Modélisation Prétopologique pour la Structuration Sémantico-Lexicale

Guillaume Cleuziou*, Gaël Dias**, Vincent Levorato*

*Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)
Université d'Orléans, Orléans, France
**Centre of Human Language Technology and Bioinformatics (HULTIG)
University of Beira Interior, Covilhã, Portugal

Résumé. Cet article présente une nouvelle méthode pour la construction d'une structure sémantico-lexicale à partir de l'observation des co-occurrences de termes en corpus. Nous utilisons le formalisme de la prétopologie pour modéliser à la fois le niveau de généralité/spécificité des termes et leurs proximités sémantiques à partir d'une matrice de similarité asymétrique. Nous montrons qu'un ordre partiel peut être défini sur le lexique et proposons un algorithme de construction d'une structure hiérarchique.

1 Introduction et motivations

Les structures sémantico-lexicales jouent un rôle essentiel dans la Recherche d'Information (RI) et le Traitement Automatique du Langage (TAL). En codant les relations sémantiques entre concepts du discours, elles permettent d'enrichir les capacités de raisonnement des applications de la RI et du TAL. Cependant, leur développement est largement limité par les efforts nécessaires pour leur construction. En effet, la plupart des structures existantes sont construites manuellement et principalement pour la langue anglaise. De nombreux travaux ont donc émergé, depuis une vingtaine d'années, pour apprendre (semi-)automatiquement des structures sémantico-lexicales, encore appelées ontologies dans leur acception la plus forte. Dans ce cadre, la plupart des travaux présentés ont privilégié des approches dépendantes du langage en prônant l'utilisation de ressources linguistiques spécifiques aux domaines et aux langues étudiées. Dans cet article, nous proposons une nouvelle approche qui évalue le degré de généralité/spécificité des termes ainsi que leurs proximités sémantiques à partir de mesures de similarité asymétriques appliquées à leurs occurrences en corpus. Ainsi, à partir d'une matrice de proximités rendant compte du degré d'attraction d'un terme vers un autre, nous utilisons le formalisme de la prétopologie pour obtenir, en une seule et unique étape, une structure hiérarchique sous forme d'une famille de graphes orientés acycliques non triangulaires correspondant aux sous-domaines du lexique indépendamment de la langue ou du domaine.

2 Formalisation et structuration prétopologiques

La prétopologie est un outil de modélisation du concept de proximité. Celle-ci permet des opérations sur les ensembles telles que l'*adhérence* ou la *fermeture* (Belmandt, 1993), utilisant

une axiomatique plus faible que la topologie. De manière générale, soit E un ensemble fini, une application $a(.)$ de $\mathcal{P}(E)$ dans $\mathcal{P}(E)$ est appelée *adhérence* ssi $\forall A \in \mathcal{P}(E) : a(\emptyset) = \emptyset$ et $A \subseteq a(A)$. Le couple (E, a) est appelé espace prétopologique. Les voisinages d'un élément x de E sont définis par la famille $\mathcal{V}(x)$ avec $\forall\, x \in E, \forall\, V \in \mathcal{V}(x)$ et $V \subset E$, x appartient à V. Dans notre étude, nous définissons l'*adhérence* (1) basée sur le voisinage $V(x)$ d'un élément x de E tel que $V(x) = \{y \in E | xRy\}$ où R est une relation binaire non-symétrique réflexive entre les éléments.

$$\forall A \in \mathcal{P}(E), a(A) = \{x \in E \mid A \cap V(x) \neq \emptyset\} \tag{1}$$

L'application successive de l'*adhérence* sur un ensemble $A \in \mathcal{P}(E)$ génère un processus de dilatation. Lorsque ce processus atteint un point fixe ($a^{k+1}(A) = a^k(A)$) l'ensemble généré est appelé *fermeture* de A, noté F_A défini dans (2) où $a^k()$ désigne l'*adhérence* k^{ieme} (e.g. $a^2(A) = a(a(A))$).

$$\forall A \in \mathcal{P}(E), \exists\, k \in \mathbb{N} \text{ tel que } F_A = a^k(A) \tag{2}$$

Dans ce cadre, nous introduisons la notion de *fermé élémentaire* F_x pour désigner la *fermeture* d'un singleton $\{x\}$ ainsi que $\mathcal{F}_e(E, a)$ l'ensemble des *fermés élémentaires* sur l'espace prétopologique (E,a) tel que $\mathcal{F}_e(E, a) = \{F_x \mid x \in E\}$. Enfin, on parlera de *fermés élémentaires maximaux* $\mathcal{F}_M(E, a)$ pour désigner l'ensemble des *fermés élémentaires* qui ne sont inclus dans aucun autre *fermé élémentaire*, c'est-à-dire $F_x \in \mathcal{F}_M(E, a) \Leftrightarrow \forall\, F_{x'} \in \mathcal{F}_e(E, a),\ F_x \not\subset F_{x'}$.

