Exploitation de l'Asymétrie entre Termes pour l'Extraction Automatique de Taxonomies à partir de Textes.

Davide Buscaldi, Guillaume Cleuziou, Gaël Dias, Vincent Levorato

Exploitation de l'asymétrie entre termes pour l'extraction automatique de taxonomies à partir de textes

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

*LIFO, Université d'Orléans
{davide.buscaldi,guillaume.cleuziou,vincent.levorato}@univ-orleans.fr
**GREYC, Université de Caen Basse-Normandie
gael.dias@unicaen.fr
***IRISE, CESI
vlevorato@cesi.fr

Résumé. Nous présentons dans cet article une nouvelle approche pour la génération automatique de structures lexicales (ou taxonomies) à partir de textes. Cette tâche est fondée sur l'hypothèse forte selon laquelle l'accumulation de faits statistiques simples sur les usages en corpus permet d'approximer des informations de niveau sémantique sur le lexique. Nous utilisons la prétopologie comme cadre de travail afin de formaliser et de combiner plusieurs hypothèses sur les usages terminologiques et enfin de structurer le lexique sous la forme d'une taxonomie. Nous considérons également le problème de l'évaluation des taxonomies résultantes et proposons un nouvel indice afin de les comparer et de positionner notre approche par rapport à la littérature.

1 Introduction

Le codage des relations sémantiques entre concepts au sein d'une structure lexico-sémantique de type « taxonomie » permet d'améliorer considérablement la pertinence des processus de recherche d'information (RI) et de traitement automatique des langues (TAL). Cependant, l'utilisation de telles ressources est fortement limitée du fait des efforts considérables à entreprendre pour les construire. Afin de limiter ces efforts, un certains nombre de recherches ont été entreprises ces dernières années pour « apprendre » des taxonomies à partir de l'observation des usages en corpus (Biemann, 2005; Cimiano et al., 2009). L'apprentissage automatique de taxonomies à partir de textes plutôt que leur construction manuelle présente des avantages indéniables. Non seulement cela permet d'ajuster la connaissance extraite à tout domaine de spécialité en choisissant le corpus adapté ; mais de plus, le coût par entrée lexicale sera considérablement réduit par rapport à une décision experte manuelle ou même assistée, permettant ainsi la génération de ressources plus importantes.
Différentes méthodologies d'apprentissage ont été proposées afin de construire automatiquement des taxonomies. Elles peuvent être organisées en trois types d'approches : les méthodes fondées sur la notion de similarités (Paaß et al., 2004; Cimiano et al., 2004), sur la théorie des

ensembles (Ganter et Wille, 1998; Cimiano et al., 2005) ou encore les méthodes associatives (Sanderson et Lawrie, 2000).
Cette étude se situe dans le cadre de travail des méthodes dites associatives. Nous proposons d'analyser la topologie des structures de graphe induites par les mesures d'associativité entre termes et d'en dériver une méthodologie non-supervisée, formalisée dans un cadre prétopologique, pour apprendre automatiquement une taxonomie. Ainsi, étant donnés un ensemble de termes (provenant éventuellement de différents domaines) et un corpus représentatif du/des domaine(s) considéré(s), (1) nous évaluons la proximité asymétrique entre chaque paire de termes, (2) nous définissons une famille de voisinages associée à chacun des termes, (3) nous modélisons par la prétopologie la diffusion des relations de subsomption dans le graphe et enfin (4) nous structurons les termes en un graphe orienté acyclique et non-triangulaire sensé approcher une taxonomie du/des domaine(s).
L'évaluation des structures apprises est une tâche encore difficile actuellement. En effet Smith (2004) note qu'il existe plusieurs manière d'envisager la conceptualisation d'un domaine selon l'utilisation qui doit en être faite ou plus simplement selon les points de vues experts. Nous proposons ici modestement, par souci de comparaison avec d'autres méthodes d'extraction automatique de taxonomies, une nouvelle mesure d'évaluation visant à comparer deux structures composées du même ensemble de termes initiaux. Cette mesure est alors utilisée pour quantifier la qualité d'une taxonomie apprise automatiquement par rapport à une taxonomie de référence construite manuellement et faisant référence dans le domaine.

2 Travaux connexes

Nous présentons, dans cette section, les principales méthodologies d'acquisition automatique de taxonomies.
Les approches utilisant les similarités procèdent en premier lieu par construction d'une matrice de similarités entre termes à partir d'une représentation de ces termes par des vecteurs de contextes numériques. Par exemple, Pereira et al. (1993) représentent les termes par des distributions sur leurs fréquences d'apparition dans des contextes syntaxiques spécifiés, la similarité est ensuite quantifiée par le calcul de l'entropie relative entre distributions. La taxonomie finale est obtenue par classification hiérarchique divisive dans laquelle chaque division est assurée par une méthode de recuit simulé visant à minimiser l'entropie moyenne au sein des classes. Dans cette approche, les nœuds de la structure sont peuplés en conservant les termes les plus représentatifs de chaque cluster comme label. De façon comparable, Caraballo (1999) utilisent des *patterns* pré-définis pour construire les vecteurs de contextes avec une similarité entre termes définie par la mesure du cosinus entre ces vecteurs de contextes. Ensuite, une classification ascendante hiérarchique est réalisée et pour chaque nœud interne de l'arbre hiérarchique un nom représentatif est sélectionné comme label. L'une des principales difficultés des approches par similarités concerne l'étiquetage des nœuds internes, d'autre part ces approches sont généralement dépendantes de la langue et de la disponibilité sur cette langue d'outils linguistiques (étiqueteur grammatical, analyseur syntaxique, etc.).
Les méthodes d'acquisition de taxonomie utilisant la théorie des ensembles reposent sur un ordonnancement partiel des objets (ici les termes) suivant des relations d'inclusions sur leurs ensembles d'attributs (Petersen, 2004; Sporleder, 2002; Cimiano et al., 2005). Par exemple, Cimiano et al. (2005) présentent une approche de type Analyse Formelle de Concepts (FCA)

