Acquisition de structures lexico-sémantiques à partir de textes : un nouveau cadre de travail fondé sur une structuration prétopologique

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

*LIFO
Université d'Orléans - Orléans, FRANCE
nom.prenom@univ-orleans.fr
**HULTIG
Université de Beira Interior - Covilhã, PORTUGAL
ddg@di.ubi.pt

Résumé. Les structures lexico-sémantiques jouent un rôle essentiel dans les processus de fouille de textes. En codant les relations sémantiques entre concepts du discours elles apportent une connaissance stratégiques pour enrichir les capacités de raisonnement. Le développement de telles structures étant fortement limité du fait des efforts nécessaires à leur construction, nous proposons un nouveau formalisme d'acquisition automatique d'ontologies terminologiques à partir de textes. Nous utilisons pour cela une formalisation prétopologique de l'espace des termes sur laquelle s'appuie un modèle générique de structuration. Nous présentons une étude empirique préliminaire rendant compte du potentiel de ce modèle en terme d'extraction de connaissances.

1 Introduction

Les structures lexico-sémantiques jouent un rôle essentiel en Recherche d'Information (RI) et en Traitement du Langage Naturel (TLN). En codant les relations sémantiques entre concepts du discours elles permettent d'enrichir les capacités de raisonnement pour des applications telles que la classification de textes (Hotho et al., 2003), l'expansion de requêtes (Bhogal et al., 2007), la recherche d'information personnalisée (Mylonas et al., 2008) pour n'en citer que quelques unes. Cependant, leur développement est fortement limité en raison des efforts nécessaires à leur construction. Ces dernières années, des études ont été menées dans le but d'apprendre de telles structures à partir des textes et ainsi réduire la quantité de travail d'ingénierie nécessaire à leur élaboration (Cimiano et al., 2009). Apprendre des structures lexico-sémantiques directement à partir des textes plutôt que de les établir manuellement présente des avantages indéniables : d'abord la possibilité d'élaborer une structure pour un domaine spécifique en ajustant le corpus de textes en conséquence ; également la réduction considérable du "coût" pour chaque entrée lexicale, permettant alors d'envisager des structures à large échelle.

Même si l'acquisition automatique de ressources sémantiques de tout type est une tâche difficile, différentes méthodologies ont été proposées pour apprendre des information de niveau conceptuel à partir de corpus de textes. On peut les organiser en trois classes majeures : (1) les méthodes à base de similarités (Pereira et al., 1993), (2) les méthodes s'appuyant sur la théorie des ensembles (Cimiano et al., 2005) et (3) les méthodes fondées sur les modèles associatifs (Sanderson et Croft, 1999; Sanderson et Lawrie, 2000; Dias et al., 2008). Les deux premières méthodes adoptent l'hypothèse distributionnelle de Harris [1] (Harris, 1968), elles s'appuient toutes les deux sur une caractérisation des termes par des vecteurs de contextes mais se différencient par le traitement qui en est fait. Les premières utilisent des mesures de similarité (e.g. mesure du cosinus) afin de quantifier le degré de similarité entre vecteurs de contextes et ainsi appréhender les données dans un espace numérique ; à l'inverse les méthodes ensemblistes procèdent par ordonnancement partiel des termes à partir des relations d'inclusion entre ensembles de contextes et réalisent un traitement symbolique des données. Enfin, les approches associatives ne suivent pas l'hypothèse Harrissienne mais organisent les termes hiérarchiquement en utilisant des mesures d'association asymétriques censées modéliser le degré (ou force) de subsomption d'un terme vis-à-vis d'un autre.

Les trois méthodologies énoncées s'appuyant sur des paradigmes différents, les structures lexico-sémantiques extraites diffèrent également. D'une part les méthodes à base de similarités conduisent à des ontologies prototypiques (Biemann, 2005) qui se caractérisent par la présence d'instances de concepts plutôt que de concepts et d'axiomes ; des catégories sont constituées en collectant de manière extensionelle les instances jugées similaires, les plus représentatives sont ensuite sélectionnées comme prototypes de catégories et sont utilisées pour étiqueter les nœuds supérieurs de la structure lexicale .

