TALN 2011

RECITAL 2011

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

Résumé des articles de la 18e conférence
sur le

Traitement Automatique des Langues Naturelles

Résumé des articles de la 15e

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

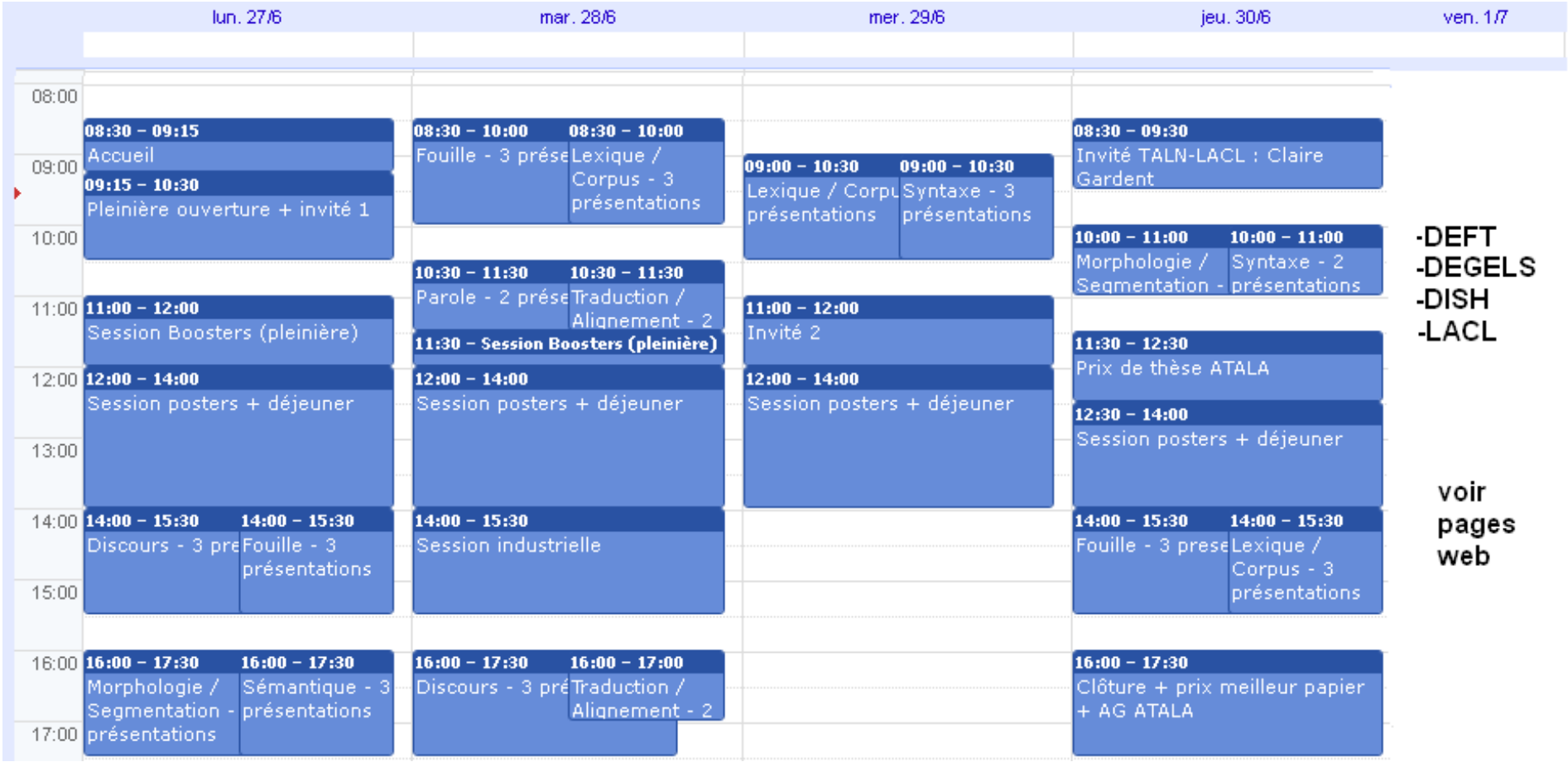
---

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 0
(Programme et résumés)



-DEFT
-DEGELS
-DISH
-LACL

voir pages web

Table des matières

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

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.

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

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

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.

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

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.

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.

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.

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

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

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.

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.

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

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

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

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

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

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

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

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

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

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

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.

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

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

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*)

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.

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.

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).

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,

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*

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.

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.

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/

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.

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.

<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

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

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.

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

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"

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`

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

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

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

TALN 2011, Montpellier, 27 *juin* –1er *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.

Evaluation de la détection des émotions, des opinions ou des sentiments : dictature de la majorité ou respect de la diversité d'opinions ?

Jean-Yves Antoine[1], Marc Le Tallec[1], Jeanne Villaneau[2]

(1) Université François Rabelais de Tours, LI, 37000 Blois
(2) Université Européenne de Bretagne, VALORIA, 56100 Lorient
Jean-Yves.Antoine@univ-tours.fr, Marc.Le-Tallec@univ-tours.fr, Jeanne.Villaneau@univ-ubs.fr

Résumé - Détection d'émotion, fouille d'opinion et analyse des sentiments sont généralement évalués par comparaison des réponses du système concerné par rapport à celles contenues dans un corpus de référence. Les questions posées dans cet article concernent à la fois la définition de la référence et la fiabilité des métriques les plus fréquemment utilisées pour cette comparaison. Les expérimentations menées pour évaluer le système de détection d'émotions *EmoLogus* servent de base de réflexion pour ces deux problèmes. L'analyse des résultats d'*EmoLogus* et la comparaison entre les différentes métriques remettent en cause le choix du vote majoritaire comme référence. Par ailleurs elles montrent également la nécessité de recourir à des outils statistiques plus évolués que ceux généralement utilisés pour obtenir des évaluations fiables de systèmes qui travaillent sur des données intrinsèquement subjectives et incertaines.

Abstract - Emotion detection, opinion identification and sentiment analysis are generally assessed by means of the comparison of a reference corpus with the answers of the system. This paper addresses the problem of the definition of the reference and the reliability of the metrics which are commonly used for this comparison. We present some experiments led with *EmoLogus*, a system of emotion detection, to investigate these two problems. A detailed analysis of the quantitative results obtained by *EmoLogus* on various metrics questions the choice of a majority vote among several human judgments to build a reference. Besides, it shows the necessity of using more sophisticated statistical tools to obtain a reliable evaluation of such systems which are working on intrinsically subjective and uncertain data.

Mots-clés : Détection d'émotion, analyse de sentiments, fouille d'opinion ; Evaluation : métrique d'évaluation, constitution de référence, analyse statistique des résultats.
Keywords: Detection of emotion, sentiment analysis, opinion mining, Evaluation: objective measures, test reference, statistical analysis of the results.

1 Evaluation en détection des émotions / opinions / sentiments

La détection d'émotions, la fouille d'opinion ou l'analyse des sentiments sont des tâches très proches qui consistent à trouver et catégoriser dans des flux langagiers oraux ou écrits des passages porteurs d'un état émotionnel ou traduisant un jugement. La granularité de la détection est variable suivant l'application : il peut s'agir d'un document ou d'un discours complet, d'un paragraphe, d'une phrase (qui peut être spécifiquement un titre, par exemple) ou d'un tour de parole dans le cas du dialogue oral homme-machine. Le grain de catégorisation recherché peut également varier d'une tâche à l'autre. On peut ainsi ne considérer que trois classes principales (valence positive, négative ou neutre) ou rechercher une caractérisation plus fine sous la forme de modalités correspondant aux émotions principales définies en psychologie : colère, joie, dégout, peur, surprise, tristesse et émotion neutre (Ekman 1999).

A Collocation-Driven Approach to Text Summarization

Violeta Seretan
Institute for Language, Cognition and Computation
Human Communication Research Centre, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, United Kingdom
violeta.seretan@gmail.com

Résumé. Dans cet article, nous décrivons une nouvelle approche pour la création de résumés extractifs – tâche qui consiste à créer automatiquement un résumé pour un document en sélectionnant un sous-ensemble de ses phrases – qui exploite des informations collocationnelles spécifiques à un domaine, acquises préalablement à partir d'un corpus de développement. Un extracteur de collocations fondé sur l'analyse syntaxique est utilisé afin d'inférer un modèle de contenu qui est ensuite appliqué au document à résumer. Cette approche a été utilisée pour la création des versions simples pour les articles de Wikipedia en anglais, dans le cadre d'un projet visant la création automatique d'articles simplifiées, similaires aux articles recensées dans Simple English Wikipedia. Une évaluation du système développé reste encore à faire. Toutefois, les résultats préalables obtenus pour les articles sur des villes montrent le potentiel de cette approche guidée par collocations pour la sélection des phrases pertinentes.

Abstract. We present a novel approach to extractive summarization – the task of producing an abstract for an input document by selecting a subset of the original sentences – which relies on domain-specific collocation information automatically acquired from a development corpus. A syntax-based collocation extractor is used to infer a content template and then to match this template against the document to summarize. The approach has been applied to generate simplified versions of Wikipedia articles in English, as part of a larger project on automatically generating Simple English Wikipedia articles starting from their standard counterpart. An evaluation of the developed system has yet to be performed; nonetheless, the preliminary results obtained in summarizing Wikipedia articles on cities already indicated the potential of our collocation-driven method to select relevant sentences.

Mots-clés : résumé de texte automatique, résumé extractif, statistiques de co-occurrence, collocations, analyse syntaxique, Wikipedia.

Keywords: text summarization, extractive summarization, co-occurrence statistics, collocations, syntactic parsing, Wikipedia.

Quel apport des unités polylexicales dans une formule de lisibilité pour le français langue étrangère

Thomas François[1, 2, 3] Patrick Watrin[2, 3]
(1) Aspirant FNRS
(2) Centre de traitement automatique du langage (CENTAL), UCLouvain
(3) Institut Langage et Communication (IL&C), UCLouvain
thomas.francois@uclouvain.be, patrick.watrin@uclouvain.be

Résumé. Cette étude envisage l'emploi des unités polylexicales (UPs) comme prédicteurs dans une formule de lisibilité pour le français langue étrangère. À l'aide d'un extracteur d'UPs combinant une approche statistique à un filtre linguistique, nous définissons six variables qui prennent en compte la densité et la probabilité des UPs nominales, mais aussi leur structure interne. Nos expérimentations concluent à un faible pouvoir prédictif de ces six variables et révèlent qu'une simple approche basée sur la probabilité moyenne des n-grammes des textes est plus efficace.

Abstract. This study considers the use of multi-words expressions (MWEs) as predictors for a readability formula for French as a foreign language. Using a MWEs extractor combining a statistical approach with a linguistic filter, we define six variables. These take into account the density and the probability of MWEs, but also their internal structure. Our experiments conclude that the predictive power of these six variables is low. Moreover, we show that a simple approach based on the average probability of n-grams is a more effective predictor.

Mots-clés : Lisibilité du FLE, unités polylexicales nominales, modèles N-grammes.

Keywords: Readability of FFL, nominal MWEs, N-grams models.

Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées

Frédéric Béchet[1], Benoît Sagot[2], Rosa Stern[2,3]
(1) Aix Marseille Université, LIF-CNRS, route de Luminy, Marseille
(2) Alpage, INRIA & Univ. Paris 7, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
(3) Agence France-Presse – Medialab, 2 place de la Bourse, 75002 Paris, France
frederic.bechet@lif.univ-mrs.fr, benoit.sagot@inria.fr, rosa.stern@afp.com

Résumé. La détection et le typage des entités nommées sont des tâches pour lesquelles ont été développés à la fois des systèmes symboliques et probabilistes. Nous présentons les résultats d'une expérience visant à faire interagir le système à base de règles NP, développé sur des corpus provenant de l'AFP, intégrant la base d'entités Aleda et qui a une bonne précision, et le système LIANE, entraîné sur des transcriptions de l'oral provenant du corpus ESTER et qui a un bon rappel. Nous montrons qu'on peut adapter à un nouveau type de corpus, de manière non supervisée, un système probabiliste tel que LIANE grâce à des corpus volumineux annotés automatiquement par NP. Cette adaptation ne nécessite aucune annotation manuelle supplémentaire et illustre la complémentarité des méthodes numériques et symboliques pour la résolution de tâches linguistiques.

Abstract. Named entity recognition and typing is achieved both by symbolic and probabilistic systems. We report on an experiment for making the rule-based system NP, a high-precision system developed on AFP news corpora and relies on the Aleda named entity database, interact with LIANE, a high-recall probabilistic system trained on oral transcriptions from the ESTER corpus. We show that a probabilistic system such as LIANE can be adapted to a new type of corpus in a non-supervized way thanks to large-scale corpora automatically annotated by NP. This adaptation does not require any additional manual anotation and illustrates the complementarity between numeric and symbolic techniques for tackling linguistic tasks.

Mots-clés : Détection d'entités nommées, adaptation à un nouveau domaine, coopération entre approches probabilistes et symboliques.

Keywords: Named entity recognition, domain adaptation, cooperation between probabilistic and symbolic approaches.

1 Introduction

La reconnaissance d'entités nommées est une des tâches les plus étudiées du domaine du traitement automatique des langues. Reconnaître dans du texte ou de la transcription de parole les mentions d'entités nommées reste un préalable nécessaire avant toute tâche plus complexe telle que l'analyse syntaxique, l'analyse sémantique ou l'extraction d'informations. Ainsi, la tâche de reconnaissance d'entités nommées fait l'objet de nombreuses campagnes d'évaluation depuis plus d'une vingtaine d'années, dont les premières ont été les campagnes MUC (Message Understanding Conference). Ces campagnes ont donné lieu à la construction de corpus de référence, notamment pour l'anglais, le chinois, l'espagnol ou le japonais. Pour le français, on peut citer notamment l'évaluation menée dans le cadre de la campagne ESTER sur des transcriptions de nouvelles, qui a donné naissance à un corpus français annoté en entités nommées selon des directives assez différentes de celles des campagnes MUC, notamment pour les emplois polysémiques et métaphoriques (Galliano *et al.*, 2009). Pour une discussion plus précise de ces questions et des différents types d'ambiguïtés rencontrées en reconnaissance des entités nommées, on pourra se reporter à (Béchet, 2011).

Toutes les campagnes d'évaluation ont montré que la tâche de reconnaissance d'entités nommées peut être traitée efficacement aussi bien avec des systèmes symboliques qu'avec des systèmes probabilistes. Elles ont également montré que les systèmes symboliques couplés à de très grands lexiques donnent de meilleurs résultats que les

Création de clusters sémantiques dans des familles morphologiques à partir du TLFi

Nuria Gala[1] Nabil Hathout[2] Alexis Nasr[1] Véronique Rey[3] Selja Seppälä[1]
(1) LIF-TALEP, 163, Av. de Luminy case 901, 13288 Marseille Cedex 9
(2) CLLE-ERSS, 5, allées Antonio Machado, 31058 Toulouse Cedex 9
(3) EHESS, 2, rue de la Charité, 13002 Marseille
{nuria.gala, alexis.nasr, selja.seppala}@lif.univ-mrs.fr, nabil.hathout@univ-tlse2.fr,
veronique.rey-lafay@univmed.fr

Résumé. La constitution de ressources linguistiques est une tâche longue et coûteuse. C'est notamment le cas pour les ressources morphologiques. Ces ressources décrivent de façon approfondie et explicite l'organisation morphologique du lexique complétée d'informations sémantiques exploitables dans le domaine du TAL. Le travail que nous présentons dans cet article s'inscrit dans cette perspective et, plus particulièrement, dans l'optique d'affiner une ressource existante en s'appuyant sur des informations sémantiques obtenues automatiquement. Notre objectif est de caractériser sémantiquement des familles morpho-phonologiques (des mots partageant une même racine et une continuité de sens). Pour ce faire, nous avons utilisé des informations extraites du TLFi annoté morpho-syntaxiquement. Les premiers résultats de ce travail seront analysés et discutés.

Abstract. Building lexical resources is a time-consuming and expensive task, mainly when it comes to morphological lexicons. Such resources describe in depth and explicitly the morphological organization of the lexicon, completed with semantic information to be used in NLP applications. The work we present here goes on such direction, and especially, on refining an existing resource with automatically acquired semantic information. Our goal is to semantically characterize morpho-phonological families (words sharing a same base form and semantic continuity). To this end, we have used data from the TLFi which has been morpho-syntactically annotated. The first results of such a task will be analyzed and discussed.

Mots-clés : Ressources lexicales, familles morphologiques, clusters sémantiques, mesure de Lesk.

Keywords: Lexical resources, morphological families, semantic clusters, Lesk measure.

1 Introduction

Les ressources linguistiques sont indispensables aussi bien dans une perspective de traitement automatique de la langue que dans le cadre d'une utilisation humaine (apprentissage des langues, thérapie orthophoniste, etc.). Une des problématiques saillantes dans ce domaine concerne leur constitution : elle s'avère longue et coûteuse, spécialement lorsqu'on vise des informations fines comme la description de l'organisation dérivationnelle du lexique. La création de ce type de ressource peut être réalisée soit manuellement comme (Gala & Rey, 2008), soit à partir d'informations morphologiques dérivationnelles acquises automatiquement. Pour le français, citons le projet MorTAL (Hathout *et al.*, 2002; Dal *et al.*, 2004), les travaux en informatique médicale qui portent sur l'apprentissage de relations morphologiques en corpus spécialisés (Zweigenbaum *et al.*, 2003; Langlais *et al.*, 2009). Signalons également que plusieurs travaux portent sur l'acquisition automatique de familles morphologiques du français comme (Gaussier, 1999; Bernhard, 2007; Hathout, 2009; Lavallée & Langlais, 2010). Notre objectif à terme est de créer une ressource comparable à la base CELEX [1] (Baayen *et al.*, 1995), c'est-à-dire, un lexique avec des informations morpho-sémantiques fines pour au moins 50 000 mots du français.

Dans cet article, nous présentons une méthodologie de création de sous-familles ou clusters sémantiques dans des

1. CELEX décrit la phonologie et la morphologie de trois langues germaniques, l'anglais, l'allemand et le néerlandais. Elle contient notamment une analyse morphologique dérivationnelle fine pour l'ensemble des lexèmes : 52 447 entrées pour l'anglais, 51 728 pour l'allemand et 124 136 pour le néerlandais

Génération automatique de questions à partir de textes en français *

Louis de Viron[1, 3] Delphine Bernhard[1] Véronique Moriceau[1, 2] Xavier Tannier[1, 2]
(1) LIMSI-CNRS, 91403 Orsay, France
(2) Université Paris Sud, 91405 Orsay, France
(3) Université Catholique de Louvain, Belgique
louis.deviron@student.uclouvain.be, {delphine.bernhard, moriceau, xtannier}@limsi.fr

Résumé. Nous présentons dans cet article un générateur automatique de questions pour le français. Le système de génération procède par transformation de phrases déclaratives en interrogatives et se base sur une analyse syntaxique préalable de la phrase de base. Nous détaillons les différents types de questions générées. Nous présentons également une évaluation de l'outil, qui démontre que 41 % des questions générées par le système sont parfaitement bien formées.

Abstract. In this article, we present an automatic question generation system for French. The system proceeds by transforming declarative sentences into interrogative sentences, based on a preliminary syntactic analysis of the base sentence. We detail the different types of questions generated. We also present an evaluation of the tool, which shows that 41 % of the questions generated by the system are perfectly well-formed.

Mots-clés : génération de questions, analyse syntaxique, transformation syntaxique.

Keywords : question generation, syntactic analysis, syntactic transformation.

1 Introduction

La génération automatique de questions à partir de textes consiste à transformer automatiquement une phrase déclarative en une phrase interrogative. Il s'agit d'une tâche complexe, qui mobilise nombre de ressources et outils du TAL tels que la détection d'entités nommées, l'analyse syntaxique, la résolution d'anaphores et la simplification de phrases. Les applications de la génération automatique de questions sont par ailleurs variées : création de tests et de questionnaires à choix multiples pour l'aide à l'apprentissage, systèmes de dialogue homme-machine ou de questions-réponses interactifs. Si le domaine a été largement traité dans le monde anglophone, et ce depuis longtemps (on retrouve un article de Wolfe (1976) qui évoque déjà le sujet), il n'existe pas, à notre connaissance, de travaux équivalents pour le français. Nous présentons dans cet article un système de génération automatique de questions pour le français qui procède par transformation d'arbre syntaxique à partir de l'analyse fournie par XIP (Ait-Mokhtar *et al.*, 2002) afin de produire à la fois des questions factuelles et des questions fermées.

2 État de l'art

La génération automatique de questions trouve son application dans deux domaines principaux :
– Les systèmes de dialogue à partir de textes expositifs (Prendinger *et al.*, 2007 ; Piwek & Stoyanchev, 2010). Ces dialogues peuvent être présentés comme des textes ou encore à l'aide de personnages virtuels.
– Les applications éducatives : génération de questions ouvertes pour la compréhension de texte (Wolfe, 1976 ; Gates, 2008) ou questions à choix multiple (Mitkov *et al.*, 2006).

Les systèmes de génération de questions développés pour l'anglais procèdent généralement selon le schéma suivant : (i) analyse morphosyntaxique et/ou syntaxique du texte source (ii) identification du syntagme cible sur lequel portera la question, (iii) déplacement du syntagme cible, (iv) remplacement du syntagme cible par le mot interrogatif approprié, (v) inversion et accord sujet-verbe, (vi) post-traitement de la question pour générer une

∗. Ces travaux ont été partiellement financés par OSEO dans le cadre du programme QUAERO.

Sélection de réponses à des questions dans un corpus Web par validation

A. Grappy[1, 2], B. Grau[1, 3], M.-H. Falco[1, 2], A.-L. Ligozat[1, 3], I. Robba[1, 4], A. Vilnat[1, 2]
(1) LIMSI-CNRS
(2) Université Paris 11
(3) ENSIIE
(4) UVSQ
prenom.nom@limsi.fr

Résumé. Les systèmes de questions réponses recherchent la réponse à une question posée en langue naturelle dans un ensemble de documents. Les collections Web diffèrent des articles de journaux de par leurs structures et leur style. Pour tenir compte de ces spécificités nous avons développé un système fondé sur une approche robuste de validation où des réponses candidates sont extraites à partir de courts passages textuels puis ordonnées par apprentissage. Les résultats montrent une amélioration du MRR (Mean Reciprocal Rank) de 48% par rapport à la baseline.

Abstract. Question answering systems look for the answer of a question given in natural language in a large collection of documents. Web documents have a structure and a style different from those of newspaper articles. We developed a QA system based on an answer validation process able to handle Web specificity. Large number of candidate answers are extracted from short passages in order to be validated according to question and passage characteristics. The validation module is based on a machine learning approach. We show that our system outperforms a baseline by up to 48% in MRR (Mean Reciprocal Rank).

Mots-clés : systèmes de questions réponses ; validation de réponses ; analyse de documents Web.

Keywords: question-answering system ; answer validation ; Web document analysis .

1 Introduction

La recherche d'informations précises dans des textes, en réponse à des questions posées en langue naturelle, constitue un domaine largement étudié depuis la première évaluation de systèmes de réponses à des questions (SQR dans la suite) lancée à TREC en 1998 (*Q&A track*). Les meilleurs systèmes (Hickl *et al.*, 2006; Bouma *et al.*, 2005; Laurent *et al.*, 2010) utilisent des connaissances et des processus avancés de TAL notamment des analyseurs syntaxiques. Ces connaissances et processus interviennent notamment lors de la phase de sélection de passages pertinents et d'extraction de réponses, qui ont fait l'objet d'études spécifiques.

L'ordonnancement de réponses ou de passages consiste à ordonner les différentes réponses extraites afin d'obtenir la meilleure réponse en première position. Là aussi, les meilleures approches se fondent sur des correspondances syntaxiques ou sémantiques entre les passages (souvent constitués d'une phrase) et la question, correspondances obtenues par calcul de similarité entre arbres syntaxiques (Kouylekov *et al.*, 2006) ou en tenant compte de chemins de dépendances communs (Cui *et al.*, 2005).

Ces différents systèmes obtiennent de bons résultats sur des documents issus d'articles de journaux, mais ne peuvent être appliqués en l'état sur des collections provenant du Web, comme celle constituée dans le cadre du projet Quæro[1] pour évaluer les SQR. Pour le français, les SQR participants ont trouvé entre 27 % et 50 % des réponses en 2009 (Quintard *et al.*, 2010) après adaptation alors que le meilleur système en obtenait 69% lors de l'évaluation CLEF 2006 sur des articles de journaux (Laurent *et al.*, 2010). Ces difficultés sont dues entre autres aux spécificités des documents Web très souvent composés de tableaux, de listes ou de menus qui mettent en défaut les analyses syntaxiques une fois le texte extrait des pages.

[1]http ://www.quaero.org - Quæro est un programme financé par OSEO

Filtrage de relations pour l'extraction d'information non supervisée

Wei Wang[1] Romaric Besançon[1] Olivier Ferret[1] Brigitte Grau[2]
(1) CEA, LIST, 18 route du Panorama, BP 6, 92265 Fontenay-aux-Roses
(2) LIMSI, UPR-3251 CNRS-DR4, Bat. 508, BP 133, 91403 Orsay Cedex
wei.wang@cea.fr, romaric.besancon@cea.fr, olivier.ferret@cea.fr, brigitte.grau@limsi.fr

Résumé. Le domaine de l'extraction d'information s'est récemment développé en limitant les contraintes sur la définition des informations à extraire, ouvrant la voie à des applications de veille plus ouvertes. Dans ce contexte de l'extraction d'information non supervisée, nous nous intéressons à l'identification et la caractérisation de nouvelles relations entre des types d'entités fixés. Un des défis de cette tâche est de faire face à la masse importante de candidats pour ces relations lorsque l'on considère des corpus de grande taille. Nous présentons dans cet article une approche pour le filtrage des relations combinant méthode heuristique et méthode par apprentissage. Nous évaluons ce filtrage de manière intrinsèque et par son impact sur un regroupement sémantique des relations.

Abstract. Information Extraction have recently been extended to new areas, by loosening the constraints on the strict definition of the information extracted, thus allowing to design more open information extraction systems. In this new domain of unsupervised information extraction, we focus on the task of extracting and characterizing new relations between a given set of entity types. One of the challenges of this task is to deal with the large amount of candidate relations when extracting them from a large corpus. We propose in this paper an approach for filtering such candidate relations, based on heuristic and machine learning methods. We present an evaluation of this filtering phase and an evaluation of the impact of the filtering on the semantic clustering of relations.

Mots-clés : Extraction d'information non supervisée, filtrage, apprentissage automatique, clustering.

Keywords: Unsupervised information extraction, filtering, machine learning, clustering.

1 Introduction[1]

Les années récentes ont vu se développer de nouveaux paradigmes dans le domaine de l'extraction d'information (EI), parmi lesquels la notion d'EI non supervisée. Cette approche prend comme point de départ des entités ou des types d'entités et se fixe comme objectif de mettre en évidence les relations intervenant entre ces entités, sans connaissance *a priori* de leur type. Cette mise en évidence est éventuellement suivie d'un regroupement de ces relations en fonction de leurs similarités pour en faire la synthèse. Les travaux effectués dans ce champ de recherche s'envisagent selon trois points de vue. Le premier est l'acquisition de connaissances, que ce soit des connaissances sur le monde collectées à vaste échelle à partir du Web, comme avec le concept d'*Open Information Extraction* développé dans (Banko *et al.*, 2007), ou dans des domaines plus spécialisés, comme le domaine biologique, où cette extraction est le moyen d'ajouter de nouveaux types de relations entre entités à une ontologie existante (Ciaramita *et al.*, 2005). Le deuxième se situe dans le cadre d'applications d'EI, où ce type d'approche correspond à la volonté d'offrir aux utilisateurs des modes d'extraction de l'information plus souples et plus ouverts quant à la spécification de leur besoin informationnel. L'approche *On-demand information extraction* (Sekine, 2006), préfigurée dans (Hasegawa *et al.*, 2004) et concrétisée par les travaux sur la *Preemptive Information Extraction* (Shinyama & Sekine, 2006), vise ainsi à induire l'équivalent d'un *template* à partir d'un ensemble de documents représentatifs des informations à extraire, obtenus par le biais d'un moteur de recherche, par le regroupement des relations qui en sont extraites (Rosenfeld & Feldman, 2007). Enfin, l'EI non supervisée peut aussi servir à compléter l'EI supervisée, qui dépend de corpus annotés qui ne sont généralement pas de grande taille, étant donné la complexité des tâches considérées. Les résultats d'une approche non supervisée peuvent alors être utilisés pour élargir la couvertures des modèles appris (Banko & Etzioni, 2008; González & Turmo, 2009).

Dans cet article, nous nous plaçons dans le cadre du deuxième point de vue exposé ci-dessus, celui d'une extrac-

[1]Ce travail a été partiellement réalisé dans le cadre du projet FILTRAR-S soutenu par le programme CSOSG 2008 de l'ANR.

Un lexique pondéré des noms d'événements en français

Béatrice Arnulphy[1, 2] Xavier Tannier[1, 2] Anne Vilnat[1, 2]
(1) Univ. Paris-Sud 11, 91405 Orsay
(2) LIMSI-CNRS, 91403 Orsay
prenom.nom@limsi.fr

Résumé. Cet article décrit une étude sur l'annotation automatique des noms d'événements dans les textes en français. Plusieurs lexiques existants sont utilisés, ainsi que des règles syntaxiques d'extraction, et un lexique composé de façon automatique, permettant de fournir une valeur sur le niveau d'ambiguïté du mot en tant qu'événement. Cette nouvelle information permettrait d'aider à la désambiguïsation des noms d'événements en contexte[1].

Abstract. This article describes a study on automatic extraction of event nominals in French texts. Some existing lexicons are used, as well as some syntactic extraction rules, and a new, automatically built lexicon is presented. This lexicon gives a value concerning the level of ambiguity of each word as an event.

Mots-clés : extraction d'information, événements nominaux, lexiques.

Keywords: information extraction, nominal events, lexicons.

1 Introduction

La plupart des événements dans la langue est exprimée par les verbes et les noms. La forme verbale a été largement traitée, dans une approche formelle notamment par Vendler (1967) ou encore en traitement automatique des langues par le biais de TimeML (Pustejovsky *et al.*, 2005). Si les événements verbaux sont plus nombreux, plus simples à identifier et à lier aux autres informations temporelles, ils expriment souvent des événements plus communs et moins pertinents, tandis que la nominalisation d'un événement indique souvent son importance.

