Mise en lumière de relations sémantiques pour la construction d'ontologies à partir de textes

Rim Bentebibel, Adeline Nazarenko, Sylvie Szulman

Laboratoire d'Informatique de l'université Paris-Nord (LIPN)
UMR 7030 Université Paris 13 & CNRS
99, avenue Jean-Baptiste Clément
93430 Villetaneuse
prénom.nom@lipn.univ-paris13.fr

Résumé :
La construction d'ontologies à partir de textes consiste à décrire des concepts par leurs relations conceptuelles et éventuellement leurs instances, à partir des matériaux textuels (termes, relations lexicales). Cet article propose une méthode pour mettre en lumière, par l'analyse de corpus, des relations lexicales susceptibles de donner naissance à des relations conceptuelles. Cette méthode ne fait aucune hypothèse sur les types de relations recherchées ni sur leur forme linguistique. Notre approche s'inspire des méthodes distributionnelles de construction de classes sémantiques mais il s'agit de construire des classes d'associations de termes et non des classes de termes. Les classes d'associations obtenues représentent des relations sémantiques candidates qui peuvent ensuite être élaborées en relations conceptuelles par l'ontologue.

Mots-clés : Extraction de relation, Les relations dans l'ontologie, Ingénierie des connaissances, TAL, analyse distributionnelle.

1 Introduction

Les méthodes de construction d'ontologies à partir de textes sont aujourd'hui bien connues : pour identifier les concepts du domaine, elles s'appuient sur l'analyse terminologique pour les unes, sur l'analyse distributionnelle et les classes de mots pour les autres.

Au-delà des concepts et de leurs instances, il est également important de repérer les relations conceptuelles qui structurent le domaine. Cette problématique est reconnue comme difficile. Rastier (2004) distingue « les liens verticaux [qui] sont des liens de catégorisation » et « les liens horizontaux [qui] sont des liens d'actance ». Des approches distributionnelles ont été proposées, mais pour l'explicitation des liens verticaux uniquement. Les approches classiques, héritées de

la terminologie traditionnelle, permettent d'extraire des liens horizontaux comme verticaux. Elles explorent les textes à l'aide de patrons mais ces patrons ou schémas de phrases sont différents pour chaque relation et ils varient souvent d'un corpus à l'autre. Nous proposons ici une méthode générique de découverte de relations sémantiques à partir de textes. Il s'agit d'explorer les textes pour identifier les relations conceptuelles qu'ils véhiculent sans idée préconçue sur le type de relations qu'on recherche. Bien entendu, un travail manuel de validation et de conceptualisation des relations candidates proposées est nécessaire. L'originalité de la méthode proposée est double : elle est guidée par le corpus lui-même (les données) plutôt que par les relations à acquérir (le but) et elle est générique par rapport au corpus et à la tâche. Il s'agit, comme pour l'extraction terminologique, de faire « émerger » la sémantique du domaine du corpus, même si les résultats ont besoin d'être retravaillés.

La section 2 situe notre travail par rapport à l'état de l'art en acquisition de relations et souligne l'intérêt d'une approche guidée par le corpus par rapport aux approches guidées par les relations à construire. La section 3 décrit les étapes de notre méthode. La section 4 présente une discussion et conclusion de ce travail.

