clés : Wikipédia, révisions, identification de paraphrases.

Keywords: Wikipedia, revisions, paraphrase identification.

1 Introduction

Wikipédia ne cesse de croître et est actuellement l'encyclopédie libre la plus volumineuse et la plus fréquentée au monde. Ses articles sont écrits et maintenus de manière collaborative et bénévole. Les énormes quantités de données présentes dans cette encyclopédie ont motivé de nombreux travaux sur l'acquisition automatique de ressources comme par exemple l'acquisition des connaissances lexico-sémantiques (Zesch *et al.*, 2008). Cependant, la majorité de ces études n'utilisent que la version la plus récente des articles de l'encyclopédie. Wikipédia met également à disposition l'historique des révisions de chacun de ses articles qui sont itérativement modifiés et affinés par de multiples utilisateurs du Web. Ces révisions rendent possible l'extraction de certains types de modifications locales reflétant l'évolution, la maturation et la correction de la forme linguistique des articles, et constituent donc une importante source de connaissances encore peu exploitée à ce jour.

Dans cet article, nous détaillons une typologie des modifications locales présentes dans le corpus WiCoPaCo [1]

1. Librement téléchargeable sur `http://wicopaco.limsi.fr`

de révisions extraites automatiquement de la version française de Wikipédia. Notre étude met l'accent sur le phénomène de *paraphrases locales*, qui sont de plus en plus utilisées pour améliorer les performances de plusieurs applications de TAL comme les systèmes de traduction automatique (Max, 2010) ou de question-réponse (Duclaye *et al.*, 2003), ainsi qu'en génération, pour aider des auteurs à trouver des formulations plus adaptées (Max, 2008).

Cet article est organisé comme suit : dans la section 2, nous passons tout d'abord en revue les principaux travaux portant sur l'utilisation de l'historique des révisions de Wikipédia, puis nous décrivons le corpus WiCoPaCo utilisé dans cette étude dans la section 3. Nous présentons la typologie des modifications locales que nous proposons dans la section 4. Dans la section 5 nous exposons nos premières expériences sur l'identification automatique de paraphrases dans ce corpus, et enfin nous présentons nos observations et décrivons nos travaux futurs dans la section 6.

2 Exploitation des révisions de Wikipédia : état de l'art

Les révisions de Wikipédia ont déjà été exploitées pour différentes tâches et applications. Nelken & Yamangil (2008) exploitent l'historique des révisions de Wikipédia pour acquérir une grande quantité de données d'apprentissage en comparant les versions adjacentes d'un même article pour trois tâches à différents niveaux de granularité linguistique : collecte des fautes d'orthographe et de leur correction (niveau du mot), données d'apprentissage pour les algorithmes de compression de phrases (niveau phrase) et d'amorçage pour les systèmes de résumé automatique (niveau document). Yatskar *et al.* (2010) utilisent l'historique des modifications dans la version anglaise simplifiée de Wikipédia pour en extraire des simplifications lexicales.

Max & Wisniewski (2010) décrivent WICOPACO (Wikipédia Correction and Paraphrase Corpus), une ressource construite en explorant automatiquement l'historique des révisions de Wikipédia et en extrayant les modifications locales effectuées par les contributeurs. Ce corpus comprend différents types de corrections et de réécritures. Le travail de Wisniewski *et al.* (2010) montre par exemple comment cette ressource peut être utilisée pour améliorer la performance d'un système de correction orthographique automatique. Zanzotto & Pennacchiotti (2010) exploitent quant à eux l'historique des révisions de Wikipédia pour extraire un grand nombre de paires d'unités textuelles en relation d'implication et appliquent des méthodes d'apprentissage semi-supervisé pour rendre l'ensemble des données extraites cohérentes par rapport aux données existantes.

Dans la mesure où l'utilisation et la popularité des wikis ainsi que d'autres systèmes collaboratifs s'accroît, les questions concernant la fiabilité de ces informations gagnent en importance. Hu *et al.* (2007) ont par exemple développé un modèle de confiance basé sur l'historique des éditions des articles de Wikipédia afin de calculer et contrôler leur fiabilité. D'autres travaux ont proposé des visualisations originales des révisions de Wikipédia, tels que History Flow (Viégas *et al.*, 2004) ou WikiDashboard (Suh *et al.*, 2008). WikipediaViz (F. Chevalier, S. Huot et J-D. Fekete, 2010) propose un ensemble de visualisations basé sur un mécanisme de collecte et d'aggrégation de données d'éditions de Wikipédia pour aider le lecteur à appréhender la maturité d'un article.

Comme nous l'avons montré, la plupart des travaux de recherche antérieurs sur l'historique des révisions de Wikipédia se concentrent sur des aspects spécifiques de la ressource et se fixent pour objectif des applications bien définies telles que la simplification de texte, la compression de phrases ou encore la visualisation des informations. À notre connaissance, il n'y a pas de vision globale des phénomènes de modifications locales disponibles dans les révisions de Wikipedia, bien qu'il existe une grande variété de types de modifications qui sont d'intérêt pour de nombreuses applications de traitement automatique des langues, et en particulier les paraphrases locales.

3 WICOPACO, un corpus de modifications locales de Wikipédia

L'acquisition de paires de segments textuels ayant le même sens (*paraphrases locales*) a été à l'origine d'un nombre important de travaux sur l'exploitation automatique de corpus de textes (voir par exemple (Madnani & Dorr, 2010)). Les corpus utilisés peuvent être organisés par le degré de correspondance entre deux unités de texte : des paires de paraphrases phrastiques, obtenues par exemple par traduction multiple (*corpus parallèles monolingues*) ; les paires de phrases ayant le même sens, obtenues à partir de corpus composés de textes dans la même langue partageant une partie du vocabulaire employé, ce qui implique généralement que les textes parlent d'un même sujet durant la même période (*corpus monolingues comparables*) ; les paires de phrases partageant des traductions dans d'autres langues (*corpus parallèles multilingues*). Les corpus monolingues parallèlles sont les corpus les plus appropriés pour observer et acquérir automatiquement des segments de texte en relation de paraphrase locale de haute qualité (Bouamor *et al.*, 2010). Or ce type de corpus existe en très faible quantité, et leur construction est une tâche compliquée et coûteuse. *A contrario*, un des principaux défauts des autres types de corpus est que les paraphrases potentielles n'y sont observées qu'*indirectement*, par exemple par l'intermédiaire d'une traduction commune ou d'un contexte jugé similaire.

Une autre source potentielle de paraphrases locales réside dans les nombreuses modifications que les rédacteurs font lors de la révision d'un texte, certaines d'entre elles étant destinées à ne pas modifier le sens du texte, mais à améliorer sa qualité, le rendre plus cohérent, ou limiter sa redondance. Des brouillons d'écrivains ont été notamment utilisés dans les critiques génétiques textuelles qui étudient les processus de création de textes (Bourdaillet & Ganascia, 2007). Ces documents annotés sont malheureusement disponibles en petites quantités et sont de plus difficiles à encoder en format électronique. En outre, ces projets contiennent souvent des réorganisations textuelles importantes qui sont très difficiles à exploiter pour l'acquisition de paraphrases. L'émergence et l'adoption des *wikis* a fait de l'écriture collaborative une pratique très courante. L'encyclopédie en ligne Wikipédia, en particulier, attire de nombreuses contributions sur un large éventail de sujets et dans de nombreuses langues. Bien que certaines contributions consistent en des changements importants (par exemple la création d'un article, la suppression d'une section, la réécriture complète d'un paragraphe), une proportion importante des modifications textuelles sont effectuées sur des textes courts pour corriger, améliorer ou enrichir le contenu de l'encyclopédie. L'historique des révisions de cette ressources constitue donc une source importante de phénomènes de réécriture *naturelle*, y compris des paraphrases locales dans leur contexte.

WICOPACO (Max & Wisniewski, 2010) est un corpus de modifications locales, extrait à partir de l'historique des révisions des articles de Wikipédia, et disponible pour le moment pour le français. Ce corpus a été construit en 4 étapes :

1. Sélection de paires de versions d'articles.
2. Normalisation du texte (segmentation, suppression du « wikitexte », etc.).
3. Alignement des modifications par une recherche de plus longues sous-séquences communes.
4. Filtrage des modifications retenues (séquences initiales de 7 mots ou plus) et extraction du contexte avant et après modification (paragraphe englobant).

Le corpus que nous avons utilisé contient 408 816 entrées uniques. Celui-ci est disponible sous forme d'un fichier au format XML associant un élément à chaque modification. Une modification y est décrite par un contexte avant modification et un contexte après modification, ainsi que par un ensemble de métadonnées donnant des informations sur le contributeur de la révision et permettant de localiser le texte extrait dans la ressource Wikipédia d'origine. Un exemple d'un tel élément, correspondant à une simplification lexicale, est donné à la figure 1.

```
<modif id="407851" wp_page_id="1830844" wp_before_rev_id="20691183" wp_after_rev_id="20691225"
wp_user_id="287861" wp_user_num_modif="81" wp_comment="">
<before>Le genre Archaeopteris possède plus de caractéristiques communes avec les plantes à graines que toute autre
<m num_words="1">ptéridophyte</m> connue et les analyses cladistiques récentes le placent en groupe-frère des
plantes à graines .</before>
<after>Le genre Archaeopteris possède plus de caractéristiques communes avec les plantes à graines que toute autre
<m num_words="2">plante fossile</m> connue et les analyses cladistiques récentes le placent en groupe-frère des
plantes à graines .</after>
</modif>
```

FIGURE 1 – Exemple d'une modification dans le corpus WICOPACO.

4 Typologie des modifications locales

Nous avons analysé le corpus WICOPACO afin de développer une typologie détaillée des modifications locales [2] dans les révisions de Wikipédia. Cette typologie permet de représenter tous les phénomènes observables dans WICOPACO et d'indiquer le degré de variation sémantique entre les segments correspondant à des paires de modifications locales. Elle se compose de deux catégories distinguant deux grandes classes de variation sémantique : la classe des *faibles variations sémantiques* et la classe des *corrections factuelles et vandalismes*. Ces deux classes peuvent contenir des modifications locales pour lesquelles il n'existe pas de relation sémantique stricte entre le segment avant la modification et le segment après modification, ce qui est par exemple le cas pour les modifications typographiques [3].

4.1 Modifications à faible variation sémantique

La classe des modifications à faible variation sémantique comporte les *corrections* et les *reformulations* (voir Table 1).

Les corrections de surface font référence aux changements de surface qui visent à améliorer le texte afin qu'il soit conforme aux normes linguistiques, et se décomposent de la manière suivante :

– *Corrections typographiques* : changement de la disposition et du format du texte, par exemple, ajout ou suppression d'espaces ou de signes de ponctuation, changement de casse d'un caractère, modification du format d'une date ou d'une heure, écriture d'un nombre en toutes lettres ou en chiffres, etc.
– *Corrections orthographiques* : corrections affectant les fautes d'orthographe. Elles se réfèrent à la transformation d'un mot inexistant en un mot attesté dans le lexique, comme la modification de diacritiques ou le remplacement d'un ou plusieurs caractères.
– *Corrections grammaticales* : résolution des fautes d'orthographe qui ne peuvent être détectées et corrigées que par la prise en compte du contexte.

2. Ici nous héritons de la définition suivie dans la ressource utilisée pour la localité des modifications observées : il s'agit de segments d'au plus 7 mots (ponctuations et autres signes non inclus).

3. La typologie complète est disponible sous forme de document technique du LIMSI (Dutrey *et al.*, 2011) et est accessible à l'adresse suivante : `http://wicopaco.limsi.fr/pub/typologie-modifications-wikipedia.pdf`

CORRECTIONS DE SURFACE
Corrections typographiques
⇒ ex. une espace remplacée par un trait d'union pour corriger une erreur typographique : *Le triceps brachial est un muscle extenseur de l' **[avant bras → avant-bras]** sur le bras.*
Corrections orthographiques
⇒ ex. un caractère alphabétique supprimé pour transformer un *non-mot* en un mot attesté dans le lexique : *Ces trois parties se **[rejoingnent → rejoignent]** pour former une épaisse masse.* ⇒ ex. un diacritique remplacé par un autre pour transformer un *non-mot* en un mot attesté dans le lexique : *L' **[ëglise → église]** gothique Sainte-Marie…*
Corrections grammaticales
⇒ ex. un diacritique remplacé par un autre pour corriger une erreur portant sur un mot attesté dans le lexique : *L'anathème pour le **[pêcheur → pécheur]** : ce dernier est privé de sépulture chrétienne.* ⇒ ex. un mot remplacé par un autre pour corriger une erreur portant sur un mot attesté dans le lexique : *Il chante avec une **[voie → voix]** de troubadour.*
REFORMULATIONS
Reformulations lexicales
⇒ ex. un emprunt remplacé par la forme correspondante en français standard : ***[L'implémentation → La mise en œuvre]** de l'algorithme…*
Reformulations syntaxiques
⇒ ex. une permutation entre deux segments sur l'axe syntagmatique : *Source : **[l'Invention de l'Europe d'Emmanuel Todd → Emmanuel Todd, L'Invention de l'Europe]**.* ⇒ ex. une proposition circonstancielle transformée en une proposition relative : *Un infomercial pseudo-scientifique **[en exposant → qui expose]** grossièrement…*
Reformulations sémantiques
⇒ ex. un mot remplacé par un autre appartenant au même champ lexical (hyponymie) : *Il fonde le **[journal → quotidien]** francophone « Le Tunisien »en 1907.* ⇒ ex. une paraphrase servant différents propos (infra précision de sens) : *Ce vers de Nuit rhénane d'Apollinaire **[qui paraît presque sans structure rythmique → dont la césure est comme masquée]**…*

TABLE 1 – Types de modifications à faible variation sémantique

Les reformulations correspondent à des changements plus importants qui modifient les choix lexicaux et syntaxiques faits par le contributeur précédent sans modifier profondément la signification du texte :

– *Reformulations lexicales* consistant, par exemple, à remplacer un acronyme avec son nom complet, traduire un mot étranger ou un emprunt, remplacer une variante régionale par sa version standard, etc.
– *Reformulations syntaxiques* permettant, par exemple, de modifier l'ordre des propositions, de transformer une phrase à la voix active ou passive ou de changer le type de proposition.
– *Reformulations sémantiques* comme l'utilisation d'hyperonymes ou d'hyponymes, la normalisation encyclopédique, l'utilisation de synonymes ou l'ajout d'informations additionnelles peu significatives.

4.2 Corrections factuelles et vandalismes

CORRECTIONS FACTUELLES
⇒ ex. un mot remplacé par son antonyme : *Un catalyseur solide (phase **[liquide → solide]**) avec de l'hydrogène (phase gazeuse).* ⇒ ex. un segment remplacé par un autre n'ayant aucun lien sémantique avec le premier : *représente pour eux **[l'Occident chrétien → la supériorité de la race celto-germanique]**.*
VANDALISMES
⇒ ex. une chaîne insérée produisant un *non-mot* (vandalisme manifeste) : *L'Autriche a été occupée **[par → psh ! ! ar]** les Romains.* ⇒ ex. un mot remplacé par un autre qui ne produit aucun sens compte tenu du contexte (vandalisme subtil) : *Devant la Cour de **[Cassation → Castration]**...*

TABLE 2 – Corrections factuelles et vandalismes

Cette classe se décompose en deux sous-types (voir la Table 2), les *corrections factuelles* et les *vandalismes*, pour lesquels le sens du texte est fortement affecté et peut être totalement changé.

Les corrections factuelles correspondent soit à une modification qui induit une forte variation de sens, soit à une modification ne présentant aucun lien sémantique avec le texte initial. Elles consistent, par exemple, à remplacer un mot par un antonyme ou à changer le temps d'un verbe de sorte à ce que le sens de la phrase soit modifié. Ce genre de modifications vise à améliorer le contenu de Wikipédia.

Le vandalisme fait référence aux modifications qui, délibérément, modifient ou détruisent le contenu afin de nuire à la qualité de Wikipédia. Le vandalisme manifeste se caractérise par l'insertion de non-mots ou d'insultes tandis que le vandalisme subtil se caractérise par des reformulations grammaticales mais dont l'interprétation est en complète contradiction avec le sens initial. Il est particulièrement important de détecter ce dernier type, puisque dans certains cas un lecteur peu attentif ou crédule pourra considérer à tort l'information décrite comme fiable.

4.3 Annotation manuelle

Nous avons conçu un schéma d'annotation basé sur la typologie décrite précédemment. L'objectif de cette annotation est d'évaluer le degré de complexité de l'identification manuelle des paraphrases dans les modifications locales. L'annotation est guidée par notre application cible qui est l'identification automatique des paraphrases dans WICOPACO. Dans notre typologie, les paraphrases correspondent aux reformulations présentes dans la classe des réécritures à faible variation sémantique. Elles doivent être distinguées des corrections de surface et des reformulations qui induisent un changement sémantique majeur. Cette annotation a donc deux buts principaux : repérer des phénomènes liés à une réécriture à faible variation sémantique et repérer les vandalismes, notamment dans une optique ultérieure d'apprentissage supervisé.

Afin de faciliter cette tâche, nous avons élaboré un schéma d'annotation composé de quatre classes principales :
– *Les corrections de surface*, qui englobent toutes les modifications visant à rendre le texte conforme aux normes de la langue.

– *Les reformulations*, qui correspondent aux différents types de paraphrases, y compris les précisions et les simplifications.
– *Les corrections factuelles et les vandalismes*
– *Les défauts d'alignement* qui correspondent aux cas où les modifications locales identifiées présentent un défaut dans leur alignement (voir Figure 2). Cependant, même avec un défaut d'alignement un segment peut contenir une modification locale.

Henri IV	**fut assassiné par Ravaillac en 1610**.
A partir de la conversion d' Henri IV	**la fidélité au roi l' a emporté sur l' appartenance religieuse**.

FIGURE 2 – Exemple d'un défaut d'alignement dans le corpus WICOPACO.

Une annotation couvre l'ensemble du segment identifié comme une modification locale (notée par une balise XML `m` dans le corpus WICOPACO, comme illustré dans la Figure 1) : l'objectif est de déterminer le type de la modification de partir d'une paire de segments, mais pas de réaligner les mots dans ces segments. En outre, il était possible d'attribuer plusieurs étiquettes à la même modification.

Pour réaliser cette annotation, nous avons utilisé l'outil d'alignement Yawat (Germann, 2008) conçu à l'origine pour l'alignement de textes parallèles bilingues au niveau du mot. Nous avons adapté le schéma d'annotation de cet outil pour notre annotation multi-niveaux. L'annotation a été réalisée par quatre annotateurs [4] sur 200 paires de segments tirées d'une version filtrée du corpus WICOPACO. Comme les modifications de ponctuation sont fréquentes, seules les modifications d'une distance d'édition (Levenshtein) d'au moins 4 ont été considérées pour l'annotation.

4.4 Résultats de l'annotation

La Table 3 décrit l'accord inter-annotateur de notre annotation, calculé à l'aide de la mesure du Kappa (κ) [5]. L'accord inter-annotateur varie de modéré à fort, en fonction de la classe. Globalement, les valeurs du κ sont proches des valeurs déclarées par Dolan & Brockett (2005) pour l'identification de paraphrases (κ de 0,62) et par Glickman *et al.* (2005) (κ de 0,6) pour l'implication textuelle.

Type	**κ moy**	**Interprétation**	**κ maximum**	**κ minimum**
Corrections factuelles et vandalismes	0,65	Accord fort	0,71	0,61
Reformulation	0,60	Accord modéré	0,71	0,51
Correction	0,54	Accord modéré	0,81	0,40
Défaut d'alignement	0,48	Accord modéré	0,62	0,28

TABLE 3 – Accord inter-annotateur pour l'annotation des révisions de Wikipédia.

Nous indiquons également le nombre d'annotations identiques attribuées par 1 à 4 annotateurs (voir la Table 4) ainsi que les annotations uniques, attribuées par un seul annotateur. Cela permet de quantifier approximativement les phénomènes présents dans le corpus. Les paraphrases ont le plus grand nombre d'occurrences, suivies par les corrections factuelles et vandalismes. Ceci montre que les révisions de Wikipédia constituent un corpus bien

4. Co-auteurs du présent article.
5. Nous avons utilisé le calculateur κ en ligne pour annotateurs et classes multiples diponible sur `http://cosmion.net/jeroen/software/kappa/`.

adapté pour l'acquisition automatique de paraphrases [6]. En outre, les défauts d'alignement sont assez rares, ce qui montre que la méthode d'alignement utilisée pour la construction de WICOPACO est suffisamment précise pour fournir des modifications utiles.

	4 ann.	3 ann.	2 ann.	unique ann.	Total
Correction de surface	9	2	7	23	41
Reformulation	60	33	24	15	132
Corrections factuelles et vandalismes	47	15	13	32	107
Défaut d'alignement	2	4	8	6	20

TABLE 4 – Nombre d'annotations identiques attribués par 1, 2, 3 ou 4 annotateurs.

L'étude des annotations a souligné certains problèmes potentiels pour l'identification automatique des classes décrites dans notre typologie. Tout d'abord, plusieurs phénomènes peuvent se produire simultanément, par exemple une transformation de diathèse (voix grammaticale) peut inclure une correction d'un non-mot (erreur). Dans ce cas, un classifieur automatique devrait être en mesure d'assigner plusieurs classes à une modification. Deuxièmement le contexte phrastique fourni par le corpus WICOPACO n'est parfois pas suffisant pour prendre une décision sur un type de modification spécifique. Un contexte plus large pourrait être utile aux classifieurs automatiques. Troisièmement, le typage correct d'une modification nécessite parfois une certaine connaissance des intentions du contributeur. Ce type d'information est parfois disponible dans les commentaires associés à une révision, mais peut être difficile à interpréter de façon automatique.

5 Identification de paraphrases : une méthode à base de règles

Nous avons mis en œuvre une méthode automatique destinée à distinguer les paraphrases des autres modifications dans WICOPACO, en adaptant l'outil de reconnaissance de variantes de termes `Fastr` (Christian Jacquemin, 1994). L'opération d'*indexation contrôlée* de ce système définit les variations acceptables par un système de métarègles s'appliquant à des règles de termes. Elles permettent d'exprimer les réécritures morphosyntaxiques possibles, ainsi que les relations d'ordre morphologique ou sémantique contenues dans des ressources préexistantes.

Nous avons dû créer un nouvel ensemble de métarègles pour la reconnaissance de paraphrases car le jeu de métarègles original s'est révélé innaproprié pour notre étude, cet ensemble ayant été développé avec l'objectif de reconnaissance de variantes de termes qui recouvrent une définition beaucoup plus permissive de la paraphrase. Nous avons cependant pu réutiliser les familles morphologiques et les familles sémantiques fournies par `Fastr`, pour exprimer des contraintes sémantiques et morphologiques. L'utilisation de `Fastr` nous permet d'évaluer si un système à base de règles est adapté pour l'identification des paraphrases dans un corpus tel que le nôtre, présentant une riche variété de phénomènes.

Nous avons utilisé un autre type de corpus pour le développement des nouvelles métarègles, afin de vérifier si les règles sont suffisamment générales pour être appliquées sur un corpus de type différent. Ce corpus est extrait de MULTITRAD (Bouamor, 2010), un corpus construit par collecte de paraphrases d'énoncés par traduction multiple multilingue, et annoté au niveau des mots. Ce corpus de développement a permis l'extraction de patrons

6. Cet article ne présente pas d'usage concret de paraphrases extraites de la ressource utilisée, mais il est évident que différentes paraphrases ne seront pas nécessairement adaptées pour les mêmes usages.

de reformulations exprimés sous forme de séquences de catégories morphosyntaxiques. La Table 5 montre par exemple les principaux patrons de reformulations observés pour le patron initial NOM VER VER.

Réécriture	Fréquence	Exemple
NOM VER VER	7	orateurs ont estimé → locuteurs ont jugé
NOM VER VER PRP NOM CONJ	2	Parlement a demandé → Parlement a fait des demandes pour
NOM VER PRP DET NOM	1	lois sont écrites→ lois restent dans les livres

TABLE 5 – Principaux patrons de réécriture associés à la séquence NOM VER VER dans MULTITRAD

La Table 6 illustre un exemple de nouvelle métarègle pour FASTR. Cette métarègle a pour nom **NAtoVASyn** car elle porte sur un segment source dont la structure est un **N**om suivi d'un **A**djectif réécrit (**to**) en un segment dont la structure est au minimum un **V**erbe suivi d'un **N**om suivi d'**A**djectif. La métarègle intègre également certaines contraintes morphologiques et sémantiques qui précisent que (i) le nom du segment source et le verbe du segment cible ont une racine morphologique commune et (ii) les adjectifs des segments source et cible sont synonymes. Il est à noter que l'utilisation d'un moteur de détection de variante est comparable aux travaux de Deléger & Zweigenbaum (2009) sur l'extraction de paraphrases de vulgarisation en langue de spécialité.

Metarule NAtoVAsyn(X1 → N1 A1) = X1 → V1 {ART ? | PRON ? | PREP ?} N A2 : <N1 root> = <V1 root> <A1 syn> = <A2 syn> <X1 metaLabel> = 'XX'.
protection constante → protéger de façon permanente

TABLE 6 – Exemple d'une métarègle de `Fastr`

Un ensemble de 83 métarègles a été développé pour la reconnaissance de paraphrases. Nous avons d'abord évalué la couverture des règles construites manuellement à l'aide de 206 paires de paraphrases d'énoncés issues du corpus MULTITRAD n'ayant pas été utilisés pour le développement des métarègles. `Fastr` a été en mesure d'identifier 185 paraphrases candidates, dont certaines sont illustrées dans la table 7.

MultiTrad	**WiCoPaCo**
décrit dans la proposition ↔ proposé	décéda ↔ mourut
objectif ultime ↔ but ultime	abritant ↔ qui abrite
reste ↔ demeure	standardisation ↔ normalisation

TABLE 7 – Exemples de paires de paraphrases locales identifiées par `Fastr` avec le jeu de métarègles développé.

Afin d'évaluer les règles sur les révisions de Wikipédia, nous avons construit manuellement un corpus de 200 paraphrases positives (paires de modifications en relation de paraphrases) et 200 négatives (paires de modifications sans lien de paraphrase) à partir du corpus WICOPACO. `Fastr` a identifié 31 paires de paraphrases candidates dans le corpus positif. Parmi elles 22 (70%) sont correctes (la modification est identifiée entièrement comme paraphrase), 7 (22,5%) correspondent à une sous-partie de la modification et 2 (6%) n'existent pas dans la référence (c'est-à-dire qu'elles couvrent une autre partie du contexte). Dans le corpus négatif, seulement 4 paraphrases candidates ont été trouvées, parmi lesquelles une seulement se trouve dans le corpus de référence.

Ces résultats préliminaires montrent que les patrons de réécriture morphosyntaxiques peuvent atteindre une bonne précision pour identifier des paraphrases locales dans les révisions de Wikipédia, mais que parfois des informations

plus fines sur le contexte syntaxique et sémantique sont nécessaires. La couverture obtenue est très limitée du fait de la grande variété de phénomènes concernés, en outre difficilement capturés par les règles développées sur un corpus de développement différent du corpus sur lequel a porté notre évaluation. Par ailleurs, l'étude de plusieurs exemples a révélé que les ressources morphologiques et sémantiques utilisées par `Fastr` pourraient être enrichies afin d'assurer une meilleure couverture pour notre tâche.

6 Conclusions et perspectives

Dans cet article, nous avons décrit une typologie des modifications locales présentes dans les révisions des articles de Wikipédia. Cette typologie pourra servir de repère utile pour des travaux ultérieurs sur cet ensemble de données. Si nous ne l'avons pas formellement démontré, nous pensons que la structure de haut niveau de notre typologie s'applique assez directement quelle que soit la langue étudiée. Nous travaillerons prochainement sur une version anglaise du corpus WICOPACO et testerons alors cette hypothèse.

Nous avons également effectué une annotation manuelle d'un sous-ensemble du corpus étudié. Cette étude a montré qu'une quantité importante de modifications correspondent à des reformulations avec de faibles variations sémantiques. Ceci constitue donc un résultat encourageant dans l'optique d'exploiter les révisions d'une ressource importante et dynamique telle que Wikipédia pour l'acquisition de paraphrases. Nos premières expériences exploitant un moteur de reconnaissance de variantes de termes adapté à nos besoins ont révélé des résultats encourageants, permettant avant tout d'obtenir une bonne précision sur les paraphrases identifiées. Les différentes limites de l'outil utilisé que nous avons identifiées nous ont permis de spécifier un nouveau moteur d'identification par règles plus adapté à nos besoins, permettant l'expression de règles avec une combinaison quelconque de contraintes portant sur le lexique, les constituants, ou les dépendances syntaxiques.

Cependant, la richesse des types de réécriture possibles rendent difficile l'obtention d'un jeu de règles suffisamment couvrant. Nous comptons donc par la suite étudier différents types de classifieurs automatiques avec différents jeux de traits, portant sur les caractéristiques linguistiques des modifications mais également sur les métadonnées des révisions. Ce type d'apprentissage se fondera vraisemblablement sur une quantité de données importante dont nous ne disposions pas encore à ce stade de notre étude. Les données qui pourraient être prochainement disponible *via* un jeu en ligne développé dans notre laboratoire [7] seraient ici particulièrement utiles, puisqu'elles incluent des paraphrases candidates en contexte proposées par des joueurs et des évaluations chiffrées attribuées par d'autres joueurs. Il pourrait donc par exemple s'agir d'un cadre d'annotation original pour les données de WICOPACO.

Le présent travail a montré que les révisions de Wikipédia contiennent de nombreuses réécritures à faible variation sémantique, incluant de nombreuses paraphrases locales. Nos propositions pour l'identification de paraphrases ainsi que nos travaux en cours produiront des listes de paires de paraphrases candidates en contexte, possiblement associées à un score de confiance. Nous comptons par la suite évaluer les paraphrases extraites par la tâche, en bénéficiant des travaux en recherche d'information précise et en traduction automatique menés dans notre laboratoire, qui correspondent à des applications du TAL très sensibles à la variation en langue.

Un autre type d'application sur lequel nous comptons travailler porte sur l'exploitation des patrons de réécriture acquis (incluant les patrons de corrections orthographiques et grammaticaux obtenus par Wisniewski *et al.* (2010)) pour l'aide à la rédaction d'articles sous Wikipédia. Une interface efficace permettrait notamment d'anticiper certaines corrections ultérieures par d'autres contributeurs et ainsi de rendre plus efficace le travail d'un contributeur,

7. Voir (Bouamor *et al.*, 2009) pour une description initiale.

et de réduire globalement le nombre de révisions nécessaires à l'échelle de l'encyclopédie. Par exemple, une normalisation fréquemment apportée aux textes de l'encyclopédie pourrait être suggérée de façon interactive au contributeur qui l'introduirait dans une nouvelle contribution.

Remerciements

Les auteurs tiennent à remercier Julien Boulet, Martine Hurault-Plantet et Guillaume Wisniewski pour leur participation à la création du corpus WICOPACO utilisé dans ce travail, ainsi que les très nombreux contributeurs à la création du corpus MULTITRAD.

Références

BOUAMOR H. (2010). Construction d'un corpus de paraphrases d'énoncés par traduction multiple multilingue. In *Actes de RÉCITAL 2010*, Montréal, Canada.

BOUAMOR H., MAX A. & VILNAT A. (2009). Amener des utilisateurs à créer et évaluer des paraphrases par le jeu. In *Actes de TALN, session de démonstrations*, Senlis, France.

BOUAMOR H., MAX A. & VILNAT A. (2010). Comparison of Paraphrase Acquisition Techniques on Sentential Paraphrases. In *Proceedings of the 7th International Conference on NLP, IceTAL 2010*, volume 6233 of *Lecture Notes in Computer Science, Advances in Natural Language Processing* : Springer Berlin / Heidelberg.

BOURDAILLET J. & GANASCIA J.-G. (2007). Machine Assisted Study of Writers' Rewriting Processes. In *Proceedings of the International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2007)*.

CHRISTIAN JACQUEMIN (1994). Recycling terms into a partial parser. In *Proceedings of the fourth conference on Applied natural language processing*, Stuttgart, Germany.

DELÉGER L. & ZWEIGENBAUM P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora*, Singapore.

DOLAN W. B. & BROCKETT C. (2005). Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

DUCLAYE F., COLLIN O. & YVON F. (2003). Apprentissage automatique de paraphrases pour l'amélioration d'un système de questions-réponses. In *Actes de TALN*, Batz-sur-mer, France.

DUTREY C., BOUAMOR H., BERNHARD D. & MAX A. (2011). *Typologie des modifications dans les révisions de Wikipédia*. Notes et documents du LIMSI 2011-01, LIMSI-CNRS.

F. CHEVALIER, S. HUOT ET J-D. FEKETE (2010). Visualisation de mesures agrégés pour l'estimation de la qualité des articles Wikipédia. In *Extraction et gestion des connaissances (EGC'2010), Actes, Hammamet, Tunisie, 26 au 29 janvier 2010*.

GERMANN U. (2008). Yawat : Yet Another Word Alignment Tool. In *Proceedings of the ACL-08 : HLT Demo Session*.

GLICKMAN O., DAGAN I. & KOPPEL M. (2005). A probabilistic classification approach for lexical textual entailment. In *Proceedings of the 20th national conference on Artificial intelligence (AAAI'05)* : AAAI Press.

HU M., LIM E.-P., SUN A. & LAUW, HADY WIRAWANAND VUONG B.-Q. (2007). Measuring article quality in Wikipedia : models and evaluation. In *CIKM '07 : Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, Lisbon, Portugal : ACM.

MADNANI N. & DORR B. J. (2010). Generating Phrasal & Sentential Paraphrases : A Survey of Data-Driven Methods. *Computational Linguistics*, **36**(3).

MAX A. (2008). Génération de reformulations locales par pivot pour l'aide à la révision. In *Actes de TALN*, Avignon, France.

MAX A. (2010). Example-based paraphrasing for improved phrase-based statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

MAX A. & WISNIEWSKI G. (2010). Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History. In *Proceedings of LREC 2010*, Valletta, Malta.

NELKEN R. & YAMANGIL E. (2008). Mining Wikipedia's Article Revision History for Training Computational Linguistic Algorithms. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence :An Evolving Synergy*.

SUH B., CHI E., KITTUR A. & PENDLETON B. (2008). Lifting the veil : improving accountability and social transparency in Wikipedia with Wikidashboard. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* : ACM.

VIÉGAS F., WATTENBERG M. & DAVE K. (2004). Studying Cooperation and Conflict Between Authors With History FlowVisualization. *Proceedings of the SIGCHI conferenceonHuman Factors in Computing Systems (CHI'04)*.

WISNIEWSKI G., MAX A. & YVON F. (2010). Recueil et analyse d'un corpus écologique de corrections orthographiques extrait des révisions de Wikipédia. In *Actes de TALN 2010*, Montréal, Canada.

YATSKAR M., PANG B., DANESCU-NICULESCU-MIZIL C. & LEE L. (2010). For the sake of simplicity : Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of the NAACL*.

ZANZOTTO F. M. & PENNACCHIOTTI M. (2010). Expanding textual entailment corpora from Wikipedia using co-training. In *Proceedings of the 2nd Workshop on Collaboratively Constructed Semantic Resources*.

ZESCH T., MÜLLER C. & GUREVYCH I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Language Resources and Evaluation(LREC'08)*, Marrakech, Morocco.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

<TextCoop>: un analyseur de discours basé sur les grammaires logiques

Patrick Saint-Dizier
IRIT-CNRS, Toulouse
stdizier@irit.fr

Résumé. Dans ce document, nous présentons les principales caractéristiques de <TextCoop>, un environnement basé sur les grammaires logiques dédié à l'analyse de structures discursives. Nous étudions en particulier le langage DisLog qui fixe la structure des règles et des spécifications qui les accompagnent. Nous présentons la structure du moteur de <TextCoop> en indiquant au fur et à mesure du texte l'état du travail, les performances et les orientations en particulier en matière d'environnement, d'aide à l'écriture de règles et de développement applicatif.

Abstract. In this paper, we introduce the main features of <TextCoop>, an environment dedicated to discourse analysis within a logic-based grammar framework. We focus on the structure of discourse rules (DisLog language) and on the features of the engine, while outlining the results, the performances and the orientations for future work.

Mots-clés : grammaire du discours, programmation en logique, grammaires logiques.

Keywords: discourse structure, logic programming, logic-based grammars.

1 Analyser quelles structures discursives ?

Lorsque l'on pense à l'analyse de structures discursives, il vient d'abord à l'esprit l'analyse des structures rhétoriques qui, d'une façon ou d'une autre, sont censées permettre de rendre compte de façon complète des diverses articulations discursives d'un texte (Marcu 97, 02). L'objectif est de relier tous les éléments d'un texte par le biais de ces relations, ce qui rend alors compte de la structure sémantico-pragmatique de ce texte. Outre le fait que ces relations existent en grand nombre et avec parfois des définitions un peu vagues et difficilement opérationalisables, il existe en fait, pour le besoin des applications, un grand nombre d'autres structures qui rentrent plus ou moins facilement dans le paradigme rhétorique.

C'est ainsi le cas des cadres du discours, initié en France par M. Charolles, pour lesquels les relations rhétoriques 'frame' ou 'background' ne sont pas tout à fait satisfaisantes. C'est aussi le cas de nombreux types de structures 'dédiées', comme par exemple les instructions dans le discours procédural. Enfin, notons toutes les structures qui relèvent de la typographie et qui ont un lien avec le contenu du texte (titres, notes, paragraphes, listes, etc.). Enfin, notons la complexité sous-jacente de certaines représentations qui forment des réseaux complexes de liens entre structures.

Dans la suite de ce document, nous proposons un environnement, <TextCoop>, dédié à l'analyse des structures discursives, basé sur la notion de grammaire logique. Nos expérimentations ayant largement tourné autour de l'analyse des diverses structures rencontrées dans les textes procéduraux, nombre d'exemples sont empruntés à ce cadre (Delpech et al 07, 08) (Aouladomar et al. 05), voir aussi (Delin 94) ou (Takechi 03). <TextCoop> désigne l'ensemble de l'architecture du système, y compris les outils d'aide à la mise au point et les ressouces linguistiques associées. DisLog (pour 'Dicourse in Logic' ou 'Discontinuities in Logic') désigne le langage qui décrit les règles d'analyse et les contraintes que l'on peut y associer.

Notre modélisation n'est pas dédiée à un cadre applicatif particulier ou à un genre textuel. Après un bref positionnement, nous présentons la syntaxe des règles de DisLog ainsi que des outils associés. Contrairement à une approche basée sur l'apprentissage (Marcu 02), l'ensemble de notre travail est positionné dans une modélisation linguistique et déclarative, typique des grammaires logiques, qui autorise le raisonnement. Notre approche est quelque peu basé sur une vision générative à base de principes. Nous présentons ensuite les fonctionalités du

moteur ainsi que son environnement actuel. Le développement de <TextCoop> est encore dans un stade expérimental : un travail est toujours en cours sur les propriétés de son environnement et des fonctions qu'il peut offir. Par contre, ses fondements sont globalement fixés, et ont été testés dans plusieurs cadres linguistiques et applicatifs.

Historiquement, <TextCoop> a initialement fait l'objet d'une étude dédiée aux procédures grand-public (Delpech et al 08) avec une implémentation simple en Perl de l'ensemble des fonctions. Toutefois, la rigidité, le peu de portabilité et les limites expressives de Perl nous ont poussé à refaire une implémentation en Java, à base de générateurs d'automates, utilisant JCUP. Cette approche a dû être abondonnée après 10 mois de programmation infructueuse. Voulant augmenter les possibilités expressives du système, en particulier au niveau raisonnement, et avoir un développement fiable et rapide, nous avons finalement opté pour une version en Prolog que nous pouvons facilement faire évoluer et maintenir. Via une collaboration avec une société, les aspects interfaces et aide à la mise au point seront développés dès que pertinent pour en faire une plateforme opérationelle. Une licence de type GPL est prévue au moins pour la partie noyau.

1.1 Le positionnement de <TextCoop>

La plateforme <TextCoop>, dédiée à l'analyse de structures du discours, doit permettre de pouvoir reconnaître une grande diversité de structures, génériques ou dédiées à des applications ou à des genres textuels. <TextCoop> vise à la fois le traitement de structures discursives génériques, dans des textes quelconques, et le traitement de structures plus spécifiques, 'métier', dans des textes plus spécialisés.

Considérant la complexité de la description des structures du discours, nous développons une vision qui s'appuie sur quelques considérations simples de la grammaire générative à savoir développer :
– des principes productifs, qui ont un bon niveau d'abstraction, linguistiquement adéquat, mais qui sur-reconnaissent dans certains cas,
– et des principes restrictifs qui viennent limiter la puissance des premiers, sur la base de contraintes de bonne formation, qui peuvent être généraux ou spécifiques.
Cette approche modulaire permet une meilleure modélisation des phénomènes, plus compartimentée, et un meilleur contrôle sur le résultat. Elle permet aussi une mise au point des règles et une évolutivité plus simple. Ces principes sont gérés par un traitement en cascade des règles, y compris de liage et de correction.

Par le biais des différentes contraintes introduites dans DisLog, il est possible de produire des représentations étiquetées complexes, sous forme d'arbres, de graphe ou de dépendances. DisLog permet d'introduire des relations de un vers plusieurs ou de plusieurs vers un, permettant ainsi qu'une structure soit en relation avec plusieurs autres structures de natures différentes. Cependant, dans la plupart des textes étudiés, ces relations sont relativement simples, le souci étant en général de préserver l'intelligibilité des documents.

Le formalisme des règles, DisLog, permet d'introduire tout type de forme de raisonnement a priori. Ceci est un point original et crucial en analyse du discours, facilité par l'implémentation réalisée en Prolog. Ces formes de raisonnements permettent entre autres (1) de réaliser des calculs, reportés dans les annotations produites, (2) de lever des ambiguïtés d'analyse, (3) de compléter l'analyse grammaticale par l'appel par exemple à des connaissances (pour inclure des données pragmatiques). Si une requête de raisonnement échoue, alors la règle échoue.

La littérature est particulièrement abondante s'agissant de l'analyse du discours. On y trouve plusieurs directions. Un mouvement théorique assez important s'est développé depuis 20 ans environ, autour de plusieurs cadres dont la DRT et ses extensions. Ce cadre demeure essentiellement abstrait et orienté vers des modes de représentations peu expressifs. Notre orientation étant à la fois plus empirique et liée à une sémantique conceptuelle plutôt que formelle, ce cadre n'est a priori pas pertinent pour nos travaux et ne sera pas évoqué ici. Plusieurs approches empiriques sont par contre d'un intérêt marqué. Il y a tout d'abord les travaux qui caractérisent la nature et la forme des relations rhétoriques. (Mann et al 88) ont proposé une formulation contemporaine de ce cadre. De nombreux travaux ont suivi dont (Delin 94), (Kosseim et al 00), (Rossner et al 92), (Saito et al 06), (Vander Linden 93), etc. qui affinent ces relations ou les étudient dans des cadres spécialisés. cependant, on assiste alors à une prolifération de ces relations, où les définitions deviennent parfois vagues.

Un courant plus profond se préoccupe du sens véhiculé par ces relations, dans une perspective cognitive, de leur aspects pragmatiques ainsi que des intentions de communication sous-jacentes (Wright 04), (Moschler 85) (Davidson 63) (Anscrombre et al 81). Ceci est particulièrement intéressant dans différents cadres tels que le dialogue, l'argumentation et la négociation (Amgoud et al. 01, 05), et la génération de langue naturelle (Rosner et

al. 92) qui, au niveau de son composant de planification, s'appuie en particulier sur des schémas rhétoriques.

Une interrogation toujours d'actualité concerne la caractérisation en langue des relations de discours afin de pouvoir les identifier automatiquement. C'est de toute évidence un défi ouvert, où des solutions parfois parallèles ont été tentées, car ces relations n'ont pas systématiquement des marques qui les identifient. Notons par exemple (Mann et al. 88), (Saito et al. 06), (Takechi et al 03) (Di Eugenio et al 96) qui soulignent bien les résultats que l'on peut escompter. Récemment, par exemple via le projet ANR Annodis, une approche à base d'apprentissage à partir d'annotations manuelles s'est développée. Une telle entreprise se heurte à deux difficultés : les désaccords importants (mais inévitables) entre annotateurs et la difficulté de développer de l'apprentissage sur des segments textuels importants où peu d'information est en fait pertinente. Enfin, notons les travaux qui développement des grammaires pour le texte par exemple à partir de TAGs (Gardent 97) (Webber 04).

Au niveau des environnements, GATE (http ://gate.ac.uk/) est une plateforme très répandue et qui intervient dans de nombreux projets. Elle est essentiellement dédiée à l'analyse de phrases ou de courts fragments de textes. Par ailleurs, Linguastream
(http ://www.linguastream.org/home.html) est une plateforme ouverte pour l'analyse du langage qui peut accepter en entrée tout type de texte XML. Il est basé sur une architecture en composants et offre plusieurs API Java utiles pour l'intégration. C'est un système largement ouvert qui laisse une grande liberté à l'utilisateur tout en lui proposant un ensemble d'outils d'aide et d'interfaces très pertinents. Linguastream est d'abord dédié à l'analyse de la phrase au sein de textes. Cette plateforme ne permet pas d'inclure de modules de raisonnement comme cela est très utile en analyse de structures discursives, par exemple pour lever des ambiguités ou pour introduire des considérations pragmatiques. Il serait toutefois intéressant de voir avec ces plateformes comment on peut écrire des analyseurs de structures de discours.

Au niveau des grammaires de discours, nous pensons qu'il est nécessaire de préserver une analyse linguistique précise, qui permet de décrire les phénomènes à un bon niveau d'abstraction, en préservant une certaine prédictibilité. Nous nous attacherons donc à un travail essentiellement manuel, même si des traitements automatiques sur corpus sont utilisés pour explorer les constructions (par exemple par bootstrapping). Ce texte étant dédié à la partie grammaire, cet aspect méthodologique est traité ailleurs.

1.2 Le langage des structures du discours

Il n'est pas dans notre objectif d'argumenter pour les différents aspects un langage élargi qui rende compte de l'ensemble des formes que peuvent prendre les structures qui relèvent du discours. Nous nous contenterons d'en observer un certain nombre, fortement récurrentes dans les situations que nous avons examinées, et qui sont à la base du langage de description DisLog introduit dans <TextCoop>. Ce langage est conçu de façon assez ouverte pour pouvoir permettre de coder de nombreuses configurations.