Les événements nominaux peuvent être construits de trois manières différentes. Certains sont construits à partir de noms déverbaux (*la **fête** de la musique* ou *l'**adoption** d'une réglementation*), sachant qu'ils peuvent désigner l'événement ou le résultat de l'action indiquée par le verbe dont il est issu (*construction*). D'autres sont formés à partir de noms autres que déverbaux et qui décrivent intrinsèquement des événements (*le **festival** de Cannes* ou *le **match** PSG-OM*), ces mots pouvant être ambigus (*le **salon** du livre*). Enfin, des syntagmes nominaux (SN) sans valeur événementielle peuvent par le résultat d'une métonymie, en contexte, référer à l'événement lié par exemple à un lieu (***Tchernobyl***), une date (*le **11 septembre***) ou l'objet d'une affaire (*les **frégates** de Taïwan*). Au vu de ces trois types d'événements, on se rend compte que le recours au lexique est nécessaire, mais pas suffisant ; une désambiguïsation par le contexte est indispensable.

L'événement est ce qui survient, le changement d'état opéré. Il peut être récurrent ou unique, prévu ou non, durer ou être instantané, se produire dans le passé, le présent ou le futur. Nous présentons une étude préalable détaillée des noms qui les caractérisent, par l'utilisation d'un corpus manuellement annoté. Après un bref état de l'art du domaine (section 2), nous présenterons les ressources dont nous disposons (section 3) et l'étude en vue de l'extraction des noms d'événements (section 4).

2 État de l'art

Quelques définitions des événements ont été proposées en philosophie, histoire, linguistique ou encore journalisme. Ces deux dernières disciplines nous intéressent tout particulièrement, parce que nous travaillons sur des

[1]Ce travail a été partiellement financé par OSEO dans le cadre du programme Quaero.

Alignement automatique pour la compréhension littérale de l'oral par approche segmentale

Stéphane Huet et Fabrice Lefèvre
Université d'Avignon, LIA-CERI, France
{stephane.huet,fabrice.lefevre}@univ-avignon.fr

Résumé. Les approches statistiques les plus performantes actuellement pour la compréhension automatique du langage naturel nécessitent une annotation segmentale des données d'entraînement. Nous étudions dans cet article une alternative permettant d'obtenir de façon non-supervisée un alignement segmental d'unités conceptuelles sur les mots. L'impact de l'alignement automatique sur les performances du système de compréhension est évalué sur une tâche de dialogue oral.

Abstract. Most recent efficient statistical approaches for language understanding require a segmental annotation of the training data. In this paper we study an alternative that obtains a segmental alignment of conceptual units with words in an unsupervised way. The impact of the automatic alignment on the understanding system performance is evaluated on a spoken dialogue task.

Mots-clés : Alignement non-supervisé, compréhension de la parole.

Keywords: Unsupervised alignment, spoken language understanding.

1 Introduction

Une des toutes premières étapes pour construire un système de compréhension de l'oral pour les systèmes de dialogue est l'extraction de concepts littéraux à partir d'une séquence de mots issue d'un système de reconnaissance de la parole. Pour résoudre ce problème d'étiquetage en concepts, un certain nombre de techniques sont disponibles. Ces techniques reposent sur des modèles classiques maintenant, qui peuvent être génératifs ou discriminants, parmi lesquels on peut citer : les modèles de Markov cachés, les transducteurs à états finis, les modèles de Markov à entropie maximale, les machines à vecteurs supports, les réseaux bayésiens dynamiques (*Dynamic Bayesian Networks*, DBN) ou encore les champs de Markov conditionnels (*Conditional Markov Random Fields*, CRF (Lafferty *et al.*, 2001)). Dans (Hahn *et al.*, 2010), il est montré que les CRF permettent d'obtenir les meilleures performances sur la tâche MEDIA (Bonneau Maynard *et al.*, 2008) en français, mais aussi sur deux corpus comparables en italien et en polonais. De même, la robustesse des CRF a pu être montrée en observant ses résultats sur la compréhension de transcriptions manuelles et automatiques.

Dans beaucoup d'approches, l'interprétation littérale se contente d'une relation lexique-concept ; c'est ainsi le cas du système PHOENIX (Ward, 1991) basé sur la détection de mots-clefs. L'approche segmentale fait une analyse plus fine en considérant la phrase comme une séquence de segments lors de son interprétation. Elle permet alors de relier correctement les différents niveaux d'analyse : lexicaux, syntaxiques et sémantiques. Toutefois, afin de simplifier la mise en œuvre, les segments ont été définis spécifiquement pour l'annotation conceptuelle et n'ont pas de relation imposée avec les unités syntaxiques (*chunks*, groupes syntaxiques...). Une autre raison est que l'objectif étant d'utiliser le module d'interprétation au sein de systèmes de dialogue oral, les données qui sont ici traitées sont fortement bruitées (langage naturel très spontané et agrammatical, erreurs dues à la reconnaissance automatique de la parole), ce qui perturbe fortement les analyseurs syntaxiques.

L'approche segmentale présente aussi l'intérêt de pouvoir découpler la détection d'une unité conceptuelle de l'estimation de sa valeur. La valeur correspond à la normalisation de la forme de surface associée au concept ; par exemple si au concept `temps-depart` est associé le segment « pas avant 11h », sa valeur est « matin ». De même pour « entre 8h et 12h » ou « dans la matinée ». L'estimation de la valeur nécessite donc un ancrage des concepts sur les mots de la phrase. Il est alors possible de traiter le problème de normalisation à partir de règles sous formes

Ajout d'informations contextuelles pour la recherche de passages au sein de Wikipédia

Romain Deveaud Eric SanJuan Patrice Bellot
LIA - Université d'Avignon
339, chemin des Meinajariès Agroparc BP 91228
84 911 Avignon Cedex 9
{romain.deveaud,eric.sanjuan,patrice.bellot}@univ-avignon.fr

Résumé. La recherche de passages consiste à extraire uniquement des passages pertinents par rapport à une requête utilisateur plutôt qu'un ensemble de documents entiers. Cette récupération de passages est souvent handicapée par le manque d'informations complémentaires concernant le contexte de la recherche initiée par l'utilisateur. Des études montrent que l'ajout d'informations contextuelles par l'utilisateur peut améliorer les performances des systèmes de recherche de passages. Nous confirmons ces observations dans cet article, et nous introduisons également une méthode d'enrichissement de la requête à partir d'informations contextuelles issues de documents encyclopédiques. Nous menons des expérimentations en utilisant la collection et les méthodes d'évaluation proposées par la campagne INEX. Les résultats obtenus montrent que l'ajout d'informations contextuelles permet d'améliorer significativement les performances de notre système de recherche de passages. Nous observons également que notre approche automatique obtient les meilleurs résultats parmi les différentes approches que nous évaluons.

Abstract. Traditional Information Retrieval aims to present whole documents that are relevant to a user request. However, there is sometimes only one sentence that is relevant in the document. The purpose of Focused Information Retrieval is to find and extract relevant passages instead of entire documents. This retrieval task often lacks of complement concerning the context of the information need of the user. Studies show that the performance of focused retrieval systems are improved when user manually add contextual information. In this paper we confirm these observation, and we also introduce a query expansion approach using contextual information taken from encyclopedic documents. We use the INEX workshop collection and evaluation framework in our experiments. Results show that adding contextual information significantly improves the performance of our focused retrieval system. We also see that our automatic approach obtains the best results among the different approach we evaluate.

Mots-clés : Recherche de passages, enrichissement de requêtes, contexte, Wikipedia, INEX, entropie.

Keywords: Focused retrieval, query expansion, context, Wikipedia, INEX, entropy.

Construction d'un lexique des adjectifs dénominaux

Jana Strnadová[1,2] & Benoît Sagot[3]
(1) LLF, CNRS & Univ. Paris 7, 5 rue Thomas Mann, 75205 Paris Cedex 13, France
(2) Univerzita Karlova, Filozofická Fakulta, nám. J. Palacha 2, 116 38 Prague, Rép. Tchèque
(3) Alpage, INRIA & Univ. Paris 7, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
strnadjana13@gmail.com, benoit.sagot@inria.fr

Résumé. Après une brève analyse linguistique des adjectifs dénominaux en français, nous décrivons le processus automatique que nous avons mis en place à partir de lexiques et de corpus volumineux pour construire un lexique d'adjectifs dénominaux dérivés de manière régulière. Nous estimons à la fois la précision et la couverture du lexique dérivationnel obtenu. À terme, ce lexique librement disponible aura été validé manuellement et contiendra également les adjectifs dénominaux à base supplétive.

Abstract. After a brief linguistic analysis of French denominal adjectives, we describe the automatic technique based on large-scale lexicons and corpora that we developed for building a lexicon of regular denominal adjectives. We evaluate both the precision and coverage of the resulting derivational lexicon. This freely available lexicon should eventually be fully manually validated and contain denominal adjectives with a suppletive base.

Mots-clés : Adjectifs dénominaux, dérivation morphologique, lexique dérivationnel.

Keywords: Denominal adjectives, morphological derivation, derivational lexicon.

1 Introduction

La morphologie constructionnelle est la partie de la morphologie qui permet de créer des lexèmes à partir d'autres lexèmes, et d'augmenter ainsi le lexique d'une langue. Il s'agit donc d'un procédé de structuration du lexique qui relie des lexèmes entre eux et un procédé d'extension du lexique. À ce double titre, la disponibilité de ressources lexicales dérivationnelles est cruciale à la fois pour la description linguistique et pour le traitement automatique des langues (TAL). En linguistique, les ressources lexicales enrichies de liens dérivationnels permettent une approche systémique du lexique et des procédés productifs de construction de lexèmes. En traitement automatique des langues, de nombreux travaux antérieurs ont ainsi montré l'importance de la prise en compte de la morphologie constructionnelle pour l'analyse des mots inconnus (néologismes, termes techniques) et l'enrichissement de ressources lexicales (Dal *et al.*, 1999; Hathout & Tanguy, 2005), l'analyse syntaxique (Bourigault & Frérot, 2004), mais également dans des contextes plus applicatifs tels que les systèmes de question-réponse (Bernhard *et al.*, 2011) ou de traduction automatique (Cartoni, 2009).

Pour le français, les noms déverbaux ont été étudiés de façon systématique et font l'objet d'une ressource lexicale librement disponible, VerbAction (Tanguy & Hathout, 2002). Cette ressource a été développée en partie grâce au système Webaffix (Hathout & Tanguy, 2005), qui utilise le Web comme un corpus pour y détecter des lexèmes construits néologiques. Nous nous penchons ici sur les adjectifs dénominaux, avec pour objectif de construire à terme une ressource du même type que VerbAction, mais qui s'en distinguera entre autres par des entrées plus détaillées (définitions, procédé morphologique dérivationnel détaillé…) et par la prise en compte de la dérivation non régulière (*école* vs *scolaire*, cf. partie 2).

Dans cet article, nous décrivons la première phase des travaux entrepris en ce sens. Nous nous sommes pour l'instant concentrés sur les adjectifs dérivés à partir d'une base nominale par affixation régulière, tout en prenant en compte les nombreuses variantes possibles selon les affixes et les bases. Après une étude linguistique préliminaire (section 2), nous décrivons nos expériences destinées à produire des liens de dérivation formels entre entrées lexicales connues de lexiques de référence (section 3.2) puis entre noms connus et dérivés adjectivaux inconnus (section 3.3). Nous estimons enfin la précision et la couverture des résultats obtenus (section 4). La ressource obtenue est librement disponible sous licence LGPL-LR comme complément au lexique Le*fff* (cf. plus bas).

Développement de ressources pour le persan : PerLex 2, nouveau lexique morphologique et MElt$_{fa}$, étiqueteur morphosyntaxique

Benoît Sagot[1] Géraldine Walther[2, 3] Pegah Faghiri[3] Pollet Samvelian[3]
(1) Alpage, INRIA & Univ. Paris 7, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
(2) LLF, CNRS & Univ. Paris 7, 5 rue Thomas Mann, 75205 Paris Cedex 13, France
(3) MII, CNRS & Univ. Paris 3, 27 rue Paul Bert, 94204 Ivry-sur-Seine, France
benoit.sagot@inria.fr, geraldine.walther@linguist.jussieu.fr,
pegah.faghiri@etud.sorbonne-nouvelle.fr, pollet.samvelian@univ-paris3.fr

Résumé. Nous présentons une nouvelle version de PerLex, lexique morphologique du persan, une version corrigée et partiellement réannotée du corpus étiqueté BijanKhan (BijanKhan, 2004) et MElt$_{fa}$, un nouvel étiqueteur morphosyntaxique librement disponible pour le persan. Après avoir développé une première version de PerLex (Sagot & Walther, 2010), nous en proposons donc ici une version améliorée. Outre une validation manuelle partielle, PerLex 2 repose désormais sur un inventaire de catégories linguistiquement motivé. Nous avons également développé une nouvelle version du corpus BijanKhan : elle contient des corrections significatives de la tokenisation ainsi qu'un réétiquetage à l'aide des nouvelles catégories. Cette nouvelle version du corpus a enfin été utilisée pour l'entraînement de MElt$_{fa}$, notre étiqueteur morphosyntaxique pour le persan librement disponible, s'appuyant à la fois sur ce nouvel inventaire de catégories, sur PerLex 2 et sur le système d'étiquetage MElt (Denis & Sagot, 2009).

Abstract. We present a new version of PerLex, the morphological lexicon for the Persian language, a corrected and partially re-annotated version of the BijanKhan corpus (BijanKhan, 2004) and MElt$_{fa}$, a new freely available POS-tagger for the Persian language. After PerLex's first version (Sagot & Walther, 2010), we propose an improved version of our morphological lexicon. Apart from a partial manual validation, PerLex 2 now relies on a set of linguistically motivated POS. Based on these POS, we also developped a new version of the BijanKhan corpus with significant corrections of the tokenisation. It has been re-tagged according to the new set of POS. The new version of the BijanKhan corpus has been used to develop MElt$_{fa}$, our new freely-available POS-tagger for the Persian language, based on the new POS set, PerLex 2 and the MElt tagging system (Denis & Sagot, 2009).

Mots-clés : Ressource lexicale, validation, étiqueteur morphosyntaxique, persan, catégories, PerLex, MElt.

Keywords: Lexical resource, validation, tagger, Persian, POS, PerLex, MElt.

1 Introduction

Les ressources lexicales et les outils de pré-traitement automatique des langues comme les étiqueteurs morphosyntaxiques sont des ressources indispensables pour développer des ressources plus complexes et progresser rapidement dans la description théorique des langues en donnant accès à un nombre plus conséquent de données. Malheureusement, ils ne sont que trop rarement librement disponibles, et ce même pour des langues ayant un grand nombre de locuteurs et donc de bénéficiaires potentiels. Pour le persan, ce n'est qu'en 2010 que les premiers lexiques librement disponibles ont commencé à apparaître. Notre lexique morphologique PerLex en est un des précurseurs.

Nous avions développé une première version de PerLex (Sagot & Walther, 2010) dont nous proposons desormais une deuxième version partiellement validée. PerLex 2 possède un nouvel inventaire de catégories fondé sur des choix linguistiques discutés au sein du projet ANR/DFG franco-allemand PerGram. Le développement de PerLex 2 s'accompagne de celui d'un étiqueteur morphosyntaxique, MElt$_{fa}$, qui s'appuie sur l'étiqueteur MElt (Denis & Sagot, 2009) et sur une nouvelle version du corpus BijanKhan (BijanKhan, 2004), résultat de corrections significatives de la tokenisation et d'un réétiquetage selon les nouvelles catégories.

Dans cet article, nous exposons les différentes facettes de ce triple travail dans leur succession et leur interaction. Après une rapide présentation des spécificités du traitement du persan, nous décrivons les améliorations de PerLex

Identification de cognats à partir de corpus parallèles français-roumain

Mirabela Navlea Amalia Todiraşcu

(1) Université de Strasbourg, 22 rue René Descartes, BP, 80010, 67084 Strasbourg, cedex
navlea@unistra.fr, todiras@unistra.fr

Résumé Cet article présente une méthode hybride d'identification de cognats français - roumain. Cette méthode exploite des corpus parallèles alignés au niveau propositionnel, lemmatisés et étiquetés (avec des propriétés morphosyntaxiques). Notre méthode combine des techniques statistiques et des informations linguistiques pour améliorer les résultats obtenus. Nous évaluons le module d'identification de cognats et nous faisons une comparaison avec des méthodes statistiques pures, afin d'étudier l'impact des informations linguistiques utilisées sur la qualité des résultats obtenus. Nous montrons que l'utilisation des informations linguistiques augmente significativement la performance de la méthode.

Abstract This paper describes a hybrid French - Romanian cognate identification method. This method uses lemmatized, tagged (POS tags) and sentence-aligned parallel corpora. Our method combines statistical techniques and linguistic information in order to improve the results. We evaluate the cognate identification method and we compare it to other methods using pure statistical techniques to study the impact of the used linguistic information on the quality of the results. We show that the use of linguistic information in the cognate identification method significantly improves the results.

Mots-clés : cognat, identification de cognats, corpus parallèles alignés au niveau propositionnel
Keywords: cognate, cognate identification, sentence-aligned parallel corpora

1 Introduction

Les cognats sont des indices lexicaux importants pour différentes applications multilingues, et notamment pour des systèmes d'alignement de corpus parallèles et de traduction automatique statistique. Nous définissons comme cognats les équivalents de traduction ayant une forme identique d'une langue à l'autre ou présentant des similarités aux niveaux orthographique ou phonétique (mots d'étymologie commune, emprunts). Les cognats sont nombreux entre des langues apparentées comme le français et le roumain, deux langues latines avec une morphologie flexionnelle riche. Mais, l'identification de cognats à partir des textes multilingues parallèles est une tâche difficile. Ceci est dû aux similarités orthographiques ou phonétiques importantes entre des mots ayant un sens différent.

Ainsi, de nombreux travaux se concentrent sur l'identification de cognats pour différentes paires de langues. Plusieurs approches exploitent les similarités orthographiques entre les mots d'une paire bilingue. Une approche simple et efficace est la méthode appelée 4-grammes: deux mots sont considérés comme cognats s'ils possèdent au moins 4 caractères et leurs premiers 4 caractères sont identiques (Simard et al., 1992). D'autres méthodes exploitent le coefficient de Dice (Adamson, Boreham, 1974 ; Brew, McKelvie, 1996). Ce score d'association calcule le rapport entre le nombre de caractères des bigrammes communs aux deux mots considérés et le nombre total des bigrammes des deux mots. Afin d'identifier les cognats, certaines méthodes calculent le rapport entre le nombre de caractères (ordonnés et pas nécessairement contigus) de la sous-chaine maximale commune aux deux mots et la longueur du mot le plus long (Melamed, 1999 ; Kraif, 1999). De manière similaire, d'autres méthodes calculent la distance entre deux mots, qui représente le nombre minimum de substitutions, insertions et suppressions utilisées pour transformer un mot dans un autre (Wagner, Fischer, 1974). D'autre part, certaines approches estiment la distance phonétique entre deux mots appartenant à une paire bilingue (Oakes, 2000). Kondrak (2009) propose des méthodes identifiant trois caractéristiques des cognats : les correspondances de sons récurrents, la similarité phonétique et l'affinité sémantique.

Le TAL au service de l'ALAO/ELAO
L'exemple des exercices de dictée automatisés

Richard Beaufort Sophie Roekhaut
CENTAL, UCLouvain, Place Blaise Pascal 1, B-1348 Louvain-la-Neuve
{richard.beaufort,sophie.roekhaut}@uclouvain.be

Résumé. Ce papier s'inscrit dans le cadre général de l'Apprentissage et de l'Enseignement des Langues Assistés par Ordinateur, et concerne plus particulièrement l'automatisation des exercices de dictée. Il présente une méthode de correction des copies d'apprenants qui se veut originale en deux points. Premièrement, la méthode exploite la composition d'automates à états finis pour détecter et pour analyser les erreurs. Deuxièmement, elle repose sur une analyse morphosyntaxique automatique de l'original de la dictée, ce qui facilite la production de diagnostics.

Abstract. This paper comes within the scope of the Computer Assisted Language Learning framework, and addresses more especially the automation of dictation exercises. It presents a correction method of learners' copies that is original in two ways. First, the method exploits the composition of finite-state automata, to both detect and analyze the errors. Second, it relies on an automatic morphosyntactic analysis of the original dictation, which makes it easier to produce diagnoses.

Mots-clés : ALAO/ELAO, exercices de dictée, alignement, diagnostic, machines à états finis.

Keywords: CALL, dictation exercises, alignment, diagnosis, finite-state machines.

1 Introduction

L'Apprentissage et l'Enseignement des Langues Assistés par Ordinateur (ALAO/ELAO) ont pour objectif premier d'améliorer l'acquisition des langues par les apprenants. Pourtant, force est de constater qu'actuellement, l'investissement dans le domaine a plus été technologique que didactique : pour l'essentiel, la numérisation des cours n'a pas modifié leur contenu ni leurs méthodes d'évaluation (Desmet & Héroguel, 2005). Selon les spécialistes, l'amélioration de l'apprentissage et de l'enseignement implique de dépasser les sempiternels exercices fermés, tels que les textes à trous et les choix multiples qui, s'ils sont faciles à corriger, limitent considérablement les possibilités d'évaluation des connaissances. Il faudrait au moins proposer des exercices semi-ouverts, qui autorisent plusieurs réponses relativement prévisibles, pour autant que la correction automatique de ces exercices soit fiable : apprentissage et enseignement, en effet, ne tolèrent pas l'approximation (Antoniadis *et al.*, 2009).
La dictée, où l'enseignant lit à haute voix un passage que les étudiants doivent copier, est typiquement l'un de ces exercices semi-ouverts qui, s'il était automatisé, pourrait considérablement améliorer l'apprentissage et l'enseignement des langues. La dictée est ainsi un très bon moyen d'estimer le niveau d'un étudiant (Coniam, 1996). Sa pratique, en outre, permet d'améliorer des compétences telles que la maîtrise de la grammaire, les capacités de lecture, la connaissance du vocabulaire et le niveau de compréhension (Rahimi, 2008). Actuellement pourtant, rares ont été les essais d'automatisation de cet exercice. Or, grâce aux synthétiseurs de la parole, lire automatiquement un texte inconnu n'est plus un problème. Mais la correction d'une copie d'apprenant, par contre, est une étape beaucoup plus délicate à automatiser, parce qu'elle pose deux questions sensibles : la détection de la place réelle des erreurs et leur classification.

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

Une analyse basée sur la S-DRT pour la modélisation de dialogues pathologiques

Maxime Amblard [1,4] Michel Musiol[2,4] Manuel Rebuschi[3,4]
(1) LORIA - UMR 7503
(2) InterPSY - EA 4432 / MSH Lorraine USR 3261
(3) Archives Poincaré - UMR 7117 / MSH Lorraine USR 3261
(4) Université Nancy 2 – 54000 Nancy
{maxime.amblard, michel.musiol, manuel.rebuschi}@univ-nancy2.fr

Résumé. Dans cet article, nous présentons la définition et l'étude d'un corpus de dialogues entre un schizophrène et un interlocuteur ayant pour objectif la conduite et le maintien de l'échange. Nous avons identifié des discontinuités significatives chez les schizophrènes paranoïdes. Une représentation issue de la S-DRT (sa partie pragmatique) permet de rendre compte des ces usages non standards.

Abstract. In this article, we present a corpus of dialogues between a schizophrenic speaker and an interlocutor who drives the dialogue. We had identified specific discontinuities for paranoid schizophrenics. We propose a modeling of these discontinuities with S-DRT (its pragmatic part).

Mots-clés : S-DRT, interaction verbale, schizophrénie, dialogue pathologique, incohérence pragmatique.

Keywords: S-DRT, verbal interaction, schizophrenia, pathological dialogue, pragmatical incoherence .

1 Contexte Scientifique

La pathologie schizophrénique constitue aujourd'hui encore une entité clinique complexe et mal définie. Elle suscite de nombreuses controverses relativement aux caractéristiques symptomatologiques (ou regroupements syndromiques) susceptibles de la définir. Il est difficile de trouver des caractéristiques ou traits partagés par les individus présentant ce diagnostic et nous ne disposons d'aucun signe pathognomonique clairement défini dans la littérature scientifique qui puisse la spécifier. Ce constat théorique est relayé par la multiplicité des manifestations cliniques présentées. Quant à la manifestation objective des symptômes, il est encore impossible de rapporter sans risque des traits de comportements manifestes et identifiables à des caractéristiques syndromiques circonscrites.

Nous estimons que les productions et manifestations comportementales de tout sujet, "normal" ou pathologique, sont nécessairement soumises à l'épreuve d'un cadre interactionnel et discursif, fût-il expérimental ou clinique. Nous formulons l'hypothèse selon laquelle le comportement verbal de tout (inter)-locuteur est susceptible de refléter des spécificités syndromiques ; ce comportement s'étaye sur un ensemble de contraintes sociales et cognitives qui sont la condition de l'usage naturel de la langue. L'interaction verbale est alors considérée comme le "lieu naturel d'expression des symptômes" (Trognon & Musiol, 1996). Nous envisageons de mettre au jour le plus objectivement possible, c'est-à-dire de manière "décisive" [1], les discontinuités apparaissant dans le discours et le dialogue, discontinuités dont on discutera ensuite la relation à de possibles spécificités syndromiques. Ce programme de recherche a aussi pour ambition de traiter de l'interprétation de ces discontinuités en termes d'incohérences ou de dysfonctionnements. Sur le plan méthodologique, nous nous proposons de spécifier la notion d'incohérence en types ou modèles de discontinuité de l'interaction verbale. Et nous proposons plus globalement un programme d'analyse de l'interaction verbale d'inspiration pragmatique, cognitive et formelle (Musiol & Rebuschi, 2007; Rebuschi & Musiol, 2010; Amblard *et al.*, 2010) tel que le repérage et la description de ces spécificités (ou discontinuités), quand il y en a, devrait améliorer à moyen terme à la fois les stratégies de diagnostic usuelles et les tentatives de spécification des troubles sur le plan des opérations cognitives et de pensée complexes. Les modalités d'expression du trouble sont appréhendées dans la structure intentionnelle du dialogue. Ainsi, à condition de disposer d'une stratégie de modélisation et de formalisation adéquate, cette structure intentionnelle du dialogue nous livre en filigrane les principales propriétés de la rationalité du trouble dans sa modalité inférentielle et computo-représentationnelle.

1. Cf plus bas. Les séquences conversationnelles à ruptures décisives ne résistent pas à l'épreuve du principe logique de non-contradiction. Les séquences conversationnelles à ruptures non décisives présentent des caractéristiques incongrues ou des formes d'incohérences étayant des infractions comportementales de type "normatives".

The Text+Berg Corpus
An Alpine French-German Parallel Resource

Anne Göhring, Martin Volk
UZH, Institute of Computational Linguistics
University of Zurich, Switzerland
lastname@cl.uzh.ch

Résumé. Cet article présente un corpus parallèle français-allemand de plus de 4 millions de mots issu de la numérisation d'un corpus alpin multilingue. Ce corpus est une précieuse ressource pour de nombreuses études de linguistique comparée et du patrimoine culturel ainsi que pour le développement d'un système statistique de traduction automatique dans un domaine spécifique. Nous avons annoté un échantillon de ce corpus parallèle et aligné les structures arborées au niveau des mots, des constituants et des phrases. Cet "alpine treebank" est le premier corpus arboré parallèle français-allemand de haute qualité (manuellement contrôlé), de libre accès et dans un domaine et un genre nouveau : le récit d'alpinisme.

Abstract. This article presents a French-German parallel corpus of more than 4 million tokens which we have compiled as part of the digitization of a large multilingual heritage corpus of alpine texts. This corpus is a valuable resource for cultural heritage and cross-linguistic studies as well as for the development of domain-specific machine translation systems. We have turned a small fraction of the parallel corpus into a high-quality parallel treebank with manually checked syntactic annotations and cross-language word and phrase alignments. This alpine treebank is the first freely available French-German parallel treebank. It complements other treebanks with texts in a new domain and genre : mountaineering reports.

Mots-clés : corpus alpin français-allemand, structures arborées parallèles, annotation morphosyntaxique du français.

Keywords: French-German alpine corpus, parallel treebank, French morphosyntactic annotation, Text+Berg, e-Humanities.

1 Introduction

Parallel corpora have become central resources for many areas in natural language processing such as word sense disambiguation, bilingual terminology extraction and machine translation. However most of the large available parallel corpora come from a limited set of domains (parliamentary proceedings, legal texts). We have compiled a sizable parallel corpus of French-German alpine texts. It also differs from previous parallel corpora in that it was built on the basis of printed books that we scanned and OCRized.

Our parallel corpus is a by-product of our effort to digitize and annotate all the yearbooks of the Swiss Alpine Club from 1864 until today. Since 1957 the books have been published in parallel language versions in French and German. We have scanned all books (more than 80,000 pages) until the year 2000. The books from 2001 to 2009 were provided by the Swiss Alpine Club as PDF documents. Overall, this resulted in a parallel corpus of more than 4 million tokens each in French and German.

We selected a small part of this corpus, 1000 sentences from mountaineering reports, to build a parallel treebank. This treebank consists of manually checked syntax structures on both the French and German sentences as well as cross-language word and phrase alignments. There are state-of-the-art automatic tree aligners : the supervised approach reported in (Tiedemann & Kotzé, 2009) outperforms the unsupervised technique described in (Zhechev, 2009). We decided to build a manually checked, high-quality alpine treebank to complete our collection.[1] These treebanks are useful to train and evaluate automatic tree annotation and alignment.

In this paper we first describe the creation of our Text+Berg corpus. We then focus on the French language parts in the mixed language period from 1864 to 1956 and the parallel language period from 1957 to 2009. We give an overview of the number of French articles. But we also look at French sentences scattered throughout the corpus in German articles. This case study of language mixture illustrates the linguistic richness in our alpine corpus. In the final section we present our steps for building the French-German parallel treebank, with particular attention to the annotation of the French treebank.

[1] We have released this treebank as part of our SMULTRON corpus which otherwise consists of parallel treebanks in English, German, Spanish and Swedish for three other text genres. We are distributing the latest version of our multilingual parallel treebank as SMULTRON v 3.0, Volk *et al.* (2010c).