2 Etat de l'art

Beaucoup de travaux sur l'acquisition d'ontologies à partir de textes s'appuient sur des patrons d'extraction pour retrouver des relations dans les textes (Auger & Barriere, 2008). Cette approche a évidemment son intérêt pour la construction d'ontologies et un module d'extraction à base de patrons est prévu dans la plateforme Dafoe. Cependant, (Jacques & Aussenac-Gilles, 2006) ont montré que ces patrons varient fortement d'une relation à l'autre mais aussi d'un corpus à l'autre pour une même relation. Il existe peu de patrons génériques et le coût de mise au point des patrons d'extraction nécessaires pour la construction d'une ontologie donnée est vite apparu prohibitif. On a donc cherché à tirer profit des recherches menées en extraction d'information sur l'apprentissage de patrons, notamment les approches semi-supervisées (Morin & Martienne, 1999; Agichtein & Gravano, 2000; Blohm *et al.*, 2007; Turney, 2006). Une autre approche est proposée par (Hasegawa *et al.*, 2004) pour enrichir des systèmes de question/réponse ou de résumé de textes : elle vise à repérer des relations sémantique entre entités nommées. La méthode consiste à faire émerger des classes homogènes de couples d'entités nommées, chaque classe étant alors considérée comme représentante d'une relation intéressante pour le domaine. En dépit de son intérêt, cette approche est limitée par le fait qu'elle se fonde exclusivement sur les entités nommées qu'elles présupposent étiquetées sémantiquement. Nous nous en inspirons mais en cherchant à en généraliser l'application à d'autres types d'unités textuelles, notamment aux termes qui servent eux aussi d'ancres pour la découverte de relations. Nous voulons aussi étendre la méthode pour permettre d'associer des patrons d'extraction, ou tout au moins des ébauches de patrons, aux relations sémantiques candidates. En réalité, la méthode que nous présentons s'apparente aux méthodes distributionnelles de construction de

classes sémantiques de mots (voir par ex. (Faure & Nédellec, 1999)) qui rapprochent les mots sur la base des éléments des contextes qu'ils partagent et qui proposent les classes obtenues comme ébauches de concepts. Sauf qu'il s'agit ici de construire des classes d'associations de termes et non pas des classes de termes.

3 Méthode

Notre méthode d'extraction de relations est constituée de quatre processus comme illustré sur la figure 1 : construction d'une représentation normalisée des textes, extraction des associations d'unités sémantiques les plus pertinentes ; regroupement des différentes occurrences d'association en classe de relations ; construction des ébauches de patrons.

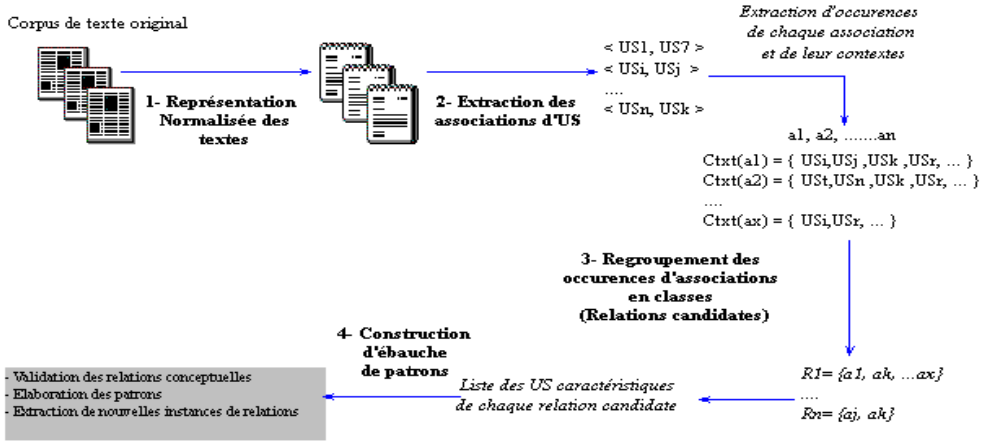

FIG. 1 – Schéma général de la méthode d'extraction de relations

3.1 Construction d'une représentation normalisée des textes

La première étape consiste à construire une représentation simplifiée et normalisée du corpus d'acquisition. Comme le but est de construire une ontologie, nous ne nous intéressons pas aux mots du texte mais aux *unités sémantiques* (US). En pratique, les US sont les termes qui relèvent du vocabulaire du domaine et qui sont souvent des unités lexicales composées[1]. Nous considérons aussi les entités nommées comme des éléments sémantiquement pertinents. Comme l'extracteur de termes que nous utilisons[2] ne reconnaît pas les termes verbaux, nous conser-

[1]Nous privilégions les termes les plus longs

[2]YaTeA (Aubin & Hamon, 2006).

vons aussi les mots « sémantiquement pleins » après élimination des mots grammaticaux et des mots athématiques qui figurent dans un antidictionnaire (ex. *sorte, faire, être*, etc.). Par souci de normalisation, nous considérons les formes lemmatisées des unités sémantiques[3]. Ce choix de représentation efface toute information syntaxique (une fois les unités sémantiques identifiées, seul l'ordre des unités est conservé) mais cela permet de simplifier les phrases et donc de faciliter leur rapprochement. Cette normalisation est importante puisque notre approche repose sur la récurrence des unités et de leurs associations dans le corpus, à la différence des méthodes à base de patrons qui s'appuient davantage sur la structure des phrases.