Pour la construction d'une structure sémantico-lexicale, E désigne l'ensemble des termes d'un domaine, également appelé vocabulaire. L'information disponible sur ce vocabulaire E est une matrice **asymétrique** de proximités P prenant ses valeurs dans $[0, 1]$. Cette matrice rend compte de l'attractivité sémantique **orientée** d'un terme x_i vis à vis d'un autre terme x_j. Ainsi, une forte valeur $P_{i,j}$ indique que la sémantique associée au terme x_i attire fortement celle du terme x_j, mais le contraire n'est pas forcément vrai (e.g. le terme "laryngite" attire son hypéronyme "larynx" alors que le contraire n'est pas forcément vrai).

Dans le cadre de notre étude, nous voulons à la fois traiter la notion de similarité sémantique ainsi que le degré de généralité/spécificité entre tous les termes du vocabulaire. Il faut donc analyser les asymétries existantes entre deux termes ($P_{i,j}$ vs. $P_{j,i}$) à partir du moment où leurs valeurs dépassent un seuil d'intérêt ϵ fixé. De façon simplifiée, une faible asymétrie ($P_{i,j} \simeq P_{j,i}$) sur un couple (x_i, x_j) pourrait correspondre à une relation de synonymie tandis qu'une forte asymétrie ($P_{i,j} \gg P_{j,i}$) pourrait symboliser une relation d'hyponymie (x_i hyponyme de x_j) c'est-à-dire x_i "est un type de" x_j. Cette observation nous conduit à proposer une définition pour la relation R où $var()$ et $moy()$ désignent respectivement la variance et la moyenne des deux valeurs de proximité observées sur le couple (x_i, x_j) :

$$\forall x_i, x_j \in E, x_i R x_j \Leftrightarrow P_{i,j} \geq \epsilon \wedge \left(P_{i,j} \geq P_{j,i} \vee var(\{P_{i,j}, P_{j,i}\}) \leq \alpha.moy(\{P_{i,j}, P_{j,i}\})^2\right) \tag{3}$$

Avec cette définition (3), deux termes x_i et x_j seront en relation mutuelle ($x_i R x_j$ et $x_j R x_i$) seulement si : i) les proximités $P_{i,j}$ et $P_{j,i}$ sont ϵ-significatives et ii) l'asymétrie n'est pas α-significative. Dans les autres cas, x_i et x_j seront en relation unilatérale s'ils excèdent le seuil ϵ. Dans le cas contraire, ils ne seront pas considérés comme proches. En résumé, l'*adhérence* d'un singleton $\{x\}$ correspondrait à l'ensemble des synonymes et/ou hyponymes du terme x, tandis que la *fermeture élémentaire* F_x généraliserait à l'ensemble des hyponymes indirects de x et de ses synonymes.

Nous basons notre méthode sur un algorithme de structuration ascendante (Largeron et Bonnevay (2002)) pour lequel nous inversons le principe afin d'obtenir une structuration descendante permettant la construction d'une structure sémantico-lexicale. Chaque étape de l'algorithme consiste à i) rechercher l'ensemble des individus (i.e. termes) à l'origine de *fermés élémentaires maximaux*, ii) les organiser en classes d'équivalence (concepts) et iii) ré-exécuter l'algorithme sur leur *fermeture* privée de leurs concepts correspondants.

L'algorithme d'origine fournit une structuration non hiérarchique mais ordonnée. Or, dans notre cas, il est facile de démontrer qu'un ordre partiel peut être défini sur le lexique. Par conséquent, celui-ci permet la construction d'une famille de graphes orientés acycliques non triangulaires correspondant aux sous-domaines du lexique (voir Figure 1). En particulier, on notera qu'un noeud du graphe peut contenir plusieurs termes, ce qui correspond à la notion de concept (ou de synset dans le cadre de l'ontologie WordNet - http ://wordnet.princeton.edu). Cette caractéristique correspond à la notion de (quasi-)synonymie puisque chacun des termes recouvre le même ensemble d'hyponymes et leur attractivité est mutuelle.