Ordonner un résumé automatique multi-documents fondé sur une classification des phrases en classes lexicales

Aurélien Bossard Émilie Guimier De Neef
Orange Labs
2 av. Pierre Marzin
22300 Lannion, France
prenom.nom@orange-ftgroup.com

Résumé. Nous présentons différentes méthodes de réordonnancement de phrases pour le résumé automatique fondé sur une classification des phrases à résumer en classes thématiques. Nous comparons ces méthodes à deux *baselines* : ordonnancement des phrases selon leur pertinence et ordonnancement selon la date et la position dans le document d'origine. Nous avons fait évaluer les résumés obtenus sur le corpus RPM2 par 4 annotateurs et présentons les résultats.

Abstract. We present several sentence ordering methods for automatic summarization which are specific to multi-document summarizers, based on sentences subtopic clustering. These methods are compared to two baselines : sentence ordering according to pertinence and according to publication date and inner document position. The resulting summaries on RPM2 corpus have been evaluated by four judges.

Mots-clés : Résumé automatique, ordonnancement de phrases.

Keywords: Automatic summarization, sentence ordering.

1 Introduction

Les systèmes de résumé automatique multi-documents par extraction fondés sur une classification préalable des phrases en classes lexicales (CL) ont récemment prouvé leur efficacité lors des campagnes d'évaluation internationales TAC [1]. Ces systèmes offrent une modélisation du corpus à résumer différente de celles traditionnellement utilisées . La redondance dans les corpus multi-documents est plus importante que dans des documents simples. Une telle modélisation le prend en compte, et s'appuie sur la redondance pour cibler l'information pertinente tout en évitant la répétition d'éléments d'information dans le résumé. Nous avons montré que cette modélisation pouvait être utile afin d'affiner la sélection de phrases (Bossard & De Neef, 2011).

La problématique du réordonnancement de phrases pour le résumé multi-documents est plus complexe que pour le résumé mono-document. En effet, dans un document unique, les phrases sont toutes issues de la même structure discursive, et peuvent être restituées dans l'ordre du document source. En revanche, en multi-documents, les phrases peuvent être extraites de documents épars, et une structure discursive doit être recomposée. Nous montrons ici que la modélisation en classes lexicales peut également servir à réordonner les phrases du résumé, et présentons une évaluation de notre méthode de réordonnancement. Dans un premier temps, nous dressons un état de l'art de l'ordonnancement de phrases pour le résumé automatique. Cet état de l'art vise à renseigner le lecteur sur les différentes stratégies de réordonnancement de résumé, non sur les méthodes de résumé en soi. Pour plus de renseignements sur les méthodes de résumé automatique, le lecteur pourra se référer à (Mani, 1999) ou à (Das & Martins, 2007). Dans un second temps, nous présentons notre méthode d'ordonnancement et son évaluation réalisée sur le corpus RPM2 (de Loupy *et al.*, 2010). Enfin, nous discutons les résultats obtenus

1. *Text Analysis Conference*, organisée par le *National Institute of Science and Technology* : `http://www.nist.gov/tac`

Construction d'une grammaire d'arbres adjoints pour la langue arabe

Férial Ben Fraj

(1) Laboratoire RIADI, École Nationale des Sciences de l'Informatique, 2010 Manouba, Tunisie.
Feriel.BenFraj@riadi.rnu.tn

Résumé. La langue arabe présente des spécificités qui la rendent plus ambigüe que d'autres langues naturelles. Sa morphologie, sa syntaxe ainsi que sa sémantique sont en corrélation et se complètent l'une l'autre. Dans le but de construire une grammaire qui soit adaptée à ces spécificités, nous avons conçu et développé une application d'aide à la création des règles syntaxiques licites suivant le formalisme d'arbres adjoints. Cette application est modulaire et enrichie par des astuces de contrôle de la création et aussi d'une interface conviviale pour assister l'utilisateur final dans la gestion des créations prévues.

Abstract. The Arabic language consists of a set of specificities. Thus, it is more ambiguous than other natural languages. Its morphology, syntax and semantic are correlated to each other. We have constructed an application for the construction of a tree adjoining grammar which respects the characteristics of the Arabic language. This tool allows constructing the grammatical rules as elementary trees enriched by different feature structures. It helps the user by its interface and control system to manage correct and uniform rules.

Mots-clés : Outil semi-automatique, grammaire d'arbres adjoints, langue arabe, traits d'unification
Keywords: semi-automatic tool, tree adjoining grammar, Arabic language, feature structures

1 Introduction

La langue arabe présente des caractéristiques qui la rendent plus ambigüe que d'autres langues naturelles, spécialement les langues indo-européennes. Pour cette langue, il y a un entrelacement fort entre les différents niveaux du traitement linguistique ; à savoir morphologique, syntaxique et aussi sémantique. Ainsi, la syntaxe ne peut être bien gérée qu'avec l'intervention de données à la fois morphologiques et aussi sémantiques. Cette corrélation se manifeste, par exemple, dans l'accord en genre et en nombre entre les noms et les adjectifs ou encore l'accord entre les noms et les verbes. Ces accords dépendent, intimement, de l'animation et/ou de l'humanité du nom en question. Citons alors l'exemple des phrases suivantes : الكلاب جائعة (*les chiens sont affamés*) : le mot الكلاب (*les chiens*) étant un nom masculin pluriel inhumain ne s'accorde pas en nombre et en genre avec l'adjectif جائعة (*est affamée*) qui est au féminin singulier. Ceci n'est pas le cas avec la phrase الأطفال جياع (*les enfants sont affamés*), où l'accord entre le nom et l'adjectif existe, vu que الأطفال (*les enfants*) est un nom animé humain, et ainsi l'adjectif جياع (*sont affamés*) est au masculin pluriel. D'autant plus, le schème (الوزن) de l'item lexical qui est une information morphologique aide à spécifier l'utilisation syntaxique de cet outil voire même son sens, tel par exemple le schème فاعـل (*Faâilun*) qui indique le réalisateur de l'action.

Par conséquent, la construction d'une grammaire formelle pour la langue arabe doit, obligatoirement, faire intervenir différents types d'informations utiles afin de mieux gérer la bonne composition des agencements syntaxiques. Dans ce papier, nous dressons le problème de cette construction. Il existe différentes grammaires formelles ; à savoir : les grammaires génératives (Chomsky, 1965) et celles

FreDist : Automatic construction of distributional thesauri for French

Enrique Henestroza Anguiano & Pascal Denis
Alpage, INRIA Paris-Rocquencourt & Université Paris 7
Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
{henestro, pascal.denis}@inria.fr

Résumé. Dans cet article, nous présentons FreDist, un logiciel libre pour la construction automatique de thésaurus distributionnels à partir de corpus de texte, ainsi qu'une évaluation des différents ressources ainsi produites. Suivant les travaux de (Lin, 1998) et (Curran, 2004), nous utilisons un corpus journalistique de grande taille et implémentons différentes options pour : le type de relation contexte lexical, la fonction de poids, et la fonction de mesure de similarité. Prenant l'EuroWordNet français et le WOLF comme références, notre évaluation révèle, de manière originale, que c'est l'approche qui combine contextes linéaires (ici, de type bigrammes) et contextes syntaxiques qui semble fournir le meilleur thésaurus. Enfin, nous espérons que notre logiciel, distribué avec nos meilleurs thésaurus pour le français, seront utiles à la communauté TAL.

Abstract. In this article we present FreDist, a freely available software package for the automatic construction of distributional thesauri from text corpora, as well as an evaluation of various distributional similarity metrics for French. Following from the work of (Lin, 1998) and (Curran, 2004), we use a large corpus of journalistic text and implement different choices for the type of lexical context relation, the weight function, and the measure function needed to build a distributional thesaurus. Using the EuroWordNet and WOLF wordnet resources for French as gold-standard references for our evaluation, we obtain the novel result that combining bigram and syntactic dependency context relations results in higher quality distributional thesauri. In addition, we hope that our software package and a joint release of our best thesauri for French will be useful to the NLP community.

Mots-clés : thésaurus distributionnel, similarité sémantique, méthodes non supervisées, lexique.

Keywords: distributional thesaurus, semantic similarity, unsupervised methods, lexicon.

1 Introduction

We present FreDist, software that implements methods for the automatic construction of distributional thesauri. Distributional lexical resources are appealing because they can be constructed automatically from raw text corpora, and are useful for alleviating data sparseness in many NLP applications (e.g. parsing and coreference resolution). Moreover, we believe that open software like FreDist can be useful to the NLP community by providing an easy way to generate distributional thesauri from any text corpus using adjustable settings.

We base our work on that of (Lin, 1998), which uses word context relations to calculate lexical distributional similarity, and the subsequent work of (Curran, 2004), which distinguishes between weight and measure functions and evaluates different functions on a semantic similarity task for English. We build on their work by considering the joint use of different types of context relations, and evaluating distributional similarity metrics for French.

Current lexical resources for French that have been semi-automatically created include the work of (Sagot, 2010) on the Le*fff*, a large-coverage morphosyntactic lexicon, and (Sagot & Fišer, 2008) on the WOLF, a semantic resource based on the Princeton WordNet. Our work differs by providing a fully automatic approach to the creation of a lexical resource. Previous work on distributional methods for French includes that of (Bourigault, 2002) on UPERY, a distributional analysis module that calculates proximities between words and their contexts, and the work of (Ferret, 2004), which uses distributional similarity to build word senses from a network of lexical co-occurrences. Our work differs by focusing on the construction and evaluation of distributional thesauri, combining different types of context relations, and making FreDist and our best distributional thesauri freely-available. [1]

1. `http://alpage.inria.fr/~henestro/fredist.html`

Using shallow linguistic features for relation extraction in bio-medical texts

Ali Reza Ebadat[1] Vincent Claveau[2] Pascale Sébillot[3]
(1) INRIA-INSA, (2) IRISA-CNRS, (3) IRISA-INSA
Campus de Beaulieu, 35042 Rennes, France
ali_reza.ebadat@inria.fr, vincent.claveau@irisa.fr, pascale.sebillot@irisa.fr

Résumé. Dans cet article[1], nous proposons de modéliser la tâche d'extraction de relations à partir de corpus textuels comme un problème de classification. Nous montrons que, dans ce cadre, des représentations fondées sur des informations linguistiques de surface sont suffisantes pour que des algorithmes d'apprentissage artificiel standards les exploitant rivalisent avec les meilleurs systèmes d'extraction de relations reposant sur des connaissances issues d'analyses profondes (analyses syntaxiques ou sémantiques). Nous montrons également qu'en prenant davantage en compte les spécificités de la tâche d'extraction à réaliser et des données disponibles, il est possible d'obtenir des méthodes encore plus efficaces tout en exploitant ces informations simples. La technique originale à base d'apprentissage « paresseux » et de modèles de langue que nous évaluons en extraction d'interactions géniques sur les données du challenge LLL2005 dépasse les résultats de l'état de l'art.

Abstract. In this paper[2], we model the corpus-based relation extraction task as a classification problem. We show that, in this framework, standard machine learning systems exploiting representations simply based on shallow linguistic information can rival state-of-the-art systems that rely on deep linguistic analysis. Even more effective systems can be obtained, still using these easy and reliable pieces of information, if the specifics of the extraction task and the data are taken into account. Our original method combining lazy learning and language modeling out-performs the existing systems when evaluated on the LLL2005 protein-protein interaction extraction task data.

Mots-clés : Extraction de relations, classification, apprentissage paresseux, modèle de langue, analyse linguistique de surface.

Keywords: Relation extraction, classification, lazy learning, langage model, shallow linguistic analysis.

[1]Ces travaux ont été réalisés dans le cadre du programme QUAERO, financé par OSEO, agence française pour l'innovation.
[2]This work was achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

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

Vers une prise en charge approfondie des phénomènes itératifs par TimeML

Julien Lebranchu, Yann Mathet
Université de Caen Basse-Normandie, UMR 6072 GREYC, F-14032 Caen, France
Prénom.Nom@unicaen.fr

Résumé. Les travaux menés ces dernières années autour de l'itération en langue, tant par la communauté linguistique que par celle du TAL, ont mis au jour des phénomènes particuliers, non réductibles aux représentations temporelles classiques. En particulier, une itération ne saurait structurellement être réduite à une simple énumération de procès, et du point de vue de l'aspect, met en jeu simultanément deux visées aspectuelles indépendantes. Le formalisme TimeML, qui a vocation à annoter les informations temporelles portées par un texte, intègre déjà des éléments relatifs aux itérations, mais ne prend pas en compte ces dernières avancées. C'est ce que nous entreprenons de faire dans cet article, en proposant une extension à ce formalisme.

Abstract. The work that has recently been done concerning the iterative phenomena in language, which was performed by the linguistic and TAL communities, has illuminated specific phenomena, not reducible to classical time representations. In particular, an iteration can not structurally be reduced to a simple listing of process, and involves simultaneously two independent referred aspectual. The TimeML formalism, which aims to annotate temporal information of a given text, includes already relative elements to iterations but does not take into account recent advances. That is the reason why in this paper, we propose to extend this formalism.

Mots-clés : TimeML, discours, sémantique, phénomènes itératifs.

Keywords: TimeML, discourse, semantics, iterative phenomena.

1 Introduction

Le travail que nous menons actuellement se place dans le contexte de la sémantique temporelle, et plus précisément dans celui des phénomènes itératifs en corpus. Si plusieurs études linguistiques se sont attachées à décrire les mécanismes itératifs autour de la proposition ou de la phrase, nos travaux se positionnent quant à eux au niveau du discours (repérage des données itératives au sein d'un texte, détermination de l'étendue de chaque itération, etc., cf. (Lebranchu, 2009)). C'est dans ce contexte que nous nous attachons à déterminer le plus précisément les informations temporelles et itératives d'un texte, et que nous portons un intérêt particulier à TimeML.

L'itération en langue est un phénomène menant à la construction de différents procès répartis dans le temps et considérés comme étant la répétition d'un même événement. De façon typique, ce phénomène peut se manifester à partir d'une seule proposition telle que *« Nous sommes allés 7 fois à la montagne. »* où le procès « nous *aller* à la montagne » est itéré par l'adverbe de quantification « 7 fois ». Elle peut être circonscrite à une seule proposition, mais peut aussi s'étaler sur plusieurs propositions ou phrases, successives ou non, comme dans l'exemple qui suit.

> *Chaque lundi matin, le brocanteur qui logeait sous l'allée étalait par terre ses ferrailles. Vers midi, au plus fort du marché, on voyait paraître sur le seuil un vieux paysan de haute taille. Peu de temps après, c'était Liébard, le fermier de Toucques, […]*

Ainsi, le circonstanciel *chaque lundi matin* déclenche l'itération du procès « le brocanteur *étaler* par terre ses ferrailles », à raison d'une fois par semaine sur une certaine période. Les deux propositions « Vers midi, …, on *voir paraître* », et « Peu de temps après, ce *être* Liébard » sont elles des constituants de cette itération, qu'elles viennent successivement enrichir. Nous noterons en revanche que la subordonnée (*qui logeait sous l'allée*) n'est pas incluse dans l'itération.

Afin de représenter les informations liées aux phénomènes itératifs, nous avons opté pour ISO-TimeML (ci-après TimeML), un formalisme XML défini par Lee *et al.* (2007), dont l'objectif est d'annoter des textes avec l'ensemble de leurs informations temporelles. Il permet d'annoter les unités linguistiques réalisant des procès, les expressions temporelles, les relations aspectuelles, temporelles ou de subordination qui peuvent exister entre événements et expressions temporelles ainsi que les marqueurs de ces relations.

Une procédure pour identifier les modifieurs de la valence affective d'un mot dans des textes

Noémi Boubel[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
noemi.boubel@uclouvain.be, yves.bestgen@uclouvain.be

Résumé : Cette recherche s'inscrit dans le champ de la fouille d'opinion et, plus particulièrement, dans celui de l'analyse de la polarité d'une phrase ou d'un syntagme. Dans ce cadre, la prise en compte du contexte linguistique dans lequel apparaissent les mots porteurs de valence est particulièrement importante. Nous proposons une méthodologie pour extraire automatiquement de corpus de textes de telles expressions linguistiques. Cette approche s'appuie sur un corpus de textes, ou d'extraits de textes, dont la valence est connue, sur un lexique de valence construit à partir de ce corpus au moyen d'une procédure automatique et sur un analyseur syntaxique. Une étude exploratoire, limitée à la seule relation syntaxique associant un adverbe à un adjectif, laisse entrevoir les potentialités de l'approche.

Abstract This research is situated within the field of opinion mining and focuses more particularly on the analysis of the opinion expressed in a sentence or a syntagm. Within this frame of research, taking into account the linguistic context in which words which carry valence appear is particularly important. We propose a methodology to automatically extract such linguistic expressions from text corpora. This approach is based on (a) a corpus of texts, or text excerpts, the valence of which is known, (b) on a valence lexicon built from this corpus using an automatic procedure and (c) on a parser. An exploratory study, focusing on the syntactic relation associating an adverb to an adjective, shows the potential of the approach.

Mots-clés : modifieurs de valence, fouille d'opinion, lexique de valence
Keywords: contextual valence shifter, opinion mining, semantic orientation lexicon

1 Introduction

Depuis une dizaine d'années, la détection d'opinion et de sentiments dans les textes est devenue un sujet de recherche important en traitement automatique du langage, ainsi qu'un enjeu stratégique pour les entreprises et les institutions (Pang, Lee, 2008). La tâche classique de ce domaine consiste à déterminer automatiquement la polarité globale d'un texte. Progressivement, l'attention des chercheurs s'est également portée vers le niveau phrastique ou syntagmatique afin d'identifier les segments d'un texte qui expriment une opinion, la valence de celle-ci et sa cible (Hatzivassiloglou, Wiebe, 2000 ; Kessler, Nicolov, 2009 ; Vernier et al., 2009). Ce deuxième axe de recherche souligne l'importance de la prise en compte du contexte linguistique dans lequel apparaissent les mots porteurs de valence. Si certains travaux en fouille d'opinion commencent à prendre en compte ce genre de phénomènes (Kennedy, Inkpen, 2006 ; Musat, Trausan-Matu, 2010 ; Wilson et al., 2005), très rares sont ceux qui en ont fait leur objet central d'étude. Une recherche de Zaenen et Polanyi (2004) constitue cependant une exception notable. Leur hypothèse de travail est que la valence de termes polarisés peut être renforcée ou affaiblie par la présence d'autres items lexicaux, par la structure du discours et le type de texte, ou enfin par des facteurs culturels. Ces chercheuses proposent d'appeler *contextual valence shifters* les différents éléments ou procédés ayant un impact sur la valeur d'un mot comme la présence d'une négation, d'un adverbe de degré, d'un verbe modal ou d'une tournure ironique. Il faut toutefois noter que les arguments empiriques présentés reposent sur l'analyse de quelques exemples

Stratégie d'exploration de corpus multi-annotés avec GlozzQL

Yann Mathet[1] Antoine Widlöcher[1]
(1) GREYC, UMR CNRS 6072, Université de Caen, 14032 Caen Cedex
{prénom.nom}@unicaen.fr

Résumé. La multiplication des travaux sur corpus, en linguistique computationnelle et en TAL, conduit à la multiplication des campagnes d'annotation et des corpus multi-annotés, porteurs d'informations relatives à des phénomènes variés, envisagés par des annotateurs multiples, parfois automatiques. Pour mieux comprendre les phénomènes que ces campagnes prennent pour objets, ou pour contrôler les données en vue de l'établissement d'un corpus de référence, il est nécessaire de disposer d'outils permettant d'explorer les annotations. Nous présentons une stratégie possible et son opérationalisation dans la plate-forme Glozz par le langage GlozzQL.

Abstract. More and more works in compuational linguistics and NLP rely on corpora. They lead to an increasing number of annotation campaigns and multi-annotated corpora, providing informations on various linguistic phenomena, annotated by several annotators or computational processes. In order to understand these linguistic phenomena, or to control annotated data, tools dedicated to annotated data mining are needed. We present here an exploration strategy and its implementation within the Glozz platform, GlozzQL.

Mots-clés, Keywords: Corpus, Annotation, Exploration, GlozzQL.

1 Besoins en exploration de corpus

De nombreux travaux sur corpus, en linguistique computationnelle et en TAL, donnent lieu à des campagnes d'annotation et à la production de corpus multi-annotés. Pour mieux comprendre les phénomènes que ces campagnes prennent pour objets, pour assister le processus d'annotation et pour contrôler son résultat, il est nécessaire de disposer d'outils et de méthodes permettant d'interroger les annotations. Avant d'introduire le langage GlozzQL, qui en permet l'exploration, nous commencerons par indiquer les caractéristiques des données auxquelles il s'applique et présenter le méta-modèle sur lequel il repose.

1.1 Corpus multi-annotés

Après avoir insisté sur le fait que nous ne visons pas la mise en place d'un environnement dédié à l'étude d'une configuration linguistique particulière, mais la définition de stratégies permettant l'exploration combinée de phénomènes hétérogènes, nous pouvons en particulier mettre en avant les propriétés suivantes. **Données multi-annotateurs :** Les données explorées sont le fruit d'annotations réalisées par plusieurs annotateurs. Elles peuvent porter sur les mêmes textes et sur les mêmes objets et on souhaitera souvent pouvoir les confronter. **Variété catégorielle :** Ces corpus annotés sont porteurs d'informations linguistiques et infra-linguistiques (par exemple typo-dispositionnelles) hétérogènes correspondant à différents points de vue sur le matériau textuel. Ainsi, par exemple, on pourra disposer d'un étiquetage morpho-syntaxique, d'indications de dépendance résultant d'une analyse syntaxique, du repérage d'occurrences d'éléments de lexiques... **Configurations structurellement variées :** Les différents phénomènes annotables relèvent de modes de structuration fortement hétérogènes. Certaines approches identifient ainsi, par exemple, des segments textuels (unités morphologiques, entités nommées, segments thématiques...), là où d'autres se concentrent sur l'identification de relations entre des objets plus ou moins distants (relations syntaxiques, relations sémantiques de cohérence...) ou sur la reconnaissance de dispositifs plus complexes (chaînes de coréférence, structures énumératives...). **Granularité variable :** Ces différentes structures opèrent à des échelles pouvant varier fortement. Si le mot constitue ainsi souvent l'échelle élémentaire, les annotations syntaxiques prendront place à l'échelle de la phrase et les annotations discursives (par exemple thématiques) pourront déborder les échelles phrastiques. **Topologies variées :** La distribution des différents phénomènes répond à des « topologies » variées : parfois juxtaposés, les objets pourront aussi être imbriqués ou se chevaucher.

Attribution de rôles sémantiques aux actants des lexies verbales

Fadila Hadouche[1] Guy Lapalme[1] Marie-Claude L'Homme[2]

(1) RALI, (2) OLST
Université de Montréal, C.P 6128 Succursale Centre-ville, Montréal, Québec, Canada H3C 3J7
hadouchf@iro.umontreal.ca, lapalme@iro.umontreal.ca, mc.lhomme@umontreal.ca

Résumé

Dans cet article, nous traitons de l'attribution des rôles sémantiques aux actants de lexies verbales en corpus spécialisé en français. Nous proposons une classification de rôles sémantiques par apprentissage machine basée sur un corpus de lexies verbales annotées manuellement du domaine de l'informatique et d'Internet. Nous proposons également une méthode de partitionnement semi-supervisé pour prendre en compte l'annotation de nouvelles lexies ou de nouveaux rôles sémantiques et de les intégrés dans le système. Cette méthode de partitionnement permet de regrouper les instances d'actants selon les valeurs communes correspondantes aux traits de description des actants dans des groupes d'instances d'actants similaires. La classification de rôles sémantique a obtenu une F-mesure de 93% pour Patient, de 90% pour Agent, de 85% pour Destination et de 76% pour les autres rôles pris ensemble. Quand au partitionnement en regroupant les instances selon leur similarité donne une F-mesure de 88% pour Patient, de 81% pour Agent, de 58% pour Destination et de 46% pour les autres rôles.

Abstract

In this paper, we discuss assigning semantic roles to actants of verbal lexical units in French specialized corpus. We propose a machine learning classification of semantic roles based on a corpus of verbal lexical units, which are annotated manually in the Informatics and Internet domain. We also propose a semi supervised clustering method to consider the annotation of new verbal lexical units or new semantic roles and integrated them in the system. Clustering is used to group instances of actants according to their common values corresponding to the features describing these actants into groups of similar instances of actants. The classification model give an F-measure of 93% for Patient, 90% for Agent, 85% for Destination and 76% for other roles. When partitioning by grouping instances according to their similarity gives an F-measure of 88% for Patient, 81% for Agent, 58% for Destination and 46% for other roles.

Mots-clés : Rôles sémantiques, traits syntaxiques, classification, partitionnement semi-supervisé
Keywords: Semantic roles, syntactic features, classification, semi supervised partitioning

Utiliser l'amorçage pour améliorer une mesure de similarité sémantique

Olivier Ferret
CEA, LIST, Laboratoire Vision et Ingénierie des Contenus,
Fontenay-aux-Roses, F-92265 France.
olivier.ferret@cea.fr

Résumé. Les travaux sur les mesures de similarité sémantique de nature distributionnelle ont abouti à un certain consensus quant à leurs performances et ont montré notamment que leurs résultats sont surtout intéressants pour des mots de forte fréquence et une similarité sémantique étendue, non restreinte aux seuls synonymes. Dans cet article, nous proposons une méthode d'amélioration d'une mesure de similarité classique permettant de rééquilibrer ses résultats pour les mots de plus faible fréquence. Cette méthode est fondée sur un mécanisme d'amorçage : un ensemble d'exemples et de contre-exemples de mots sémantiquement liés sont sélectionnés de façon non supervisée à partir des résultats de la mesure initiale et servent à l'entraînement d'un classifieur supervisé. Celui-ci est ensuite utilisé pour réordonner les voisins sémantiques initiaux. Nous évaluons l'intérêt de ce réordonnancement pour un large ensemble de noms anglais couvrant différents domaines fréquentiels.

Abstract. Work about distributional semantic similarity measures has now widely shown that such measures are mainly reliable for high frequency words and for capturing semantic relatedness rather than strict semantic similarity. In this article, we propose a method for improving such a measure for middle and low frequency words. This method is based on a bootstrapping mechanism : a set of examples and counter-examples of semantically related words are selected in an unsupervised way from the results of the initial measure and used for training a supervised classifier. This classifier is then applied for reranking the initial semantic neighbors. We evaluate the interest of this reranking for a large set of english nouns with various frequencies.

Mots-clés : Extraction de voisins sémantiques, similarité sémantique, méthodes distributionnelles.

Keywords: Semantic neighbor extraction, semantic similarity, distributional methods.

1 Introduction

Le travail présenté ici prend place dans le domaine de la sémantique lexicale et plus particulièrement de la similarité sémantique au niveau lexical. La notion de *similarité sémantique* couvre, aussi bien du point de vue de sa définition que de sa caractérisation, une pluralité d'approches. Concernant sa définition, la dichotomie principale se fait entre une similarité reposant sur des relations sémantiques de nature paradigmatique (hyperonymie, synonymie, etc) et une similarité reposant sur des relations sémantiques de nature syntagmatique (relations de cohésion lexicale au statut théorique plus flou). Cette dichotomie recouvre celle faite entre les notions de *semantic similarity* et de *semantic relatedness*. Bien que justifiée par la différence de nature des relations impliquées, cette différenciation n'est pas en pratique toujours très nette, en particulier au niveau de l'évaluation. Dans le cadre du travail présenté ici, nous nous focalisons plus spécifiquement sur une caractérisation distributionnelle de la similarité sémantique. Les recherches la concernant ont montré que les relations sémantiques couvertes par une telle approche relèvent à la fois de l'axe paradigmatique et de l'axe syntagmatique. À défaut donc de nous restreindre à un seul type de relations, nous nous efforcerons de distinguer au niveau des évaluations les proximités sémantiques relevant de relations comme la synonymie de celles impliquant un ensemble plus large de relations sémantiques.

Au-delà d'une mise en œuvre « classique » de l'approche distributionnelle telle qu'elle est incarnée par (Curran & Moens, 2002), un certain nombre de propositions ont été faites pour améliorer les résultats dans le cadre de ce paradigme. Une part significative de ces propositions portent sur la pondération des éléments constitutifs des contextes associés aux mots mais un certain nombre impliquent des changements plus profonds. L'utilisation de techniques de réduction de dimensions, en l'occurrence l'analyse sémantique latente dans (Padó & Lapata, 2007), ou la redéfinition de l'approche distributionnelle dans le cadre bayésien dans (Kazama *et al.*, 2010), se classent

Un calcul de termes typés pour la pragmatique lexicale: chemins et voyageurs fictifs dans un corpus de récits de voyage

Richard Moot[1], Laurent Prévot[2], Christian Retoré[1]
(1) Université de Bordeaux, LaBRI & INRIA
(2) Université de Provence, LPL
richard.moot@labri.fr, laurent.prevot@lpl-aix.fr, christian.retore@labri.fr

Résumé. Ce travail s'inscrit dans l'analyse automatique d'un corpus de récits de voyage. À cette fin, nous raffinons la sémantique de Montague pour rendre compte des phénomènes d'adaptation du sens des mots au contexte dans lequel ils apparaissent. Ici, nous modélisons les constructions de type *'le chemin descend pendant une demi-heure'* où ledit chemin introduit un voyageur fictif qui le parcourt, en étendant des idées que le dernier auteur a développé avec Bassac et Mery. Cette introduction du voyageur utilise la montée de type afin que le quantificateur introduisant le voyageur porte sur toute la phrase et que les propriétés du chemin ne deviennent pas des propriétés du voyageur, fût-il fictif. Cette analyse sémantique (ou plutôt sa traduction en lambda-DRT) est d'ores et déjà implantée pour une partie du lexique de Grail.

Abstract. This work is part of the automated analysis of travel stories corpus. To do so, we refine Montague semantics, to model the adaptation of word meaning to the context in which they appear. Here we study construction like *'the path goes down for half an hour'* in which the path introduces a virtual traveller following it, extending ideas of the last author with Bassac, Mery. The introduction of a traveller relies on type raising satisfies the following requirements : the quantification binding the traveller has the widest scope, and properties of the path do not apply to the traveller, be it virtual. This semantical analysis (actually its translation in λ-DRT) is already implemented for a part of the Grail lexicon.

Mots-clés : Sémantique lexicale, pragmatique, sémantique compositionnelle.

Keywords: Lexical semantics, pragmatics, compositional semantics.