(Ganter et Wille, 1998) dans laquelle chaque nom est caractérisé par un ensemble d'attributs composés par les verbes pour lesquels le nom apparaît comme argument. Les attributs d'un nom forment son contexte formel, ces contextes sont ensuite utilisés pour construire un treillis de concepts formels (ensemble de noms partageant les mêmes contextes formels). Malgré le support théorique intéressant, on constate à nouveau que ce type d'approche requiert l'utilisation d'outils linguistiques propres à la langue considérée (par exemple les relations Verbe/Objet ou Verbe/Sujet).

Les méthodes associatives (Sanderson et Croft, 1999; Sanderson et Lawrie, 2000; Dias et al., 2008) se concentrent sur les distributions des termes dans le corpus de documents afin d'extraire des relations du type « est plus général que » entre deux termes qui constituent l'information structurante des taxonomies. Ces relations de généralité/spécificité sont formalisées par la notion de subsomption dans un corpus selon laquelle un terme t_1 est subsumé par un autre terme t_2 si t_2 apparaît dans la plupart des documents où t_1 apparaît et si l'inverse n'est pas vérifié. Sanderson et Croft (1999) ont été les premiers à utiliser la subsomption pour dériver une taxonomie de termes à partir d'un corpus de textes. Leur définition de la subsomption revient à considérer que t_2 subsume t_1 si la présence de t_1 dans un document s'accompagne de la présence de t_2 avec une probabilité élevée fixée à $P(t2|t_1) \geq 0.8$. En collectant l'ensemble des relations ainsi définies, ils construisent une structure sémantique correspondant à un graphe orienté sans cycle (DAG) puis à un DAG non-triangulaire après suppression des transitivités. La définition de la subsomption est rendue paramétrable dans (Sanderson et Lawrie, 2000) à travers l'expression suivante $P(t_2|t_1) \geq P(t_1|t_2)$ et $P(t_2|t_1) > t$ où t [1] désigne le paramètre à ajuster en fonction du corpus. Dias et al. (2008) proposent une méthodologie d'ordonnancement complet de l'ensemble des termes. Ils construisent un graphe orienté à partir d'une relation de subsomption selon laquelle t_1 est plus général que t_2 si la proximité sémantique de t_1 vers t_2 est plus forte que de t_2 vers t_1, étant donné une mesure de proximité asymétrique (une liste de sept mesures issues de la littérature étant établie). Le graphe est pondéré par la valeur de proximité calculée, puis un score de généralité est calculé pour chaque nœud (terme) du graphe via l'algorithme TextRank (Mihalcea et Tarau, 2004) afin d'en dériver un ordre total sur les termes, susceptible d'aider à la structuration de type taxonomie.

D'autres approches associatives visent à apprendre des patterns permettant d'extraire des relations de subsomption plus précises telles que les relations *part-of* ou *is-a* entre termes. Ces patterns peuvent êtres composés par une séquence de mots, une séquence de caractères ou encore une structure syntaxique. Dans (Snow et al., 2004), un classifieur est construit afin d'extraire les relations d'hyperonymie en utilisant les chemins de dépendance extraits d'arbres syntaxiques. Les relations extraites peuvent être utilisées afin d'enrichir des taxonomies existantes telles que WordNet. Dans Navigli et Velardi (2010), les définitions présentes dans Wikipedia sont utilisées pour extraire des treillis de classes de mots composés d'étiquettes morpho-syntaxiques et de mots clé. Enfin, Kozareva et Hovy (2010) utilisent le Web comme corpus pour calculer la *confidence* sur les relations d'hyperonymie en comparant le nombre de séquences du type « *X <is-a pattern> Y* » par rapport au nombre de séquences inverses « *Y <is-a pattern> X* ». Ils tentent de reconstruire ainsi la taxonomie WordNet sur les animaux, les plantes et les véhicules. Cependant, Yang et Callan (2009) montrent que les statistiques issues de l'analyse des co-occurrences sont aussi efficaces que l'utilisation de patterns lexico-syntaxiques.

1. t reste fixé à 0.8 dans les expériences réalisées par Sanderson et Lawrie (2000).

