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

détermination des mots représentatifs à partir de contextes. Différentes améliorations ont été proposées par la suite (Zhang *et al.*, 2005) (He *et al.*, 2006) (Bollegala *et al.*, 2010). La méthode que nous proposons se fonde sur le même principe. Mais elle se distingue de ces travaux antérieurs notamment par l'exploitation des résultats de l'analyse syntaxique permettant d'extraire non pas les contextes linéaires de surface mais structurels. Par ailleurs, afin d'améliorer l'efficacité de la classification, nous avons pleinement exploité la dualité de la représentation des relations sémantiques, comme proposé dans (Bollegala *et al.*, 2010), et nous avons mené une expérience comparative de différentes méthodes de calcul de similarité.

Nous allons d'abord introduire deux notions principales : dualité de la représentation des relations et chemins de relations (§ 2) avant de décrire notre méthode non-supervisée d'extraction de relations (§ 3) avec les différentes similarités que nous avons employées. La section suivante sera consacrée à la présentation du résultat d'évaluation (§ 4), avant de conclure notre exposé par les perspectives de nos travaux (§ 5).

2 Deux notions principales

Une relation sémantique existant entre deux éléments peut être définie de deux manières : définition extensionnelle et définition intensionnelle. Une définition extensionnelle de la relation R consiste à créer la liste complète des instances de cette relation, notée $C(R)$. Par exemple, la définition extensionnelle de la relation *EstPrésident* entre un individu X et une organisation Y énumère toutes les paires d'EN entretenant cette relation comme : C(*EstPrésident*) $=$ {(*Joseph S. Blatter, FIFA*), (*Martin Hirsch, Emmaüs France*), ...}. Une définition intensionnelle de la relation R spécifie un ensemble de propriétés permettant d'identifier cette relation, noté $P(R)$. Ainsi, la même relation *EstPrésident* peut être définie de manière intensionnelle par des patrons lexicaux tels que : P(*EstPrésident*) $=$ {*"X est le président de Y"*, *"X, président de Y"*, ...}. Comme nous utilisons des résultats d'analyse syntaxique, cette liste est constituée, non pas de ces patrons textuels, mais de ce que nous appelons des « chemins syntaxiques de relations ».

Nous considérons les chemins reliant deux EN dans un arbre syntaxique comme représentant leur relation et nous les appelons chemins syntaxiques de relations. Comme dans notre méthode semi-supervisée (Nakamura-Delloye & Villemonte de La Clergerie, 2010), la première opération de notre extraction consiste à identifier tous les couples d'EN et les chemins qui les relient dans les arbres syntaxiques de dépendance. Par exemple, dans la phrase « *Xavier Bertrand succède officiellement à Patrick Devedjian* », on identifie un couple d'EN (*Xavier Bertrand*, *Patrick Devedjian*), et le chemin de relation qui les relie, $(X) \rightarrow^{suj\text{-}v}$ *succéder*$_v$ $\leftarrow^{cpl\text{-}v(à)}(Y)$.

3 Méthode non-supervisée d'acquisition des relations entre EN

Tous les couples d'EN qui partagent le même chemin de relation p_i sont regroupés pour constituer des couples tels que (P, C), P un ensemble de chemins $\{p_i\}$ (contenant initialement un seul chemin) et C l'ensemble des couples d'EN regroupés $\{c_1, ...c_n\}$. On considère ces ensembles comme des sous-ensembles de relations et l'opération d'acquisition de relations se définit alors comme la constitution, par fusions de ces sous-ensembles, des ensembles complets pour différentes relations, $R_i = (P(R_i), C(R_i))$. Ces fusions sont réalisées par la classification de ces sous-ensembles à l'aide de leur similarité. Étant donné la dualité de représentation de ces relations, leur similarité peut également être évaluée sur deux plans, en termes de définitions intensionnelle (c'est-à-dire la similarité des chemins partagés) ou extensionnelle (c'est-à-dire la similarité des instances de relations).

Calcul des similarités La similarité des chemins est calculée selon la similarité lexicale des mots composant les chemins. Chaque chemin est d'abord représenté dans un espace vectoriel par ses composants lexicaux pondérés avec la mesure *tf.idf*. Dans nos travaux, seuls les verbes (sauf *être* et *avoir*) et les noms sont pris en compte. La valeur du mot i est donc calculée comme suit : $m_i = tf_i \cdot idf_i$ où tf_i correspond à la fréquence du mot, et idf_i est calculé à partir du nombre de chemins où le terme i n'apparaît pas, compte tenu du nombre total de chemins. Les similarités entre les vecteurs représentant les chemins sont ensuite calculées selon la similarité cosinus.