Les relations rhétoriques (Mann et al 88) sont structurées sous forme de deux types de relations :
– une relation hiérarchique dite de noyau vers satellite. Ainsi dans *insérez verticalement la carte mère sinon vous risquez d'endommager les connecteurs'*, la première partie de l'énoncé est habituellement appelée conclusion d'argument (ici de type avertissement) tandis que la seconde partie, qui explique les risques encourus, est appelée support de l'argument. La conclusion peut apparaitre seule, avec un support vide, mais le support n'a de sens que s'il est relié à au moins une conclusion.
– une relation non hiérarchique de noyau vers noyau. Ainsi la relation 'parallèle' associe-t-elle deux structures de même niveau, comme dans l'ellipse : *Jean est reçu à son permis, Marie aussi*.
L'analyse discursive des textes s'applique souvent sur ces deux types de schémas. Cette structure est bien entendu récursive ou emboitée : un satellite peut être lui même une structure composite complexe.

On observe cependant des situations plus complexes. Ainsi un noyau peut-il gouverner plusieurs satellites, éventuellement de statuts différents (une définition suivie d'un conseil, un avertissement au milieu d'instructions). Un cas courant de multiplicité de satellites dans les textes procéduraux est celui d'un titre (énoncé d'un but) et des instructions qui permettent de réaliser ce but. Dans un texte, le réseau de relations devient alors très complexe, suivant une structure de graphe orienté. de plus, les relations noyau-satellite(s) sont souvent floues et ambigues et varient selon le point de vue. Enfin, le caractère hiérarchique de certaines relations est difficile à établir et peut dépendre du contexte.

Comme cela est souvent indiqué dans la littérature sur les relations du discours, celles-ci peuvent être en très

grand nombre, et de définitions variables selon les auteurs ou les annotateurs. Cette situation est nettement plus complexe, pour faire un parallèle, que dans le cas de la relation prédicat arguments ou ajouts, où des relations thématiques sont souvent employées et relativement bien maitrisées. Si l'on considère par exemple des textes techniques, on observe que chaque genre peut avoir quelques structures spécifiques. C'est le cas par exemple des instructions, du sommaire et des pré-requis dans les procédures.

1.3 Caractériser les structures discursives

De nombreux auteurs se sont attaqués au problème difficile qui consiste à définir un formalisme grammatical pour reconnaitre les structures discursives. Nous faisons ici en quelque sorte une synthèse des principales difficultés. Si l'on veut caractériser la structure grammaticale d'une structure discursive, il convient de définir avec précision :

– Comment identifier un objet discursif, sur quelles bases linguistiques, pragmatiques, typographiques, etc. (DiEugenio et al. 96). Certaines relations sont relativement bien marquées dans la plupart des cas, alors que d'autres sont rarement marquées ou ne s'y prêtent pas. Les discours en langue contrôlée ou à visée finalisée, visant l'efficacité et la clarté, développent en général des marques nettement plus évidentes, comme, par exemple, dans le discours procédural ou didactique. Les marques peuvent de surcroit être ambiguës entre plusieurs relations. Assez souvent, enfin, on remarque qu'un noyau se trouve identifié parce qu'un satellite a été identifié et peut lui être associé : les satellites sont souvent mieux marqués que leur noyau.
– Comment délimiter l'objet discursif une fois identifié ? D'autres marques (ponctuation, typographie, connecteurs, etc.) peuvent être considérées en complément des marques identifiantes indiquées ci-dessus. Une approche intéressante vise à développer la notion de structure de discours élémentaire (EDU (Schauer 06)), comparable, dans la phrase, à la proposition. Il convient d'en évaluer l'utilisabilité. Des groupes d'EDUs peuvent parfois considérées et évaluées comme étant une structure discursive autonome (par exemple sur la base de la théorie du centrage).
– Comment, une fois une structure discursive identifiée, la relier à une ou plusieurs autres structures ? La difficulté est ici d'identifier les structures exactes à mettre en relation, par exemple une énumération (satellite) doit être liée exactement à l'élément initiateur de cette énumération, comme par exemple un terme ou une expression plus générique (un titre). Les composants sont souvent contigüs, mais leur délimitation peut s'avérer très difficile.

On se trouve donc devant une triade : **identification, délimitation, liage** (des différents protagonistes de la relation).

2 Le formalisme grammatical de <TextCoop>

Le formalisme grammatical de <TextCoop>, DisLog, étend les possibilités expressives des approches à base d'expressions régulières et les adapte aux besoins de l'analyse de discours, en y intégrant un composant de raisonnement souvent utile dans l'analyse de telles structures. Notre approche s'appuie aussi sur les travaux, maintenant assez anciens, mais toujours actuels, des grammaires logiques (DCGs, XGs, MGs, etc.), qui s'appuient sur des modèles d'exécution inspirés des programmes logiques, dont Prolog. D'autres schémas de stratégie, par exemple à base de contraintes ou utilisant du parallélisme ET-OU sont possibles. Cette approche nous parait intéressante pour l'analyse du discours en raison de son caractère déclaratif marqué, de son indépendance relative aux stratégies de traitement, et aussi de son aptitude à intégrer naturellement des modules de raisonnement et des structures de contraintes puissantes (par exemple des structures de traits typés, des contraintes d'arbres).

Le formalisme que nous proposons ici peut aussi bien permettre de coder des règles d'analyse de structures discursives conçues par des linguistes et codées manuellement que des règles issues de mécanismes d'apprentissage à partir de textes annotés, qui produisent en sortie des formes contraintes. Il est aussi possible de coder globalement une forme noyau-satellite que ces mêmes formes séparemment. Ce dernier choix est nécessaire lorsque la combinatoire entre noyau et satellite est élevée, ou que ces constituants sont discontinus, ce qui est assez fréquent.

2.1 DisLog : le formalisme grammatical

Le langage DisLog offert par <TextCoop> comprend les symboles suivants, ils suivent en général la syntaxe de Prolog et des DCGs (grammaires à clauses définies) :

– des **symboles pré-terminaux et non terminaux**. Les pré-terminaux se dérivent directement en des entrées lexicales ou des expressions (caractéristiques de structures rhétoriques ou de domaines, par exemple) ou bien encore en des marques typographiques ou des marques d'annotations (html, XML, ...). Ces marques d'annotations peuvent faire référence à des structures déjà identifiées. Les symboles non terminaux font appel à des grammaires, essentiellement à caractère local (par exemple grammaire des expressions temporelles) ou, plus rarement, des constructions standard de la langue (SN, SV, etc.). Les règles ne s'appellent pas entre-elles. Les liens entre structures sont réalisés par des opérations de liage sélectif (voir ci-dessous). Les symboles non terminaux et préterminaux peuvent être associés à des structures de traits attribut-valeur, ceci ne sera pas développé ici, tant bien connu. Enfin, ces symboles sont utilisés soit pour identifier un type de structure discursive soit comme élément de délimitation (inclus ou exclu). Lorsqu'ils sont exclus, ils apparaissent dans un prédicat 'borne'.
– des symboles terminaux indiqués entre crochets, cette possibilité est utile lorsqu'il y a peu de choix sur ces terminaux au sein d'une règle, dans le cas contraire, il est péférable de faire appel à un préterminal,
– des indications d'optionalité ou d'itérativité sur les symboles préterminaux et non terminaux,
– des symboles permettant d'exprimer la précédence linéaire (la ','), ainsi que la co-occurence de symboles (le ';') si l'on veut utiliser la forme abrégée des règles (non développée ici),
– des 'gaps' qui représentent des séquences finies de mots qui ne présentent pas d'intérêt pour la règle en cours de description. La condition d'arrêt est constituée par le symbole explicite qui suit le gap. Dès qu'un tel symbole est rencontré, le gap s'arrête. Les gaps peuvent être associés à des contraintes, en particulier des symboles terminaux ou non-terminaux qui ne doivent pas être ignorés. Si un gap rencontre un tel symbole avant d'atteindre sa condition d'arrêt alors il y a échec de la règle à reconnaitre la structure. Un gap ne peut ni commencer ni terminer une règle, il doit toujours être borné explicitement par un symbole ou un terminal.
– des appels à des prédicats qui introduisent des contraintes, des connaissances à intégrer ou des calculs divers. Ceux-ci sont représentés entre accolades comme dans les DCGs.
– des fonctions d'assignation, explicites ou par défaut, d'étiquettes dédiées permettant d'étiqueter les structures reconnues avec d'éventuels attributs, calculés par les prédicats ci-dessus.

A priori, les règles sont de type 2, avec la syntaxe des DCGs. Toutefois des règles de type 1 peuvent aussi être construites. Le symbole en partie gauche de règle contient une variable qui représente le résultat : en général il s'agit de la structure complète telle que lue avec des marques d'annotation, éventuellement avec des attributs, au début et à la fin correspondant à la structure reconnue. Il est aussi possible de repositionner des composants du texte lus en entrée.

A titre d'exemple, des expressions d'avertissement du type *il est conseillé de ne jamais ACTION parce que....* se représentent simplement (i.e. sans faire de généralisation) comme ci-dessous (Fontan et al. 08). On considère ici que la structure commençant par *parce que* ne fait pas partie de l'avertissement (en fait c'est un support de l'avertissement selon les théories de l'argumentation, celui-ci est reconnu séparemment) :
avertissement(R) –> [il, est], expr([type :conseil]), [de], negation, gap([connecteur([type :cause])), borne([parce, que]).

Cette règle débute par la mention de deux terminaux, qui font partie de la structure à reconnaitre, suivie d'un pré-terminal de type conseil (indiqué ici par la structure attribut valeur dans l'argument). Elle se poursuit par un autre terminal, un non terminal qui reconnait la négation, puis un gap dont on indique qu'il ne doit pas ignorer les connecteurs de cause sur son parcours qui se termine sur la borne (une marque externe à la règle, voir ci-dessous) qui est le terminal [parce,que]. La variable R représente la structure étiquetée, qui est ici de la forme :
<avertissement> ... texte lu ... < /avertissement>.

2.2 L'insertion d'étiquettes XML

DisLog prévoit la possibilité (1) de spécifier d'autres types d'étiquettes que celles liées au symbole en partie gauche, (2) d'inclure des attributs, et (3) d'insérer à tout endroit du segment de texte lu tout autre type d'étiquette. Considérons :

<avertissement> il est <exp-conseil force="modéré"> recommandé < /exp-conseil> de ne jamais ouvrir la boite < /avertissement>

Dans cet exemple, on a inséré une balise <exp-conseil> avec un attribut (déduit de propriétés lexicales) qui

indique la force du conseil, ici 'modéré'.

L'insertion d'étiquettes XML se fait comme suit :
– par défaut en début et fin de séquence reconnue, en utilisant le non terminal donné en partie gauche de règle,
– si l'on veut insérer une autre étiquette, alors celle-ci est spécifiée explicitement par une variable d'insertion. Ceci vaut aussi si on veut ajouter un ou plusieurs attributs.
– si l'on veut insérer des étiquettes complémentaires dans la séquence reconnue, celles-ci sont aussi spécifiées par des variables d'insertion dans la partie droite de règle.
– si, d'aventure, on ne veut rien insérer en début et fin de séquence, on emploie la notation $noinsert.

Les variables d'insertion sont représentées par : $insert1, $insert2, qui sont instanciées explicitement en fin de règle dans une section 'calculs' entre accolades. On peut employer des variables qui proviennent soit de déductions soit de caractéristiques héritées de données lexicales. Pour l'exemple ci-dessus, on doit ajouter dans la règle :
avertissement(R) –> [il, est], $insert1, expr([type :conseil, force :F]), $insert2, [de], negation,
gap([connecteur([type :cause])), borne([parce, que]), { $insert1= <exp-conseil, force :F>, $insert2=< /exp-conseil> }.

avec la donnée lexicale : expr([type :conseil, force :modéré]) –> [recommandé].

2.3 Les règles de liage sélectif

Les règles de liage sélectif permettent de lier deux structures ou plus comme évoqué dans l'introduction. L'objectif est de lier noyau et satellite(s), ou tout autre lien que l'on souhaite établir (par exemple, connecteur - EDU, etc.). Les règles de liage ont donc un statut de non-terminaux : elle lient entre-elles des règles, permettant de construire des arbres partiels dans un texte, indiquant les structures discursives qui sont en relation. Par exemple si l'on veut lier les structures discursives a et b pour former c, on peut définir une règle de liage comme suit :
c(R) –> [<a>], gap, [< /a>], gap, [<b>], gap, [< /b>].
ce qui produira : <c> <a>, ... < /a>, ... <b>, ... < /b> < /c>.
où les structures a et b sont reproduites telles quelles.

De façon plus concrète, si l'on considère la structure duale des arguments : conclusion-support, comme dans :
Il est capital d'insérer verticalement la carte mère car vous riquez d'endommager les connecteurs.,
on aurait la règle de liage sélectif :
argument(R) –> [<conclusion>], gap, [< /conclusion>], connecteur([type :cause]), [<support>], gap, [< /support>].
et la structure résultante est :
<argument> <conclusion> Il est capital d'insérer verticalement la carte mère < /conclusion>, car <support> vous riquez d'endommager les connecteurs < /support> < /argument>.

2.4 Les règles de correction

<TextCoop> permet la définition de règles de réécriture sur les balises permettant de repositionner certaines balises qui pourraient ne pas être positionnées correctement.

Un emploi immédiat est lié au ré-équilibrage des balises qui peuvent se chevaucher, en particulier lorsque les conditions de délimitation d'une règle sont trop peu contraintes. Typiquement, ces règles permettent de corriger une situation telle que :
<a>, ...<b> < /a>, ... < /b> en <a>, ... < /a>, ... <b>, ... < /b> .

Les règles de correction ont la même forme que celle ci-dessus, la variable R contenant la structure corrigée :
corriger([<A>], gap, [< /A>], gap, [<B>], gap, [< /B>]) –> [<A>], gap,[<B>],gap, [< /A>], gap, [< /B>].
On notera qu'ici A et B sont des variables représentant a priori n'importe quel identifiant de balise.

2.5 L'art d'écrire des règles

Dans notre approche, les règles sont écrites pour l'instant totalement de façon manuelle, à partir d'analyse de documents et de repérage de marques. Toutefois, le formalisme a été conçu pour accueillir les spécifications,

relativement ouvertes, de systèmes basés sur l'apprentissage. Nous souhaitons aussi introduire des outils d'aide à l'écriture de règles, par exemple basés sur des techniques de bootstrapping. Ceci reste toutefois à approfondir car l'analyse de discours a des caractéristiques très différentes de celle de la phrase où bootstrapping et apprentissage ont été largement testés. Notre expérience est que l'écriture de règles qui reconnaissent des structures de discours et les lient sont complexes : elles présentent peu de marques explicites, ce qui les rend ambiguës, elles couvrent aussi très souvent des fragments de texte conséquents. Ces raisons font que nous avons privilégié dans notre approche une écriture manuelle des règles que nous associerons à un ensemble d'outils de visualisation de façon à guider les auteurs. Cette écriture, nous pensons, permet d'accéder à une meilleure adéquation linguistique et un meilleur niveau de généralisation.

Par exemple, c'est à ce niveau que des mécanismes de raisonnement peuvent être introduits. Dans le cas ci-dessous :
La confiture se prépare avec des fruits rouges (cassis, fraises, framboises) afin de ...
pour identifier que la structure entre parenthèses est une illustration, faute de marques explicites, via le contrôle : cassis est_un 'fruit rouge', etc. une terminologie simple est nécessaire. La règle suivante intègre noyau et satellite et s'écrit, par exemple :
illustration(R) –> Nom(Type), ['('], liste_Noms(Type1), [')'], { subsume(Type,Type1) }.

Enfin, l'analyse de structures du discours mène à de nombreuses ambiguités. Les règles décrites par les auteurs reflètent ces ambiguités. Notre système est conçu soit pour refléter toutes les analyses possibles (mode mise au point) soit pour privilégier un choix a priori (via le mécanismes de cascades de règles ou des heuristiques). Toutefois, les techniques d'interprétation des programmes logiques pourraient permettre des modes intermédiaires, ainsi qu'une interprétation à base de contraintes 'actives' permettant de produire l'ensemble des analyses possibles.

2.6 Gestion de la concurrence entre règles

Nous proposons ici quelques contraintes qui gèrent la concurrence entre règles. Les règles étant parfois trop permissives (ou pourrait parler de principe productif comme pour la syntaxe X-bar), il est nécessaire d'ajouter des contraintes qui en limitent la puissance (on pourrait parler de principes restrictifs). Notre système fonctionne par segments appelés unités textuelles. Celles-ci peuvent être des paragraphes, des sections, des arbres (pour des documents semi-structurés, etc.). Nous ferons une présentation ici illustrée de ces contraintes.

Certaines structures discursives ont des formes très proches, difficiles à distinguer. <TextCoop> offre actuellement deux possibilités pour gérer la concurrence au niveau de la reconnaissance de structures. Le moteur de <TextCoop>, présenté ci-dessous, exécute les règles en cascades. Le langage de <TextCoop> permet de spécifier l'ordre dans lesquels les règles sont exécutées.

Nous appellerons ci-dessous *paquet de règles* l'ensemble des règles qui sont liées à la reconnaissance d'une structure donnée (par exemple, support d'argument, illustration, reformulation) identifiée par l'emploi d'un symbole identique en partie gauche de règle. L'analogie avec les paquets de clauses en Prolog est immédiate. <TextCoop> permet de spécifier une structure d'ordre (éventuellement partiel) qui indique dans quel ordre les paquets de règles doivent être exécutés. Ainsi dans :
titre < prérequis < sommaire.
les règles reconnaissant les titres seront exécutées d'abord puis celles liées aux pré-requis, etc. sans possibilité de retour arrière.

Associé à ce mécanisme de cascades, il est possible de définir des *zones fermées* où une fois une zone reconnue, aucune règle ne pourra être appliquée sur cette zone. Une zone fermée est définie sur un segment de texte inclus dans une balise ouvrante et fermante du même type. Lorsque la zone est identifiée, il n'est pas possible de tenter d'appliquer d'autres règles à l'intérieur de cette zone. Par exemple, la balise <titre> définie dans le traitement des procédures introduit une telle zone :
zone_fermee([titre]).
Un titre ressemble en effet, quant à sa structure, à une instruction : *monter votre mezzanine*, on ne veut pas, une fois un titre reconnu, l'étiqueter aussi comme une instruction. <TextCoop> permet de définir une liste de zones fermées. Cette liste comprend des structures de deux types : (1) des structures qui peuvent être reconnues par plusieurs paquets de règles, mais où l'on veut privilégier un choix et ne pas provoquer de double étiquetage, afin de limiter les problèmes d'ambiguités (cut 'rouge' en Prolog), (2) des structures discursives 'terminales', c'est à

dire qui ne doivent pas être davantage décomposées, tout au moins par rapport à la grammaire de discours telle qu'elle est écrite. Par exemple, la structure de pré-requis (liste d'ingrédients ou d'équipements) est analysée dans les procédures comme une structure terminale (mais elle pourrait être plus finement analysée dans un autre cadre). Ce second cas est défini d'abord pour des raisons d'efficacité (comme un cut 'vert' en Prolog).

Outre la possibilité de zones fermées, DisLog offre la possibilité de spécifier d'autres types de contraintes. Tout d'abord on peut indiquer si une structure doit en dominer une autre :
dom(instruction,but).
indique que toute structure de type but est dominée par une instruction. De la même façon, on peut indiquer que deux structures doivent être dans deux branches différentes de la structure reconnue, sans aucune dominance possible :
not_dom(instruction,avertissement).
Une instruction ne peut contenir un avertissement. Enfin, nous permettons de spécifier directement une relation rhétorique :
rel_rhetorique(noyau,satellite,structure_englobante)., comme dans :
rel_rhetorique(conclusion_avt,support_avt,avertissement).
comme décrit ci-dessus. A ce stade aucune contrainte de précédence n'est donnée. Si l'on veut que le noyau soit toujours avant le satellite, il faut ajouter :
prec(conclusion_avt,support_avt). La contraite rel_rhetorique inclut donc la contrainte 'sister' et un liage sélectif des deux premiers éléments spécifiés.

3 Le moteur de <TextCoop>

Nous décrivons ici le fonctionnement du moteur <TextCoop> qui est, pour l'heure, un interpréteur. Nous indiquons d'abord la forme des règles telles que traitées par cet interpréteur puis le fonctionnement du moteur lui-même. D'autres fonctionnements peuvent être envisagés, sur le même schéma que celui des DCGs qu'il ne fait que généraliser. Nous ne nous étendrons pas ici sur les aspects théoriques du système. Les textes analysés étant composés de phrases à nombre de mot finis, les gaps ignorant eux aussi des suites finies de mots, on peut, via énumérabilité récursive, appliquer le théorème du point fixe pour donner une sémantique déclarative 'simple', comme dans les programmes logiques en général.

3.1 Traduction des règles

Le moteur fonctionne comme un interpréteur, les règles ainsi que les divers dispositifs présentés ci-dessus sont donc traduits sous forme de structures de données directement utilisables par le moteur. Pour les règles, cette structure est proche de celle définie pour les DCGs dans un mode interprété, elle a la forme suivante :
forme(Identifiant,Entrée,Sortie,Partie_droite,Résultat).
où :
– Identifiant est le nom du symbole en partie gauche de règle,
– Entrée et Sortie contiennent le texte en cours de traitement, ceci traduit la technique des listes de différences des DCGs, dont nous avons un besoin explicite ici pour reconstruire le texte étiqueté,
– Partie_droite est la partie droite de règle, développée ci-dessous,
– Résultat est la variable R, résultat de l'analyse (avec étiquetage) comme présenté ci-dessus.

Les symboles en partie droite sont représentés sous forme de liste, après développement complet de la règle lorsque la forme abrégée est utilisée. Les symboles terminaux et non terminaux sont augmentés de trois arguments :
– un argument qui représente la chaine de mots couverte par ce symbole,
– les deux variables qui représentent la liste de différence propre à ce symbole lorsqu'il est développé.

Ainsi, le symbole préterminal : expr([type :conseil]) qui représente une expression (terminale) de type conseil est-il traduit en : expr(EXPR, [type :conseil],E0,E1). En ce qui concerne le symbole gap, les trois mêmes variables sont ajoutées ainsi qu'une marque qui indique le symbole d'arrêt du gap, cette marque est l'identifiant du symbole qui suit le gap dans la règle et ses contraintes.

Plus globalement, la règle suivante, qui traite de la structure d'une conclusion de conseil :
concl_conseil –> pro(_), aux([ty :etre]), gap([supconseil]), expr([type :conseil]), gap([supconseil]), mfin(_).

est traduite de la façon suivante :

```
forme(c-cons-fr, E, S, [  pro(PRO,_,E,E2), aux(AUX,etre,E2,E3),
     gap(supconseil, [expr,conseil], E3,E4,Saute1),expr(EXPR,conseil,E4,E5),
     gap(supconseil, [mfin,_], E5,E6,Saute2),   mfin(MFIN,_,E6,S)], [], % no reasoning
    [ '<concl-cons>' ,PRO, AUX, Saute1, EXPR, Saute2, '</concl-cons>', MFIN ]).
```

'aux' demande l'auxiliaire être, 'mfin' est une marque de fin, répertoriée dans le lexique. Les deux symboles gap contiennent les variables Saute1 et Saute2 qui indiquent la chaine de mots qu'ils ont parcourue, et qui sera ignorée, jusqu'à rencontrer le symbole suivant dans la règle. L'argument qui contient les restrictions (le seul présent dans la structure initiale) peut être soit une constante Prolog, dont l'interprétation est directe soit une structure de traits, qui est interprétée de façon standard, avec la subsomption sur les traits sémantiques.

3.2 Le fonctionnement global du moteur

Le moteur de <TextCoop> est écrit en Prolog SWI, et tous les composants linguistiques attachés sont aussi implémentés dans ce cadre. Nous avons cherché à optimiser les traitements sans que ceci soit la priorité : l'objectif sont les traitements de structures du discours en mode batch.

Comme indiqué ci-dessus, le moteur fonctionne par cascades de règles, suivant les priorités énoncées. Il n'y a pas de retour arrière sur les étapes précédentes d'une cascade. Une étape de cascade est identifiée par le ou les symboles en partie droite de règle : des paquets de règles sont donc exécutés à chaque étape. Au sein d'une étape, les règles dans un paquet sont exécutées dans l'ordre dans lequel elle sont écrites. Une option du moteur permet d'éviter les retours arrière (couteux) sur ces règles. Cette option est désactivée en phase de mise au point. les contraints énoncées sont vérifiées à chaque étape de l'exécution des règles.

A titre expérimental, l'interpréteur est pourvu, outre la stratégie de traitement de la gauche vers la droite, d'une stratégie inverse, de la droite vers la gauche. Celle-ci est utilisée lorsque les marques identifiantes d'une relation rhétorique sont placées à la fin de la structure, alors que la structure ne débute que par une marque de délimitation. C'est le cas, par exemple des structures d'illustration du type : *(a,b,c,.., par exemple)*, ou la parenthèse ouvrante sert de délimiteur (elle est peu discriminante de la relation) et où la marque 'par exemple' est située en fin de chaine. Les résultats que nous obtenons indiquent un meilleur taux de reconnaissance et une meilleure efficacité. Cette stratégie sera automatisée, à partir de l'analyse de la position des marques identifiantes.

Le résultat final de l'analyse est le texte donné en entrée augmenté de balises XML qui délimitent et caractérisent les structures analysées.

3.3 Evaluation

La première version du moteur est à présent disponible avec son environnement. Celle-ci a été réalisée dans l'objectif de valider les idées et de définir précisémment les fonctionnalités internes et externes utiles en analyse discursive. L'optimisation du code n'y est que partielle. Les performances du système actuel sont les suivantes, sur la base de 30 patrons (liés au traitement des procédures), avec un lexique de 1300 mots ou expressions et un accès à un analyseur morphologique, lui aussi en Prolog. Le calcul se fait sur un PC standard. Nous traitons 60 Mo de texte (hors balises) par heure. La taille du lexique dans cette expérience est assez importante pour le traitement du discours. Dans de nombreuses applications, le lexique utile peut être nettement plus réduit. Ceci a un impact important sur les performances. Le lexique qui a servi à l'expérimentation contient 800 verbes, cependant, dans une application d'analyse de procédures, ce nombre descend à environ 150 verbes, et à des variations morphologiques très réduites. On peut alors traiter jusqu'à 200 Mo de texte par heure. Bien entendu la complexité des patrons a aussi une influence sur les performances : les gaps, la longueur des règles, les restrictions mais aussi le non déterminisme introduit par les règles d'un même paquet sont des facteurs importants, mais qui restent difficiles à analyser finement.

Il est nettement plus difficile d'évaluer la qualité de reconnaissance des règles. Cela dépend beaucoup des types de textes traités et de la qualité de conception des règles, comme dans tout type de programmation. En ce qui concerne les procédures, des résultats détaillés sont donnés dans (Fontan et al. 08). Les textes professionnels

donnent des résultats très bons, du fait de leur qualité de rédaction et de leur régularité, aussi bien au niveau de l'écriture des instructions que de celle des conseils, avertissements, définitions, commentaires ou illustrations. Nous travaillons actuellement au dévelopement de techniques basées sur le bootstrapping, mais adaptées à la problèmatique du discours, visant à améliorer la définition des règles pour un type de phénomène donné. Cela inclut les aspects structurels aussi bien que les aspects lexicaux, en particulier la caractérisation fine des marques. Ces marques sont de types très diversifiés. Pour bien organiser nos ressources lexicales, il convient d'élaborer, au sein du système, une architecture souple et modulaire, mais non redondante, des ressources lexicales (marqueurs, classes sémantiques de verbes, expressions dédiées, etc.).

4 L'environnement linguistique de <TextCoop>

Nous avons présenté essentiellement dans ce document le moteur de <TextCoop> ainsi que le formalisme des règles. Il est clair qu'un tel système ne peut fonctionner qu'en s'appuyant sur un ensemble de ressources lexicales, et ne peut être véritablement intégré dans des traitements de grande échelle que si, outre ses performances, ses modules de ressources (lexique, grammaires locales, ontologies et terminologies, etc.) et formats d'entrée sortie des documents suivent des formats normalisés. Il est essentiel aussi qu'il dispose d'un module de mise au point des règles, de visualisation des résultats et des recommandations et une méthode pour son intégration dans des applications.

Pour l'heure, nous avons développé et intégré différents modules de ressources qui sont plus particulièrement pertinents pour l'analyse du discours que l'on peut utiliser tels quels dans les règles, comme par exemple :
- des listes de connecteurs typés : de cause, concession, temps, etc.
- des listes de termes spécifiques pouvant être utilisés comme marques : marques de l'illustration, de la reformulation, du développement, etc.
- des listes de verbes par classes et sous-classes sémantiques, inspirées des classes de WordNet,
- des listes de termes avec polarité négative ou positive, ceci est utile en argumentation et en analyse d'opinions.
- des grammaires locales : expression du temps, de la quantité, expressions évaluatives simples, etc.
- enfin, des modules qui contiennent quelques règles simples pour reconnaitre des structures telles que : illustration, reformulation, but, condition, définition et élaboration (une relation nettement plus complexe et composite). De façon plus générale, nous souhaitons étudier comment on peut greffer un analyseur de phrases au sein de <TextCoop> pour le compléter.

En matière de visualisation des résultats, nous avons conduit une expérimentation très concluante en utilisant NAVITEXTE (http ://panini.u-paris10.fr/jlm/?Start :projets :NaviTexte), même si cette interface est un peu trop élaborée pour nos besoins. Nous envisageons à présent une visualisation qui soit davantage sous forme d'abre, tout en préservant la structure du document original. En matière de normalisation des données et de certains traitements, nous étudions le cadre de UIMA (http ://www.uima-fr.org/) qui est certainement une direction forte. L'aide à la mise au point des règles est difficile à concevoir dans le cadre du discours. Dans notre cadre, elle se limite actuellement à quelques recommandations sur leur forme et la façon de les simplifier ou de les généraliser. Nous envisageons de développer un analyseur qui détecterait certains types d'erreurs ou d'incompatibilités entre règles, comme cela existe dans certains compilateurs de règles. Etant dans une étape qui demeure expérimentale, il nous faut bien identifier les besoins 'raisonnables' en matière de mise au point de règles et de ressources avant d'en réaliser une implémentation. L'architecture des ressources linguistiques utilisées dans les règles, pour éviter les doubles et les incohérences est aussi un sujet d'étude actuel.

5 Les applications

Le projet <TextCoop> est à l'origine dédié à l'analyse des procédures (Delpech et al. 2008). Nous avons donc conduit nos premières expérimentations sur les structures liées aux procédures, qui ont des caractéristiques assez diversifiées et qui permettent de bien tester à la fois le moteur et l'écriture des règles. Un élément intéressant est que, dans ce projet, nous avons analysé à la fois des structures linguistiques connues et des structures dédiées (titres, instructions, etc.). Les caractéristiques globales de ces structures sont les suivantes :
- titres : ressemblent à des instructions, bien que souvent très elliptiques (verbe ou objet sous-entendu), les règles de reconnaissance s'appuient sur de nombreuses considérations typographiques.

– instructions : nous traitons ici en fait de composés instructionnels qui peuvent contenir plusieurs instructions élémentaires. Dans ces règles, les difficultés sont la délimitation des instructions ainsi que le grand nombre de symboles gaps introduits. L'identification des instructions se fait simplement sur la base de verbes et d'indications morphosyntaxiques.
– prérequis : longue structure sous forme de liste, pauvre en verbes (les gaps excluent les verbes), elle se situe en général au début du document, elle est parfois difficile à identifier à cause de conflits avec des sommaires ou de la publicité. La structure typographique, codée dans les règles, est un élément essentiel.
– conseils et avertissements, ont la forme d'arguments : séquence d'une conclusion (une instruction particulière) suivie d'un ou plusieurs supports (les difficultés à prévoir si on ne fait pas ce qui est demandé). Ces deux structures sont liées et demandent un traitement conjoint via deux paquets de règles. Des règles de liage lient les deux composants. Conseils et avertissements ressemblent à la structure des instructions, mais sont pourvus de marqueurs spécifiques (Fontan et al. 2009) riches et diversifiés. Ces structures sont traitées dans une étape de cascade qui précède celle des instructions. Des zones fermées évitent les doubles analyses.

Par ailleurs, selon le même schéma, nous avons développé une analyse d'avis consommateurs où nous faisons ressortir la structure du discours, globalement : type de produit, circonstances d'achat, liste d'arguments (propriétés avec des avis positifs et négatifs), recommandation, commentaires. Bien que ce travail soit encore dans un stade expérimental, <TextCoop> est bien adapté pour rendre compte de la structure globale de ce type de texte.

Enfin, <TextCoop> a été utilisé ponctuellement comme outil d'exploration à l'enrichissement de documents semi-structurés, comme par exemple l'ajout d'une zone de pré-requis dans les procédures qui n'en ont pas.

En perspective, via un projet ANR, LELIE, qui débute, <TextCoop> sera utilisé pour l'analyse et la prévention des risques industriels tels qu'ils peuvent apparaitre dans les procédures (mauvaise rédaction, non suivi d'exigences réglementaires ou métier). Dans ce cadre pré-industriel, <TextCoop> recevra un environnement plus professionnel.

6 Conclusion

Dans ce document, nous avons présenté les principales caractéristiques de <TextCoop>, un environnement basé sur les grammaires logiques dédié à l'analyse de structures discursives via le langage offert par DisLog. Nous avons détaillé en particulier la structure des règles et du moteur. Nous avons indiqué au fur et à mesure du texte l'état du travail, les performances et les orientations en particulier en matière d'environnement, d'aide à l'écriture de règles et de développement applicatif.

S'il existe plusieurs plateformes très élaborées pour l'analyse de la phrase et de structures plus petites, il y a peu de plateformes qui soient dédiées à l'analyse de structures du discours, qu'elles soient rhétoriques ou dédiées à des applications. Outre cette originalité relative, un élément important dans sa viabilité est le développement d'aide à l'écriture de règles et des éléments du langage qui y sont associés.

Le code <TextCoop> et des données linguistiques associées seront prochainement mis à disposition sous licence GPL.

Remerciements

Ce projet est soutenu par un projet de coopération franco-indien (IFCPAR) ainsi que par l'ANR (projet terminé TextCoop, et projet actuel LELIE).

Références

Amgoud, L., Bonnefon, J.F., Prade, H., *An Argumentation-based Approach to Multiple Criteria Decision*, in 8th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU'2005, Barcelona, 2005.

Amgoud, L., Parsons, S., Maudet, N., *Arguments, Dialogue, and Negotiation*, in : 14th European Conference on Artificial Intelligence, Berlin, 2001.

Anscombre, J.-Cl. Ducrot, O., *Interrogation et Argumentation*, in Langue francaise, no 52, L'interrogation, 5 - 22, 1981.

Aouladomar, F., Saint-Dizier, P., *An Exploration of the Diversity of Natural Argumentation in Instructional Texts*, 5th International Workshop on Computational Models of Natural Argument, IJCAI, Edinburgh, 2005.

Bouffier, A., Poibeau, T., *Re-engineering free texts to obtain XML documents : a discourse based approach*, RANLP 2007. Carberry, S., Plan Recognition in natural language dialogue, Cambridge university Press, MIT Press, 1990.

Cruse, A., *Lexical Semantics*, Cambridge Univ. Press, 1986.

Delin, J., Hartley, A., Paris, C., Scott, D., Vander Linden, K., *Expressing Procedural Relationships in Multilingual Instructions*, Proceedings of the Seventh International Workshop on Natural Language Generation, pp. 61-70, Maine, USA, 1994.

Delpech, E., Saint-Dizier, P., Investigating the Structure of Procedural Texts for Answering How-to Questions, LREC 2008, Marrakech.

Davidson, D., *Actions, Reasons, and Causes*, Journal of Philosophy, 60, 1963

Di Eugenio, B. and Webber, B.L., Pragmatic Overloading in Natural Language Instructions, International Journal of Expert Systems, 1996.

Lionel Fontan, Patrick Saint-Dizier. Constructing a Know-How Repository of Advices and Warnings from Procedural Texts. Dans ACM International Conference on Document Engineering, Sao Paolo, Dick Bulterman, Luiz Soares (Eds.), ACM, p. 234-240, september 2008.

Gardent, C., Discourse tree adjoining grammars, report nb. 89, Univ. Saarlandes, Saarbrucken, 1997.

Kosseim, L., Lapalme, G., *Choosing Rhetorical Structures to Plan Instructional Texts*, Computational Intelligence, Blackwell, Boston, 2000.

Mann, W., Thompson, S., Rhetorical Structure Theory : Towards a Functional Theory of Text Organisation, TEXT 8 (3) pp. 243-281, 1988.

Marcu, D., The Rhetorical Parsing of Natural Language Texts, ACL 1997.

Marcu, D., Au unsupervised approach to recognizing Discourse relations, ACL 2002.

Moschler, J., *Argumentation et Conversation, Eléments pour une Analyse Pragmatique du Discours*, Hatier - Crédif, 1985.

Rosner, D., Stede, M., *Customizing RST for the Automatic Production of Technical Manuals*, in R. Dale, E. Hovy, D. Rosner and O. Stock eds., Aspects of Automated Natural Language Generation, Lecture Notes in Artificial Intelligence, pp. 199-214, Springler-Verlag, 1992.

Saito, M., Yamamoto, K., Sekine, S., Using Phrasal Patterns to Identify Discourse Relations, ACL, 2006.

Schauer, H., From Elementary Discourse Units to Complex Ones, ACL 2006.

Takechi, M., Tokunaga, T., Matsumoto, Y., Tanaka, H., *Feature Selection in Categorizing Procedural Expressions*, The Sixth International Workshop on Information Retrieval with Asian Languages (IRAL2003), pp.49-56, 2003.

Vander Linden, K., *Speaking of Actions Choosing Rhetorical Status and Grammatical Form in Instructional Text Generation* Thesis, University of Colorado, 1993.

Webber, B., D-LTAG : extending lexicalized TAGs to Discourse, Cognitive Science 28, pp. 751-779, Elsevier, 2004.

Wright, von G.H., Explanation and understanding, Cornell university Press, 2004.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Vers une algèbre des relations de discours pour la comparaison de structures discursives

Charlotte Roze
Alpage, INRIA Paris–Rocquencourt & Université Paris 7
charlotteroze@linguist.jussieu.fr

Résumé. Nous proposons une méthodologie pour la construction de règles de déduction de relations de discours, destinées à être intégrées dans une algèbre de ces relations. La construction de ces règles a comme principal objectif de pouvoir calculer la fermeture discursive d'une structure de discours, c'est-à-dire de déduire toutes les relations que la structure contient implicitement. Calculer la fermeture des structures discursives peut permettre d'améliorer leur comparaison, notamment dans le cadre de l'évaluation de systèmes d'analyse automatique du discours. Nous présentons la méthodologie adoptée, que nous illustrons par l'étude d'une règle de déduction.

Abstract. We propose a methodology for the construction of discourse relations inference rules, to be integrated into an algebra of these relations. The construction of these rules has as main objective to allow for the calculation of the discourse closure of a structure, i.e. deduce all the relations implicitly contained in the structure. Calculating the closure of discourse structures improves their comparison, in particular within the evaluation of discourse parsing systems. We present the adopted methodology, which we illustrate by the study of a rule.

Mots-clés : Relation de discours, fermeture discursive, évaluation, déduction.

Keywords: Discourse relation, discourse closure, evaluation, inference.

1 Introduction

L'analyse rhétorique (ou discursive) d'un texte a pour but de représenter sa structure globale, c'est-à-dire les liens qui s'établissent entre les différentes parties du texte, permettant à son lecteur de l'interpréter comme formant un tout cohérent, et pas comme une simple succession de phrases indépendantes les unes des autres. Ces liens sont appelés *relations rhétoriques* ou relations de discours. Ils s'établissent entre des *segments de discours*, qui couvrent des propositions, des phrases et/ou de plus larges portions du texte. Différentes théories et formalismes, comme la RST (*Rhetorical Structure Theory*, Mann & Thompson, 1988), la SDRT (*Segmented Discourse Representation Theory*, Asher & Lascarides, 2003), D–LTAG (*Discourse Lexicalized Tree Adjoining Grammar*, Webber, 2004), et D–STAG (*Discourse Synchronous Tree Adjoining Grammar*, Danlos, 2009), proposent de représenter ce type de structures. Dans le travail présenté ici, le cadre théorique adopté est la SDRT.

Le traitement automatique du discours vise principalement à développer des systèmes permettant de générer des analyses de la structure discursive d'un texte. Dans cette perspective, la constitution de corpus de référence et l'évaluation des annotations produites par les systèmes d'analyse automatique sont des tâches primordiales. Les corpus de référence fournissent des données aux systèmes basés sur des méthodes d'apprentissage et permettent d'évaluer les annotations en sortie d'un analyseur. La constitution de ces corpus nécessite bien souvent la « fusion » de différentes annotations d'un même texte, donc la comparaison de structures discursives. L'évaluation des annotations générées par un système implique elle aussi la comparaison de structures discursives : les structures contenues dans les annotations du système et les structures contenues dans les annotations de référence.

Les questions qui se posent dans un objectif de construction d'une référence ou d'évaluation sont donc les suivantes : comment comparer deux annotations en discours ? quelles structures de discours sont équivalentes ou compatibles ? En effet, deux annotations discursives d'un même texte peuvent différer sans que l'une ou l'autre soit pour autant « fausse » ou « incomplète ». Considérons par exemple le discours en (1), qui contient trois segments de discours, que nous nommons (π_1), (π_2) et (π_3). On peut avoir deux annotations différentes et néanmoins équivalentes pour ce discours : une première annotation A_1, contenant les relations *Result*(π_1, π_2) et

Elaboration(π_2, π_3) ; une seconde annotation A_2 contenant ces deux mêmes relations et la relation *Result*(π_1, π_3). Les deux annotations diffèrent, mais sont équivalentes : dans l'annotation A_1, il est renseigné que (π_3) élabore (π_2), qui lui-même décrit un résultat de (π_1) ; il est donc implicitement renseigné que (π_3) décrit un résultat de (π_1), ce qui signifie que l'on peut déduire l'annotation A_2 à partir de l'annotation A_1.

1. (a) Il a beaucoup plu aujourd'hui. (π_1)
   (b) *Du coup*, Jean n'a pas pu faire ce qu'il avait prévu. (π_2)
   (c) Il n'a pas pu faire son footing, *notamment*. (π_3)

Dans le cadre de la construction d'une annotation de référence pour un texte donné, on veut pouvoir intégrer les informations présentes dans les différentes annotations de ce texte, si bien sûr elles sont compatibles. Dans le cadre de l'évaluation, on veut savoir si certaines informations absentes dans une annotation A_1 (que ce soit la référence ou non) et présentes dans une annotation A_2 sont en réalité implicitement renseignées dans l'annotation A_1. Autrement dit, on voudrait pouvoir déduire toutes les informations (les relations de discours) implicitement présentes dans les deux annotations à comparer.

Étant donnée une structure discursive associée à un texte, composée de relations de discours entre les segments qui le constituent, notre objectif est donc de pouvoir calculer, à l'aide de règles de déduction, la *fermeture discursive* de la structure, c'est-à-dire toutes les relations de discours qui peuvent être déduites à partir des relations déjà annotées. Pour calculer la fermeture discursive d'une structure, des règles de déduction de relations de discours sont nécessaires. Par exemple, pour calculer la fermeture discursive de l'annotation A_1 du discours en (1), nous avons besoin de la règle suivante : *Result*$(\pi_1, \pi_2) \wedge$ *Elaboration*$(\pi_2, \pi_3) \rightarrow$ *Result*(π_1, π_3). Or, les théories du discours ne définissent pas (ou peu) de règles de ce type. Nous proposons d'étudier et de construire des règles de déduction de relations de discours, destinées à être intégrées dans une algèbre des relations de discours, similaire à celle construite par Allen (1983) pour les relations temporelles.

L'ensemble de relations rhétoriques utilisé varie selon les théories, ainsi que la façon dont ces relations sont définies. Néanmoins, il existe un consensus sur un certain nombre de relations, comme par exemple les relations *Elaboration* et *Narration*. De plus, les relations peuvent généralement être mises en correspondance d'une théorie à une autre : par exemple, la relation *Result* de la SDRT recouvre les relations *Volitional Result* et *Non-Volitional Result* de la RST. Nous avons choisi de placer ce travail dans le cadre théorique de la SDRT (Asher & Lascarides, 2003), parce que : d'une part, cette théorie se trouve, en ce qui concerne la définition des relations de discours, à un niveau de granularité intermédiaire entre les approches multiplicatrices comme celle de la RST, qui construisent des listes étendues de relations, et les approches réductionnistes, comme celle de Grosz & Sidner (1986), qui proposent de ne distinguer que deux relations structurelles : *dominates* (dominance) et *satisfaction-precedence* (satisfaction de la précédence) ; d'autre part, elle rend explicites les contraintes sémantiques établies par une relation de discours sur ses arguments, et ces contraintes sont le point de départ de l'étude des règles de déduction. Dans la SDRT, l'établissement d'une relation donnée impose des contraintes (temporelles, causales, structurelles) sur les deux segments qu'elle relie : par exemple, la relation *Narration* implique une précédence temporelle entre les éventualités qu'elle relie, la relation *Result* implique une relation causale entre deux éventualités, etc. (voir section 3.2). La structure hiérarchique du discours est représentée à l'aide d'une distinction entre relations coordonnantes et relations subordonnantes, ce qui permet également de restreindre, dans un discours donné, l'ensemble des segments disponibles pour l'attachement de nouveaux segments dans un discours.

Dans la section 2, nous justifions l'utilisation de règles de déduction dans l'évaluation des structures discursives, en montrant pourquoi les métriques d'évaluation utilisées en analyse syntaxique ne permettent pas d'évaluer correctement des analyses discursives. Dans un second temps, nous présentons des travaux qui s'intéressent aux relations temporelles, et qui proposent d'améliorer l'évaluation de graphes temporels à l'aide de règles de déduction. Dans la section 3, nous présentons une méthodologie pour la construction des règles de déduction. Dans la section 4, nous illustrons cette méthodologie par l'étude et la construction complète d'une règle. Pour terminer, dans la section 5, nous concluons en apportant des perspectives dans la construction des règles de déduction.

