Détection de relations sémantiques à partir de texte

Guillaume Tisserant, Violaine Prince, Mathieu Roche

Détection de relations sémantiques à partir de texte

Guillaume Tisserant*, Violaine Prince*, Mathieu Roche*

* LIRMM, CNRS UMR 5506, Université Montpellier 2

Résumé : Cet article présente une approche de prédiction de liens sémantiques à partir de textes en appliquant des méthodes d'apprentissage supervisé. Ces travaux se focalisent sur l'utilisation de nouveaux descripteurs linguistiques afin d'améliorer la prédiction.

1 Introduction

Une ontologie est une conceptualisation d'un domaine partagé par une communauté d'acteurs, c'est-à-dire un ensemble de concepts et de relations définis à l'aide d'un langage formel compréhensible par un ordinateur. Les ontologies sont souvent représentées sous forme de graphe où chaque sommet représente un concept appartenant au domaine modélisé par l'ontologie, et chaque arrête une relation entre deux concepts.

La construction d'ontologies est un domaine de recherche très largement étudié. Certaines constructions sont entièrement manuelles, d'autres semi-automatiques, et d'autres encore complètement indépendantes d'interventions humaines. Le travail présenté se situe dans cette dernière catégorie, et se fonde sur l'extraction de relations sémantiques dans des corpus textuels portant sur des domaines de spécialité.

2 L'approche

La construction d'ontologie est un domaine de recherche complexe où deux branches principales cohabitent. Une partie considère que l'information nécessaire à la création d'ontologies sur des domaines spécifiques se trouve dans les textes appartenant au domaine. Le processus est souvent fondé sur l'apprentissage et la reconnaissance de schémas lexicaux et/ou syntaxiques marquant les relations (Morin, 1999) avec l'utilisation de connaissances expertes (Bourigault et Aussenac Gilles, 2003 ; Jacques et Aussenac Gilles, 2006). Une seconde partie s'appuie directement sur des connaissances issues d'experts du domaine (Roche, 2006). Ces types de méthodes sont plus précis mais les constructions nécessitent une ressource en main d'œuvre experte plus importante. La méthode proposée dans cet article repose sur des méthodes de fouille de texte tout en cherchant à minimiser le besoin d'experts.

La construction d'ontologie à partir d'un corpus se fonde souvent sur les trois phases suivantes. Il est d'abord nécessaire de détecter les termes appartenant au domaine contenus dans le corpus, trouver ceux qui sont en relations, puis identifier les types de relations

(hyperonymie[1], synonymie, méronymie, etc.). Nous nous intéressons ici à la troisième étape, c'est-à-dire l'identification des relations entre les termes à partir du texte.

Les méthodes proposées s'appuient sur des techniques d'apprentissage automatique. La contribution ne se situe pas sur l'amélioration des algorithmes d'apprentissage en eux-mêmes mais sur l'identification fine des descripteurs fournis en entrée. Les descripteurs peuvent être des mots, des mots couplés avec leur positionnement dans la phrase, le nombre de mots présents entre les deux termes en relation, etc. Comme nous nous situons dans le cadre de l'apprentissage supervisé, nous devons utiliser des exemples étiquetés. Dans notre cas, ces derniers sont des couples de termes dont la relation sémantique est connue.

L'algorithme de classification permet ainsi d'apprendre un modèle permettant d'indiquer une relation sémantique probablement liée à un groupe de descripteurs donnés en entrée. La contribution portant sur les descripteurs, nous avons choisi de ne pas réimplanter les algorithmes de classification, mais d'utiliser ceux fournis par Weka (http ://www.cs.waikato.ac.nz/ml/weka/).

3 Extraction des descripteurs linguistiques

Cette section décrit le processus d'identification des descripteurs à associer aux couples *Xi*, *Yi* en relation sémantique *Rj*. Afin d'apprendre un modèle, il est d'abord nécessaire de constituer un ensemble d'entrainement conséquent (exemples étiquetés). Dans ce cadre, notre système récupère dans un corpus donné toutes les phrases contenant les termes *Xi*, *Yi*. Par exemple, dans le domaine biomédical, le couple *X1* : leukocytes et *Y1* : basophils est en relation d'hyperonymie (cf. Thésaurus MeSH). Nous pouvons donc retrouver des phrases contenant ces deux mots dans un corpus donné (par exemple MEDLINE).

L'objectif de notre travail consiste à apprendre un modèle permettant de prédire le type de relation *Rj* à partir des descripteurs linguistiques situés entre *Xi* et *Yi* issus de notre corpus. Dans la suite, nous allons proposer et discuter l'utilisation de nouveaux descripteurs linguistiques.

Généralisation des descripteurs linguistiques (Approche GenDesc).

Certains mots qui apparaissent peu comme descripteurs ne sont pas toujours bien pris en compte par les systèmes d'apprentissage automatique. Nous proposons donc de les généraliser en utilisant leur fonction grammaticale (Nom, Adjectif, Adverbe, etc) comme descripteur plutôt que le mot en lui même. Ces informations grammaticales sont obtenues via un étiqueteur (TreeTagger ou Brill). Au contraire, des mots déjà très généraux (très présents dans le corpus) ne doivent pas être davantage généralisés. Ils apportent des informations sémantiques tout à fait cruciales. Par exemple la présence du mot "is" peut dénoter un type de relation particulier. Ainsi, avec l'approche *GenDesc*, un tel mot n'est pas généralisé contrairement à d'autres descripteurs. Le choix du niveau de généralisation s'appuie sur des méthodes statistiques. La méthode de base consiste à ne généraliser que les mots ayant un nombre d'occurrences inférieur à *n*. Ce seuil *n* sera discuté dans la section 4 de cet article.