La similarité des instances de relations est calculée selon le nombre de couples d'EN communs. Deux mesures ont été expérimentées. La première est basée sur l'indice de Jaccard. La similarité entre deux ensembles C_i et C_j est alors calculée comme : $sim(C_i, C_j) = \frac{nb(C_i,C_j)}{nb(C_i)+nb(C_j)}$ où $nb(X)$ est le nombre de couples d'EN dans l'ensemble X et $nb(X, Y)$ le nombre de couples d'EN communs dans les ensembles X, Y. Le second type est basé sur la similarité cosinus avec les vecteurs des couples d'EN. La valeur du couple i de l'ensemble C_j est calculée comme : $c_i^j = tf_i^j \cdot idf_i$ où tf_i^j correspond à la fréquence du couple dans cet ensemble, et idf_i est calculé à partir du nombre de chemins qui ne relient pas le couple i, compte tenu du nombre total de chemins.

Classification des chemins et étiquetage des classes Avec ces similarités, la classification des sous-ensembles est réalisée par une méthode de classification en deux temps. Dans un premier temps, les sous-ensembles sont fusionnés selon la similarité lexicale des chemins. À ce stade, les couples (P_i, C_i) dont les chemins appartenant à l'ensemble P_i ont une similarité significative sont regroupés par une méthode de classification ascendante hiérarchique (CAH). Dans un deuxième temps, les éléments de l'ensemble C_i sont examinés pour regrouper ces couples selon la similarité des instances calculée par une des deux méthodes décrites précédemment.

À la fin de cette classification en deux temps, nous obtenons un ensemble de couples représentant les relations R_i, constitués d'un ensemble de chemins $P(R_i)$ correspondant à la définition intensionnelle, et d'un ensemble d'instances $C(R_i)$ correspondant à la définition extensionnelle. Ces classes ainsi construites sont ensuite étiquetées par le terme le plus représentatif des chemins partagés. La représentabilité du terme i dans cette classe j est la somme des valeurs de ses n occurrences : $rep(t_i^j) = \Sigma_{i=1}^{n} val(t_i^s)$, où la valeur de chaque occurrence $val(t_i^s)$ correspond au nombre de couples d'EN que relie le chemin s dans lequel le terme apparaît.

4 Évaluation

Une évaluation de notre méthode a été réalisée avec le résultat d'extraction d'un corpus constitué d'un mois de dépêches AFP (janvier 2007) contenant 41 731 phrases. Le corpus comporte les résultats non vérifiés de l'analyse syntaxique avec étiquetage des EN, fournis par l'analyseur FRMG (Villemonte de La Clergerie *et al.*, 2009), utilisant lui-même pour l'étiquetage d'entités nommées SxPipe (Sagot & Boullier, 2008).

4.1 Méthode d'évaluation

La longueur maximum de chemins est fixée à cinq. Cette valeur a été définie suite à l'étude de la productivité des chemins (i.e. nombre de couples d'EN qu'ils relient) selon la longueur, et du temps de calcul selon la longueur maximum définie. De plus, nous n'avons traité que les chemins reliant au moins deux couples d'EN différents. Nous avons également éliminé les chemins sans nœud intermédiaire, correspondant à la relation syntaxique réalisée par une juxtaposition. En effet, les relations sémantiques représentées par ces chemins telles que « appartenance » sont tellement larges qu'elles englobent des sous-classes telles que « *est président de* », « *est porte-parole*

de », qui sont utiles pour le peuplement de notre référentiel ontologique. Contrairement aux travaux comme (Hasegawa *et al.*, 2004) qui ne prenaient en compte que des couples ayant plus de trente cooccurrences, nous n'avons défini aucun seuil pour les couples, en supposant que l'utilisation des chemins syntaxiques permettait de repérer de manière efficace et fiable les couples d'EN entretenant une certaine relation sémantique. L'extraction des relations a été réalisée pour deux types de relations : relations *individu-organisation* (IND-ORG) et *individu-individu* (IND-IND). Pour la mesure de dissimilarité inter-classe lors de la classification, nous avons testé deux stratégies : l'une consiste à définir la similarité de la classe nouvellement constituée avec chaque autre classe par la plus petite similarité des deux classes regroupées (Min), et l'autre par la plus grande similarité des deux (Max). Le seuil pour l'arrêt des agrégations est défini à la similarité 0 comme dans les travaux antérieurs. Nous avons également testé trois types de classifications : un premier type réalisé uniquement avec la similarité des chemins, les deux autres réalisés en deux temps avec la similarité des chemins puis celle des instances calculée par indice de Jaccard ou par similarité cosinus.