2 Évaluation de graphes discursifs

Les travaux en traitement automatique du discours, et plus précisément le développement de systèmes d'analyse automatique de la structure discursive, posent, comme toutes les tâches de TAL, la question cruciale de l'évaluation : il faut évaluer les analyses produites par les systèmes, en prenant en compte les particularités des structures

discursives. L'analyse de la structure discursive d'un texte comprend deux tâches distinctes : la *segmentation* en unités minimales de discours (appelés *segments élémentaires*), et l'*annotation des relations* existant entre les différents segments (certains de ces segments recouvrent plusieurs segments élémentaires, et sont appelés *segments complexes*). Nous ne traiterons ici que de l'évaluation de la seconde étape.

Les structures discursives sont, dans la SDRT, représentées par des graphes, dont les noeuds sont des segments de discours (qui recouvrent des portions plus ou moins larges d'un texte : propositions, phrases, paragraphes, etc.), et dont les arcs sont des relations de discours (par exemple : *Narration*, *Explanation*, etc.). Dans la RST, les structures de discours sont représentées par des arbres binaires étiquetés (Marcu, 1996), dans lesquels les feuilles sont des segments élémentaires, les autres noeuds sont des relations rhétoriques, les étiquettes des arcs décrivant le type des arguments de relations (*Nucleus* ou *Satellite*).

Dans cette section, nous décrivons succinctement certaines métriques utilisées dans l'évaluation de structures proches des structures discursives : les arbres de constituants syntaxiques (proches des arbres de la RST, si l'on fait un parallèle entre les segments de discours et les mots, et un autre entre les relations rhétoriques et les constituants), et les graphes de dépendances syntaxiques (proches des graphes de la SDRT). Nous verrons pourquoi ces métriques ne permettent pas d'évaluer correctement des analyses discursives. Dans un second temps, nous présentons des travaux qui s'intéressent aux relations temporelles, et qui proposent d'améliorer l'évaluation de graphes temporels à l'aide de règles de déduction. Pour terminer, nous proposons d'utiliser des règles de déduction pour l'évaluation de graphes discursifs, et discutons de la forme de ces règles.

Les systèmes de TAL sont généralement évalués en termes de *rappel*, *précision* et *F-score* (adaptés à l'évaluation de la classification d'objets indépendants), mais diverses métriques sont utilisées, selon la tâche et le type de structure à évaluer : métrique Parseval en analyse syntaxique, métrique BLEU (qui s'applique à des séquences de mots) en traduction automatique, etc. En analyse syntaxique, les arbres de constituants sont le plus souvent évalués suivant la métrique Parseval (Black *et al.*, 1991), qui calcule, pour un arbre syntaxique, la précision et le rappel des constituants, en partant du principe suivant : un constituant dans l'arbre calculé par l'analyseur est correct s'il existe un constituant dans l'arbre correspondant du corpus de référence qui domine la même séquence de symboles terminaux et possède le même label. Elle calcule également la moyenne des constituants dans un arbre qui « croisent » des frontières de constituants dans l'autre arbre (*crossing brackets*). Dans l'évaluation de dépendances syntaxiques, les métriques utilisées (Nivre & Scholz, 2004) sont généralement : UAS (*unlabelled attachment score*), qui calcule la proportion de mots pour lesquels le gouverneur assigné est correct ; LAS (*labelled attachment score*), qui calcule la proportion de mots pour lesquels le gouverneur assigné est correct et le type de dépendance est correct ; et LabAcc (*labelled accuracy score*), qui calcule la proportion de mots pour lesquels le type de dépendance est correct.

Si l'on tente d'adapter les métriques utilisées dans l'évaluation de dépendances syntaxiques pour évaluer des graphes discursifs (en remplaçant les mots par les segments de discours, et les dépendances syntaxiques par les relations de discours), on constate qu'elles ne permettent pas toujours une évaluation satisfaisante. Considérons par exemple le discours en (2). Dans ce discours, le segment (π_1) est élaboré par les segments (π_2) et (π_3) : ces deux segments décrivent une partie du repas de Jean mentionné en (π_1). La description du repas est faite dans l'ordre chronologique, les segments (π_2) et (π_3) forment donc une narration. Si l'on adapte les métriques UAS, LAS et LabAcc pour évaluer une annotation A de ce discours contenant $\textit{Elaboration}(\pi_1, \pi_2) \wedge \textit{Narration}(\pi_2, \pi_3)$, en prenant comme référence l'annotation R de ce même discours contenant $\textit{Elaboration}(\pi_1, [\pi_2, \pi_3]) \wedge \textit{Narration}(\pi_2, \pi_3)$, on n'obtient pas un score de 1. Pourtant, étant donné la sémantique des relations en jeu, on peut déduire que les deux structures annotées sont équivalentes. En effet, l'annotation A permet de déduire l'annotation R : les informations a priori « manquantes » en A sont en réalité implicitement présentes.

2. (a) Jean a fait un excellent repas. (π_1)
   (b) Il a mangé un délicieux saumon, (π_2)
   (c) puis s'est régalé d'un copieux plateau de fromages. (π_3)

Il est donc nécessaire d'utiliser des règles de déduction pour procéder à une évaluation précise de structures discursives. Concernant le discours en (2), la SDRT possède une règle qui permet d'établir l'équivalence entre l'annotation A et la référence R : $\textit{Elaboration}(\alpha, \beta) \wedge \textit{Narration}(\beta, \gamma) \rightarrow \mathit{Elaboration}(\alpha, [\beta, \gamma])$. Cependant, la question des équivalences entre structures discursives reste très peu étudiée dans les théories du discours, et pour une majorité de couples de relations de discours (R_1, R_2), nous ne savons pas (par exemple) si la structure $R_1(\alpha, \beta) \wedge R_2(\beta, \gamma)$ contient une information implicite, et si oui, laquelle. Nous proposons donc d'étudier et de définir des règles de déduction permettant une meilleure évaluation des graphes discursifs. L'idée d'utiliser des

règles de déduction pour compléter et évaluer des annotations a déjà été exploitée en ce qui concerne les relations temporelles (Setzer *et al.*, 2003) et les relations de coréférence (Vilain *et al.*, 1995). Par exemple, si une annotation temporelle contient les informations : l'événement e_1 a lieu avant e_2 ($e_1 < e_2$) et l'événement e_2 a lieu avant e_3 ($e_2 < e_3$), alors il est implicite que l'événement e_1 a lieu avant e_3 (et l'on peut déduire $e_1 < e_3$). De la même façon, si dans une annotation en chaînes de coréférence les informations suivantes sont présentes : l'expression e_1 coréfère avec e_2 et l'expression e_2 coréfère avec e_3, alors il est implicite que l'expression e_1 coréfère avec e_3.

En ce qui concerne les relations temporelles, Allen (1983) définit une algèbre temporelle complète, avec des règles de la forme : $r_1(A, B) \wedge r_2(B, C) \rightarrow r_3(A, C)$[1]. L'algèbre utilise 13 relations (*before*, *during*, *overlaps*, etc.). Dans beaucoup de cas, il existe plus d'une relation r_3 déductible entre les intervalles A et C. Par exemple : *overlaps*$(A, B) \wedge$ *overlaps*$(B, C) \rightarrow$ *before*$(A, C) \vee$ *overlaps*$(A, C) \vee$ *meets*(A, C). Setzer *et al.* (2003) introduisent la notion de *fermeture temporelle* d'un graphe temporel, qui est la représentation complète des conséquences temporelles du graphe. Pour aider à la création de corpus de référence, ils proposent de comparer deux annotations temporelles d'un même texte en termes d'équivalence ou de recouvrement de leurs fermetures temporelles, qui sont calculées à partir de règles d'inférences similaires à celles définies par Allen (1983). Ces règles peuvent également servir d'aide à l'annotation. Tannier & Muller (2008) proposent une autre méthode de comparaison de graphes temporels. Ils distinguent deux types de relations dans un graphe temporel : les relations essentielles, et les relations qui peuvent être déduites à partir d'autres relations. Ils proposent d'effectuer la comparaison et l'évaluation d'annotations temporelles uniquement à partir des relations essentielles, ce qui nécessite également l'exploitation de règles de déduction.

Nous proposons de construire une algèbre des relations de discours, inspirée de l'algèbre des relations temporelles construite par Allen (1983), afin d'améliorer la qualité de l'évaluation de graphes discursifs : étant donnée une annotation discursive d'un texte, composée de relations de discours entre les segments qui le constituent, notre objectif est de calculer, à l'aide de règles de déduction, la *fermeture discursive* de l'annotation, c'est-à-dire toutes les relations de discours qui peuvent être déduites à partir des relations déjà annotées. Pour construire une algèbre des relations de discours, au moins deux types de règles semblent nécessaires. En effet, si l'on considère un discours quelconque à trois segments (α), (β) et (γ), il y a potentiellement deux structures (présentées à la Figure 1) pour lesquelles une relation reste indéterminée (en pointillés dans la figure) : dans la première structure, la relation entre (α) et (γ) n'est pas explicite ; dans la seconde, c'est la relation entre (β) et (γ) qui ne l'est pas[2]. Considérant ces deux structures, nous proposons d'utiliser deux types de règles de déduction[3] (représentés à la Figure 1). Notons que la déduction peut être une disjonction de relations, comme dans l'algèbre de Allen.

FIG. 1 – Schémas de déduction pour les deux types de règles : déduction de R_z ou déduction de R_y (où (α), (β) et (γ) représentent trois segments de discours successifs)

3 Méthodologie pour la construction d'une algèbre des relations de discours

Nous détaillons dans cette section la méthodologie adoptée dans l'étude et la construction des règles de déduction. Pour plus de lisibilité, nous ne présentons la méthodologie que pour le premier type de règle proposé (à gauche dans la Figure 1). La construction des règles se fait prémisse par prémisse. Dans la section 3.1, nous décrivons

[1] A, B et C représentent des intervalles temporels, et r_1, r_2 et r_3 des relations temporelles.

[2] Dans la SDRT, la seconde structure n'est valide que si R_x est une relation subordonnante : selon la *contrainte de la frontière droite*, si R_x est coordonnante, (γ) ne peut théoriquement pas être « attaché » à (α). Cependant, ces structures en théorie invalides ne sont pas nécessairement exclues des annotations en discours, nous les prenons donc en compte dans les règles de déduction à étudier.

[3] Si les relations traitées étaient des relations temporelles, le cas $R_x(\alpha, \beta) \wedge R_z(\alpha, \gamma)$ pourrait être ramené au cas : $R_x^{-1}(\beta, \alpha) \wedge R_z(\alpha, \gamma)$, où R_x^{-1} est la relation « inverse » de R_x. En effet, dans l'algèbre de Allen, il n'existe que des règles dont la prémisse est $r_1(A, B) \wedge r_2(B, C)$. En ce qui concerne les relations de discours, il nous faut distinguer les deux cas, puisqu'il n'existe pas de relation « inverse » à chaque relation de discours. De plus, l'ordre des segments dans un discours a un impact sur sa structure et son interprétation.

le travail d'extraction de règles candidates à partir du corpus ANNODIS (Péry-Woodley *et al.*, 2009)[4]. Dans la section 3.2, nous présentons les éléments qui nous permettent de mettre en évidence les différentes déductions possibles pour une prémisse de règle donnée. Dans la section 3.3, nous montrons comment les règles sont validées par l'annotation de données empiriques, extraites à partir de corpus non annotés en discours.

3.1 Extraction de déductions candidates pour la construction des règles

Pour construire une algèbre des relations de discours utilisant environ 20 relations, il y a potentiellement 8000 règles du type $R_x(\alpha,\beta) \wedge R_y(\beta,\gamma) \rightarrow R_z(\alpha,\gamma)$ à étudier (il y a 20×20 prémisses de règles possibles, et pour chaque prémisse 20 déductions possibles), plus 8000 règles du type $R_x(\alpha,\beta) \wedge R_z(\alpha,\gamma) \rightarrow R_y(\beta,\gamma)$, donc environ 16000 règles en tout. Chaque règle nécessitant une étude linguistique, il est impératif de faire une sélection dans les règles à étudier en priorité, et de dégager les règles candidates les plus pertinentes. Pour extraire des déductions candidates, nous exploitons le corpus ANNODIS, constitué de 100 textes annotés en discours, provenant notamment de Wikipédia et du corpus de l'Est Républicain. Après avoir été segmenté en unités minimales de discours, chaque texte du corpus a reçu deux annotations distinctes, avec la SDRT comme point de départ au guide d'annotation. 19 relations de discours sont utilisées dans le corpus. Nous les présentons dans le tableau 1, avec leur nombre d'occurrences dans les annotations.

Relation	Nombre d'occurrences	Relation	Nombre d'occurrences
Elaboration	1662	*Parallel*	154
Entity Elaboration	1169	*Attribution*	151
Continuation	658	*Background*	134
Narration	567	*Flashback*	106
Frame	416	*Description Continuation*	54
Contrast	334	*Conditional*	53
Result	303	*Alternation*	35
Explanation	259	*Source*	18
Goal	238	*Explanation**	12
Commentary	222	*Result**	0

TAB. 1 – Relations utilisées dans le corpus ANNODIS

Nous dégageons à partir de ce corpus des règles de déduction candidates : pour chaque prémisse de règle de la forme $R_x(\alpha,\beta) \wedge R_y(\beta,\gamma)$ nous extrayons la ou les relation(s) $R_z(\alpha,\gamma)$ les plus probables. Nous nous intéressons donc aux annotations contenant des triplets de segments de discours (α, β, γ) pour lesquels les relations sont saturées, c'est-à-dire telles que : une relation (au moins) a été annotée entre (α) et (β), de même entre (β) et (γ), ainsi qu'entre (α) et (γ). Les triplets « saturés » du corpus permettent de calculer, pour toute relation R_z, la probabilité que la relation $R_z(\alpha,\gamma)$ soit établie sachant que les relations $R_x(\alpha,\beta)$ et $R_y(\beta,\gamma)$ sont présentes :

$$P(R_z(\alpha,\gamma) \mid R_x(\alpha,\beta) \wedge R_y(\beta,\gamma)) = \frac{count(R_x(\alpha,\beta),R_y(\beta,\gamma),R_z(\alpha,\gamma))}{count(R_x(\alpha,\beta),R_y(\beta,\gamma),R(\alpha,\gamma))}$$ [5].

Ces probabilités nous donnent, pour une prémisse de règle donnée, une idée de la plausibilité des déductions : plus la probabilité $P(R_z(\alpha,\gamma) \mid R_x(\alpha,\beta) \wedge R_y(\beta,\gamma))$ est grande, plus la règle $R_x(\alpha,\beta) \wedge R_y(\beta,\gamma) \rightarrow R_z(\alpha,\gamma)$ est plausible. Dans l'exploitation du corpus, nous réécrivons les relations impliquant des segments complexes[6] avec les règles de réécriture suivantes : $R_1(\alpha,[\beta,\gamma]) \wedge R_2(\beta,\gamma) \rightarrow_{rew} R_1(\alpha,\beta) \wedge R_1(\alpha,\gamma) \wedge R_2(\beta,\gamma)$; et $R_1(\alpha,\beta) \wedge R_2([\alpha,\beta],\gamma) \rightarrow_{rew} R_1(\alpha,\beta) \wedge R_2(\alpha,\gamma) \wedge R_2(\beta,\gamma)$. La réécriture permet d'exploiter un plus grand nombre de données du corpus. Par exemple, nous obtenons 0 occurrence du triplet (*Result*(α,β), *Contrast*(β,γ), *Contrast*(α,γ)) sans la réécriture, et 127 occurrences avec la réécriture. La réécriture des segments complexes permet donc de dégager une règle candidate comme *Result*$(\alpha,\beta) \wedge$ *Contrast*$(\beta,\gamma) \rightarrow$ *Contrast*(α,β)[7] à partir

[4] Les données du corpus ANNODIS ne sont pas encore disponibles, mais nous ont été gentiment fournies par les membres du projet.

[5] $R(\alpha,\gamma)$ signifie qu'il existe une relation entre (α) et (γ).

[6] Dans un discours, certains segments élémentaires sont à la fois arguments d'une relation de discours, et à la fois partie d'un segment de discours plus large (appelé segment complexe) qui est lui-même argument d'une relation de discours. Par exemple, la structure du discours en (2) de la section 2 est *Elaboration*$(\pi_1,[\pi_2,\pi_3]) \wedge$ *Narration*(π_2,π_3), où (π_2) et (π_3) sont les deux arguments de la relation *Narration*, mais sont aussi inclus dans un segment *complexe* plus large ($[\pi_2,\pi_3]$) qui élabore (π_1).

[7] Voir la section 4 pour une étude détaillée de la prémisse *Result*$(\alpha,\beta) \wedge$ *Contrast*(β,γ).

des structures annotées *Result*$(\alpha, \beta) \wedge$ *Contrast*$([\alpha, \beta], \gamma)$ dans le corpus. Notons que la sémantique de $R_1(\alpha, \beta)$ devient alors : « la relation R_1 est établie entre (α) (ou un segment contenant (α)) et (β) (ou un segment contenant (β)) ».

3.2 Mise en évidence des déductions possibles

Pour construire les règles de déduction, nous effectuons une première étude de la prémisse de règle considérée, qui nous permet de dégager la ou les déduction(s) possible(s), et dans certains cas de dégager des paramètres influant sur la déduction. Cette étude s'appuie sur différents éléments : les déductions candidates extraites à partir du corpus ANNODIS, la sémantique des relations présentes dans la prémisse et dans les déductions candidates, et l'analyse d'exemples construits et attestés[8].

Les probabilités calculées sur le corpus ANNODIS nous permettent de dégager la/les déduction(s) plausible(s) : plus la probabilité $P(R_z(\alpha, \gamma) \mid R_x(\alpha, \beta) \wedge R_y(\beta, \gamma))$ est grande, plus la règle $R_x(\alpha, \beta) \wedge R_y(\beta, \gamma) \rightarrow R_z(\alpha, \gamma)$ est plausible. Cependant, nous ne pouvons pas nous baser sur ces seules informations pour dégager des règles de déduction, en partie parce que ces calculs exploitent des discours où toute l'information est explicitée par l'annotateur, et que nous voulons mettre au jour l'information *implicite* établie par la présence des relations des prémisses de règles. Pour cela, nous devons examiner, pour la prémisse de règle considérée, les conséquences de l'établissement des relations $R_x(\alpha, \beta)$ et $R_y(\beta, \gamma)$. En effet, lorsqu'une relation de discours est établie, elle impose certaines contraintes sur ses arguments et sur les liens qui existent entre eux. La nature de ces contraintes varie selon les relations de discours. Certaines relations, comme *Narration*, établissent des contraintes temporelles sur les éventualités qu'elles relient : si l'on a *Narration*(α, β), alors l'éventualité e_α décrite dans le premier segment (α) a lieu avant l'éventualité e_β décrite dans le second segment. D'autres relations, comme *Explanation*, établissent des relations causales : si l'on a *Explanation*(α, β), alors l'éventualité e_β décrite dans le second segment est la cause de l'éventualité e_α décrite dans le premier segment. Par exemple, dans le discours : *Jean est tombé. Max l'a poussé.*, l'éventualité décrite dans la seconde phrase est la cause de l'éventualité dans la seconde phrase.

Observer les contraintes établies par les relations R_x et R_y permet de mieux caractériser le lien existant entre les segments (α) et (γ), en exploitant la définition théorique des relations de discours en jeu. Les contraintes établies par les relations contenues dans la prémisse de la règle nous permettent ainsi dans certains cas de prédire la déduction (voir section 4.2.3), d'un point de vue théorique au moins. De plus, les contraintes établies par les relations contenues dans la prémisse nous permettent d'exclure certaines relations des déductions possibles, car les conséquences de $R_x(\alpha, \beta)$ et $R_y(\beta, \gamma)$ sont incompatibles avec l'établissement de $R_z(\alpha, \gamma)$. Par exemple, lorsque les relations *Narration*(α, β) et *Narration*(β, γ) sont établies, la relation *Flashback* ne peut lier (α) et (β). Cette incompatibilité est illustrée dans le discours en (3), où l'on a *Narration*(π_1, π_2) et *Narration*(π_2, π_3), avec pour conséquences temporelles : l'événement décrit en (π_1) a lieu avant l'événement décrit en (π_2) $(e_{\pi_1} < e_{\pi_2})$; l'événement décrit en (π_2) a lieu avant l'événement décrit en (π_3) $(e_{\pi_2} < e_{\pi_3})$; donc l'événement décrit en (π_1) a lieu avant l'événement décrit en (π_3) $(e_{\pi_1} < e_{\pi_3})$, ce qui est incompatible avec l'établissement de la relation *Flashback*(π_1, π_3), car sa conséquence temporelle est : $e_{\pi_3} < e_{\pi_1}$.

3. (a) Aujourd'hui, Julie est allée voir une expo. (π_1)
   (b) *Ensuite*, elle a déjeuné avec des amis. (π_2)
   (c) *Puis* elle a fait des courses au marché. (π_3)

L'étude des règles de déduction nécessite également un travail d'introspection, consistant à construire des discours comportant la prémisse de règle étudiée. Ce travail permet de vérifier les hypothèses formulées à partir des calculs effectués sur le corpus et des contraintes établies par les relations de discours. Dans la construction des discours, on essaie de couvrir le plus grand nombre de cas de figure. Pour cela, selon les relations présentes dans la prémisse, nous faisons varier certains paramètres. Par exemple, lors de la construction de discours impliquant la relation *Contrast*, on peut établir des contrastes où : une même entité présente deux propriétés distinctes, deux entités distinctes présentent chacune une propriété, une négation est présente dans une des propositions, etc. ; pour les discours contenant la relation *Explanation*(α, β), l'éventualité décrite dans le segment (α) peut être un événement, un état, une cause future, etc. Pour analyser les données construites ou attestées, et vérifier la déduction d'une relation entre (α) et (γ), on utilise les tests d'insertion d'un connecteur et de réorganisation du discours, décrits dans les deux paragraphes suivants.

[8] Ces exemples attestés proviennent du corpus ANNODIS et des corpus de discours extraits grâce à l'outil présenté à la section 3.3.

Insertion d'un connecteur L'insertion d'un connecteur permet de tester la présence d'une relation de discours entre (α) et (γ) dans l'étude de règles de la forme : $R_x(\alpha,\beta) \wedge R_y(\beta,\gamma) \rightarrow R_z(\alpha,\gamma)$. Si l'on veut vérifier la présence d'une relation R_z donnée entre les segments (α) et (γ) d'un discours, on peut utiliser le test suivant : si, après avoir inséré dans le segment (γ) un connecteur adverbial lexicalisant R_z, le discours reste cohérent et que son interprétation est inchangée, alors la présence de la relation $R_z(\alpha,\gamma)$ est vérifiée. Observons les discours en (4), dans lesquels les relations *Result*(π_1,π_2) et *Explanation*(π_2,π_3) sont établies. On constate à l'aide du test d'insertion que : pour (4c-i), le connecteur de résultat *du coup* peut être inséré sans rendre le discours incohérent, et sans en modifier l'interprétation, donc la relation *Result*(π_1,π_3) peut être inférée ; pour le discours en (4c-ii), en revanche, l'insertion d'un connecteur de résultat rend le discours incohérent, donc la relation *Result*(π_1,π_3) ne peut pas être établie.

4. (a) L'électricité est revenue ce matin. (π_1)
   (b) Les habitants sont très contents, (π_2)
   (c) i. *car* ***du coup*** ils ont pu regagner leurs appartements. (π_3)
   ii. *car* (**# *ainsi* / # *du coup***) ils ont besoin de chauffage. (π_3)

Réorganisation du discours Une autre méthode pour mettre au jour la relation établie entre les segments (α) et (γ) est de réorganiser le discours en échangeant la position des segments (β) et (γ). Le connecteur qui lexicalise la relation $R_y(\beta,\gamma)$ est remplacé par un connecteur lexicalisant la relation « inverse » ${R_y}^{-1}(\gamma,\beta)$ si elle existe : par exemple, si l'on a la relation *Result*(β,γ), on utilise un marqueur de la relation *Explanation*(γ,β). Dans le discours ainsi formé, si un connecteur lexicalisant R_z peut être inséré entre les segments (α) et (γ) sans rendre le discours incohérent et en conservant l'interprétation d'origine, alors on peut déduire que la relation $R_z(\alpha,\gamma)$ s'établit dans le discours d'origine. Par exemple, en (5b-i), la réorganisation du discours en (4c-i) nous permet de mettre en évidence la présence de la relation *Result*(π_1,π_3), car le connecteur de résultat *du coup* peut être inséré. En revanche, la réorganisation du discours (4c-ii) nous permet d'exclure la présence de la relation *Result* entre les segments (π_1) et (π_3), car le discours en (5b-ii) est incohérent.

5. (a) L'électricité est revenue ce matin. (π_1)
   (b) i. *Du coup*, les habitants ont pu regagner leurs appartements. (π_3)
   ii. # *Du coup*, ils ont besoin de chauffage. (π_3)
   (c) *Donc* ils sont très contents. (π_2)

3.3 Annotation pour la validation des règles construites

Après avoir mis en évidence les déductions possibles pour une prémisse donnée, on réalise une étude systématique sur un corpus de plusieurs centaines de discours contenant la prémisse étudiée : pour chaque discours du corpus collecté, on annote la relation déduite[9]. Cette annotation permet d'une part de vérifier la validité des hypothèses formulées quant aux déductions possibles, et aux paramètres ayant un impact sur la déduction ; d'autre part, elle permet, dans le cas où plusieurs déductions sont possibles pour une prémisse donnée, de connaître la fréquence de chacune des déductions.

Pour mener une telle étude, le corpus ANNODIS ne contient généralement pas suffisamment de triplets de segments (α,β,γ) au sein desquels les prémisses étudiées (et seulement les prémisses) sont établies. Pour constituer un corpus d'exemples suffisamment grand, nous avons donc développé un outil permettant d'extraire des discours contenant les prémisses de règles à partir de données non-annotées en discours. L'identification de la présence des relations de discours en jeu dans la prémisse de règle considérée est effectuée grâce à la présence de marques de surface : les connecteurs de discours. L'extraction est effectuée sur le corpus de l'Est Républicain, annoté en dépendances syntaxiques par l'analyseur BONSAI (Candito *et al.*, 2009), et exploite un lexique des connecteurs discursifs du français, LEXCONN (Roze *et al.*, 2010), qui contient 330 connecteurs auxquels sont associées une catégorie syntaxique et la relation de discours qu'il établissent. Afin de mieux exploiter le lexique, nous avons complété dans celui-ci un certain nombre de contraintes concernant les positions pouvant être occupées par les différents connecteurs lorsqu'ils lexicalisent une relation donnée. En effet, certains connecteurs n'établissent une relation de discours que lorsqu'ils occupent certaines positions, ou bien peuvent établir des relations différentes selon la position occupée. Pour les adverbes, c'est la position dans la proposition qui peut avoir une importance ; pour les conjonctions de subordination, c'est la position de la proposition subordonnée par rapport à la principale.

[9] Dans la phase d'annotation, nous utilisons les tests d'insertion d'un connecteur et de réorganisation du discours.

Pour extraire des occurrences d'une prémisse de règle $R_x(\alpha, \beta) \wedge R_y(\beta, \gamma)$, on recherche dans le corpus des contextes dans lesquels on a : p_1 *[conn$_x$]*p_2 *[conn$_y$]*p_3, où *conn$_x$* est un connecteur accueilli par la proposition p_2 et lexicalise la relation R_x, et *conn$_y$* est un connecteur accueilli par la proposition p_3 et lexicalise la relation R_y. Soit $conn_y$ se trouve dans la même phrase que $conn_x$, soit dans la phrase suivante. Par exemple, pour la prémisse de règle *Explanation*$(\alpha, \beta) \wedge$ *Result*(β, γ), on extrait des discours comme en (6), où la conjonction *car* marque la présence de la relation *Explanation*, et l'adverbe *alors* marque la présence de la relation *Result*. De la même façon, pour la prémisse de règle *Result*$(\alpha, \beta) \wedge$ *Contrast*(β, γ), on extrait des discours comme en (7).

6. Malgré l'annonce de la fin possible des combats, ils n'ont plus du tout confiance *car*, lors des années passées, ils ont vu la guerre et la paix se succéder. *Alors*, ils se disent que, cette fois encore, la guerre pourrait revenir...
7. Mme Mulot, assistante sociale DVIS, est en absence de longue durée. Ses permanences sont *donc* annulées. La prise en charge des urgences reste *néanmoins* assurée : joindre la circonscription DVIS centre-Vosges...

4 Construction d'une règle de déduction

Dans cette section, nous déroulons pour la prémisse de règle *Result*$(\alpha, \beta) \wedge$ *Contrast*(β, γ) la méthodologie présentée à la section 3. Il nous a paru intéressant d'étudier cette prémisse de règle, et d'observer les conséquences de l'établissement des deux relations relations en jeu, car elles appartiennent à deux groupes de relations distincts : les relations causales (pour *Result*) et les relations adversatives (pour *Contrast*), selon la classification de (Halliday & Hasan, 1976). La première relie deux éventualités dont l'une est la cause de l'autre, et c'est une relation de cohérence, c'est-à-dire une relation liée au contenu sémantique des propositions reliées. La seconde, en revanche, relie des éventualités présentées par le locuteur comme étant en opposition ou en contradiction. Elle est généralement considérée comme une relation intentionnelle, liée aux buts communicatifs.

4.1 Extraction de déductions candidates

L'exploitation du corpus ANNODIS permet d'obtenir les probabilités présentées dans le tableau 2. Les calculs sont effectués à partir des 150 contextes du corpus où l'on a : *Result*(α, β), *Contrast*(β, γ) et une relation $R_z(\alpha, \gamma)$, après réécriture des relations entre segments complexes. Le calcul des probabilités nous permet de dégager la règle candidate : *Result*$(\alpha, \beta) \wedge$ *Contrast*$(\beta, \gamma) \rightarrow$ *Contrast*$(\alpha, \gamma) \vee$ *Result*(α, γ). En effet, on ne retient pas la déduction de *Goal* dans la règle candidate, car la probabilité observée pour cette règle provient d'une seule annotation du corpus, au sein de laquelle les relations annotées ne sont pas cohérentes. On constate que la déduction candidate la plus probable est *Contrast*.

Relation $R_z(\alpha, \gamma)$	$p(R_z(\alpha, \gamma) \mid$ *Result*(α, β), *Contrast*$(\beta, \gamma))$
$R_z =$ *Contrast*	0.847
$R_z =$ *Result*	0.127
$R_z =$ *Goal*	0.027
$R_z \notin \{$*Result*, *Contrast*, *Goal*$\}$	0.0

TAB. 2 – Probabilités $p(R_z(\alpha, \gamma) \mid$ *Result*(α, β), *Contrast*$(\beta, \gamma))$ pour toute relation R_z

4.2 Mise en évidence des déductions possibles

Nous cherchons ici à mettre en évidence les déductions possibles pour la prémisse de règle *Result*$(\alpha, \beta) \wedge$ *Contrast*(β, γ). Dans la section 4.2.1, nous fournissons une définition de ces deux relations, et distinguons deux sous-cas pour la relation *Contrast* : un premier où elle est de type *opposition sémantique*, et un second où elle de type *concession* ou *violation d'attente*. Cette distinction nous amène à présenter l'étude des déductions possibles dans deux sections séparées : 4.2.2 pour le premier sous-cas, et 4.2.3 pour le second.

4.2.1 Définition des relations *Result* et *Contrast*

Au sein de la SDRT, la relation *Result* peut être établie par : des marques linguistiques, comme certains connecteurs de discours (*donc*, *du coup*, *alors*, etc.) et des verbes causatifs (*provoquer*, *entraîner*, etc.) ; des connaissances (extra-)linguistiques, comme $pousser(x, y) > tomber(y)$[10] (qui permet par exemple l'interprétation de : *Léa a poussé Max. Il est tombé.*). Elle a pour conséquence sémantique l'établissement d'une relation causale entre les deux éventualités reliées, à savoir que l'éventualité décrite dans le premier argument de la relation est la cause de l'éventualité décrite dans le second argument : $cause(e_\alpha, e_\beta)$ (pour $Result(\alpha, \beta)$). Cette relation causale recouvre l'implication non monotone $K_\alpha > K_\beta$ (K_α désigne le contenu sémantique du segment α). *Result* possède également des effets temporels : l'éventualité du premier argument a lieu avant l'éventualité du second argument.

La relation *Contrast* relie des segments qui présentent une dissimilarité sémantique. Cette relation recouvre dans la SDRT trois relations qui sont parfois distinguées dans la littérature (Busquets, 2007) : *opposition sémantique* ou *contraste formel*, *concession*, et *violation d'attente*. L'*opposition sémantique* (OS) est définie par Spooren (1989) comme une relation entre deux propositions qui ont deux sujets distincts, auxquels sont attribués des propriétés qui s'excluent mutuellement dans le contexte. Selon Oversteegen (1997), l'opposition sémantique ne nécessite pas la présence de deux entités distinctes : il peut aussi n'y avoir qu'une seule entité, à laquelle différentes propriétés sont assignées, à des localisations temporelles ou spatiales distinctes, où dans différents mondes possibles. Dans la SDRT, cette relation implique une similarité structurelle entre les segments reliés. Busquets (2007) résume la relation d'opposition sémantique comme un contraste ou une dissimilarité entre les éléments comparés (des états, des événements ou des individus) sans contradiction entre eux. En revanche, les éléments impliqués dans une relation de *concession* ou de *violation d'attente* sont en contradiction. Ils s'inscrivent dans le schéma suivant : $(A > C) \wedge (B > \neg C)$. En ce qui concerne la *concession* (CS), les propositions A et B sont explicites (Gröte *et al.*, 1995). Par exemple, dans le discours en (8), on infère de (8a) la proposition *nous avons mangé* $(= C)$, et de (8b) la proposition *nous n'avons pas mangé* $(= \neg C)$. En ce qui concerne la *violation d'attente* (VA), la proposition $\neg C$ est présente. Par exemple, en (9) on infère *Pierre n'aime pas le foot* $(= C)$ de la proposition en (9a) et l'on a $\neg C$ en (9b). On note que les deux éléments en relation de VA peuvent apparaître dans l'ordre inverse.

8. (a) Nous avions faim, $>$ *Nous avons mangé.*
   (b) *mais* les restaurants étaient fermés. $>$ *Nous n'avons pas mangé.*
9. (a) Pierre n'aime pas le sport, $>$ *Pierre n'aime pas le foot.*
   (b) *mais* il aime le foot.

Il nous faut noter que deux segments de discours peuvent à la fois entretenir une relation de *Result* et une relation de *Contrast*. Par exemple, dans le discours en (10), on infère que l'accident de Marie est la cause de ses fractures, ce qui correspond à l'établissement de la relation *Result*. On infère également que l'accident aurait normalement dû causer des blessures plus graves que des fractures, mais qu'il n'en a pas causé. Dans ce discours, l'établissement de la relation *Result* a pour conséquence sémantique $K_{\pi_1} > K_{\pi_2}$, et l'établissement de la relation *Contrast* (de type CS) a pour conséquence sémantique $(K_{\pi_1} > P) \wedge (K_{\pi_2} > \neg P)$ (où $P =$ *Marie a des blessures graves*). Cet exemple nous montre qu'il peut exister des cas dans lesquels on déduira à la fois la relation *Result* et la relation *Contrast*.

10. (a) Marie a eu un accident de voiture, (π_1)
    (b) *mais* elle n'a que quelques fractures. (π_2)

4.2.2 Déduction dans le cas où *Contrast* est de type *opposition sémantique*

Dans le cas où la relation $Contrast(\beta, \gamma)$ est de type OS il semble que trois déductions soient possibles : soit on déduit $Result(\alpha, \gamma)$ comme dans le discours en (11), où l'on infère que Julie est aux anges *parce que* la France a perdu le match, et où l'éventualité en (α) a des effets opposés sur deux entités distinctes (*Marie* et *Julie*) ; soit on déduit $Contrast(\alpha, \gamma)$ comme dans le discours en (12a), où l'on infère que *malgré* l'accident, Julie n'est pas blessée ; soit on déduit $Result(\alpha, \gamma) \wedge Contrast(\alpha, \gamma)$ comme dans le discours en (12b), où, par le contenu de (α) et le contenu de (β), on infère que Julie aurait *normalement* dû être plus gravement blessée, ce qui est contredit par le segment (γ), qui décrit parallèlement un état causé par l'accident décrit en (α).

[10] L'opérateur conditionnel non monotone permet d'exprimer des règles défaisables (ou révisables) (Asher & Lascarides, 2003). Par exemple, $A > B$ signifie : « si A est vrai, alors normalement, B est vrai ».

11. (α) La France a perdu le match.
    (β) *Du coup*, Marie est très déçue.
    (γ) *Par contre*, Julie est aux anges.
12. (α) Marie et Julie ont eu un accident de voiture.
    (β) Marie a une jambe cassée.
    (γ) a. *En revanche*, Julie est indemne.
    b. *En revanche*, Julie n'a que quelques égratignures.

Notons que lorsque la déduction de *Contrast* a lieu, le type de contraste établi entre (α) et (γ) n'est pas le même que celui établi entre (β) et (γ). Par exemple, dans le discours en (13), le contraste entre (β) et (γ) est de type OS car deux entités distinctes (*Julie* et *Léa*) présentent des propriétés opposées. En revanche, entre les segments (α) et (γ), c'est un contraste de type VA qui s'établit. En effet, on a : $K_\alpha > \neg K_\gamma$ (si Marie se fait agresser, alors normalement Léa devrait intervenir). D'ailleurs, si l'on réorganise le discours en (13) en intervertissant les positions de (β) et (γ), on lexicalise la relation entre (α) et (γ) par le connecteur *mais*, qui établit la relation *Contrast* (*Marie s'est fait agresser.* ***Mais*** *Léa n'a pas bougé.* ***Par contre****, Julie a tenté de la défendre.*).

13. (α) Marie s'est fait agresser.
    (β) *Du coup*, Julie a tenté de la défendre.
    (γ) *Par contre*, Léa n'a pas bougé.

4.2.3 Déduction dans le cas où *Contrast* est de type *concession* ou *violation d'attente*

Dans le cas où la relation *Contrast*(β, γ) est de type CS ou VA, il semble que l'on puisse déduire la relation *Contrast*(α, γ) directement à partir des conséquences sémantiques de l'établissement des relations de la prémisse. Notons que la présence de la relation *Result*(α, β) a toujours pour conséquence $K_\alpha > K_\beta$. Si le contraste est de type CS, comme dans le discours en (14), alors on a de plus : $(K_\beta > P)$ et $(K_\gamma > \neg P)$. On a donc $K_\alpha > K_\beta > P$, et l'on peut déduire : $K_\alpha > P$. Comme $K_\gamma > \neg P$, on retrouve les conséquences sémantiques de l'établissement d'une relation de CS entre (α) et (γ).

14. (α) Nous n'avions pas mangé de la journée.
    (β) *Donc* nous avions très faim. $> P$ (*Nous avons mangé.*)
    (γ) *Mais* les restaurants étaient fermés. $> \neg P$ (*Nous n'avons pas mangé.*)

Si le contraste est de type VA, on a deux cas possibles, représentés par les discours en (15) et (16). Dans le premier cas, on a $(K_\beta > P)$ et $(K_\gamma = \neg P)$. On a donc $K_\alpha > K_\beta > P$, et l'on peut déduire : $K_\alpha > P$. Comme $K_\gamma = \neg P$, on retrouve les conséquences de l'établissement d'une relation de VA entre (α) et (γ). Dans le second cas, illustré en (16), on a $(K_\beta = P)$ et $(K_\gamma > \neg P)$. On a $K_\alpha > K_\beta = P$, et l'on peut déduire $K_\alpha > P$. On a donc : $(K_\alpha > P)$ et $(K_\gamma > \neg P)$, ce qui correspond à l'établissement d'une relation de CS entre (α) et (γ).

15. (α) Pierre n'aime pas courir.
    (β) *Du coup* il n'aime pas le sport. $> P$ (*Pierre n'aime pas le foot.*)
    (γ) *Mais* il aime le foot. $(\neg P)$
16. (α) Tous les copains de Pierre jouent au foot.
    (β) *Du coup*, il s'est inscrit avec eux. (P)
    (γ) *Pourtant* il n'aime pas le sport. $> \neg P$ (*Pierre ne s'est pas inscrit au foot.*)

4.3 Résultats de l'annotation

Nous présentons dans cette section les résultats de l'annotation effectuée sur les discours extraits automatiquement en utilisant les connecteurs de discours comme marques des relations recherchées. Parmi les 360 discours analysés, 189 (soit 52.5%) ne contiennent pas la prémisse[11] ou sont difficilement analysables, par manque de contexte (extra-)linguistique. En revanche, 171 discours contiennent bien la prémisse à étudier (soit 47.5%). Nous présentons dans

Relation(s) déduite(s)	Pourcentage	Nombre
Contrast	73.7	126
Result	12.3	21
Unknown	5.8	10
None	5.3	9
Contrast et *Result*	2.9	5
Total	100	171

TAB. 3 – Pourcentages des relations déduites entre (α) et (γ) pour les discours extraits contenant la prémisse *Result*$(\alpha, \beta) \wedge$ *Contrast*(β, γ)

le tableau 3 les résultats obtenus pour ces discours. On regroupe les cas où la relation déduite est ambiguë sous *Unknown*, et les cas où aucune relation n'est déduite à partir de la prémisse sous *None*.

Ces résultats nous montrent que dans une majorité de cas, la relation *Contrast* est déduite (76.6%). Nous proposons donc d'établir une règle défaisable : *Result*$(\alpha, \beta) \wedge$ *Contrast*$(\beta, \gamma) >$ *Contrast*(α, γ). De façon plus générale, l'étude de la prémisse permet de formuler une règle dure : *Result*$(\alpha, \beta) \wedge$ *Contrast*$(\beta, \gamma) \rightarrow$ *Contrast*$(\alpha, \gamma) \vee$ *Result*$(\alpha, \gamma) \vee$ *None*(α, γ). Pour que les règles puissent s'inscrire dans une algèbre des relations de discours, il nous faut définir une relation artificielle *None*, exprimant le fait qu'il n'existe aucune relation de discours entre deux segments, qui est exclusive de toutes les autres relations de la déduction. Concernant les cas où l'on déduit *None*(α, γ) pour la prémisse étudiée, nous avons observé que d'autres relations que celles contenues dans la prémisse sont généralement présentes. Par exemple, pour le discours en (17), deux relations temporelles (de recouvrement temporel plus précisément) viennent s'ajouter aux relations de la prémisse étudiée : on a $Background_{forward}(\pi_1, \pi_2)$ et $Background_{forward}(\pi_2, \pi_3)$. Nous formulons l'hypothèse que la présence d'autres relations que celles contenues dans la prémisse peuvent « bloquer » la déduction. Les interactions entre les différentes règles de déduction restent donc à étudier, ainsi que les relations incompatibles.

17. (a) Le généraliste était bien connu dans sa petite localité pour ses problèmes d'alcool. (π_1)
    (b) Entre avril et août 2001 , on lui interdisait *donc* de continuer à exercer. (π_2)
    (c) *Or* la CPAM a continué à recevoir des feuilles de remboursement de patients. (π_3)

Pour raffiner la règle générale formulée, on peut distinguer des sous-cas nous permettant de mieux prédire la déduction, en nous basant sur le type de contraste établi, comme nous l'avons montré dans la section 4.2 : *Result*$(\alpha, \beta) \wedge$ (*Contrast*$_{\text{CS}}(\beta, \gamma) \vee$ *Contrast*$_{\text{VA}}(\beta, \gamma)) \rightarrow$ *Contrast*(α, γ).

5 Conclusion et perspectives

Nous avons proposé une méthodologie pour la construction de règles de déduction de relations de discours, destinées à être intégrées dans une algèbre (complète) de ces relations. La construction d'une telle algèbre a comme principal objectif de permettre une meilleure comparaison des structures dans le cadre de l'évaluation de systèmes d'analyse automatique du discours et dans le cadre de la construction de corpus de référence. Elle peut également aider à la détection d'incohérences dans des structures discursives, ce qui peut servir à améliorer l'annotation discursive manuelle ou automatique. Nous avons présenté l'étude complète d'une prémisse de règle, qui a servi à l'élaboration de la méthodologie proposée. Cette étude nous amène à formuler : des règles dures, établies grâce à des éléments théoriques sur les relations de discours, des données construites et des données attestées ; des règles molles, établies grâce à des probabilités de déduction calculées à partir de l'annotation manuelle de données extraites automatiquement. Pour extraire ces données, nous avons développé un outil qui exploite les connecteurs de discours pour détecter la présence des relations de discours recherchées.

La construction des règles de déduction est en cours, et nous avons pour objectif de dégager des traits linguistiques permettant de prédire la relation déduite lorsqu'une règle donne lieu à la déduction d'une disjonction de relations, en nous basant sur l'étude linguistique des règles et sur l'exploitation des données annotées par des méthodes statistiques. Au fur et à mesure de l'étude des règles, nous allons chercher à établir des généralisations sur les règles,

[11] Certains de ces exemples impliquent les relations *Result* et *Contrast* sans contenir précisément la séquence recherchée : les segments (α), (β) et (γ) ne sont pas consécutifs, et l'on a par exemple la structure *Result*$(\alpha, \beta_1) \wedge R(\beta_1, \beta_2) \wedge$ *Contrast*(β_2, γ).

et tenter de déterminer si le type de relation (coordonnante ou subordonnante) a un impact sur la déduction, et si des relations partageant certaines conséquences sémantiques ont un comportement similaire dans la déduction.

Références

ALLEN J. (1983). Maintaining knowledge about temporal intervals. In *Communications of the* ACM : ACM Press.

ASHER N. & LASCARIDES A. (2003). *Logics of Conversation*. Cambridge University Press.

BLACK E., ABNEY S., FLICKENGER D., GDANIEC C., GRISHMAN R., HARRISON P., HINDLE D., INGRIA R., JELINEK F., KLAVANS J., LIBERMAN M., MARCUS M., ROUKOS S., SANTORINI B. & STRZALKOWSKI T. (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Speech and Natural Language : Proceedings of a Workshop Held at Pacific Grove, California*.

BUSQUETS J. (2007). Discourse contrast : Types an tokens. *Language, Representation and Reasoning. Memorial Volume to Isabel Gómez Txurruka*, p. 103–123.