D'autre part, les méthodes ensemblistes construisent des ontologies fondées sur la sémantique en structurant les données en "concepts formels", correspondant à des abstractions de concepts issus de la pensée humaine ; ces concepts formels facilitent l'interprétabilité de la structure (Ganter et Wille, 1998). Ces méthodologies peuvent être vues comme des techniques de clustering conceptuel produisant des descriptions intensionnelles de concepts abstraits.

Enfin, les modèles associatifs aboutissent à la construction d'ontologies terminologiques (Biemann, 2005) qui spécifient partiellement des relations sémantiques d'hyperonymie/hyponymie et décrivent chaque concept par une étiquette terminologique ou un ensemble de termes synonymes plutôt qu'une instance prototype (Figure 1). Les ontologies terminologiques les plus connues sont par exemple le thésaurus générique WordNet (Miller et al., 1990) ou l'UMLS [2] spécialisé pour le domaine médical.

Dans cet article, nous proposons d'utiliser le formalisme de la prétopologie pour l'apprentissage d'ontologies terminologiques. À partir d'un ensemble de termes [3] provenant de (sous-) domaines potentiellement différents d'une part et d'un corpus de textes associés d'autre part (corpus spécialisé attesté ou corpus Web), le dégré de généralité/spécificité des termes ainsi que leur proximité sémantique sont évalués au moyen de mesures d'association conduisant à une matrice non-symétrique des proximités. Nous utilisons ensuite le formalisme prétopologique pour structurer l'ensemble de termes en un graphe orienté acyclique (DAG) correspondant à la structure sémantique du/des (sous-) domaine(s) étudié(s). L'algorithme de structuration que

1. Par cette hypothèse, deux termes sont similaires s'ils partagent des contextes linguistiques similaires.
2. http ://www.nlm.nih.gov/research/umls/
3. Nous considérons ici que l'extraction et la structuration terminologique sont deux étapes bien distinctes et indépendantes dans le processus d'acquisition et étudions uniquement l'étape de structuration.

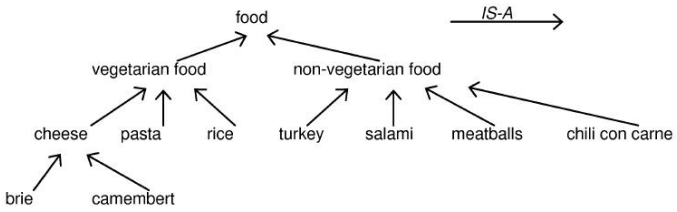

FIG. 1 – *Ontologie terminologique (Biemann, 2005).*

nous présentons fournit un premier cadre de structuration dont nous cherchons dans cette étude à évaluer le potentiel. Nous montrons les connexions qui existent avec d'autres approches de structuration terminologique (Sanderson et Lawrie, 2000) et proposons un cadre d'évaluation des ces approches.

L'article est donc organisé de la manière suivante : la prochaine section précise la classe des méthodes associatives en présentant l'état actuel des recherches dans ce domaine. La Section 3 introduit l'algorithme générique de structuration que nous utiliserons après avoir présenté les notions de prétopologies nécessaires à sa compréhension. Enfin, les deux dernières sections de l'article visent d'une part à analyser le potentiel de l'approche proposée au regard de premiers résultats expérimentaux obtenus sur le domaine médical et d'autre part à présenter différentes pistes d'études susceptibles de déboucher sur des instanciations prometteuses du cadre de structuration proposé.

2 Méthodes associatives

L'objectif visé par ce travail étant la construction d'ontologies terminologiques, nous nous focalisons ici sur les modèles associatifs de Sanderson et Croft (1999), Sanderson et Lawrie (2000) et Dias et al. (2008).