1 Le sens compositionnel en contexte

Suivant une tradition initiée par (Pustejovsky, 1995) et poursuivie par (Asher & Pustejovsky, 2005; Asher, 2009), et d'autres (Nunberg, 1995; Jacquey, 2006; Jayez, 2008) nous souhaitons rendre compte du sens des mots en contexte dans un cadre compositionnel, et analyser correctement les prédications acceptées tout en rejetant celles qui ne le sont pas. Les phénomènes pris en compte jusqu'ici sont illustrés par les exemples suivants :

(1) *Le dîner était sympathique, pourtant l'entrée était brûlée.*
Ici, on réfère à deux aspects d'un événement complexe, le dîner.

(2) *Ce livre est volumineux mais intéressant.*
Coprédication correcte entre les deux facettes de livre : contenu informationnel et objet physique.

(3) *J'ai mis les livres au grenier, je les avais tous lus.*
Les livres sont comptés en tant qu'objets physiques, puis repris par *les* en tant que contenus informationnels par le second prédicat.

(4) *Washington borde le Potomac et a attaqué l'Irak.*
Coprédication incorrecte (sauf trait d'humour) sur le président siégeant à Washington et le lieu géographique de cette même ville.

Pour ces phénomènes, nous avons conçu une structure de lexique et un algorithme qui permettent de calculer les représentations sémantiques de telles phrases, de rendre compte des coprédications correctes (1, 2, 3) et d'échouer

Catégoriser les réponses aux interruptions dans les débats politiques

Brigitte Bigi[1] Cristel Portes[1] Agnès Steuckardt[1] Marion Tellier[1]
(1) Laboratoire Parole & Langage, CNRS & Aix-Marseille Universités
5, avenue Pasteur, BP 80975, 13604 Aix en Provence, France
brigitte.bigi@lpl-aix.fr, cristel.portes@lpl-aix.fr, Agnes.Steuckardt@univ-provence.fr, marion.tellier@lpl-aix.fr

Résumé. Cet article traite de l'analyse de débats politiques selon une orientation multimodale. Nous étudions plus particulièrement les réponses aux interruptions lors d'un débat à l'Assemblée nationale. Nous proposons de procéder à l'analyse via des annotations systématiques de différentes modalités. L'analyse argumentative nous a amenée à proposer une typologie de ces réponses. Celle-ci a été mise à l'épreuve d'une classification automatique. La difficulté dans la construction d'un tel système réside dans la nature même des données : multimodales, parfois manquantes et incertaines.

Abstract. This work was conducted to analyze political debates, with a multimodal point of view. Particularly, we focus on the answers produced by a main speakers after he was disrupted. Our approach relies on the annotations of each modality and on their review. We propose a manual categorization of the observed disruptions. A categorization method was applied to validate the manual one. The difficulty is to deal with multimodality, missing values and uncertainty in the automatic classification system.

Mots-clés : corpus, annotations, multimodalité, classification supervisée.

Keywords: corpus, annotations, multimodality, classification.

1 Introduction

Quelque préparé qu'il ait pu être, le discours d'un député comporte, lorsqu'il est prononcé au sein de l'Assemblée nationale, une part d'improvisation. L'orateur y est exposé aux interruptions, rarement amènes, de ses collègues, prises de parole qui ne bénéficient pas d'une autorisation formelle du Président de séance. Le regard du député, alors, quitte les notes, l'intonation change, la parole spontanée prend le relai de la déclamation. Si l'analyse de discours s'est intéressée traditionnellement aux débats politiques retransmis par la télévision (Bonnafous & Tournier, 2001), en revanche les interactions à l'œuvre dans les assemblées délibératives (Assemblée nationale et Sénat), ont été moins étudiées, faute de matériau adéquat. Or l'Assemblée nationale met aujourd'hui à disposition sur internet la retransmission des vidéos des séances. Cette ressource nouvelle ouvre à la recherche l'opportunité de mieux comprendre les mécanismes de la délibération politique. Nous nous proposons d'aborder, selon une approche multimodale, les stratégies mises en œuvre pour contrer la répartie dans les débats politiques.

On retrouve dans de nombreux domaines les termes modalité et multimodalité. Cependant, leur usage et leur définition sont variables tant par le sens que par le formalisme de la définition. Par essence, cependant, la multimodalité est l'utilisation d'au moins deux des cinq sens pour l'échange d'informations. De cette définition, on peut directement déduire l'importance de la mutimodalité dans la question du langage et de la parole. Les annotations multimodales de corpus ont soulevé de nombreuses questions. La création et l'annotation de corpus multimodaux présentent des difficultés liées aux choix de logiciel, à des décisions concernant la découpe et l'alignement parole/transcription, à la pertinence des méthodes automatiques, etc. Nous avons donc été confrontés aux grands problèmes posés par l'annotation de ce type de ressource. Cet article présente en premier lieu la méthodologie qui a été conduite pour l'annotation multimodale d'une partie de la séance publique du 4 mai 2010 à l'Assemblée nationale durant laquelle Yves Cochet intervient. Le corpus contient de nombreux niveaux d'annotations automatiques (éventuellement révisés manuellement), ou manuels. Cette approche s'appuie sur celle qui a été proposée pour le corpus CID - Corpus of Interactional Data (Blache *et al.*, 2010). À ces annotations, nous ajoutons l'indication des segments durant lesquels le locuteur répond (ou choisit de ne pas répondre) à une interpellation. En effet, c'est non pas aux interruptions elles-mêmes, peu audibles dans la vidéo et incomplètement enregistrées par

Mesure non-supervisée du degré d'appartenance d'une entité à un type

Ludovic Bonnefoy[1,2], Patrice Bellot[1], Michel Benoit[2]
(1) Université d'Avignon - CERI/LIA, Agroparc – B.P. 1228, 84911 Avignon Cedex 9
(2) iSmart, Le Mercure A, 13851 Aix-en-Provence Cedex 3
patrice.bellot@univ-avignon.fr, {ludovic.bonnefoy,michel.benoit}@ismart.fr

Résumé. La recherche d'entités nommées a été le sujet de nombreux travaux. Cependant, la construction des ressources nécessaires à de tels systèmes reste un problème majeur. Dans ce papier, nous proposons une méthode complémentaire aux outils capables de reconnaître des entités de types larges, dont l'objectif est de déterminer si une entité est d'un type donné, et ce de manière non-supervisée et quel que soit le type. Nous proposons pour cela une approche basée sur la comparaison de modèles de langage estimés à partir du Web. L'intérêt de notre approche est validé par une évaluation sur 100 entités et 273 types différents.

Abstract. Searching for named entities has been the subject of many researches. In this paper, we seek to determine whether a named entity is of a given type and in what extent it is. We propose to address this issue by an unsupervised Web oriented language modeling approach. The interest of it is demonstrated by our evaluation on 100 entities and 273 different types.

Mots-clés : typage d'entités nommées, comparaison de distribution de mots, divergence de Kullback-Leibler.

Keywords: named entity identification, language modeling approach, Kullback-Leibler divergence.

1 Introduction

Depuis les années 1990, les entités nommées sont au centre de nombreux travaux en traitement de la langue naturelle écrite (résumé automatique, ontologies, ...). Un tel développement est, en grande partie, dû à l'impulsion donnée par de multiples campagnes d'évaluation, qui ont accordé une part importante à leur identification et utilisation au sein de leurs pistes tels que MUC (*Named Entity task* [1]), TREC (avec la tâche *Question Answering* (Voorhees, 1999))...

En l'absence de corpus d'apprentissage, les premières méthodes de recherche d'entités nommées, se basaient sur l'utilisation de larges ensembles de patrons d'extraction (Nadeau & Sekine, 2007) et aujourd'hui encore il est conseillé de procéder de la sorte si un corpus d'entraînement n'est pas disponible pour les types souhaités (Sekine & Nobata, 2004). Lorsque les premiers corpus d'apprentissage pour certains types (personne, lieu, organisation et date) firent leur apparition, la plupart des méthodes d'apprentissage automatique furent utilisées pour ce problème telles que les modèles de Markov cachés (Bikel *et al.*, 1997), les arbres de décision (Sekine, 1998) ou encore les SVMs (Asahara & Matsumoto, 2003) et les CRFs (McCallum, 2003). Des méthodes dites semi-supervisées ont aussi été étudiées telle le *bootstrapping* qui consiste à démarrer d'un petit jeu d'exemples et de l'agrandir par itérations successives en ayant recours à divers critères comme les relations syntaxiques (Cucchiarelli & Velardi, 2001) ou synonymiques (Pasca *et al.*, 2006).

La reconnaissance des entités nommées est centrale dans bon nombre de problématiques en recherche d'information comme par exemple Questions-Réponses (QR). Cette tâche a connu un fort engouement ces dernières années. En effet, on a pu voir plusieurs campagnes d'évaluation internationales en faire un sujet important (TREC, CLEF, INEX, Equer, ...). Un système QR présente au moins deux différences par rapport à un système de recherche d'information (RI). La première est la formulation de la requête qui est une phrase interrogative en langage naturel (par exemple *"Je veux connaître les spécifications techniques du nouveau Blackberry"*). Cela a de l'intérêt pour les utilisateurs (la formulation de requêtes efficaces sous forme de mots clés est une tâche difficile) et pour les systèmes (apport d'un contexte et d'informations supplémentaires). La seconde principale différence est la forme

1. http ://cs.nyu.edu/faculty/grishman/NEtask20.book_1.html

Traduction (automatique) des connecteurs de discours

Laurence Danlos et Charlotte Roze
ALPAGE, Université Paris Diderot (Paris 7), 175 rue du Chevaleret, F-750013 Paris
Laurence.Danlos@linguist.jussieu.fr et Charlotte.Roze@linguist.jussieu.fr

Résumé. En nous appuyant sur des données fournies par le concordancier bilingue TransSearch qui intègre un alignement statistique au niveau des mots, nous avons effectué une annotation semi-manuelle de la traduction anglaise de deux connecteurs du français. Les résultats de cette annotation montrent que les traductions de ces connecteurs ne correspondent pas aux « transpots » identifiés par TransSearch et encore moins à ce qui est proposé dans les dictionnaires bilingues.

Abstract. On the basis of data provided by the bilingual concordancer TransSearch which propose a statistical word alignment, we made a semi-manual annotation of the English translation of two French connectives. The results of this annotation show that the translations of these connectives do not correspond to the "transpots" identified by TransSearch and even less to the translations proposed in bilingual dictionaries.

Mots-clés : Traduction (automatique), TransSearch, Discours.

Keywords: (Machine) Translation, TransSearch, Discourse.

1 Introduction

Les connecteurs de discours n'appartiennent pas à une catégorie morpho-syntaxique unique : ce sont principalement des conjonctions de coordination ou de subordination (*mais, parce que*) et des adverbiaux (*ainsi, après tout*). Ils se définissent fonctionnellement comme des prédicats dont les arguments sont des « segments de discours » (pour simplifier des phrases) et qui indiquent au niveau sémantico-discursif la « relation de discours » connectant ces phrases. Ils facilitent la compréhension d'un texte par rapport aux cas où deux phrases sont simplement juxtaposées sans connecteur explicite pour les relier (ou, ce qui revient au même, avec le connecteur vide, noté ϵ) : avec le connecteur ϵ, le lecteur doit inférer la relation de discours en jeu, alors que ce travail d'inférence n'est pas nécessaire avec un connecteur explicite. Ainsi en (1) traduit de (Wilson & Sperber, 1993), si la première phrase *Pierre n'est pas idiot* est suivie de la phrase (a) avec le connecteur ϵ, on ne sait pas si (a) est relié à la première phrase par la relation de discours $Résultat$ ou par $Évidence$. En revanche, avec la phrase (b) introduite par le connecteur *du coup*, on sait que la relation de discours est $Résultat$ et avec (c) introduit par *après tout*, on sait que c'est $Évidence$.

(1) Pierre n'est pas idiot.
(a) ϵ Il peut trouver son chemin tout seul. [$Résultat$ ou $Évidence$]
(b) *Du coup*, il peut trouver son chemin tout seul. [$Résultat$]
(c) *Après tout*, il peut trouver son chemin tout seul. [$Évidence$]

Les connecteurs de discours forment une classe semi-fermée (semi-ouverte) : ainsi le français compte environ 330 connecteurs selon la base lexicale LexConn (Roze, 2009; Roze *et al.*, 2010), ce qui contraste avec les classes ouvertes (V, N, Adj, Adv) qui comptent des milliers d'éléments chacune et les classes fermées (Prép, Pro, Det, ...) qui comptent moins d'une centaine d'éléments chacune. LexConn [1] est une base lexicale des connecteurs discursifs du français, dans laquelle est renseignée, pour chaque connecteur, sa catégorie morpho-syntaxique et la ou les relation(s) de discours qu'il exprime [2]. Les connecteurs ont été identifiés grâce à des critères syntaxiques, sémantiques et discursifs, appliqués à une liste de conjonctions de subordination fournie par Eric Laporte et une

1. LexConn est librement accessible à `www.linguist.univ-paris-diderot.fr/~croze/D/Lexconn.xml`
2. Un connecteur peut exprimer plusieurs relations de discours. C'est pourquoi les 328 connecteurs donnent lieu à 428 emplois.

Découverte de patrons paraphrastiques en corpus comparable: une approche basée sur les n-grammes

Bruno Cartoni[1] Louise Deléger[2]
(1)Département de Linguistique, Université de Genève
(2)Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center
bruno.cartoni@unige.ch, louise.deleger@cchmc.org

Résumé. Cet article présente l'utilisation d'un corpus comparable pour l'extraction de patrons de paraphrases. Nous présentons une méthode empirique basée sur l'appariement de n-grammes, permettant d'extraire des patrons de paraphrases dans des corpus comparables d'une même langue (le français), du même domaine (la médecine) mais de registres de langues différents (spécialisé ou grand public). Cette méthode confirme les résultats précédents basés sur des méthodes à base de patrons, et permet d'identifier de nouveaux patrons, apportant également un regard nouveau sur les différences entre les discours de langue générale et spécialisée.

Abstract. This paper presents the use of a comparable corpus for extracting paraphrase patterns. We present an empirical method based on n-gram matching and ordering, to extract paraphrase pattern in comparable corpora of the same language (French) and the same domaine, but of two different registers (lay and specialised). This method confirms previous results from pattern-based methods, and identify new patterns, giving fresh look on the difference between specialised and lay discourse.

Mots-clés : Identification de paraphrases, extraction de patrons, type de discours, domaine médical, corpus comparable monolingue.

Keywords: paraphrase identification, lexico-syntactic pattern discovery, discourse type, medical domain, monolingual comparable corpora.

1 Introduction

Cet article présente une étude basée sur un corpus comparable composé de textes de deux types de discours différents (spécialisé ou grand public) mais d'un même domaine (médecine) et dans une même langue (français) permettant d'explorer empiriquement le phénomène de la paraphrase et de valider des travaux antérieurs. On peut définir les paraphrases comme des expressions linguistiques possédant une signification similaire. La compréhension des mécanismes de paraphrase est un élément-clé de nombreuses applications du TALN comme l'extraction d'information (Shinyama & Sekine, 2003), le résumé automatique (Barzilay, 2003) et la simplification de textes (Elhadad & Sutaria, 2007). Différentes approches ont été employées pour la détection de paraphrases : elles peuvent varier au niveau du type de corpus utilisés (parallèle (Barzilay & McKeown, 2001), comparable (Shinyama & Sekine, 2003)), du type de paraphrases recherchées (phrastique (Barzilay & Lee, 2003), sous-phrastique (Elhadad & Sutaria, 2007)) et du type de technique employée (alignement de graphes lexicaux (Barzilay & Lee, 2003), similarité distributionnelle (Elhadad & Sutaria, 2007), patrons lexico-syntaxiques (Jacquemin, 1999)). Parmi les approches en corpus comparable, beaucoup s'appuient sur des corpus journalistiques relatant les mêmes informations mais provenant de différentes sources (Barzilay & Lee, 2003; Shinyama & Sekine, 2003). Dans le domaine médical, (Elhadad & Sutaria, 2007) travaillent sur un corpus comparable presque parallèle, composé d'articles scientifiques et de leur version "grand public", pour extraire des paraphrases entre deux types de discours, grand public vs. spécialisé. Dans nos précédents travaux (Deléger & Zweigenbaum, 2008; Deléger & Cartoni, 2010), nous avons également étudié l'extraction de paraphrases entre ces deux types de discours, à l'aide de patrons lexicaux pré-définis basés sur des ancrages de type morphosémantique (verbe / nom déverbal, adjectif relationnel / nom). Dans la présente étude, nous prolongeons ces travaux en adoptant une approche moins "supervisée" permettant de découvrir de nouveaux patrons de paraphrases, et de confirmer la pertinence des paraphrases utilisées dans nos approches à base de patrons. Cette approche se fonde sur le repérage d'identité entre des n-grammes

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

Prise en compte de la sous-catégorisation verbale dans un lexique bilingue anglais-japonais

Alexis Kauffmann
LATL, Université de Genève
2, Rue de Candolle, 1211 Genève, Suisse
alexis.kauffmann@unige.ch

Résumé. Dans cet article, nous présentons une méthode de détection des correspondances bilingues de sous-catégorisation verbale à partir de données lexicales monolingues. Nous évoquons également la structure de ces lexiques et leur utilisation en traduction automatique (TA) à base linguistique anglais-japonais. Les lexiques sont utilisés par un programme de TA fonctionnant selon une architecture classique dite "à transfert", et leur structure permet une classification précise des sous-catégorisations verbales. Nos travaux ont permis une amélioration des données de sous-catégorisation des lexiques pour les verbes japonais et leurs équivalents anglais, en utilisant des données linguistiques compilées à partir d'un corpus de textes extrait du web. De plus, le fonctionnement du programme de TA a pu être amélioré en utilisant ces données.

Abstract. In this paper, we present a method for the detection of bilingual correspondences of verb subcategorization from monolingual lexical data. We also mention the structure of the lexicons and examples making use of such data in linguistics-based English-Japanese machine translation (MT). The lexicons are used by a MT system with a classical transfer-based architecture, and their structure allow an accurate classification of verb subcategorization. Our work has improved the lexical data about subcategorization of Japanese verbs and their English equivalents, using linguistic data compiled from a corpus of web extracted texts. Furthermore, the MT system could also be improved by the use of this data.

Mots-clés : bases de données lexicales, sous-catégorisation verbale, traduction automatique à base linguistique, japonais.

Keywords: lexical databases, verb subcategorisation, linguistics-based machine translation, Japanese.

1 Introduction

Connaître la sous-catégorisation des verbes [1] peut être utile en traduction automatique afin d'améliorer la traduction des verbes et de leurs différents arguments. Ce sujet a souvent donné lieu a la création de fichiers monolingues (Raza, 2010), (Kawahara & Kurohashi, 2010) et aussi de fichiers multilingues (Mangeot & Kuroda, 2003).

Nous allons aborder ici le problème de la détection automatique et de l'enregistrement de telles données dans des bases de données lexicales, à un niveau monolingue et surtout au niveau bilingue. Nous verrons aussi comment de telles données peuvent être utilisées en TA, avec le programme de TA à base linguistique Its-2, en traduction anglais-japonais.

L'article est organisé ainsi : dans la deuxème partie, nous décrirons brièvement l'architecture du programme de TA Its-2 et la structure de ses bases de données lexicales ; dans la troisième partie, nous présenterons le fichier lexical "Case Frames" qui décrit des structures argumentales de verbes japonais ; ensuite, dans la quatrième partie, nous présenterons notre méthode de détection des correspondances bilingues de sous-catégorisations verbales et la mise à jour de nos lexiques ; enfin, en cinquième partie, nous évoquerons deux améliorations apportées au programme de TA grâce aux données sur la sous-catégorisation.

1. La sous-catégorisation verbale (ou "cadres de valence syntaxique") décrit le comportement des verbes à un niveau syntaxique, en connaissant l'ensemble de leurs compléments. Cela permet de savoir si un verbe est transitif ou intransitif, s'il peut prendre une phrase comme complément, s'il peut prendre un complément d'objet indirect, etc.

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

Extraction non-supervisée de relations basée sur la dualité de la représentation

Yayoi Nakamura-Delloye[1]
(1) ALPAGE, INRIA-Rocquencourt
Domaine de Voluceau Rocquencourt B.P.105 78153 Le Chesnay
yayoi@yayoi.fr

Résumé. Nous proposons dans cet article une méthode non-supervisée d'extraction des relations entre entités nommées. La méthode proposée se caractérise par l'utilisation de résultats d'analyses syntaxiques, notamment les chemins syntaxiques reliant deux entités nommées dans des arbres de dépendance. Nous avons également exploité la dualité de la représentation des relations sémantiques et le résultat de notre expérience comparative a montré que cette approche améliorait les rappels.

Abstract. We propose in this paper an unsupervised method for relation and pattern extraction. The proposed method is characterized by using parsed corpora, especially by leveraging syntactic paths that connect two named entities in dependency trees. We also use the dual representation of semantic relations and the result of our comparative experiment showed that this approach improves recall.

Mots-clés : Extraction des connaissances, relations entre entités nommées, dualité relationnelle.

Keywords: Knowledge extraction, named entity relationships, relational duality.

1 Introduction

L'extraction de relations entre entités nommées (EN ci-après) est une opération importante pour beaucoup d'applications et de nombreuses études ont été proposées dans différents cadres de travail tels que la conception d'un système de question-réponse (Iftene & Balahur-Dobrescu, 2008), l'extraction d'information (Banko *et al.*, 2007) ou l'extraction de réseaux sociaux (Matsuo *et al.*, 2006). Nous proposons dans cet article une méthode non-supervisée d'extraction de relations entre EN à partir de résultats d'analyses syntaxiques. Nos travaux ont été menés en vue du peuplement du référentiel ontologique constitué dans le cadre d'expériences d'enrichissement sémantique de dépêches de l'Agence France Presse (Stern & Sagot, 2010).

De nombreuses méthodes supervisées d'acquisition de relations basées sur des grands corpus annotés telles que (Zelenko *et al.*, 2002), ont été proposées. Un de leurs plus grands défauts est le coût élevé pour la réalisation de l'annotation. Les approches semi-supervisées se fondent généralement sur un principe d'« induction » qui recourt à un petit ensemble d'exemples de relations (Hearst, 1992) (Brin, 1998) (Agichtein & Gravano, 2000). Mais, les difficultés de déterminer préalablement des relations intéressantes, et de trouver des exemples pertinents pour ces relations constituent des inconvénients de ces méthodes semi-supervisées. Les travaux de (Hasegawa *et al.*, 2004) ont proposé une méthode non-supervisée qui écarte ces problèmes de prédéfinition des relations à extraire. Elle est constituée de deux grandes étapes : *clustering* selon les contextes partagés et étiquetage des *clusters* par

Vers la détection des dislocations à gauche dans les transcriptions automatiques du Français parlé

Corinna Anderson[1] Christophe Cerisara[1] Claire Gardent[1]
(1) LORIA-CNRS UMR 7503, Campus Scientifique, Vandoeuvre-les-Nancy
{andersoc,cerisara,gardent}@loria.fr

Résumé. Ce travail prend place dans le cadre plus général du développement d'une plate-forme d'analyse syntaxique du français parlé. Nous décrivons la conception d'un modèle automatique pour résoudre le lien anaphorique présent dans les dislocations à gauche dans un corpus de français parlé radiophonique. La détection de ces structures devrait permettre à terme d'améliorer notre analyseur syntaxique en enrichissant les informations prises en compte dans nos modèles automatiques. La résolution du lien anaphorique est réalisée en deux étapes : un premier niveau à base de règles filtre les configurations candidates, et un second niveau s'appuie sur un modèle appris selon le critère du maximum d'entropie. Une évaluation expérimentale réalisée par validation croisée sur un corpus annoté manuellement donne une F-mesure de l'ordre de 40%.

Abstract. Left dislocations are an important distinguishing feature of spoken French. In this paper, we present a hybrid approach for detecting the coreferential link that holds between left-dislocated elements and the coreferential pronoun occurring further on in the sentence. The approach combines a symbolic graph rewrite step with a maximum entropy classifier and achieves around 40% F-score. We conjecture that developing such approaches could contribute to the general anaphora resolution task and help improve parsers trained on corpora enriched with left dislocation anaphoric links.

Mots-clés : Détection des dislocations à gauche, Maximum Entropy, français parlé.

Keywords: Left dislocation detection, Maximum Entropy, spoken French.

1 Introduction

Les dislocations sont considérées depuis longtemps comme des caractéristiques importantes du Français parlé (De Cat, 2007; Delais-Roussarie *et al.*, 2004; Hirschbuhler, 1975; Lambrecht, 1994). Bien qu'elles apparaissent plus fréquemment dans la parole spontanée, il n'est pas rare de les trouver également dans de nombreux autres registres du français oral, et ce depuis au moins le XVII^{ème} siècle (De Cat, 2007; Blanche-Benveniste, 1997).

Une caractéristique liée à la définition des dislocations à gauche est qu'elles impliquent un lien coréférentiel entre l'élément disloqué and un pronom situé plus loin dans la phrase. Ainsi, le traitement des dislocations à gauche contribue au domaine plus général qui est celui de la résolution des anaphores. De plus, l'annotation automatique ou semi-automatique de ces liens coréférentiels constitue une information potentiellement utile pour améliorer la qualité des analyseurs syntaxiques automatiques de l'oral.

Nous nous intéressons dans cet article à ce problème et présentons une approche hybride de résolution des liens de coréférence entre les éléments disloqués à gauche et le pronom correspondant. Nous décrivons au paragraphe 2 la syntaxe des dislocations à gauche en l'illustrant avec des exemples extraits du corpus radiophonique ESTER du français parlé, initialement conçu pour les campagnes d'évaluation nationales des systèmes de transcription automatique de la parole (Gravier *et al.*, 2004). Notre méthodologie est ensuite décrite au paragraphe 3, ainsi que les résultats expérimentaux. Le paragraphe 4 conclut l'article en présentant notamment quelques perspectives.

2 Dislocation à gauche : syntaxe et coréférence

Une dislocation à gauche peut être définie comme une expression située dans un voisinage gauche d'une proposition qui contient un pronom présentant un lien de coréférence avec cette expression, comme l'illustre l'exemple (Ex.1) dans le cas d'une phrase simple, et les exemples (Ex.2,3) pour une proposition subordonnée. Notons qu'une séparation hiérarchique entre le constituant disloqué et le pronom comme dans (Ex.3) n'a aucun effet sur la coréférence.

Règles et paradigmes en morphologie informatique lexématique

Nabil Hathout(1) Fiammetta Namer(2)

(1) UMR CLLE-ERSS, Toulouse
(2) UMR ATILF-Université Nancy2, Nancy
nabil.hathout@univ-tlse2.fr, fiammetta.namer@univ-nancy2.fr

Résumé.

Les familles de mots produites par deux analyseurs morphologiques, DériF (basé sur des règles) et Morphonette (basé sur l'analogie), appliqués à un même corpus lexical, sont comparées. Cette comparaison conduit à l'examen de trois sous-ensembles :

- un sous-ensemble commun aux deux systèmes dont la taille montre que, malgré leurs différences, les approches expérimentées par chaque système sont valides et décrivent en partie la même réalité morphologique.

- un sous-ensemble propre à DériF et un autre à Morphonette. Ces ensembles (a) nous renseignent sur les caractéristiques propres à chaque système, et notamment sur ce que l'autre ne peut pas produire, (b) ils mettent en évidence les erreurs d'un système, en ce qu'elles n'apparaissent pas dans l'autre, (c) ils font apparaître certaines limites de la description, notamment celles qui sont liées aux objets et aux notions théoriques comme les familles morphologiques, les bases, l'existence de RCL « transversales » entre les lexèmes qui n'ont pas de relation d'ascendance ou de descendance.

Abstract.

The word families produced by two morphological analyzers of French, DériF (rule-based) and Morphonette (analogy-based), applied on the same lexical corpus have been compared. The comparison led us to examine three classes of relations:

- one subset of relations that are shared by both systems. It shows that, despite their differences, the approaches implemented in these systems are valid and describe, to some extent, one and the same morphological reality.

- one subset of relations specific to DériF and another one to Morphonette. These sets (a) give us informations on the characteristics proper to each system, and especially on what the other system is unable to produce; (b) they highlight the errors of one system, in so that they are absent from the results of the other; (c) they reveal some of the limits of the description, especially the ones related to theoretical objects and concepts such as morphological family, base or the existence of transverse LCR (lexeme construction rules) between lexemes that are not ascendant nor descendant of each other.

Mots-clés : morphologie constructionnelle ; analyse automatique ; règles ; analogie ; familles morphologiques ; comparaison ; synergie
Keywords: Word formation ; automatic analysis ; rules ; analogy ; morphological families ; comparison ; synergy.

1 Introduction

Cet article présente une comparaison entre deux ressources morphologiques très différentes : l'analyseur DériF et le réseau morphologique Morphonette. DériF (Namer, 2009) est un analyseur qui implémente des règles de construction de lexèmes (RCL) établies et mises au point manuellement. Leur application est contrôlée par un ensemble d'exceptions qui permettent de prendre en compte efficacement l'ensemble des irrégularités qui se sont accumulées au cours de

] **Inga Gheorghita** *Méthodologie de construction automatique du thésaurus pour l'indexation et la recherche des images (RECITAL)*

Méthodologie de construction automatique du thésaurus pour l'indexation et la recherche des images

Inga Gheorghita[1,2]

(1) ATILF-CNRS, Nancy-Université (UMR 7118), France
(2) XILOPIX, 37 rue de la Plaine, 75020 Paris, France
inga.gheorghita@atilf.fr

Cet article présente une méthodologie de construction automatique du thésaurus à l'aide du Trésor de la Langue Française informatisé (TLFi). Nous utilisons les définitions du TLFi pour désambiguïser et enrichir les mots-clés présents dans les descriptions textuelles associées aux images, en construisant un arbre hiérarchique. L'approche proposée peut être utilisée pour la catégorisation très précise d'images, pour l'indexation de grandes quantités d'images et pour la recherche.

This article presents a methodology for automatic thesaurus construction using the "Trésor de la Langue Française informatisé" (TLFi). We use the definitions of TLFi to disambiguate and expand the keywords of image's textual descriptions, by building a hierarchical tree. The proposed approach can be used for accurate categorization of images, the indexation of large amounts of images and search.