CANDITO M., CRABBÉ B., DENIS P. & GUÉRIN F. (2009). Analyse syntaxique du français : des constituants aux dépendances. In *Proceedings of TALN'09*, Senlis, France.

DANLOS L. (2009). D-STAG : un formalisme d'analyse automatique de discours basé sur les TAG synchrones. *Revue TAL*, **50**, 1–30.

GROSZ B. & SIDNER C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, **12**, 175–204.

GRÖTE B., LENKE N. & STEDE M. (1995). Ma(r)king concessions in English and German. In LEIDEN, Ed., *Proceedings of the Fifth European Workshop on Natural Language Generation*, Netherlands.

HALLIDAY M. A. K. & HASAN R. (1976). *Cohesion in English*. London : Longman.

MANN W. & THOMPSON S. (1988). Rhetorical structure theory : Towards a functional theory of text organization. *Text*, **8**, 243–281.

MARCU D. (1996). Building up rhetorical structure trees. In *Proceedings of 13th National Conference on Artificial Intelligence*, volume 2, p. 1069–1074, Portland, Oregon.

NIVRE J. & SCHOLZ M. (2004). Deterministic dependency parsing of english text. In *COLING 2004*, p. 64–70, Geneva, Switzerland.

OVERSTEEGEN L. E. (1997). On the pragmatic nature of causal and contrastive connectives. *Discourse Processes*, **24**, 51–85.

PÉRY-WOODLEY M.-P., ASHER N., ENJALBERT P., BENAMARA F., BRAS M., FABRE C., FERRARI S., HO-DAC L.-M., DRAOULEC A. L., MATHET Y., MULLER P., PRÉVOT L., REBEYROLLE J., TANGUY L., COURET M. V., VIEU L. & WIDLÖCHER A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives (poster). In *Traitement Automatique des Langues Naturelles (TALN 2009)*, Senlis, France.

ROZE C., DANLOS L. & MULLER P. (2010). LEXCONN : a French Lexicon of Discourse Connectives. In *Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010)*, Moissac, France.

SETZER A., GAIZAUSKAS R. & HEPPLE M. (2003). Using semantic inferences for temporal annotation comparison. In *Proceedings of the Fourth International Workshop on Inference in Computational Semantics (ICoS-4)*.

SPOOREN W. (1989). *Some aspects of the form and interpretation of global contrastive coherence relations*. PhD thesis, K.U. Nijmegen.

TANNIER X. & MULLER P. (2008). Evaluation metrics for automatic temporal annotation of texts. In *Language Resources and Evaluation Conference* (LREC 2008), Marrakech.

VILAIN M., BURGER J., ABERDEEN J., CONNOLLY D. & HIRSCHMAN L. (1995). A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference* (MUC–6), Columbia, Maryland.

WEBBER B. (2004). D–LTAG : Extending lexicalized TAG to discourse. *Cognitive Science*, **28**, 751–779.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Integration of Speech and Deictic Gesture in a Multimodal Grammar

Katya Alahverdzhieva & Alex Lascarides
School of Informatics, University of Edinburgh
K.Alahverdzhieva@sms.ed.ac.uk, alex@inf.ed.ac.uk

Résumé. Dans cet article, nous présentons une analyse à base de contraintes de la relation forme-sens des gestes déictiques et de leur signal de parole synchrone. En nous basant sur une étude empirique de corpus multimodaux, nous définissons quels énoncés multimodaux sont bien formés, et lesquels ne pourraient jamais produire le sens voulu dans la situation communicative. Plus précisément, nous formulons une grammaire multimodale dont les règles de construction utilisent la prosodie, la syntaxe et la sémantique de la parole, la forme et le sens du signal déictique, ainsi que la performance temporelle de la parole et la deixis afin de contraindre la production d'un arbre de syntaxe combinant parole et gesture déictique ainsi que la représentation unifiée du sens pour l'action multimodale correspondant à cet arbre. La contribution de notre projet est double : nous ajoutons aux ressources existantes pour le TAL un corpus annoté de parole et de gestes, et nous créons un cadre théorique pour la grammaire au sein duquel la composition sémantique d'un énoncé découle de la synchronie entre geste et parole.

Abstract. In this paper we present a constraint-based analysis of the form-meaning relation of deictic gesture and its synchronous speech signal. Based on an empirical study of multimodal corpora, we capture generalisations about which multimodal utterances are well-formed, and which would never produce the intended meaning in the communicative situation. More precisely, we articulate a multimodal grammar whose construction rules use the prosody, syntax and semantics of speech, the form and meaning of the deictic signal, as well as the relative temporal performance of the speech and deixis to constrain the production of a single syntactic tree of speech and deictic gesture and its corresponding meaning representation for the multimodal action. In so doing, the contribution of our project is two-fold: it augments the existing NLP resources with annotated speech and gesture corpora, and it also provides the theoretical grammar framework where the semantic composition of an utterance results from its gestural and speech synchrony.

Mots-clés : Deixis, parole et geste, grammaires multimodales
.

Keywords: Deixis, speech and gesture, multimodal grammars.

1 Introduction

Through the physical co-location of people known as *co-presence* (Goffman, 1963), individuals convey information to each other using various meaningful and visibly accessible channels such as the arrangements of the bodies in the shared space, the bodily orientations, the pointing signals of their hands, etc. In recent years, it has become commonplace to integrate input from different modalities of interaction, such as natural language and deictic gesture, in multimodal systems for the purposes of human-robot interaction (Giuliani & Knoll, 2007), or pen-based applications (Oviatt *et al.*, 1997), (Johnston, 1998).

In this paper, we show that co-speech deictic gesture can be integrated into a constraint-based grammar using purely linguistic information such as the prosody, syntax, semantics of speech, the form and meaning of the deictic signal, and their relative temporal performance. Our overall aim is to articulate the mapping from the form of multimodal signals to their (underspecified) meaning, using established methods from linguistics such as constraint-based syntactic derivation and semantic composition. To specify this mapping, we develop a grammar for speech and co-speech deictic gesture (referred to as *deixis*) which captures generalisations about well-formed multimodal actions and about multimodal actions that cannot convey the intended meaning in the specific context. We have already captured constraints on depicting dimensions via a constraint-based grammar (Alahverdzhieva & Lascarides, 2010). Here were are going to demonstrate that constraint-based grammars are expressive enough to represent the form-meaning mapping for deictic dimensions too.

2 Data

We start with an overview of deictic gesture and its relation to other co-speech gestures, and we then present the major challenges arising from the range of ambiguities and distinct performances of the pointing hand.

2.1 Deixis Background

Our focus of study are spontaneously performed co-speech deictic gestures. Compared to, say, depicting gestures where the hand literally or metaphorically *depicts* its denotation, deictic gestures designate spatial reference in Euclidean space marked by the projection of the pointing medium (finger, hand, arm, head, etc.) to a region that is proximal or distant in relation to the speaker's origo. Deictic gestures are thus anchored to the space and time of the communicative act, and so their propositional content is understood as a function that maps from a world in its contextually-specific time and space to truth values. The same is not necessarily valid for depicting gestures: uttering "What a big cake" while performing a circular motion with both hands in the frontal centre is not related to the spatial and temporal context in which the utterance occurs. We therefore argue that whereas depicting gestures provide qualitative characteristics of the referent, deictic gestures are at heart quantitative. This is the diametrical distinction that sets apart depicting and deictic gestures, and that prevails in how their semantics is defined.

Note that by "gesture" we mean the expressive part of the whole movement, the kinetic peak of the excursion that carries the gesture's meaning—the so called *stroke*. What is intuitively recognised as a gesture, is known as a *gesture phrase*. It contains the following *phases*: a non-obligatory *preparation* (the hands are lifted from the rest position to the frontal space to perform the semantically intended motion), a non-obligatory *pre-stroke hold* (the hands are sustained in a position before reaching the kinetic peak), an obligatory stroke, and a non-obligatory *post-stroke hold* (the hands sustain their expressing position). The deictic stroke might be static (the pointing forelimbs are stationary in the expressive position) or dynamic (gesture's meaning is derived from a movement of the pointing forelimbs).

2.2 Range of Deictic Use

The deictic signal on its own is ambiguous with respect to the region pointed out and the syntactic and semantic relation between speech and deixis. To clarify the region's ambiguity, let's consider the following example: when pointing in the direction of a book, does the space demarcated by the deictic gesture identify with the physical object book, the location of the book—e.g., the table—or with the cover of the book? Often there is not an exact

correspondence between the region identified by the pointing hand, the so called 'pointing cone' (Kranstedt *et al.*, 2006) and the reference. Our formal model does not intend to solve this ambiguity since it has no effects on multimodal perception, and certain ambiguities remain unresolved in context similarly to unimodal input. Based on Lascarides & Stone (2009), we formally regiment the location of the pointing hand with the constant $\vec{c}$, that marks the physical location of the tip of the index finger. This combines with the hand's shape, orientation and movement to determine the region $\vec{p}$ actually marked the gesture—e.g., a stationary stroke with hand shape 1-index will make $\vec{p}$ a line (or even a cone) that starts at $\vec{c}$ and continues in the direction of the index finger. We will also be using a function v to map the physical space $\vec{p}$ designated by the gesture to the actual space it denotes.

We further stated that there is a range of distinct relations between the speech signal and the pointing signal. An example from Clark (1996) illustrates this: George points at a copy of Wallace Stegner's novel *Angle of Repose* and says: 1. "That book is mine"; 2. "That man was a friend of mine"; 3. "I find that period of American history fascinating". In 1., there is one-to-one correspondence between the gesture space and the physical space (so v is identity), and the speech referent for "man" and the deictic referent are also bound by *identity*. In 2., the denotation of the deictic gesture and that of the synchronous speech are not identical since the individual pointed at is not present at the exact coordinates projected by the pointing fingers, and so the relation would be rather *virtual counterpart*. Finally in 3., the deictic gesture's denotation is again not equal to that in speech, and they are connected through *depiction*. Further ambiguity arises even in the context of the co-occurring speech: does the pointing gesture while uttering "We turn right" identify the event e of turning or the direction x? Our formal model fully supports ambiguity and partial meaning since we map deictic form to an underspecified meaning representation whose main variable can resolve to either e or x in context, and we also connect speech and deictic referents in the grammar through an underspecified relation *deictic_rel* that is resolvable in context to several possible values, among them *identity, virtual counterpart, depiction*, and even *paraphrase*.

We have also observed that depending on how the hand is used in the pointing act, deictic gestures can designate regions of the visible space in two distinct ways: first, the form of the hand, including the location $\vec{c}$ of the tip of the index finger, identifies the region $\vec{p}$ in visible space that is designated by the gesture as exactly that region that is taken up by the hand itself. This use of deixis is common in living space descriptions and in direction giving dialogues; e.g., (1).[1]

(1) There's like a [$_{NN}$little] [$_{N}$hallway]
Hands are open, vertical, parallel to each other. The speaker places them between the centre and the left periphery.

Second, the hand marks a distant region in the visible space to establish a real or virtual identity between the individual pointed at and the individual referred to in speech as in (2), or to perform a meta-narrative function such as offering up an instance of an object or acknowledging the addressee's statement. In this case, the form of the hand, including the physical coordinate $\vec{c}$, establish a region $\vec{p}$ in visible space that does *not* overlap with the hand.

(2) …[$_{PN}$You] guys come from tropical [$_{N}$countries]
Speaker C turns slightly to the right towards speaker A pointing at him using Right Hand (RH) with palm open up.

§ 3.2 details how these two meanings are reflected in the formal semantic representation of deictic gesture.

It is generally assumed in the literature that deictic gesture combines with the temporally co-occurring speech signal without considering synchrony *outside* the temporal alignment (McNeill, 2005). For depicting gestures, we have shown elsewhere that synchrony is also possible beyond the strict temporal alignment of both signals (Alahverdzhieva & Lascarides, 2010). For deictic gestures, we have observed that the synchronous semantically related speech phrase can be a few milliseconds before or after the deictic stroke. In (3), for instance, the gesture is produced while uttering "Thank you" when obviously the denotation of the hand is identical to that of the computer mouse.

(3) [$_{N}$Thank] you. [$_{NN}$I'll] take the [$_{N}$mouse]
RH is loosely closed, index finger is loosely extended, pointing at the computer mouse

[1] For the utterance transcription, we have adopted the following convention: the speech signal aligned with the stroke is underlined, and the signal aligned with a post-stroke hold is underlined with a curved line. Here we have also included those words that start/end at midpoint in relation to the gesture phase boundaries. The pitch accented words are shown in square brackets with the accent type in the left corner: PN (pre-nuclear), NN (non-nuclear) and N (nuclear).

Upon our empirical study of the temporally misaligned occurrences, we learnt that the temporal relaxation is applicable in cases where the visible space $\vec{p}$ that is designated by the gesture is identical to the space $v(\vec{p})$ that it denotes (in other words, v is identity). Otherwise, any synchronicity between a deictic gesture and an individual not present at the exact coordinates of the gesture space would fail to produce the intended logical form in the specific context. We shall therefore equip our grammar with rules that apply only when there is an identity function mapping the visible space to space in denotation. This will support an analysis of (3) where the deixis does *not* denote the same individual as the pronoun "I". An alternative interpretation would be where the gesture is synchronous with the temporally co-occurring speech "Thank you" in which case the gesture complements on the speech by introducing a causal relationship of the sort "Thank you for handing me the mouse".

Having introduced the main challenges that we are dealing with, we now turn to the problem of how deixis and speech interact at the level of linguistic form (prosody) and meaning.

3 Speech-Deixis Interaction

Our motivation for unifying speech and gesture into a grammar stems from the descriptive accounts that gesture takes an integral part in language production and language comprehension (McNeill, 2005). We thus analyse deixis in *synchrony* with speech, as a mapping from form to some (underspecified) meaning in the final logical form of the utterance. Due to the controversial findings concerning the temporal alignment of speech and gesture, Alahverdzhieva & Lascarides (2010) proposed the following definition of synchrony, which considers only qualitative factors coming from form and meaning:

Definition 1 Synchrony. *The choice of which linguistic phrase a gesture stroke is synchronous with is guided by: i. the final interpretation of the gesture in specific context-of-use; ii. the speech phrase whose content is semantically related to that of the gesture given the value of (i); and iii. the syntactic structure that, with standard semantic composition rules, would yield an underspecified logical formula supporting (ii) and hence also (i).*

The gestural signal and the spoken signal are closely related on both the level of form and of meaning. We view form as a matter of temporal co-occurrence between the two modalities: there is increasing evidence in the literature that gesture performance is constrained by the prosody of speech, both speech and gesture are integrated into a common rhythmical system, and the perception of one mode is dependent on the performance of the other—e.g., Loehr (2004), Giorgolo & Verstraten (2008). We shall perform some experiments to validate these claims, and hence equip our grammar with the constraints on the mapping between form and meaning of co-speech deictic actions that stem from the relative temporal performance of gesture and speech, and prosody (among other factors), where these constraints model our empirical findings in multimodal corpora.

3.1 Prosody

In this project, we adopt the Autosegmental-Metrical (AM) theory (term coined by Ladd (1996)) for the analysis of speech prosody. This theory views prosodic prominence as a relational property between two juxtaposed units where the prominence of unit A is determined by its (strong or weak) relation to unit B.

Based on the findings of a previous prosody study (Calhoun, 2006), we argue that it is not the lower or higher tune but rather the nuclear accenting that constrains the alignment between gesture and speech. We view nuclear accenting as the perception of phrase-level prominence which is relative to the metrical structure, and not to the acoustic properties of the syllables. In the AM model, nuclear prominence results from the following operations: (a). mapping a syntactic structure to a binary metrical tree; (b). assigning *strong (s)* or *weak (w)* prosodic weight to the nodes in the metrical tree according to the metrical formulation of the Nuclear Stress Rule (Liberman & Prince, 1977, p.257) as shown in Definition 2; and (c). tracing the path dominated by s nodes.

Definition 2 Nuclear Stress Rule. *In a configuration $[_C AB]$, if C is a phrasal category, B is strong.*

In the default case of broad focus, the metrical structure is right-branching—that is, the nuclear accent is associated with the right-most word. For instance, Figure 1[2] illustrates the metrical tree for "fasten a cloak" in its broad

[2]The example is taken from Klein (2000)

focused reading with the nuclear accent being on the word entirely dominated by s nodes—"cloak". Liberman & Prince (1977) call the most prominent element of a given constituent *Designated Terminal Element* (DTE).

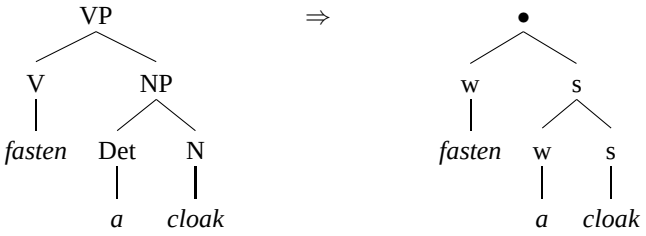


Figure 1: Syntactic Tree and Metrical tree

Strong nodes on the left of the nuclear accent can also appear, and these are known as pre-nuclear accents. Unlike the nuclear accents, pre-nuclear accents are signalled by their acoustic properties rather than their relative position in the metrical tree.

3.1.1 Empirical Study

We used empirical data to determine constraints on the interaction between deictic gestures and speech signals.

Hypothesis 1 *Deictic gestures align with the nuclear pitch accents in speech both in the default case of broad focus, and in case of narrow focus. In case of early pre-nuclear rise, deictic gestures align with the pre-nuclear pitch accents.*

To test the validity of our hypothesis, we used two multimodal corpora: a 5.53 min recording from the Talkbank Data,[3] and observation IS1008c, speaker C from the AMI corpus.[4] The domain of the former is living-space descriptions and navigation giving, and the latter is a multi-party face-to-face conversation among four people discussing the design of a remote control. Annotation on both corpora proceeded in two separate stages: annotation of speech which included word transcription, pitch accents pointing to words and prosodic phrases; and gesture annotation which included marking of gesture phrases, gesture phases, and also formless moves that beat along the speech rhythm known as beats. Both annotations were performed independently from each other.

Prosody Annotation As an annotation tool, we used Praat (Boersma & Weenink, 2003). Our annotation schema is largely based on the guidelines of the prosody annotation of the Switchboard corpus (Brenier & Calhoun, 2006). We marked the following layers:

1. *Orthographic Transcription.*
2. *Pitch Accents.* Words were unambiguously associated with at least one accent of the following type: *nuclear:* the accent of the whole prosodic phrase that is structurally, and not phonetically perceived as the most important one; *pre-nuclear:* an early emphatic high rise characterised by a high pitch contour; *non-nuclear:* unlike nuclear accents, non-nuclear accents are perceived on the basis of their phonetic properties, and the rhythm of the sentence (they correspond to 'plain' or 'regular' accents in Brenier & Calhoun (2006) and Calhoun (2006)); *none*: a non-discernible accent in a phrase (it corresponds to a 'Z' accent in Brenier & Calhoun (2006)); *?*: uncertainty concerning the presence of an accent.
3. *Prosodic Phrases.* A group of words form a prosodic phrase whose type is determined by the break type after the last word in the phrase. We annotated the following phrases: *disfluent*: phrase where the break after the last word would be marked in ToBI with the *p* diacritic, that is *1p, 2p, 3p* correspond to disfluent phrases; *minor*: phrase where the break after the last word corresponds to ToBI break 3; *major*: phrase where the break after the last word corresponds to ToBI break 4; *backchannel*: short phrases containing only fillers such as "er", "um", "you know", etc.

[3]The video clip can be found here `http://www.talkbank.org/media/Gesture/Cassell/kimiko.mov`
[4]`http://corpus.amiproject.org`

Gesture Annotation We used the Anvil labelling tool (Kipp, 2001) to annotate the gesture phrases, gesture phases and beats. Along the lines of Loehr (2004), we annotated gestures for the dominant *H1* hand, and for the non-dominant *H2* hand. Bi-handed gestures where the movement of *H1* was symmetrical to *H2* were coded in *H1*.

1. *Hand Movement.* The annotation of the hand movement proceeded in two main passes. The first pass aimed at marking the temporal boundaries of all hand movements, and performing a binary classification on them in terms of *communicative–non-communicative* signals. The second pass determined what dimensions the communicative signals belong to, they being *literally depicting, metaphorically depicting* or *deictic*. To stay consistent with the findings in the literature that a single gesture can have dimensions of, say, depicting and deictic gestures (McNeill, 2005), our annotation schema permitted for marking gestures belonging to more than one dimension.
2. *Gesture Phases.* This step involved annotating the phases comprising each hand movement: *preparation, pre-stroke hold, stroke, post-stroke hold* and *retraction*. The distinction between pre-stroke holds and post-stroke holds was often not clear, that is, the form of the hand itself was ambiguous as to whether the signal belonged to the new gesture phrase and it was thus a pre-stroke hold, or it belonged to the previous gesture phrase, and it was thus a post-stroke hold. We observed that pre-stroke holds tend to appear with hesitation pauses while the speaker is looking for some stable verbal form, and so recovery of the temporal cohesion is anticipated; contrarily, post-stroke holds are more likely to occur with fluent speech when the speaker elaborates on the content reached during the stroke.
3. *Beat.* Beat movements were marked in a separate layer so as to study whether they always superimpose other gestural dimensions, or pure beats also occur.

Past annotation tasks of the Switchboard corpus (Calhoun, 2006) and of the multimodal corpus of Loehr (2004) have shown that the annotation of accents and boundaries is reliable (see Table 1), and also the annotation of gesture dimensions (see Table 2).

	All Types	**+/-**
Accents	0.800	0.800
Boundaries	0.889	0.910
Words	(752)	

Table 1: Inter-coder reliability for accents and phrase boundaries & for the presence/absence (=/-) of an accent/boundary in *kappa* (Calhoun, 2006)

	Coding	**Segmentation**
Hand movement	0.8536/0.8994	0.8502/0.8659
Deictic gesture	0.8605/0.8994	0.8502/0.8659
Literally depicting	0.8663/0.8916	0.8502/0.8659
Metaphorically depicting	0.8221/0.8623	0.8502/0.8659
Gesture phase	0.662/0.7	0.8864/0.8971
Beat	0.6599/0.8203	

Table 2: Inter-coder reliability for gesture coding agreement & segmentation agreement in *Cohen's kappa/corrected kappa*

Multimodal Corpora in NXT The annotated corpora were converted into Nite XML Technology (NXT) format (Carletta *et al.*, 2005), (Calhoun *et al.*, 2010) which allows for querying a corpus as a coherent set and extracting information from it by exploring the relations between the annotation layers. A corpus in NXT consists of 'observations'—our two video recordings—and annotations associated with it—orthographic transcriptions, pitch accents, prosodic phrases, gesture phrases, gesture phases and beats. Each data object is necessarily equipped with timestamps which are synchronised with the video and/or audio signal.

Data objects can be bound either by structural or by temporal relations which is specified in a meta-data file containing the annotation schema of the corpus. The type of relation also determines the query that can be executed onto these objects. The annotation of each data object is stored in a separate XML file; any relations between the annotation objects are defined in terms of stand-off links between the elements. Figure 2 illustrates the relation between the 'accents' and 'words' layers: the accent's attribute nite:pointer serves as a pointer to the unique nite:id of the relevant word. In this way, we can elegantly capture accents not overlapping a word, accents associated with two words, and also words associated with two accents.

We further specified the relationships between gestures and gesture phases, and between prosodic phrases and words as parent-child relations. This choice of representation is consistent with the essence of prosodic phrases and gesture phrases: prosodic phrases are made up by a certain number of words, and so the beginning of the first word aligns with the beginning of the prosodic phrase, and the end of the last word aligns with the end of the prosodic phrase. The same mechanism applies to gestures which are made up by at least one gesture phase. We

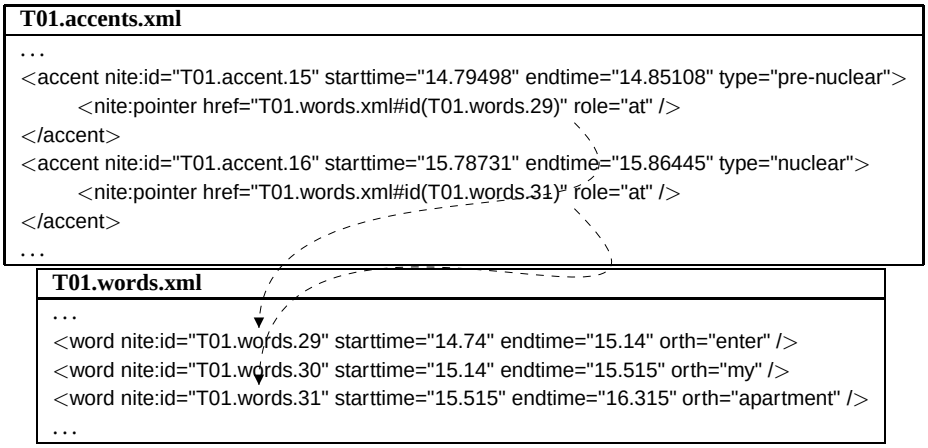


Figure 2: NXT Coding of Accents Associated with Words

forego any details about the specification of beats since they are not represented in a structural relationship with other layers.

3.1.2 Results and Discussion

In relation to our hypothesis, we searched for the types of accents overlapping a deictic gesture stroke. The corpora contained 87 deictic strokes (65 for the Talkbank, and 22 for AMI). 86 of them—that is, 98.85%—overlapped a nuclear and/or a pre-nuclear accented word. Strokes overlapping a combination of non-nuclear and nuclear accented words were also common. Essentially, the empirical analysis confirmed the expected alignment between the nuclear prominent word (not simply the nuclear accent) and the gesture stroke both in case of broad focus, and in case of narrow-focused utterances. The following two utterances illustrate our findings: (4) is a broad-focused utterance with the nuclear accent being on the right-most word. Utterance (5), which is a continuation of (4), displays narrow focus with the nuclear accent pointing to the first word of the prosodic phrase–"left". The interaction between prosodic prominence and gesture stroke appears to be on the level of information structure: nuclear prominence, along with gesture stroke aligns with the focused (kontrastive)[5] elements that push the communication forward, and not with those available from the background. This prediction has its grounds in the descriptive literature of gesture where "a break in the continuity" (Givón, 1985) of the narrative implies "highest degree of gesture materialisation" (McNeill, 2005, p.55).

(4) I keep [$_N$going] until I [$_{NN}$hit] Mass [$_N$Ave], I think
Right arm is bent in the elbow at a 90-degree angle, RH is loosely closed and relaxed, fingers point forward. Left arm is bent at the elbow, held almost parallel to the torso, palm is open vertical facing forward, finger tips point to the left

(5) And then I [$_N$turn] *[pause]* [$_N$left] on [$_{NN}$Mass] Ave
Hands are held in the same position as in (4), then along with "left" RH moves to the left periphery over LH, RH is vertically open

The single counterexample concerns the first gesture in (6): at this stage we remain agnostic as to why this misalignment occurred. As long as it is not a recurrent feature found over a larger amount of data, we would rather attribute it to impreciseness of annotation than to a general phenomenon to be considered in a model of multimodal actions.

(6) [$_{NN}$Between] the living [$_N$room] and *[pause]* the [$_N$study] and the *[pause]* [$_N$bedroom]
Hands are in the front centre, bent in elbows, palms are open, vertical, facing each other; along with "between", they perform a loose sweeping movement to the right periphery, then LH moves away to the

[5] In the Information Structure literature kontrast designates "parts of the utterance—actually, words—which contribute to distinguishing its actual content from alternatives the context makes available." (Kruijff-Korbayová & Steedman, 2003)

left upper centre with palm vertical, finger tips oriented forward; along with "the study", RH is moved in parallel to LH, as if both hands place a rectangular object in space

Further to this, we looked at all-new information utterances with an initial strong acoustic pitch and then a nuclear accent on the right-most element. In these utterances, the stroke was performed along with the initial pre-nuclear accent, and there might have been a post-stroke hold on the other components of the utterance. This is exemplified in (7) where an initial meaningful speech segment aligns with the stroke, and then the content is elaborated while holding the hands in an expressive position.

(7) I [$_{PN}$enter] my [$_{N}$apartment]
Hands are in centre, palms are open vertically, finger tips point forward; along with "enter" they move briskly downwards.

We use the results of this statistical analysis to define constraints on the temporal overlap between deictic gesture and speech. Also, we need to explore whether any semantic relation can be established between the temporally aligned signals.

3.2 Mapping Deixis Form to Deixis Meaning

Following previous research (Johnston, 1998), (Kopp *et al.*, 2004), the form of the pointing hand is represented using typed feature structures, where each feature value pair corresponds to an aspect of form. We use fine-grained an analysis as possible: we consider that the shape of the hand, the orientation of the palm and the fingers, the hand movement, and also the location of the hand in the spatio-temporal coordinates $\vec{c}$ are the distinct classes of form that potentially have semantic effects, e.g., the shape of the hand influences the mapping from $\vec{c}$ to $\vec{p}$. Moreover, this form representation captures the fact that the different attributes composing deictic gesture's form are not hierarchically ordered, but are rather a flat list. Figure 3 gives the form representation of the gesture in (1)—the value $\vec{c}$, which identifies the spatio-temporal coordinates of the hand, together with the other values, serve to identify the region $\vec{p}$ designated by the gesture's content (Lascarides & Stone, 2009); as explained in § 2.2, a pointing gesture (with hand shape 1-index) will make $\vec{p}$ denote a cone or line that starts at *hand-location* $\vec{c}$ and whose direction is the same value as *finger-direction* (Kranstedt *et al.*, 2006). Note also that the gesture is typed as *communicative_gesture_deictic* to distinguish between form features contributed by depicting gestures, and those contributed by pointing gestures.

communicative_gesture_deictic	
HAND-SHAPE:	open-flat
PALM-ORIENTATION:	vertical
FINGER-ORIENTATION:	forward
HAND-MOVEMENT:	away-centre-left
HAND-LOCATION:	$\vec{c}$

Figure 3: TFS Representation of Form of Deictic Gesture

As a semantics description language we use Robust Minimal Recursion Semantics (RMRS) since it is highly flexible about the semantic underspecification it supports: in RMRS, one can leave the main predicate underspecified until resolved by further context. In this way, we can elegantly capture the fact that the form of a deictic gesture alone does not fully determine its content — e.g., it does not determine whether the gesture denotes an individual or an event, but rather contextual information is needed as well to infer this aspect of the gesture's (pragmatic) interpretation. Defining flat semantics in RMRS involves defining a set of *Elementary Predications (*EPs). Each EP is associated with a label l_i that ultimately identifies the scopal position of the predicate in the context-resolved logical form. Shared labels are also possible, and they mark implicit conjunction as in intersective modifiers. Each EP is also associated with a unique anchor a_i, which serves as its locus for specifying arguments to the EP — e.g., $ARG2(a, x)$ means that the second argument to the EP whose anchor is a is the individual x. The absence of such ARG relations in the RMRS thus serves to underspecify the arguments to predicates and even the predicate's arity. Holes (h_i) are used to represent scopal arguments whose value is not fully determined by syntax. The admissible pluggings are constrained by equality conditions ($=_q$) between holes and labels ($h_i =_q l_i$ means that only 0 or more quantifiers intervene between the scopal positions). Finally, a top label h_0 is added for the whole formula.

§ 2.2 detailed the two distinct functions of deictic gestures. We will now present their compositional semantics as follows:

1. **Hand as reference.** The speaker here points to an individual/event represented by the hand which is located at the spatial coordinate $\vec{p}$ designated by the finger tips often, but not necessarily, in relation to another individual available from the discourse. The form features of the pointing hand further constrain the set of possible relations between gesture and speech, e.g., an open hand supine used for turn coordination can resolve to a *metatalk* relation (Lascarides & Stone, 2009) — roughly put, the gesture can have a meaning that can be paraphrased by the parenthetical phrase "I am telling you".

   The RMRS representation of the gesture in (1) is shown in Figure 4. Following Lascarides & Stone (2009), this RMRS semantics says that the pointing hand provides the spatial reference of an underspecified referent i (an individual or an event) at some position in the physical space $v(\vec{p})$. In context, the underspecified variable i may resolve to an individual x as in (1), or to an event e as in (7). To stay consistent with the findings in the descriptive literature, namely that the shape of the pointing hand is associated with a specific meaning (Kendon, 2004), we map each form feature-value pair to a two-place predicate. Their formal treatment is similar to the treatment of intersective modifiers in the English Resource Grammar (ERG) in that they share labels with the main predicate sp_ref. Again for consistency with ERG where individuals are bound by quantifiers, there is a quantifier outscoping the referent introduced by the deictic gesture. Following Lascarides & Stone (2009), we use the $\mathcal{G}$ operator so as to guarantee that individuals referred in speech cannot be co-referred to individuals introduced in gesture. To obtain this, $\mathcal{G}$ must outscope all gesture predications (formalised in terms of $=_q$ equality conditions).
2. **Reference is the region marked by the hand.** The hand here also points to an underspecified reference i located at $v(\vec{p})$ but unlike the previous function, the hand shape denotes not the reference itself but the region marked by it. The semantics of the gesture in (2) is shown in Figure 5, and it is similar to the one displayed in Figure 4 with the only difference being that it is the region that is modified by the various gesture form-features. Since the rest of the predications remain the same, we forego any details about them.

$l_0 : a_0 : [\mathcal{G}](h_1)$
$l_1 : a_1 : deictic\_q(i)\ RSTR(a_1, h_2)\ BODY(a_1, h_3)$
$l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p}))$
$l_2 : a_3 : hand\_shape\_open\_flat(e_0)\ ARG1(a_3, i)$
$l_2 : a_4 : palm\_orient\_vertical(e_1)\ ARG1(a_4, i)$
$l_2 : a_5 : finger\_orient\_forward(e_3)\ ARG1(a_5, i)$
$l_2 : a_6 : hand\_move\_away\_centre\_left(e_5)\ ARG1(a_6, i)$
$h_1 =_q l_1;\ h_1 =_q l_2;\ h_2 =_q l_2$

Figure 4: RMRS for Hand as Reference

$l_0 : a_0 : [\mathcal{G}](h_1)$
$l_1 : a_1 : deictic\_q(i)\ RSTR(a_1, h_2)\ BODY(a_1, h_3)$
$l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p}))$
$l_2 : a_3 : RH\_palm\_orient\_vertical(e_1)\ ARG1(a_3, \vec{p})$
$l_2 : a_4 : RH\_finger\_orient\_forward(e_2)\ ARG1(a_4, \vec{p})$
$l_2 : a_5 : RH\_hand\_move\_away\_body\_left(e_3)\ ARG1(a_5, \vec{p})$
$h_1 =_q l_1;\ h_1 =_q l_2;\ h_2 =_q l_2$

Figure 5: RMRS for the Region Marked by the Hand

4 Rules for Combining Deixis and Speech in the Grammar

We intend to augment the existing wide-coverage grammar for English—the English Resource Grammar ERG—with construction rules for combining speech and gesture. This task involves specifying the prosodic component in the grammar (we shall be using the AM theory), and also interfacing it with the syntax-semantics component.

We formally regiment our findings about the deixis-prosody interaction (§ 3.1) into the following basic construction rules:

Definition 2.1 *Deictic gesture attaches to the temporally overlapping nuclear/pre-nuclear head word.*

Definition 2.2 *Deictic gesture attaches to the temporally overlapping nuclear/pre-nuclear head word after it had been combined with the arguments and/or modifiers to the head.*

The motivation to include the latter stems from the fact that semantically the deictic signal is not strictly constrained to its temporally co-occurring word but rather it can be linked to a larger phrase. For instance, in (7), there is no information coming from the form of the hand, nor from its relative timing that it should be attached to "enter" only, and not to "enter my apartment", in which case the form of the hand would be related to the rectangular shape of, say, an entrance door to an apartment. Intuitively in this case, the gesture directs not only to the point of entering the house, but also to the entrance door which by the hand shape is rectangular.

The syntactic structure is derived in parallel with the prosodic one, and so the syntactic component would consist either of a single head word without further constraints on its syntactic category, or a larger phrase it being a head-argument, a head-modifier phrase, or an entire utterance. Generally speaking, a deictic gesture cannot be

combined with a non-prosodic word. We will come back to this point a bit later, when we will see that some exceptions to this rule can also arise. Finally, the semantic component uses the RMRS representation in § 3.2.

Semantic composition with RMRS (Copestake *et al.*, 2001), (Copestake *et al.*, 2005) is monotonic, ensuring that the semantic representations of the daughters are always subsumed by that of the mother. For each phrase, one specifies semantic entities (*sement*s) of the following parts: (1). *Top*: the global label containing the whole formula. During composition, the top labels of the daughters are equated with the top label of the mother to demonstrate the derivation of a single logical form; (2). *Hook*: placeholder that records the semantic value of the formula. It contains (a). *local top*: the label containing an EP. For instance, in Figure 4 the local top of $l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p}))$ is l_2; (b). *semantic index* ($i_1, i_2 \ldots i_n$): it indicates what the phrase is about and has two subtypes: events ($e_1, e_2 \ldots e_n$) and individuals ($x_1, x_2 \ldots x_n$). It is obtained by co-indexation with the topmost EP. For instance, in Figure 4, the semantic index of the phrase is i obtained by co-indexing it with the main variable of $sp\_ref(i)$; (3). *Slots*: resources that need to be consumed so that a functor becomes semantically saturated; (4). *Rels*: a bag of EPs; (5). *Equality constraints* ($=_q$): scopal constraints indicating the admissible plugging of a subformula into a hole.

To summarise, an RMRS sement is: $\langle Top\,[ltop, i]\,\{slots\}\{rels\}\,[=_q]\rangle$. Semantic composition of a $sement_M = op(sement_{D1}, sement_{D2})$ involves the following operations: 1. making Top of $sement_M$ = $sement_{D1}$ = $sement_{D2}$; 2. making the hook of $sement_M$ the hook of $sement_{D1}$; 3. making the remaining slots of $sement_{D1}$ and $sement_{D2}$ the slots of $sement_M$; 4. making the $rels$ and $hcons$ of $sement_M$ the union of those of the daughters.

As argued in § 2.2, deictic gesture always relates with the synchronous speech signal through some sort of relation; e.g., *identity, virtual counterpart* or a *paraphrase* relation. Based on Lascarides & Stone (2009), the construction rule therefore introduces an underspecified relation $deictic\_rel(i_1, i_2)$ between the semantic index i_1 of the deictic gesture and the semantic index i_2 of speech. How this relation resolves is a matter of discourse context. Similarly to the treatment of intersective modification in language, *deictic_rel* shares the same label as the speech head daughter since it further restricts the individual/event introduced in speech. In so doing, any quantifier outscoping the head would also outscope this relation. This is similar to the treatment of appositives in ERG.

With this machinery at hand, let us now turn to the derivation of utterance (1): the deictic gesture overlaps the nuclear-accented prosodic word "hallway", and hence we could build a single situated noun out of "hallway" and deixis as demonstrated in Figure 6. In semantics, we need to extend the RMRS representation in Figure 4 with a top label, hook and slots as follows (for the sake of readability, we gloss the semantic predications contributed by the deictic form features as $l_2 : a_3 : deictic\_eps(e_0)\ ARG1(a_3, i)$):

$< h_0, [l_0, a_0, i]\,, \{\}$
$\{l_0 : a_0 : [\mathcal{G}](h_1)$
$l_1 : a_1 : deictic\_q(i)\ RSTR(a_1, h_2)\ BODY(a_1, h_3)$
$l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p}))$
$l_2 : a_3 : deictic\_eps(e_0)\ ARG1(a_3, i)\}$
$[h_1 =_q l_1;\ \ h_1 =_q l_1;\ \ h_2 =_q l_2] >$

The top label of the whole formula is h_0 and in derivation it is made identical with that of the mother. Further, the local top is identified with the label of the $\mathcal{G}$ operator which contains all other predications. The semantic index of the deictic gesture is the underspecified variable i introduced by the sp_ref predicate that in composition resolves to x_1. Finally, we assign no slots to the formula.

As argued above, the form of the deictic signal is not sufficient to decide whether the hand refers to "hallway", to "little hallway" or even to "a little hallway", and so our grammar does not impose constraints on the syntactic phrase, and is thus able to generate all these combinations. Prosodically, we integrate deixis into a metrical tree where the prosodically prominent element, the DTE, is "hallway" , and syntactically into a head-modifier construction with "hallway" being the head daughter. Since *deictic_rel* shares the same label as the head, when combining the NP "a little hallway" and deixis, both the head noun and the deictic relation would appear within the restrictor of the quantifier. The semantic composition remains the same as above.

We shall now look at some exceptions that are not covered by the temporal alignment constraint and the nuclear prominence constraint. We saw that in (3) there exists an obvious misalignment between the semantically related speech and deixis signals. Similarly in (8), the deixis denotation is identical to the denotation of "she" despite it not being prosodically prominent.

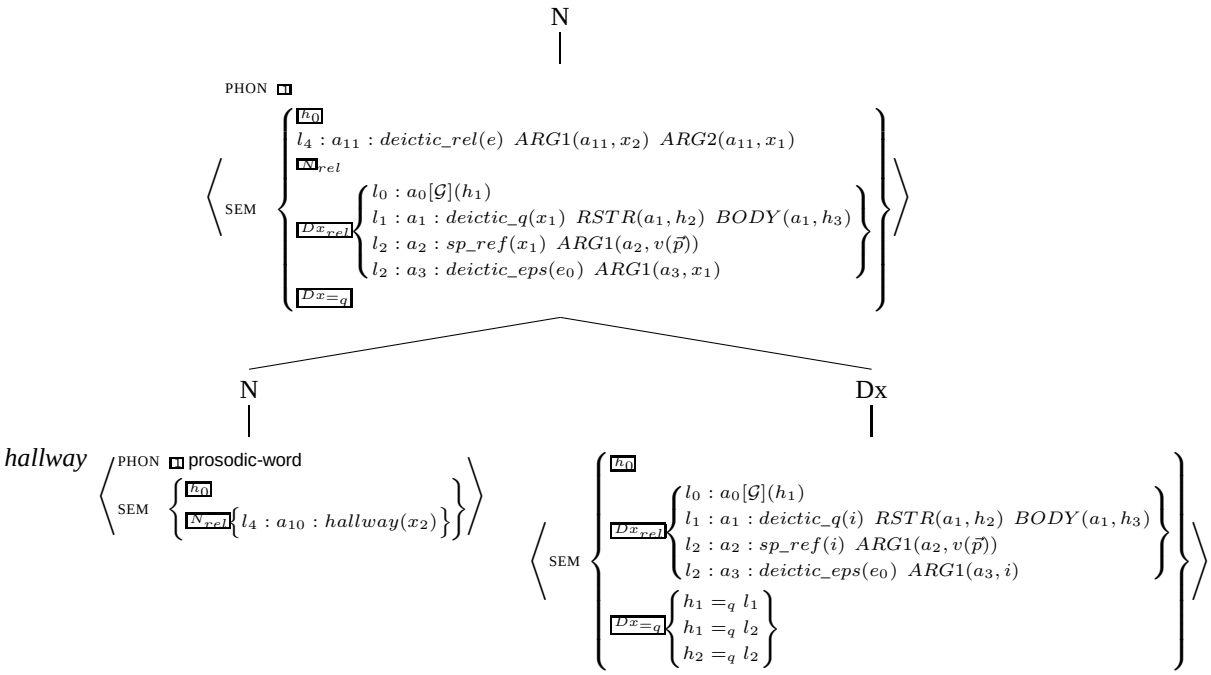


Figure 6: Derivation Tree for Deictic Gesture and the N "hallway"

(8) And a as she [$_N$said], it's an environmentally friendly uh material
Speaker C extends right hand palm supine towards the speaker B

To cope with these exceptions, we studied all utterances where the semantically preferred attachment of the deictic gesture is an element beyond its temporal performance and/or a non-prosodically prominent element. This temporal/prosodic relaxation is a matter of making individuals in the surrounding space salient and it is thus necessary only in utterances where the gesture's denotation is physically present in the visible space, that is, there is an identity between the physical space that the hand points at and the actual denotation of the gesture's referent. Of course, this does not mean that *deictic_rel* would always in this case resolve to identity. Let us illustrate this by reusing the example from Clark (1996) in § 2.2: when pointing to the novel while uttering "This man was a friend of mine" there is an identity between the visible space that the hand points at and the denotation of the gesture since the novel is salient in the physical space gesture points at. However, the denotation of the gesture is not identical to the one of speech, and we therefore claimed that the relation between speech and deixis is rather *depiction*. In our grammar, we therefore spell out the following rule:

Definition 2.3 *Deictic gesture attaches to an item (prosodically prominent or non-prominent) whose temporal performance is adjacent to that of the gesture if the mapping v resolves to identity.*

Importantly, this relaxation is not applicable in cases where the hand serves as an abstract reference pointing to an object not present in the communicative act. If the gesture in (4) was related to the speech head daughter "I", the logical form would fail to resolve.

5 Conclusions

In this paper, we demonstrated that well-established methods from linguistics are sufficient to provide the form-meaning mapping of multimodal communicative actions consisting of speech and deictic gesture. This goal was achieved by integrating them into a multimodal grammar thereby using constraints coming from the form of the speech signal, the form of the gesture signal and their relative temporal performance so as to map them to a single meaning representation in the final logical form of the utterance. This paper contributed to the existing resources by setting the theoretical framework for a multimodal grammar, and also by extending the existing corpora with prosody and gesture annotation in the NXT format which can further be used for various studies of multimodal communication.

ALAHVERDZHIEVA K. & LASCARIDES A. (2010). Analysing speech and co-speech gesture in constraint-based grammars. In S. MÜLLER, Ed., *The Proceedings of the 17th International Conference on Head-Driven Phrase Structure Grammar*, p. 6–26, Stanford: CSLI Publications.

BOERSMA P. & WEENINK D. (2003). *'Praat:doing phonetics by computer'*. `http://www.praat.org`.

BRENIER J. & CALHOUN S. (2006). Switchboard prosody annotation scheme. Department of Linguistics, Stanford University and ICCS, University of Edinburgh. Internal publication.

CALHOUN S. (2006). *Information Structure and the Prosodic Structure of English: a Probabilistic Relationship*. University of Edinburgh. PhD Thesis.

CALHOUN S., CARLETTA J., BRENIER J., MAYO N., JURAFSKY D., STEEDMAN M. & BEAVER D. (2010). The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, **44**, 387–419.

CARLETTA J., EVERT S., HEID U. & KILGOUR J. (2005). The nite xml toolkit: Data model and query language. *Language Resources and Evaluation*, **39**, 313–334.