Bien que les modèles associatifs présentent des propriétés intéressantes, notamment leur indépendance partielle ou totale vis-à-vis du domaine ou de la langue, ils souffrent de limitations évidentes comme le précisent d'ailleurs Sanderson et Croft (1999); Sanderson et Lawrie (2000); Dias et al. (2008). En effet, tous ces modèles ont tendance à sur-générer les relations de subsomption entre termes et par conséquent conduisent à la création de structures lexico-sémantiques difficiles à exploiter lorsque le nombre de termes augmente. Afin de traiter ce problème, nous introduisons une nouvelle approche utilisant le formalisme de la prétopologie pour générer une structure de DAG non-triangulaire à partir d'une matrice de proximités asymétriques. En particulier, un fort accent est mis sur l'exploitation de la topologie de l'espace de représentation des termes induit par cette matrice.
A ce stade de notre argumentation, il est nécessaire de préciser que les structures lexicales qui seront extraites automatiquement doivent être vues comme des approximations de taxonomies d'un domaine, dans la mesure où certaines relations de subsomption qui seront extraites du corpus sont purement accidentelles dans le sens où elles ne correspondent pas à une relation sémantique attestée.

3 Algorithme prétopologique de structuration

Les liens entre les éléments d'une population peuvent être modélisés de diverses manières, e.g. par un espace topologique. Cependant, les axiomes et les propriétés d'un espace topologique sont trop restreints pour modéliser concrètement un espace lexical. Grâce à la prétopologie, on peut modéliser la notion de proximité d'une manière plus générale. Nous proposons donc d'utiliser ce formalisme pour modéliser un « espace lexical » basé sur des relations définies par une prétopologie afin d'en extraire une structure permettant des stratégies de propagation.

3.1 Notions prétopologiques

Nous définissons un espace prétopologique par une famille de voisinages. Soit (E, a) un espace prétopologique (Belmandt, 2011) où $a(.)$ est une fonction adhérence et E un ensemble non vide. Un voisinage $N(x)$ de $x \in E$ est un sous-ensemble de E contenant x, et une famille de voisinages $\mathcal{N}(x)$ pour x est définie comme l'union des voisinages de x telle que $\mathcal{N}(x) = \{N \subseteq E | x \in N\}$. On construit une adhérence basée sur une famille de voisinages telle que :

$$\forall A \in \mathcal{P}(E), a(A) = \{x \in E | (\bigcap_{N \in \mathcal{N}(x)} N(x)) \cap A \neq \emptyset\}$$

Dans notre contexte d'étude, un espace prétopologique est défini par un vocabulaire E (ensemble de termes) et un opérateur adhérence $a(.)$ modélisant la propagation des dépendances sémantiques à travers des ensembles de termes. La façon dont on définit la famille de voisinages est primordiale pour notre modélisation. Par exemple, l'approche proposée par Sanderson et Lawrie (2000) peut être décrite dans un cadre prétopologique par deux voisinages $\mathcal{N}_{OHC}(x) = \{N_O(x), N_{HC}(x)\}$ correspondant aux deux propriétés (indice de confiance élevé et ordonnancement) utilisées dans la définition de la subsomption de Sanderson et Lawrie (2000) i.e. $N_{HC}(x) = \{y \in E | P(y|x) > t\}$ et $N_O(x) = \{y \in E | P(y|x) \geq P(x|y)\}$.

Sachant que la fonction adhérence n'est pas idempotente, on peut l'appliquer successivement jusqu'à obtenir un ensemble fermé. Les fermés représentent des sous-ensembles en relation avec la fonction adhérence. Une structure induite par les fermés élémentaires (construits à partir des singletons de E) et par les fermés maximaux (en terme d'inclusion) peut être considérée comme l'ensemble de groupes les moins homogènes de E. La nature de ces sous-ensembles est intéressante en terme d'analyse car nous considérons une relation d'inclusion entre ces sous-ensembles, permettant d'appliquer un algorithme de structuration. Une telle structure peut être obtenue par un algorithme prétopologique proposé par Largeron et Bonnevay (2002). Nous proposons ici une version « top-down » de cet algorithme qu'on retrouve dans Cleuziou et al. (2011). En utilisant $\mathcal{N}_{OHC}$, l'algorithme nous fournit donc un graphe acyclique non-triangulaire qui correspond exactement à la structure obtenue par Sanderson et Lawrie (2000).

3.2 Analyse de l'algorithme de structuration

De manière à correspondre le plus possible aux structures lexicales existantes (e.g. WordNet), la structure finale $\mathcal{S}$ doit satisfaire les propriétés suivantes :

- structure de type arbre : chaque noeud C doit être caractérisé par deux ensembles disjoints de prédecesseurs $Pred(C)$ et de successeurs $Succ(C)$, sans aucun cycle.
- noeuds aggrégateurs : chaque noeud C doit contenir un ou plusieurs termes du vocabulaire E ; du point de vue appliqué, les noeuds de la structure seront appelés de manière générale des « concepts ».