Mots-clés : thésaurus automatique, TLFi, indexation, recherche, images
Keywords: automatic thesaurus, TLFi, indexation, search, images

1 Introduction

Avec l'arrivée de l'Internet le marché de l'image numérique a progressé de manière exponentielle. L'offre d'illustrations n'a jamais été aussi grande. Sachant qu'une agence de photo gère habituellement entre un et vingt millions d'images, qu'un satellite météo envoie plusieurs giga-octets de données chaque jour, qu'un possesseur d'appareil photo numérique actif prendra de l'ordre de cents mille photos en trente ans (Gros, 2007), l'accès et la recherche dans cette masse d'informations énorme posent de nouveaux défis. L'organisation non structurée des images provoque un très grand désordre et une extraordinaire confusion lorsque l'on cherche à les identifier ou les repérer. L'une des causes est que les descriptions textuelles associées aux images ne sont souvent pas suffisantes pour permettre leur structuration. Afin de gérer et d'utiliser efficacement ces bases d'images, un système d'indexation et de recherche est donc nécessaire. C'est pour cette raison que la recherche d'images est devenue un sujet très actif dans la communauté internationale et a connu un véritable engouement au cours des deux dernières décennies.

lundi 27 juin, 10h45-12h15

La complexité linguistique Méthode d'analyse

Adrien Barbaresi
ICAR, ENS LYON

Résumé. La complexité linguistique regroupe différents phénomènes dont il s'agit de modéliser le rapport. Le travail en cours que je décris ici propose une réflexion sur les approches linguistiques et techniques de cette notion et la mise en application d'un balayage des textes qui s'efforce de contribuer à leur enrichissement. Ce traitement en surface effectué suivant une liste de critères qui représentent parfois des approximations de logiques plus élaborées tente de fournir une image « raisonnable » de la complexité.

Abstract. Linguistic complexity includes various linguistic phenomena which interaction is to be modeled. The ongoing work described here tackles linguistic and technical approaches of this idea as well as an implementation of a parsing method which is part of text enrichment techniques. This chunk parsing is performed according to a list of criteria that may consist in logical approximations of more sophisticated processes in order to provide a « reasonable » image of complexity..

Mots-clés : Complexité, lisibilité, allemand, analyse de surface.

Keywords: Complexity, lisibility, German, chunk parsing.

1 Enjeux

L'analyse de la complexité se situe dans le cadre de l'assistance à la compréhension. Il s'agit ici de déterminer la lisibilité d'un texte pour des humains ou pour des machines, c'est-à-dire d'une part le niveau de maîtrise et de pratique de la langue requis et d'autre part le modèle formel et les instruments à utiliser.

Ce thème très riche invite à penser les langues avant de les analyser à différentes échelles et selon différents modes opératoires. Du point de vue disciplinaire, on peut le situer à la croisée de la linguistique, des études sur la lisibilité, des sciences cognitives et de la théorie de l'information. Le traitement envisagé aborde des notions d'informatique et l'intégration d'un marquage des textes étudiés, approches qui sont en prise directe avec la réflexion actuelle chez les chercheurs et les entrepreneurs sur les données, leur statut, leur forme et leur traitement.

Cette démarche s'inscrit en ce sens entre la réflexion en sciences humaines et l'exploitation technique. Elle est également en rapport avec la transmission d'une langue et son « outillage » (Auroux, 1994). En effet, au-delà d'une tentative consistant à modéliser les processus à l'œuvre lors du déchiffrage d'un texte, il s'agit d'équiper une langue, de l'enrichir d'une description utile.

De fait, pour (Gibson, 1998), étudier la complexité linguistique, c'est expliquer les étapes de l'apprentissage de sa langue maternelle par un enfant, donner des éléments pour aborder les problèmes syntaxiques chez les aphasiques, et fournir des applications dans lesquelles la compréhensibilité de la langue est importante, comme les correcteurs grammaticaux ou la génération automatique de textes.

L'intérêt premier porte sur la complexité de phrases, de paragraphes ou de textes écrits en allemand. L'attention portera spécifiquement sur un standard de cette langue, considéré comme une *koinè*, une langue commune qui dépasse des disparités régionales.

Il ne s'agit donc pas d'un travail de comparaison entre différentes langues. En revanche, la pertinence de la notion de sous-langage pourra être examinée. De même, dans un deuxième temps, une adaptation de la démarche et des outils à l'anglais et au français apportera peut-être quelques éléments qui viendront enrichir la compréhension du sujet en infirmant ou confirmant des hypothèses.

Bidirectional Sequence Classification for Tagging Tasks with Guided Learning

Andrea Gesmundo
Université de Genève, route de Drize 7, 1227 Genève
andrea.gesmundo@unige.ch

Résumé. Dans cet article nous présentons une série d'adaptations de l'algorithme du "cadre d'apprenstissage guidé" pour résoudre différentes tâches d'étiquetage. La spécificité du système proposé réside dans sa capacité à apprendre l'ordre de l'inférence avec les paramètres du classifieur local au lieu de la forcer dans un ordre pré-défini (de gauche à droite). L'algorithme d'entraînement est basé sur l'algorithme du "perceptron". Nous appliquons le système à différents types de tâches d'étiquetage pour atteindre des résultats au niveau de l'état de l'art en un court temps d'exécution.

Abstract. In this paper we present a series of adaptations of the Guided Learning framework to solve different tagging tasks. The specificity of the proposed system lies in its ability to learn the order of inference together with the parameters of the local classifier instead of forcing it into a pre-defined order (left-to-right). The training algorithm is based on the Perceptron Algorithm. We apply the system to different kinds of tagging tasks reaching state of the art results with short execution time.

Mots-clés : Bidirectionnel, Classification de Séquence, Apprentissage Guidé.

Keywords: Bidirectional, Sequence Classification, Guided Learning.

1 Introduction

The system described in this paper carries out tagging tasks with semi-supervised training.We extend to the Guided Learning (GL) framework presented in (Shen *et al.*, 2007). This approach has been applied in the past to POS tagging task with excellent results. One of the aims of this paper is to show that GL can be adapted to solve a wide set of tagging and chunking tasks obtaining good performances with short execution time. This framework is more complex than supervised learning. The system can learn the parameters for the local classifier from gold standard labels, but has no indications on the order of inference. Basing the learning algorithm on the Perceptron scheme allows one to keep a low system complexity and moderate execution time, without sacrificing learning capability and quality of the results. Compared to other systems that use a Perceptron algorithm, such as (Collins, 2002), GL introduces a bidirectional search strategy. Instead of forcing the order of the tagging in a left-to-right fashion, any tagging order is allowed. GL follows an easiest-first approach and incorporates the learning of the order of inference in the training phase. In this way right-context and bidirectional-context features can be used at little extra cost. In a direct comparison with (Collins, 2002) we show that it is possible to achieve better accuracy with shorter execution time allowing the inference order to be predicted by the system instead of using an exhaustive search strategy.

We test the effectiveness of this approach applying it to different tagging tasks, taking part in shared tasks or experimenting on widely used corpora, this allows us to make a comparison between our system and the state of the art. The tasks we focus on are : Part of Speech Tagging, Noun Phrase Chunking, and Named Entity Recognition. NP chunking and NER are defined as chunking tasks, but following the general guidelines of (Ramshaw & Marcus, 1995) we can solve these problems as tagging tasks. For the chunking tasks, we apply a voting system between multiple data representations of text chunks (Shen & Sarkar, 2005).

2 Bidirectional Guided Classification

The input of the Inference Algorithm is a sequence of tokens $t_1 t_2 \cdots t_n$. For each token t_i, we have to assign a label $l_i \in L$, with L being the label set. A subsequence $t_i \cdots t_j$ is called a span, and is denoted $[i, j]$. To each span

Calcul de réseaux phrastiques pour l'analyse et la navigation textuelle

Dominique Legallois[1] Peggy Cellier[2] Thierry Charnois[3]
(1) CRISCO Université de Caen Basse-Normandie, Campus 1, 14000 Caen
(2) IRISA-INSA de Rennes, Campus Beaulieu 35042 Rennes cedex
(3) GREYC Université de Caen Basse-Normandie, Campus 2, 14000 Caen

Résumé. Le travail présente une méthode de navigation dans les textes, fondée sur la répétition lexicale. La méthode choisie est celle développée par le linguiste Hoey. Son application manuelle à des textes de grandeur conséquente est problématique. Nous proposons dans cet article un processus automatique qui permet d'analyser selon cette méthode des textes de grande taille ; des expériences ont été menées appliquant le processus à différents types de textes (narratif, expositif) et montrant l'intérêt de l'approche.

Abstract. In this paper, we present an automatic process based on lexical repetition introduced by Hoey. The application of that kind of approaches on large texts is difficult to do by hand. In the paper, we propose an automatic process to treat large texts. We have conducted some experiments on different kinds of texts (narrative, expositive) to show the benefits of the approach.

Mots-clés : Réseau phrastique, Appariement de phrases, Analyse textuelle, Navigation textuelle.

Keywords: Sentence network, Bonds between sentences, Textual analysis, Textual navigation.

1 Introduction

Notre travail propose une analyse des textes, fondée à la fois sur la « réduction textuelle » et la répétition lexicale. Nous nous inspirons du modèle linguistique de Hoey (Hoey, 1991). Par « réduction », nous entendons une méthode linguistique qui vise à déterminer dans un texte, quelles sont les phrases les plus pertinentes informationnellement. L'ensemble de ces phrases délesté des phrases informationnellement marginales, constitue un ensemble théoriquement cohérent. Contrairement aux travaux sur les résumés automatiques, [1] nous définissons une méthode qui s'inscrit dans une démarche expérimentale visant, en même temps, à mieux comprendre l'organisation textuelle et à proposer un mode de navigation. Nous ne connaissons que peu d'analyses linguistiques proposant une telle démarche. Outre Hoey, mentionnons les travaux de Thomas (Thomas, 1999) dont l'analyse est fondée sur la proéminence sémantique des propositions, et de Toolan (Toolan, 2009) qui travaille exclusivement sur les textes narratifs afin de conserver les phrases constituants les nœuds importants de l'histoire. Le modèle de Hoey, que nous avons appliqué au français (Legallois, 2004, 2006), ne prend en compte que les textes expositifs (par ex. les textes scientifiques, philosophiques) et exclut en principe le genre narratif [2]. De plus, les appariements de phrases ayant des lexèmes en commun ne peuvent être repérés que sur des textes courts. En effet, l'application manuelle de la méthode sur des textes de plusieurs milliers de lignes est extrêmement problématique.

Dans cet article, nous proposons d'implémenter l'adaptation au français de la méthode de Hoey dans un processus automatisé afin de pouvoir analyser des corpus variés en genre et en taille. Cette mise en œuvre informatique permet notamment au linguiste d'observer et de tester l'application du modèle en corpus. L'implémentation concerne uniquement la répétition lexicale au sens strict (i.e., le même lemme), qui constitue, d'après nos observations, le cas de répétition le plus fréquent. La répétition par anaphore et la synonymie ne sont pas prises en compte [3]. Toutefois, cette limitation n'entrave en rien la pertinence du modèle pour la recherche en linguistique comme les expériences le montrent. Dans la suite de l'article nous présentons le modèle linguistique (Section 2). Puis nous

1. *Cf.* (Knight & Marcu, 2002) ou les récents systèmes présentés aux compétitions TAC 2008-2010 (`www.nist.gov/tac/tracks/`)
2. Nous verrons plus bas que nous avons fait une entorse à ce principe.
3. Dans l'état actuel des techniques, l'identification de la répétition par anaphore reste un problème non résolu sur un texte long.

Exploitation d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques

Achille Falaise (1), Agnès Tutin (2), Olivier Kraif (2)

(1) GETALP-LIG
(2) LIDILEM
achille.falaise@imag.fr, agnes.tutin@u-grenoble3.fr, olivier.kraif@u-grenoble3.fr

Résumé

L'exploitation de corpus analysés syntaxiquement (ou corpus arborés) pour le public non spécialiste n'est pas un problème trivial. Si la communauté du TAL souhaite mettre à la disposition des chercheurs non-informaticiens des corpus comportant des annotations linguistiques complexes, elle doit impérativement développer des interfaces simples à manipuler mais permettant des recherches fines. Dans cette communication, nous présentons les modes de recherche « grand public » développé(e)s dans le cadre du projet Scientext, qui met à disposition un corpus d'écrits scientifiques interrogeable par partie textuelle, par partie du discours et par fonction syntaxique. Les modes simples sont décrits : un mode libre et guidé, où l'utilisateur sélectionne lui-même les éléments de la requête, et un mode sémantique, qui comporte des grammaires locales préétablies à l'aide des fonctions syntaxiques.

Abstract

The exploitation of syntactically analysed corpora (or treebanks) by non-specialist is not a trivial problem. If the NLP community wants to make publicly available corpora with complex annotations, it is imperative to develop simple interfaces able to handle advanced queries. In this paper, we present queries methods for the general public developed during the Scientext project, which provides a searchable corpus of scientific texts searchable from textual part, part of speech and syntactic relation. The simple query modes are described: a guided query mode, where the user easily selects the elements of the query, and a semantic mode which includes local pre-established grammars using syntactic functions.

Mots-clés : environnement d'étude de corpus, corpus étiquetés et arborés, création de grammaires assistée, visualisation d'information linguistique
Keywords: corpus study environment, treebanks, assisted grammars creation, visualization of linguistic information

1 Introduction

Les outils d'exploration de corpus annotés, en particulier de corpus arborés (c'est-à-dire comportant des relations syntaxiques), sont souvent complexes à utiliser, *a fortiori* pour des utilisateurs non initiés à la linguistique-informatique. L'ergonomie et la facilité d'utilisation des outils sont cependant des enjeux majeurs en TAL, surtout si l'on souhaite diffuser des traitements et des annotations linguistiques complexes dans la communauté des linguistes. Pour élargir le nombre d'utilisateurs des corpus annotés, il est essentiel de développer des outils d'exploration de corpus faciles à manipuler mais puissants. C'est ce qui nous a amenés à proposer un environnement de recherche simple, adapté aux linguistes, didacticiens, lexicographes ou épistémologues.

Communautés Internet comme sources de préterminologie

Mohammad Daoud, Christian Boitet

Laboratoire LIG — Université Joseph Fourier — 385, rue de la Bibliothèque, 38041 Grenoble, France
{Mohammad.Daoud, Christian.Boitet }@imag.fr

Résumé

Cet article décrit deux expériences sur la construction de ressources terminologiques multilingues (preterminologies) préliminaires, mais grandes, grâce à des communautés Internet, et s'appuie sur ces expériences pour cibler des données terminologiques plus raffinées venant de communautés Internet et d'applications Web 2.0. La première expérience est une passerelle de contribution pour le site Web de la Route de la Soie numérique (DSR). Les visiteurs contribuent en effet à un référentiel lexical multilingue dédié, pendant qu'ils visitent et lisent les livres archivés, parce qu'ils sont intéressés par le domaine et ont tendance à être polygottes. Nous avons recueilli 1400 contributions lexicales en 4 mois. La seconde expérience est basée sur le JeuxDeMots arabe, où les joueurs en ligne contribuent à un réseau lexical arabe. L'expérience a entraîné une croissance régulière du nombre de joueurs et de contributions, ces dernières contenant des termes absents et des mots de dialectes oraux.

Mots-clés: terminologie, préterminologie, approches collaboratives, réseaux lexicaux, DSR, jeux sérieux.

Abstract

This paper describes two experiments on building preliminary but large multilingual terminological resources (preterminologies) through Internet communities, and draws on these experiments to target more refined terminological data from Internet communities and Web 2.0 applications. The first experiment is a contribution gateway for the Digital Silk Road (DSR) website. Visitors indeed contribute to a dedicated multilingual lexical repository while they visit and read the archived books, because they are interested in the domain and tend to be multilingual. We collected 1400 lexical contributions in 4 months. The second experiment is based on the Arabic JeuxDeMots, where online players contribute to an Arabic lexical network. The experiment resulted in a steady growth of number of players and contributions, the latter containing absent terms and spoken dialectic words.

Keywords: terminology, preterminology, collaborative approaches, lexical networks, DSR, serious games.

1 Introduction

Construire des ressources terminologiques multilingue pour un domaine est une tâche difficile et compliquée, car la terminologie représente la structure conceptuelle d'un domaine en utilisant un ensemble d'unités lexicales dans une langue particulière. Cette structure conceptuelle est plus dynamique et change plus vite que sa représentation symbolique (terminologie). En outre, toutes les communautés de langues différentes ne partagent pas le même intérêt dans un domaine donné. Par exemple, l'arabe est considéré comme une langue pauvrement dotée en ressources linguistiques (Yassin, 2003) (Diab, Habash, 2009), en particulier dans les domaines de la science et de la technologie, tandis que l'anglais et le français ont une terminologie beaucoup plus riche dans ces domaines. Classiquement, pour construire des ressources terminologiques, une banque de termes est construite par des terminologues: le domaine considéré est étudié et des documents sont consultés pour en extraire les termes pertinents. Cette approche dépend fortement de terminologues et de financement par de grandes organisations, elle est donc très coûteuse.

De nombreux chercheurs ont été à la recherche d'alternatives (Kageura, Umino, 1998) (Joubert, Lafourcade, 2008) (Nagata et al., 2001). Les approches contributives (Ahn, 2005) à la collecte de connaissances semblent prometteuses à

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

Évaluation de G-LexAr pour la traduction automatique statistique

Wigdan Mekki[(1)], Julien Gosme[(1)], Fathi Debili[(2)], Yves Lepage[(3)], Nadine Lucas[(1)]
(1) GREYC, UMR 6072, CNRS, Université de Caen Basse-Normandie, Caen, France
(2) LLACAN, UMR 8135, CNRS, Villejuif, France
(3) IPS, Université Waseda, Japon

Résumé. G-LexAr est un analyseur morphologique de l'arabe qui a récemment reçu des améliorations substantielles. Cet article propose une évaluation de cet analyseur en tant qu'outil de pré-traitement pour la traduction automatique statistique, ce dont il n'a encore jamais fait l'objet. Nous étudions l'impact des différentes formes proposées par son analyse (voyellation, lemmatisation et segmentation) sur un système de traduction arabe-anglais, ainsi que l'impact de la combinaison de ces formes. Nos expériences montrent que l'utilisation séparée de chacune de ces formes n'a que peu d'influence sur la qualité des traductions obtenues, tandis que leur combinaison y contribue de façon très bénéfique.

Abstract. G-LexAr is an Arabic morphological analyzer that has recently been improved for speed. This paper gives an assessment of this analyzer as a preprocessing tool for statistical machine translation. We study the impact of the use of its possible outputs (vocalized, lemmatized and segmented) through an Arabic-English machine translation system, as well as the impact of the combination of these outputs. Our experiments show that using these outputs separately does not influence much translation quality. However, their combination leads to major improvements.

Mots-clés : traduction automatique statistique, analyse morphologique, pré-traitement de l'arabe.

Keywords: statistical machine translation, morphological analysis, arabic preprocessing.

1 Introduction

L'arabe est une langue à morphologie riche dont la complexité présente des défis pour la traduction automatique (voir (Habash, 2007) pour une description des problèmes morphologiques relatifs à cette tâche).

Des expériences utilisant l'analyse morphologique pour améliorer la traduction automatique ont déjà été menées pour l'allemand (p. ex. Nießen & Ney, 2004) ou le turc (p. ex. Bisazza & Federico, 2009). Ces travaux utilisent diverses sortes de segmentation, lemmatisation et étiquetage grammatical. Dans le cas de l'arabe, les travaux de Lee (2004), utilisant une approche de segmentation en racines et affixes, puis de Habash & Sadat (2006), avec une approche de *tokenization* linguistiquement motivée, ont montré que le pré-traitement morphologique peut être utile à la traduction automatique statistique. D'un autre côté, Diab *et al.* (2007) ont montré que l'utilisation de la voyellation seule ne conduit à aucune amélioration (voyellation partielle), voire à de moins bons résultats (voyellation complète).

Dans cet article, nous évaluons l'analyseur morphologique de l'arabe G-LexAr (Debili *et al.*, 2002) sur des tâches de traduction automatique statistique arabe-anglais, en utilisant l'analyse morphologique comme étape de pré-

Enrichir la notion de patron par la prise en compte de la structure textuelle - Application à la construction d'ontologie

Marion laignelet[1] Mouna Kamel[2] Nathalie Aussenac-Gilles[2]
(1) CLLE-ERSS, Université de Toulouse 2, 5 allée A. Machado, 31058 Toulouse Cedex 9
(2) IRIT, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse Cedex 9
marion.laignelet@univ-tlse2.fr, kamel@irit.fr, aussenac@irit.fr

Résumé. La projection de patrons lexico-syntaxiques sur corpus est une des manières privilégiées pour identifier des relations sémantiques précises entre éléments lexicaux. Dans cet article, nous proposons d'étendre la notion de patron en prenant en compte la sémantique que véhiculent les éléments de structure d'un document (définitions, titres, énumérations) dans l'identification de relations. Nous avons testé cette hypothèse dans le cadre de la construction d'ontologies à partir de textes fortement structurés du domaine de la cartographie.

Abstract. Matching lexico-syntactic patterns on text corpora is one of the favorite ways to identify precise semantic relations between lexical items. In this paper, we propose to rely on text structure to extend the notion of pattern and to take into account the semantics that the structure (definitions, titles, item lists) may bear when identifying semantic relations between concepts. We have checked this hypothesis by building an ontology via highly structured texts describing spatial, i.e. geographical information.

Mots-clés : Construction d'ontologie, patron lexico-syntaxique, structure textuelle.

Keywords: Ontology engineering, lexico-syntactic patterns, textual structure.

1 Introduction

La projection de patrons lexico-syntaxiques sur corpus, utilisée pour l'extraction d'informations, s'applique également sur des corpus spécialisés pour la construction d'ontologies de domaines : on s'attend à trouver des traces linguistiques de concepts et de relations sémantiques binaires entre ces concepts. Des travaux en linguistique ont mis en évidence le rôle de la structure dans l'interprétation d'un texte (Luc & Virbel, 2001; Pascual & Péry-Woodley, 1997; Rebeyrolles *et al.*, 2009). Nous supposons que des informations issues de la structure textuelle peuvent être intégrées aux patrons de recherche de relations sémantiques, en plus des éléments lexicaux et syntaxiques. Nous avons testé cette hypothèse dans le cadre de la construction d'ontologies à partir de textes, en nous focalisant sur des éléments textuels particulièrement favorables à la recherche de relations ontologiques : les titres, les zones définitoires et les énumérations. Ces éléments permettent non seulement d'extraire des traces linguistiques pour définir des concepts et des relations pertinents pour notre domaine d'application mais également de résoudre certaines situations elliptiques, fréquentes dans nos données.

Le domaine d'étude de nos travaux s'inscrit dans le traitement automatique de textes structurés dans lesquels l'organisation même des éléments textuels reflète celle des concepts du domaine. Dans ces types de textes, la structure peut être explicite comme c'est le cas avec les versions électroniques de thésaurus ou de dictionnaires (Hearst, 1992) ou inférable à partir de leur mise en forme. En ce qui nous concerne, c'est à partir de la mise en forme matérielle des documents que les différentes classes (ou concepts) du domaine sont nommées et mises en relation les unes avec les autres. On trouve de tels types de textes dans des domaines spécifiques, comme la botanique qui se prête naturellement à la représentation de taxinomies : dans les travaux de (Role & Rousse, 2006), la structure des titres reflète un découpage en genres et en espèces et permet d'initialiser une hiérarchie de classes de l'ontologie.

Les textes de notre corpus appartiennent au domaine géographique [1] : ils décrivent les objets susceptibles d'intégrer des bases de données géographiques et cartographiques. Dans un premier temps, nous nous intéressons à

1. Projet Géonto, http ://geonto.lri.fr/

La traduction automatique des séquences clitiques dans un traducteur à base de règles. *

Lorenza Russo, Éric Wehrli
Laboratoire d'Analyse et de Technologie du Langage (LATL)
Département de linguistique – Université de Genève
2, rue de Candolle – CH-1211 Genève 4
{Lorenza.Russo, Eric.Wehrli}@unige.ch

Résumé. Dans cet article, nous discutons la méthodologie utilisée par Its-2, un système de traduction à base de règles, pour la traduction des pronoms clitiques. En particulier, nous nous focalisons sur les séquences clitiques, pour la traduction automatique entre le français et l'anglais. Une évaluation basée sur un corpus de phrases construites montre le potentiel de notre approche pour des traductions de bonne qualité.

Abstract. In this paper we discuss the methodology applied by Its-2, a rule-based MT system, in order to translate clitic pronouns. In particular, we focus on French clitic clusters, for automatic translation between French and English. An evaluation based on a corpus of constructed sentences shows the potential of this approach for high-quality translation.

Mots-clés : Analyseur syntaxique, traduction automatique, pronom clitique, séquences clitiques.

Keywords: Syntactic parser, automatic translation, clitic pronoun, clitic clusters.

1 Introduction

Le phénomène de la cliticisation des pronoms a suscité l'attention de très nombreux linguistes, en particulier suite aux travaux de Kayne (1975). Mais les pronoms clitiques sont importants aussi du point de vue du traitement automatique du langage naturel (TALN) en général et de la traduction automatique en particulier. Les systèmes de traduction automatique actuellement disponibles ne sont pas toujours capables de reconnaître les pronoms clitiques dans la langue source [1]. Ainsi, par exemple, dans le cas de séquences de pronoms clitiques, c'est-à-dire de l'occurrence de deux (ou plus) pronoms clitiques attachés au même hôte verbal, les systèmes de traduction automatique actuellement disponibles génèrent fréquemment des phrases cibles dans lesquelles seulement un des deux clitiques est traduit (1). Google Translate [2], par exemple, atteint un pourcentage très bas de traductions correctes des séquences clitiques sur un corpus d'exemples construits [3].

Au vu de ces résultats et compte tenu aussi de l'absence de travaux de recherche à ce sujet, notre but est celui de souligner ici l'importance de l'information lexicale et syntaxique pour le traitement de telles constructions afin d'obtenir des traductions automatiques syntaxiquement correctes et complètes. Pour cela, nous présentons dans cet article la méthodologie utilisée par Its-2 – un traducteur automatique multilingue à base de règles développé dans notre laboratoire - en nous focalisant en particulier sur la paire de langues français–anglais et sur les séquences clitiques.

∗. Le travail de recherche présenté ici a bénéficié du support du Fond National Suisse de la Recherche Scientifique (No 100015-130634). Cet article a été en partie adapté de l'article de Russo (2010).

1. Considérons, de plus, que dans la traduction de deux langues typologiquement différentes, comme c'est le cas pour le français et l'anglais ou aussi pour le français et l'allemand, par exemple, le problème principal est dû au fait que l'anglais et l'allemand standard n'ont pas de pronoms clitiques à proprement parler. Dans ce cas, un système de traduction automatique doit transformer le pronom clitique français (ia) dans un complément du verbe en anglais (ib) et en allemand (ic).
(i) a. Je lui parle. b. I talk to him. c. Ich spreche mit ihm.
2. En français Google traduction (`http://translate.google.com`).
3. Pour plus de détails sur cette évaluation et sur le type de corpus utilisé, nous renvoyons le lecteur à la section 3 de cet article.

Étude inter-langues de la distribution et des ambiguïtés syntaxiques des pronoms. *

Lorenza Russo, Yves Scherrer, Jean-Philippe Goldman,
Sharid Loáiciga, Luka Nerima, Éric Wehrli

Laboratoire d'Analyse et de Technologie du Langage
Département de Linguistique – Université de Genève
2, rue de Candolle – CH-1211 Genève 4

{lorenza.russo, yves.scherrer, jean-philippe.goldman,
sharid.loaiciga, luka.nerima, eric.wehrli}@unige.ch

Résumé. Ce travail décrit la distribution des pronoms selon le style de texte (littéraire ou journalistique) et selon la langue (français, anglais, allemand et italien). Sur la base d'un étiquetage morpho-syntaxique effectué automatiquement puis vérifié manuellement, nous pouvons constater que la proportion des différents types de pronoms varie selon le type de texte et selon la langue. Nous discutons les catégories les plus ambiguës de manière détaillée. Comme nous avons utilisé l'analyseur syntaxique Fips pour l'étiquetage des pronoms, nous l'avons également évalué et obtenu une précision moyenne de plus de 95%.

Abstract. This paper compares the distribution of pronouns according to the text genre (literary or news) and to the language (French, English, German and Italian). On the basis of manually verified part-of-speech tags, we find that the proportion of different pronoun types depends on the text and on the language. We discuss the most ambiguous cases in detail. As we used the Fips parser for the tagging of pronouns, we have evaluated it and obtained an overall precision of over 95%.

Mots-clés : Pronoms, ambiguïté pronominale, étiquetage morpho-syntaxique.

Keywords: Pronouns, pronominal ambiguity, part-of-speech tagging.

1 Introduction

En traitement automatique du langage (TAL), la plupart des recherches sur les pronoms se sont focalisées sur la résolution des anaphores. Dans ce domaine, de très nombreux travaux traitent d'algorithmes capables de détecter des chaînes anaphoriques inter- et intra-phrastiques, de leur implémentation et de leur évaluation (Lappin & Leass, 1994; Mitkov *et al.*, 2002; Trouilleux, 2002). Beaucoup moins nombreux sont les travaux qui ont étudié l'impact de la résolution anaphorique sur des systèmes de TAL (Mitkov *et al.*, 2007; Hardmeier & Federico, 2010). Enfin, des études sur corpus ont été effectuées afin de quantifier la fréquence des pronoms anaphoriques dans différents types de texte (Tutin, 2002; Laurent, 2001).

Le travail que nous présentons dans cet article vise à répondre à trois questions plus générales, indépendantes du caractère anaphorique des pronoms : à quelle fréquence rencontre-t-on les pronoms dans les textes ? Est-ce que la distribution des pronoms change par rapport au type de texte et aussi par rapport à la langue utilisée ? Quelles sont les ambiguïtés pour chaque type de pronom ? Nous avons effectué une étude sur corpus afin d'étudier la distribution des pronoms dans deux textes différents (un texte littéraire et un corpus de communiqués de presse) et dans quatre langues (français, anglais, allemand et italien). Notre but principal est de mieux comprendre la distribution des pronoms en fonction du style du texte [1] et de la langue concernée et d'évaluer l'étiquetage de notre système

∗. Le travail de recherche ici présenté a bénéficié du support du Fonds National Suisse de la Recherche Scientifique (No 100015-130634).

1. Nous préférons parler ici de style de texte plutôt que de genre de texte, car le texte littéraire que nous analysons ne peut pas être considéré comme un échantillon représentatif du genre littéraire.