CLARK H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

COPESTAKE A., FLICKINGER D., SAG I. & POLLARD C. (2005). Minimal recursion semantics: An introduction. *Journal of Research on Language and Computation*, **3**(2–3), 281–332.

COPESTAKE A., LASCARIDES A. & FLICKINGER D. (2001). An algebra for semantic construction in constraint-based grammars. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL/EACL 2001)*, p. 132–139, Toulouse.

GIORGOLO G. & VERSTRATEN F. (2008). Perception of speech-and-gesture integration. In *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, p. 31–36.

GIULIANI M. & KNOLL A. (2007). Integrating multimodal cues using grammar based models. In *HCI (6)*, p. 858–867.

GIVÓN T. (1985). Iconicity, Isomorphism and Non-arbitrary Coding in Syntax. In J. HAIMAN, Ed., *Iconicity in Syntax*, p. 187–219. Amsterdam: John Benjamins.

GOFFMAN E. (1963). *Behavior in Public Places: Notes on the Social Organization of Gatherings*. The Free Press.

JOHNSTON M. (1998). Multimodal language processing. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

KENDON A. (2004). *Gesture. Visible Action as Utterance*. Cambridge: Cambridge University Press.

KIPP M. (2001). Anvil — a generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg: Georgetown University.

KLEIN E. (2000). Prosodic constituency in hpsg. In *Grammatical Interfaces in HPSG, Studies in Constraint-Based Lexicalism*, p. 169–200: CSLI Publications.

KOPP S., TEPPER P. & CASSELL J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. In *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*, p. 97–104, New York, NY, USA: State College, PA, USA ACM.

KRANSTEDT A., LÜCKING A., PFEIFFER T., RIESER H. & WACHSMUTH I. (2006). Deixis: How to determine demonstrated objects using a pointing cone. In S. GIBET, N. COURTY & J.-F. KAMP, Eds., *Gesture in Human-Computer Interaction and Simulation*, volume 3881 of *Lecture Notes in Computer Science*, p. 300–311. Springer Berlin / Heidelberg.

KRUIJFF-KORBAYOVÁ I. & STEEDMAN M. (2003). Discourse and information structure. *Journal of Logic, Language and Information*, **12**, 249–259.

LADD R. D. (1996). *Intonational Phonology (first edition)*. Cambridge University Press.

LASCARIDES A. & STONE M. (2009). A formal semantic analysis of gesture. *Journal of Semantics*.

LIBERMAN M. & PRINCE A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, **8**(2), 249–336.

LOEHR D. (2004). *Gesture and Intonation*. Washington DC: Georgetown University. Doctoral Dissertation.

MCNEILL D. (2005). *Gesture and Thought*. Chicago: University of Chicago Press.

OVIATT S. L., DEANGELI A. & KUHN K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. *CHI*, p. 415–422.

Traduction et Alignement

TALN 2011, Montpellier, 27 juin –1er juillet 2011

Généralisation de l'alignement sous-phrastique par échantillonnage

Adrien Lardilleux[1] François Yvon[1,2] Yves Lepage[3]
(1) LIMSI-CNRS, BP 133, 91403 Orsay Cedex
(2) Université Paris-Sud
(3) IPS, université Waseda, Japon
Adrien.Lardilleux@limsi.fr, Francois.Yvon@limsi.fr, Yves.Lepage@aoni.waseda.jp

Résumé. L'alignement sous-phrastique consiste à extraire des traductions d'unités textuelles de grain inférieur à la phrase à partir de textes multilingues parallèles alignés au niveau de la phrase. Un tel alignement est nécessaire, par exemple, pour entraîner des systèmes de traduction statistique. L'approche standard pour réaliser cette tâche implique l'estimation successive de plusieurs modèles probabilistes de complexité croissante et l'utilisation d'heuristiques qui permettent d'aligner des mots isolés, puis, par extension, des groupes de mots. Dans cet article, nous considérons une approche alternative, initialement proposée dans (Lardilleux & Lepage, 2008), qui repose sur un principe beaucoup plus simple, à savoir la comparaison des profils d'occurrences dans des sous-corpus obtenus par échantillonnage. Après avoir analysé les forces et faiblesses de cette approche, nous montrons comment améliorer la détection d'unités de traduction longues, et évaluons ces améliorations sur des tâches de traduction automatique.

Abstract. Sub-sentential alignment is the process by which multi-word translation units are extracted from sentence-aligned multilingual parallel texts. Such alignment is necessary, for instance, to train statistical machine translation systems. Standard approaches typically rely on the estimation of several probabilistic models of increasing complexity and on the use of various heuristics that make it possible to align, first isolated words, then, by extension, groups of words. In this paper, we explore an alternative approach, originally proposed in (Lardilleux & Lepage, 2008), that relies on a much simpler principle, which is the comparison of occurrence profiles in sub-corpora obtained by sampling. After analyzing the strengths and weaknesses of this approach, we show how to improve the detection of long translation units, and evaluate these improvements on machine translation tasks.

Mots-clés : alignement sous-phrastique, traduction automatique par fragments.

Keywords: sub-sentential alignment, phrase-based machine translation.

1 Introduction

L'alignement sous-phrastique consiste à extraire des traductions d'unités textuelles de grain inférieur à la phrase à partir de corpus multilingues parallèles, c'est-à-dire dont les phrases ont préalablement été mises en correspondance. Cette tâche constitue la première étape de la plupart des systèmes de traduction automatique fondés sur les données (traduction statistique et traduction par l'exemple). Les systèmes qui concentrent aujourd'hui les efforts de recherche sont majoritairement des systèmes statistiques par fragments (*phrases* en anglais), qui utilisent comme principale ressource une table de traductions, dérivée d'alignements sous-phrastiques. Un telle table consiste en une liste pré-calculée de couples de traductions associant à chaque couple de fragments (*source*, *cible*) un certain nombre de scores reflétant la probabilité que *source* se traduise par *cible*.

On peut globalement inscrire les méthodes d'alignement sous-phrastique dans l'un des deux courants suivants : l'approche estimative, introduite par Brown *et al.* (1988), et l'approche associative, introduite par Gale & Church (1991). La première est la plus utilisée à ce jour, principalement parce qu'elle est parfaitement intégrée à la traduction automatique statistique, dont elle constitue un pilier depuis l'apparition des modèles IBM (Brown *et al.*, 1993). Cette approche consiste à définir un modèle probabiliste du corpus parallèle dont les paramètres sont estimés selon un processus de maximisation globale sur l'ensemble des couples de phrases disponibles. Pratiquement, le but est de déterminer les meilleurs appariements possibles entre les mots sources et cibles dans chacun des couples de phrases parallèles. Dans la seconde approche, on établit une liste de traductions candidates soumises à un test d'indépendance statistique, tels que l'information mutuelle (Fung & Church, 1994) ou le rapport de vraisemblance

(Dunning, 1993) — voir (Melamed, 2000; Moore, 2005) pour des travaux récents dans cette lignée. Il s'agit ici d'un processus de maximisation *locale* : chaque segment est traité indépendamment des autres. Cette approche est plus souvent utilisée pour extraire directement des couples de traductions, tandis que la première cherche avant tout à établir des *liens* de traduction entre les mots sources et cibles de chacun des couples de phrases du corpus d'entrée. Ces liens permettent, dans un deuxième temps, d'extraire des couples de traductions.

Nous avons récemment proposé une méthode d'alignement sous-phrastique (Lardilleux & Lepage, 2008, 2009; Lardilleux, 2010), apparentée aux méthodes associatives, s'attaquant à un certain nombre de problèmes souvent négligés dans le domaine : traitement simultané de multiples langues, parallélisme massif, passage à l'échelle au cœur de la méthode, et simplicité de mise en œuvre. En moyenne, cette méthode s'est révélée meilleure que l'état de l'art sur des tâches de constitution de lexiques bilingues, mais en retrait sur des tâches de traduction automatique par fragments (Lardilleux *et al.*, 2009). Nous n'avions émis jusqu'alors que des *hypothèses* pour expliquer ces résultats a priori contradictoires. Dans cet article, nous proposons une analyse fine du comportement de notre méthode afin de déterminer l'origine de ces différences, ainsi qu'une généralisation destinée à améliorer ses performances en traduction automatique par fragments.

Cet article est organisé de la façon suivante : la section 2 présente une vue d'ensemble de la méthode d'alignement d'origine ; la section 3 présente des expériences mettant en évidence l'origine de ses faiblesses ; nous décrivons dans la section 4 une généralisation, et évaluons ses performances ; et la section 5 conclut ces travaux.

2 Vue d'ensemble de la méthode d'alignement d'origine

2.1 Principes de base

Notre méthode d'alignement peut être vue comme une émulation des méthodes associatives, à la différence (majeure) près qu'elle ne se restreint pas à aligner des *couples de mots*[1] (*source*, *cible*). Elle permet, en effet, de considérer des *séquences de mots* de taille variable, éventuellement discontinues, qui partagent strictement la même distribution (répartition) dans les phrases du corpus parallèle d'entrée, indépendamment de leur langue. Ces séquences constituent en fait un sous-ensemble des candidats de traduction qui obtiendraient un score maximal par des tests d'association statistiques. Le nombre de séquences de mots ayant exactement la même distribution étant réduit, nous ne recherchons pas ces séquences dans le corpus d'entrée même, mais dans des sous-corpus de celui-ci, l'idée étant que plus un sous-corpus est petit, plus les mots qu'il contient ont de chances de partager la même distribution, et que par conséquent plus le nombre de mots alignés dans ce sous-corpus est élevée.

Le cœur de la méthode consiste donc à extraire des alignements à partir de multiples sous-corpus indépendants construits par échantillonnage. En pratique, nous privilégions les sous-corpus de petite taille car ils sont plus rapides à traiter et semblent donner de meilleurs résultats (Lardilleux, 2010). Pour chaque séquence de mots de même distribution dans un sous-corpus, deux alignements sont extraits : la séquence elle-même, d'une part, et son complémentaire, d'autre part. Le nombre de sous-corpus à traiter n'étant pas défini à l'avance, le processus est *anytime*, c'est-à-dire qu'il peut être interrompu à tout moment par l'utilisateur, ou selon des critères tels que le temps écoulé ou le taux de couverture du corpus de départ. Plus le nombre de sous-corpus traités est élevé, plus la couverture du corpus de départ est grande et plus les mesures d'association sont précises. Les alignements extraits sont collectés à partir de l'ensemble des sous-corpus traités, et sont évalués par divers scores (probabilité de traduction et poids lexicaux (Koehn *et al.*, 2003)) à proportion du nombre de fois qu'ils ont été extraits. Le résultat est une table de traductions directement utilisable, par exemple, pour des tâches de traduction automatique.

2.2 Algorithme complet

L'algorithme d'extraction complet est schématisé dans le tableau 1.

La figure 1 illustre les principales étapes de l'algorithme sur un exemple d'alignement d'un texte trilingue. Dans la suite de cet article consacré aux applications de l'alignement en traduction automatique, nous nous limiterons à une application bilingue de la méthode, bien que son caractère multilingue en constitue un atout majeur.

[1] Nous employons le terme « mot » pour désigner toute forme graphique identifiée par un programme de *tokenisation*.

Entrée : un corpus multilingue, ici arabe-français-anglais.

1	قهوة ، من فضلك .	↔ Un café , s'il vous plaît .	↔ One coffee , please .
2	هذه قهوة ممتازة .	↔ Ce café est excellent .	↔ This coffee is excellent .
3	شاي ثقيل .	↔ Un thé fort .	↔ One strong tea .
4	قهوة ثقيلة .	↔ Un café fort .	↔ One strong coffee .

⇓

Transformation en corpus alingue *(= monolingue) en concaténant les traductions d'une même phrase et distinguant les mots en fonction de leur langue d'origine. Sélection d'un sous-corpus aléatoire (ici, les trois premières lignes du corpus d'origine).*

1	$_{1}$. $_{1}$فضلك $_{1}$من $_{1}$، $_{1}$قهوة	Un$_2$ café$_2$,$_2$ s'il$_2$ vous$_2$ plaît$_2$.$_2$	One$_3$ coffee$_3$,$_3$ please$_3$.$_3$
2	$_{1}$. $_{1}$ممتازة $_{1}$قهوة $_{1}$هذه	Ce$_2$ café$_2$ est$_2$ excellent$_2$.$_2$	This$_3$ coffee$_3$ is$_3$ excellent$_3$.$_3$
3	$_{1}$. $_{1}$ثقيل $_{1}$شاي	Un$_2$ thé$_2$ fort$_2$.$_2$	One$_3$ strong$_3$ tea$_3$.$_3$

⇓

Indexation des mots (calcul des vecteurs de présence). Les mots ayant même distribution sont regroupés.

	$_{1}$.	.$_2$	.$_3$	$_1$قهوة	café$_2$	coffee$_3$	One$_3$	Un$_2$	$_{1}$، $_{1}$فضلك	$_1$من	,$_3$	,$_2$	plaît$_2$	please$_3$	s'il$_2$	vous$_2$	$_1$هذه	$_1$ممتازة	Ce$_2$	This$_3$	est$_2$	excellent$_2$	…
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	…
2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	…
3	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	…

⇓

Chaque groupe de mots permet d'extraire deux alignements par phrase où il apparaît.

Les mots :	apparaissent dans les phrases :	d'où sont extraits :
$_1$قهوة café$_2$ coffee$_3$	1	$_1$قهوة café$_2$ coffee$_3$ $_{1}$. $_{1}$فضلك $_{1}$من $_{1}$، Un$_2$ _ ,$_2$ s'il$_2$ vous$_2$ plaît$_2$.$_2$ One$_3$ _ ,$_3$ please$_3$.$_3$
	2	$_1$قهوة café$_2$ coffee$_3$ $_{1}$. $_{1}$ممتازة $_1$_ $_1$هذه Ce$_2$ _ est$_2$ excellent$_2$.$_2$ This$_3$ _ is$_3$ excellent$_3$.$_3$

⋮

⇓

Décompte des alignements et rétablissement des limites entre langues.

Arabe		Français		Anglais	Décompte
قهوة	↔	café	↔	coffee	2
. فضلك من ،	↔	Un _ , s'il vous plaît .	↔	One _ , please .	1
. ممتازة _ هذه	↔	Ce _ est excellent .	↔	This _ is excellent .	1

⋮

FIG. 1 – Vue d'ensemble de la méthode d'alignement. C'est la phase d'indexation et de constitution des groupes de mots (troisième étape sur la figure) que nous généraliserons dans la suite de l'article.

Transformer le corpus parallèle d'entrée, multilingue, en corpus *alingue* (= monolingue)
Initialiser un tableau associatif *CompteurAlignements*
Faire
    Sélectionner un sous-corpus par échantillonnage
    Indexer les mots par leur vecteur de présence dans les *phrase*s du sous-corpus
    Les mots de même distribution sont rassemblés dans un même *groupe*
    Pour chaque *groupe* de mots :
        Pour chaque *phrase* où le *groupe* apparaît :
            Rétablir l'ordre des mots du *groupe*
            CompteurAlignements[groupe] ++
            CompteurAlignements[phrase - groupe] ++
Jusqu'à interruption par l'utilisateur **ou** temps imparti écoulé **ou** plus aucun alignement obtenu **ou** tout autre critère
Calculer les scores des alignements

TAB. 1 – Les étapes de la méthode d'alignement.

2.3 Résultats

Dans cette section, nous résumons les principaux résultats et conclusions de (Lardilleux, 2010). Nous avons évalué cette méthode d'alignement sur deux tâches : en traduction automatique statistique par fragments et en constitution de lexiques bilingues. L'implémentation de notre méthode, *Anymalign*[2], est comparée à MGIZA++[3] (Gao & Vogel, 2008), l'implantation la plus récente des modèles IBM. Anymalign étant *anytime*, nous commençons en pratique par exécuter MGIZA++ avec ses paramètres par défaut (5 itérations de chacun des modèles IBM1, HMM, IBM3 et IBM4), mesurons son temps d'exécution, et exécutons Anymalign pendant la même durée. Les corpus parallèles utilisés dans les expériences sont principalement Europarl (Koehn, 2005) et des extraits du BTEC (Takezawa *et al.*, 2002), distribués lors des campagnes d'évaluation de traduction automatique IWSLT (Fordyce, 2007). Les extraits du BTEC sont constitués de 20 000 à 40 000 couples de phrases courtes alignées (10 mots anglais en moyenne) et ceux d'Europarl de 100 000 couples de phrases longues (30 mots anglais).

Dans la tâche de traduction automatique statistique par fragments, nous comparons les scores obtenus par Moses (Koehn *et al.*, 2007) avec sa table de traductions par défaut, construite à partir des alignements de MGIZA++, et celle produite par Anymalign. En moyenne, Anymalign est en retrait de deux points BLEU (Papineni *et al.*, 2002) sur l'ensemble des expériences que nous avons menées. Dans le meilleur des cas, nous avons obtenu un gain d'un point par rapport à MGIZA++ (BTEC, japonais-anglais) ; dans le pire, une perte de huit points (Europarl, finnois-anglais). Dans l'ensemble, les écarts sont plus prononcés sur Europarl que sur le BTEC.

Dans la tâche de constitution de lexiques bilingues, nous comparons les tables de traductions produites par les deux aligneurs avec un lexique bilingue de référence[4]. Dans un premier temps, ce lexique est filtré de façon qu'il ne contienne que des couples de traductions qui peuvent effectivement être extraits par les aligneurs à partir du corpus parallèle d'entrée. En pratique, un couple de traductions du lexique de référence est conservé s'il s'agit d'une sous-séquence d'un couple de phrases du corpus parallèle. Nous définissons alors le score d'une table de traductions relativement à ce lexique de référence filtré comme la somme des probabilités de traduction source $\rightarrow$ cible des alignements de la table de traductions présents dans la référence, divisée par le nombre d'entrées distinctes dans la référence. Le résultat s'interprète comme un score de rappel, entre 0 et 1. En moyenne, Anymalign est meilleur de 7 % relativement à MGIZA++ sur l'ensemble des expériences que nous avons menées. Dans le meilleur des cas, nous avons obtenu un gain relatif de 70 % (Europarl, finnois-français) ; dans le pire une perte de 18 % (Europarl, suédois-finnois). Le genre de textes constituant le corpus ne semble pas avoir d'influence majeure sur ces scores.

En résumé, notre méthode est en retrait sur les tâches de traduction automatique par fragments, mais produit de meilleurs alignements de mots, comme l'attestent les résultats de comparaison avec lexiques de référence, dont les entrées sont majoritairement des mots simples (le nombre moyen de mots par entrée est 1,2). Nous avons montré (Lardilleux *et al.*, 2009) que cela est en fait principalement dû à la faible capacité de cette méthode à produire des alignements de n-grammes de mots avec $n \geqslant 2$, comme l'illustre la figure 2. Le but de la section suivante est de mettre en évidence l'origine de ces différences.

[2] `http://users.info.unicaen.fr/~alardill/anymalign`
[3] `http://geek.kyloo.net/software/doku.php/mgiza:overview`
[4] Nos lexiques proviennent principalement du site XDXF : `http://xdxf.sourceforge.net`

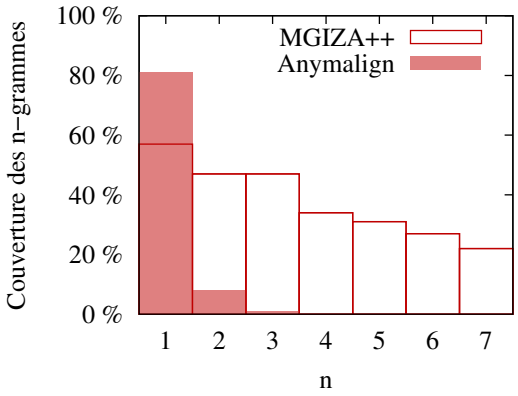


FIG. 2 – Couverture de la partie source d'un échantillon d'Europarl français-anglais par les tables de traductions de MGIZA++ et d'Anymalign. Anymalign aligne plus d'unigrammes, mais peu de n-grammes plus longs.

3 Une analyse du comportement de la méthode

Dans cette section, nous présentons des expériences montrant que deux causes principales sont à l'origine des résultats apparemment contradictoires présentés ci-dessus : les différences de fréquences des mots qui composent les séquences à aligner (cause propre à la méthode), et les fréquences de mots utiles à ces tâches (cause propre à la tâche). Les expériences présentées ici sont réalisées sur un extrait d'environ 320 000 phrases d'Europarl, avec les couples de langues portugais-espagnol (cas extrêmes de langues proches dans nos expériences) et finnois-anglais (cas extrême de langues éloignées : le finnois est une langue ouralienne agglutinante, l'anglais une langue germanique d'influence romane isolante, ce qui s'exprime par une grande différence de taille des vocabulaires). Le tableau 2 présente le nombre de mots de chaque partie de nos corpus.

Langue	Nombre de mots *(tokens)*	Taille du vocabulaire
portugais	9 249 177	87 341
espagnol	9 330 199	85 366
finnois	6 472 649	274 958
anglais	8 955 995	53 704

TAB. 2 – Caractéristiques des corpus utilisés pour nos analyses.

3.1 Différences de fréquences

Nous avons précédemment montré (Lardilleux *et al.*, 2009) qu'en pratique, la contrainte d'identité des distributions qui est au cœur de la méthode empêche d'extraire des séquences composées de mots de fréquences différentes. Par exemple, un bigramme constitué d'un mot hapax suivi du point de fin de phrase (assimilé à un mot typographique) ne peut être produit, car en supposant que le point apparaisse dans toutes les phrases du corpus d'entrée, la seule configuration dans laquelle ces deux mots partageraient la même distribution serait un sous-corpus constitué d'une seule phrase. Dans une telle configuration, presque tous les mots seraient hapax, et la séquence extraite consisterait donc en l'unique phrase de ce sous-corpus. Le bigramme attendu serait donc « masqué » et ne pourrait pas être extrait isolément.

Nous faisons un pas supplémentaire en étudiant la taille des sous-corpus d'où les mots sont extraits en fonction de la fréquence de ces mots. Étant donné un mot source m_s à aligner isolément, trois cas peuvent se produire :

1. dans un sous-corpus « trop petit », d'autres mots sources ont la même distribution que m_s. Il n'est donc pas possible d'aligner m_s isolément.
2. dans un sous-corpus de taille « idéale », aucun autre mot source n'a la même distribution que m_s, et au moins un mot cible a cette distribution. m_s peut donc être aligné isolément.

3. dans un sous-corpus « trop grand », aucun autre mot source n'a la même distribution que m_s, mais aucun mot cible non plus. m_s ne peut donc pas être aligné du tout.

Il existe ainsi une plage de tailles de sous-corpus qui permet d'extraire un mot isolément. Cette plage dépend bien entendu du mot à extraire et plus particulièrement de sa fréquence. Ces plages sont déterminées empiriquement en mesurant, pour chaque mot source d'un corpus parallèle, la taille moyenne des sous-corpus à partir de laquelle il peut être aligné isolément, ainsi que celle à partir de laquelle il ne peut plus être aligné du tout. Pour cela, nous commençons par tirer aléatoirement un sous-corpus d'une seule phrase contenant ce mot, testons si le mot peut y être aligné, puis recommençons ce test en augmentant le sous-corpus d'une nouvelle phrase tirée aléatoirement. Le processus s'arrête lorsque plus aucun mot cible n'a la même distribution que le mot source testé.

Chaque expérience produit deux nombres : la taille à partir de laquelle le mot peut être aligné isolément (passage du cas 1 au cas 2 ci-dessus), et celle à partir de laquelle le mot ne peut plus être aligné du tout (du cas 2 au cas 3). Ce test est répété 1 000 fois pour chaque mot source, et nous effectuons la moyenne des mesures recueillies sur l'ensemble des 1 000 tirages. Les résultats sont présentés à la figure 3, par classes de mots de fréquences proches.

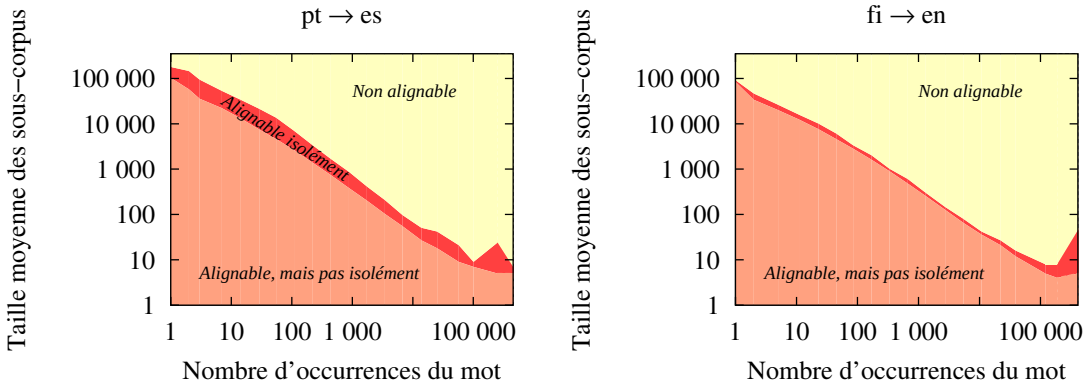


FIG. 3 – Tailles moyennes des sous-corpus à partir desquelles un mot source peut être extrait en fonction de la fréquence de ce mot. Dans la zone inférieure, le mot ne peut pas être aligné isolément (cas 1). Dans la zone du milieu, le mot peut être aligné isolément (cas 2). Dans la zone supérieure, le mot ne peut pas être aligné du tout (cas 3). Le petit sursaut de la limite supérieure à l'extrémité droite des deux graphiques est dû au point de fin de phrase, qui s'aligne plus facilement que les autres mots fréquents : il peut être aligné isolément dans des sous-corpus de 5 à 80 phrases environ.

Ces graphiques nous permettent de faire deux remarques. D'abord, la plage des tailles « idéales » des sous-corpus, autrement dit la largeur de la zone du milieu, varie grandement d'un couple de langues à l'autre. Notons que l'échelle logarithmique fait paraître cette plage plus étroite qu'elle ne l'est en réalité : le rapport moyen entre sa limite supérieure et sa limite inférieure est de 2,2 pour le couple espagnol-portugais et 1,2 pour le couple finnois-anglais. Cette différence de rapport s'explique aisément par les différences de morphologie des langues dans chacun de ces couples. Nous pouvons donc nous attendre à ce que l'alignement d'un mot donné par Anymalign nécessite le traitement de davantage de sous-corpus avec le couple finnois-anglais qu'avec le couple portugais-espagnol, puisqu'il est alors plus difficile de tirer aléatoirement un sous-corpus de la « bonne » taille.

La seconde remarque nous intéresse tout particulièrement dans le cadre de cet article : plus un mot est fréquent, plus les sous-corpus à partir desquels il est extrait sont petits, et réciproquement. Les mots rares (partie gauche des graphiques) sont donc alignés à partir de grands sous-corpus, tandis que les mots fréquents (partie droite des graphiques) sont alignés à partir de petits sous-corpus, constitués par exemple de 5 à 9 phrases pour la virgule. Ces résultats valident nos premières hypothèses : s'il est difficile de tirer un sous-corpus dans lequel deux mots source de fréquences différentes partagent la même distribution, c'est avant tout parce que ces mots ne peuvent pas être alignés à partir du même sous-corpus. Pour aligner des mots de fréquences différentes, il est nécessaire de les extraire à partir de sous-corpus de tailles différentes. Nous proposerons une alternative dans la section suivante.

3.2 Fréquences utiles

La seconde explication des différences de résultats d'Anymalign sur les deux tâches sur lesquelles il a été évalué provient en fait de la tâche elle-même, ou pour être plus précis du couple (aligneur, tâche).

Notre méthode et les modèles IBM reposent sur des intuitions opposées : la première tire parti de la rareté des mots pour les aligner (on réduit artificiellement et temporairement la fréquence de tous les mots en se plaçant dans un sous-corpus), tandis que les seconds sont estimés à partir des observations mesurées sur l'ensemble du corpus. En conséquence, Anymalign aligne mieux les mots rares, tandis que MGIZA++ aligne mieux les mots fréquents, comme l'illustre la figure 4.

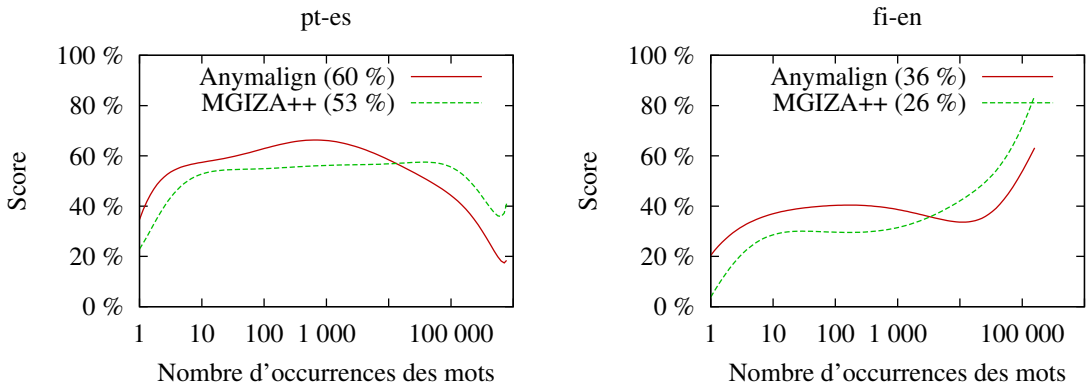


FIG. 4 – Scores obtenus par les tables de traductions produites par Anymalign et MGIZA++ sur la tâche de constitution de lexiques bilingues. Les scores entre parenthèses sont les scores globaux, calculés comme décrits au 3e paragraphe de la section 2.3. Les courbes présentent le détail de ces scores, en fonction du nombre d'occurrences du mot source de chacun des alignements : un score a été calculé localement pour chaque effectif de mot. Les courbes ont été lissées pour améliorer leur lisibilité.

Ce qui nous intéresse ici n'est pas tant l'allure générale des courbes que leur position relative : la courbe correspondant à Anymalign est au-dessus de celle de MGIZA++ pour les mots d'effectif 1 à 5 000 environ, et en-dessous pour les effectifs supérieurs. Cela montre qu'Anymalign aligne mieux non seulement les mots rares, mais également les mots de fréquence intermédiaire. Cette observation a été corroborée sur d'autres couples de langues (de-en, es-en, fr-en).

Or, les mots rares étant beaucoup plus nombreux dans tout texte — cf. loi d'Estoup-Zipf (Zipf, 1965; Mandelbrot, 1954; Montemurro, 2004) —, *a fortiori* dans notre corpus parallèle ainsi que dans les tables de traductions produites, et notre protocole d'évaluation par comparaison avec lexiques de référence traitant les mots indépendemment de leur fréquence, il est attendu que notre méthode obtienne de meilleurs scores en constitution de lexiques bilingues, puisque les mots qu'elle aligne le mieux sont au total les plus nombreux. À l'opposé, les mots fréquents sont beaucoup moins nombreux, mais autrement plus importants en traduction automatique car ils y sont beaucoup plus sollicités : un mot fréquent a plus de chances d'apparaître dans un jeu de test qu'un mot rare. Cela peut expliquer, au moins pour partie, les scores plus faibles d'Anymalign en traduction automatique. Idéalement, nous aimerions pouvoir utiliser les alignements de tel ou tel aligneur en fonction de la fréquence des mots, par exemple en combinant les tables de traductions produites par les aligneurs. Des expériences préliminaires utilisant les probabilités de traduction d'Anymalign comme fonction de trait supplémentaire dans la table de traduction par défaut de Moses ont donné des résultats prometteurs. Cela sort cependant du cadre de cet article, et nous nous consacrons par la suite à l'alignement de mots de fréquences différentes. Nous garderons néanmoins à l'esprit que, pour bien faire en traduction automatique, notre méthode devra également aligner plus efficacement les mots fréquents, ce que nous gardons pour des recherches futures.

4 Généralisation de la méthode à toutes les chaînes de mots

Dans cette section, nous présentons une généralisation de la méthode destinée à améliorer ses performances en traduction automatique statistique par fragments. En conformité avec la méthode d'origine, nous travaillerons

toujours sur les formes surfaciques des mots et sans ressource autre que le corpus d'entrée (traitement endogène). Notre but est d'extraire davantage d'alignements de n-grammes (chaînes de mots) avec $n \geqslant 2$ (cf. figure 2), tout en contournant le problème de l'extraction des mots de fréquences différentes (section 3.1).

4.1 Phase d'indexation

Nous introduisons le traitement à un grain variable en indexant des n-grammes plutôt que des mots. Nous ne chercherons pas à effectuer une segmentation particulière des phrases, par exemple en chunks, dont Vergne (2009) a montré qu'ils pouvaient être déterminés de façon endogène, mais traiterons plus simplement tous les n-grammes de mots se chevauchant. Considérons le (sous-)corpus d'entrée alingue[5] suivant, constitué de trois phrases :

1 a b c
2 a b d e
3 a c

L'indexation sur l'ensemble des n-grammes de ce corpus, *avant* recensement des groupes de même distribution servant de base à l'extraction des alignements, produit le résultat suivant :

	n = 1					n = 2					n = 3			n = 4
	a	b	c	d	e	ab	ac	bc	bd	de	abc	abd	bde	abde
1	1	1	1	0	0	1	0	1	0	0	1	0	0	0
2	1	1	0	1	1	1	0	0	1	1	0	1	1	1
3	1	0	1	0	0	0	1	0	0	0	0	0	0	0

Dans l'étape suivante, le recensement des groupes de même distribution, nous introduisons un changement majeur : si des n-grammes de même distribution se chevauchent, le groupe de mots résultant est constitué de l'union de ces n-grammes. Par exemple, les bigrammes de même distribution *bd* et *de* formeront le groupe de mots *bde*. Autrement dit, les groupes ne sont plus constitués de mots de même distribution, mais de mots issus de n-grammes de même distribution. Un même mot peut désormais apparaître dans plusieurs groupes, ce qui n'était pas le cas dans la méthode d'origine.

Ce changement soulève un problème qui ne pouvait pas se produire avec la méthode d'origine : des n-grammes peuvent masquer des (n−1)-grammes, et ce récursivement. L'unigramme *b* est par exemple masqué par le bigramme de même distribution *ab*, car l'union de *b* et *ab* donne *ab*, et *b* ne peut plus être aligné isolément. Il est donc nécessaire de traiter l'introduction de chaque longueur de n-gramme de façon spécifique.

4.2 Stratégie de constitution des groupes de mots

Nous avons testé trois stratégies :

1. traiter séparément les n-grammes en fonction de leur longueur. Ainsi, les groupes de mots ne sont construits qu'à partir de n-grammes de même longueur en source *et* en cible. Cela est bien entendu d'efficacité limitée sur des couples de langues tels que finnois-anglais : il serait préférable d'autoriser l'extraction d'un seul mot d'une langue agglutinante avec plusieurs mots d'une langue isolante.
2. permettre le mélange de toutes les longueurs de n-grammes, mais en ajoutant progressivement chaque longueur. L'ensemble initial ne contient que des unigrammes (méthode d'origine). Dans un deuxième temps, nous ajoutons les bigrammes et recréons tous les groupes de mots : certains sont identiques (les décomptes des alignements correspondants sont renforcés), d'autres sont nouveaux, d'autres enfin sont masqués mais cela n'a pas d'importance car ils ont déjà été extraits à partir des unigrammes. On ajoute ensuite les trigrammes, etc. Les alignements sont extraits à chaque fois que des n-grammes sont ajoutés.
3. forcer l'alignement de n-grammes de longueurs différentes, à contrepied de la première stratégie, en traitant séquentiellement tous les couples de longueurs (*source*, *cible*) possibles (produit cartésien). Cela permet l'alignement de n-grammes de longueurs très différentes en source et en cible, voire *trop* : puisque nous n'avons recours à aucune connaissance extérieure, Anymalign ne sait pas *a priori* quelle langue est traitée, et rien ne l'empêche par exemple de vouloir aligner des unigrammes en anglais avec de longs n-grammes en finnois, quand bien même il est peu probable que le moindre alignement puisse être produit à partir d'une telle configuration. En outre, la complexité de cette approche est bien plus importante que celle des deux précédentes, et ne passe pas à l'échelle lorsque nous traitons plus de deux langues simultanément.

[5]Comme décrit à la section 2, notre principal algorithme ne fait pas de différence entre corpus multilingues et corpus monolingues.

Pour comparer ces trois stratégies, nous préparons un ensemble de 100 000 sous-corpus aléatoires issus d'Europarl (français-anglais) et en extrayons les alignements selon chacune de ces stratégies. Nous réalisons l'expérience pour des longueurs maximales de n-grammes allant de 1 à 5. Les tables de traductions ($3 \times 5 = 15$ tables au total), obtenues à partir de ce *même* ensemble de sous-corpus, sont évaluées sur les mêmes tâches que précédemment : en traduction automatique statistique par fragments (les critères d'évaluation sont BLEU et TER (Snover *et al.*, 2006)) et en constitution de lexiques bilingues. Les résultats sont présentés dans le tableau 3.

Stratégie	*n* max.	Score en lexique (%)	BLEU (%)	TER (%)	Nombre d'entrées	Long. moy. des entrées
	1	36,19	21,12	63,57	83 967	1,92
	2	**36,71**	22,62	61,93	277 858	2,79
1.	3	36,66	23,08	62,06	366 971	3,13
	4	36,60	**23,23**	**61,43**	393 453	3,24
	5	36,58	22,92	62,14	399 810	3,27
	1	36,19	21,12	63,57	83 967	1,92
	2	37,08	23,63	60,68	290 631	2,78
2.	3	37,35	**24,72**	**59,86**	398 880	3,12
	4	37,45	24,47	60,69	436 760	3,25
	5	**37,56**	24,25	59,94	448 212	3,31
	1	**36,19**	21,12	63,57	83 967	1,92
	2	31,71	23,85	60,41	312 273	2,86
3.	3	30,90	**24,50**	60,68	453 429	3,24
	4	30,48	24,47	**59,96**	507 359	3,39
	5	30,25	24,26	60,03	524 091	3,45

TAB. 3 – Qualité et caractéristiques des tables de traductions produites selon chacune des trois stratégies de constitution de groupes de mots, pour différente longueurs maximales de n-grammes indexés. Les lignes où *n* max. = 1 sont identiques pour les trois stratégies et correspondent à la méthode d'origine.

Comme il était attendu, plus la longueur maximale des n-grammes indexés est grande, plus le nombre d'entrées dans la table de traductions et la longueur de ces entrées sont également élevés, car les alignements produits avec un *n* max. donné contiennent ceux produits avec un *n* max. inférieur (inclusion des tables). Les scores en constitution de lexiques augmentent de façon négligeable lorsque *n* max. augmente avec les deux premières approches, mais se dégradent de façon significative avec la troisième. Le gain en traduction automatique est significatif avec les trois approches. La seconde semble néanmoins fournir des résultats très légèrement meilleurs selon les trois critères d'évaluation. Son temps d'exécution est légèrement supérieur à celui de la première (au pire deux fois plus lent avec les 5-grammes), mais bien en-deçà de celui de la troisième (de l'ordre de l'heure à celui de la journée avec les 5-grammes).

La stratégie que nous utiliserons par la suite sera donc la deuxième. Elle constitue sur le fond un bon compromis entre les deux autres. La figure 5 présente le détail de la colonne « Nombre d'entrées » du tableau 3 pour cette deuxième stratégie, et est à confronter avec la figure 2.

Dans l'ensemble, l'ajout d'une longueur de n-grammes indexés, autrement dit le passage d'une courbe à celle immédiatement au-dessus, augmente considérablement la quantité de l'ensemble des n-grammes produits (y compris, de façon marginale, les n-grammes de taille inférieure, mais cela n'est dû qu'à l'extraction des complémentaires des groupes de mots). Le cas le plus significatif est celui de l'indexation des bigrammes (*n* max. = 2), qui fait exploser la quantité de bigrammes en sortie, et dans une moindre mesure de toutes les tailles de n-grammes supérieures. Le phénomène se produit également en indexant des n-grammes encore plus longs, mais cela est de moins en moins significatif à mesure que *n* max. augmente. Le graphique semble montrer qu'il n'est pas utile d'indexer des n-grammes de plus de 3 ou 4 mots, car cela se révèle peu productif. Les n-grammes qui nous intéressent le plus sont de toute façon ceux de longueur 1 à 3, parce que ce sont généralement les plus utiles en traduction automatique par fragments.

4.3 Expériences et nouveaux résultats

Nous comparons à présent notre méthode généralisée (indexation des n-grammes + constitution des groupes de mots selon la deuxième stratégie testée) à MGIZA++ sur des tâches de traduction automatique statistique par

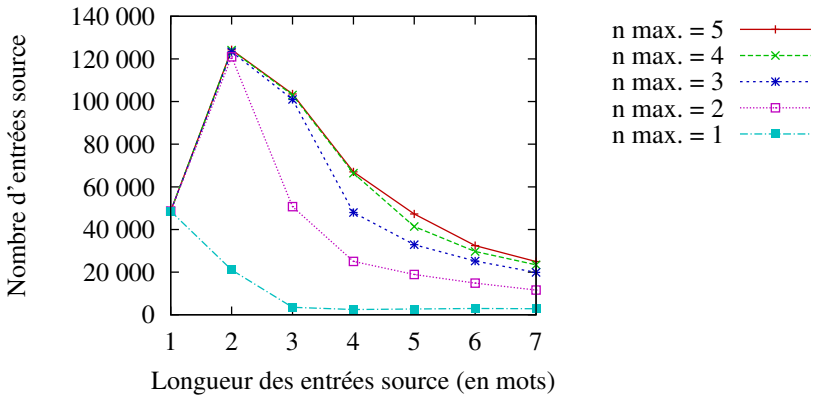


FIG. 5 – Distribution des n-grammes dans les cinq tables de traductions obtenues par la deuxième stratégie de constitution de groupes de mots. Chaque courbe correspond à une ligne du tableau 3, et la somme des ordonnées de ses points reportés est égale à la valeur indiquée dans la colonne « Nombre d'entrées » du tableau. La courbe la plus basse (*n* max. = 1) correspond à la méthode d'origine (cf. figure 2).

Tâche	Entraînement	Développement	Test	Références par phrase de test
BTEC : ar-en	19 972	1 512	489	7
BTEC : zh-en	19 972	1 512	989	7
Europarl : fi-en, fr-en, pt-es	318 804	500	1 000	1

TAB. 4 – Caractéristiques des corpus utilisés pour notre évaluation.

	Aligneur	*n* max.	BLEU (%)	TER (%)	Nombre d'entrées
	MGIZA++		*33,68*	*46,17*	*217 512*
	Anymalign	1	26,33	51,17	170 521
ar-en	-	2	30,88	49,70	269 454
	-	3	31,81	51,48	273 197
	-	4	**33,75**	**48,80**	258 141
	MGIZA++		*15,46*	*70,49*	*141 773*
	Anymalign	1	14,77	**68,97**	158 904
zh-en	-	2	16,35	71,70	263 315
	-	3	16,54	70,62	250 292
	-	4	**16,84**	69,45	269 353

TAB. 5 – Résultats des tâches de traduction sur le BTEC.

			Même temps de traitement que MGIZA++			*Temps théorique = 20 × MGIZA++*		
	Aligneur	*n* max.	BLEU (%)	TER (%)	Nombre d'entrées	BLEU (%)	TER (%)	Nombre d'entrées
	MGIZA++		*21,68*	*65,50*	*5 241 325*			
	Anymalign	1	13,73	77,57	1 871 639	13,54	74,34	5 178 683
fi-en	-	2	14,39	**76,59**	890 644	16,21	**71,18**	5 948 094
	-	3	**14,64**	77,15	696 420	**17,44**	72,63	4 001 816
	-	4	12,79	78,46	279 437	16,80	71,34	2 266 448
	MGIZA++		*29,39*	*54,37*	*10 783 083*			
	Anymalign	1	22,74	61,85	1 755 334	23,58	61,09	7 882 822
fr-en	-	2	**24,68**	60,22	1 805 297	24,55	58,42	8 317 221
	-	3	24,40	**59,77**	1 074 258	**25,29**	**57,66**	6 943 421
	-	4	23,01	61,86	492 530	24,78	58,11	5 121 617
	MGIZA++		*38,22*	*47,47*	*17 828 592*			
	Anymalign	1	34,63	50,25	1 532 520	34,84	50,35	6 730 554
pt-es	-	2	**36,03**	**49,63**	987 884	36,72	49,10	7 295 581
	-	3	35,72	49,95	744 947	35,98	49,02	6 126 896
	-	4	35,18	50,34	342 168	**37,01**	**48,71**	3 926 578

TAB. 6 – Résultats des tâches de traduction sur Europarl.

fragments. Le tableau 4 présente les caractéristiques des données utilisées pour chacune de ces expériences, et les tableaux 5 et 6 présentent les résultats.

Les lignes où *n* max. = 1 correspondent à la version d'origine d'Anymalign. Comme décrit précédemment (section 2.3), Anymalign étant *anytime*, la condition d'arrêt que nous lui imposons dépend du temps d'exécution de MGIZA++. Ce temps est constant quelle que soit la valeur de *n* max. Le temps de traitement augmentant avec ce paramètre, plus ce paramètre est élevé et plus le nombre de sous-corpus traités est *faible*, contrairement aux expériences présentées dans la section 4.2 où l'ensemble des sous-corpus à traiter était fixé à l'avance, impliquant un temps de traitement dépendant de *n* max. Théoriquement, les tables produites pour un *n* max. donné sont plus grandes que pour un *n* max. inférieur, à condition que l'aligneur soit exécuté suffisamment longtemps. Cela explique pourquoi les tables de traductions des tableaux 5 et 6 peuvent contenir moins d'entrées pour de plus grandes valeurs de *n* max. En pratique, ces tables contiennent tout de même davantage de longs n-grammes, ce qui permet une amélioration très significative des scores, malgré une table de traductions plus petite.

Sur les tâches impliquant le BTEC, les lignes où *n* max. = 1 montrent que la version d'origine d'Anymalign obtient des scores BLEU comparables à MGIZA++ en chinois-anglais, et est loin derrière en arabe-anglais. La généralisation aux n-grammes lui permet de devancer MGIZA++ de plus d'un point BLEU en chinois-anglais, et de l'égaliser en arabe-anglais, soit un gain spectaculaire de 7 points BLEU.