Ainsi donc, le corpus est représenté à l'issue de cette étape de normalisation par une séquence de documents qui sont eux-mêmes représentés récursivement comme des séquences de phrases puis d'unités sémantiques.

3.2 Extraction des associations d'unités sémantiques

Notre méthode repose sur le repérage dans le corpus de couples d'unités sémantiques fortement associées, l'idée étant que ces associations sont potentiellement des indices de relations sémantiques du domaine. Pour extraire les associations d'unités sémantiques, nous nous appuyons sur un calcul de cooccurrence : plus les unités apparaissent ensemble (dans les mêmes phrases), plus elles sont considérées comme fortement associées. La force de cette association est donnée par la mesure de l'information mutuelle (formule 1) qui est comprise entre 0 et 1 et qui est calculée pour tous les couples d'unités sémantiques présentes dans le corpus.

$$IM(US_i, US_j) = log_2 \frac{P(US_i US_j)}{P(US_i)P(US_j)} \quad (1)$$

Nous conservons au final comme « associations » tous les couples d'unités sémantiques dont la valeur IM est supérieure à un seuil donné[4].

3.3 Regroupement des occurrences d'association en classes

On applique ensuite un processus de classification ascendante (ou regroupement, *clustering* en anglais) sur l'ensemble des occurrences a_k des associations $< US_i, US_j >$ qui ont été identifiées à l'étape précédente[5]. Ce processus consiste à regrouper en classes les occurrences les plus similaires. Pour calculer la similarité entre les occurrences d'association, on représente chaque occurrence a_k par son contexte $ctxt(a_k)$, c'est-à-dire par l'ensemble des unités sémantiques qui figurent dans la même phrase que a_k. Considérons par exemple l'association

[3] Ces formes lemmatisées sont obtenues à partir des résultats de l'application préalable du TreeTager (Schmid, 1994) sur notre corpus. Pour la reconnaissance des entités nommées, nous utilisons l'extracteur de Gate.

[4] Pour la mise au point de la méthode, ce seuil a volontairement été fixé très bas.

[5] *A priori* si deux unités sémantiques sont considérées comme associées, on trouve en effet plusieurs occurrences de cette association dans le corpus.

$< US_1, US_{12} >$ et l'une de ses occurrences a_1. Si on a $ctxt(a_1) = \{US_{20}, US_{33}, US_{45}\}$, cela signifie que ces trois unités sémantiques apparaissent aux côtés de US_1 et US_{12} dans la phrase où cette occurrence particulière a_1 de l'association apparaît. Plus formellement, $ctxt(a_k)$ est un vecteur dans l'espace du vocabulaire des unités sémantiques du corpus. La coordonnée de $ctxt(a_k)$ sur l'axe de l'unité US_i est 1 si US_i figure dans le contexte de a_k et 0 sinon.

Pour déterminer à quel point deux contextes $ctxt(a_x)$ et $ctxt(a_y)$ sont similaires, nous considérons le cosinus de l'angle $A(X, Y)$ formé par leurs vecteurs X et Y, en calculant le cosinus comme le produit scalaire des vecteurs normalisé par leurs longueurs (formule 2). Le résultat est compris entre 0 et 1 : si la valeur du cosinus est 0, cela signifie que les deux vecteurs sont perpendiculaires, donc que les contextes ne sont pas similaires ; si la valeur de cosinus est égale à 1, les deux vecteurs pointent dans la même direction et les contextes sont jugés similaires.

$$\cos(A(X, Y)) = \frac{X.Y}{|X||Y|} \tag{2}$$

L'intuition sous-jacente est que ces occurrences d'association qui ont été jugées similaires et regroupées partagent un même « sens », c'est-à-dire une même relation sémantique. Comme toutes les méthodes distributionnelles, le risque d'obtenir des classes bruitées est réel : il faut analyser les classes obtenues, en supprimer ou en redécouper certaines, y supprimer des intrus et, au final, nommer la relation sous-jacente si la classe apparaît suffisamment cohérente.