La traduction automatique des pronoms. Problèmes et perspectives. *

Yves Scherrer, Lorenza Russo, Jean-Philippe Goldman,
Sharid Loáiciga, Luka Nerima, Éric Wehrli

Laboratoire d'Analyse et de Technologie du Langage
Département de Linguistique – Université de Genève
2, rue de Candolle – CH-1211 Genève 4

{yves.scherrer, lorenza.russo, jean-philippe.goldman,
sharid.loaiciga, luka.nerima, eric.wehrli}@unige.ch

Résumé. Dans cette étude, notre système de traduction automatique, Its-2, a fait l'objet d'une évaluation manuelle de la traduction des pronoms pour cinq paires de langues et sur deux corpus : un corpus littéraire et un corpus de communiqués de presse. Les résultats montrent que les pourcentages d'erreurs peuvent atteindre 60% selon la paire de langues et le corpus. Nous discutons ainsi deux pistes de recherche pour l'amélioration des performances de Its-2 : la résolution des ambiguïtés d'analyse et la résolution des anaphores pronominales.

Abstract. In this work, we present the results of a manual evaluation of our machine translation system, Its-2, on the task of pronoun translation for five language pairs and in two corpora : a litterary corpus and a corpus of press releases. The results show that the error rates reach 60% depending on the language pair and the corpus. Then we discuss two proposals for improving the performances of Its-2 : resolution of source language ambiguities and resolution of pronominal anaphora.

Mots-clés : Pronoms, traduction automatique, analyse syntaxique, anaphores pronominales.

Keywords: Pronouns, machine translation, parsing, pronominal anaphora.

1 Introduction

Il y a un quart de siècle, Kay (1986) affirmait : « We know a good deal more about programming techniques and have larger machines to work with ; we have more elegant theories of syntax and what modern linguists are pleased to call semantics ; and there has been some exploratory work on anaphora. But, we still have little idea how to translate into a closely related language like French or German, English sentences containing such words as *he*, *she*, *it*, *not*, *and*, and *of*. » La recherche que nous présentons dans cet article prend comme point de départ l'affirmation de M. Kay et se donne comme objectif l'évaluation de la traduction automatique des pronoms. La plupart des recherches récentes portant sur la résolution des anaphores – pas forcément dans une perspective de traduction (Mitkov *et al.*, 2007) –, nous avons lancé une évaluation plus vaste concernant plusieurs paires de langues. Notre objectif principal est d'améliorer notre système de traduction automatique sur ce phénomène : nous en avons donc repéré les principaux problèmes ainsi que des pistes de recherche pour les résoudre.

Dans la section 2, nous présentons rapidement une pré-étude sur la distribution des pronoms dans la langue source, pour ensuite donner, dans la section 3, les détails de l'étude des pronoms dans la langue cible. La section 4 présente les résultats obtenus ; enfin, dans la section 5, nous discutons les pistes de recherche pour de futures implémentations.

∗. Le travail de recherche ici présenté a bénéficié du support du Fonds National Suisse de la Recherche Scientifique (No 100015-130634).

Ressources lexicales pour une sémantique inférentielle : un exemple, le mot « quitter »

Daniel Kayser

LIPN – UMR 7030 du CNRS
Institut Galilée - Université Paris-Nord
93430 Villetaneuse
Daniel.Kayser@lipn.univ-paris13.fr

Résumé. On étudie environ 500 occurrences du verbe « quitter » en les classant **selon les inférences** qu'elles suggèrent au lecteur. On obtient ainsi 43 « schémas inférentiels ». Ils ne s'excluent pas l'un l'autre : si plusieurs d'entre eux s'appliquent, les inférences produites se cumulent ; cependant, comme l'auteur sait que le lecteur dispose de tels schémas, s'il veut l'orienter vers une seule interprétation, il fournit des indices permettant d'éliminer les autres. On conjecture que ces schémas présentent des régularités observables sur des familles de mots, que ces régularités proviennent du fonctionnement d'opérations génériques, et qu'il est donc sans gravité de ne pas être exhaustif, dans la mesure où ces opérations permettent d'engendrer les schémas manquants en cas de besoin.

Abstract. Around 500 occurrences of the French verb "quitter" are scrutinized and sorted **according to the inferences** they trigger in the reader's mind. This yields 43 so-called inferential schemata. They are not exclusive from one another: when several of them are applicable, their conclusions add together; however, as the author knows that the reader possesses this kind of schema, if s/he wants to direct the reader towards a given interpretation, s/he provides some clues to block the other ones. The schemata reveal regularities across families of similar words, and these regularities are conjectured to be due to the operation of generic procedures: omitting some schemata is thus harmless, insofar as these procedures have the ability to generate the missing ones in case of need.

Mots-clés : Sémantique lexicale. Inférence. Glissements de sens.
Keywords: Lexical Semantics. Inference. Shifts in Meaning.

1 Introduction

Je préconise depuis longtemps (Kayser, 1997) une sémantique qui ne soit pas basée sur une *référence* au monde (selon laquelle les mots ont pour fonction principale de décrire des objets, des événements, etc.), mais qui ne soit référentielle que par effet dérivé de sa fonction première : déclencher des *inférences* chez le lecteur / auditeur.

Dans le cas de la sémantique lexicale, cette thèse devrait signifier que chaque mot possède un pouvoir inférentiel (Small, 1981). Cependant, coupé de tout contexte, un mot engendre assez peu d'inférences et ce pouvoir reste latent (cf. l'idée de « signifié de puissance » de (Guillaume, 1964)) quoique, comme l'ont remarqué Schank et Abelson (1977), certains mots ont la propriété d'installer des "scripts" qui déclenchent par eux-mêmes de nombreuses inférences. Pour les autres, ce n'est qu'une fois connus le co-texte et certains éléments de la situation d'énonciation que la potentialité du mot se concrétise. La façon de représenter cette potentialité latente et, au fur et à mesure que les caractéristiques du contexte se dessinent, l'actualisation de cette potentialité en de véritables inférences restent à élucider. Pour cela, il est nécessaire de répertorier les inférences qui paraissent légitimes dans les différentes circonstances où un même mot est utilisé.

À ma connaissance, il n'existe pas d'étude explicitement orientée vers la collecte des inférences liées à l'usage d'un mot. Cette contribution est une première tentative en ce sens, avec tous les tâtonnements et imperfections que cela comporte.

Repérer les phrases évaluatives dans les articles de presse à partir d'indices et de stéréotypes d'écriture

Mathias Lambert

Université Paris IV-Sorbonne, Laboratoire STIH (LaLIC) - 28 rue Serpente, 75006 Paris
Mathias.Lambert@paris-sorbonne.fr

Résumé. Ce papier présente une méthode de recherche des phrases évaluatives dans les articles de presse économique et financière à partir de marques et d'indices stéréotypés, propres au style journalistique, apparaissant de manière concomitante à l'expression d'évaluation(s) dans les phrases. Ces marques et indices ont été dégagés par le biais d'une annotation manuelle. Ils ont ensuite été implémentés, en vue d'une phase-test d'annotation automatique, sous forme de grammaires DCG/GULP permettant, par filtrage, de matcher les phrases les contenant. Les résultats de notre première tentative d'annotation automatique sont présentés dans cet article. Enfin les perspectives offertes par cette méthode relativement peu coûteuse en ressources (à base d'indices non intrinsèquement évaluatifs) font l'objet d'une discussion.

Abstract. This paper presents a method to locate evaluative sentences in financial and economic newspapers, relying on marks and stereotyped signs. Peculiar to journalese, these are present concomitantly with the expression of evaluation(s) in sentences. These marks or signs have been found by means of a manual annotation. Then, in preparation for an automatic annotation phase, they have been implemented in the form of DCG/GULP grammars which, by filtering, allows to locate the sentences containing them. The results of our first automatic annotation attempt are shown in this article. Furthermore, the prospects offered by this method, which relies on non-intrinsically evaluative marks and therefore does not require long lists of lexical resources, are discussed.

Mots-clés : Opinion, évaluation, repérage de phrases évaluatives, presse économique et financière, style journalistique, indices/marques/stéréotypes d'écriture.
Keywords: Opinion, appraisal, detection of evaluative sentences, financial and economic newspapers, journalese, writing signs/marks/stereotypes.

Corpus-Based methods for Short Text Similarity

Prajol Shrestha
LINA-UFR Sciences, 44322 Nantes Cedex 3
prajol.shrestha@etu.univ-nantes.fr

Résumé. Cet article concerne la détermination de la similarité entre des textes courts (phrases, paragraphes, ...). Ce problème est souvent abordé dans la littérature à l'aide de méthodes supervisées ou de ressources externes comme le thesaurus Wordnet ou le British National Corpus. Les méthodes que nous proposons sont non supervisées et n'utilisent pas de connaissances à priori. La première méthode que nous présentons est basée sur le modèle vectoriel de Salton auquel nous avons apporté des modifications pour prendre en compte le contexte, le sens et la relation entre les mots des textes. Dans un deuxième temps, nous testons les mesures de Dice et de ressemblance pour résoudre ce problème ainsi que l'utilisation de la racinisation. Enfin, ces différentes méthodes sont évaluées et comparées aux résultats obtenus dans la littérature.

Abstract. This paper presents corpus-based methods to find similarity between short text (sentences, paragraphs, ...) which has many applications in the field of NLP. Previous works on this problem have been based on supervised methods or have used external resources such as WordNet, British National Corpus etc. Our methods are focused on unsupervised corpus-based methods. We present a new method, based on Vector Space Model, to capture the contextual behavior, senses and correlation, of terms and show that this method performs better than the baseline method that uses vector based cosine similarity measure. The performance of existing document similarity measures, Dice and Resemblance, are also evaluated which in our knowledge have not been used for short text similarity. We also show that the performance of the vector-based baseline method is improved when using stems instead of words and using the candidate sentences for computing the parameters rather than some external resource.

Mots-clés : Similarité, Modèle Vectoriel, Mesure de Similarité.

Keywords: Similarity, Vector Space Model, Similarity metric.

Développement d'un système de détection des infections associées aux soins à partir de l'analyse de comptes-rendus d'hospitalisation

Caroline Hagège1 Denys Proux1 Quentin Gicquel2 Stefan Darmoni3
Suzanne Pereira4 Frédérique Segond1 Marie-Helène Metzger2

(1) XRCE, 6 Chemin de Maupertuis, 38240 Meylan, France
(2) UCBL-CNRS, UMR 5558 Lyon, France
(3) CISMEF, Rouen, France
(4) VIDAL, Issy les Moulineaux, France
Caroline.Hagege@xrce.xerox.com, Denys.Proux@xrce.xerox.com, Quentin.Gicquel@chu-lyon.fr,
Stefan.Darmoni@cismef.fr, Suzanne.Pereira@vidal.fr, Frederique.Segond@xrce.xerox.com,
Marie-Helene.Metzger@chu-lyon.fr

Résumé

Cet article décrit la première version et les résultats de l'évaluation d'un système de détection des épisodes d'infections associées aux soins. Cette détection est basée sur l'analyse automatique de comptes-rendus d'hospitalisation provenant de différents hôpitaux et différents services. Ces comptes-rendus sont sous forme de texte libre. Le système de détection a été développé à partir d'un analyseur linguistique que nous avons adapté au domaine médical et extrait à partir des documents des indices pouvant conduire à une suspicion d'infection. Un traitement de la négation et un traitement temporel des textes sont effectués permettant de restreindre et de raffiner l'extraction d'indices. Nous décrivons dans cet article le système que nous avons développé et donnons les résultats d'une évaluation préliminaire.

Abstract

This paper describes the first version and the results obtained by a system which detects occurrences of healthcare-associated infections. The system automatically analyzes hospital discharge summaries coming from different hospitals and from different care units. The output of the system consists in stating for each document, if there is a case of healthcare-associated infection. The linguistic processor which analyzes hospital discharge summaries is a general purpose tool which has been adapted for the medical domain. It extracts textual elements that may lead to an infection suspicion. Jointly with the extraction of suspicious terms, the system performs a negation and temporal processing of texts in order to refine the extraction. We first describe the system that has been developed and give then the results of a preliminary evaluation.

Mots-clés : Extraction d'information médicale, compte-rendus d'hospitalisation, infection nosocomiale, analyse syntaxique
Keywords: Information extraction in medical domain, hospital discharge summaries, hospital acquired infections, parsing

Un système de détection d'opinions fondé sur l'analyse syntaxique profonde

Caroline Brun

Xerox Research Centre Europe, 6 chemin de Maupertuis, 38240 Meylan, France
Caroline.Brun@xrce.xerox.com

Résumé : Dans cet article, nous présentons un système de détection d'opinions construit à partir des sorties d'un analyseur syntaxique robuste produisant des analyses profondes. L'objectif de ce système est l'extraction d'opinions associées à des produits (les concepts principaux) ainsi qu'aux concepts qui leurs sont associés (en anglais «features-based opinion extraction»). Suite à une étude d'un corpus cible, notre analyseur syntaxique est enrichi par l'ajout de polarité aux éléments pertinents du lexique et par le développement de règles génériques et spécialisées permettant l'extraction de relations sémantiques d'opinions, qui visent à alimenter un modèle de représentation des opinions. Une première évaluation montre des résultats très encourageants, mais de nombreuses perspectives restent à explorer.

Abstract: In this paper, we present an opinion detection system built on top of a deep robust syntactic parser. The goal of this system is to extract opinions associated to products but also to characteristics of these products, i.e. to perform feature-based opinion extraction. To carry out this task, and following the results of a target corpus study, the robust syntactic analyzer is enriched by the association of polarity to pertinent lexical elements and by the development of generic rules extracting semantic relations of opinions, in order to feed an opinion representation model. A first evaluation gave very encouraging results, but many perspectives remain to be explored.

Mots-clés : détection d'opinions, analyse de sentiments, analyse syntaxique robuste, extraction d'information
Keywords: opinion detection, sentiment analysis, robust parsing, information extraction

1 Introduction

La fouille d'opinions (parfois aussi qualifiée d'analyse de sentiments) fait l'objet d'un engouement tout particulier que ce soit dans les milieux académiques ou dans l'industrie. En effet, avec l'émergence de groupes de discussions, forums, blogs, sites compilant des avis consommateur, on trouve une masse très importante de documents contenant des informations exprimant des opinions, constituant une source énorme de données pour des applications de veille diverses (technologique, marketing, concurrentielle, sociétale). De nombreux travaux de recherche, à la croisée du TALN et de la fouille de données, se penchent sur le problème de la détection d'opinions. Dans cet article, nous présentons le système de détection d'opinions développé pour l'anglais dans le cadre du projet européen Scoop[1], système basé sur l'utilisation d'un analyseur syntaxique robuste adapté à l'analyse des opinions. Après une brève revue des travaux du domaine, nous décrirons l'analyse de corpus que nous avons réalisée, sur un premier corpus cible constitué de revues sur des imprimantes, photocopieurs et scanners extraits du site de revues grand public "Epinion", et qui a conduit à la conception et au développement du système, lui-même décrit en détails dans la section suivante. Nous présenterons ensuite l'évaluation préliminaire de ce système et conclurons sur le bilan et les perspectives envisagées.

[1] http://www.scoopproject.eu/overview.html

PLATON
Plateforme d'apprentissage et d'enseignement de l'orthographe sur le Net

Richard Beaufort Sophie Roekhaut
CENTAL, UCLouvain, Place Blaise Pascal 1, B-1348 Louvain-la-Neuve
{richard.beaufort,sophie.roekhaut}@uclouvain.be

La plateforme PLATON s'inscrit dans le cadre général de l'apprentissage et de l'enseignement des langues assistés par ordinateur (ALAO/ELAO). Dédiée à l'amélioration de la maîtrise de l'orthographe, cette plateforme s'adresse aussi bien à des apprenants natifs qu'à des allophones, pour autant que ceux-ci présentent déjà un niveau de maîtrise avancé de la langue à l'oral et à l'écrit [1]. Sur ce point, PLATON se distingue des autres plateformes d'ALAO/ELAO, classiquement dédiées aux langues secondes.

PLATON est une plateforme en ligne, accessible aux enseignants et à leurs apprenants. Chaque enseignant gère un ou plusieurs cours, divisés en leçons. Un cours étant vu comme un niveau de maîtrise, un apprenant est normalement inscrit à un seul cours à un moment donné.
Actuellement, PLATON gère principalement la partie « exercices et corrigés » : l'enseignant accède à PLATON pour ajouter un exercice à une leçon ou pour visualiser les résultats des apprenants concernés, tandis que l'apprenant y accède pour réaliser un exercice ou visualiser ses résultats. A terme, PLATON proposera également le contenu didactique des leçons, directement produit par l'enseignant. L'adresse de la plateforme sera rendue publique lors de son lancement, aux alentours de septembre 2011.

Dans l'ensemble, le développement de cette plateforme tâche de répondre aux différents besoins relevés par les acteurs de l'ALAO/ELAO (Desmet, 2006). L'un d'eux, un véritable défi, a particulièrement retenu notre attention : dépasser les exercices classiques que sont le texte à trous et le choix multiple, qui limitent considérablement l'éventail des connaissances testées. Pour ce faire, l'idée est de proposer des exercices de type semi-ouvert, qui évitent de signaler trop explicitement le lieu de la difficulté et stimulent la spontanéité des réponses, tout en maintenant l'éventail des variations possibles dans les limites d'un ensemble gérable automatiquement.
La dictée, exercice de type semi-ouvert du fait de la présence d'un original, est l'exercice central de la plateforme, qui en gère automatiquement tous les aspects : sa vocalisation par synthèse de la parole lors de son ajout par l'enseignant, les différentes étapes de sa réalisation par l'apprenant (écoute, copie, relecture) et, bien sûr, sa correction. La phase de correction propose un diagnostic automatique des erreurs, basé sur des méthodes d'alignement et d'analyse linguistique automatique présentées dans un article de la conférence (Beaufort & Roekhaut, 2011). A terme, la plateforme proposera aussi d'autres types d'exercices (textes à trous, jeu des 7 erreurs, etc.), mais également des exercices générés automatiquement, sur la base des lacunes de chaque apprenant.

Dans le cadre de la session de démonstration, nous montrerons les deux pans de la plateforme. Du côté de l'enseignant, nous nous focaliserons sur l'ajout d'une nouvelle dictée. Du côté de l'apprenant, nous nous concentrerons sur la réalisation d'une dictée complète. Accessoirement, nous montrerons également comment visualiser et modifier une dictée existante du côté enseignant, et comment visualiser les copies que ce soit du côté enseignant ou apprenant. Les participants pourront eux-mêmes interagir avec le système.

Références

BEAUFORT R. & ROEKHAUT S. (2011). Le TAL au service de l'ALAO/ELAO. L'exemple des exercices de dictée automatisés. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, 27 juin–1er juillet, Montpellier, France. A paraître.

DESMET P. (2006). L'enseignement/apprentissage des langues à l'ère du numérique : tendances récentes et défis. *Revue française de linguistique appliquée*, **11**(1), 119–138.

1. Les niveaux C1 et C2 du Cadre européen commun de référence pour les langues.
Voir `http://eduscol.education.fr/cid45678/cadre-europeen-commun-de-reference.html`.

SpatiAnn, un outil pour annoter l'utilisation de l'espace dans les corpus vidéo

Annelies Braffort, Laurence Bolot

LIMSI-CNRS, Campus d'Orsay Bt. 508, BP 133, F-91403 Orsay cx, France
annelies.braffort@limsi.fr, laurence.bolot@limsi.fr

SpatiAnn (Spatial Annotator) est un logiciel développé au LIMSI pour l'annotation de l'utilisation de l'espace par les gestes dans les corpus vidéo de langue des signes ou multimodaux. Les gestes s'expriment dans le temps, mais aussi dans l'espace, nommé « espace de signation » pour la langue des signes et « espace de gestualisation » pour le multimodal. Des études de plus en plus nombreuses portent sur l'utilisation linguistique de cet espace et nécessitent des annotations. Les logiciels d'annotation actuels (Elan, Anvil, iLex…) ne permettent pas d'annoter de manière directe des informations de nature tridimensionnelles. Actuellement, les annotations sont basées sur une segmentation arbitraire de l'espace, par exemple dans un plan vertical tel que le « gesture space » de McNeill (1992), ou sous forme de cubes (Lenseigne, Dalle 2005), ce qui peut limiter les analyses qui s'appuient sur ces annotations. C'est pourquoi nous développons actuellement un logiciel qui permet d'annoter directement en 3d. Il se présente sous la forme d'un cube, où la vidéo est projetée sur l'une des faces. On peut aussi projeter plusieurs vidéos, en fonction des différentes vues dont on dispose, ce qui permet d'annoter dans un contexte d'interaction. La figure 1 montre un exemple avec le corpus DEGELS1 (Boutora, Braffort 2011) pour lequel on dispose de trois vues. L'utilisateur peut manipuler le cube afin d'annoter « devant » la vidéo de son choix. Cette annotation peut prendre n'importe quelle forme, le vocabulaire contrôlé et sa forme graphique sont libres. Pour nos études, nous utilisons un vocabulaire contrôlé constitué d'un ensemble fini de formes géométriques simples (point, segment, plan, tore, volume…), qui catégorise la trace du geste dans l'espace. Par exemple la trace d'un pointage associé à un mouvement circulaire (description d'un rond-point) est catégorisée par un tore placé à l'endroit pointé (figure 1). Ces traces sont, comme toutes les autres annotations, synchronisées avec la vidéo. Elles peuvent être plus ou moins actives selon leur utilisation dans le discours. Une trace réalisée précédemment dans la vidéo s'atténue au cours du temps mais peut être réactivée par un pointage anaphorique. Cela se visualise en faisant varier sa transparence.

Figure 1 : Utilisation de SpatiAnn avec un corpus constitué de trois vues

Le logiciel est développé actuellement sous la forme d'un prototype autonome en vue d'évaluer et d'améliorer son ergonomie. Dans un deuxième temps, ce prototype sera intégré dans un système distribué permettant son emploi avec le logiciel d'annotation AnCoLin développé dans le cadre du projet européen Dicta-Sign (Collet, Gonzalez, Milachon 2010). Dans un troisième temps, on envisage d'en développer une version sous forme d'un plugin utilisable avec certains logiciels courants, tels qu'Anvil et Elan, afin de le rendre accessible à une plus grande partie de la communauté scientifique concernée.

MCNEILL, D. (1992). *Hand and Mind: What gestures reveal about thought*. Chicago: Univ. of Chicago Press.

LENSEIGNE, B., DALLE, P. (2005). A Signing space model for the interpretation of sign language interactions. Actes de *Sign Language Linguistics and the Application of Information Technology to Sign Languages*.

BOUTORA L., BRAFFORT A. (2011). *DEGELS1*. oai:crdo.fr:crdo000767.

COLLET, C., Gonzalez M., Milachon F. (2010). Distributed system architecture for assisted annotation of video corpora. Actes de *International workshop on the Representation and Processing of Sign Languages : Corpora and Sign Language Technologies (LREC 2010)*.

Libellex : une plateforme multiservices pour la gestion des contenus multilingues

François Brown de Colstoun[1], Estelle Delpech[1,2], Étienne Monneret[1]

(1) LINGUA ET MACHINA, Laval Technopole, 6 rue Léonard de Vinci, 53001 Laval Cedex
et c/o Inria, Rocquencourt BP 105, 78 153 Le Chesnay Cedex
(2) LINA FRE CNRS 2729, 2 rue de la Houssinière BP 92208, 44322 Nantes Cedex 3
fbc(a)lingua-et-machina.com, ed(a)lingua-et-machina.com, em(a)lingua-et-machina.com

Cette démonstration industrielle présente *Libellex*, un prototype de plateforme multiservices pour la gestion des contenus multilingues. Cette plateforme vise à la fois des professionnels de la langue (traducteurs, terminologues) et un public plus général amené à communiquer en langue étrangère dans un contexte professionnel. *Libellex* assiste ces utilisateurs dans la production, la compréhension et la traduction de documents dans des langues ou des domaines qu'ils maîtrisent bien, imparfaitement ou pas du tout. Pour cela, *Libellex* propose une palette de services exploitant une base de connaissances linguistiques constituée automatiquement à partir de textes fournis par l'utilisateur. Ces services vont de la simple recherche d'expression à la pré-traduction et la constitution automatique de terminologies bilingues. Le processus d'extraction de connaissances linguistiques s'appuie sur des algorithmes issus des recherches en traduction automatique (Gale et Church 1993 ; Lardilleux, 2010), traduction assistée par ordinateur (Planas, 200), terminologie computationnelle (Bourigault, 1994) et exploitation des corpus comparables (Fung, 1997 ; Morin et Daille, 2009).

Remerciements

Libellex a été en partie financé par l'ANR (subvention no. ANR-08-CORD-009) et par Oséo (subvention no. A1010034Z). Nous remercions également Guillaume Pelluau et Mikaël Morardo pour leur participation au développement de *Libellex*.

Références

Bourigault, D. (1994). LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes. *Thèse en Mathématiques, Informatique Appliquée aux Sciences de l'Homme*. École des Hautes Études en Sciences Sociales.

Fung, P. (1997). Finding Terminology Translations from Non-parallel Corpora. Dans *5th Workshop on Very Large Corpora*, Hong Kong, p. 192-202.

Gale, W. A., et K. W. Church. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19(1): 85-102.

Lardilleux, A. (2010). Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle. *Thèse en Informatique*. Université de Caen Basse-normandie.

Morin, E., et B. Daille. (2009). Compositionality and lexical alignment of multi-word terms. Dans *Language Resources and Evaluation (LRE), Multiword expression: hard going or plain sailing*, P. Rayson, S. Piao, S. Sharoff, S. Evert, B. Villada Moirón, p. 79–95.

Planas, E. (2000). Multi-level Similar Segment Matching Algorithm for Translation Memories and Example-Based Machine Translation. Dans *Proceedings of the 18th Conference on Computational Linguistics*, Saarbrücken, Germany, p. 622-627.

Une application de la grammaire structurelle: L'analyseur syntaxique du français SYGFRAN

Jacques Chauché[1]
(1) LIRMM, 161, rue ADA 34392 Montpellier Cedex 5
jacques.chauche@lirmm.fr

Résumé. La démonstration présentée produit une analyse syntaxique du français. Elle est écrite en SYGMART, fournie avec les actes, exécutable à l'adresse : http ://www.lirmm.fr/ chauche/ExempleAnl.html et téléchargeable à l'adresse : http ://www.sygtext.fr.

Abstract. The software produces a syntactic analysis of french. It is written in SYGMART, including acts, runable at http ://www.lirmm.fr/ chauche/ExempleAnl.html and downlodable at : http ://www.sygtext.fr.

Mots-clés : Analyse syntaxique.

Keywords: syntactic analysis.

1 Modèle de traitement

Les algorithmes de Markov ont la puissance d'une machine de Turing. Ils peuvent donc représenter n'importe quel algorithme. Ces traitements se font à partir d'un traitement de chaîne par la substitution d'infixes. La grammaire structurelle transforme des structures arborescentes. Elle correspond donc à une extension des algorithmes de Markov appliquée aux structures arborescentes : au lieu de remplacer un infixe d'une chaîne on remplace une sous-structure d'une arborescence. La grammaire structurelle permet de définir des récurrences ou des compositions d'applications. Le système SYGMART est un outil qui permet l'écriture et l'exécution de grammaires structurelles. Ce système a comme entrée soit une structure déjà produite par ailleurs, soit un texte quelconque. Il produit soit un texte soit une structure. Par exemple une application récente concernait la manipulation de fichiers xml afin de modifier et/ou de composer leurs structures. D'autres applications textuelles permettent d'écrire un traducteur, un taggeur, un compresseur de texte, un classifieur, ...

2 Caractéristiques de SYGFRAN

L'analyse s'effectue en deux temps :
– une analyse morphologique produisant l'ensemble des solutions possibles pour chaque mot dans une abrorescence simple. Cette arborescence comporte deux niveaux. Dans le premier niveau chaque point correspond à un mot du texte. Chaque point de ce niveau est la racine des différentes solutions pour le mot associé. Cette étape est basée sur un automate d'états finis.
– Une analyse syntaxique basée sur la grammaire structurelle. La grammaire comprend environ 250 parties et 20000 règles.

Sur un PC ou un Mac comprenant au moins 2 GO de mémoire l'analyse supporte le traitement d'un texte représentant une arborescence syntaxique de 200000 points. L'analyse du livre "Le petit prince" de Saint Exupery traite et produit une structure d'environ 35000 points et mets 3mn sur un ordinateur MacBookPro Intel core 2 Duo, 4GO de mémoire.

Proxem Ubiq : une solution d'e-réputation par analyse de feedbacks clients

François-Régis Chaumartin
Proxem, 19 bd de Magenta, 75010 Paris
frc@proxem.com

Mots-clés : e-réputation, reconnaissance d'entités nommées, classification, clustering, analyse syntaxique, apprentissage

Être à l'écoute de ses clients est un enjeu majeur pour toute grande marque. Les verbatims d'expression spontanée des consommateurs se trouvent le plus souvent sur des sources externes (blogs, forums, news, RSS, tweets…) et internes (mails envoyés spontanément, réponses aux questions ouvertes de sondages). Ubiq permet aux entreprises de calculer leur e-réputation en analysant ces différents feedbacks. Ubiq identifie les attentes des consommateurs, détecte les tendances, analyse les opinions et permet d'anticiper des problèmes. En un coup d'œil, on visualise les « sujets chauds » du moment.

La plateforme de TAL Antelope est au cœur d'Ubiq. L'analyse sémantique effectuée enchaîne plusieurs opérations.

(1) La qualité des documents traités étant très variable, une correction orthographique est souvent nécessaire ; néanmoins, cette opération doit être effectuée avec une connaissance du contexte métier ; par exemple, les noms de marques qui viennent d'apparaître (et ne figurent pas encore dans un lexique) ne doivent pas être « corrigés » vers un mot proche.

(2) La reconnaissance d'entités nommées vise classiquement à identifier des personnes, lieux et organisation. Dans un contexte d'enseigne de grande distribution, les entités intéressantes à détecter sont plutôt les produits, marques et concurrents cités, ainsi que des concepts liés au métier (le risque sanitaire ou le risque juridique, par exemple). Nous avons développé une nouvelle approche d'acquisition à large échelle d'entités nommées. (2a) Une première phase d'extraction terminologique permet d'amorcer la liste des concepts du domaine. (2b) Une seconde phase utilise deux ressources de large couverture (la Wikipédia et un WordNet pour le français) pour créer des gazettes ; en cas d'ambiguïté possible (*orange* fruit ou *Orange* marque), les termes des gazettes sont automatiquement associés à des mots clés activateurs ou inhibiteurs (pour les deux sens d'orange : jus, fruit, pulpe… ou internet, contrat, carte sim, opérateur…). (2c) L'application de ces gazettes permet de constituer un premier corpus annoté selon les entités nommées du domaine. Un apprentissage (par CRF) est alors effectué sur le corpus, pour identifier de nouvelles instances d'entités. (2d) Chaque document fait aussi l'objet d'une classification multi-motifs (dont une analyse d'opinion pour en déterminer la valence).