Sur les tâches impliquant Europarl, les scores de la version d'origine d'Anymalign sont en retrait de façon significative par rapport à MGIZA++, ce qui est conforme aux expériences que nous avions menées précédemment. Cela dit, la différence n'était pas aussi prononcée dans nos anciennes expériences : nous observions une différence de 2 à 3 points BLEU en moyenne, alors qu'elle est ici de 6 points. Nous pensons que ce changement est dû à la taille de notre corpus qui est désormais beaucoup plus élevé : 320 000 couples de phrases contre 100 000 précédemment. La taille des tables de traductions d'Anymalign, très petites par rapport à celles de MGIZA++, semble indiquer que le temps d'exécution de notre méthode n'est pas suffisant. Pour cette raison, le tableau 6 contient dans sa partie droite une deuxième série de résultats, qui correspondent à l'exécution d'Anymalign pendant une durée totale égale à 20 fois le temps d'exécution de MGIZA++. En pratique, Anymalign étant massivement parallélisable, nous avons découpé les traitements en 140 processus et les avons exécutés sur un cluster, pour finalement profiter d'un temps de traitement 7 fois plus rapide qu'avec les résultats présentés dans la partie gauche du tableau. Les tailles des tables de traductions dans la partie droite du tableau sont plus proches de celles de MGIZA++, ce qui confirme que le temps d'exécution n'était pas suffisant[6], mais le gain en BLEU de la version d'origine d'Anymalign n'est pas significatif pour autant. Il l'est par contre lorsque nous augmentons *n* max. : nous gagnons jusqu'à 3 points BLEU en finnois-anglais (*n* max. = 3) simplement en exécutant Anymalign plus longtemps. Dans tous les cas de la partie droite du tableau, l'indexation des n-grammes permet un gain en BLEU allant d'1,7 point en français-anglais à près de 4 points en finnois-anglais. En moyenne, les meilleurs scores d'Anymalign sont désormais en retrait de 3,5 points BLEU par rapport à MGIZA++, divisant pratiquement par deux son retard initial.

5 Conclusion

Cet article a présenté une généralisation de notre méthode d'alignement sous-phrastique afin d'améliorer ses résultats en traduction automatique. La méthode d'origine obtient de meilleurs résultats que l'état de l'art sur des tâches de constitution de lexiques bilingues, mais des résultats inférieurs en traduction automatique statistique par fragments. Nous avons montré que ces différences ont principalement deux causes : les différences de fréquences des mots qui composent les séquences à aligner (cause propre à la méthode), et les fréquences de mots utiles à ces tâches (cause propre à la tâche). Pour pallier le premier problème, nous avons proposé une généralisation de la phase d'indexation de notre méthode, en ne considérant non plus le mot comme unité, mais le n-gramme. Le résultat de cette généralisation est un fort accroissement du nombre de n-grammes en sortie, qui mène à des gains très significatifs en traduction automatique par fragments (jusqu'à +7 points BLEU sur le couple arabe-anglais). Notre méthode fait désormais jeu égal avec l'état de l'art sur des tâches « simples » de traduction automatique (BTEC), et nous avons pratiquement divisé son retard par deux sur des tâches plus difficiles (Europarl). Pour aller plus loin, nous envisageons d'étudier le cas de l'alignement des mots fréquents, dont nous avons montré qu'ils étaient moins bien alignés que les mots rares par notre méthode, ainsi que la question de sa condition d'arrêt.

[6]Cela soulève une autre question, qui est celle de la condition d'arrêt d'Anymalign. Les présentes expériences montrent que nos critères actuels sont insuffisants, ne serait-ce que pour effectuer une juste comparaison avec d'autres outils.

Remerciements

Les travaux présentés dans cet article ont été partiellement financés par le projet Cap Digital SAMAR.

Références

BROWN P., COCKE J., DELLA PIETRA S., DELLA PIETRA V., JELINEK F., MERCER R. & ROOSSIN P. (1988). A Statistical Approach to Language Translation. In *Proceedings of Coling'88*, p. 71–76, Budapest.

BROWN P., DELLA PIETRA S., DELLA PIETRA V. & MERCER R. (1993). The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics*, **19**(2), 263–311.

DUNNING T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

FORDYCE C. S. (2007). Overview of the IWSLT 2007 Evaluation Campaign. In *Proceedings of IWSLT 2007*, p. 1–12, Trente.

FUNG P. & CHURCH K. (1994). K-vec : A New Approach for Aligning Parallel Texts. In *Proceedings of Coling'94*, volume 2, p. 1096–1102, Kyōto.

GALE W. & CHURCH K. (1991). Identifying Word Correspondences in Parallel Texts. In *Proceedings of the fourth DARPA workshop on Speech and Natural Language*, p. 152–157, Pacific Grove.

GAO Q. & VOGEL S. (2008). Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, p. 49–57, Columbus (Ohio, USA).

KOEHN P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, p. 79–86, Phuket.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL 2007*, p. 177–180, Prague.

KOEHN P., OCH F. & MARCU D. (2003). Statistical Phrase-Based Translation. In *Proceedings of HLT-NAACL 2003*, p. 48–54, Edmonton.

LARDILLEUX A. (2010). *Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle*. PhD thesis, université de Caen Basse-Normandie. 204 pages.

LARDILLEUX A., CHEVELU J., LEPAGE Y., PUTOIS G. & GOSME J. (2009). Lexicons or phrase tables ? An investigation in sampling-based multilingual alignment. In *Proceedings of EBMT3*, p. 45–52, Dublin.

LARDILLEUX A. & LEPAGE Y. (2008). A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method. In *Proceedings of AMTA 2008*, p. 125–132, Waikiki.

LARDILLEUX A. & LEPAGE Y. (2009). Sampling-based multilingual alignment. In *Proceedings of RANLP 2009*, p. 214–218, Borovets.

MANDELBROT B. (1954). Structure formelle des textes et communication. *Word*, **10**, 1–27.

MELAMED D. (2000). Models of Translational Equivalence among Words. *Computational Linguistics*, **26**(2), 221–249.

MONTEMURRO M. (2004). A generalization of the Zipf-Mandelbrot Law in Linguistics. *Nonextensive Entropy : interdisciplinary applications*. 12 pages.

MOORE R. (2005). Association-Based Bilingual Word Alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, p. 1–8, Ann Arbor.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*, p. 311–318, Philadelphie.

SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L. & MAKHOUL J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA 2006*, p. 223–231, Cambridge.

TAKEZAWA T., SUMITA E., SUGAYA F., YAMAMOTO H. & YAMAMOTO S. (2002). Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World. In *Proceedings of LREC 2002*, p. 147–152, Las Palmas de Gran Canaria.

VERGNE J. (2009). Defining the chunk as the period of the functions length and frequency of words on the syntagmatic axis. In *Proceedings of LTC'09*, p. 85–89, Poznań.

ZIPF G. (1965). *The Psycho-Biology of Language : An Introduction to Dynamic Philology*. Classic Series. Cambridge, USA : The MIT Press. Fist edition 1935.

TALN 2011, Montpellier, 27 juin –1er juillet 2011

Estimation d'un modèle de traduction à partir d'alignements mot-à-mot non-déterministes

Nadi Tomeh Alexandre Allauzen François Yvon
Université Paris Sud et LIMSI/CNRS
BP 133 91 403 Orsay
{nadi,allauzen,yvon}@limsi.fr

Résumé. Dans les systèmes de traduction statistique à base de segments, le modèle de traduction est estimé à partir d'alignements mot-à-mot grâce à des heuristiques d'extraction et de valuation. Bien que ces alignements mot-à-mot soient construits par des modèles probabilistes, les processus d'extraction et de valuation utilisent ces modèles en faisant l'hypothèse que ces alignements sont déterministes. Dans cet article, nous proposons de lever cette hypothèse en considérant l'ensemble de la *matrice d'alignement*, d'une paire de phrases, chaque association étant valuée par sa probabilité. En comparaison avec les travaux antérieurs, nous montrons qu'en utilisant un modèle exponentiel pour estimer de manière discriminante ces probabilités, il est possible d'obtenir des améliorations significatives des performances de traduction. Ces améliorations sont mesurées à l'aide de la métrique BLEU sur la tâche de traduction de l'arabe vers l'anglais de l'évaluation *NIST MT'09*, en considérant deux types de conditions selon la taille du corpus de données parallèles utilisées.

Abstract. In extant phrase-based statistical translation systems, the translation model relies on word-to-word alignments, which serve as constraints for further heuristic extraction and scoring processes. These word alignments are infered in a probabilistic framework ; yet, only one single best word alignment is used as if alignments were deterministically produced. In this paper, we propose to take the full probabilistic alignment matrix into account, where each alignment link is scored by its probability score. By comparison with previous attempts, we show that using an exponential model to compute these probabilities is an effective way to achieve significant improvements in translation accuracy on the *NIST MT'09* Arabic to English translation task, where the accuracy is measured in terms of BLEU scores.

Mots-clés : traduction statistique, modèles de traduction à base de segments, modèles d'alignement mot-à-mot.

Keywords: statistical machine translation, phrase based translation models, word alignment models.

1 Introduction

Dans les systèmes de traduction statistique à base de segments (*phrase-based systems*), le *modèle de traduction* sert de pont entre les langues source et cible. Sur la base d'hypothèses de segmentation de la phrase source à traduire, il permet de proposer, pour chacun des segments, des traductions candidates en langue cible. Ces hypothèses de traduction sont sélectionnées dans un inventaire qui enregistre des appariements valués entre segments de longueur variable. Ces associations et les scores qui les accompagnent constituent la table de traductions (*phrase-table*).

Ce modèle est estimé en deux temps à partir d'un corpus parallèle : (i) extraction d'un ensemble de couples de segments candidats, (ii) valuation des couples retenus dans la phase (i). Faute de disposer de méthodes d'estimation théoriquement bien fondées, chacune de ces deux étapes repose sur un ensemble d'heuristiques. Il s'avère en effet impossible d'estimer directement les valuations calculées en (ii), ni même de recencer tous les appariements possibles en (i). En effet, estimer de façon non-supervisée un modèle probabiliste des alignements de segments demanderait de pouvoir calculer des sommes sur tous les alignements de segments possibles, à défaut, de savoir calculer un alignement optimal utilisant des segments de taille variable. Ces deux procédures posent des problèmes combinatoires NP-difficiles (DeNero & Klein, 2008) et ne peuvent être effectuées de manière exacte. De manière plus subtile, construire des modèles d'alignements de segments demande de mettre en compétition des segmentations conjointes de taille variable des phrases source et cible, au risque de toujours préférer les alignements impliquant des segments longs. Enfin, ne considérer qu'une seule segmentation lors de l'apprentissage semble avoir un effet négatif sur la capacité de généralisation du modèle (DeNero *et al.*, 2006).

La solution pratique qui s'est progressivement imposée contourne le problème en considérant en premier lieu une segmentation

maximale et en effectuant un alignement préalable au niveau des mots ; des procédures efficaces fondées sur l'algorithme EM (*Expectation-Maximisation*) pour effectuer cet alignement de manière efficace existent depuis le début des années 90 (Brown *et al.*, 1993; Och & Ney, 2003). Ces alignements de mots sont ensuite ré-analysés pour en déduire des alignements de segments, la méthode la plus répandue consistant à extraire les alignements de segments *compatibles* avec les contraintes posées par les alignements de mots.

Dans un troisième temps, les statistiques d'occurrence de ces alignements de segments sont collectées et utilisées pour attribuer des scores de confiance à ces groupes bilingues. Ces trois étapes successives de la construction du modèle de traduction sont usuellement abordées et optimisées séparément les unes des autres. Le risque est naturellement que les erreurs s'accumulent le long de cette séquence de traitements. Ainsi, des erreurs précoces dans les calculs des alignements mot-à-mot viennent bruiter le processus d'extraction des couples de segments appariés et biaiser les calculs de scores afférents.

Pour pallier ce problème, les auteurs de (Liu *et al.*, 2009) proposent d'extraire davantage d'informations de la phase d'alignement des mots, sous la forme d'une *matrice d'alignements pondérés*, qui représente de manière compacte un ensemble d'alignements de mots potentiels. Cette matrice est utilisée dans les étapes ultérieures de l'apprentissage. Dans une matrice pondérée, chaque lien d'alignement potentiel est nanti d'une probabilité qui mesure la confiance dans l'alignement de ces deux mots. Dans (Liu *et al.*, 2009), ces probabilités sont estimées à partir du calcul des n-meilleurs alignements de mots tels que produits par les modèles d'alignement standards. À l'aide de cette technique, ces auteurs parviennent à améliorer de façon modeste leurs systèmes de traduction automatique. Une des limites de cette approche est toutefois l'utilisation d'une liste de n-meilleurs, qui ne représente que très imparfaitement la diversité et la variabilité des alignements de mots potentiels, et conduit à des mauvais estimateurs des probabilités *a posteriori* des liens d'alignement.

Dans ce travail, nous soutenons qu'une meilleure estimation des probabilités des liens d'alignement est susceptible de donner lieu à de meilleurs modèles. Nous étudions donc une méthode alternative pour réaliser cette estimation, fondée sur des modèles discriminants pour l'alignement de mots (Ayan & Dorr, 2006; Tomeh *et al.*, 2010, 2011) et analysons les performances qu'elles permettent d'obtenir. La principale contribution de ce travail est donc de nature empirique : en comparant différentes manières de calculer et d'exploiter ces matrices d'alignement pondérées, nous montrons qu'il peut être bénéfique, en particulier quand les données d'apprentissage du modèle de traduction sont réduites, de prendre en compte l'information contenue dans ces matrices, au-delà du meilleur alignement mot-à-mot.

Cet article est organisé comme suit. Après avoir brièvement posé le cadre de la construction du modèle de traduction dans l'approche standard, nous présentons à la section 2 les principes de construction et d'exploitation de matrices d'alignements pondérées. Nous introduisons, à la section 3 une approche alternative permettant d'estimer directement la matrice d'alignement pondérée. Les résultats expérimentaux sont ensuite décrits à la section 4. Enfin, nous explicitons le positionnement de notre approche par rapport aux travaux existants, avant de conclure et d'évoquer diverses pistes vers lesquelles nous comptons nous orienter dans le futur.

2 Matrices pondérées pour la construction de modèles de traduction

Pour un système de traduction à base de segments (Zens *et al.*, 2002), le modèle de traduction est la source de connaissance principale qui établit une correspondance entre les deux langues (source et cible). Son rôle est de guider la construction, pour chaque phrase source, d'un ensemble d'hypothèses de traduction en langue cible. L'unité de traduction est le segment, qui correspond à un groupe de mots contigus. L'association entre un segment source et une traduction possible en cible forme un bi-segment. Notons qu'il est possible qu'un segment admette plusieurs traductions alternatives, donnant lieu à plusieurs bi-segments partageant le même segment source. Afin de faire un bon usage de ces bi-segments, il est nécessaire de leur associer des mesures, par exemple statistiques, qui quantifient la confiance en l'association ainsi réalisée.

Dans la suite de cet article, nous utilisons les notations suivantes : un couple de phrases est désigné par $(\mathbf{e}, \mathbf{f})$, où la phrase source $\mathbf{f} = f_1, ..., f_i, ..., f_I$ est une séquence de I mots et la phrase cible $\mathbf{e} = e_1, ..., e_j, ..., e_J$ est une séquence de J mots. De plus, une sous-séquence de mots extraite d'une phrase sera notée $f_{i_1}^{i_2} = f_{i_1} \dots f_{i_2}$ et donc $\mathbf{f} = f_1^I$.

2.1 Cadre général

Les méthodes décrites dans la littérature pour construire le modèle de traduction peuvent se résumer par l'algorithme présenté dans la partie gauche de la figure 1. Le point de départ est un couple de phrases accompagné d'un alignement mot-à-mot représenté par une *matrice d'alignement*. Chaque cellule de cette matrice booléenne $\mathbf{A} = \{a_{i,j} : 1 \leq i \leq I, 1 \leq j \leq J\}$ représente un lien

1: **POUR** toutes les paires de phrases (f_1^J, e_1^I) **FAIRE**
2: **POUR** tous les segments f_{j1}^{j2} **FAIRE**
3: Construire l'ensemble des bi-segments $E_A = \{f_{j1}^{j2}, e_{i1}^{i2}\}$ satisfaisant le jeu de contraintes $\mathcal{C}_\mathcal{A}$
4: Trier E_A selon la fonction f_R
5: Appliquer le critère de sélection $\mathcal{C}_\mathcal{S}$ définissant l'ensemble E_{AS} des bi-segments à extraire
6: Assigner une fonction de compte f_C à chaque bi-segments $(f_{j1}^{j2}, e_{i1}^{i2})$ de E_{AS}
7: **end POUR**
8: **end POUR**
9: **POUR** chaque bi-segments extraite $\{(e, f)\}$ **FAIRE**
10: Calcul des scores :

$$\phi(e|f) = \frac{f_C(e,f)}{\sum_{f_i} f_C(e,f_i)},$$

$$lex(e|f,\mathbf{A}) = \prod_{i=1}^{length(e)} \frac{1}{|\{j:(i,j)\in\mathbf{A}\}|} \sum_{\forall(i,j)\in a} w(e_i|f_j),$$

où $\mathbf{A}$ désigne la matrice d'alignement, et w une probabilité de traduction lexicale (IBM1 ou fréquence relative).
11: **end POUR**

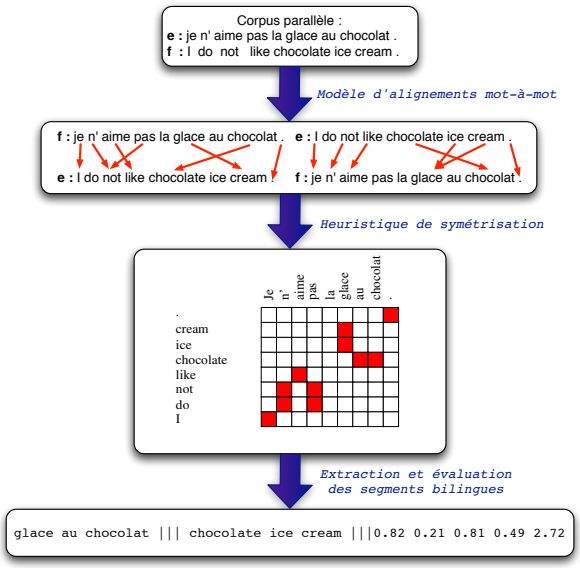


FIGURE 1 – Algorithme générique pour la construction du modèle de traduction et un exemple de son application fréquement utilisé

d'alignement potentiel ; la variable binaire $a_{i,j}$ vaut 1 si le lien entre le $i^{\text{ème}}$ mot de **f** et le $j^{\text{ème}}$ mot de **e** est actif, et 0 sinon.

Un jeu de contraintes $\mathcal{C}_\mathcal{A}$ permet de définir, parmi tous les bi-segments potentiellement contenus dans une paire de phrases, ceux qui sont « acceptables » ou cohérents avec la matrice d'alignement. Les contraintes apportées par les alignements de mots permettent l'énumération conjointe de toutes les segmentations de la paire de phrases avec tous les alignements de segments autorisés. Une fois cet ensemble de bi-segments identifié, il est possible de le trier (f_R) et de lui appliquer un critère de séléction $\mathcal{C}_\mathcal{S}$ afin d'éliminer les bi-segments qui semblent *a priori* les moins plausibles. La dernière étape concerne la valuation des bi-segments ainsi extraits. Les fonctions de valuation les plus communément utilisées sont :
– la fréquence d'observation du segment e connaissant le segment f notée $\phi(e|f)$ ainsi que le terme symétrique $\phi(f|e)$;
– les poids lexicaux ou *lexical weights* dans les deux directions ($lex(e|f,\mathbf{A})$ et $lex(f|e,\mathbf{A})$), qui utilisent, le plus souvent, les probabilités de traduction lexicale du modèle IBM1.
Ces fonctions sont définies dans l'algorithme détaillé sur la figure 1 (ligne 10).

L'instanciation standard de cet algorithme correspond aux travaux de (Zens *et al.*, 2002; Koehn *et al.*, 2003) (voir partie droite de la figure 1), qui se déduit du cadre général en utilisant les définitions suivantes :
– $\mathcal{C}_\mathcal{A}$ représente des contraintes de cohérence qui s'appliquent à un alignement mots-à-mots symétrisé d'une paire de phrases. Ces alignements se déduisent des deux meilleures hypothèses données par le modèle *IBM4* (une pour chaque direction de traduction), symétrisées par l'heuristique *grow-diag-final-and* (Koehn *et al.*, 2003).
– La fonction de compte et celle de tri sont les mêmes : $f_R = f_C = 1$
– la contrainte $\mathcal{C}_\mathcal{S}$ est définie par un seuil portant sur la longueur relative des segments source et cible et permet de filtrer les bi-segments trop longs.
Les hypothèses simplificatrices utilisées dans l'approche standard permettent d'obtenir une procédure efficace et robuste ; elles soulèvent néanmoins quelques critiques. Tout d'abord, le choix du modèle IBM4 pose problème puisque sa complexité interdit d'utiliser des procédures exactes lors de l'inférence et du calcul des probabilités *a posteriori* de chacun des liens d'alignement. Ainsi, les contraintes de cohérence des bi-segments portent sur des alignements qui ne sont pas forcément les meilleurs et pour lesquels les approximations des probabilités *a posteriori* ne reflètent qu'imparfaitement la confiance du modèle. Ce dernier point implique naturellement le choix des fonctions de compte et de tri $f_C = f_R = 1$, puisqu'en l'absence de mesure de confiance, une décision binaire s'impose. Enfin, ces simplifications entraînent que l'exploration de la matrice d'alignement est restreinte à la sous-partie sélectionnée par les alignements IBM4 et ne prend pas en considération la plus grande partie de la matrice d'alignement.

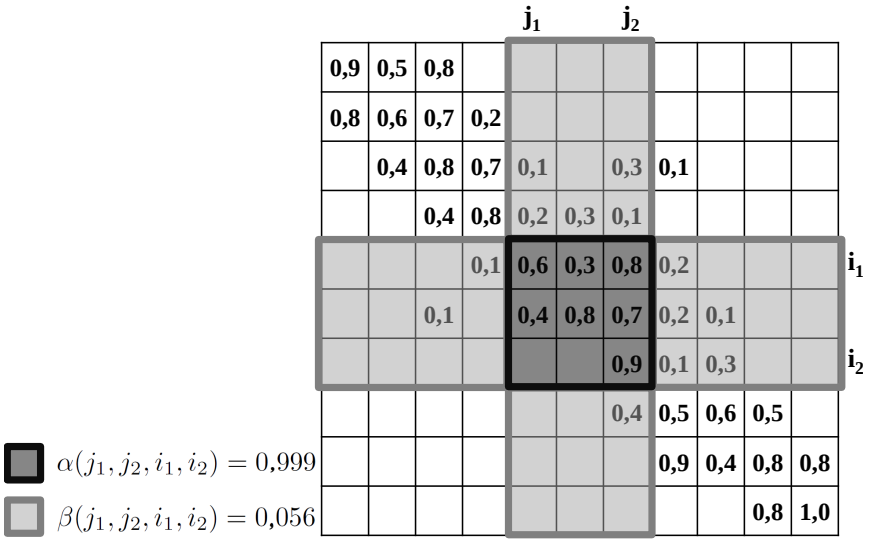


FIGURE 2 – Illustration du calcul des comptes fractionnaire pour un bi-segment donné. Dans cet exemple, le calcul des comptes fractionnaires se fait de la manière suivante : $f_C(f_{j_1}^{j_2}, e_{i_1}^{i_2}) = \alpha(j_1, j_2, i_1, i_2) \times \beta(j_1, j_2, i_1, i_2)$.

2.2 La matrice d'alignement pondérée

Dans (Liu *et al.*, 2009), les auteurs proposent d'augmenter le nombre des alignements mot-à-mot qui sont impliqués dans l'estimation des modèles de traduction et introduisent, à cet effet, la notion de *matrice d'alignement pondérée* : $\mathbf{A_p} = \{p(a_{i,j}|\mathbf{e}, \mathbf{f}) : 1 \leq i \leq I, 1 \leq j \leq J\}$. Dans cette matrice, chaque lien d'alignement est pondéré par sa probabilité *a posteriori* $p(a_{i,j}|\mathbf{e}, \mathbf{f})$. Ces probabilités sont calculées à partir des n-meilleurs alignements symétrisés proposés par le modèle IBM4. Partant de cette matrice, l'algorithme représente à la figure 1 est modifié de la manière suivante :

– Les contraintes de cohérence $\mathcal{C}_\mathcal{A}$ stipulent qu'un bi-segment est acceptable si au moins un lien d'alignement $a_{i,j}$ à l'intérieur du bi-segment est tel que $p(a_{i,j}|\mathbf{e}, \mathbf{f})$ est supérieur à un certain seuil.
– Les fonctions de compte $f_C = f_R$ prennent en compte le caractère non-déterministe des liens d'alignement de la manière suivante. Pour un bi-segment $f_C(f_{j_1}^{j_2}, e_{i_1}^{i_2})$:

$$f_C(f_{j_1}^{j_2}, e_{i_1}^{i_2}) = \alpha(j_1, j_2, i_1, i_2) \times \beta(j_1, j_2, i_1, i_2) \text{ avec} \tag{1}$$

$$\alpha(j_1, j_2, i_1, i_2) = 1 - \prod_{(j,i) \in in(j_1, j_2, i_1, i_2)} \bar{p}(a_{i,j}|\mathbf{e}, \mathbf{f}), \tag{2}$$

$$\beta(j_1, j_2, i_1, i_2) = \prod_{(j,i) \in out(j_1, j_2, i_1, i_2)} \bar{p}(a_{i,j}|\mathbf{e}, \mathbf{f}) \tag{3}$$

où $\bar{p}(a_{i,j}|\mathbf{e}, \mathbf{f}) = (1 - p(a_{i,j}|\mathbf{e}, \mathbf{f}))$, le coefficient α correspond à la confiance accordée au lien à l'intérieur (in) du bi-segment et β quantifie la masse totale de probabilité des liens situés à l'extérieur (out) de ce bi-segment. L'estimation de cette fonction est illustrée à la figure 2.

Avec ces nouvelles définitions, l'évaluation des bi-segments doit être modifiée pour également prendre en compte les probabilités des alignements. La fonction ϕ ne nécessite pas de modification, puisqu'elle utilise la fonction f_C, qui a été redéfinie. En revanche, les poids lexicaux sont maintenant définis comme suit :

$$lex(e|f, \mathbf{A_p}) = \prod_{i=1}^{|e|} \left(\left(\frac{1}{\{j|p(a_{i,j}|\mathbf{e}, \mathbf{f}) > 0\}} \sum_{\forall j: p(a_{i,j}|\mathbf{e}, \mathbf{f}) > 0} w(e_i|f_j) p(a_{i,j}|\mathbf{e}, \mathbf{f}) \right) + w(e_i|f_0) \prod_{j=1}^{|f|} \bar{p}(a_{i,j}|\mathbf{e}, \mathbf{f}) \right). \tag{4}$$

L'une des hypothèses explorée dans notre travail est que les gains modestes obtenus par (Liu *et al.*, 2009) sont dus à la méthode utilisée pour estimer cette matrice pondérée, qui s'appuie sur un petit ensemble d'alignements calculés par le modèle IBM4. En

effet l'échantillonnage des alignements en ne considérant que les n-meilleures hypothèses des modèles IBM4 ($n = 10$ en pratique) revient à considérer qu'un sous-ensemble qui ne contient que peu de variation et beaucoup de redondance. Ainsi, l'exploration de la matrice d'alignement est par construction très limitée et l'estimation approximative. Par ailleurs, le calcul de la matrice d'alignement s'appuie sur une procédure *ad hoc* de recombinaisons des probabilités *a posteriori* des alignements initialement calculés séparément pour chaque direction de traduction.

L'alternative que nous proposons d'explorer consiste à estimer cette matrice en utilisant une modélisation directe de la probabilité d'un lien d'alignement en utilisant des modèles conditionnels exponentiels qui seront décrits à la section 3.

3 Modélisation de la matrice d'alignement

Un alignement mot à mot entre une phrase source, et sa traduction (la phrase cible) regroupe un ensemble de liens décrivant une relation de traduction entre mots. Ainsi, prédire la matrice d'alignement peut être envisagé comme un problème de classification supervisée pour des données structurées. Lorsque des données étiquettées sont disponibles, la solution proposée dans (Ayan & Dorr, 2006; Tomeh *et al.*, 2010, 2011) consiste à estimer de manière indépendante la probabilité de chaque lien dans la matrice à l'aide d'un modèle de régression logistique défini par :

$$p\left(y|\boldsymbol{x}\right) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_{k=1}^{K} \lambda_k f_k\left(y, \boldsymbol{x}\right)\right), \tag{5}$$

où y désigne la variable aléatoire binaire qui indique si un lien est actif, $\boldsymbol{x}$ l'observation, $Z(\boldsymbol{x})$ le facteur de normalisation, $(f_k)_{k=1}^K$ définit un ensemble de fonctions caractéristiques, et $(\lambda_k)_{k=1}^K$ les poids associés. Dans l'équation (5), l'observation $\boldsymbol{x}$ désigne la paire de phrases augmentée de son étiquetage morphosyntaxique et des liens d'alignement produits par les modèles génératifs.

Cette formulation du problème permet de modéliser directement chaque cellule de la matrice d'alignement. Mais elle peut être également perçue comme une manière de fusionner différents alignements d'une paire de phrases. Cette approche permet donc également de remplacer l'étape heuristique de symétrisation, nommée *grow-diag-final-and* (Koehn *et al.*, 2003) dans l'approche standard, par un modèle d'apprentissage statistique pouvant prendre en compte un nombre arbitraire d'alignements en entrée.

Estimer ce modèle à partir d'exemples demande néanmoins de prendre en considération le caractère très creux de la matrice d'alignement, conséquence du fait qu'une forte majorité de liens sont inactifs. La tâche de classification considérée est donc très déséquilibrée. Afin d'éviter d'apprendre un classifieur trop biaisé en faveur de la prédiction de liens inactifs, l'ensemble des liens à étiqueter peut être au préalable réduit à un sous-ensemble de la matrice. Pour définir ce sous-ensemble, les modèles génératifs classiques sont utilisés (modèles de Markov cachés et/ou IBM4 dans les deux directions) : tout lien qui n'apparait pas dans un des alignements génératifs est considéré comme inactif ; les autres liens sont évalués par le modèle de classification. Dans ce cadre, les alignements génératifs sont utilisés pour réduire l'espace de recherche et permettent de limiter l'effet potentiellement néfaste de données déséquilibrées (Ayan & Dorr, 2006; Elming & Habash, 2007).

Ce modèle est utilisé pour estimer la matrice pondérée d'alignement $\mathbf{A_p}$ décrite à la section 2.2. Le classifieur supervisé estime donc la probabilité $p(a_{i,j}|\mathbf{e}, \mathbf{f})$ pour chaque cellule de la matrice.

Apprentissage L'estimation des paramètres du modèle (les λ_k dans l'équation (5)) est faite de manière à maximiser la vraisemblance conditionnelle régularisée à partir d'un corpus d'entraînement. La régularisation utilisée est connue sous le nom d' *elastic-net* (Zou & Hastie, 2005) et combine un terme de régularisation ℓ^1, qui aide à sélectionner les fonctions caractéristiques les plus utiles et ainsi réduire la taille du modèle, et un terme de régularisation ℓ^2, qui garantit que le Hessien de la fonction objectif n'est jamais trop proche de zéro, et permet ainsi d'éviter les problèmes d'instabilité numérique. Ce choix de régularisation nous permet d'envisager de nombreuses fonctions caractéristiques, sachant que certaines d'entre elles seront éliminées lors de l'entraînement car jugées inutiles.

Les caractéristiques Les fonctions caractéristiques utilisées pour le classifieur sont décrites en détail dans (Tomeh *et al.*, 2010) et reprennent en partie celles proposées par (Ayan & Dorr, 2006). Elles prennent en compte les multiples sources d'informations : la paire de phrases augmentée de son étiquetage morphosyntaxique et les liens d'alignement produits par les différents modèles génératifs considérés. Ainsi, pour un lien d'alignement donné, ces fonctions binaires indiquent par exemple : l'association entre les mots source/cible et de même pour les étiquettes morphosyntaxiques associées ; quel modèle génératif propose ce lien comme actif

ainsi que le nombre total de modèles génératifs proposant ce lien comme actif ; combien de liens sont proposés par les modèles génératifs dans le voisinage ; la fertilité du mot source (et resp. du mot cible) considérant l'ensemble des alignements d'entrée ; l'écart du lien à la diagonale afin de favoriser ou non les alignements monotones ; la distance du lien avec le mot aligné le plus proche (en source et en cible) afin de caractériser si ce lien est isolé des autres.

À ces caractéristiques s'ajoutent celles que nous allons décrire. Une première famille de fonctions caractéristiques décrit les mots source et cible relatifs à un lien d'alignement (*i.e* une case de la matrice) :
– Probabilité de traduction lexicale pour le couple de mots utilisé : $p(f_i|e_j)$ et $p(e_i|f_j)$ estimées par le modèle IBM1.
– La fréquence des mots source et cible ainsi que leur ratio.
– Un test sur tous les préfixes et suffixes de longueur 3.
– La similarité entre les mots source et cible calculée par la distance d'édition. Cette caractéristique tente de capturer la propension qu'ont les noms propres à être traduits de manière similaire, comme par exemple : *SdAm Hsyn* et *Saddam Hussein*.
– Un test portant sur l'égalité entre les mots source et cible.
– Un test indiquant si l'un des mots est une ponctuation associé avec un mot qui n'est pas une ponctuation.
Nous avons également définit un ensemble de fonctions caractéristiques permettant de décrire la structure de la matrice et les liens qui la composent. En plus des fonctions décrites dans (Tomeh *et al.*, 2010), nous ajoutons la fonction qui indique si un lien d'alignement implique un mot dupliqué dans l'une des phrases. Cette caractéristique permet de pallier une faible modélisation de la distorsion. Par exemple le mot arabe *fy* peut apparaître plusieurs fois dans une même phrase et être ainsi toujours aligné avec le même mot *in* en cible. Cette fonction retourne la distance du lien considéré à la diagonale.

4 Expériences

Pour évaluer les différentes approches, nous utilisons la tâche de traduction de l'arabe vers l'anglais de l'évaluation *NIST MT'09*. Nous comparons quatre méthodes d'estimation de la matrice pondérée : l'approche standard qui utilise les modèles d'alignement IBM4 et les heuristiques d'extraction et de valuation usuelles ; la méthode décrite dans le premier article sur les matrices pondérées (Liu *et al.*, 2009) ; le système *PostCAT* (Graça *et al.*, 2010) (décrit brièvement à la section 4.1) ; et l'estimation directe de la matrice via un modèle de régression logistique. Le système de traduction utilisé est MOSES(Koehn *et al.*, 2007), un outil sous licence libre.

4.1 Corpus et outils

Pour entraîner le modèle logistique, nous avons utilisé *Wapiti* (Lavergne *et al.*, 2010) [1], avec comme corpus d'apprentissage et de développement les données alignées manuellement du corpus IBMAC (Ittycheriah *et al.*, 2006), contenant respectivement 10 000 et 663 paires de phrases. Nous avons construit 2 sous-ensembles, de taille différente, de données parallèles pour entraîner le système de traduction, afin d'évaluer l'impact du volume de données disponibles sur les résultats obtenus. Ces deux corpus ont été constitués à partir des données autorisées dans la tâche *contrainte* de l'évaluation *NIST MT'09*. Elles sont toutes disponibles via le *Linguistic Data Consortium* [2].

Nous avons ainsi défini 2 tâches, l'une avec un corpus parallèle de 30 000 phrases (30k) et l'autre avec 130 000 phrases (130k). Les systèmes de traduction sont construits avec MOSES [3] en utilisant la configuration par défaut. Les paramètres de ces systèmes sont optimisés de manière usuelle avec l'outil MERT (*Minimum Error Rate Training*) avec comme données de développement le corpus *NIST MT'06* contenant 1 800 phrases arabes et 4 traductions anglaises. Les traductions produites sont évaluées avec la métrique BLEU (Papineni *et al.*, 2002) sur les données d'évaluation *NIST MT'08*, qui contiennent 1 400 phrases et 53k mots.

Pour le système *PostCAT* [4] et l'extraction des unités de traduction [5], nous avons utilisé les outils libres disponibles sur la toile. Enfin les modèles de langue cible ont été appris avec la boîte à outils *SRILM* [6] en utilisant toutes les données monolingues autorisées dans le cadre de l'évaluation *NIST MT'09* (pour plus de détails, on se reportera à (Allauzen *et al.*, 2009)).

La partie anglaise des données est pré-traitée de manière classique (les méthodes utilisées sont décrites dans (Allauzen *et al.*, 2009)).

1. http ://wapiti.limsi.fr/
2. La description complète est disponible à l'adresse http ://www.itl.nist.gov/iad/mig/tests/mt/2009/
3. http ://www.statmt.org/moses/
4. http://www.seas.upenn.edu/~strctlrn/CAT/CAT.html
5. http://www.nlp.org.cn/~liuyang/wam/wam.html.
6. http://www-speech.sri.com/projects/srilm/.

Pour la partie arabe, toutes les phrases sont analysées et segmentées avec l'outil MADA+TOKAN [7]. Nous avons utilisé le schéma de segmentation D2 afin de tenir compte de la morphologie riche de l'arabe et ainsi segmenter les mots arabes en des unités qui correspondent approximativement aux mots anglais.

4.2 Construction des modèles de traduction

Dans la section 2, nous avons décrit un algorithme générique pour la construction d'un modèle de traduction. Cet algorithme fonctionne en trois étapes séparées : construction des matrices d'alignement pondérées, extraction puis évaluation des bi-segments. Nous allons maintenant évaluer l'impact de ces trois étapes sur les résultats en traduction automatique.

Pour la première étape, nous expérimentons deux manières de construire les matrices pondérées :
(i) la méthode standard qui ne considère que les meilleurs alignements
(ii) la matrice pondérée par les probabilités qui est utilisée dans le processus d'extraction et de valuation.
Notons qu'il est possible de passer de la configuration (ii) à (i) par un simple seuillage sur les probabilités. Dans toutes nos expériences, nous utilisons un seuil de $0,5$. Ainsi, pour chaque modèle d'alignement, deux types de systèmes sont construits : *standard* (configuration (i)) et *WAM* pour la matrice pondérée (configuration (ii)). Le corpus de référence *IBMAC* contient également un jeu de test qui est utilisé pour calculer le taux d'erreur d'alignement (ou AER, pour *Alignment Error Rate*).

Les deux autres étapes (extraction et valuation des bi-segments) dépendent du mode de construction de la matrice d'alignement. Dans le cas *standard*, les bi-segments sont extraits et évalués selon les heuristiques décrites à la section 2.1. Lorsque l'on utilise des matrices pondérées, nous utilisons les méthodes d'extraction et de valuation décrites à la section 2.2, qui prennent en compte la probabilité des liens d'alignement. Pour cette dernière approche, seuls les bi-segments dont la probabilité est supérieure à un seuil sont conservés. Ceci permet, comme le préconisent les auteurs de (Liu *et al.*, 2009), de restreindre le nombre de bi-segments qui sont extraits. De plus, comme cela est fait dans l'approche standard, les bi-segments comprenant un segment de longueur supérieur à 7 sont également rejetés. Comme il est d'usage, les performances en traduction automatique sont évaluées par la métrique BLEU (Papineni *et al.*, 2002).

4.3 Modèles d'alignement mot-à-mot

En plus des deux méthodes de construction du modèle de traduction, nous avons également considéré plusieurs modèles d'alignement mot-à-mot, que nous allons décrire brièvement.

MGIZA++ [8] propose une implémentation efficace et parallèle (Gao & Vogel, 2008) des modèles génératifs les plus utilisés : les modèles IBM1 à IBM4 (Brown *et al.*, 1993) et HMM (Vogel *et al.*, 1996). Ces modèles sont utilisés par la suite pour construire des modèles de traduction selon la configuration *standard* et pour entraîner notre système d'alignement discriminant (voir section 3).

N-best WAM construit la matrice pondérée en effectuant une moyenne des occurences des liens d'alignement à partir des n-meilleures séquences d'alignement produites par le modèle IBM4. Cette méthode correspond à l'article original sur les matrices pondérées (Liu *et al.*, 2009). Comme ces auteurs, nous avons utilisé la valeur $n = 10$.

PostCAT (Posterior Constrained Alignment Toolkit) propose une implémentation des modèles HMM permettant d'injecter des contraintes lors de l'apprentissage via l'algorithme EM. Ces contraintes portent sur les probabilités *a posteriori* des variables latentes (Graça *et al.*, 2010) et permettent de corréler les deux directions d'alignement. Deux types de contraintes simples (*symmétrie* et *bijectivité*) permettent au modèle HMM d'atteindre des performances comparables au modèle IBM4. Le fait d'utiliser des modèles HMM permet de pouvoir calculer de manière exacte et efficace les probabilités *a posteriori* et ainsi construire la matrice pondérée en considérant l'ensemble des liens d'alignement. Dans cet article, nous avons utilisé la boite à outils Geppetto [9] (Ling *et al.*, 2010), une implémentation de PostCAT et des matrices d'alignement pondérées.

7. http ://www1.ccls.columbia.edu/ cadim/MADA.html
8. http ://geek.kyloo.net/
9. http ://code.google.com/p/geppetto/

MaxEntWA est le système décrit à la section 3. Il s'agit d'un classifieur *MaxEnt* qui prédit pour chaque lien de la matrice sa probabilité *a posteriori*.

Exception faite du modèle noté *MGIZA++*, il est possible pour tous les modèles d'extraire et de valuer les bi-segments selon les deux méthodes. Pour appliquer la méthode (i), nous avons appliqué pour toutes les expériences un seuil de $0,1$ comme les auteurs de (Liu *et al.*, 2009).

4.4 Résultats

Les résultats expérimentaux pour les différentes configurations et les différents modèles d'alignement sont rassemblés dans le tableau 1. Examinons pour commencer, la partie *30k* du tableau qui correspond aux expériences où MOSES a été entraîné sur un corpus de 30 000 phrases parallèles. La partie *MGIZA++* présente les résultats obtenus en utilisant l'approche standard : utilisation des meilleures hypothèses d'alignement IBM4 symmetrisés pour extraire et valuer les bi-segments via les heuristiques usuelles (Koehn *et al.*, 2003). Ainsi sur la tâche 30k, le système standard obtient un score BLEU de 35,9. La partie 10-*best WAM* correspond au matrice pondérée où les probabibilités *a posteriori* sont estimées à partir des 10 meilleurs alignements de IBM4. Cette approche permet d'obtenir un faible gain de 0,3 points BLEU par rapport à l'approche standard, soit (36,2). Ce résultat est cohérent avec ceux publiés dans (Liu *et al.*, 2009).

La partie *PostCAT* introduit par rapport aux travaux de (Liu *et al.*, 2009) l'utilisation des modèles HMM pour les alignements de mot et donc la possibilité d'estimer les probabilités *a posteriori* de manière exacte pour l'ensemble de la matrice. Cet apport permet d'augmenter le BLEU de manière significative : de 35,9 à 36,9 ou 37,0 selon la variante du modèle HMM utilisée. Enfin la partie *MaxEntWA* présente les résultats obtenus en utilisant un modèle exponentiel pour prédire la matrice d'alignement. Les résultats montrent un gain en BLEU supplémentaire et conséquent : 1,5 points par rapport à l'approche standard et 0,5 points par rapport à l'approche *PostCAT*. Notons également, que même si les méthodes standard d'extraction et de valuation sont utilisées, les matrices d'alignements engendrées par *PostCAT* et *MaxEntWA* permettent d'obtenir de meilleurs résultats et que *MaxEntWA* est à nouveau la méthode donnant le meilleur résultat.

Sur la tâche *130k* (MOSES est entrainé sur 130 000 phrases parallèles), nous observons les mêmes tendances, avec cependant des gains en BLEU moindres. Notons que le gain modeste obtenu avec la méthode 10-*best* pour estimer la matrice pondérée est similaire à celui obtenu sur la tâche *30k*. Pour les autres méthodes de calcul de la matrice pondérée, les gains restent significatifs, bien que moins importants. De nouveau, nous pouvons observer que le calcul de la matrice d'alignement avec le modèle de régression logisitique (*MaxEntWA*) permet d'obtenir de meilleurs résultats en termes de score BLEU.

La colonne *PT* du tableau 1 indique la taille du modèle de traduction en nombre de bi-segments extraits. Nous observons, tout naturellement, que quand on considère l'intégralité de la matrice pondérée (*PostCAT* et *MaxEntWA*), la taille du modèle de traduction augmente considérablement, puisqu'elle se trouve multipliée par plus de 4, alors même que le seuil de filtrage est resté constant à 0,1. Le risque était, en multipliant les entrées du modèle de traduction, d'ajouter un bruit pouvant affecter le comportement global du système. Toutefois, il apparaît que la valuation des bi-segments par les probabilités (voir la section 2.2) est un moyen effectif pour filtrer les bi-segments les moins utiles lors de l'étape de traduction.

Ainsi, l'amélioration de la valuation des bi-segments a un impact significatif sur les résultats en BLEU. Si cette amélioration peut être imputée en partie à l'utilisation des matrice pondérée, la colonne *AER (Alignment Error Rate)* montre que cette amélioration peut provenir également d'alignements mot-à-mot de meilleure qualité. Partant d'un l'*AER* obtenu avec les modèles IBM4 symétrisés d'une valeur de 25,0%, on note tout d'abord que l'usage des 10-meilleurs alignements ne permet pas d'améliorer la qualité intrinsèque des alignements. En revanche, l'utilisation d'un modèle plus approprié tel que *PostCAT* entraîne une amélioration sensible des alignements, avec un *AER* de 22,5%. Cette tendance est encore plus affirmée avec la méthode *MaxEntWA*, qui introduit dans le processus des alignements de qualité nettement accrue, puisque la réduction absolue de l'*AER* est de plus de 10 points.