Ces modèles exploitent les distributions des mots dans les documents afin d'en extraire des relations du genre "mot_1 est plus général/spécifique que mot_2" ; ces relations peuvent être assimilées à des relations sous-type/super-type qui constituent la base des ontologies terminologiques en décrivant les positions relatives des concepts entre eux sans pour autant les définir complètement. La méthodologie associative repose sur une définition de la subsomption orientée documents, selon laquelle un terme t_1 et plus spécifique qu'un terme t_2 si t_2 apparaît souvent dans les documents qui contiennent t_1 et si l'inverse n'est pas vérifié. Sanderson et Croft (1999) ont été les premiers à extraire des mots et expressions à partir de documents puis à les organiser de manière hiérarchique en utilisant la subsomption. Dans un premier temps, ils ont considéré pour cela qu'un terme t_1 subsume un autre terme t_2 si l'ensemble des documents contenant t_1 est un sous-ensemble des documents contenant t_2 ; cette contrainte forte d'inclusion stricte a été ensuite relâchée pour devenir une contrainte seuillée sur le degré d'inclusion : $P(t_2|t_1) \geq 0.8$. Les mêmes auteurs généraliseront dans (Sanderson et Lawrie, 2000) la définition de la subsomption via l'expression suivante : $P(t_2|t_1) \geq P(t_1|t_2) \ \wedge \ P(t_2|t_1) > t$ avec t un paramètre de seuil à ajuster[4]. En rassemblant l'ensemble des relations de subsomption

4. Dans (Sanderson et Lawrie, 2000), t est fixé à 0.8

ainsi extraites ils déduisent une structure de graphe orienté que l'on peut aisément montrer être sans cycles (DAG) ; un ultime traitement sur ce graphe consiste à supprimer les transitivités superflues de manière à éviter les subsomptions directes ($t_1 \rightarrow t_3$) lorsqu'il existe un chemin simple ($t_1 \rightarrow t_2 \rightarrow t_3$). La structure finale devient ainsi un DAG non-triangulaire.

Plus récemment, Dias et al. (2008) ont proposé une méthodologie associative visant à dériver un ordre (total) sur l'ensemble des termes à partir des relations de subsomption observées, puis à structurer par clustering cet ensemble en groupes de termes d'un même niveau de généralité. La stratégie d'ordonnancement des termes du plus général au plus spécifique, repose sur le graphe des subsomptions [5] sur lequel est exécuté l'algorithme TextRank (Mihalcea et Tarau, 2004) (inspiré de l'approche plus connue du PageRank) qui associe à chaque nœud du graphe un score de généralité.

Les méthodologies associatives proposées jusqu'ici ont l'avantage considérable de ne reposer que sur les distributions des mots dans les documents, assurant ainsi à l'étape de structuration terminologique, une indépendance totale vis-à-vis de la langue et du domaine. En revanche, ces approches souffrent d'une forte sensibilité au paramétrage (seuils de subsomption, paramètres de clustering, etc.) et de limitations dans l'exploitation des graphes de subsomption construits. Par exemple on peut supposer, et on observera, que le seuil t utilisé par (Sanderson et Lawrie, 2000) pour définir la subsomption entre deux termes n'est pas universel :

1. il dépend de l'adéquation entre le domaine à modéliser et le corpus de documents utilisé (niveau de spécialité et taille du corpus),
2. il varie en fonction du niveau de généralité des termes au sein d'une même structure.

Ces observations constituent la motivation première de la nouvelle méthodologie de structuration que nous proposons dans cette étude. Celle-ci consiste en un algorithme reposant sur des opérateurs génériques issus du formalisme de la prétopologie. Nous montrerons qu'une instanciation simple de ces opérateurs correspond aux approches associatives de Sanderson et Lawrie (2000) mais que cette formalisation ouvre également la voie à toute une série d'instanciations qui permettraient d'exploiter un graphe de subsomption de manière très fine.

3 Méthode de structuration prétopologique

Les relations entre les éléments d'une population peuvent être modélisées de plusieurs manières, par exemple par des arcs en théorie des graphes ou par des voisinages en topologie. Si les travaux mentionnés précédemment (Sanderson et Croft, 1999; Dias et al., 2008) se sont plutôt fondés sur la théorie des graphes, nous choisissons ici de modéliser notre problématique par le formalisme topologique. Ce dernier s'appuie en particulier sur un opérateur de fermeture dont les propriétés (axiomes de fermeture de Kuratowski) peuvent être parfois trop restrictives (e.g. la propriété d'idempotence) ou imprécises (e.g. la propriété d'homomorphisme) pour permettre de modéliser de manière concrète un espace d'objets sur lequel est défini une mesure de proximité [6]. La prétopologie offre un cadre plus favorable à ce type de modélisation par l'introduction d'opérateurs plus précis tels que l'adhérence (*pseudo-closure*) et les notions de fermetures qui en découlent.