L'idée globale de l'algorithme de structuration version « top-down » de Largeron et Bonnevay (2002) suit les étapes suivantes :

1. Construire la famille des fermés élémentaires $\mathcal{F}e(E, a)$, c'est-à-dire l'ensemble des fermés des singletons x de E.
2. Extraire la famille des fermés élémentaires maximaux $\mathcal{F}M(E, a)$. Cela revient à énumérer tous les fermés élémentaires qui sont maximaux par inclusion de $\mathcal{F}e(E, a)$. Chaque élément de $F \in \mathcal{F}M(E, a)$ est un noyau.
3. A partir de chaque noyau, on détermine les fermés élémentaires les plus grands jusqu'à ne plus en trouver aucun. Ce processus récursif nous permet de générer, à partir de chaque noyau, un ensemble de parties homogènes par réductions successives, et d'en sortir la structure sémantico-lexicale finale.

Quand on utilise les voisinages de *la confidence la plus élevée* $\mathcal{N}_{OHC}$ comme défini précédemment, l'algorithme structurant nous fournit un DAG (Directed Acyclic Graph) non-triangulaire qui correspond à la structure obtenue par Sanderson et Lawrie (2000). Le cadre général de la prétopologie nous permet de proposer des voisinages permettant d'aggréger les noeuds dans une structure pertinente pour la modélisation d'espace lexical.

3.3 Modélisation d'espaces lexicaux

3.3.1 k-plus proches voisins

Un espace prétopologique des k-plus proches voisins (k-NN) consiste en la définition du voisinage d'un élément x par le sous-ensemble composé des k éléments les plus proches de x. Dans

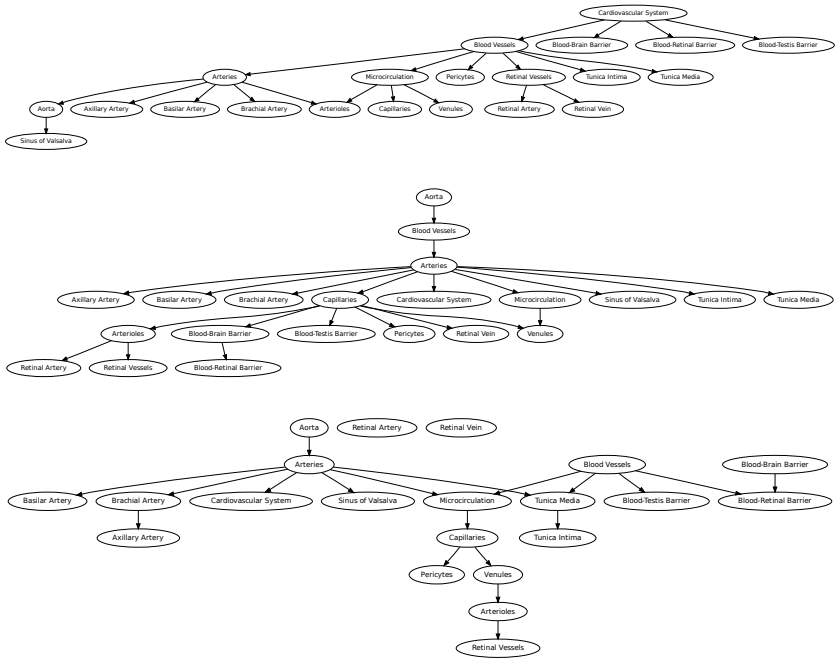

FIG. 1 – *La structure de référence pour le sous-domaine Cardio (en haut), la structure obtenue par l'algorithme de structuration utilisant les voisinages 2NN (au milieu) et HC (en bas).*

le cadre de notre problème, nous choisissons comme voisin de x les termes y qui ont la confidence la plus élevée selon $P(y|x)$. Ainsi, nous définissons la famille de voisinage suivante : $\mathcal{N}_{kNN}(x) = \{N_{kNN}(x), N_O(x)\}$ avec $N_{kNN}(x) = \{y \in E | y \in kNN_E(x)\}$. La famille $\mathcal{N}_{kNN}$ permet de définir l'opérateur adhérence $a(.)$ tel que $a(x)$ correspond au singleton $\{x\}$ étendu aux termes plus généraux (N_O) qui possèdent x dans leurs « meilleurs » prédécesseurs (N_{kNN}).

3.3.2 Voisins relatifs orientés (DRN)

La seconde modélisation est basée sur une propriété statistique observée sur les structures lexicales. Etant donnés une structure de référence $\mathcal{S}_r$ et un corpus du domaine de $\mathcal{S}_r$, nous avons effectué une analyse de la distribution de la confidence par rapport aux chemins que l'on trouve dans la référence, de la racine aux feuilles. Nous avons testé plusieurs hypothèses, et l'une d'entre elles est apparue comme pertinente au regard des tests statistiques effectués. Soit $x_1, x_2, \ldots, x_n$ un chemin dans la structure de référence tel que x_i subsume x_{i+1}. On constate que le terme x_i du chemin possède une confidence plus petite avec ses prédécesseurs que ses successeurs ont avec leurs propres prédécesseurs. Cette constatation peut être formalisée par : $\forall i,\ min\{P(x_j|x_i)\}_{j=1}^{i-1} \geq min\{P(x_j|x_{i+1})\}_{j=1}^{i}$. En appliquant cette propriété localement sur un triplet (w, x, y), y est un voisin de x ssi un w quelconque satisfait la propriété d'être