Globalement, les résultats expérimentaux montrent que l'utilisation de la matrice pondérée pour extraire et valuer les bi-segments permet d'améliorer les performances des systèmes de traduction, quand cette méthode est associée à un mode de calcul pertinent pour les valuations de la matrice pondérée. Ce dernier point recouvre d'une part la manière dont sont calculées les probabilités d'alignement, et d'autre part la fraction de cette matrice qui est effectivement explorée. La différence de résultats entre les deux tâches (*30k* et *130k*) suggère que l'utilisation d'un modèle de régression logistique pour estimer la matrice pondérée conduit à des gains bien plus importants sur la petite tâche (*30k*). Une explication de cette différence est que cette approche permet, lorsque l'on dispose de peu de données parallèles, d'extraire plus de bi-segments : lorsque les données manquent pour estimer le modèle de traduction, il est en effet important de pouvoir malgré tout engendrer un grand nombre de bi-segments potentiels. De surcroît, on note que la valuation par des probabilités permet effectivement de limiter, au moment du décodage, les effets de l'introduction d'entrées bruitées dans la table de traduction.

Tâche de traduction :		**30K**					**130K**				
Construction du MT :			**Standard(i)**		**WAM(ii)**			**Standard(i)**		**WAM(ii)**	
Alignement		**AER**	**BLEU**	**PT**	**BLEU**	**PT**	**AER**	**BLEU**	**PT**	**BLEU**	**PT**
MGIZA++	HMM	28,4	35,0	3,6	-	-	26,8	39,2	9,7	-	-
	IBM4	25,0	35,9	2,4	-	-	23,3	40,2	6,5	-	-
10-**best**	IBM4	24,9	35,8	2,4	36,2	3,0	23,3	40,0	6,6	40,4	8,5
PostCAT	Bijective	22,5	36,6	3,3	36,9	10,2	20,5	40,1	9,1	40,6	29,5
	Symmetric	22,5	36,7	2,9	37,0	10,7	20,8	40,2	8,5	40,4	30,2
MaxEntWA	HMM	17,6	36,9	6,7	37,5	11,7	16,4	40,5	17,7	40,8	30,0
	IBM4	15,6	37,2	5,5	37,5	9,6	14,3	41,0	14,5	41,1	25,0
	HMM+IBM 1,3,4	14,7	37,1	5,2	37,9	8,6	13,9	40,8	13,4	41,1	22,2

TABLE 1 – Comparaison de 4 modèles d'alignement (MGIZA++, 10-best, PostCAT and MaxEntWA) et de leurs interactions avec la méthode d'extraction et de valuation de la table de traduction en termes de taux d'erreur d'alignement (*AER*), de score BLEU et de la taille de la table de traduction exprimée en millions de bi-segments (*PT*). Les deux méthodes de construction du modèle de traduction (*MT*) sont l'approche standard (*standard*) et celle utilisant les matrices pondérées (*WAM*). Deux tailles de données parallèles d'apprentissage sont considérées (*30K* et *130K*).

5 Discussion

De nombreux travaux récents se sont intéressés aux méthodes d'extraction d'unités de traduction à partir de corpus parallèles. Que ce soit dans le cadre des systèmes hiérarchiques ou à base de segments, le processus d'extraction (Koehn *et al.*, 2003; Chiang, 2007) repose sur les matrices d'alignement mot-à-mot construites à partir des modèles d'alignement IBM4 (Brown *et al.*, 1993) symétrisés. Comme nous l'avons évoqué à la section 2.1, ce choix de la première étape se justifie par un souci d'efficacité puisqu'il restreint considérablement l'espace des unités qui sont explorées, puis sélectionnées. Néanmoins, ce choix favorise la propagation d'erreurs dues à des décisions (d'accepter ou de rejeter des liens d'alignement) qui sont prises trop tôt dans le processus, sans qu'il soit de surcroit possible d'affecter de réels scores de confiance à ces décisions.

Lorsqu'il s'agit d'étendre l'espace des unités qui sont explorées, la première difficulté est la complexité qui résulte de l'énumération puis de la valuation de toutes les unités de traduction possible. Ainsi, une partie des travaux récents s'intéresse à l'élaboration d'une représentation efficace. Dans (Mi & Huang, 2008), le processus d'extraction des règles pour un système hiérarchique est étendu en considérant l'ensemble composé des n-meilleurs arbres d'analyse syntaxique au lieu de tenir compte uniquement du meilleur. Afin de représenter puis de manipuler efficacement ces n-meilleurs arbres, les auteurs utilisent une représentation efficace (*packed forest*) (Billot & Lang, 1989) ayant également démontré son utilité (Galley *et al.*, 2006; Wang *et al.*, 2007) en traduction automatique.

De manière similaire, les n-meilleurs alignements peuvent être utilisés afin d'enrichir la matrice d'alignement, que ce soit pour extraire les bi-segments (Xue *et al.*, 2006), ou les règles d'un système hiérarchique (Venugopal *et al.*, 2008). Dans ce dernier article comme dans (Mi & Huang, 2008), les auteurs définissent une distribution de probabilité sur les alignements à partir des n-meilleurs alignements et des n-meilleurs arbres d'analyse syntaxique. Cette approche par échantillonage permet aux auteurs d'introduire des comptes fractionnaires pour les règles extraites et ainsi de pouvoir estimer le modèle de traduction.

Ce recours à l'échantillonnage pour l'inférence des probabilités *a posteriori* des d'alignement se justifie par la complexité d'inférence du modèle IBM4. Il existe en revanche, pour les modèles plus simples, tels que ceux qui s'inspirent des modèles de Markov cachés (souvent désignés de manière générique sous le nom de « modèle *HMM* ») (Vogel *et al.*, 1996) ou pour le modèle IBM1 (Brown *et al.*, 1993), des algorithmes d'inférence exacts et efficaces (Venugopal *et al.*, 2003; Deng & Byrne, 2005). Une des limitations du modèle HMM est son absence de modélisation de la fertilité. Pour pallier cette limitation, les auteurs de (Deng & Byrne, 2005) définissent un HMM permettant d'aligner des mots avec des segments qui rivalise en termes de performances avec le modèleIBM4, tout en préservant la possibilité d'un calcul exact des probabilités *a posteriori* des alignements de mots et qui s'étend au calcul de distributions *a posteriori* des segments ou des règles. Les expériences montrent que cette approche améliore significativement le processus d'extraction d'unités de traductions pour les systèmes à base de segments (Deng & Byrne, 2005) et hiérarchiques (de Gispert *et al.*, 2010).

L'introduction des matrices pondérées (Liu *et al.*, 2009) que nous décrivons à la section 2 peut être considérée comme l'adaptation

des *packed forests* des systèmes hiérarchiques au systèmes à base de segments : une exploration plus exhaustive de la matrice d'alignement, l'usage des probabilités des alignements de mots pour dériver des scores de confiance sur les bi-segments extraits. Pour ce dernier point, les auteurs s'inspirent d'ailleurs des travaux de (Mi & Huang, 2008).

Comme mentionné à la section 2, un des problème des matrices pondérées est l'estimation des probabilités *a posteriori* des alignements. Dans (Liu *et al.*, 2009), cette estimation est faite en échantillonnant les n-meilleurs alignements des modèles IBM4, alors que dans (Deng & Byrne, 2005; de Gispert *et al.*, 2010; Ling *et al.*, 2010) le modèle HMM ou une de ses variante est utilisé pour les estimer de manière exacte. Cependant, dans ce dernier type d'approche, il est encore nécessaire de fusionner les alignements correspondant aux deux directions (un modèle d'alignement de source vers cible et réciproquement). Les solutions envisagées semblent peu satisfaisantes : soit la fusion est heuristique et consiste simplement à prendre la moyenne arithmétique des distributions *a posteriori* (Graça *et al.*, 2010; Ling *et al.*, 2010) ; soit de manière beaucoup plus coûteuse, deux systèmes de traduction indépendants sont utilisés utilisant chaque modèle HMM, la fusion se fait alors sur les treillis engendrés par chaque système (de Gispert *et al.*, 2010).

Dans cet article, nous introduisons donc une extension du travail de (Liu *et al.*, 2009) en proposant une nouvelle méthode de construction de la matrice d'alignement. Pour cela, nous proposons d'utiliser un classifieur au maximum d'entropie décrit dans (Ayan & Dorr, 2006; Tomeh *et al.*, 2010, 2011). Cette approche permet en effet de calculer directement la matrice pondérée sans avoir recours ni à une fusion heuristique des distributions *a posteriori*, ni à une coûteuse étape de fusion de système. Faute de données étiquettées permettant de mettre en œuvre cette démarche, l'approche de (Graça *et al.*, 2010) semble fournir des performances proches de nos meilleurs résultats.

6 Conclusion

Dans cet article, nous avons abordé le problème de l'estimation des modèles de traduction à partir d'alignements mot-à-mot non-déterministes. En effet, dans l'approche considérée comme standard, les modèles de traduction sont estimés à partir d'alignements mot-à-mot grâce à des heuristiques d'extraction et de valuation. Bien que ces alignements mot-à-mot soient construits par des modèles probabilistes, les processus d'extraction et de valuation utilisent ces modèles comme produisant des alignements déterministes. À la suite (Liu *et al.*, 2009), la solution que nous avons envisagée lève cette limitation en considérant une matrice d'alignement pondérée, dans laquelle chaque lien d'alignement est valué par sa probabilité. Les premiers travaux dans cette direction étaient, selon nos hypothèses, limités par la méthode d'estimation de la matrice pondérée, et nous avons proposé une méthode permettant d'estimer directement cette matrice à l'aide d'une méthode de classification supervisée.

Afin de valider cette approche, nous avons effectué des expériences sur la tâche de traduction automatique de l'Arabe vers l'Anglais de l'évaluation *NIST MT'09*. Dans ce cadre expérimental, nous avons comparé 4 méthodes de construction du modèle de traduction, contrastant ainsi l'approche standard avec l'usage des matrices pondérées, et évaluant différents estimateurs de cette matrice. Les résultats ont montré que l'usage des matrices pondérées impliquait une extraction plus importante de bi-segments et que leur valuation adaptée permettait au système de traduction d'obtenir de meilleurs résultats mesurés en terme de BLEU. En particulier, des gains significatifs (entre 2 et 0,9 point BLEU, selon la tâche considérée) ont été obtenus par notre méthode, qui semble la mieux à même de produire des alignements de bonne qualité (au sens de l'*AER*). Ces résultats nous ont permis de conclure que le choix de l'estimateur des matrices pondérés a un impact net sur les performances en traduction et que notre méthode est nettement plus pertinente que celles proposées dans les travaux antérieurs.

Contrairement aux heuristiques standard, notre méthode permet de contrôler et d'adapter le nombre de bi-segments extraits à la taille des données parallèles d'entraînement. Nous souhaitons donc à l'avenir explorer cet aspect. L'approche envisagée consiste à extraire le plus de bi-segments possibles et à travailler sur leur filtrage. L'intérêt de cette approche est que nous pensons ainsi limiter l'impact des erreurs commises par les modèles d'alignement. De plus, l'étape de filtrage peut se faire en prenant en compte l'utilité des bi-segments lors de l'étape de traduction et ainsi ne pas se limiter à des tests statistiques qui ne prennent pas en compte la finalité des modèles de traduction. Des articles récents comme (Wuebker *et al.*, 2010) montrent l'importance d'une valuation des bi-segments qui améliorerait les simples calculs de fréquences, et qui serait plus directement en rapport avec la finalité des modèles de traduction.

Remerciements

Ces travaux ont été en partie financé par l'agence OSEO dans le cadre du programme Quaero. Les auteurs tiennent à remercier Thomas Lavergne pour son aide précieuse concernant la mise en œuvre de *Wapiti*.

Références

ALLAUZEN A., CREGO J., MAX A. & YVON F. (2009). LIMSI's statistical translation systems for WMT'09. In *Proc. of the 4th Workshop on Statistical Machine Translation*, p. 100–104, Athens, Greece : Association for Computational Linguistics.

AYAN N. F. & DORR B. J. (2006). A maximum entropy approach to combining word alignments. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, p. 96–103 : Association for Computational Linguistics.

BILLOT S. & LANG B. (1989). The structure of shared forests in ambiguous parsing. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, ACL '89, p. 143–151.

BROWN P. F., PIETRA V. J. D., PIETRA S. A. D. & MERCER R. L. (1993). The mathematics of statistical machine translation : parameter estimation. *Comput. Linguist.*, **19**, 263–311.

CHIANG D. (2007). Hierarchical phrase-based translation. *Comput. Linguist.*, **33**(2), 201–228.

DE GISPERT A., PINO J. & BYRNE W. (2010). Hierarchical phrase-based translation grammars extracted from alignment posterior probabilities. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, p. 545–554, Morristown, NJ, USA : Association for Computational Linguistics.

DENERO J., GILLICK D., ZHANG J. & KLEIN D. (2006). Why generative phrase models underperform surface heuristics. In *Proceedings on the Workshop on Statistical Machine Translation*, p. 31–38, New York City : Association for Computational Linguistics.

DENERO J. & KLEIN D. (2008). The complexity of phrase alignment problems. In *Proceedings of ACL-08 : HLT, Short Papers*, p. 25–28, Columbus, Ohio : Association for Computational Linguistics.

DENG Y. & BYRNE W. (2005). Hmm word and phrase alignment for statistical machine translation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, p. 169–176, Morristown, NJ, USA : Association for Computational Linguistics.

ELMING J. & HABASH N. (2007). Combination of statistical word alignments based on multiple preprocessing schemes. In *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Companion Volume, Short Papers*, NAACL-Short '07, p. 25–28, Stroudsburg, PA, USA : Association for Computational Linguistics.

GALLEY M., GRAEHL J., KNIGHT K., MARCU D., DENEEFE S., WANG W. & THAYER I. (2006). Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, p. 961–968, Sydney, Australia : Association for Computational Linguistics.

GAO Q. & VOGEL S. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, p. 49–57, Stroudsburg, PA, USA : Association for Computational Linguistics.

GRAÇA J. A. V., GANCHEV K. & TASKAR B. (2010). Learning tractable word alignment models with complex constraints. *Comput. Linguist.*, **36**, 481–504.

ITTYCHERIAH A., AL-ONAIZAN Y. & ROUKOS S. (2006). *The IBM Arabic-English Word Alignment Corpus*. Rapport interne RC24024, IBM.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, p. 177–180, Prague, Czech Republic : Association for Computational Linguistics.

KOEHN P., OCH F. J. & MARCU D. (2003). Statistical phrase-based translation. In *NAACL '03 : Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, p. 48–54 : Association for Computational Linguistics.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

LING W., LUÍS T., GRAÇA J., COHEUR L. & TRANCOSO I. (2010). Towards a General and Extensible Phrase-Extraction Algorithm. In M. FEDERICO, I. LANE, M. PAUL & F. YVON, Eds., *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, p. 313–320.

LIU Y., XIA T., XIAO X. & LIU Q. (2009). Weighted alignment matrices for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 2 - Volume 2*, EMNLP '09, p. 1017–1026, Morristown, NJ, USA : Association for Computational Linguistics.

MI H. & HUANG L. (2008). Forest-based translation rule extraction. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, p. 206–214, Honolulu, Hawaii : Association for Computational Linguistics.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, **29**, 19–51.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 311–318, Stroudsburg, PA, USA : Association for Computational Linguistics.

TOMEH N., ALLAUZEN A., WISNIEWSKI G. & YVON F. (2010). Refining word alignment with discriminative training. In *Proceedings of the ninth Conference of the Association for Machine Translation in the America (AMTA)*, Denver, CO.

TOMEH N., LAVERGNE T., ALLAUZEN A. & YVON F. (2011). Designing an improved discriminative word aligner. In *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, Tokyo, Japan.

VENUGOPAL A., VOGEL S. & WAIBEL A. (2003). Effective phrase translation extraction from alignment models. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, p. 319–326, Stroudsburg, PA, USA : Association for Computational Linguistics.

VENUGOPAL A., ZOLLMANN A., SMITH N. A. & VOGEL S. (2008). Wider pipelines : N-best alignments and parses in MT training. In *Proceedings of the Association for Machine Translation in the Americas (AMTA)*.

VOGEL S., NEY H. & TILLMANN C. (1996). Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, p. 836–841 : Association for Computational Linguistics.

WANG W., KNIGHT K. & MARCU D. (2007). Binarizing syntax trees to improve syntax-based machine translation accuracy. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 746–754, Prague, Czech Republic : Association for Computational Linguistics.

WUEBKER J., MAUSER A. & NEY H. (2010). Training phrase translation models with leaving-one-out. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 475–484, Uppsala, Sweden : Association for Computational Linguistics.

XUE Y.-Z., LI S., ZHAO T., YANG M. & LI J. (2006). Bilingual phrase extraction from n-best alignments. In *ICICIC (3)*, p. 410–414.

ZENS R., OCH F. J. & NEY H. (2002). Phrase-based statistical machine translation. In *KI '02 : Proceedings of the 25th Annual German Conference on AI*, p. 18–32, London, UK : Springer-Verlag.

ZOU H. & HASTIE T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, **67**, 301–320.

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Combinaison d'informations pour l'alignement monolingue

Houda Bouamor Aurélien Max Anne Vilnat
LIMSI-CNRS, Univ. Paris-Sud
Orsay, F-91403, France
{prénom.nom}@limsi.fr

Résumé. Dans cet article, nous décrivons une nouvelle méthode d'alignement automatique de paraphrases d'énoncés. Nous utilisons des méthodes développées précédemment afin de produire différentes approches hybrides (hybridations). Ces différentes méthodes permettent d'acquérir des équivalences textuelles à partir d'un corpus monolingue parallèle. L'hybridation combine des informations obtenues par diverses techniques : alignements statistiques, approche symbolique, fusion d'arbres syntaxiques et alignement basé sur des distances d'édition. Nous avons évalué l'ensemble de ces résultats et nous constatons une amélioration sur l'acquisition de paraphrases sous-phrastiques.

Abstract. In this paper, we detail a new method to automatic alignment of paraphrase of statements. We also use previously developed methods to produce different hybrid approaches. These methods allow the acquisition of textual equivalence from a parallel monolingual corpus. Hybridization combines information obtained by using advanced statistical alignments, symbolic approach, syntax tree based alignment and edit distances technique. We evaluated all these results and we see an improvement on the acquisition of sub-sentential paraphrases.

Mots-clés : Paraphrase sous-phrastique, corpus parallèle monolingue, hybridation.

Keywords: Phrasal paraphrase, monolingual parallel corpora, hybridization.

1 Introduction

Le traitement de corpus monolingues et multilingues constitue un champ d'investigation très animé dans le domaine du traitement automatique des langues. Ils sont souvent constitués d'unités de texte ayant des liens sémantiques forts, une information qui peut être exploitée pour acquérir des équivalences entre des mots ou des groupes de mots et construire des ressources linguistiques importantes pour diverses applications. Ces resssources peuvent être utilisées par la suite pour extraire des réponses à des questions (Duclaye *et al.*, 2003), par exemple, ou autoriser des formulations différentes en évaluation de la traduction automatique (Russo-Lassner .G & .P, 2005; Kauchak & Barzilay, 2006), ainsi qu'en génération, pour aider des auteurs à trouver des formulations plus adaptées (Max, 2008).

De nombreuses techniques ont été proposées pour l'acquisition de segments en relation de paraphrase. Ces techniques ont en commun d'être directement liées aux types de ressources sur lesquelles elles s'appliquent. Les plus nombreuses exploitent des corpus monolingues comparables disponibles en grandes quantités, et se fondent sur l'hypothèse que des unités linguistiques apparaissant de nombreuses fois dans des contextes similaires peuvent avoir la même signification. Restreindre les corpus utilisés à des textes comparables, sélectionnés sur la base d'un genre ou de thèmes communs, permet d'augmenter la probabilité que les correspondances obtenues seront effectivement valides grâce aux contextes plus restreints.

Peu de travaux ont, en comparaison, porté sur l'exploitation de corpus monolingues parallèles, constitués de phrases alignées en relation de paraphrase. Cela peut certainement s'expliquer par la faible disponibilité de telles ressources engendrée par le coût de leur construction. Mais elles présentent des caractéristiques qui en font les candidates les plus naturelles pour l'étude de la paraphrase sous-phrastique : les phrases parallèles étant issues de la volonté d'exprimer la même idée, les équivalences apprises apparaissent comme beaucoup plus fiables que celles extraites indirectement via des textes comparables ou des équivalences de traduction. En outre, le contexte de ces équivalences peut être extrait de façon directe, ce qui est particulièrement important pour caractériser les

conditions de leur validité.

Ce travail porte sur l'acquisition de paraphrases sous-phrastiques depuis des corpus monolingues parallèles, et vise en particulier à extraire des paraphrases de qualité. Dans cet article, nous présentons DIST une nouvelle méthode symbolique optimisée pour l'alignement de bi-segments exploitant un corpus monolingue parallèle. Puis nous décrivons une approche hybride d'extraction de paraphrases sous-phrastiques par la combinaison d'informations issues de différentes techniques. Cet article est organisé comme suit : dans la section 2, nous passons en revue les travaux portant sur l'acquisition automatique de paraphrases puis nous détaillons, dans la section 3, le cadre expérimental de notre travail, l'approche suivie pour combiner des informations issues de différentes techniques et extraire des bi-segments à partir de corpus monolingues parallèles ainsi que les résultats obtenus. Nous terminerons par une description de nos prochains travaux (section 4).

2 Travaux précédents en acquisition de paraphrases

L'acquisition de paraphrases peut être réalisées à l'aide de diverses méthodologies. Langkilde & Knight (1998) se sont basés sur les connaissances sémantiques de WordNet (Miller, 1995) pour exploiter les relations de synonymie entre termes et les utiliser ensuite lors de la génération de paraphrases. Cependant, ces ressources ne sont pas nécessairement disponibles dans toutes les langues et ne comportent que des équivalences textuelles au niveau des mots. C'est la raison pour laquelle de nombreux autres travaux se sont basés sur des corpus monolingues et multilingues parallèles ou comparables.

La majorité des travaux menés sur des corpus monolingues parallèles se basent essentiellement sur l'hypothèse de distributionnalité (Harris, 1954), selon laquelle les mots apparaissant dans le même contexte tendent à avoir des sens similaires. Cette hypothèse a été appliquée, par exemple, à des chemins dans des arbres de dépendance pour la découverte de règles d'inférence à partir de textes (Lin & Pantel, 2001). Barzilay & McKeown (2001) utilisent des informations contextuelles basées sur des similarités lexicales pour extraire des paraphrases à partir d'un ensemble de corpus alignés. De manière similaire, Pang *et al.* (2003) exploitent la structure syntaxique d'un ensemble de phrases issues de corpus parallèles monolingues pour construire de nouvelles paraphrases d'énoncés par fusion syntaxique et regénération. Ibrahim *et al.* (2003) présentent eux une méthode non supervisée d'acquisition de paraphrases qui consiste à extraire des paraphrases structurelles, ou des fragments d'arbres syntaxiques sémantiquement équivalents, à partir de corpus monolingues parallèles.

Puisque les corpus monolingues parallèles sont des ressources rares et difficiles à obtenir, d'autres techniques ont été implémentées en se basant sur des corpus monolingues comparables, corpus composés de textes dans la même langue partageant une partie du vocabulaire employé, ce qui implique généralement que les textes parlent d'un même sujet, durant la même période, afin d'obtenir des paraphrases. Notamment, certains travaux exploitent des corpus monolingues comparables, comme ceux de Deléger & Zweigenbaum (2009) dans le domaine médical visant la construction d'un corpus de paraphrases de segments opposant les langues de spécialité et de vulgarisation. Barzilay & Lee (2003) introduisent une technique d'alignement multi-séquence factorisant des phrases ayant la même structure syntaxique, extraites à partir d'un corpus comparable, sous forme de treillis contenant des équivalences locales. Quirk *et al.* (2004) proposent une approche consistant à apprendre un système de traduction statistique sur un corpus monolingue de phrases alignées automatiquement à partir d'un corpus comparable qui opère par reformulations locales.

Outre les corpus monolingues, des corpus multilingues parallèles ont été exploités pour l'extraction des paraphrases en se basant sur l'hypothèse que des segments partageant des traductions dans une autre langue peuvent être des paraphrases dans certains contextes. Bannard & Callison-Burch (2005) ont décrit une approche par pivot exploitant plusieurs corpus parallèles. De la même manière, Max (2009) utilise des traductions de segments en pivot pour produire des reformulations et sélectionner parmi celles-ci celles qui sont préférées par différents types de modèles. La majorité de ces approches s'attaque au problème d'acquisition de paraphrases d'énoncés complets. Or, il est également intéressant de pouvoir extraire des reformulations pour des unités de texte plus petites à partir de plusieurs corpus quel que soit leur degré de parallélisme.

3 Combiner des informations pour l'alignement

Différentes approches peuvent être utilisées pour faire l'acquisition de paraphrases sous-phrastiques depuis des corpus monolingues parallèles (Bouamor *et al.*, 2010). Outre l'amélioration individuelle de ces techniques, il est possible de parvenir à une amélioration des performances obtenues en exploitant utilement les résultats de chacune. Dans cette section, nous commençons par décrire le cadre expérimental dans lequel s'ancre notre étude sur l'alignement monolingue dans des paires de paraphrases, puis nous présentons brièvement quatre techniques que nous utilisons pour cette tâche. Nous décrivons ensuite un cadre de combinaison des résultats qu'elles produisent et détaillons les résultats de nos expériences.

3.1 Cadre expérimental

Les paraphrases d'énoncés sont relativement rares à l'état naturel, car peu d'activités humaines en gardent la trace lorsqu'elles existent. En outre, certains types de réécritures, comme le résumé, altèrent de façon significative le contenu des textes. Des solutions pour l'acquisition de paraphrases ont cependant été proposées, par exemple à partir de corpus comparables (Dolan & Brockett, 2005) ou de traces d'éditions (Dutrey *et al.*, 2010), mais l'identification de ce qui constitue des paraphrases acceptables reste une difficulté majeure. Une solution plus directe consiste à faire produire de telles paraphrases par des humains dans le cadre naturel d'une traduction où une même phrase est traduite plusieurs fois indépendamment. Le corpus MultiTrad (Bouamor, 2010) a été construit selon ce principe en obtenant des traductions vers le français d'extraits du corpus des débats parlementaire européen.

Pour l'étude présentée ici, nous avons sélectionné un corpus de développement issu de MultiTrad constitué de 50 énoncés traduits 4 fois de l'anglais vers le français. Pour chaque groupe de quatre paraphrases, la paraphrase la plus similaire en moyenne aux autres paraphrases a été identifiée et associée aux trois autres. Cette similarité est calculée par une valeur moyenne d'édition mesurée par TER (*Translation Error Rate*) (Snover *et al.*, 2009). Les 150 paires de paraphrases obtenues ont alors été annotée au niveau des mots par 3 annotateurs à l'aide de YAWAT (Germann, 2008), un outil qui permet d'utiliser, au choix, une vue parallèle entre énoncés présentés sous forme de paragraphes ou de matrices d'alignement. Chaque paire a été annotée par un seul annotateur : Callison-Burch (2008) mentionne un accord inter-annotateur acceptable sur une telle tâche [1], mais l'ensemble des annotations a par la suite été vérifié par le même annotateur. À partir des matrices d'alignement produites, l'ensemble des bi-segments de référence est extrait en respectant la contrainte suivante : tous les mots du segment contenu dans la première paraphrase sont alignés avec au moins un mot du segment de la seconde paraphrase et ne sont alignés qu'avec des mots de ce segment, et réciproquement.

Pour évaluer la performance de nos techniques d'alignement monolingue, nous utilisons l'approche PARAMETRIC (Callison-Burch *et al.*, 2008), dans laquelle un ensemble de *bi-segments* (correspondant à des paires de paraphrases sous-phrastiques) de référence est comparé aux bi-segments produits par la méthode évaluée. La mesure PARAMETRIC se décompose en des valeurs usuelles de *précision* et de *rappel*, définies respectivement comme la proportion des candidats proposés appartenant à la référence et la proportion des éléments de la référence proposés, ainsi qu'en une *F-Mesure* combinant les deux à égalité. Notre évaluation portera sur un extrait du corpus de traductions multiples issus de la campagne CESTA [2] contenant 375 paires de paraphrases (comportant entre 15 et 25 mots) et obtenues par traduction de l'anglais vers le français. L'alignement de référence a été réalisé en suivant la même procédure que pour le corpus de développement avec 2 annotateurs. Notre étude a révélé un taux d'accord inter-annotateur global de 88,96% qu n'est plus, cependant, que de 67,35% lorsque les paraphrases "identité" ne sont pas prises en compte.

3.2 Techniques individuelles

Nous avons implémenté dans ce travail quatre techniques, développées pour des besoins différents. Nous les avons choisies parce qu'elles opèrent à différents niveaux ce qui devrait permettre de tirer parti de leur complémentarité potentielle. La première est fondée sur l'apprentissage statistique d'alignements entre mots (MOT), et requiert

1. Il faut cependant noter que les travaux de Callison-Burch (2008) portait sur des textes journalistiques en anglais et qu'un guide d'annotation avait été fourni aux annotateurs.
2. Corpus de la Campagne d'Evaluation de Systèmes de Traduction Automatique : `http://www.elda.org/article125.html`

donc des quantités de données d'apprentissage en nombre relativement important. La seconde exploite des règles de description de variantes de termes et des connaissances *a priori* sur la variation lexicale (TERME). La troisième utilise la structure syntaxique des énoncés pour mettre en correspondance des segments (SYNT), et requiert par conséquent un analyseur syntaxique. La quatrième, calcule une transformation au niveau des mots pour transformer une séquence de mots en une autre en mettant en jeu des opérations de transformation dont le coût est appris automatiquement (DIST). Une étude comparative des trois premières techniques a été faite dans (Bouamor *et al.*, 2010). Elle a, en particulier, mis en évidence des différences de performance notables sur deux types de corpus parallèles monolingues obtenus par traductions multiples à partir d'une même langue d'une part, et de plusieurs langues d'autre part. Dans cet article, une nouvelle technique est introduite et utilisée de façon originale, et une combinaison efficace sous forme d'adaptation de cette dernière technique est proposée.

3.2.1 Approche fondée sur l'apprentissage d'alignements entre mots (MOT)

La technique MOT consiste à apprendre des alignements entre mots en utilisant des modèles d'alignement statistique appliqués sur deux phrases parallèles, initialement conçus pour la tâche d'alignement bilingue entre mots en traduction automatique statistique. Une telle technique requiert typiquement des quantités de données importantes pour apprendre des alignements fiables [3]. Dans nos expériences, nous mettrons à disposition de MOT toutes les paires de paraphrases possibles (pour des groupes constitués de 4 paraphrases) afin d'améliorer ses capacités d'alignement, ce qui constitue pour elle un avantage car les autres techniques ne considèrent les paires de paraphrases qu'isolément (en d'autres termes, pour les autres techniques l'information acquise sur une paire de paraphrases n'est pas directement exploitée pour les alignements ultérieurs). Par ailleurs, ce type de technique fonctionne d'autant mieux que les phrases des corpus d'apprentissage utilisées sont *parallèles*, signifiant ici qu'un alignement mot à mot est facile à réaliser. Dans le cas bilingue, ce n'est évidemment pas le cas de langues très différentes, et dans le cas monolingue, nos expériences précédentes ont montré que MOT obtenait des résultats sensiblement meilleurs lorsque les paraphrases utilisées sont obtenues par traduction depuis une même langue.

Nous avons utilisé le programme GIZA++ (Och & Ney, 2003) pour réaliser l'alignement entre mots et les heuristiques du système de traduction statistique MOSES (Koehn *et al.*, 2007) pour extraire des bi-segments à partir des matrices d'alignement obtenues. Un exemple d'une matrice d'alignement produite par MOT est donné dans la figure 1. À partir de cette matrice, 12 bi-segments différents sont extraits en appliquant les critères décrits ci-dessus.

3.2.2 Approche fondée sur l'expression symbolique de la variation (TERME)

Pour chaque paire d'énoncés en relation de paraphrase, il est possible d'exprimer des règles régissant les variations syntagmatiques et paradigmatiques acceptables au niveau des segments. Les nombreux travaux qui ont porté sur les notions de *termes* et de *variantes de termes* offrent ainsi une solution assez directe à ce problème de mise en correspondance. L'approche symbolique TERME que nous utilisons exploite l'opération d'*indexation contrôlée* du système FASTR (Jacquemin, 1999) pour trouver les alignements sous-phrastiques possibles entre deux paraphrases d'une paire donnée. Cette opération définit les variations acceptables pour un terme par un système de métarègles décrivant ses réécritures morphosyntaxiques possibles. Les métarègles peuvent également mettre en jeu des relations lexicales définissant des variations morphologiques (mots d'une même famille morphologique) et sémantiques (synonymie). Ces ressources constituent donc des connaissances *a priori* utilisées par TERME qui ne sont pas accessibles aux autres techniques.

L'outil FASTR utilisé a été conçu pour rechercher efficacement des termes et leurs variantes dans de grands corpus de textes. Pour nos besoins, considérant une paire de paraphrases d'énoncés, nous recherchons dans la première phrase (notre « corpus ») des variantes pour chacun des segments possibles de l'autre phrase (à concurrence d'une certaine taille), puis nous inversons la recherche et retenons l'intersection des résultats. L'usage que nous faisons du moteur de détection de variantes de termes semble favorable à l'obtention d'une bonne précision. À l'inverse, les métarègles définies pour le repérage de variantes de termes ne sont pas nécessairement les mieux adaptées pour assurer une bonne couverture des phénomènes paraphrastiques entre segments de nature quelconque (Dutrey *et al.*, 2010).

3. La technique développée par Lardilleux (2010) constitue une exception notable adaptée aux événements de basse fréquence, et sera naturellement considérée dans la suite de nos travaux.

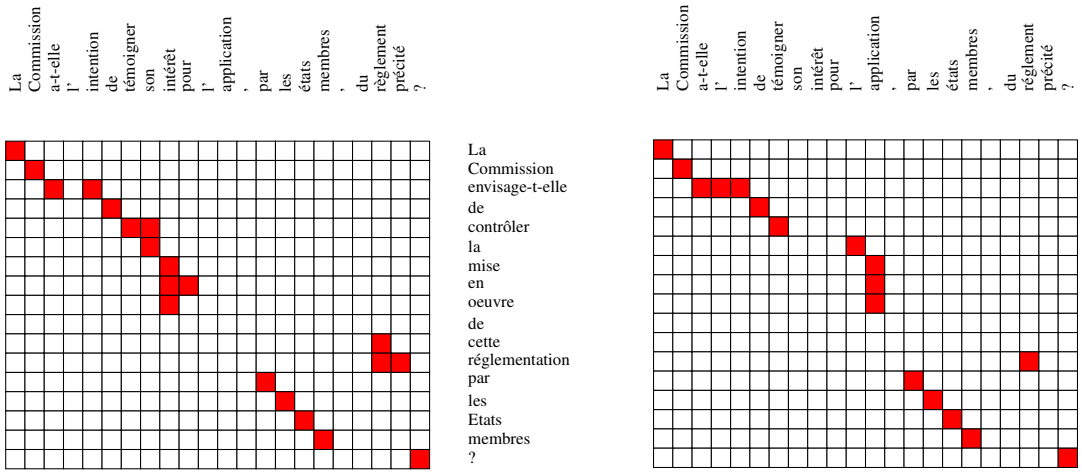


FIGURE 1 – Matrice d'alignement d'une paire de phrases dans MOT (à gauche), et sa matrice correspondante dans la base de référence.

3.2.3 Approche fondée sur l'alignement de structures syntaxiques (SYNT)

Lorsque deux énoncés en relation de paraphrase partagent une même structure syntaxique, il est possible de réaliser un alignement fin guidé par la syntaxe permettant de faire apparaître des correspondances sous-phrastiques fines. L'algorithme de Pang *et al.* (2003) décrit une *fusion syntaxique* consistant essentiellement à fusionner des arbres de constituants de deux énoncés là où les listes de catégories filles sont compatibles et qu'aucune évidence de non parallélisme syntaxique (via un mécanisme de *blocage lexical*) n'est détectée. La forêt d'arbres syntaxiques ainsi obtenue permet de construire un treillis de mots représentant des formulations alternatives qu'il est possible d'extraire par simple parcours du treillis.

Pour la méthode SYNT nous avons réimplémenté l'algorithme originel et avons amélioré sa robustesse et sa correction en ajoutant un mode de fusion flexible dans lequel les parties de la phrase non concernées par un blocage lexical sont tout de même fusionnées. Par ailleurs, étant donné que l'algorithme est très dépendant de la qualité des analyses syntaxiques produites, nous avons également ajouté un mode exploitant les k meilleures analyses produites par un analyseur probabiliste. La combinaison retenue entre une analyse du premier énoncé et une analyse du second parmi les k^2 combinaisons possibles est celle minimisant le nombre de nœuds dans le treillis obtenu avant réduction. Un exemple de treillis obtenu par application de SYNT est donné dans la figure 2.

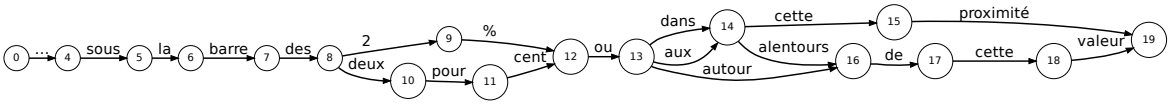


FIGURE 2 – Exemple d'un treillis obtenu par application de SYNT

Tout comme TERME, cette technique semble *a priori* plus adaptée à l'extraction précise de bi-segments monolingues, mais contrairement à TERME il est attendu qu'elle ne parvienne pas à extraire de correspondance lorsque les structures syntaxiques de haut niveau des paraphrases d'énoncés ne sont pas compatibles.

3.2.4 Approche fondée sur la distance d'éditions sur des séquences de mots (DIST)

Une relation entre deux paraphrases peut également s'exprimer sous forme de la séquence d'éditions la plus directe permettant de transformer l'une en l'autre. Une telle séquence d'éditions sur les mots est, par exemple,

implémentée dans la technique TERp (Translation Edit Rate plus) (Snover *et al.*, 2009), originellement développée pour le calcul d'une distance d'édition servant de mesure en traduction automatique pour évaluer une hypothèse de traduction relativement à une traduction de référence. Ce calcul met en jeu des opérations de transformation de chaîne incluant l'insertion, la suppression et la substitution de mots, ainsi que le déplacement et la substitution de segments. Chaque type d'opération est associé à une pondération optimisée sur un corpus de développement pour une mesure particulière, et l'algorithme effectue une recherche de la séquence d'opération la moins coûteuse. Les substitutions de mots ou segments sont optionnelles, mais peuvent exploiter des listes fournies à l'algorithme [4], et les substitutions de segments ont une probabilité associée.

Pour son calcul, TERp produit donc un alignement au niveau des mots entre deux énoncés. Pour nos besoins, nous avons implémenté une méthode DIST qui extrait l'ensemble des bi-segments (à concurrence d'une taille maximale) dérivables des alignements produits par TERp. Nous avons exploité la possibilité d'optimiser TERp pour nos besoins, en optimisant ses paramètres par la méthode du *hill climbing* [5]. Par la suite, nous dénoterons DIST_A l'optimisation originelle réalisée par Snover *et al.* (2009) pour l'évaluation de la traduction automatique (le « A » est pour « *adequacy* »). Les variantes DIST_P, DIST_R et DIST_{F_1} correspondent à des optimisations réalisées sur un corpus de développement maximisant respectivement la précision, le rappel et la F-mesure de PARAMETRIC exploitant des annotations de référence. L'ensemble de ces configurations n'utilisent pas de substitutions de segments, mais nous ferons appel à cette possibilité dans un cadre d'hybridation décrit plus loin. Un exemple de résultat d'alignement fourni par TERp est donné dans la figure 3.

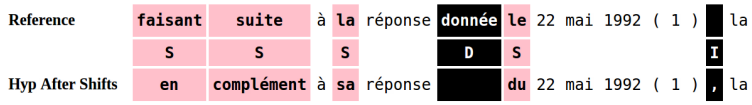


FIGURE 3 – Exemple d'un alignement résultat de DIST

3.2.5 Résultats expérimentaux et analyse

Nous avons évalué chacune des méthodes présentées ci-dessus sur le corpus de test décrit dans la section 3.1. Les techniques MOT, TERME et SYNT ont été utilisées telles que décrites. Pour DIST, nous avons exploité la possibilité d'optimiser la mesure selon nos propres objectifs. La variante DIST_A, évaluée pour référence, correspond à la version de TERp optimisée pour les besoins de l'évaluation de la traduction automatique. Les autres variantes DIST_P, DIST_R et DIST_{F_1} correspondent à TERp optimisée pour maximiser respectivement la précision, le rappel et la f-mesure de PARAMETRIC. La première partie de la table 1 donne les résultats obtenus sur les trois sous-mesures de PARAMETRIC. On constate tout d'abord que les résultats pour les 3 premières techniques sont cohérents avec ceux obtenus dans (Bouamor *et al.*, 2010). La seule différence notable est l'amélioration de la précision des deux techniques symboliques TERME et SYNT. La technique statistique d'alignement entre mots MOT obtient un rappel beaucoup plus important que les deux autres techniques qui se distinguent par une précision relativement forte (60,87 pour TERME et 66,96 pour SYNT). La précision de MOT reste toutefois dans une zone raisonnable à 47,02. Comme expliqué précedemment, MOT tire avantage des 3 paires de paraphrases sur lesquelles il peut réaliser son apprentissage, alors que les deux autres techniques, telles qu'implémentées, ne peuvent améliorer l'alignement à l'intérieur d'une phrase en exploitant des informations dérivées d'autres phrases.

L'ajout original pour notre tâche de DIST, technique fondée sur une distance d'édition sur des séquences de mots, révèle de nouveaux résultats intéressants. Tout d'abord, on constate qu'au niveau de la f-mesure, il n'existe qu'une faible différence entre DIST_A et la variante optimisée sur la f-mesure, DIST_{F_1}. Ceci signifie que nos objectifs sont très similaires à ceux de l'évaluation en traduction automatique tels que décrits par (Snover *et al.*, 2009). On constate cependant que des optimisations spécifiques en faveur de la précision ou du rappel mènent ici à des gains très importants de +8,69 en précision et de +7,42 en rappel. Ces résultats montrent que la technique DIST peine à améliorer simultanément la précision et le rappel, même si celle-ci obtient globalement des performances très proches de la meilleure technique envisagée jusque-là, MOT, avec une précision légèrement meilleure et un rappel

4. La version standard de TERp fournit des techniques de racinisation ainsi que des ressources de synonymie ainsi que de paraphrases locales pour l'anglais. TERp utilise jusqu'à 11 paramètres.

5. La première itération d'optimisation se fait avec des poids uniformes, puis nous réalisons 10 itérations avec des valeurs initiales tirées aléatoirement afin de diminuer le risque d'utiliser un minimum local.

légèrement inférieur. Il est possible que les modèles mis en jeu pour le calcul de la distance d'édition ne soient pas suffisamment expressifs pour nos besoins, et qu'en particulier, la non prise en compte de critères linguistiques pour opérer des transformations de séquences de mots soit à mettre en cause.

	Précision	**Rappel**$_{/13532}$	$\mathbf{F}_1$
Mot	47,02	61,42	53,26
Terme	60,87	4,19	7,85
Synt	66,96	13,11	21,92
Dist$_A$	49,85	54,14	51,91
Dist$_P$	58,54	2,68	5,13
Dist$_R$	39,48	61,56	48,11
Dist$_{F_1}$	49,03	56,21	52,37
union(Mot,Terme,Synt,Dist$_{F_1}$**)**	38,99	**73,55**	50,97
intersection(Mot,Dist$_{F_1}$**)**	**70,38**	32,31	44,29

TABLE 1 – Résultats obtenus pour chaque technique

La dernière partie de la table 1 donne les résultats obtenus en opérant une combinaison élémentaire des résultats visant à maximiser d'une part la précision, et d'autre part le rappel. L'union sur le résultat de l'ensemble des techniques obtient un maximum de valeur de rappel de 73,55 (+12,13 relativement à MOT), avec une précision légèrement affectée (-2,29 relativement à MOT). Par ailleurs, réaliser l'intersection entre les différentes techniques peut raisonnablement mener à une précision améliorée. Cependant, le peu de résultats produits par TERME et SYNT nous ont fait préférer une mesure sur l'intersection de MOT et DIST$_{F_1}$: nous obtenons alors une valeur maximale de précision de 70,38 (+23,36 relativement à MOT et +21,35 relativement à DIST$_{F_1}$). Ces résultats montrent bien la complémentarité qui existe entre ces différentes techniques, et servent donc ici de motivation pour la recherche d'un mode de combinaison plus efficace des informations issues de chaque technique.

3.3 Approche hybride d'extraction de paraphrases locales

3.3.1 Observations et motivations

Les expériences présentées dans la section 3.2.5 ont révélé que les différentes techniques ont des performances variées, ce qui permet aussi de faire l'hypothèse qu'il est possible d'opérer une combinaison efficace de leurs résultats. Nous faisons ici une synthèse des points forts et des limitations de chacune de ces techniques orientée par la recherche d'un mode de combinaison plus efficace :

– MOT : très sensible à la fréquence de ses observations de mots et de cooccurrences entre mots, cette technique peut être informée par des connaissances d'association *a priori*, qui peuvent par exemple être transmises sous forme de données d'apprentissage additionnelles. En outre, il est possible, avec des données d'entraînement annotées, d'améliorer les performances des alignements statistiques par apprentissage discriminant (Tomeh *et al.*, 2010).
– TERME : cette technique est spécialisée dans l'extraction d'un type de bi-segments contraints par des règles de réécriture et de variation lexicale. Les métarègles, qui ont été développées manuellement, sont assez précises et ne peuvent couvrir tous les phénomènes de paraphrase. Leur apprentissage automatique peut améliorer la couverture, mais au détriment de la précision. L'enrichissement automatique des familles morphologiques et sémantiques devrait également permettre d'augmenter le rappel.
– SYNT : cette technique est très sensible au degré de parallélisme des énoncés qui décide de la fusion de constituants syntaxiques. Nous avons déjà pris en compte la qualité des analyses syntaxiques en autorisant la fusion à opérer sur les *k*-meilleures analyses syntaxiques. Le blocage lexical empêche une fusion lorsqu'un mot présent dans le constituant d'une phrase est attesté dans un constituant non aligné de l'autre phrase. Il pourrait être amélioré par la connaissance *a priori* de paraphrases locales, ce qui, néanmoins, ne pourrait bénéficier qu'à la précision.
– DIST : cette technique transforme une séquence de mots en une autre en un coût minimal, en utilisant des pondérations optimisées pour les différentes opérations utilisées. L'algorithme manipule des segments qui n'ont pas nécessairement de motivation linguistique, ce qui peut mener à des transformations aberrantes. En outre, des

opérations d'insertion et de suppression peuvent être utilisées à tort lorsque des correspondance au niveau des mots ou des segments ne sont pas connues. Ainsi, si de telles correspondances peuvent être fournies à TERp, il est possible d'espérer diminuer le nombre d'opérations de transformation aberrantes et ainsi d'augmenter la performance.

