Reconnaissance automatique de concepts à partir d'une ontologie

Valentina Ceausu, Sylvie Desprès

Université René Descartes
CRIP5 – Equipe IAA – Groupe SBC
UFR Mathématiques et Informatique
45 rue des Saints-Pères
75006 PARIS
valentina.ceausu@math-info.univ-paris5.fr
sd@math-info.univ-paris5.fr

Résumé Ce papier présente une approche qui s'appuie sur une ontologie pour reconnaître automatiquement des concepts spécifiques à un domaine dans un corpus en langue naturelle. La solution proposée est non-supervisée et peut s'appliquer à tout domaine pour lequel une ontologie a été déjà construite. Un corpus du domaine est utilisé dans lequel les concepts seront reconnus. Dans une première phase, des connaissances sont extraites de ce corpus en faisant appel à des fouilles de textes. Une ontologie du domaine est utilisée pour étiqueter ces connaissance. Le papier donne un aperçu des techniques de fouilles employées et décrit le processus d 'étiquetage. Les résultats d'une première expérimentation dans le domaine de l'accidentologie sont aussi présentés.

1 Introduction

L'important volume de documents disponibles en langue naturelle et leur évolution rapide font émerger la nécessité de définir des approches permettant de retrouver rapidement des informations pertinentes dans ces documents.

Ce papier présente une approche qui utilise une ontologie de domaine pour identifier automatiquement des concepts du domaine dans un corpus en langue naturelle. Cette identification de concepts peut servir dans différents contextes : annotation des documents, indexation d'une collection de documents, etc. L'approche proposée est complètement automatique et non-supervisée, mise à part l'utilisation d'une ontologie de domaine. Etant donnés une ontologie *O* et un corpus *C*, le but est de retrouver dans *C* des termes *w* qui sont l'expression linguistique des concepts de l'ontologie *O*. On peut ainsi étiqueter les termes retrouvés dans le corpus par des concepts de l'ontologie. Cet étiquetage est réalisé en trois étapes : (1) une première étape emploie des techniques de fouille de textes pour identifier des termes du domaine dans le corpus; (2) pour chaque terme *w* retrouvé, le voisinage sémantique *V(w)* est identifié ; (3) en supposant que les relations dans le voisinage du terme *w* soient déjà dans l'ontologie, le positionnement des relations dans l'ontologie et des mesures statistiques sont utilisés pour étiqueter le terme *w*.

Le papier présente l'approche adoptée en répondant à un certain nombre de questions : Comment extraire des termes à partir du corpus ? Comment identifier le voisinage sémantique des termes extraits ? Ces questions sont traitées dans le paragraphe 2. Etant donnés le terme et son voisinage sémantique, quelles sont les stratégies d'étiquetage ? Une réponse est apportée dans le paragraphe 3. Le paragraphe 4 présente les résultats d'une première expérimentation en accidentologie; les perspectives à donner à ce travail sont discutées dans le paragraphe 5.

2 Extraction des termes et du voisinage sémantique

La fouille de textes est employée pour retrouver des termes du domaine qui représentent l'expression linguistique des concepts (Ville-Ometz et *al.*, 2004). La technique adoptée consiste à rechercher dans le corpus des associations de catégories lexicales susceptibles d'engendrer des regroupements de mots valides. Une telle association de catégories lexicales constitue un patron lexical, par exemple (***Verbe***, ***Préposition***, ***Nom***).

Un algorithme de reconnaissance (Ceausu et Desprès, 2005) identifie, dans le corpus annoté par l'analyseur syntaxique TreeTagger (Schmid, 1994), les instances des patrons prédéfinis : les patrons verbaux contenant un verbe et les patrons nominaux qui ne contiennent pas de verbe. Les instances des patrons nominaux permettent l'identification de termes du domaine, par exemple {***balise de priorité, priorité du passage***}. Les instances des patrons verbaux mettent en évidence des relations entre les termes, par exemple {***tourner sur droite***}.