un successeur de x dans le chemin (x, y) i.e. $N_{DRN}(x) = \{y \in E | \forall w \in E,\ P(y|x) \geq min\{P(x|w), P(y|w)\}\}$. Comme conséquence, une nouvelle famille de voisinages est proposée : $\mathcal{N}_{DRN}(x) = \{N_{DRN}(x), N_O(x)\}$. Plus précisément, l'adhérence dérivée de $\mathcal{N}_{DRN}$ étend le terme singleton $\{x\}$ avec ses termes plus généraux (N_O) qui satisfont la propriété ultramétrique (N_{DRN}). Une propriété intéressante de $\mathcal{N}_{DRN}$ est qu'il n'y a besoin d'aucun paramètre. Cependant, cela mène en pratique à des voisinages sur-dimensionnés. Une manière d'ajuster ces voisinages consiste en l'introduction d'un paramètre basé sur la confidence tel que $\mathcal{N}_{HC\_DRN}(x) = \{N_{DRN}(x), N_O(x), N_{HC}(x)\}$. Une autre solution permettant d'éviter le paramétrage du voisinage est présentée dans la section suivante.

3.3.3 k-plus proches voisins relatifs orientés

Comme mentionné précédemment, le voisinage type DRN produit des voisinages sur-dimensionnés. Pourtant, il permet de jouer un rôle de « filtre », et ainsi de forcer la sélection d'un voisinage pertinent par rapport aux structures que l'on souhaite obtenir. Pour cela, on définit un nouveau voisinage qui combine d'une part les bénéfices structurels ainsi que le paramétrage simple de l'approche kNN, et d'autre part la propriété statistique qu'on obtient par l'approche DRN tel que $N_{kN\_DRN}(x) = \{y \in E | y \in kNN_{N_{DRN}(x)}(x)\}$. Au final, la nouvelle famille de voisinages s'écrit :

$$\mathcal{N}_{kN\_DRN}(x) = \{N_{kN\_DRN}(x), N_O(x)\}$$

4 Expérimentations

Deux jeux de tests ont été utilisés comme référence pour mettre en évidence deux types de relations sémantiques : la synonymie et la méronymie. Tout d'abord, nous avons utilisé UMLS [2] duquel quatre sous-domaines distincts ont été choisis (*cardiovascular* (CS), *digestive* (DS), *respiratory* (RS) et *nervous* (NS)). Chaque sous-domaine est représenté par sa propre structure lexicale présente dans le meta-thesaurus utilisant la relation hyperonyme/hyponyme. La seconde ontologie de référence a été extraite de WordNet en considérant tous les lieux géographiques dérivés du concept « United States of America » (relation de méronymie). Nous la nommerons GEO-WordNet. Pour chaque référence, nous avons récupéré les proximités entre les termes à partir de deux corpus différents. Pour UMLS, nous avons utilisé (1) PubMed [3] et (2) BioMed [4]. Pour GEO-WordNet, nous avons utilisé le Glasgow Herald (GH95) et le Los Angeles Times (LAT94) qui sont tout deux utilisés dans les campagnes d'évaluations GeoCLEF [5], où les toponymes ont été identifiés par la Stanford Named Entity Recognition (NER) (Finkel et al., 2005) et désambigüisés par une méthode basée sur la densité conceptuelle.
Maedche et Staab (2002) ont proposé de comparer les ontologies par leurs structures (la mesure J_1). Soit un ensemble de termes E et deux ontologies $\mathcal{O}_1$ et $\mathcal{O}_2$ qui strucuturent E, l'idée générale est de comparer pour chaque élément $x \in E$ la correspondance entre les super/-subconcepts de x dans $\mathcal{O}_1$ et les super/subconcepts de x dans $\mathcal{O}_2$. Ce type d'évaluation est

2. http ://www.nlm.nih.gov/research/umls/
3. http ://www.ncbi.nih.gov/pubmed/
4. http ://www.biomedcentral.com/
5. http ://ir.shef.ac.uk/geoclef

applicable dans notre cas en quantifiant la correspondance qu'il y a entre les prédécesseurs $Pred_{\mathcal{S}}(x)$ et les successeurs $Succ_{\mathcal{S}}(x)$ d'un terme x dans les deux structures lexicales $\mathcal{S}_1$, $\mathcal{S}_2$. Cependant, le principal inconvénient de J_1 est qu'il n'est pas sensible à l'orientation des relations dans ces structures. Ainsi, si on inverse toutes les relations de celles-ci, on aura tout de même une correspondance totale. Pour éviter ce problème, nous proposons d'évaluer séparément les prédecesseurs et les successeurs dans l'évaluation de la mesure. Un nouvel indice de mesure J_2 est proposé comme étant la moyenne (géométrique) de deux indices de Jaccard pour lequel un score de 1 correspond à deux structures strictement identiques dont la définition est :

$$J_2(\mathcal{S}_1, \mathcal{S}_2) = \frac{1}{|X|} \sum_{x \in E} \left(\frac{|Pred_{\mathcal{S}_1}(x) \cap Pred_{\mathcal{S}_2}(x)|}{|Pred_{\mathcal{S}_1}(x) \cup Pred_{\mathcal{S}_2}(x)|} \right)^{1/2} \cdot \left(\frac{|Succ_{\mathcal{S}_1}(x) \cap Succ_{\mathcal{S}_2}(x)|}{|Succ_{\mathcal{S}_1}(x) \cup Succ_{\mathcal{S}_2}(x)|} \right)^{1/2}$$