(3) L'ensemble des documents est partitionné en sous-ensemble homogènes, pour déterminer les tendances du moment ; l'utilisation de techniques de clustering spectral permet de traiter en quelques minutes plusieurs milliers de documents.

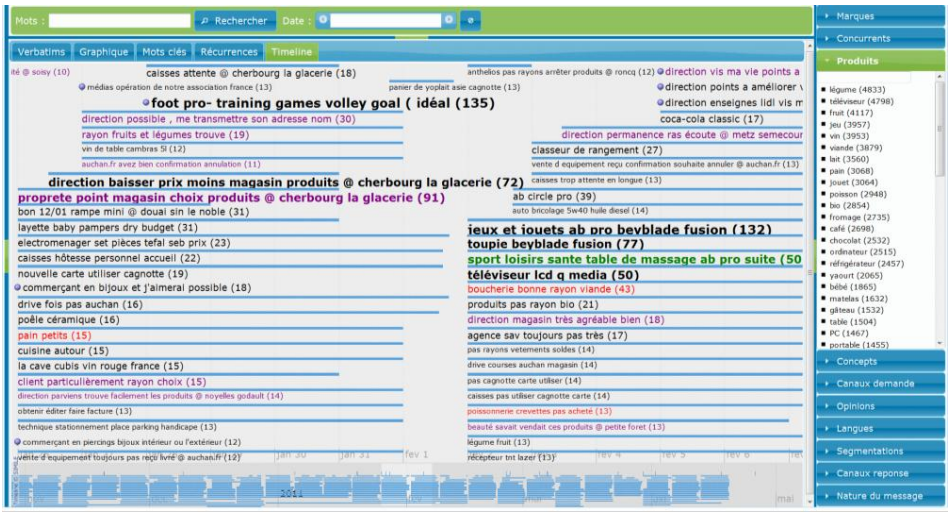


Figure 1 : Une capture d'écran d'Ubiq, montrant d'une façon synthétique ce qui s'est passé pendant deux semaines dans une enseigne de la grande distribution. La partie centrale est le « résumé sémantique » de plus de 10 000 feedbacks.

TTC TermSuite : une chaîne de traitement pour la fouille terminologique multilingue

Béatrice Daille Christine Jacquin Laura Monceaux Emmanuel Morin Jérome Rocheteau
Université de Nantes - LINA – 2 rue de la Houssinière – BP 92208 – 44322 Nantes cedex 3, France
{prenom.nom}@univ-nantes.fr

Le projet européen TTC [1] vise à exploiter les possibilités offertes par les corpus comparables pour améliorer les performances des outils informatiques de traduction. Il s'agit de traiter des domaines techniques dans un contexte massivement multilingue où il est nécessaire de traduire un même document dans plusieurs langues. TTC TermSuite est un ensemble de composants logiciels pour l'extraction et l'alignement terminologique multilingue à partir de corpus comparables dans 5 langues européennes - Anglais, Français, Allemand, Espagnol et une langue peu dotée, le Letton, ainsi qu'en Chinois et en Russe.
TTC TermSuite adopte la plate-forme Apache UIMA [2] conçue pour faciliter l'assemblage de composants, leur intégration au sein d'une chaîne de traitement ainsi que le passage à l'échelle dans un contexte industriel.
TTC TermSuite procède à une extraction terminologique monolingue pour les 7 langues, puis à son alignement par paire de langues. En entrée, sont fournis plusieurs corpus comparables dont les documents sont composés de deux types de fichiers : le texte du document et les métadonnées associées au format Dublin Core [3]. Ces métadonnées recensent la langue, la source du document, la date d'extraction s'il s'agit d'un fichier extrait du web, le format (.txt, .html, .pdf, etc.), le sujet. Seule la langue est une métadonnée obligatoire. En sortie, sont produites des listes terminologiques monolingues et bilingues sous la forme d'une fichier XML au format TermBase eXchange [4].
TTC TermSuite effectue les traitements informatiques dédiés à l'acquisition terminologique en 4 phases :

Traitements préliminaires : identification et conversion des encodages de caractères, détection de la langue ;

Analyses linguistiques découpage du texte en mots, analyse morphosyntaxique et lemmatisation et conversion au format Multext ;

Extraction terminologique monolingue détection d'occurrences de termes simples et complexes, normalisation et regroupement des termes en fonction de leurs variations, filtrage statistique ;

Alignement terminologique bilingue alignement contextuel par paires de langues.

Chacune des unités fonctionnelles qui composent les 4 phases de cette architecture logicielle est réalisée par un composant UIMA dédié. Chacun de ces composants gère le multilinguisme et, au besoin, répartit le document en cours de traitement à un sous-composant dédié au traitement de la langue de ce document.

TTC TermSuite est librement distribué [5]accompagné d'une vidéo sur Youtube [6] expliquant comment l'utiliser.

La démonstration présentera l'extraction et l'alignement des termes simples sur l'Anglais, Français, Allemand, Espagnol, Chinois et Russe, ainsi que l'extraction et alignement de termes complexes sur le Français et l'Anglais.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 248005.

1. http://www.ttc-project.eu
2. http://uima.apache.org
3. http://dublincore.org
4. http://www.lisa.org/Term-Base-eXchange.32.0.html
5. http://code.google.com/p/ttc-project
6. http://www.youtube.com/watch?v=Vi6yoXaFZ44

An Interaction Mining Suite Based On Natural Language Understanding

Rodolfo Delmonte[1,2], Vincenzo Pallotta[1], Violeta Seretan[3], Lammert Vrieling[1], David Walker[1]

(1) Interanalytics, Geneva, Switzerland
(2) Department of Language Science, University of Venice, Italy
(3) School of Informatics, University of Edinburgh, United Kingdom
delmonte@unive.it, <firstname.surname>@internalytics.ch, violeta.seretan@gmail.com

We introduce Interanalytics™ *Interaction Mining suite*, a collection of tools that performs the analysis, summarization and visualization of conversational content. Interaction Mining is an emerging Business Intelligence (BI) application whose main goal is the discovery and automatic extraction of useful information from human conversational interactions for analytical purposes. Conversations – interactions that serve specific purposes and whose participants contribute to the achievement of a shared goal – are ubiquitous in our real and digital life. Especially on the Internet, people interact in natural language using several technologies such as social networks, instant messaging, VoIP, discussion forum, or (micro)blogs.

Turning conversational data into meaningful information leads to better business decisions through appropriate visualization and navigation techniques. Interanalytics™ leveraged an advanced Natural Language Understanding (NLU) technology to a BI tool enabling analysts to understand and generate insights from conversational content in selected business applications such as Speech Analytics, Social Media monitoring, and Market Research (Pallotta 2010). In order to achieve this, Interanalytics™ has tailored a sophisticated (NLU) technology for mining universal facets of digital conversations (Delmonte et al. 2009). We will demonstrate the main features of the Interaction Mining suite by showcasing two business applications:

1. Advanced abstractive summarization for multi-party discussions (e.g. meetings, focus groups, political debates) that highlights the processes of opinion negotiation and decision-making. The tool produces high-quality memos and insightful visualization of the participants' behavior in terms of their collaborative participation to the discussion (Pallotta et al. 2011).

2. Contact Centers conversations analysis that enables the implementation of novel practical metrics for contact center quality management. The tool allows quality managers to assess agents' performance and predict customer-rating outcomes.

More generally, we will discuss how Interanalytics™ technology enables a wider spectrum of Business Intelligence for unstructured and conversational data. More information about Interanalytics™ can be found at www.interanalytics.ch.

References

PALLOTTA V., DELMONTE R., BISTROT A. Abstractive Summarization of Voice Communications. In Vetulani Z. (Ed.) *Human Language Technology: challenges for the information society*. LNCS n. 6562, Springer Verlag. April, 2011.

PALLOTTA V. Content-based retrieval of distributed multimedia conversational data. In A. Soro, E. Vargiu, G. Armano, G. Paddeu (eds.) Information Retrieval and Mining in Distributed Environments, Springer Verlag series: Studies in Computational Intelligence, 2011, Volume 324/2011, pp. 183-212, 1st Edition, ISBN: 978-3-642-16088-2.

DELMONTE R., BRISTOT A., VOLTOLINA G., PALLOTTA V. Scaling up a NLU system from text to dialogue understanding. Proceedings of the *NAACL HLT Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 40–41, Boulder, Colorado, June 2009. Association for Computational Linguistics.

Démonstration de l'API de NLGbAse

François-Xavier Desmarais, Éric Charton

École Polytechnique de Montréal, 2900 boul. Edouard Montpetit, Montréal, Canada H3T 1J4
{francois-xavier.desmarais, eric.charton}@polymtl.ca

Description

Le système NLGbAse transforme des contenus encyclopédiques en métadonnées. Il utilise ensuite ces métadonnées pour entrainer et faire fonctionner des systèmes d'étiquetage de textes [1]. NLGbAse permet d'étiqueter les entités nommées (EN) d'un texte en reprenant la taxonomie ESTER [2]. Il peut ensuite établir un lien sémantique entre l'EN identifiée et sa représentation sur le web sémantique, notamment son point d'entrée dans le réseau LinkedData [3]. Le système NLGbAse est multilingue et peut étiqueter un texte en français, anglais ou espagnol. Il peut être utilisé en utilisant soit une interface en ligne, soit un API. Notre démonstration consiste à présenter ses fonctionnalités.

Dans un texte, chaque entité nommée est associée à son étiquette de classe et à des liens. Les liens relient l'EN avec sa page descriptive de Wikipedia *(*libellé *wp)* et à un point d'entrée sur le réseau LinkedData, au format « Resource Description Framework » (libellé *rdf*).

L'utilisation de l'interface en ligne n'est appropriée que pour l'étiquetage de courts textes et n'autorise qu'une étude rudimentaire. Pour l'obtention d'un corpus plus large et étiqueté, une API est disponible. Cette API permet d'envoyer une séquence de texte depuis un programme PERL et de recevoir en retour une sortie étiquetée par NLGbAse, fournie sous la forme de lignes composées de cinq colonnes, séparées par des tabulations.

Le mot (ou ponctuation), la nature du mot (POS), son étiquette taxonomique (EN), son lien avec sa métadonnées représentative de NLGbAse et un lien vers Dbpedia (qui permet de collecter tous les points d'entrées disponibles pour un terme sur le réseau LinkedData).

L'API de NLGbAse est disponible gratuitement pour étiqueter, par période de 24 heures, un maximum de 100 documents d'au plus 10000 caractères chacun. Des volumes plus importants peuvent être alloués sur demande pour des recherches académiques.

Bibliographie :

[1] Charton, E. & Torres-Moreno, J. (2010). NLGbAse: a free linguistic resource for Natural Language Processing systems. (Eds.)*English*, (1), 2621-2625. Proceedings of LREC 2010.

[2] Charton, E. & Torres-Moreno, J. (2009). Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées. Dans Taln 2009 , volume 1, pages 24–26. TALN.

[3] Charton, E., Gagnon, M., & Ozell, B. (2010). Extension d'un système d'étiquetage d'entités nommées en étiqueteur sémantique. *Actes de TALN 2010, 1*(1), 19-23.

Système d'analyse de la polarité de dépêches financières

Michel Généreux
Centro de Linguística da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal
genereux@clul.ul.pt

Résumé. Nous présentons un système pour la classification *en continu* de dépêches financières selon une polarité positive ou négative. La démonstration permettra ainsi d'observer quelles sont les dépêches les plus à même de faire varier la valeur d'actions cotées en bourse, au moment même de la démonstration. Le système traitera de dépêches écrites en anglais et en français.

Abstract. We present a system for classifying *on-line* financial news items into a positive or negative polarity. The demonstration will therefore allow us to observe which news are most likely to influence the price of shares traded at the time of the demonstration. The system will cover news items written in English and French.

Mots-clés : Analyse de Sentiments, Linguistique de Corpus, Dépêches Financières.

Keywords: Sentiment Analysis, Corpus Linguistics, Financial News Items.

Description

Nous avons mené des travaux sur l'analyse de sentiments en utilisant comme domaine d'investigation celui de la finance (Généreux *et al.*, 2011). Nous faisions l'hypothèse que la réaction du marché suite à la publication d'une dépêche reliée à une action particulière est un bon indicateur de la polarité de la nouvelle, et qu'un algorithme d'apprentissage à partir de ces dépêches permet de construire un système qui donne à l'investisseur une source d'information supplémentaire qui peut être exploitée de façon avantageuse dans une stratégie d'investissement. Dans cette démonstration, nous voulons donner un aperçu des possibilités offertes par ce genre d'analyse dans une situation concrète où des dépêches sont examinées et classées dès leur publication. Notre système effectue l'apprentissage d'un modèle pour la classification polaire (positif ou négatif) à partir de dépêches déjà annotées selon qu'elles précèdent une hausse ou une baisse de la valeur d'une action particulière. Selon cette perspective, toute dépêche précédent une hausse significative de la valeur d'une action sera annotée comme positive, négative dans le cas contraire. Ces dépêches annotées nous permettent d'établir un lexique de termes financiers discriminants servant à construire un classifieur SVM. Notre démonstrateur classifie les dépêches *en continu*, i.e. qu'il analyse les flots de dépêches en format *RSS* telles que fournies par des sites spécialisés dans la finance (e.g. *Financial Times*, *Les Échos*). Lorsqu'une dépêche présente une orientation non-neutre, elle est mise en évidence par le système, permettant à un investisseur potentiel de se concentrer sur les dépêches ayant, en principe, une plus grande probabilité de faire varier la valeur de l'action visée par la dépêche. Certes, la taille restreinte de certaines dépêches que l'on retrouve sur les flux RSS limite la portée d'une approche comme la nôtre basée sur la sélection de traits lexicaux, et les coûts de transaction requièrent que les gains potentiels soient significativement *au-delà* de la mise de départ. Néanmoins, la démonstration permettra à l'utilisateur de se faire une idée du potentiel d'exploitation de l'analyse de sentiments comme stratégie (partielle) d'investissement financier.

Références

GÉNÉREUX M., POIBEAU T. & KOPPEL M. (2011). *Sentiment analysis using automatically labelled financial news items*, In *Affective Computing and Sentiment Analysis : Metaphor, Ontology, Affect and Terminology*, chapter 9, p. 111–126. Springer Verlag.

Babouk – exploration orientée du web pour la constitution de corpus et de terminologies

Clément de Groc[1,2] Javier Couto[1,3] Helena Blancafort[1,4] Claude de Loupy[1]

(1) Syllabs, 15 rue Jean-Baptiste Berlier, 75013 Paris
(2) Univ. Paris Sud et LIMSI-CNRS, F-91405 Orsay
(3) MoDyCo, UMR 7114, CNRS-Université Paris Ouest Nanterre, La Défense
(4) Universitat Pompeu Fabra Roc Boronat,138, 08018 Barcelona, Spain
{cdegroc, jcouto, blancafort, loupy}@syllabs.com

Babouk est un crawler orienté (Chakrabarti et al., 1999) : son objectif est le rapatriement efficace de documents pertinents pour un domaine défini. Comparativement au crawling traditionnel, le crawling orienté permet un accès rapide à des données spécialisées tout en évitant le coût prohibitif d'un parcours en largeur du web. L'exploitation du web comme source de données linguistiques a permis de créer de nombreux corpus généralistes et spécialisés par le biais de requêtes à un moteur de recherche (Baroni, Bernardini, 2004) ou d'un crawl du web (Baroni & Ueyama, 2006). Babouk ne requiert qu'un petit ensemble de termes ou URLs amorces en entrée. Le reste de la procédure est automatique. L'utilisateur peut régler le crawler par un ensemble de paramètres et reprendre la main sur la procédure à tout moment.

Babouk doit trouver un maximum de documents pertinents en téléchargeant le minimum de pages. Le crawler s'appuie sur un catégoriseur qui filtre les documents non pertinents et ordonne par pertinence les pages à télécharger. Le catégoriseur est basé sur un lexique pondéré construit durant la première itération du crawling : une extension de l'entrée utilisateur est effectuée en utilisant la procédure BootCaT (Baroni, Bernardini, 2004). Le lexique est ensuite pondéré à l'aide d'une mesure de « représentativité » s'appuyant sur le web. Une phase de calibration automatique permet de déterminer un seuil pour la catégorisation. Pour guider le crawler en priorité vers les pages les plus pertinentes, le score fourni par le catégoriseur est utilisé de manière analogue au critère OPIC (Abiteboul et al., 2003).

Plusieurs critères d'arrêt on été implémentés tels qu'un nombre maximal de tokens ou de documents à télécharger, une profondeur ou une durée de crawl maximale. Plusieurs filtres sont appliqués dans le but d'améliorer la qualité des corpus constitués. L'utilisateur peut ainsi choisir de ne conserver que des pages d'une certaine taille ou appartenant à un certain format de fichier (parmi Microsoft Office, Adobe PDF, ou HTML). Il peut également limiter le crawl à certains domaines/sites ou, au contraire, les filtrer.

Babouk est basé sur Nutch et distribué sur une grappe de machines (optionnellement sur le « cloud »), ce qui assure un passage à l'échelle en termes de puissance de calcul nécessaire pour la réalisation de nombreux crawls simultanément. Enfin, les documents et méta-informations résultants du crawling peuvent être stockés dans une base de données distribuée assurant, encore une fois, la *scalabilité* du système. Les utilisateurs peuvent configurer et lancer leurs crawls à partir d'une interface web dynamique. Cette dernière offre également un suivi (logs) du crawl en temps réel.

ABITEBOUL M., PREDA M., COBENA G. (2003). Adaptive on-line page importance computation. Actes de *12th international conference on the World Wide Web – WWW*. 280-290.

BARONI M., BERNARDINI S. (2004). BootCaT : Bootstrapping Corpora and Terms from the Web. Actes de *4th international conference on language resources and evaluation – LREC*. 1313-1316.

BARONI M., UEYAMA M. (2006). Building general- and special-purpose corpora by Web crawling. Actes de *13th NIJL International Symposium, Language Corpora: Their Compilation and Application.* 31-40.

CHAKRABARTI S., DEN BERG M.V., DOM B. (1999). Focused crawling : a new approach to topic-specific Web resource discovery. Actes de *Computer Networks, vol. 31*. 1623-1640.

Extraction d'informations médicales au LIMSI

Cyril Grouin[1], Louise Deléger[1], Anne-Lyse Minard[1],
Anne-Laure Ligozat[1], Asma Ben Abacha[1], Delphine Bernhard[1],
Bruno Cartoni[2], Brigitte Grau[1], Sophie Rosset[1], Pierre Zweigenbaum[1]
(1) LIMSI-CNRS, BP133, 91403 Orsay Cedex, France
(2) Département de Linguistique, Université de Genève, Suisse
{prenom.nom}@limsi.fr, bruno.cartoni@unige.ch, louise.deleger@cchmc.org

Les textes présents dans les dossiers de patients dans les hôpitaux contiennent des informations précieuses pour diverses tâches médicales comme par exemple la prise de décision diagnostique et thérapeutique. Il est de ce fait utile de chercher à extraire automatiquement ces informations. Récemment, des évaluations internationales annuelles de systèmes d'extraction d'informations à partir de documents cliniques en anglais ont été organisées par l'institut i2b2 aux États-Unis (Uzuner *et al.*, 2010). Cette démonstration présente de façon concrète les outils mis en place au LIMSI pour participer aux évaluations i2b2 2009 et i2b2/VA 2010. Ces outils prennent en entrée les phrases d'un texte en anglais, et utilisent au besoin les annotations réalisées par des outils précédemment appliqués, pour détecter les informations requises, fournies sous la forme d'annotations déportées. Résultats en 2009 : F=0.773, 8^e/20 ; en 2010 : F=0.773, 12^e/22 (piste 1), F=0.931, 5^e/21 (piste 2), F=0.709, 3^e/16 (piste 3).

La tâche 2009 consistait à extraire les prescriptions médicamenteuses (médicament, dosage, fréquence, quantité, mode d'administration, durée, et raison de la prescription). Notre méthode repose sur des lexiques et l'observation des régularités d'expression de ces prescriptions, implémentées sous la forme de patrons fondés sur les classes sémantiques correspondant aux types d'informations recherchées (Deléger *et al.*, 2010; Grouin *et al.*, 2011).

Trois tâches étaient proposées en 2010 : (i) détection de trois types de concepts médicaux : problèmes (maladies, symptômes), traitements (y compris médicamenteux) et examens ; (ii) détermination du statut d'un problème médical (présent, absent, hypothétique, etc.) ; et (iii) détection de sept types de relations entre problèmes, traitements et examens. Nous avons mis en place des méthodes à base de connaissances expertes et de TAL (lexiques, patrons, étiquetage morphosyntaxique, analyse syntaxique) ainsi que des méthodes à base d'apprentissage supervisé (SVM, CRF) entraînées sur les corpus annotés fournis par les organisateurs de la campagne (Bernhard & Ligozat, 2011; Grouin *et al.*, 2011; Minard *et al.*, 2011a,b), et cherché à combiner ces deux types de méthodes.

Ce travail a été partiellement financé par les projets Akenaton (ANR-07-TecSan-001), InterSTIS (ANR-07-TecSan-010) et Quæro (financement Oseo, agence française pour l'innovation et la recherche).

Références

BERNHARD D. & LIGOZAT A.-L. (2011). Analyse automatique de la modalité et du niveau de certitude : application au domaine médical. In *TALN 2011*, Montpellier. *À paraître*.

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**, 555–558.

GROUIN C., DELÉGER L., CARTONI B., ROSSET S. & ZWEIGENBAUM P. (2011). Accès au contenu sémantique en langue de spécialité : extraction des prescriptions et concepts médicaux. In *TALN 2011*. *À paraître*.

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. *J Am Med Inform Assoc*. *À paraître*.

MINARD A.-L., LIGOZAT A.-L. & GRAU B. (2011b). Extraction de relations dans des comptes rendus hospitaliers. In *Actes des 22emes Journées francophones d'Ingénierie des Connaissances*.

UZUNER O., SOLTI I. & CADAG E. (2010). Extracting medication information from clinical text. *J Am Med Inform Assoc*, **17**(5), 514–518.

Système d'analyse catégorielle *ACCG* : adéquation au traitement de problèmes syntaxiques complexes

Juyeon Kang, Jean-Pierre Desclés

LaLIC, 28, Rue Serpente, 75006 Paris, France

kjuyeon79@yahoo.fr, jean-pierre.desclés@paris-sorbonne.fr

Le système catégoriel, *ACCG (Applicative Combinatory Cateogrial Grammar)*, est conçu pour objectif de produire des calculs catégoriels et d'engendrer des structures opérateur/opérande des phrases du français et du coréen. Il s'agit d'un système motivé pour l'analyse de bonne qualité linguistique pouvant traiter de phénomènes langagiers précis, particulièrement, le système casuel, le double cas, la flexibilité de l'ordre des mots dans la langue coréenne, et la coordination, la subordination ainsi que la thématisation dans les langues française et coréenne. Pour cela, nous nous sommes basés sur la Grammaire Catégorielle Combinatoire Applicative (GCCA) (Desclés et Biskri, 2005 ; Kang et Desclés , 2008) qui est sous-jacente à un modèle mettant en œuvre le calcul fonctionnel des types de Church, les combinateurs de la Logique Combinatoire (LC) (Curry et Feys, 1958) et des méta-règles (deux types de méta-règles : a/ méta-règles pour contrôler l'intervention des combinateurs ; b/ méta-règles « contextuelles » pour lever l'ambiguïté dans l'assignation des types à certaines unités polysémiques). Ces méta-règles contextuelles sont des sortes de règles contextuelles au sens de J.-P Desclés (Desclés, 2006).
Ce système *ACCG* (cf. Figure 1) est composé d'une interface, d'un lemmatiseur, de deux dictionnaires des mots typés et d'un analyseur catégoriel : 1) le lemmatiseur que nous avons développé permet essentiellement de segmenter des phrases en entrée en morphèmes ou en mots, ce qui est une étape pour une analyse morpho-syntaxique des langues agglutinantes comme le coréen ; 2) les deux dictionnaires préalablement définis contiennent environ milles unités linguistiques typées pour le coréen et le français[1] ; 3) l'analyseur catégoriel implémenté en OCaml.

Figure 1. Interface de l'analyseur *ACCG*

Le module « Demo » de l'*ACCG* propose des analyses syntaxiques de phénomènes précis tels que le double cas, l'ordre des mots, la coordination, la subordination et la topicalisation.
Le module « Parser » est une composante essentielle de l'*ACCG* en effectuant des analyses syntaxiques. Le lemmatiseur et les dictionnaires prédéfinis interviennent dans ce module. Le processus d'analyse est le suivant : 1) choisir une langue spécifique ; 2) entrer une phrase à analyser ; 3) cliquer sur le bouton '1. Lemmatiser' si on veut vérifier le résultat de lemmatisation ; 4) cliquer sur le bouton '2. French Parser' ou '3. Korean Parser' pour l'analyse catégorielle de la phrase segmentée.

BISKRI I., DESCLÉS J.-P. (2005). Applicative and Combinatory Categorial Grammar and Subordinate Constructions in French, *International Journal on Artificial Intelligence Tools*, Vol.14, N°1&2, 125-136.
DESCLÉS J.-P. (2006). Contextual Exploration Processing for Discourse Automatic Annotations of Texts. *FLAIRS-19 (International Florida Artificial Intelligence Research Society Conference)*, 281-284.
KANG J.Y., DESCLÉS J.-P. (2008). Korean Parsing based on the Applicative Combinatory Categorial Grammar. *The 22nd Pacific Asia Conference on Language, Information and Computation*, 215-224.

[1] Ces deux dictionnaires sont chacun construits pour le français et le coréen. Ils sont accompagnés de dictionnaires restreints pour l'anglais et le japonais : le dictionnaire pour le français contient une centaine d'entrées pour l'anglais et celui pour le coréen une centaine d'entrées pour le japonais, de façon à tester la pertinence des analyses linguistiques de problèmes syntaxiques particulièrement complexes.

RefGen, outil d'identification automatique des chaînes de référence en français

Laurence Longo Amalia Todirascu

Université de Strasbourg, 22 avenue René Descartes, 67084 Strasbourg Cedex, France
longo@unistra.fr, todiras@unistra.fr

Nous présentons *RefGen*, un outil d'identification automatique des chaînes de référence (CR) en français. Les CR sont composées d'au moins trois expressions référentielles (Schnedecker, 1997). Développé dans un cadre industriel[1], *RefGen* est un prototype (développé en Perl et en Java) pouvant être intégré dans un système de détection automatique de thèmes. L'architecture de *RefGen* est modulaire et composée d'un étiquetage fin, d'un module d'annotation des expressions référentielles (groupes nominaux simples et complexes, entités nommées) et d'un module de calcul de la référence. *RefGen* utilise aussi une série de paramètres spécifiques au genre textuel pour calculer les relations de référence (distance entre les maillons d'une CR, nombre de maillons d'une CR, etc).

Pour l'étiquetage, *RefGen* utilise le catégoriseur TTL[2] (Ion, 2007) dans sa version française. Développé en Perl, TTL utilise le jeu d'étiquettes morphosyntaxiques fin proposé dans le projet Multext (Ide et Véronis, 1994), permettant de préciser des informations comme le genre, le nombre, le mode, le temps, etc. En plus de cet étiquetage fin, TTL identifie certains noms propres et fournit une analyse syntaxique partielle en chunks (groupes nominaux, groupes prépositionnels, groupes adjectivaux et groupes verbaux). Les sorties étiquetées sont disponibles en format XML.

Pour identifier les relations de référence entre les différentes entités du discours, *RefGen* annote d'abord les diverses expressions référentielles contenues dans les CR (groupes nominaux, entités nommées) avec RefAnnot. Ainsi, le module d'annotations RefAnnot (développé en Java) applique un ensemble de règles morphosyntaxiques pour identifier les expressions référentielles. Ces règles sont définies dans un format XML facilement transformable vers un autre format. RefAnnot identifie les groupes nominaux complexes (CNp, groupes nominaux modifiés par deux groupes prépositionnels au plus), plus informatifs, qui introduisent de nouveaux éléments dans le discours ainsi que certaines entités nommées (noms de personnes, organisations, lieux et fonctions). Pour faciliter le tri des divers candidats anaphoriques, les emplois impersonnels du pronom « il » sont aussi annotés. Ainsi, *RefGen* ne les prendra pas en compte lors de son calcul. Les sorties de RefAnnot sont disponibles en format xml mais aussi en html (pour faciliter la lisibilité).

Le texte enrichi en annotations passe alors dans le module de calcul de la référence CalcRef (développé en Java). L'algorithme mis en place utilise des paramètres liés au genre textuel pour sélectionner les premiers maillons des chaînes de référence mais aussi un score d'accessibilité global calculé à partir de l'échelle d'accessibilité d'(Ariel, 1990). Puis, la sélection des paires antécédent-anaphore s'effectue par la validation d'une série de contraintes (lexicales, syntaxiques, sémantiques) fortes et faibles. Une fois les paires identifiées, *RefGen* construit les CR suivant la propriété de transitivité.

Outre son utilisation première (l'identification des CR), *RefGen* peut aussi être utilisé comme outil de pré-annotation de corpus. D'autres règles morphosyntaxiques peuvent être facilement ajoutées à RefAnnot suivant les besoins (identification de dates, d'évènements, ou typage plus fin des entités nommées (fonction administrative, fonction politique, etc.). Une interface graphique est en cours de réalisation pour faciliter l'utilisation de l'outil.

ARIEL M. (1990). *Accessing Noun-Phrase Antecedents*. Londres : Routledge.

IDE N., VERONIS J. (1994). MULTEXT (Multilingual Tools and Corpora). *Actes de IAACL*, Kyoto.

ION R. (2007). Word Sense Disambiguation Method Applied to English and Romanian. Thèse de doctorat, Bucharest.

SCHNEDECKER C. (1997). « Nom propre et chaînes de référence », *Recherches Linguistiques*, 21, Paris, Klincksiek.

[1] L'outil a été développé dans le cadre d'une convention CIFRE avec la société RBS, Strasbourg (www.rbs.fr).