Le voisinage sémantique d'un terme *w* est composé de relations comportant comme argument le terme *w*. La méthode adoptée pour déterminer ce voisinage exploite l'ensemble des regroupements engendrés par les patrons verbaux pour construire des classes de verbes. Une classe de verbes est constituée des regroupements de termes engendrés par le même verbe. Par exemple, la classe diriger est composée des regroupements : *diriger vers* ; *diriger vers* ***square*** *; diriger vers* ***esplanade*** ; ***véhicule*** *diriger vers* ; ***automobile*** *diriger vers* ***esplanade***. Les termes sujets des constructions verbales d'une même classe (***véhicule et automobile)*** et les compléments des regroupements verbaux du type ***« Verbe, Préposition »*** sont, en général, sémantiquement proches (square, esplanade). L'algorithme suivant permet d'assigner les termes aux concepts de l'ontologie en utilisant cette heuristique : (1) identification des classes de verbes dans l'ensemble d'instances des patrons verbaux ; (2) identification des arguments - sujet et complément - des constructions verbales du type « verbe, préposition » ; (3) utilisation de la représentations des relations dans l'ontologie pour étiqueter les termes. Un pré-traitement de regroupement des arguments des constructions verbales est utilisé pour réduire la variance linguistique entre les arguments dont le sens est voisin. Des mesures statistiques entre les chaînes de caractères, présentées *infra,* sont utilisées.

2.1 Mesures de similarité lexicale

Une mesure de similarité associe un nombre réel ***r*** à une paire de chaînes de caractères (***S1,S2***). Une valeur importante de ***r*** indique une similarité importante entre (***S1,S2***). Différentes approches permettent de calculer les similarités entre chaînes de caractères (Cohen et *al*., 2003). Les mesures de Jaccard, Jaro, Jaro-Winkler, Monge-Elkan ont été implémentées dans le cadre de ce travail.

La mesure de Jaccard estime la similarité entre deux chaînes ***S*** et ***T*** :

$$Jaccard(S,T) = |S \cap T| / |S \cup T| \quad (1)$$

Cette mesure est le rapport entre le nombre des sous chaînes communes à ***S*** et ***T*** et le nombre total de sous chaînes de ***T*** et de ***S***. Si les sous chaînes sont des caractères, la mesure de Jaccard correspond au nombre de caractères communs aux deux chaînes.

Les mesures de Jaro et Jaro-Winkler prennent simultanément en compte le nombre et la position des sous chaînes communes de ***S*** et ***T***.

Soient deux chaînes $s = a_1 \ldots a_k$ et $t = b_1 \ldots b_l$. Un caractère $a_i \in s$ **sera considéré commun** aux deux chaînes si, il existe un caractère $b_j \in t$ satisfaisant les conditions suivantes : $a_i = b_j$ et $i - H \leq j \leq i + H$, où $H = \frac{\min(|s|, |t|)}{2}$.

Soient $s^1 = a^1_1 \ldots a^1_k$ les caractères de ***s*** communs à ***t*** (même ordre que dans s) et $t^1 = b^1_1 \ldots b^1_l$ les caractères de ***t*** communs à ***s***. Une transposition entre s^1 et t^1 est définie par un indice i tel que $a^1_i \neq b^1_i$. T_{s^1,t^1} est la moitié du nombre de transpositions de s^1 à t^1. **La mesure de Jaro** exprime la similarité entre s et t selon la formule :

$$Jaro(s,t) = \frac{1}{3}\left(\frac{|s^1|}{|s|} + \frac{|t^1|}{|t|} + \frac{|s^1| - T_{s^1,t^1}}{|s^1|} \right) \quad (2)$$

La mesure de Jaro-Winkler (1999) est une extension de la mesure de Jaro qui utilise la taille P du plus long préfixe commun aux deux chaînes. En posant $P^1 = \max(P, 4)$, on écrit :

$$Jaro - Winkler(s,t) = Jaro(s,t) + \frac{P^1}{10}(1 - Jaro(s,t)) \quad (3)$$

Il existe aussi des approches hybrides qui calculent les similarités entre deux chaînes de manière récursive, en analysant des sous chaînes des chaînes initiales. Ainsi, la mesure de **Monge-Elkan** estime la similarité entre $s = a_1 \ldots a_k$ et $t = b_1 \ldots b_l$ selon :

$$sim(s,t) = 1/k \left(\sum_{i=1}^{k} \max_{j=1}^{L} \left(sim^1(a_i, b_j) \right) \right) \quad (4)$$

où les valeurs $sim^1(a_i, b_i)$ sont données par une fonction de similarité dite de base. Les mesures de Jaccard, Jaro et Jaro-Winkler ont été implémentées comme fonctions de base.

2.2 Pré-traitement des classes de verbes : regroupement des arguments

Le rôle de cette étape de prétraitement est d'identifier des similarités entre les arguments des relations du type « verbe, proposition » pour les regrouper. Pour une relation donnée, les arguments présentent différents niveaux de granularité, par exemple : ***partie*** - *partie gauche* -

partie droite ; ***rétroviseur*** *- rétroviseur extérieur - rétroviseur intérieur*. Un algorithme de regroupement des arguments, fondé sur l'utilisation de la plus grande spécificité des arguments composés de plusieurs mots sur les arguments mono mot, permet de construire des clusters de termes similaires. Un cluster est composé d'un terme central appelé centroïde ***c*** et ses ***k*** plus proches voisins.