5. Obtenu avec une mesure d'association asymétrique quelconque parmi 7 mesures identifiées.

6. Dans l'étude des relations entre termes, la proximité peut ne pas satisfaire les propriétés d'une mesure de similarité telles que la symétrie.

3.1 Notions de prétopologie

Un espace prétopologique est défini par un un ensemble non vide d'éléments E, muni d'un opérateur d'adhérence $a(.)$ (Brissaud, 1975).

Définition 3.1.1 (Adhérence)**.** *Soit E un ensemble non vide et $\mathcal{P}(E)$ l'ensemble des parties de E, un opérateur d'adhérence sur E est une fonction de $\mathcal{P}(E) \longrightarrow \mathcal{P}(E)$ vérifiant les propriétés suivantes :*

- $a(\emptyset) = \emptyset$.
- $\forall A \in \mathcal{P}(E), A \subseteq a(A)$.

Une façon usuelle de structurer un espace prétopologique consiste à choisir une adhérence fondée sur une famille de voisinages. Soit (E, a) un espace prétopologique, on définit une famille de voisinages $\mathcal{N}(x)$ pour tout élément $x \in E$ en choisissant un sous-ensemble de parties de E qui contiennent x :

$$\mathcal{N}(x) = \{N \subseteq E | x \in N\}.$$

On peut ensuite construire l'adhérence d'un sous ensemble $A \subset E$ en considérant les éléments de E dont tous les voisinages intersectent A :

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

Différents types d'espaces prétopologiques ont été caractérisés ; ils diffèrent par des propriétés supplémentaires vérifiées par l'opérateur d'adhérence et permettent d'ajuster la modélisation au domaine d'applications (Belmandt, 1993; Brissaud, 1975). Nous considérerons dans cette étude des espaces prétopologiques de type $\mathcal{V}_S$ vérifiant les propriétés suivantes :

- $\forall A, B \in \mathcal{P}(E), A \subseteq B \Rightarrow a(A) \subseteq a(B)$.
- $\forall A \in \mathcal{P}(E), a(A) = \bigcup_{x \in A} a(x)$.

Les espaces prétopologiques de type $\mathcal{V}_S$ sont intéressants pour l'implémentation d'algorithmes dans la mesure où ils permettent de calculer l'adhérence d'une partie à partir de celle de ses éléments.

L'adhérence ne vérifie pas a priori la propriété d'idempotence ($\forall A \in \mathcal{P}(E), a(a(A)) = a(A)$) propre aux espaces topologiques ; c'est seulement son application répétée qui conduira à un sous-ensemble fermé, comme le montre la Figure 2. Les (sous-ensembles) fermés représentent des sous-ensembles homogènes ou interdépendant relativement à l'adhérence.

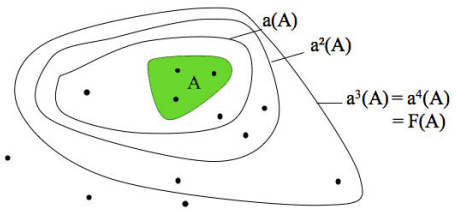

FIG. 2 – *Fermeture d'un sous-ensemble A.*

Définition 3.1.2 (Fermés)**.**
- *un sous-ensemble $F \subseteq E$ tel que $a(F) = F$ est un "fermé" dans l'espace prétopologique (E, a). On note $\mathcal{F}(E, a)$ la famille des fermés de cet espace.*
- *un "fermé élémentaire" F_x correspond à la fermeture d'un élément singleton $\{x\}$ de E. On note $\mathcal{F}e(E, a)$ la famille des fermés élémentaires : $\mathcal{F}e(E, a) = \{F_x | x \in E\}$.*
- *un "fermé élémentaire maximal" de (E, a) est un élément maximal de $\mathcal{F}e(E, a)$ au sens de l'inclusion. La famille des fermés élémentaires maximaux sera notée $\mathcal{F}M(E, a)$.*

La nature des fermés (élémentaires) maximaux et la structuration de l'espace qu'ils induisent est intéressante à exploiter. Nous proposons dans ce qui suit un algorithme générique de structuration d'un espace prétopologique, dont la généricité repose sur la liberté de définition de l'adhérence.