3.4 Construction d'ébauches de patrons

Pour faciliter l'analyse des classes obtenues, il est utile de comprendre les éléments qui ont permis le rapprochement des occurrences d'association qui la composent et, une fois la relation sémantique identifiée et sélectionnée, il est précieux de pouvoir lui associer un patron d'extraction pour en repérer de nouvelles occurrences.

Nous cherchons donc à identifier les éléments de contexte qui caractérisent le mieux les classes obtenues et à s'en servir pour construire des ébauches de patrons. Ces éléments caractéristiques sont en réalité donnés par le processus de regroupement précédent : ce sont les unités sémantiques que les différents contextes des occurrences d'association regroupées partagent et les unités sémantiques qui entrent dans l'assocation.

Il suffit alors de mettre ces unités sémantiques caractéristiques dans un ordre textuel plausible (dans les vecteurs de contexte, les unités sémantiques sont en effet ordonnées dans un ordre unique et arbitraire). On regarde, pour ce faire, comment les unités sémantiques caractéristiques sont ordonnées dans les phrases correspondant aux différentes occurrences d'association de la classe et on prend l'ordre majoritaire s'il existe, n'importe quel ordre attesté à défaut.

Sur cette base, nous construisons des ébauches de patrons en distinguant le rôle des unités sémantiques associées et des unités sémantiques contextuelles.

4 Discussion et conclusion

Pour l'instant, l'approche proposée n'a été testée que sur un corpus d'une centaine de phrases pour épouvrer les différentes étapes de notre méthode et analyser son comportement en détail.

Nous proposons dans cet article une méthode générique pour mettre en lumière les relations sémantiques d'un domaine à partir de textes de ce domaine. En tant que telle, cette méthode est indépendante du domaine et de la langue. Elle doit à terme s'intégrer dans la plateforme Dafoe de construction d'ontologies à partir de textes. Il s'agit d'extraire tout type de relation sans connaitre *a priori* les relations à extraire. La méthode est fondée sur une représentation normalisée des textes comme séquences d'unités sémantiques du domaine (essentiellement des termes et des entités nommées) et l'idée maîtresse consiste à construire des classes d'associations d'unités sémantiques de manière distributionnelle. L'hypothèse sous-jacente est que les occurrences d'association réunies sur la base des éléments de contexte qu'elles partagent ont des chances de relever d'une même relation sémantique et que les relations candidates ainsi proposées peuvent aider le travail de conceptualisation de l'ontologie.

Références

Agichtein E. & Gravano L. (2000). Snowball : Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*.

Aubin S. & Hamon T. (2006). Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, p. 380–387 : 5th International Conference on NLP.

Auger A. & Barriere C. (2008). Pattern-based approaches to semantic relation extraction : A state-of-the-art. *Terminology*, **14**(1), 1–19.

Blohm S., Cimiano P. & Stemle E. (2007). Harvesting relations from the web - quantifiying the impact of filtering functions. In *AAAI*, p. 1316–1321.

Faure D. & Nédellec C. (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning : the system asium. In *Proceedings of the 11th International Conference on Knowledge Engineering and Knowledge Management*, p. 329–334.

Hasegawa T., Sekine S. & Grishman R. (2004). Discovering Relations among Named Entities from Large Corpora. *Proc. of ACL-2004*, p. 415–422.

Jacques M.-P. & Aussenac-Gilles N. (2006). Variabilité des performances des outils de tal et genre textuel. cas des patrons lexico-syntaxiques. *Traitement Automatique des Langues (TAL)*, **47**(1), 11–32.

Morin E. & Martienne E. (1999). Raffinement de patrons lexico-syntaxiques par un système d'apprentissage. In *Actes de ic-99*, Palaiseau, France.

Rastier F. (2004). Ontologie(s). *Article paru dans la revue des sciences et technologies de l'information, série : Revue d'Intelligence artificielle*, (vol. 18, num1), 15–40.

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

Turney P. D. (2006). Expressing implicit semantic relations without supervision. In *proceedings of Acl-44*, p. 313–320, Morristown, NJ, USA.