L'algorithme construit une liste de *centroïdes* composée des arguments mono-mot et utilise la fonction Monge-Elkan pour ajouter des termes aux clusters. Cette fonction est utilisée car elle a la capacité d'agglomérer autour d'un mot les termes dérivés de ce mot.

3 Etiquetage des termes en utilisant l'ontologie

Dans ce paragraphe on présente l'étiquetage des termes extraits par des concepts de l'ontologie. On dispose d'une ontologie ***O***, contenant un ensemble ***C*** de concepts liés par des relations appartenant à un ensemble ***R*** et des classes de verbes contenant des constructions grammaticales de type : (sujet), (verbe, préposions), (complément objet). Le prétraitement des classes a regroupé les arguments en clusters homogènes.

L'hypothèse de travail est que les relations verbales appartiennent à ***R*** . Pour chaque relation verbale du type (verbe, préposition), la relation ***r*** qui lui correspond dans l'ontologie est retrouvée. Les concepts de l'ontologie liés par ***r*** sont identifiés et utilisés pour étiqueter les termes en adoptant une des stratégies d'étiquetage décrites ci-dessous.

3.1 Stratégies d'étiquetage

Les stratégies d'étiquetage définissent la manière dont les termes seront assignés aux concepts de l'ontologie. Les termes extraits sont déjà organisés en clusters, chaque cluster ayant un centroïde et des termes qui lui sont similaires (cf. 2.2.3).

Une première stratégie traite un cluster comme un ensemble non hiérarchisé de termes. Pour chaque terme du cluster ses similarités avec les concepts de l'ontologie sont évaluées en utilisant les mesures (1) à (3). Le terme est étiqueté par le concept qui maximise la valeur de cette similarité, si cette valeur dépasse un seuil imposé. Si toutes les valeurs des similarités se situent au-dessous du seuil, le terme sera étiqueté comme « inconnu ».

Les stratégies suivantes prennent en compte la structure hiérarchique de chaque cluster. Ainsi, la stratégie top-down identifie d'abord les concepts de l'ontologie qui vont étiqueter les centroïdes. Si le centroïde d'un cluster est étiqueté comme inconnu, la même étiquette est attribuée à chaque terme du cluster. Si le centroïde d'un cluster est étiqueté par un concept ***c***, les étiquettes pour les termes du cluster seront cherchées parmi les sous-concepts de ***c***. Cette stratégie a l'avantage de réduire l'espace de recherche.

Une troisième stratégie adopte une approche bottom-up. Pour chaque cluster, on évalue d'abord les similarités entre ses termes et les concepts de l'ontologie. Une des mesures (1) à (3) est utilisée et les termes sont étiquetés selon le principe de la première stratégie. Ensuite, la similarité du centroïde avec un concept de l'ontologie est donnée par :

$$sim(centroid, c) = (1/k) \sum_{i=1}^{k} sim(t_i, c) \qquad (5)$$

où $sim(t_i, c)$ exprime la similarité entre t_i et c, le concept qui étiquette k termes.

4 Expérimentations en accidentologie et premiers résultats

Le corpus utilisé pour cette expérimentation est composé de 250 procès verbaux d'accidents. L'ontologie du domaine est éditée en Terminae (Biébow et Szulman, 2004) et exprimée en OWL (Szulman et Biébow 2004).

Les termes extraits de ce corpus sont étiquetés et les résultats obtenus seront analysés selon deux points de vue : pour le même coefficient de similarité, comparer les résultats de chaque stratégie d'étiquetage ; pour une même stratégie d'étiquetage, comparer les résultats de chaque coefficient. Les arguments objets du verbe «circuler avec » sont étiquetés.

La stratégie bottom-up permet d'éliminer le centroïde « feu », qui est étiqueté comme « inconnu ». Elle pénalise les centroïdes ayant engendrés des clusters de taille réduite, qui sont assignés aux concepts de l'ontologie avec un faible coefficient, ou sont étiquetés «inconnu». Les résultats des trois stratégies sont similaires pour les coefficients Jaro et Jaro-Winkler, Cette similarité est normale car Jaro-Winkler représente juste une variation de Jaro. Dans le cas du coefficient Jaccard, la stratégie bottom-up montre une défaillance, en assignant le terme « véhicule » au concept «véhicule de service».

Quel que soit le coefficient choisi, l'étiquetage top down est plus rapide et donne de meilleurs résultats.