3.2 Algorithme de structuration

L'algorithme que nous proposons est une version descendante de celui proposé par Largeron et Bonnevay (2002) et procède en recherchant d'abord les éléments "centraux" de l'espace c'est-à-dire pour lesquels les fermés élémentaires associés sont maximaux, puis par raffinements successifs sur le même espace privé des éléments centraux déjà identifiés. Plus précisément, l'algorithme produira une structure ordonnée d'éléments de $\mathcal{P}(E)$ où chaque étape de raffinement consiste à :

1. Déterminer l'ensemble des fermés élémentaires $\mathcal{F}e(E, a)$,
2. Identifier ceux qui sont maximaux $\mathcal{F}M(E, a)$,
3. Et pour chaque fermé élémentaire maximal F,
   - retenir le(s) élément(s) générateur(s) $G(F) = \{x \in E | F_x = F\}$,
   - considérer $G(F)$ comme un élément de la structure ordonnée dont les descendants seront construits par raffinement du sous-espace $(F \setminus G(F), a)$

Nous présentons le pseudo-code de la méthode de structuration globale dans l'Algorithme 1 :

Algorithm 1 Algorithme de structuration descendante.

```
Method structure((E,a) : set)
vars : FN : family, FM : family
(note : pas de doublons dans une famille)
Begin
  FN = Fe(E,a) − FM(E,a)
  FM = FM(E,a)
  while FM ≠ ∅ do
    take F of FM
    remove F of FM
    successor(F)
  end while
  return extracted_structure
End
```

```
Method successor(F : set)
vars : FF : family
Begin
  FF = {G ∈ FN|/G ⊂ F}
  if FF ≠ ∅ then
    FM_F = MaxClosedSubsets(FF)
    for each V ∈ FM_F do
      V is a successor of F
      successor(V)
    end for
  end if
End
```

Nous illustrons le fonctionnement de l'algorithme sur un exemple jouet dont les fermés élémentaires sont listés dans le Tableau 1 et le détail de la structuration présenté en Figure 3.

$x \in E$	F_x	$x \in E$	F_x
1	{1,2,3,4,5,6}	7	{7,8}
2	{1,2,3,4,5,6}	8	{7,8}
3	{1,2,3,4,5,6}	9	{4,5,6,7,8,9}
4	{4,5,6}	10	{10}
5	{4,5,6}	11	{1,2,3,4,5,6,7,8,9,10,11}
6	{4,5,6}		

TAB. 1 – *Exemple de fermés élémentaires.*

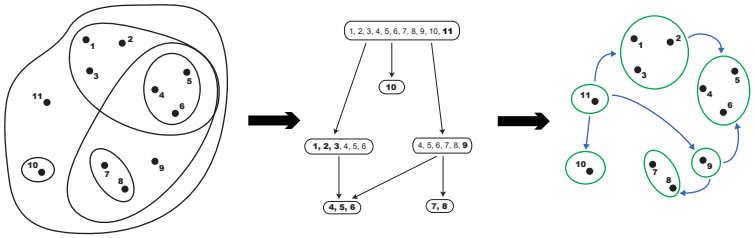

FIG. 3 – *Détail de la structuration par l'algorithme proposé.*

3.3 Application à la structuration lexico-sémantiques

L'utilisation du formalisme prétopologique pour la tâche d'acquisition d'une structure lexico-sémantique consiste à définir un espace prétopologique à partir des termes à structurer E, puis à appliquer l'algorithme défini précédement sur cet espace. La difficulté réside alors dans la définition de l'opérateur d'adhérence qui est l'élément essentiel de modélisation des relations de voisinages entre les termes. Nous proposons deux formes d'adhérences : fixe et dynamique. L'adhérence fixe correspond à une définition traditionnelle de l'opérateur prétopologique et repose sur les mesures d'association asymétriques décrites précédemment ; en revanche la forme dynamique correspond à une adhérence paramétrée susceptible d'évoluer au cours de l'algorithme de structuration.