Afin d'évaluer les structures qu'on obtient avec la mesure J_2, il est nécessaire de fixer le paramètre k ou le seuil t selon l'espace prétopologique que l'on utilise. Dans nos expériences, nous avons noté deux phénomènes : (1) les meilleures structures ont été obtenues avec des voisinages de petite taille, et (2) la taille des voisinages doit pouvoir être comparable afin d'être en mesure de confronter les résultats d'espaces prétopologiques différents. La première observation corrobore l'idée de Sanderson et Lawrie (2000) de garder uniquement les valeurs de confidence élevées en utilisant un seuil élevé (e.g. $t = 0.8$). Ceci est illustré dans la Figure 2 qui donne les scores du J_2 obtenu par l'espace prétopologique $\mathcal{N}_{OHC}$ où les valeurs de confidence retenues augmentent (i.e. le seuil t décroît). En outre, un tel seuil ne peut pas être universel et doit être ajusté pour chaque corpus. Par exemple, la proportion de liens dont les valeurs de confidence sont supérieures à 0.8 est d'environ 1% pour le sous-domaine CS de l'UMLS avec le corpus BioMed, et seulement de 0.07% pour GEO-WordNet basé sur le corpus LAT94.

Partant de ces constatations, nous avons introduit deux heuristiques permettant le paramétrage des espaces prétopologiques. Soit n la taille du vocabulaire E que l'on souhaite structurer, le seuil t est ajusté de manière à ce que seulement n valeurs de confidence soient supérieures à t (première heuristique) ou que $2n$ valeurs de confidence soient supérieures à t (deuxième heuristique). Ces deux heuristiques sont utilisées dans les espaces prétopologiques nécessitant des valeurs de confidence élevées (i.e. $\mathcal{N}_{OHC}$ et $\mathcal{N}_{HC\_DRN}$). Pour les espaces prétopologiques type plus proches voisins, (i.e. $\mathcal{N}_{kNN}$ et $\mathcal{N}_{kN\_DRN}$) les voisinages possédant des tailles comparables sont obtenus avec les paramètres $k = 1$ et $k = 2$. La table 1 montre l'évaluation de chaque structure obtenue comparée à la structure de référence.

D'importantes variations de la mesure J_2 sont observées selon les jeux de test et le corpus utilisés. Par exemple, une faible correspondance est obtenue pour RS avec PubMed où les scores sont parfois inférieurs à 0.10 alors qu'une forte correspondance est obtenue sur NS et pour le domaine géographique où les scores peuvent dépasser 0.40. De telles variations peuvent être expliquées par la nature des structures de référence utilisées et plus particulièrement le type de relations sémantiques que l'on considère. Certains des domaines étudiés sont structurés par des relations de type *Part-of* (e.g. Geo-WordNet) ou de type *Is-a* (e.g. sous-domaine NS), alors que d'autres structures de référence mélangent les deux types comme le sous-domaine CS par exemple (et la référence UMLS globale par extension). Il semble incontestable qu'une telle hétérogénéité dans la structure sémantique du vocabulaire est un problème que nos approches ne traitent pas.