3 Illustration et discussion

Afin d'illustrer les résultats de notre approche, nous avons sélectionné l'ensemble des termes du Système Cardio-vasculaire de l'UMLS (*Unified Medical Language System* - http ://www.nlm.nih.gov/research/umls/) qui fait oeuvre de référence en matière d'ontologie médicale pour la langue anglaise. Nous avons ensuite construit la matrice asymétrique de proximités P à partir du calcul des co-occurrences de termes sur le corpus médical MEDLINE/PubMed (http ://pubmed.gov) composé de plus de 17 millions de résumés d'articles de revues du domaine des Sciences de la vie et de l'Information Biomédicale. Plusieurs mesures de similarités asymétriques ont été proposées dans la littérature (Dias et al., 2008) démontrant que dans le cadre de la construction de structures sémantico-lexicales, la Probabilité Conditionnelle (Equation 4) donnait en général les meilleurs résultats. C'est cette mesure de proximité que nous avons choisie pour cette première étude en imposant la co-occurrence au niveau du document.

$$\forall x_i, x_j \in E, P_{i,j} = p(x_j|x_i) \tag{4}$$

Les résultats comparatifs entre la taxonomie extraite de l'UMLS et la structure prétopologique ($\epsilon = \alpha = 0.05$) pour le système cardio-vasculaire sont présentés dans la figure (1). Notre approche basée sur la formalisation prétopologique montre des résultats très encourageants en tenant compte du fait qu'elle ne dépend ni de la langue ni du domaine. Mais surtout, elle propose un traitement en une seule et unique étape à partir de la simple analyse des co-occurrences de termes dans un corpus. A notre connaissance, il n'existe aucun travail similaire. Cependant, il reste de nombreux points à étudier. D'abord, le corpus médical MEDLINE/PubMed n'est manifestement pas assez représentatif du langage médical. En effet, dans le domaine cardio-vasculaire, 6 termes présents dans l'ontologie UMLS ne le sont pas dans le corpus. Afin de remédier à ce problème, nous travaillons actuellement sur la toile et projetons de mener une évaluation automatique d'ontologies. Ensuite, nous travaillerons sur la définition de mesures de similarité asymétriques informatives qui nous permettront d'atteindre des niveaux de confiance plus élevés dans la définition de l'*adhérence*.

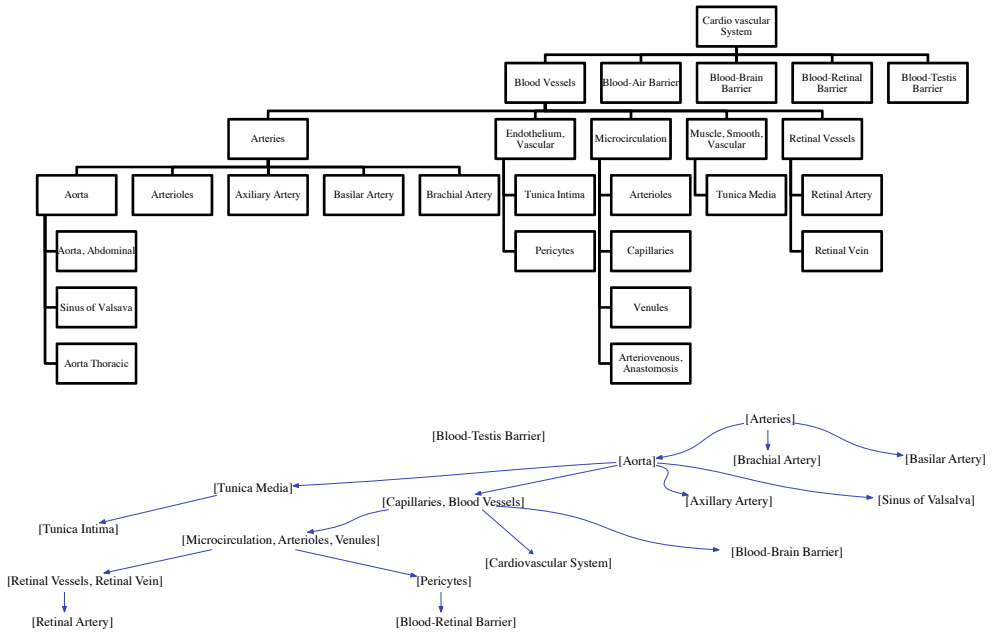

FIG. 1 – *L'ontology UMLS (en haut) et la structure prétopologique (en bas).*

Références

Belmandt, Z. (1993). *Manuel de prétopologie et ses applications : sciences humaines et sociales, réseaux, jeux, reconnaissance des formes, processus et modèles, classification, imagerie, mathématiques*. Paris : Hermès Sciences Publications.

Dias, G., R. Mukelov, et G. Cleuziou (2008). Unsupervised graph-based discovery of general-specific noun relationships from web corpora frequency counts. In *12th International Conference on Natural Language Learning*, Manchester, UK.

Largeron, C. et S. Bonnevay (2002). A pretopological approach for structural analysis. *Information Sciences 144*, 169 – 185.

Summary

This paper presents a new approach to build a prototype-based ontology from the observation of co-occurrences of terms inside a corpus. We propose to use the formalism of Pretopology to model the degree of generality/specificity as well as the semantic closeness between terms based on an asymmetric similarity matrix. In particular, we show that a partial order can be defined over the vocabulary and build a hierarchical structure.