3.3.1 Adhérence fixe

A partir d'un corpus représentatif d'un domaine dont on cherche à établir une ontologie terminologique, on peut extraire facilement - à la manière de Sanderson et Lawrie (2000) - une mesure d'association asymétrique de type probabilité conditionnelle (ou *confidence*). Cette mesure reflète l'attraction sémantique d'un terme x vers un terme y de telle sorte qu'une valeur élevée pour $P(y|x)$ (noté $P_{y,x}$ dans la suite) révèle une relation de subsomption dans le sens y *subsume* x. Ainsi une définition naturelle de l'adhérence correspond à la relation de subsomption de Sanderson et Lawrie (2000) :

$$\forall A \subseteq E, a(A) = A \cup \{x \in E | \exists y \in A \; t.q. \; P_{y,x} \geq P_{x,y} \wedge P_{y,x} \geq \epsilon\} \quad (1)$$

Cette modélisation présente peu d'intérêt dans la mesure où l'on peut montrer qu'elle conduit, après structuration par l'algorithme, au graphe des subsomptions duquel sont supprimés tous les arcs transitifs directs ou non.

Nous formalisons en (2) une seconde définition de l'adhérence qui consiste en quelquesorte à autoriser une "double subsomption" [7] dans le cas où deux termes s'attirent mutuellement *i.e.* dont les deux mesures d'associations ($P_{x,y}$ et $P_{y,x}$) sont élevées et leur variance faible (relativement à leur moyenne).

$$\forall A \subseteq E, a(A) = A \cup \{x \in E | \exists y \in A \; t.q. \; P_{y,x} \geq \epsilon \wedge \{P_{y,x} \geq P_{x,y} \vee \frac{var(P_{x,y}, P_{y,x})}{moy(P_{x,y}, P_{y,x})^2} \leq \alpha\}\} \quad (2)$$

Cette modélisation présente cette fois l'intérêt de permettre l'apparition d'éléments non-uniquement singletons dans la structure finale ; l'ontologie terminologique possédera alors des concepts extensionels constitués de plusieurs termes se rapprochant ainsi par exemple de la notion de *synsets* utilisée dans le thésaurus WordNet.

La forme fixe de l'opérateur d'adhérence nécessite de déterminer les paramètres (ϵ et α) a priori, avec beaucoup de précision, sous peine d'obtenir une structure finale : de très faible connexité (pour ϵ trop élevé), de trop grande profondeur (pour ϵ trop faible) ou avec peu d'éléments (pour α trop élevé).

3.3.2 Adhérence dynamique

La forme dynamique de l'adhérence trouve ses motivations dans l'observation que les valeurs d'association sont globalement plus faibles lorsque l'on compare des termes assez généraux que lorsque l'on compare des termes spécifiques. De façon plus générale, on aimerait pouvoir adapter localement le paramétrage des opérateurs de manière à exploiter au mieux les sous-espaces prétopologiques considérés successivement par l'algorithme de structuration. Nous pouvons par exemple considérer une adhérence dynamique $a_{\epsilon,\alpha}(.)$ sur la base de celle définie en (2). L'ajustement des paramètres de l'opérateur devient alors une tâche déterminante pour l'algorithme. On peut envisager toute une série de méthodologies d'ajustement telles que :

- la proposition d'heuristiques : recherche de valeurs médianes dans chaque sous-espace, proposition de lois de probabilités pour les paramètres, etc.
- la définition de critères objectifs (à optimiser) : critères non-supervisés structuraux ou bayésiens,
- la définition de critères semi-supervisés permettant d'intégrer des connaissances expertes (e.g. satisfaction de contraintes)

Ces méthodologies d'ajustement sont autant de pistes à explorer sous l'hypothèse que cette forme d'adhérence dynamique renferme un réel potentiel. L'objectif des premières expérimentations que nous présentons dans la suite est donc d'évaluer le potentiel de l'algorithme de structuration couplé avec un forme dynamique d'adhérence.

7. On parlerait plutôt de relation d'équivalence.

4 Étude expérimentale

Les résultats d'acquisition automatique d'ontologies sont par nature difficiles à évaluer autrement que par l'intervention experte humaine. Néanmoins, lorsque l'on dispose d'une ontologie de référence il est toujours possible de définir des mesures pour évaluer quantitativement la correspondance entre une structure de référence et un structure acquise (Maedche et Staab, 2002). Aussi audacieuses que puissent être les interprétations données aux résultats obtenus avec ce type de mesure, nous l'utiliserons comme indicateur d'une certaine tendance à se rapprocher ou au contraire à s'éloigner d'une ontologie de référence construite manuellement.