3.3.2 Présentation de l'hybridation des méthodes

Dans la section précédente nous avons montré qu'il existait plusieurs voies pour améliorer la performance de l'alignement monolingue auquel nous nous intéressons à partir des techniques décrites. Sans considérer davantage, à ce stade, l'amélioration individuelle de chacune des techniques, nous pouvons décrire les deux grandes familles d'approches possibles pour l'hybridation de la manière suivante : 1) les résultats produits indépendamment par chaque technique sont combinés *a posteriori*, et 2) une technique est *adaptée* par la connaissance des résultats produits par les autres techniques.

Nous avons déjà montré le résultat de l'évaluation d'une approche élémentaire par combinaison *a posteriori* dans la section 3.2.5, illustrée sur la partie gauche de la figure 4, qui révèle que la précision et le rappel peuvent ainsi être facilement améliorés. Nous considérons désormais la seconde approche. D'après nos observations, DIST est un candidat assez naturel pour l'adaptation. En effet, la connaissance d'alignements au niveau des mots ou des segments peut diminuer le nombre d'opérations effectuées à tort. Il s'agit précisément de la motivation majeure pour l'évolution de TER à TERp (Snover *et al.*, 2009), liée à la possibilité d'utiliser une base de paraphrases locales connues *a priori* et ainsi d'être plus robustes quant aux hypothèses de traduction acceptées par le système lorsqu'elle ne correspondent pas exactement à une traduction de référence.

Au contraire de ce qui est fait dans TERp, nous n'utiliserons pas une base de connaissances externe, même si nous ne rejetons pas cette hypothèse pour de futures expériences, mais nous adaptons dynamiquement la base de paraphrases utilisées en fonction des hypothèses extraites par les autres techniques, à savoir des hypothèses nombreuses et relativement précises pour MOT, et peu nombreuses mais précises pour TERME et SYNT. De plus, comme nous l'avons déjà montré, ces techniques peuvent être complémentaires quant aux types de paraphrases locales qu'elles permettent d'identifier, ce qui rejoint nos intuitions initiales liées à la nature de chacune d'elles.

Cette approche est illustrée sur la partie droite de la figure 4. Les bi-segments obtenus par MOT, TERME et SYNT sont combinés pour construire une table de paraphrases utilisée ensuite par DIST, que nous pouvons optimiser en fonction d'un besoin particulier (précision, rappel ou f-mesure). On remarque ici une analogie assez directe avec d'autres scénarios de combinaisons d'informations en TAL : en traduction automatique, l'approche de la partie gauche de la figure 4 correspond à la définition classique de la combinaison de systèmes (Matusov *et al.*, 2009), alors que l'approche de la partie droite correspond à l'adaptation d'un système par des sources externes telles que d'autres systèmes de traduction (Crego *et al.*, 2010).

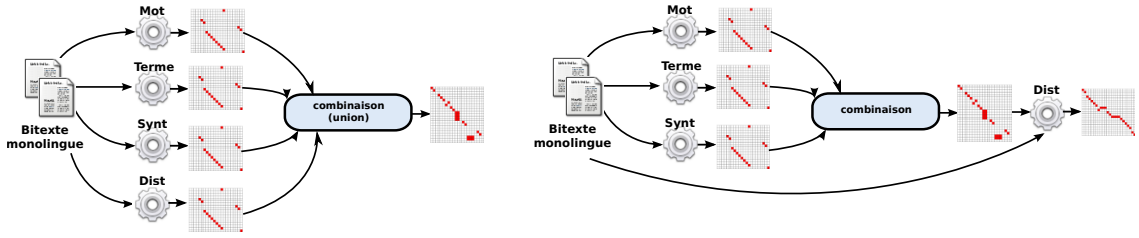


FIGURE 4 – Principales approches de combinaisons d'informations pour l'alignement multilingue. À gauche, plusieurs techniques produisent des résultats combinés pour produire une nouvelle sortie. À droite, un sous-ensemble des techniques fournissent leurs résultats à une dernière technique adaptée à l'exploitation de ces connaissances.

Un problème important à considérer concerne la manière dont la table de paraphrases utilisée par TERp est construite à partir des hypothèses produites par les différentes techniques. À ce stade de nos travaux, nous ne disposons pas de *mesures de confiance* données par chaque technique pour chacune de ses hypothèses, et nous sommes donc contraints de les considérer initialement comme équiprobables. De plus, pour assurer une comparaison plus directe avec la combinaison correspondant à la partie gauche de la figure 4, nous réalisons une combinaison simple

à base d'union : chaque hypothèse apparaissant au moins une fois parmi les hypothèses des différents systèmes est retenue et est associée à un poids constant uniforme. [6]

Un autre aspect important concerne là encore la pondération associée à chacune des paires de paraphrases *a priori* fournies à TERp. Considérons le cas où deux paraphrases sont fournies à TERp et où l'une est un sous-segment de l'autre : par exemple, (*ce degrévement* ↔ *cet allègement*) inclut (*dégrèvement* ↔ *allègement*). Si ces deux paraphrases sont fournies avec le même score à TERp, celui-ci préfèrera, dans de nombreux cas, utiliser la plus couvrante des deux, car cela minimisera souvent la quantité d'opérations de transformation restant à faire, et donc le coût global de transformation (voir partie gauche de la figure 5). Cela peut ne pas être un défaut en soi, car l'identification des plus longues sous-unités paraphrastiques peut être utile. Cependant, PARAMETRIC base ces mesures sur l'ensemble des bi-segments pouvant être extraits à partir d'alignement sur les mots. Ainsi, si dans l'exemple précédent l'alignement de référence inclut deux points d'alignement pour (*ce* ↔ *cet*) et (*dégrèvement* ↔ *allègement*), l'ensemble des bi-segments de référence sera constitué des deux bi-segments précédents et de leur combinaison ou « extension » (*ce degrévement* ↔ *cet allègement*). Si ce dernier est utilisé par TERp, il n'existe pas de moyen immédiat pour retrouver l'alignement sous-phrastique, et donc le rappel de la technique adaptée sera pénalisé.

Plusieurs solutions sont envisageables pour pallier ce problème. La pondération des paraphrases pourraient prendre en compte le nombre de mots/tokens couverts en favorisant les courts segments. Ne disposant néanmoins pas de solutions génériques applicables à toutes les techniques ni de moyen d'intégrer des scores de confiance motivés, nous préférons nous en remettre à une solution initiale plus simple, qui consiste à ne conserver que les sous-segments minimaux parmi l'union de ceux proposés par chacune des techniques. Ainsi, ne seront gardés pour construire la table de paraphrases utilisée par TERp que les bi-segments n'étant inclus dans aucun autre bi-segment, que nous appellons *bi-segments minimaux*.

FIGURE 5 – Exemple de deux alignements résultats de DIST_{F_1}, avec à gauche l'ensemble des bi-segments non filtrés, et à droite un ensemble de bi-segments minimaux

3.3.3 Résultats expérimentaux et analyse

Les résultats que nous obtenons en optimisant TERp sur les trois mesures de PARAMETRIC et en utilisant différentes sources de bi-segments sont présentés dans la table 2. Le résultat principal de ces expériences est la nouvelle f-mesure de 55,27 obtenue en optimisant sur cette mesure et en exploitant les bi-segments provenant des trois autres techniques. C'est la valeur la plus élevée sur l'ensemble de nos expériences, et elle correspond notamment à un gain de +4,3 par rapport à la combinaison par union des résultats de toutes les techniques, ou encore à un gain de +2,01 par rapport à MOT, la meilleure technique individuelle pour la f-mesure, et à un gain de +2,9 par rapport à DIST_{F_1}, la technique utilisée sans adaptation et optimisée selon le même critère. Ces résultats viennent confirmer notre hypothèse que TERp a pu ici tirer utilement profit des connaissances *a priori* qui lui ont été fournies.

Nous constatons de plus que des valeurs de précision et de rappel encourageantes peuvent être atteintes : une précision de 69,66 est obtenue en exploitant les prédictions de TERME et en optimisant sur la précision (+2,7 par rapport à la meilleure technique individuelle SYNT), et un rappel de 62,38 est obtenu en exploitant les prédictions de MOT en optimisant sur le rappel (+0.82 par rapport à la meilleure technique individuelle DIST_{R}).

Les cas de combinaisons où une seule technique est utilisée pour alimenter la base de paraphrases de TERp peuvent également être étudiés en comparant les valeurs des tables 1 et 2. Hormis les valeurs de rappel obtenues pour

6. Il serait bien sûr possible de pondérer *a priori* chaque hypothèse par le nombre de techniques l'ayant proposée, et/ou par la performance mesurée des techniques en question, dérivée par exemple de leur performance individuelle dans les différentes valeurs de PARAMETRIC. En outre, la contribution de chacune des techniques pourrait faire l'objet d'un paramètre optimisé simultanément aux paramètres de TERp. Toutes ces possibilités seront considérées dans notre travail futur.

Source de bi-segments	Critère d'optimisation								
	DIST$_P$			DIST$_R$			DIST$_{F_1}$		
	P	R$_{/13532}$	F$_1$	P	R$_{/13532}$	F$_1$	P	R$_{/13532}$	F$_1$
MOT	67,83	13,21	22,11	41,49	**62,38**	49,83	55,11	54,51	54,81
TERME	**69,66**	6,82	12,42	40,51	55,6	46,87	53,16	49,84	51,45
SYNT	68,08	8,11	14,48	29,99	56,84	39,26	51,25	50,14	50,69
comb(MOT, SYNT, TERME)	66,02	13,15	21,93	38,46	61,09	47,2	55,01	55,54	**55,27**

TABLE 2 – Résultats obtenus pour différentes optimisations et différentes sources de bi-segments. La fonction *comb* correspond à l'union avec pondération uniforme des bi-segments ne retenant que les bi-segments minimaux.

DIST$_R$ avec les paraphrases de TERME et SYNT, toutes les autres combinaisons de DIST avec les données d'une autre technique et optimisées selon un critère particulier améliorent la meilleure des deux valeurs précédentes. Par exemple, DIST$_P$ adapté avec les paraphrases de TERME obtient une précision de 69,66, qui est meilleure que celle de DIST$_P$ (58,54) et celle de TERME (60,87). Il est à noter qu'en combinaison de systèmes, comme c'est par exemple le cas en traduction automatique, des gains sont plus généralement obtenus lorsque un certain nombre de systèmes sont combinés. La complémentarité de nos sources d'information et l'impact assez immédiat d'une amélioration des informations *a priori* utilisées par TERp semblent donc ici avoir un rôle très bénéfique pour notre tâche.

Il est finalement instructif de considérer la performance des différentes configurations testées en fonction d'une certaine difficulté *a priori*. Celle-ci pourrait se mesurer par le degré d'accord inter-annotateurs pour chaque phrase, mais nous avons choisi d'utiliser un résultat en lien avec TERp : $(1 - TER(paraphrase_1, paraphrase_2))$, qui est donc d'autant plus grand que les phrases sont proches. Le résultat pour nos quatre techniques individuelles est présenté dans la figure 6. Pour la précision, on constate tout d'abord que MOT est très sensible à la difficulté telle que nous la définissons, et que les alignements que cette technique produit sont d'autant moins bons que les phrases sont différentes. De façon un peu plus surprenante, SYNT et DIST$_P$ ne semblent pas trop affectés par cette difficulté. Cependant, ceci est peut-être dû au fait que les valeurs des barres, pour chaque intervalle discrétisé, sont une moyenne qui ne rend pas compte du nombre d'éléments. Il est possible que SYNT extraie peu de bi-segments sur des paires de phrases difficiles, mais que lorsqu'elle parvient à trouver des structures syntaxiques compatibles, celles-ci permettent un alignement précis. Enfin, TERME est lui insensible à cette difficulté, ce qui était attendu puisqu'elle fonctionne sur de courts patrons morphosyntaxiques pouvant impliquer des mots différents. Nous déduisons donc de ces remarques que ces différentes techniques peuvent être utilisées à bon escient pour différents niveaux de parallélisme des corpus d'acquisition. Le rappel fait apparaître une tendance beaucoup plus marquée : MOT, DIST$_R$ et SYNT extraient d'autant moins de bi-segments de la référence que les phrases sont difficiles. À nouveau, TERME y semble insensible. On retiendra de cette analyse qu'il est préférable d'avoir des paraphrases d'énoncés les plus « parallèles » possibles pour obtenir une bonne performance en acquisition, mais que les techniques symboliques sont utiles pour extraire des paraphrases sous-phrastiques précises dans des paraphrases d'énoncés de formes très différentes.

4 Conclusion et travaux futurs

Dans cet article nous avons poursuivi deux objectifs. D'une part, nous avons présenté quatre méthodes d'acquisition de paraphrases sous-phrastiques à partir de corpus monolingues parallèles. Trois d'entre elles avaient déjà été évaluées, la dernière est nouvelle. Ces méthodes reposent sur des caractéristiques linguistiques différentes : MOT sur l'apprentissage statistique, TERME sur un approche symbolique de la variation de termes, SYNT sur des proximités syntaxiques et enfin DIST sur des distances d'édition. En évaluant ces méthodes, nous avons constaté qu'effectivement leurs résultats semblent complémentaires, ce qui nous a mené à notre second objectif, à savoir l'hybridation de ces méthodes. Plutôt que de combiner les résultats *a posteriori*, nous avons choisi d'utiliser les résultats de certaines méthodes comme données d'entrée d'une autre. Les résultats de cette approche ont confirmé notre hypothèse en montrant que la complémentarité de ces techniques donne un gain significatif.

De nombreuses pistes s'ouvrent à nous à la suite de ce travail. Nous souhaitons explorer toutes celles évoquées au cours de cet article. À court terme, nous comptons attribuer des scores de confiance à chacune des techniques afin

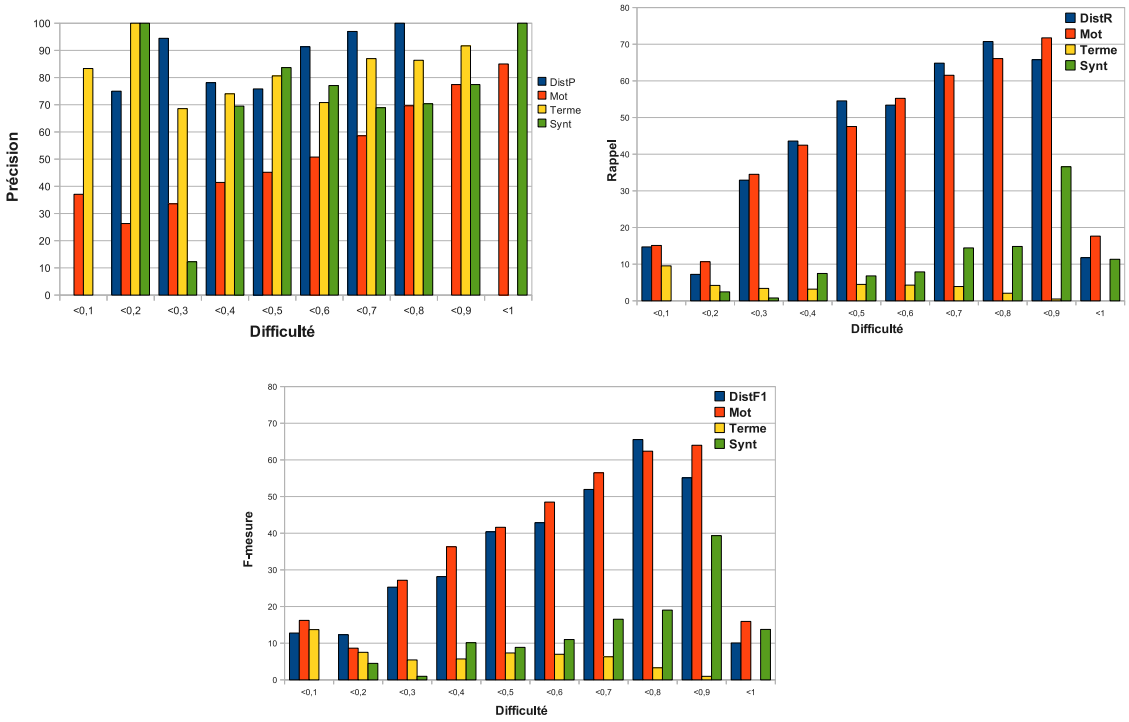


FIGURE 6 – Performance selon les différents critères de PARAMETRIC de différentes techniques. La valeur de chaque barre dans les intervalles discrétisés est une moyenne des éléments de cet intervalle, et ne rend pas compte du nombre de ces éléments. Pour la précision, une valeur de 0 peut indiquer soit l'absence de proposition pour les phrases de cet intervalle, soit de propositions toutes incorrectes.

de mieux tirer parti de leur complémentarité. Nous allons également utiliser des connaissances complémentaires. Il est important de noter que cette méthode peut s'adapter à la tâche requérant des paraphrases. Ainsi, on peut souhaiter en obtenir de nombreuses, au détriment de leur qualité pour de la recherche d'information, alors que la correction sera privilégiée pour le résumé automatique.

Références

BANNARD C. & CALLISON-BURCH C. (2005). Paraphrasing with bilingual parallel corpora. In *Actes de ACL*, Ann Arbor, USA.

BARZILAY R. & LEE L. (2003). Learning to paraphrase : an unsupervised approach using multiple-sequence alignment. In *Actes de NAACL-HLT*, Edmonton, Canada.

BARZILAY R. & MCKEOWN K. (2001). Extracting paraphrases from a parallel corpus. In *Actes de ACL*, Toulouse, France.

BOUAMOR H. (2010). Construction d'un corpus de paraphrases d'énoncés par traduction multilingue multi-source. In *Récital-TALN*, Montréal, Canada.

BOUAMOR H., MAX A. & VILNAT A. (2010). Acquisition de paraphrases sous-phrastiques depuis des paraphrases d'énoncés. In *Actes de TALN 2010*, Montréal, Canada.

CALLISON-BURCH C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. In *Actes de EMNLP*, Hawai, USA.

CALLISON-BURCH C., COHN T. & LAPATA M. (2008). Parametric : An automatic evaluation metric for paraphrasing. In *Actes de COLING*, Manchester, UK.

CREGO J. M., MAX A. & YVON F. (2010). Local lexical adaptation in machine translation through triangulation : SMT helping SMT. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China.

DELÉGER L. & ZWEIGENBAUM P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora*.

DOLAN W. B. & BROCKETT C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, Jeju Island, South Korea.

DUCLAYE F., COLLIN O. & YVON F. (2003). Apprentissage automatique de paraphrases pour l'amélioration d'un système de questions-réponses. In *Actes de TALN*, Batz-sur-mer, France.

DUTREY C., BOUAMOR H., BERNHARD D. & MAX A. (2010). Local modifications and paraphrases in wikipedia's revision history. In *Workshop on Corpus-Based Approaches to Paraphrasing and Nominalization,CBA 2010*, Barcelone, Espagne.

GERMANN U. (2008). Yawat :Yet Another Word Alignment Tool. In *Proceedings of the ACL-08 : HLT Demo Session*, Columbus, Ohio.

HARRIS Z. (1954). Distributional structure. *Word*.

IBRAHIM A., KATZ B. & LIN J. (2003). Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the second international workshop on Paraphrasing* : Association for Computational Linguistics.

JACQUEMIN C. (1999). Syntagmatic and paradigmatic representations of term variation. In *Actes de ACL*, College Park, États-Unis.

KAUCHAK D. & BARZILAY R. (2006). Paraphrasing for automatic evaluation. In *Actes de NAACL-HLT*, New York, États-Unis.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of ACL, demo session*, Prague, Czech Republic.

LANGKILDE I. & KNIGHT K. (1998). Generations that Exploits Corpus-based Statistical Knowledge. In *Proceedings of the 36th International Conference on Computational Linguistics*.

LARDILLEUX A. (2010). *Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle*. PhD thesis, Université de Caen, France.

LIN D. & PANTEL P. (2001). Discovery of inference rules for question answering. *Natural Language Engineering*, **7**(4).

MATUSOV E., LEUSCH G. & NEY H. (2009). *Learning To Combine Machine Translation Systems*. MIT Press.

MAX A. (2008). Génération de reformulations locales par pivot pour l'aide à la révision. In *Actes de TALN*, Avignon, France.

MAX A. (2009). Sub-sentencial paraphrasing by contextual pivot translation. In *Proceedings of the ACL 2009 Workshop on Applied Textual Inference*, Singapore.

MILLER G. A. (1995). Wordnet : a lexical database for english. *Commun. ACM*, **38**(11).

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*.

PANG B., KNIGHT K. & MARCU D. (2003). Syntax-based alignement of multiple translations : Extracting paraphrases and generating new sentences. In *Actes de NAACL-HLT*, Edmonton, Canada.

QUIRK C., BROCKETT C. & DOLAN W. B. (2004). Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP*, volume 149, Barcelona, Spain.

RUSSO-LASSNER .G L. J. & .P R. (2005). *A Paraphrase-Based Approach to Machine Translation Evaluation*. Rapport interne TR-2005-57, UMIACS.

SNOVER M., MADNANI N., DORR B. & SCHWARTZ R. (2009). Fluency, adequacy, or HTER ? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece.

TOMEH N., ALLAUZEN A., WISNIEWSKI G. & YVON F. (2010). Refining word alignment with discriminative training. In *Proceedings of The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*.

TALN 2011, Montpellier, 27 *juin –*1^{er} *juillet 2011*

Alignment of Monolingual Corpus by Reduction of the Search Space

Prajol Shrestha
Prajol.Shrestha@etu.univ-nantes.fr

Résumé. Les corpus comparables monolingues, alignés non pas au niveau des documents mais au niveau d'unités textuelles plus fines (paragraphe, phrases, etc.), sont utilisés dans diverses applications de traitement automatique des langues comme par exemple en détection de plagiat. Mais ces types de corpus ne sont pratiquement pas disponibles et les chercheurs sont donc obligés de les construire et de les annoter manuellement, ce qui est un travail très fastidieux et coûteux en temps. Dans cet article, nous présentons une méthode, composée de deux étapes, qui permet de réduire ce travail d'annotation de segments de texte. Cette méthode est évaluée lors de l'alignement de paragraphes provenant de dépêches en langue anglaise issues de diverses sources. Les résultats obtenus montrent un apport considérable de la méthode en terme de réduction de temps d'annotation. Nous présentons aussi des premiers résultats obtenus à l'aide de simples traitements automatiques (recouvrement de mots, de racines, mesure cosinus) pour tenter de diminuer encore la charge de travail humaine.

Abstract. Monolingual comparable corpora annotated with alignments between text segments (paragraphs, sentences, etc.) based on similarity are used in a wide range of natural language processing applications like plagiarism detection, information retrieval, summarization and so on. The drawback wanting to use them is that there aren't many standard corpora which are aligned. Due to this drawback, the corpus is manually created, which is a time consuming and costly task. In this paper, we propose a method to significantly reduce the search space for manual alignment of the monolingual comparable corpus which in turn makes the alignment process faster and easier. This method can be used in making alignments on different levels of text segments. Using this method we create our own gold corpus aligned on the level of paragraph, which will be used for testing and building our algorithms for automatic alignment. We also present some experiments for the reduction of search space on the basis of stem overlap, word overlap, and cosine similarity measure which help us automatize the process to some extent and reduce human effort for alignment.

Mots-clés : corpus comparable monolingue, alignement, similarité.

Keywords: monolingual comparable corpus, alignment, similarity.

1 Introduction

Monolingual comparable corpora from its name can be understood to be a collection of electronic text documents in a single language collected on the basis of comparability. The characteristics of comparability has not been explained and analyzed in many literature and the ones that explain are made for bilingual comparable corpora (Maia, 2003). Comparability depends on the application for which the documents are collected for instance, for information retrieval, comparable corpora would be documents that are related to a set of other documents whereas for detecting and measuring reused text, the comparable corpora are documents that have been rewritten from a set of previously written documents (Gaizauskas *et al.*, 2001).

In the field of Natural Language Processing (NLP), monolingual comparable corpora are used to build and test a wide range of applications such as information retrieval, summarization, plagiarism detection, dictionary building and so on. With a number of application to be built, the field of NLP requires a wide range of monolingual comparable corpus with different annotations. The annotations can be at different levels starting from word to document level and different types of annotations such as Part-of-Speech annotation of words to annotations of related documents. In this paper, our focus of annotation is the annotation of alignment between text segments based on similarity and from here onwards annotation will refer to alignment. There are some standard monolingual comparable corpus available, created manually by collecting existing documents written by humans or artificially by generating machine made text. These corpora are annotated on the level of :

– Documents, TDT corpus [1], for Topic detection and tracking applications created manually.
– Text segments, PAN-PC-09 (Barrón-Cedeño *et al.*, 2010), for plagiarism detection and are artificially created due to the fact that natural text on plagiarism are hard to collect.
– Documents as well as text segments, METER corpus (Gaizauskas *et al.*, 2001), for the detection of text reuse created manually.

From the list above, we can realize that for each NLP task, that use comparable corpus, we will require a corpus with specific annotations. There aren't many corpus like these that are available and as the field of NLP expands, more corpus with specific annotations would be required. Other NLP activities that do not make use of these annotations made in the standard monolingual comparable corpus make their own specific annotations. These corpus are not available for the public and therefore for NLP tasks that does not have a standard annotated corpus, a new corpus with annotations has to be built.

Most of the annotation are made manually by two or more professional annotators who search for alignments on all possible alignment pairs. In monolingual corpora, the alignments annotated are very few compared to the total possible alignment searched by the annotators. We propose a method to reduce this search space of the total possible alignments so that this task becomes much less time consuming and costly. This method can be used to align a wide range of text segments as we show how it helps in aligning similar paragraphs in monolingual comparable corpus and also present some methods that will help automatize it to some extent.

We will start in section 2 by describing the alignment process in which the reduction of space is explained in section 2.1 and the creation of the gold corpus in section 2.2. In section 3, we present the experiments carried out and their results. We finally conclude and present our future work in section 4.

2 Alignment

Alignment, in general, could be seen as a problem in which segments of text are grouped or linked to each other based on some relations like synonymy, translation, similarity and so on (Indurkhya & Damerau, 2010). The granularity of the segments for alignment may vary : characters, words, phrases, sentences, paragraphs, or documents. The relation for alignment and the granularity of the segments depend on the application for which alignment is being done. Alignment is used in many different applications of NLP : machine translation, dictionary building, summarization, information retrieval and many more.

Our alignment process focuses between text segments within a document, for instance paragraphs or sentences, and is based on the similarity between these text segments. The alignment of these text segments is a tedious work even for a corpus containing few hundred of text segments. For n number of text segments the total number of comparisons between them, to decide if an alignment exists or not, equals to :

$$\frac{n(n-1)}{2} \tag{1}$$

For instance, the corpus we use has 239 paragraphs, explained in section 3.1, for alignment which makes a total of 28,441 compar-

1. http ://projects.ldc.upenn.edu/TDT-Pilot/

isons. Usually, all of these comparisons are done manually by two or more annotators which is time consuming and therefore, we propose a method to reduce this effort by reducing the number of comparisons before giving them to the annotators.

2.1 Reducing the search space

Annotating is a tedious task searching for alignments through all the possible pairs of text unit. To reduce the amount of search we make the alignment process in two phases by breaking the problem into two parts. The first phase is the part in which the total combination of alignment pairs are reduced by selecting candidate alignment pairs such that the actual alignment pairs are the subset of this candidate alignment pairs.

$$ActualAlignmentPairs \subset CandidateAlignmentPairs \tag{2}$$

With these candidate alignments the annotators have a smaller set of paragraph pairs to work with and can be more effective as well as efficient. In the next phase, the annotators select the actual alignments from these candidate alignment pairs. This is possible because in comparable corpora there are many text segment pairs that are not similar and therefore, it is possible to filter these pairs of text segments.

Selecting the candidate alignment pairs is the first phase where, we select all the pairs of text segments that has the possibility of being similar and therefore aligned. The main objective of this phase is to reduce the size of the initial total amount of alignment pairs in such a way that the actual alignments are not missed out. This selection is done when a criteria is met. To follow the criteria, we first divide each text segments into entities. According to these entities the candidate alignment pairs are selected. The entities are listed and explained below :

– *Noun Entities* : These are parts of the text segments that represents the important nouns or noun phrases of the text segments. Importance depends on how much meaning does this entity bear to convey the meaning of the whole text segment.
– *Verb Entities* : These are the parts of the sentences that represents the main intransitive verbs of the text segments. Intransitive verbs are verbs that shows action of some sort but does not have a direct object (Loberger & Shoup, 2009). As the verb is related to one noun or noun phrase the importance of this verb tends to be higher for selecting similar text segments which will be evident when alignment is done.

Here is an example of two text segments that are compared using their entities :

Text Segment I :

William and Harry, with their father Prince Charles and their grandmother Queen Elizabeth, are thought likely to remain in seclusion at Balmoral Castle in Scotland until Saturday's ceremony.

Entities :
– William
– Harry
– Father Prince Charles
– grandmother Queen Elizabeth
– seclusion
– Balmoral Castle
– Scotland
– Saturday's ceremony

Similarly, we extract entities from the second text segment with which we want to compare the previous text segment. The second text segment is given below :

Text Segment II :

One is the state funeral, normally staged only for sovereigns, although the reigning king or queen, with the approval of Parliament, can order one for others. Churchill, Britain's prime minister during World War II, is one who received such treatment in 1965.

Entities :
– state funeral

– sovereigns
– reigning king or queen
– approval of Parliament
– Churchill
– Britain's prime minister
– World War II

Once this is done for both the text segments we select this alignment as one of the candidate alignments if the following condition is satisfied :

The concept of at least one entity should be common to the entity set of the text segments.

Comparing text segments I and II, we can see that we have a common element between these text segments which are the entities 'Queen Elizabeth' and 'king or queen'. These two elements are same obviously not because of common terms but because of the concept of 'queen' which is Queen Elizabeth. This concept can be easily known using the context in the paragraph text segments. This comparison of common entities in the text segments are easier to determine than to decide if these two text segments are similar or not and therefore can be done faster than the traditional method of directly finding similar pair of text segments.

As we can see that the text segments aren't similar with any logical definition of similarity, given that the text segments convey different information, yet these are selected because of our selection criteria. This criteria that we present will theoretically guarantee that the actual aligned text segment pairs will be present in the candidate alignments.

To make it clearer about the concept of the element here are some examples of concept of entities :

word	possible concepts
crashed	rammed into a wall, fatal impact
prince Charles	heir to the British throne
grief	sadness, mourn
high speed	121 mph, flying by

This method of selection is easy and could be done by a non professional annotator on different lengths of text segments. Once these candidate alignments are collected, they can be given to the annotators for annotating the alignments by finding the actual alignments. The number of candidate alignments will be less than the original combination of pair of text segments and therefore many annotators can work efficiently on the small set in less human hours. The next section describes how we created our gold corpus from the set of candidate alignments selected by our selection criteria.

2.2 Creation of Gold Corpus

We built the gold corpus annotated with aligned paragraphs to build and test algorithms for automatic alignment of paragraphs. To build the gold corpus, two annotators selected the actual alignment pairs from the candidate alignments which were selected as explained in the previous section 2.1. Two paragraphs are annotated as aligned in the gold corpus on the basis of similarity. To show how the annotators annotated the gold corpus we first define the term similarity.

Similarity is a difficult concept to define in general because the definition of this term depends on the application for which this measure is intended. Even with this difficulty, there are many similarity measures (Barron-Cedeno *et al.*, 2009) like cosine similarity measure, with which similar texts are measured but these measures do not define similarity, they rather assign a value of similarity. Here are some of the definitions of similarity between two texts :

1. Two sentences are similar if they contain at least one clause that expresses the same information. (Barzilay & Elhadad, 2003)
2. Two paragraphs are similar if they contain "common information". This was defined to be the case if the paragraphs referred to the same object and the object either (a) performed the same action in both paragraphs, or (b) was described in the same way in both paragraphs. (Hatzivassiloglou & Klavans, 2001) (Hatzivassiloglou *et al.*, 1999)
3. Two text are similar on the basis of these intuitions :(Lin, 1998)
   – **Intuition 1 :** The similarity between A and B is related to their commonality. The more commonality they share, the more similar they are.

- **Intuition 2 :** The similarity between A and B is related to the differences between them. The more differences they have, the less similar they are.
- **Intuition 3 :** The maximum similarity between A and B is reached when A and B are identical, no matter how much commonality they share.

All the definitions presented above focus on what is common between the text segments to call them similar. This focus on what is common is also the difference between them. Definition 1 and 2 states what should be common where as definition 3 gives no information about it and therefore, is the most general definition among them. The more general the definition is, the more difficult the annotation process becomes because of the different interpretation of the definition and in turn more disagreements between annotators. Definition 1 and 2 is difficult to apply to all the paragraphs that we see in our corpus. Definition 1 is specific to sentences and paragraphs with more than one sentence cannot be considered similar on the basis of the same information within clauses. Definition 2 considers similarity on the basis of objects and there may exist paragraphs for which the information about the objects alone do not represent the meaning of the paragraph as in the following paragraph :

French television said Diana was being pursued by paparazzi when the crash occurred, and French Interior Minister Jean-Pierre Chevenement said police were questioning seven photographers as part of a criminal investigation into the accident.

In this paragraph, the object, paparazzi, doesn't perform any action nor does the description of the paparazzi as photographers represents the paragraph.

We will define our similarity definition with the intuition 1 of definition 3 by defining the term commonality. We define the term 'commonality' in the definition on the basis of common sub-topic. The explication of sub-topic is also intuitive as we define sub-topic as the main ideas that the paragraph gives. Intuition 2 considers the differences being a basis of similarity but we only require how similar they are and the differences between paragraphs could be considered as how much the paragraphs are not similar and so we ignore this intuition. Intuition 3 is partially correct as identical paragraphs are definitely similar to it's maximum as they will share the same sub-topic but we ignore it because it is possible that two non-identical paragraphs may consist of the same sub-topic. A paragraph may have more than one sub-topic and for us a minimal of one common sub-topic would make the paragraph similar. Here is an example where we compare two paragraphs for similarity :

Paragraph I (PI) :

At Kensington Palace, the flowers covered an area estimated at 50 by 30 feet. There were more flowers as well at Harrods department store, which is owned by Dodi Fayed's father, billionaire Egyptian businessman Mohamed Fayed.

Paragraph II (PII) :

Mounds of flowers marked the sidewalk near one gate at Kensington Palace, where Diana resided, and along the main gates of the palace. There were flowers as well at Buckingham Palace, at St. James' Palace, at Harrod's department store, which is owned by Fayed's father, the Egyptian born business tycoon Mohamed Al Fayed. There were even flowers outside the gym where Diana regularly worked out.

The main ideas from Paragraph I are :

1. Flowers were present at Kensignton Palace
2. Flowers were present at Harrods department store

The main ideas from Paragraph II are :

1. Flowers were present at Kensington Palace
2. Flowers were placed at the palace
3. Flowers were placed at Buckingham Palace
4. Flowers were present at Harrods department store
5. Flowers were present at the gym

Once we have the main ideas that are present, we try to find at least one overlap between them. In paragraphs I and II we find the following overlaps between them :

PI.1 with PII.1
PI.2 with PII.4

Here is another example of finding the similarity between paragraphs which is less intuitive at the first glance : Paragraphs to compare are :

Paragraph III (PIII) :

Dodi Al Fayed's father, Harrods Department Store owner Mohammed Al Fayed, arrived here immediately after learning of his son's death.

Paragraph IV (PIV) :

Bernard Dartevelle, a lawyer for Mohamed Al Fayed, Dodi Fayed's wealthy businessman father and also the owner of the Hotel Ritz, said the revelation "changes absolutely nothing." He spoke of an "ambience of harassment" created around Diana and Fayed by the constant presence of paparazzi.

The main ideas from Paragraph III are :

– Dodi Al Fayed's father arrived here after learning his son's death.

The main ideas from Paragraph IV are :

– Bernard Dartevelle said the revelation "changes absolutely nothing."
– He spoke of an "ambience of harassment" created around Diana and Fayed by the constant presence of paparazzi.

In these two paragraphs, there is no common idea and therefore they are not selected for the actual alignment. In all of the four paragraphs in the examples given above, there is an information that are in common about Dodi Al Fayed's father but is not placed as the main idea because we believe these informations are present to support the main idea given by the paragraph and not the main idea itself. For this paper, we consider our similarity to be a binary relationship showing two paragraphs have at least one sub-topic in common or none. This binary measure can be easily changed into a continuous measure as the number of sub-topic present in the paragraphs can be counted.

3 Experiments and Results

3.1 Corpus

The corpus we used to run our experiments are taken from the Linguistic Data Consortium, LDC [2]. LDC is an organization that has a collection of a wide range of corpus for different purposes. We have selected the LDC's North American News Text Corpus [3] which is a monolingual corpus that consists of news articles from a spectrum of sources which includes New York Times ,Word Press, Associated Press, Washington post and some more. Among all these articles we selected 12 articles which were published within two consecutive days and which share the same topic to make a small monolingual comparable corpus. Our characteristics of comparability lies on the topic of the articles selected. The topic is the death of Princess Diana. These articles are from The Washington Post, Los Angeles Times, and New York Times. In these 12 articles, there are in total 239 paragraphs which will be aligned to each other on the basis of similarity.

3.2 Manual Alignment

We have manually aligned 28,441 paragraph pairs that have been selected from the corpus as explained in the previous section. The alignment process is of two parts, the first in which we select the candidate alignments and in the second part we select the actual

2. http ://www.ldc.upenn.edu/Catalog/byType.jsp
3. LDC Catalog number : LDC95T21

alignments from the candidate alignments. The first part of the alignment process was done by a single annotator to reduce the total inital alignments to only 3,416 candidate alignments. In this phase, the annotator can be flexible to decide if a candidate alignment pair is really helpful to be a candidate alignment pair or not. If this decision can be taken easily and with out doubt then the candidate alignments that is valid by our selection criteria can be ignored. This flexibility is possible because of the fact that some paragraphs may have some common concepts of the entities and yet not have anything in common other than that and as it is manually done the annotator can decide not to align them. This phase is easy and therefore fast to do. It took about 71 hours to find the set of 3,416 candidate alignments from the set of 28,441 paragraph pairs.

The second phase of finding the actual alignment was done by two annotators independently and any differences among these selection of alignments were discussed together and a decision was taken with reasoning. The alignment task took about 20 hours for both annotators and a total of 429 actual alignments were selected. The total time that took us to annotate our corpus was 91 hours. If we had directly tried to find the actual alignments without phase one, with an assumption that the time taken per paragraph pair (about 21 sec) is the same as in this second phase, it would take about 166 hours. The total time saved is 75 hours of work.

These actual alignments will be used as our gold corpus for building our algorithm for similarity. We used kappa statistics (Cohen, 1960) (Carletta, 1996) to evaluate our second phase annotation. Kappa statistics is defined as :

$$k = \frac{P_A - P_E}{1 - P_E} \tag{3}$$

where P_A is the probability of two annotators agreeing in practice and P_E is the expected probability of the two annotators agreeing. In our case P_A = 0.959, P_E = 0.780 and K = 0.813, indicating the agreement on annotations are significant.

3.3 Automatizing the Alignment

Our manual alignment method is still time consuming and difficult as manual effort has to be done so we tried to use some simple automatic methods to see how they do against the manual process. We tried stem overlap, word overlap and cosine similarity measures (Barron-Cedeno *et al.*, 2009) to find the actual alignments between similar paragraphs on the corpus we manually annotated. Before the experiments on each method we removed the stop words using a stop word list and except when using word overlap method we stemmed the remaining words using a snowball stemmer [4]. We used two types of weights for the cosine similarity measure, one the frequency of the stem within the paragraph, TF, and the second one was the TF-IDF, which is used in information retrieval (Salton & McGill, 1983). We considered a pair of paragraphs to be aligned if the threshold value was crossed. Table 1 gives the best results from these methods along with their threshold value.

Methods	Threshold	Aligned from the method	Actual Alignments included
Stem Overlap	> 0	15,276	**420**
Word Overlap	> 0	12,116	407
Cosine Similarity (weight as TF)	>0.025	14,989	**420**
Cosine Similarity (weight as TF*IDF)	>0.025	7,351	376

TABLE 1 – The table gives the number of actual alignments included in the candidate alignment which were selected by the different methods along with their threshold

These methods that have been used isn't enough to automatically select the actual alignments as the precision of these methods are very low with none of them reaching a recall of 1. The best result in terms of including the actual alignments was given by the stem overlap and cosine similarity measure, which uses TF as weights, with 420 of the actual alignments retrieved. The cosine similarity measure is based on stem overlap as shown in the equation 4

$$cos(p_1, p_2) = \frac{\Sigma_{s \in p_1 \cap p_2}(TF_{s,p_1}.TF_{s,p_2})}{\sqrt{\Sigma_{s \in p_1}(TF_{s,p_1})^2.\Sigma_{s \in p_2}(TF_{s,p_2})^2}} \tag{4}$$

where, s is the stem, p is the paragraph and TF is the stem frequency of that stem. As this method is based on stem overlap it won't do better than stem overlap in terms of the number of actual alignments included but is better in improving the precision.

4. http ://snowball.tartarus.org/

Even though these methods can't be used to find actual alignments we can use them to reduce the total initial paragraph pairs that have to be checked for candidate alignments. From the table 4 we can determine that the method which uses stem overlap and cosine similarity measure, which uses TF as weights, include 420 actual alignments and can be used to reduce the search space for finding candidate alignments if the few alignments that was not included can be ignored. Considering the time taken for the first phase, if we could ignore these undetected alignments we could save a considerable amount of time as the initial set of paragraphs are almost halved and so is the time taken to select the candidate alignments.

Some of the concepts of the alignments that were not captured by the one stem overlap method are presented below in table 2 and can be seen that some similar concepts like flower and bouquets could be shown they are similar using some knowledge source such as a dictionary while others would be a complex task.

Concept I	Concept II
100 miles per hour	flying by
Flower	Bouquets
spun into the wall	a tragic end
spun into the wall	crashed
following	pursuit
William,Charles, and Queen Elizabeth	Royal Family
causal	cause

TABLE 2 – Concepts that could not be found similar using stem overlap

Looking into the actual alignments that were captured by the method of stem overlap and the cosine similarity measure, which uses TF as weights, we saw that most of the detected alignment pairs was because of the overlap of the context of the concept rather than the concept itself which puts a question about how effective this method would be while using a small size of text segments. In the news paper corpus that we use, a large portion of paragraphs have a length of a sentence which indicates that the context of the same concept within a sentence is enough for finding the candidate alignments.

4 Conclusions and Future Work

The total number of actual aligned text segments in the gold corpus shows that only 1.5% of the 28,441 initially paired paragraphs are aligned and the effort to check the other 98.5% of the paired paragraphs is wasted in terms of the difference between the end number of actual alignments and the total number of initial paired paragraphs. Our manual alignment method can reduce this wasted effort and have saved us about 75 hours of work. This method of manual alignment by reducing the search space is better than existing methods of manual annotation as the annotators have less candidate alignment pairs to annotate. These candidate alignments are easier and faster to select than the actual similar pair of text segments because of the complex nature of the definition of similarity in terms of analyzing the text according to the definition.

Even though our method is easier, it still requires much effort to select the candidate alignments. We can further reduce this effort by reducing the original set of text segment pairs using stem overlap or cosine similarity measure for choosing the candidate alignments from which the original alignments are selected to make the gold corpus. The different methods we used also showed that simple method of stem or word overlap and even cosine similarity measure are not enough to capture text segment similarities but has given a view that cosine similarity measure increases the precision. The context is also an important part in finding the actual alignments as we see with the stem overlap method that some actual alignments are captured using the context which gives us motivation in trying to use the context, like co-occurrence, in a vector space model (Kaufmann, 2000). This hypothesis will be used in the future to make our automatic alignment algorithm.

Références

BARRON-CEDENO A., EISELT A. & ROSSO P. (2009). Monolingual text similarity measures : A comparison of models over wikipedia articles revisions. In *Proceedings of the ICON : 7th International Conference on NLP*, p. 29–38.

BARRÓN-CEDEÑO A., POTTHAST M., ROSSO P., STEIN B. & EISELT A. (2010). Corpus and Evaluation Measures for Automatic Plagiarism Detection. In N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER

& D. TAPIAS, Eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 10)* : European Language Resources Association (ELRA).

BARZILAY R. & ELHADAD N. (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the conference on Empirical methods in NLP*, p. 203–212.

CARLETTA J. (1996). Assessing agreement on classification tasks : The kappa statistic. In *Computational Linguistics*, p. 249–254.

COHEN J. (1960). A coefficient of agreement for nominal scales. In *Educational and Psyhological Measurement*, p. 37–46.

GAIZAUSKAS R., FOSTER J., WILKS Y., ARUNDEL J., CLOUGH P. & PIAO S. (2001). The meter corpus : A corpus for analysing journalistic text reuse. p. 214–223.

HATZIVASSILOGLOU V. & KLAVANS J. L. (2001). Simfinder : A flexible clustering tool for summarization. In *Proceedings of NAACL Workshop of Automati Summarization*, p. 203–212.

HATZIVASSILOGLOU V., KLAVANS J. L. & ESKIN E. (1999). Detecting text similarity over short passages : exploring linguistic feature combinations via machine learning. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, p. 203–212.

INDURKHYA N. & DAMERAU F. J. (2010). *Handbook of natural language processing*. Taylor and Francis.

KAUFMANN S. (2000). Second-order cohesion. *Computational Intelligence*, **16**, 511–524.

LIN D. (1998). An information-theoretic definition of similarity. In *ICML*, p. 296–304.

LOBERGER G. & SHOUP K. (2009). *Websters New World English Grammar Handbook*. Wiley, Hoboken.

MAIA B. (2003). What are comparable corpora ? In *Proceedings of pre-conference workshop Multilingual Corpora : Linguistic Requirements and Technical perspectives, at Corpus Linguistics*, p. 27–34 : Lancaster U.K.

SALTON G. & MCGILL M. J. (1983). *Introduction to Modern Informational Retrieval*. McGraw-Hill.

Index