[2] TTL est disponible comme service Web sur la plate-forme Weblicht (https://weblicht.sfs.uni-tuebingen.de/). Un code d'accès est nécessaire (disponible sur simple demande).

LOL : Langage objet dédié à la programmation linguistique

Jimmy Ma, Mickaël Mounier, Helena Blancafort, Javier Couto, Claude de Loupy

(1) Syllabs, 15 rue Jean-Baptiste Berlier, 75013 Paris, France
{ma, mounier, blancafort, couto, loupy}@syllabs.com

LOL (*Linguistic Object Language*) est un langage dédié à la description d'objets linguistiques développé par la société Syllabs. Ce langage s'intègre dans une plateforme industrielle et permet aux linguistes d'écrire des règles d'extraction d'information ainsi que des règles de correction d'étiquetage morphosyntaxique. Lors de la conception de ce langage, l'idée était de proposer un vrai langage de programmation qui soit à la fois puissant au niveau de l'expression et à la fois simple à utiliser par des linguistes. De plus, ce langage permet une manipulation de plus haut niveau sans nuire aux performances du système produit.

LOL est un langage déclaratif qui permet de visualiser la langue sous la forme d'un langage de description de connaissances linguistiques plutôt que d'un langage de programmation. LOL est aussi un langage objet avec des objets prédéfinis comme les *tokens* (mots, préfixes, etc.), les phrases, etc. Les linguistes peuvent définir leurs propres objets, des listes d'éléments, ou des objets plus complexes reconnus à l'aide de patrons. Les spécifications écrites par les linguistes sont interprétées et mises en relation avec l'ensemble des ressources et outils de base développés par Syllabs (un lexique morphosyntaxique, un segmenteur, un étiqueteur morphosyntaxique, un *guesser*). Cette plateforme inclut également un outil de visualisation en html. La sortie peut actuellement être fournie sous format txt, XML ou JSON. Les analyseurs ainsi produits sont mis à disposition via des APIs REST (web services) sur les plateformes de Syllabs.

Concernant la manipulation des objets linguistiques, le linguiste manipule des *tokens* et ses différents attributs. Il peut ainsi accéder à différentes informations du *token*: 1) classe du *token* (mot, url, etc.), 2) information lexicales et morphosyntaxiques, 3) information graphique (typographie, nombre de caractères, etc.), 4) informations sur le *guessing* (s'il s'agit d'un *token* inconnu et si oui, s'il a été deviné) ; 5) positionnelles pour faire des conditions en fonction de la position dans le texte (ex : début de phrase ou de paragraphe). Ci-dessous un exemple de règle de correction et de règle d'extraction.

```
correction_rule  {
// correction d'erreur
//due à l'ambiguïté Nom-Adjectif
        [conditions]
        token.POS(D) | token.POS(X)
        token.POS(A)& token.ambig(N)
        token.POS(Sp)
        [actions]
        match[2].POS=N
}
```

```
extraction_rule {
[conditions]
!left_filter_potentialPN | token.class(begin)
f: FirstNameCap{1,2}
l: (PREMOD_NOM)? FamilyNameCap //(token.string("-")? FamilyNameCap)?
[actions]
        create Person[f,l] : priority(1); confidence = 1.0 {
                firstname = match[f]
                lastname  = match[l]
        }
}
```

Figure 1 : Exemples de règles

Aujourd'hui LOL est utilisé dans plusieurs applications industrielles commercialisées, notamment pour des applications basée sur l'extraction d'information (par ex. analyse de tonalité) et pouvant s'intégrer dans un processus de veille ou de tagging automatique de textes. Le linguiste peut également utiliser l'outil pour améliorer l'étiqueteur morphosyntaxique, voire le spécialiser sur le corpus du client en jouant sur les différentes propriétés du token (par ex. : conditions d'application d'une règle en fonction de la classe du token, la position dans le texte, le contexte et graphie). LOL est indépendant de la langue et est utilisé dans des applications industrielles ou de recherche en 8 langues, dont le chinois et le russe. Nous prévoyons d'ouvrir la plateforme de manière à permettre à des utilisateurs de créer leurs propres analyseurs.

Aligner : un outil d'alignement et de mesure d'accord inter-annotateurs

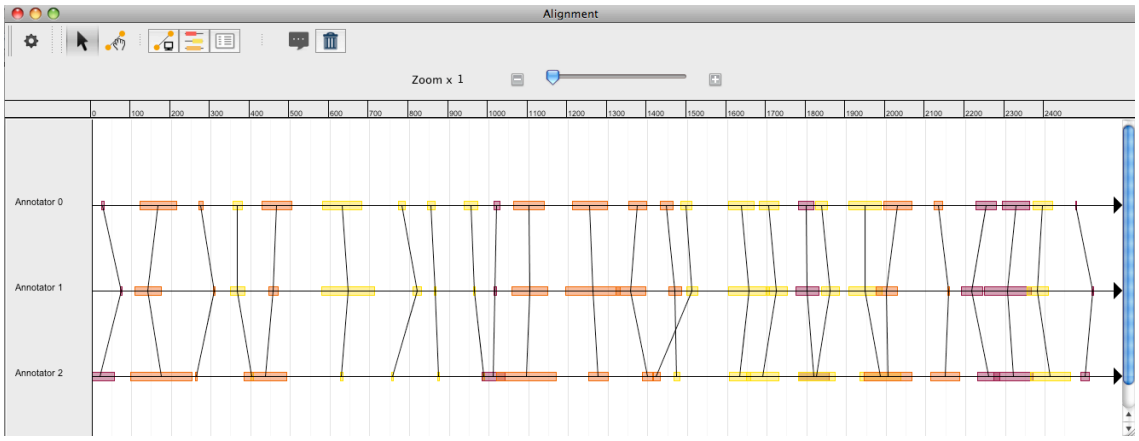
Yann Mathet[1], Antoine Widlöcher[1]

(1) GREYC, UMR CNRS 6072, Université de Caen, 14032 Caen Cedex
{prenom.nom}@unicaen.fr

Une méthode unifiée d'alignement et de mesure d'accord inter-annotateurs a été développée par (Mathet&Widlöcher 2011), consistant à considérer les annotations concurrentes faites par plusieurs annotateurs comme un facteur de désordre que le choix du meilleur **alignement** va minimiser. Cette notion de désordre, appelée entropie, est par ailleurs considérée par rapport au désordre induit par un processus aléatoire sur les mêmes textes pour établir une **mesure d'accord**, suivant la formule (1) :

$$\text{accord} = (\text{entropieHasard} - \text{entropie}) / \text{entropieHasard} \qquad (1)$$

Cette méthode a été implémentée (dans une version légèrement relaxée) et intégrée à la plateforme d'annotation GlozzQL sous forme d'un outil nommé Aligner, objet de la présente démonstration.



Lorsqu'un jeu d'annotations est chargé, l'outil crée une ligne horizontale pour chaque annotateur trouvé (3 annotateurs dans notre exemple ci-dessus), et empile ces dernières verticalement. Une ligne représente l'intégralité du texte (graduée en caractères), et porte, sous forme de segments colorés (selon leurs catégories), les différentes unités annotées par un annotateur donné. Il est alors possible de lancer le module automatique d'alignement qui crée, lorsqu'il les juge pertinents au regard de la méthode proposée, des alignements inter-annotateurs sous forme de lignes (relativement) verticales. Une valeur d'entropie correspondante est automatiquement calculée. S'il le souhaite, l'utilisateur peut modifier les alignements proposés directement à la souris. Par ailleurs, le système est capable de générer automatiquement des annotations aléatoires judicieusement disposées (en s'inspirant de ce qu'il constate en corpus), permettant de créer une référence d'entropie aléatoire de bon niveau pour un corpus donné. Pour cela, il suffit de placer un certain nombre d'exemples d'annotation du corups concerné dans un répertoire et de le soumettre à l'application. Dès lors, l'utilisateur dispose, à chaque fois qu'il charge un jeu d'annotations du corpus en question, non seulement de la génération de ses alignements, mais aussi de la mesure d'accord telle que donnée en formule (1). Un exemple complet sur corpus réel sera montré lors de la démonstration, ainsi que les déductions qui peuvent en être tirées sur la qualité des annotations d'un annotateur par rapport à l'ensemble des annotateurs.

Références

MATHET Y. ,WIDLÖCHER A. (2011), Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. *TALN 2011*, à paraître (soumission 78).

GlozzQL : un langage de requêtes incrémental pour les textes annotés

Yann Mathet[1], Antoine Widlöcher[1]

(1) GREYC, UMR CNRS 6072, Université de Caen, 14032 Caen Cedex
{prenom.nom}@unicaen.fr

GlozzQL est un langage de requêtes couplé à un moteur d'interrogation dédié et intégré à la plateforme d'annotation Glozz. Ses deux points forts sont : (1) son immersion au coeur du processus d'annotation, permettant de requêter en même temps que l'on annote, et vice-versa (la figure 2 montre comment on peut naviguer entre les résultats de requêtes et les annotations du texte en deux clics successifs) ; (2) son caractère incrémental, au sens où des requêtes peuvent être construites sur la base d'autres requêtes (on cherche les annotations répondant à telle contrainte parmi celles répondant déjà à telle autre contrainte).

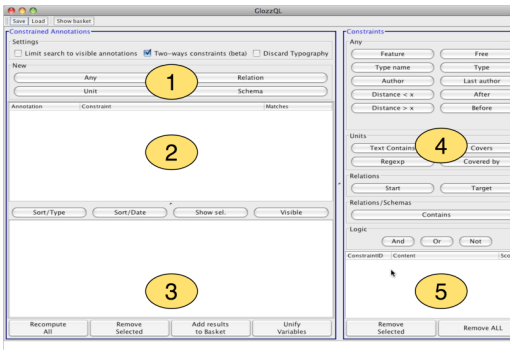


figure 1

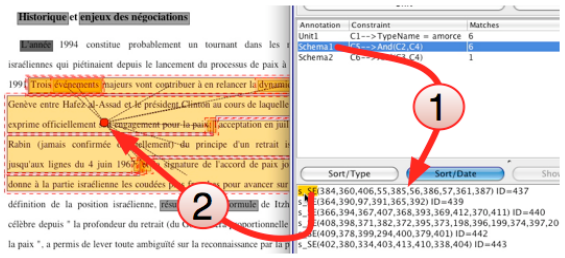


figure 2

Ce langage repose sur deux concepts inter-dépendants, les **Constraints**, définissant chacune une condition qu'une annotation doit vérifier pour être sélectionnée, et les **ConstrainedAnnotations**. ensemble d'annotations satisfaisant une contrainte donnée. C'est parce qu'une Constraint peut elle-même s'appuyer sur une ConstrainedAnnotation que des constructions incrémentales peuvent être peu à peu constituées, et produire ainsi un pouvoir expressif fort. S'adressant à un public large, ce langage est intégré sous forme **entièrement graphique** au sein de Glozz, permettant de constituer des requêtes aussi complexes que souhaité à la souris, comme illustré en figure 1 : les fenêtre 1 et 4 permettent respectivement de créer des ConstrainedAnnotations et des Constraints, lesquelles apparaissent respectivement en 2 et 5, tandis que l'ensemble des résultats relatifs à une ConstrainedAnnotation de 2 apparaît au-dessous en fenêtre 3.

La démonstration abordera GlozzQL de façon pratique et s'appuiera sur 3 exemples concrets de constitution de requêtes. Le premier, la simple détection d'annotations disposant d'une valeur de trait particulière, à savoir genre=féminin. Le second mettra en oeuvre la constitution progressive de contraintes complexes, comme la recherche d'une annotation constituée d'une relation allant d'une annotation de genre=féminin à une annotation de genre=masculin. Le troisième permettra d'entrevoir des concepts avancés de Glozz tels que **l'unification**, permettant de considérer les ConstrainedAnnotations comme des variables d'un système d'équation, et ainsi de pouvoir détecter des configuration complexes telles que des relations en triangle, ou des relations orientées dans le sens inverse du sens de lecture. Enfin, la notion de **panier** permettant de collecter sélectivement des résultats de requêtes afin soit de les enregistrer, soit au contraire de les retirer du texte annoté, sera illustrée à partir des résultats précédents.

Références

MATHET Y. ,WIDLÖCHER A. (2011), Stratégie d'exploration de corpus multi-annotés avec GlozzQL. *TALN 2011*, à paraître (soumission 186).

EASYTEXT : un système opérationnel de génération de textes

Frédéric Meunier[1] Laurence Danlos[2] Vanessa Combet[1]
(1) Watch System Assistance
(2) Université Paris Diderot, ALPAGE
frederic.meunier@watchsystance.com, laurence.danlos@linguist.jussieu.fr,
vanessa.combet@watchsystance.com

EASYTEXT est un système de génération de textes opérationnel auprès de Kantar Media, une filiale de TNS-Sofres. Cette société compile pour ses clients des données chiffrées sur leurs investissements publicitaires et envoie à chaque client sept tableaux tous les mois, comme celui de la Figure 1. Avant EASYTEXT, ces tableaux étaient accompagnés d'un commentaire général rédigé par un chargé d'étude. Le besoin s'est fait sentir d'assortir ce commentaire général de commentaires spécifiques à chaque tableau. La charge de rédaction étant alors trop lourde pour les chargés d'étude, l'idée a surgi de faire générer ces commentaires spécifiques par un système automatique.

EASYTEXT repose sur le formalisme G-TAG, un formalisme lexicalisé reposant sur les grammaires d'arbres adjoins (TAG) (Danlos, 1998). Ce formalisme a été étendu en amont pour les tâches de détermination du contenu et de structuration du texte, en suivant l'architecture décrite dans (Danlos & Ghali, 2002). La détermination du contenu revient à surligner certaines cellules du tableau. Cette tâche a été guidée par les règles métier indiquées par les chargés d'étude de TNS-Sofres. La structuration du texte consiste à introduire des relations rhétoriques (e.g. $Contraste$ ou $Parallèle$) entre le contenu sémantique des cellules surlignées, voir (Danlos *et al.*, 2001)).

G-TAG avait été implémenté dans les années 90' en ADA par F. Meunier (1997). Celui-ci a re-implémenté G-TAG (avec les extensions en amont décrites ci-dessus) en .NET, ce qui permet à EASYTEXT d'être intégré au système d'information de TNS-Sofres. La génération d'un commentaire comme celui de la Figure 1 demande en moyenne 400ms de CPU-Times. Cette implémentation intègre des outils ergonomiques pour renseigner les bases lexicales TAG et pour visualiser les différentes structures arborescentes dynamiquement construites ou en cours de construction.

Les bases lexicales ont été renseignées par une linguiste, V. Combet, qui a travaillé en étroite collaboration avec les chargés d'étude de TNS-Sofres. Une attention particulière a été portée sur la variation linguistique afin de ne pas produire des textes monotones qui auraient lassé les clients de TNS-Sofres. Cette variation concerne principalement les choix lexicaux (e.g. *augmenter, être en (forte/moyenne/faible) augmentation/hausse, (presque/plus que) doubler/tripler*) et l'ordre des syntagmes à position plus ou moins libre.

TNS-Sofres a été satisfait des résultats de EASYTEXT (même au delà de ses espérances), et donc commercialise ce service à ses clients depuis Avril 2010. EASYTEXT est donc un des tous premiers systèmes de génération de textes opérationnel en France, et il en existe peu en dehors de l'hexagone.

Figure 1 : Exemple de tableau et de commentaire généré automatiquement

Références

DANLOS L. (1998). G-TAG : un formalisme lexicalisé pour la génération de textes inspiré de TAG. *Revue TAL*, **39**(2).

DANLOS L., GAIFFE B. & ROUSSARIE L. (2001). Document structuring à la SDRT. In *International workshop on text generation - ACL*, p. 94–102, Toulouse.

DANLOS L. & GHALI A. E. (2002). A completed and integrated NLG system using NLU and AI tools. In *Proceedings of COLING'02*, Tapei, Taiwan.

MEUNIER F. (1997). *Implémentation du formalisme G-TAG*. Thèse de doctorat en informatique, Université Denis Diderot, Paris 7.

Restad : un logiciel d'indexation et de stockage relationnel de contenus XML

Yoann Moreau Eric SanJuan Patrice Bellot
LIA, 339, chemin des Meinajaries 84911 AVIGNON Cedex 9
{yoann.moreau,eric.sanjuan,patrice.bellot}@univ-avignon.fr

Restad [1] est un outil pour charger de grands nombres de documents XML dans une base de données PostgreSQL [2]. La structure XML ainsi que le contenu, y compris celui des attributs, est stocké sous forme relationnelle. Cela est fait sans aucun présupposé sur les DTDs des fichiers et permet de gérer tous les standards XML. Il est le format le plus courant pour des données semi-structurées (issues d'applications de bureautique, annotées par des analyseurs syntaxiques, enrichies de multiples annotations sémantiques...).

Restad reprend ainsi les fonctionnalités d'importation de TopX (Theobald *et al.*, 2005) et XRel (Yoshikawa *et al.*, 2001) en l'adaptant au SGBDR libre PostgreSQL et surtout, en intégrant la gestion des attributs de balises.
La structure hiérarchique des documents (balises et attributs) est représentée de manière relationnelle dans des tables. Le texte du document est enregistré en tant que bloc de texte nettoyé de toute balise XML. La table des balises conserve les positions de début et de fin de chaque balise, tandis que l'index plein-texte conserve pour chaque mot sa position.

Cette approche permet de stocker la forme et le contenu de documents XML, sans perte d'information. On peut ensuite utiliser les index de la base pour effectuer des recherches plein-texte en réduisant les requêtes à un ou plusieurs sous ensembles de l'arborescence des documents. L'utilisation d'un SGBD offre les performances du langage SQL pour effectuer des requêtes complexes.

La première version de Restad est écrite en Ruby et disponible sous licence GPL. Elle a été testée sur le corpus XML de la campagne INEX 2010 de 52Go comprenant l'ensemble du wikipedia en anglais enrichi de multiples annotations sémantiques (Schenkel *et al.*, 2007).
La base de données finale, après création de tous les index occupe un espace disque inférieur à 5 fois la taille du corpus. L'outil permet de re-générer le contenu XML d'un document très rapidement grâce aux index de la base. Différents tests sont prévus pour évaluer les performances avec des requêtes utilisant les balises XML. L'outil pourrait par la suite être facilement adapté à tout format de document arborescent et à tout autre SGBD.

Remerciements

Ces recherches ont bénéficié du soutien financier de l'Agence Nationale de la Recherche (ANR 2010 CORD 001 02) en faveur du projet CAAS.

Références

SCHENKEL R., SUCHANEK F. M. & KASNECI G. (2007). Yawn : A semantically annotated wikipedia xml corpus. In A. KEMPER, H. SCHÖNING, T. ROSE, M. JARKE, T. SEIDL, C. QUIX & C. BROCHHAUS, Eds., *BTW*, volume 103 of *LNI*, p. 277–291 : GI.

THEOBALD M., SCHENKEL R. & WEIKUM G. (2005). An efficient and versatile query engine for topx search. In K. BÖHM, C. S. JENSEN, L. M. HAAS, M. L. KERSTEN, P.-Å. LARSON & B. C. OOI, Eds., *VLDB*, p. 625–636 : ACM.

YOSHIKAWA M., AMAGASA T., SHIMURA T. & UEMURA S. (2001). XRel : A Path-Based Approach to Storage and Retrieval of XML.

1. Relational Storage for Tagged Documents (https ://github.com/ymoreau/Restad)
2. http ://www.postgresql.org

Une chaîne d'analyse des e-mails pour l'aide à la gestion de sa messagerie

Gaëlle Recourcé[1]
(1) Kwaga SAS, 15 rue J-B. Berlier, 75013 Paris, France
recource@kwaga.com

Au sein de la société Kwaga, une équipe d'ingénieurs a réalisé une chaîne d'analyse des e-mails utilisée dans des applications d'aide à la gestion de sa messagerie pour les particuliers et les professionnels. Ce cœur technologique intègre au travers d'un chaînage UIMA, plusieurs composants de TAL issus de la recherche universitaire ou réalisés en interne.

1 Une chaîne d'analyse des mails

1.1 Import des messages

La première étape de la chaîne d'analyse des mails consiste à se connecter à un serveur IMAP pour importer une boîte mail (messages entrants et sortants). Les mails sont caractérisés par une en-tête (expéditeur(s), destinataire(s) directs ou en copie, date), un contenu (sujet et corps) et par leur organisation en conversations : un ensemble de messages échangés sous un même sujet par un ensemble de participants peut être considéré comme une séquence de répliques dans une conversation intégrant des apartés (transferts).

1.2 UIMA – annotation des mails

Le corps de la chaîne d'analyse de Kwaga est implémenté dans le cadre d'une séquence d'annotations UIMA (*Unstructured Information Management Architecture)*. Cette chaîne se subdivise en trois étapes, la détection du corps textuel, l'analyse linguistique, et l'interprétation.

1. Par le jeu des réponses et transferts, le corps d'un e-mail se structure par des niveaux de reprises successifs, indiquant le degré de nouveauté dans la conversation. Cette première annotation (CAS) consiste à repérer le texte nouveau, i.e. effectivement produit par le dernier expéditeur dans la conversation. Ce texte est par ailleurs soumis à la reconnaissance de langue (grâce à un module adapté de TextCat) et diverses expressions régulières sont appliquées sur les champs structurés et sur le corps de message permettant de calculer des informations caractérisant le message qui seront utilisées dans la phase d'interprétation.

2. La phase d'analyse linguistique est réalisée par l'application de graphes sur le corps et le sujet du mail par le biais d'une librairie JNI d'Unitex. Ces grammaires locales permettent d'une part de repérer les éléments de la structure du message (formules introductives, salutations finales, et signatures) et d'autre part des éléments du texte utilisés pour l'interprétation : phrases prototypiques (demandes d'action, proposition de rencontre, facturation…) ou éléments d'information assimilables à des entités (dates, mots de passe...). Ces automates sont appliqués en une passe unique, les sorties constituant une nouvelle annotation du CAS UIMA.

3. Dernière phase d'annotation UIMA, l'interprétation exploite les informations contextuelles (data, expéditeurs, destinataires, …) et les combine avec les indices linguistiques découverts dans le corps du texte, la signature ou le sujet pour calculer la catégorie du mail et les éventuelles informations associées telles les informations de contact. Ces informations dépendent aussi du contexte de conversation : un mail en réponse peut, dans certaines conditions, hériter certaines propriétés du mail qui l'a précédé.

2 Démonstrations (en ligne)

- Analyse d'e-mails à la demande (en français ou en anglais).
- Présentation de la création à la volée de corpus d'e-mails (serveur IMAP Gmail)
- Extraction d'information dans les e-mails – factures, mots de passe, fiches contact
- Catégorisation des e-mails : mail importants et Bac'n.

Démonstration d'un outil de « Calcul Littéraire »

Jean Rohmer 1

(1) Ecole Supérieure d'Ingénieurs Léonard de Vinci 92916 Paris La Défense Cedex
jean.rohmer@devinci.fr

Sous le nom de « Calcul Littéraire » (ou « Litteratus Calculus ») nous avons développé un démonstrateur de représentation de connaissances en langage naturel, qui peut être vu à la fois comme un gestionnaire de connaissances exprimées en langage naturel, et comme une infrastructure utile pour l'analyse de documents ou de corpus de textes. Ce travail fait suite à la réalisation d'un produit industriel à base de réseaux sémantiques, Ideiance, qui a été commercialisé par la société éponyme à partir de 1996, et utilisé dans de grandes entreprises, ainsi que par les Armées pour le Renseignement Militaire, en particulier à partir du moment où il a été repris par la société Thales en 2004. Précurseur du Web Sémantique, Idéliance permet de créer et d'exploiter collectivement, sur un serveur, des réseaux sémantiques, c'est-à-dire des énoncés de la forme Sujet / Verbe / Complément, le tout assorti de la notion de Catégorie. L'expérience a montré qu'un formalisme aussi élémentaire reste rebutant pour au moins 95% des utilisateurs, qui refusent l'effort minimal de modélisation correspondant, même habillé d'éditeurs et visualiseurs graphiques ergonomiques.

Ceci nous a conduit à imaginer un outil où les énoncés ne seraient en rien contraints, mais exprimés sous forme de phrases en langage naturel, sans autre restriction. Il sera en effet difficile à un utilisateur potentiel de dire « faire une phrase, c'est trop compliqué ou trop abstrait pour moi », comme ils le disent dès qu'il s'agit de créer un graphe sémantique et ses catégories. Le principe du « calcul littéraire » est le suivant :

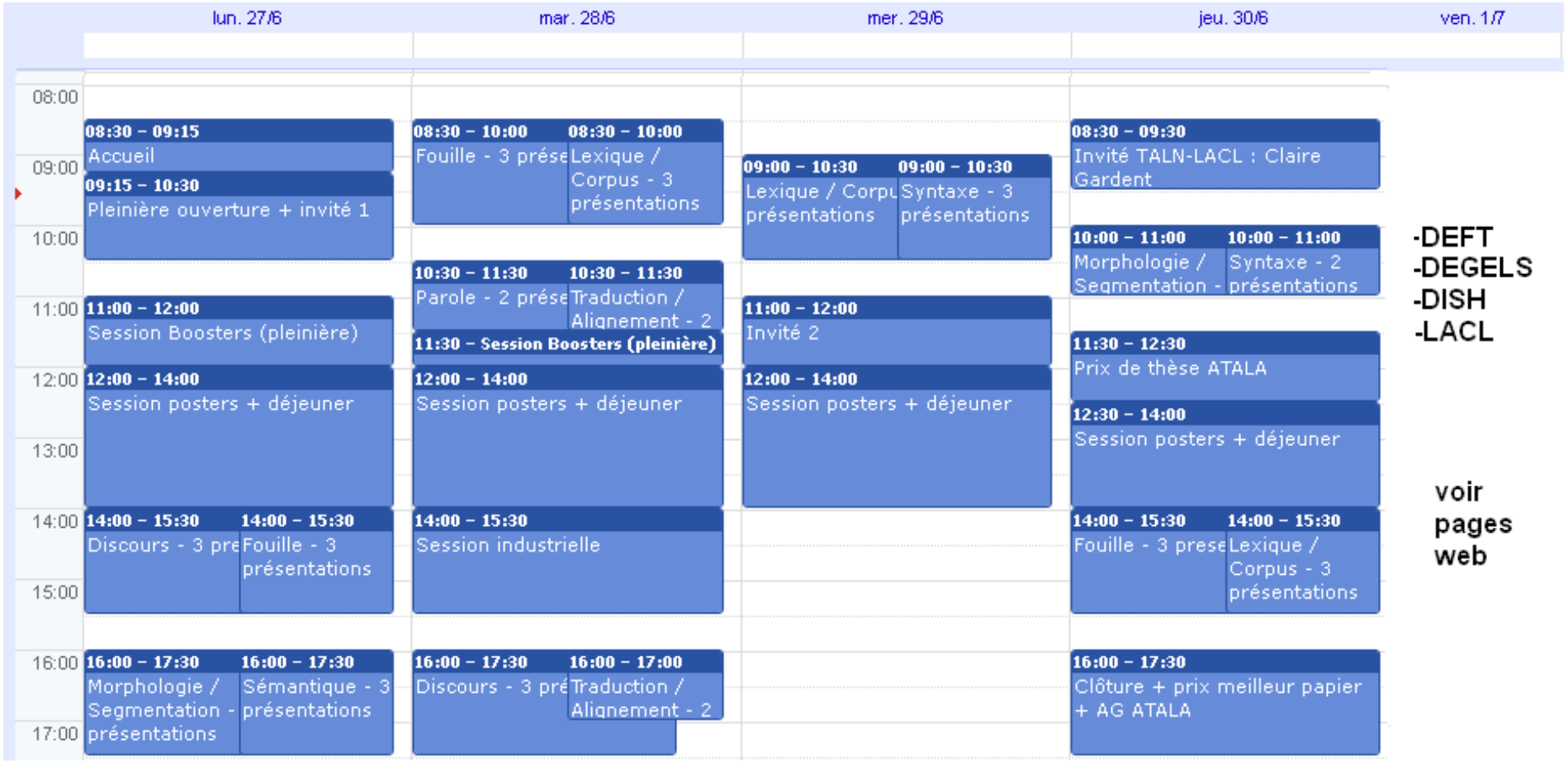
On constitue un ensemble d'énoncés ou phrases indépendantes les unes des autres, appelées ***inférons***. Plus précisément un ***inféron*** est une phrase minimale et autonome compréhensible par une certaine communauté de personnes. Pour tout couple d'***inférons***, on construit leur intersection en terme de mots –après éventuelle lemmatisation-. Ces intersections s'appellent des ***interlogos***. On constitue ainsi automatiquement un graphe biparti d'***inférons*** et d'***interlogos***. A ce graphe, on peut applique un ensemble d'opérateurs visuels de navigation et contraintes ensemblistes, que nous appelons « azimuts », qui se rapprochent de la projection des graphes conceptuels, et plus généralement de la logique du premier ordre. Il faut noter que le calcul automatique des ***interlogos*** s'apparente à un mécanisme d'extraction d'entités nommées, mais sans la contrainte de disposer au préalable d'ontologies. Une fois le graphe d'***inférons*** et d'***interlogos*** constitué, on peut lui appliquer des outils de requête, de génération de tableaux et de rapports, de mise à jour habituels pour des informations structurées, aboutissant à une sorte de « ***tableur littéraire*** ».

Nous démontrerons sur une base de plus de 50 000 énoncés qu'une utilisation « brutale » du langage naturel comme format de représentation apportait un « retour sur investissement » très significatif à l'utilisateur.

Une seconde approche introduisant plus de composants linguistiques a été expérimentée et sera démontrée : elle consiste à enrichir chaque énoncé d'une analyse syntaxique et sémantique avec des analyseurs comme XIP de Xerox XRCE, ou ceux utilisés dans le projet ANR PASSAGES. On obtient des ***interlogos*** plus riches, qui permettent des navigations de phrase en phrase plus sophistiquées, reposant sur la sélection des rôles syntaxiques ou sémantiques des ***interlogos.***

Enfin, l'outil présenté peut aussi être vu comme un outil de génie linguistique, en particulier en linguistique des corpus, pour explorer un ensemble de phrases et/ou de documents, et y découvrir, par émergence, des régularités ou des différences.

Notre objectif à court terme est de développer un produit industriel, dans un cadre approprié.



-DEFT
-DEGELS
-DISH
-LACL

voir pages web