Après avoir obtenu les résultats avec des configurations différentes, nous avons préparé manuellement une référence des instances pour 6 relations qui ont été extraites par toutes les méthodes expérimentées : *président, directeur, secrétaire* pour la relation IND-ORG et *avocat, porte-parole, remplacer* pour IND-IND. Les nombres d'instances repérées et vérifiées manuellement sont 336 pour IND-ORG et 83 pour IND-IND. Les instances de relations extraites par nos méthodes sont ensuite évaluées manuellement en correcte ou incorrecte. Les rappel, précision et F-mesure sont calculés comme suit :
Rappel : nombre de couples détectés parmi ceux qui figurent dans la liste des couples de référence, calculé par la formule suivante : $R = \frac{N_{correct}}{N_{couples\_de\_référence}}$
Précision : nombre de couples corrects parmi l'ensemble des couples détectés, calculé par la formule suivante : $P = \frac{N_{correct}}{N_{correct}+N_{incorrect}}$
F-mesure : score calculé par la combinaison du rappel et de la précision comme suit : $F = \frac{2RP}{R+P}$

4.2 Résultats et discussion

Le tableau 1 présente les résultats de l'évaluation de nos quatre méthodes : 1) similarité des chemins avec critère d'agrégation Max (Max-Chem), 2) similarité des chemins avec critère d'agrégation Min (Min-Chem), 3) similarité des chemins puis similarité des instances calculée avec l'indice de Jaccard (Min-Jac), et 4) similarité des chemins puis similarité des instances calculée avec similarité cosinus des vecteurs de couples d'EN (Min-Vec).

Toutes les méthodes ont été expérimentées avec critère d'agrégation Max, mais comme on peut le constater dans le résultat (Max-Chem) cette mesure entraîne un bruit trop important, constaté également dans des résultats non-présentés ici. Les rappels supérieurs à 1 signifient que ces méthodes ont identifié certaines instances de relations que nous n'avions pas pu repérer à la main lors de la constitution de la référence. La bonne précision que nous avons eue, tout en traitant des couples de fréquence très peu élevée, semble confirmer notre hypothèse et montrer la fiabilité des relations entre deux EN reliées par un chemin syntaxique. Un autre avantage de l'utilisation des chemins syntaxiques est que malgré la longueur maximum fixée à cinq nœuds intermédiaires, les instances de relations très éloignées (comme par une incise) dans la représentation linéaire peuvent être extraites.

La méthode Min-Chem a fourni de très bonnes précisions, mais les regroupements réalisés par cette méthode ne dépassaient jamais ceux entre les classes des chemins partageant les mêmes unités lexicales, ce qui explique ses taux de rappel limités. Les deux autres méthodes permettaient les fusions des classes dont les chemins ne présentaient aucune similarité lexicale, comme par exemple les classes comportant le terme *leader* et le terme *chef* ou encore les classes portant les termes *patron* et *PDG*. De même, les classes dont la similarité était nulle à

Type	IND-ORG				IND-IND			
Méthode	Max-Chem	Min-Chem	Min-Jac	Min-Vec	Max-Chem	Min-Chem	Min-Jac	Min-Vec
Rappel	0,732	0,711	0,937	**1,022**	0,835	0,973	**1,038**	**1,038**
Précision	0,682	**0,964**	0,944	0,834	0,926	**0,944**	0,887	0,887
F-mesure	0,686	0,732	**0,934**	0,911	0,872	**0,958**	0,952	0,952

TABLE 1 – Résultats d'évaluation

cause d'une mauvaise analyse syntaxique ont également été fusionnées par ces méthodes.

Les erreurs d'analyse syntaxique étaient en effet la source principale des erreurs. Par exemple, les syntagmes « *Le président du Sinn Féin,* ***Gerry Adams*** », « *le président palestinien* ***Mahmoud Abbas****, du Fatah* » ont été analysés par la même structure. Mais *Mahmoud Abbas* n'est pas président du *Fatah* et ce dernier aurait dû être rattaché directement à ce premier et non au terme *président*. Le terme *président*, lorsqu'il est utilisé dans le sens de chef de l'état, constitue souvent un prédicat unaire et n'attend pas deux arguments.