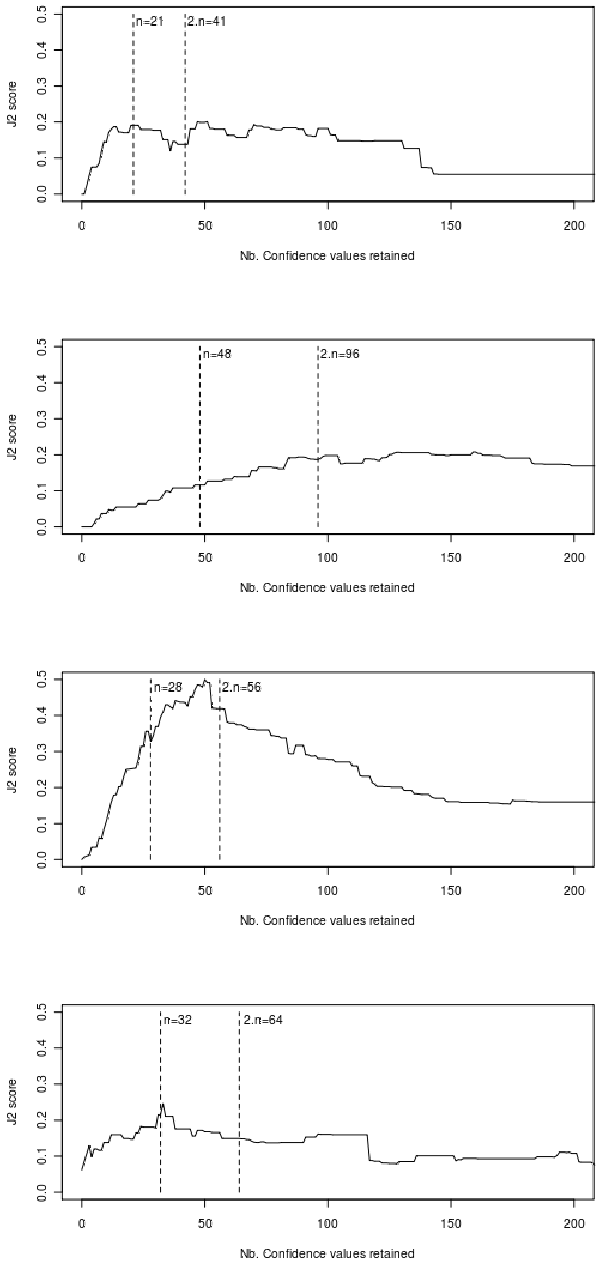

FIG. 2 – *Scores J_2 basés sur l'approche de Sanderson et Lawrie (2000) sur le corpus BioMed : les sous-domaines CS, DS, NS et RS sont représentés de haut en bas.*

Corpus	Domaine	n	$\mathcal{N}_{OHC}$		$\mathcal{N}_{kNN}$		$\mathcal{N}_{HC\_DRN}$		$\mathcal{N}_{kN\_DRN}$	
			n	$2n$	k=1	k=2	n	$2n$	k=1	k=2
BioMed	Cardiovascular system	21	0.191	0.138	0.189	**0.304**	0.192	0.133	0.144	0.136
	Digestive system	48	0.116	**0.187**	0.104	0.137	0.116	0.175	0.110	0.116
	Nervous system	28	0.328	0.419	**0.428**	0.382	0.344	0.392	**0.428**	0.414
	Respiratory system	32	0.215	0.149	0.188	0.240	0.220	0.138	0.154	**0.251**
	UMLS (4 sous-domaines)	128	0.172	**0.218**	0.151	0.162	0.184	0.213	0.180	0.173
PubMed	Cardiovascular system	21	0.100	**0.173**	0.133	0.147	0.097	0.162	0.133	0.166
	Digestive system	48	0.130	0.107	0.123	0.111	0.117	0.138	**0.243**	0.188
	Nervous system	28	0.196	0.258	**0.440**	0.401	0.208	0.257	0.429	0.381
	Respiratory system	32	0.095	0.119	**0.143**	0.139	0.092	0.127	0.131	0.101
	UMLS (4 sous-domaines)	128	0.102	0.132	0.165	0.142	0.096	0.145	0.169	**0.171**
LAT94	GEO-WordNet (USA)	150	0.207	0.312	**0.392**	0.332	0.183	0.276	0.386	0.347
GH95		131	0.305	0.372	**0.399**	0.382	0.289	0.312	0.391	0.382

TAB. 1 – *Evaluation de la correspondance de chaque structure obtenue avec la structure de référence basée sur la mesure J_2.*

Le corpus utilisé et la manière d'exploiter une collection de textes ont un impact significatif sur la qualité des statistiques obtenues, et donc sur la structure lexicale générée. Il est évident que les scores sont plus faibles en utilisant le corpus PubMed que le corpus BioMed. Ceci est principalement dû au fait que seuls les résumés des textes ont été utilisés pour calculer les valeurs de confidence sur PubMed alors les textes complets ont été traités sur BioMed.
La table 1 donne également la comparaison entre les structures obtenues par les différents types d'espaces prétopologiques. Les valeurs en gras dans la table permettent de distinguer la modélisation qui donne le meilleur résultat pour un jeu de test et un corpus donnéés. Il est intéressant de noter que même si l'espace prétopologique $\mathcal{N}_{HC\_DRN}$ basé sur une ultramétrique ne donne jamais le meilleur score, le filtrage effectué par cet espace profite aux résultats de l'espace prétopologique $\mathcal{N}_{kN\_DRN}$ basé sur les plus proches voisins qui obtient les meilleurs résultats dans nos quatre expérimentations. En résumé, on peut constater que la modélisation basée sur $\mathcal{N}_{kN\_DRN}$ permet de donner des meilleurs résultats que Sanderson et Lawrie (2000) sur deux tiers des contextes expérimentés, et même parfois avec de significatives performances comme dans le cas des jeux de test géographiques avec un score amélioré de 87% dans le meilleur des cas.

5 Conclusion

Dans ce papier, nous avons présenté un nouveau cadre de travail qui construit automatiquement des ontologies à des fins terminologiques basé sur le formalisme prétopologique. Plus particulièrement, la prétopologie propose un cadre théorique mathématique adapté pour modéliser le degré de généralité/spécificité des termes ainsi que la proximité sémantique (asymétrique) entre les termes. A la différence des approches classique, nous traitons le cas de l'asymétrie, ce qui permet de se détacher du domaine et de la langue des textes traités. Nous nous focalisons sur la topologie de la structure obtenue à partir de la mesure de proximité en évitant d'avoir des termes isolés et en simplifiant l'aspect paramétrisation. Nous proposons également une évaluation intrinsèque des structures sémantico-lexicales apprises en nous basant sur l'indice J_2 que nous avons défini, résolvant le problème de l'inversion des termes dans l'ontologie, lequel est présent dans le travail de Maedche et Staab (2002). Nous proposons une validation de notre modèle basée sur deux jeux de test : l'ontologie de référence du domaine

médical UMLS avec l'utilisation des corpus PubMed et BioMed, puis l'ontologie de référence GEO-WordNet avec l'utilisation des corpus GH95 et LAT94. Nous avons comparé nos résultats avec la méthode de Sanderson et Lawrie (2000) qui est une des plus connue du domaine, nous permettant de constater que notre approche donne de meilleurs résultats dans la majorité des cas.