5 Conclusion et perspectives

Nous avons présenté une approche permettant d'assigner des termes extraits d'un corpus en langue naturelle aux concepts d'une ontologie. Des métriques pour calculer la similarité entre chaînes de caractères ont été implémentées et interviennent dans différentes étapes de l'approche. Une première expérimentation dans le domaine de l'accidentologie montre que les coefficients de Jaro et Jaro-Winkler donnent des estimations de similarités plus fines que Jaccard. Parmi les stratégies d'étiquetage, l'étiquetage top-down est plus rapide et engendre de meilleures assignations des termes aux concepts de l'ontologie.

En perspective, des ressources terminologiques telles que WordNet peuvent être prises en compte afin d'améliorer l'estimation des similarités entre les termes du corpus et les concepts de l'ontologie. Cela permettra d'enrichir le voisinage sémantique du terme par d'autres types de relations, comme la synonymie. Une autre perspective peut être l'ajout d'un feed-back dans le processus décrit, permettant à l'utilisateur non seulement d'étiqueter les termes du domaine, mais aussi d'intégrer dans l'ontologie certains des termes découverts.

Références

Alfonseca, E. and S. Manandhar. (2001). Improving an ontology refinement method with hyponymy patterns. *Proceedings of Third International Conference on Language Resources and Evaluation, 235–239.*

Biébow B. et S. Szulman (1999) TERMINAE: A linguistic-based tool for the building of a domain ontology. 11th European Workshop, Knowledge Acquisition, Modeling and Management 49-66.

Ceausu V. et S. Desprès (2005), Fouilles de textes pour orienter la construction d'une ressource terminologique. *Actes de la conférence Extraction et Gestion des Connaissances , Cépaduès éditions*,

Cohen, W., P. Ravikumar, S. Fienberg (2003) . A Comparison of String Distance Metrics for Name-Matching Tasks. *In Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03 , 73–78.*

Faatz A., and R. Steinmetz (2002). Ontology enrichment with texts from the WWW. *Semantic Web Mining 2nd Workshop at ECML/PKDD.*

Gagliardi, H.,O. Haemmerlé, N. Pernelle, F. Saïs (2005). An automatic ontology-based approach to enrich tables semantically , *The first International Workshop on Context and Ontologies : Theory, Practice and Applications.*

Monge, A., and C. Elkan (1996). The field-matching problem: algorithm and applications. In *Second International Conference on Knowledge Discovery and Data Mining,* .

Parekh, V.,P. Jack , and T. Finin. (2004). Mining domain specific texts and glossries to evaluate and enrich domain ontologies. In *Proceedings of the 2004 InternationalConference on Information and Knowledge Engineering.*

Roux, C., D. Proux, F.Rechenmann, and L. Julliard (2000). An Ontology Enrichment Method for a Pragmatic Information Extraction System gathering Data on Genetic Interactions, *Proceedings of the ECAI'2000 Ontology Learning Workshop*, S. Staab *et al.*

Schmid. H. (1994). Probabilistic part-of-speech tagging using decision trees. *In International Conference on New Methods in Language Processing*

Szulman, S. et B.Biebow (2004). "OWL et Terminae", *Actes d' IC'2004*, Lyon, 41-52.

Valarakos, A., G. Paliouras, V. Karkaletsis, and G. Vouros (2004) A Name Matching Algorithm for Supporting Ontology Enrichment. *3rd Hellenic Conference on Artificial Intelligence (SETN04), LNAI, Vol. 3025, 381-589.*

Ville-Ometz, F., J. Royauté, A. Zasadzinski (2004) Filtrage semi-automatique des variantes de termes dans un processus d'indexation contrôlée, *Colloque International sur la Fouille de extes.*

Xiaomeng., S (2004), *Semantic Enrichment for Ontology Mapping*, Ph.D Thesis, Norwegian University of Science and Technology.

Warin, M., H. Oxhammer and M.Volk (2005). Enriching an Ontology with WordNet based on Similarity Measures. *Proceedings of the MEANING Workshop.*

Widdows, D. (2003), Unsupervised methods for developing taxonomies by combining syntactic and statistical information. *Proceedings of HTL-NAACL, 197-204.*

Summary

This paper presents an ontology supported approach to automatically recognize concepts of a specific field in a natural language corpora. This in a non-supervised solution that can be applied to any field for which an ontology was already created. A natural language corpora of the field is used in which specific concepts are recognized. In a first phase of the process, terms are extracted from the corpora using text mining techniques. Then, a domain ontology is used to label these terms. A label is assign to each term according to his semantic neighborhood and statistic measures. This paper gives a brief overview of employed text mining techniques and then it focus on the labeling process. A first experimentation of our approach in the field of accidentology was done and his results are also presented.