Par ailleurs, l'extraction des relations entre EN du même type pose des problèmes spécifiques. En effet, les classes formées pour IND-IND sont pour l'instant symétriques, c'est-à-dire qu'on sait seulement que les instances de la classe *remplacer*, par exemple, sont des couples en relation de *remplaçant et remplacé* sans savoir le rôle exact de chacun. Afin de résoudre ce problème il faudrait introduire une opération qui examine la similarité en inversant le sens de la relation de l'une des deux classes à comparer. Mais il existe aussi d'autres types de problèmes. Par exemple, dans la plupart des cas, le sujet est le remplaçant et le COD le remplacé, comme dans « *Alain Méar remplace Philippe Levrier, qui préside la commission Audiovisuel numérique terrestre* », mais dans « *Bush remplace Garner par Paul Bremer* » le sujet et le COD ne sont pas en cette relation et les informations contenues dans le chemin qui les relie ne suffit pas pour traiter correctement leur relation. Il existe également d'autres cas nous faisant ressentir la nécessité de plus d'information. Dans « *Nonce Paolini, directeur général de Bouygues Telecom* » « *Nicolas Boutaud, directeur commercial de Brittany Ferries* », les EN individu et organisation sont reliées par le même chemin. Ces couples d'EN sont donc considérés comme appartenant à la même classe de relation, alors que le *directeur général* peut tout à fait appartenir à la classe *patron* mais pas le *directeur commercial*.

5 Conclusion

Nous avons proposé une méthode non-supervisée d'extraction des relations entre entités nommées. Nous avons également présenté le résultat d'une expérience d'évaluation permettant de comparer quatre méthodes combinant différentes similarités. Le résultat montre l'efficacité des méthodes exploitant la dualité de la représentation des relations, mais pour déterminer la meilleure méthode du calcul de la similarité, il nous faudra réaliser l'évaluation d'un plus grand nombre de résultats. Nos prochains travaux porteront également sur l'intégration des relations extraites dans notre ontologie et l'amélioration de la méthode à cet effet.

Remerciements

Les travaux décrits dans cet article ont débuté avec le projet SCRIBO (Semi-automatic and Collaborative Retrieval of Information Based on Ontologies), labellisé par le pôle de compétitivité System@tic et financé par la DGE, et se sont poursuivis dans le cadre du projet EDyLex (Enrichissement Dynamique de ressources Lexicales multilingues

en contexte multimodal), financé par l'ANR (ANR-09-CORD-008).
Site internet du projet : `http://sites.google.com/site/projetedylex/`.

Références

AGICHTEIN E. & GRAVANO L. (2000). Snowball : Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries*, p. 85–94.

BANKO M., CAFARELLA M. J., SODERLAND S., BROADHEAD M. & ETZIONI O. (2007). Open information extraction from the web. In *IJCAI'07*, p. 2670–2676.

BOLLEGALA D., MATSUO Y. & ISHIZUKA M. (2010). Relational duality : Unsupervised extraction of semantic relations between entities on the web. In *Proc. of the 19th International Conference on World Wide Web (WWW 2010)*, p. 151–160.

BRIN S. (1998). Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*, p. 172–183.

HASEGAWA T., SEKINE S. & GRISHMAN R. (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, p. 415–422, Barcelona, Spain.

HE T., ZHAO J. & LI J. (2006). Discovering relations among named entities by detecting community structure. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, p. 42–48.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, p. 539–545.

IFTENE A. & BALAHUR-DOBRESCU A. (2008). Named entity relation mining using wikipedia. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco : European Language Resources Association (ELRA).

P. LANGLAIS & M. GAGNON, Eds. (2010). *Actes de TALN 2010 (Traitement automatique des langues naturelles)*, Montréal. ATALA, RALI-WeST.

MATSUO Y., MORI J., HAMASAKI M., ISHIDA K., NISHIMURA T., TAKEDA H., HASIDA K. & ISHIZUKA M. (2006). Polyphonet : An advanced social network extraction system from the web. In *Proc. of the 15th International Conference on World Wide Web (WWW 2006)*.

NAKAMURA-DELLOYE Y. & VILLEMONTE DE LA CLERGERIE E. (2010). Exploitation de résultats d'analyse syntaxique en vue d'acquisition de relations entre entités nommées. In (Langlais & Gagnon, 2010).

SAGOT B. & BOULLIER P. (2008). Sxpipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues*, **49**(2), 155–188.

STERN R. & SAGOT B. (2010). Détection et résolution d'entités nommées dans des dépêches d'agence. In (Langlais & Gagnon, 2010).

VILLEMONTE DE LA CLERGERIE E., SAGOT B., NICOLAS L. & GUÉNOT M.-L. (2009). FRMG : évolutions d'un analyseur syntaxique tag du français. In *Journée ATALA "Quels analyseurs syntaxiques pour le français ?"*.

ZELENKO D., AONE C. & RICHARDELLA A. (2002). Kernel methods for relation extraction. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, p. 71–78.

ZHANG M., SU J., WANG D., ZHOU G. & TAN C. L. (2005). Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. In *IJCNLP*, p. 378–389.