Pour cette évaluation nous utiliserons la ressource UMLS (*Unified Medical Language System*) produite par l'institut Américain de la santé et qui propose entre autre une ressource terminologique structurant les termes du domaine médical dans une ontologie terminologique ; pour cette première étude expérimentale nous nous limiterons à l'étude du sous domaine *cardiovascular system* composé d'une vingtaine de termes. Le corpus PubMed [8] sera utilisé pour extraire les relations de subsomption ; ce corpus est géré par le même institut et contient plus de 17 millions de références scientifiques dans le domaine biomédical.

Nous utilisons la mesure LTO (*Local Taxonomic Overlap*) proposée par Maedche et Staab (2002) afin d'évaluer la correspondance entre ontologies terminologiques. Cette mesure consiste, pour chaque termes, à comparer leurs parentés (ascendants ou descendants) respectives dans les deux structures par un indice du type Jaccard. Observant que cette mesure produit un score maximal de correspondance entre deux ontologies totalement inversées (en terme de subsomptions), nous utiliserons une version légèrement améliorée du LTO - que nous ne détaillons pas ici - qui compare pour chaque terme les ensembles d'ascendants d'une part puis les descendants d'autre part.

La Figure 4 présente les résultats d'une première étude empirique visant à évaluer l'algorithme de structuration de l'espace prétopologique défini avec une adhérence de forme fixe de type 2. Nous avons balayé l'espace des paramètres, chacun discrétisé en déciles, et observé pour chaque paramétrage, la correspondance de la structure générée et la structure UMLS de référence (Figure 5).

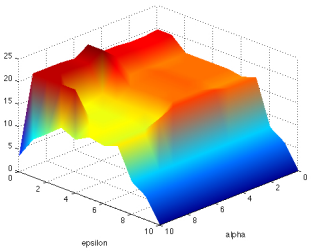

FIG. 4 – *Évolution du score LTO en fonction des paramètres de l'adhérence (fixe).*

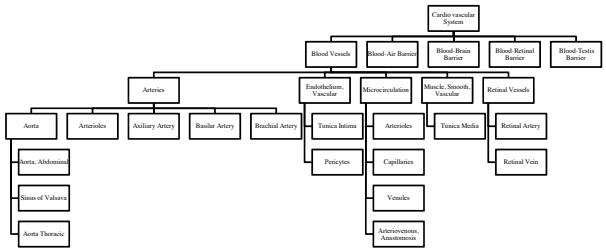

FIG. 5 – *Ontologie UMLS du sous-domaine cardiovascular system.*

8. http ://www.ncbi.nih.gog/pubmed

Les résultats obtenus montrent d'abord une faible influence du paramètre α qui s'explique par le fait que l'ontologie de référence dans l'UMLS n'autorise pas les concepts extensionels. Le paramètre ϵ joue quant à lui un rôle plus déterminant dans la structuration ; deux paramétrages (0 et 7ème déciles) permettent d'obtenir une correspondance sensiblement meilleure. Le décile 0 correspond à une valeur nulle de ϵ, il conduit à conserver toutes les relations de subsomption se traduisant par une structure très profonde (de type chaînage) et par effet de bord à un score de LTO élevé. En revanche le 7ème décile correspond à une sélection de seulement 30% des subsomptions les plus fiables et semble être un paramétrage plus naturel [9]. De manière globale, quelquesoit son paramétrage, la forme fixe de l'adhérence de type (2) permet difficilement de dépasser un score LTO de 0.20.

La seconde expérience consiste cette fois à utiliser une adhérence dynamique de type (2) pour structurer l'espace prétopologique associé. Afin d'évaluer le potentiel de ce formalisme, nous choisissons à titre expérimental une méthodologie d'ajustement supervisée. Plus précisément, à chaque étape de raffinement de l'algorithme de structuration, on recherche de manière exhaustive le paramétrage (ϵ, α) qui génère la meilleure correspondance avec l'ontologie de référence, au sens de la mesure LTO. Cette méthodologie n'est biensûr pas envisageable en situation réelle - où une telle référence ne sera pas disponible - mais elle présente l'intérêt de révéler dans quelle mesure le formalisme proposé pourrait permettre d'améliorer la structuration des termes ; il s'agira en quelque sorte d'une borne supérieure de scores envisageables.