Positionnement des descripteurs linguistiques (Approche PosDesc).

[1] L'hyperonymie est la relation sémantique hiérarchique d'un lexème à un autre selon laquelle l'extension du premier terme, plus général, englobe l'extension du second, plus spécifique.

Outre la présence ou non des descripteurs linguistiques, il semble pertinent d'ajouter les informations liées à leur distance par rapport au début de la séquence (c.-à-d. *Xi*). Ceci peut caractériser un lien sémantique fort typique de la synonymie. Le descripteur linguistique associé à cet indice de positionnement par rapport à *Xi* constituera un nouveau descripteur appelé *pos1*. Cependant, la distance par rapport à *Xi* ne fournit qu'une information partielle ; elle ne permet pas de déduire la distance relativement à *Yi*. Nous proposons donc de rajouter un nouveau descripteur appelé *pos2* lié à un tel positionnement. La pertinence de ces nouveaux descripteurs sera discutée dans la section suivante.

4 Expérimentations

Nos expérimentations ont été menées à partir d'un corpus issu de la base bibliographique MEDLINE constitué à partir de relations de synonymie et d'hyperonymie extraites de l'ontologie MeSH. 492 couples d'hyperonymes et 470 couples de synonymes ont été constitués. Les phrases contenant ces couples constituent les exemples de nos algorithmes d'apprentissage. Nous utilisons *Naïve Bayes* et les *Arbres de Décision* dans leur version implantée dans Weka.

L'analyse préliminaire du corpus montre que les mots utilisés rarement se révèlent inappropriés pour les algorithmes d'apprentissage. La recherche d'un seuil par rapport au nombre d'apparition d'un mot dans le corpus nous a montré qu'il existait une faible proportion de mots apparaissant fréquemment. Le tableau ci-dessous présente la répartition des descripteurs linguistiques selon le nombre d'occurrences à partir de notre corpus.

Nb d'occurrences ≤ 10	873
10 < Nb d'occurrences < 50	110
Nb d'occurrences ≥ 50	20

Dans nos expérimentations, nous avons dans un premier temps évalué la pertinence de l'approche *GenDesc* (cf. section 3). Les résultats selon les différents critères sont donnés dans le tableau ci-dessous. Bien que les écarts soient relativement faibles, nous pouvons noter que la généralisation partielle des mots (c'est-à-dire lorsque *n=7*) donne de meilleurs résultats que la généralisation totale ou l'absence de généralisation.

Pas de généralisation	86.6
Généralisation grammaticale partielle (avec *n=7*)	87.3
Généralisation totale	85.3

Pourcentage d'instances correctement classées en fonction de la généralisation effectuée en utilisant le classifieur Naïve Bayes (par validation croisée).

Dans un second temps, nous avons évalué l'approche *PosDesc*. La place d'un marqueur par rapport aux deux mots peut multiplier le nombre de marqueurs ce qui peut dégrader la performance des algorithmes d'apprentissage. Des expérimentations préliminaires ont tout de même montré que l'utilisation de *pos1 et pos2* permet une amélioration absolue du taux de classification de 2% et 3% respectivement.

5 Conclusions et perspectives

Dans cet article, nous avons proposé de nouveaux descripteurs textuels qui permettent d'améliorer la prédiction des types de relations sémantiques par une méthode d'apprentissage supervisé. Cependant, nous pouvons expliquer le faible gain obtenu par le fait que les marqueurs les plus fréquents sont des ponctuations (principalement des parenthèses, qui indiquent presque toujours une relation de synonymie) et qui ne sont donc jamais généralisés.

Des expérimentations sont en cours avec d'autres types de données (en français) et en prenant en compte davantage de relations sémantiques.

Dans nos futurs travaux, nous souhaitons privilégier les descripteurs considérés comme des marqueurs de gloses. Les gloses sont des commentaires en situation parenthétique, souvent introduits par des marqueurs tels que *appelé*, *c'est-à-dire*, *ou* qui signent la relation de sémantique lexicale mise en jeu : équivalence avec *c'est-à-dire*, *ou* ; spécification du sens avec *au sens* ; nomination avec *dit*, *appelé* ; hyponymie avec *en particulier*, *comme* ; hyperonymie avec *et/ou autre(s)*, etc. En effet, ce type de marqueur linguistique permet de mettre en relief de manière explicite des liens sémantiques entre les termes (Mela *et al.*, 2012).

Références

Bourigault D., Aussenac Gilles N. (2003). Construction d'ontologies à partir de textes. Proceedings of TALN 2003, 27-47.

Brill E. (1992). A simple rule-based part of speech tagger. Proceedings of the workshop on Speech and Natural Language - HLT

Jacques M., Aussenac Gilles N. (2006). Variabilité des performances des outils de TAL et genre textuel. TAL, 47(1), 11-32

Mela A., Roche M., Bekhtaoui M. (2012). Lexical Knowledge Acquisition Using Spontaneous Descriptions in Texts. . Proceedings of NLDB 2012, LNCS, 366-371

Morin E. (1999). Extraction de liens sémantiques entre termes à partir de corpus de textes techniques. Thèse de Doctorat.

Roche C. (2006). Dire n'est pas concevoir. Proceedings of the 18es Journées Francophones d'Ingénierie des Connaissances, Grenoble

Summary

This paper presents an approach for predicting semantic relationships from texts by applying supervised machine learning. This work focuses on the use of new features in order to improve prediction.