Références

Belmandt, Z. T. (2011). *Basics of pretopology*. Hermann.

Biemann, C. (2005). Ontology Learning from Text – a Survey of Methods. *LDV-Forum 20*(2), 75–93.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, Morristown, NJ, USA, pp. 120–126. Association for Computational Linguistics.

Cimiano, P., A. Hotho, et S. Staab (2004). Comparing conceptual, partitional and agglomerative clustering for learning taxonomies from text. In *Proceedings of the European Conference on Artificial Intelligence (ECAI'04)*, Valencia, Spain, pp. 435–439. IOS Press.

Cimiano, P., A. Hotho, et S. Staab (2005). Learning concept hierarchies from text corpora using formal concept anaylsis. *Journal of Artificial Intelligence Research 24*, 305–339.

Cimiano, P., A. Mädche, S. Staab, et J. Völker (2009). Ontology learning. In *Handbook of Ontologies*, pp. 245–267. Springer Verlag.

Cleuziou, G., G. Dias, et V. Levorato (2011). Acquisition de structures lexico-sémantiques à partir de textes : un nouveau cadre de travail fondé sur une structuration prétopologique. In *Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'2011)*, pp. 107–118.

Dias, G., R. Mukelov, et G. Cleuziou (2008). Fully unsupervised graph-based discovery of general-specific noun relationships from web corpora frequency counts. In *CoNLL '08 : Proceedings of the Twelfth Conference on Computational Natural Language Learning*, Morristown, NJ, USA, pp. 97–104. Association for Computational Linguistics.

Finkel, J. R., T. Grenager, et C. Manning (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *In ACL*, pp. 363–370.

Ganter, B. et R. Wille (1998). *Formal Concept Analysis : Mathematical Foundations* (1 ed.). Springer.

Kozareva, Z. et E. H. Hovy (2010). A semi-supervised method to learn and construct taxonomies using the web. In *Empirical Methods in Natural Language Processing*, pp. 1110–1118.

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

Maedche, A. et S. Staab (2002). Measuring similarity between ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, London, UK, pp. 251–263. Springer-Verlag.

Mihalcea, R. et P. Tarau (2004). TextRank : Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Navigli, R. et P. Velardi (2010). Learning word-class lattices for definition and hypernym extraction. In *Meeting of the Association for Computational Linguistics*, pp. 1318–1327.

Paaß, G., J. Kindermann, et E. Leopold (2004). Learning prototype ontologies by hierachical latent semantic analysis. In *15th ECML/PKDD conference*.

Pereira, F., N. Tishby, et L. Lee (1993). Distributional clustering of english words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, Morristown, NJ, USA, pp. 183–190. Association for Computational Linguistics.

Petersen, W. (2004). A set-theoretical approach for the induction of inheritance hierarchies. *Electronic Notes in Theoretical Computer Science 53*, 296 – 308. Proceedings of the joint meeting of the 6th Conference on Formal Grammar and the 7th Conference on Mathematics of Language.

Sanderson, M. et B. Croft (1999). Deriving concept hierarchies from text. In *SIGIR '99 : Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 206–213. ACM.

Sanderson, M. et D. Lawrie (2000). Building, testing, and applying concept hierarchies. In W. B. Croft (Ed.), *Advances in Information Retrieval*, pp. 235–266. Dordrecht : Kluwer Academic Publishers.

Smith, B. (2004). Ontology. In L. Floridi (Ed.), *The Blackwell Guide to Philosophy of Computing and Information*, pp. 155–166. Malden : Blackwell.

Snow, R., D. Jurafsky, et A. Y. Ng (2004). Learning syntactic patterns for automatic hypernym discovery. In *Neural Information Processing Systems*.

Sporleder, C. (2002). A Galois Lattice based Approach to Lexical Inheritance Hierarchy Learning. In *15th European Conference on Artificial Intelligence (ECAI'02) : Workshop on Machine Learning and Natural Language Processing for Ontology Engineering, Lyon, France*.

Yang, H. et J. Callan (2009). A metric-based framework for automatic taxonomy induction. In *Meeting of the Association for Computational Linguistics*, pp. 271–279.

Summary

We present in this paper a new approach for the automatic generation of lexical structures from texts. This task is based on the strong hypothesis that simple statistical observations on textual usages can provide pieces of semantics about the lexicon. Using such "naive" observations only, we propose a (pre)-topological framework to formalize and combine various hypotheses on textual data usages and then to derive a structure similar to usual lexical knowledge databases such as WordNet. In addition, we also consider the evaluation problem for the resulting lexical structures; we propose a new measure to compare two structures and use it to quantify the contribution of the new structuring approach with respect to the corresponding solution proposed by Sanderson et Lawrie (2000) on two case studies that differ on the domain and the size of the lexicon.