La Figure 6 montre l'ontologie (a) obtenue par cette méthode ainsi que l'ontologie obtenue par l'approche de Sanderson et Lawrie (2000) (b) avec le meilleur paramétrage (7ème décile).

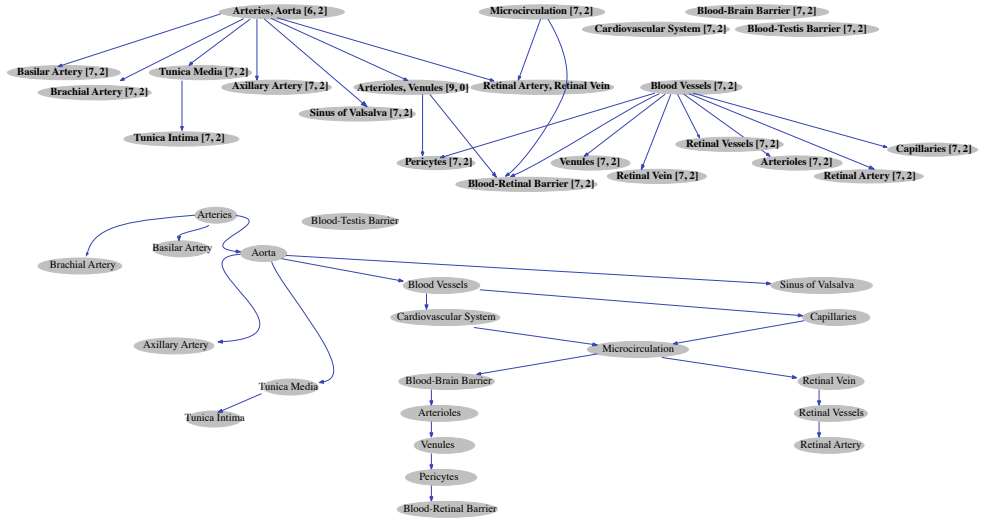

FIG. 6 – *(En haut - a) Ontologie acquise par une structuration prétopologique dynamique, (en bas - b) ontologie obtenue par Sanderson et Lawrie (2000).*

On observe une structuration très différente pour les deux ontologies acquises automatiquement, l'approche prétopologique a permis par une adhérence dynamique d'adapter la topologie de l'espace à chaque nœud de la structure via un paramétrage (ϵ, α) indiqué dans

9. Des expérimentations du même type sur d'autres sous-domaines de l'UMLS montrent également ce paramétrage comme remarquable.

chaque concept ; la structure globale est plus proche de celle de référence (faible profondeur, beaucoup de fils par concept) ce qui peut naturellement s'expliquer par l'ajustement supervisé des paramètres. L'objectif était cependant de montrer qu'une telle structure est potentiellement "atteignable" sans pour autant remettre en cause l'information élémentaire - commune aux deux méthodologies d'acquisition comparées ici - à savoir les relations subsomption, mais uniquement en exploitant différemment cette information de base. La valeur finale du score LTO obtenu pour l'approche prétopologique est de 0.37, tandis que la structure générées par l'approche de Sanderson et Lawrie (2000) reste à un niveau de 0.20.

5 Conclusion

Nous nous sommes concentrés dans cette étude sur la tâche de structuration d'un ensemble de termes d'un domaine, sur la base d'observations sur leur utilisation dans un corpus, de manière à extraire automatiquement une ontologie terminologique du domaine considéré. Nous nous sommes inspirés pour cela d'approches de structuration dites associatives qui reposent sur la définition d'une relation de subsomption. Plutôt que de considérer cette relation comme une information terminale de structuration, nous avons choisi de l'utiliser comme un point de départ et avons proposé une formalisation prétopologique de l'espace des termes. Cette formalisation permet de définir un algorithme de structuration de cet espace conduisant à une structure de DAG non-triangulaire. L'algorithme proposé doit être vu comme un modèle générique de structuration qui, par des instanciations simples (e.g. adhérence de type (1)), se ramènent à des structures connues (e.g. le graphe de subsomption seuillé de Sanderson et Lawrie (2000)) mais qui offre également un potentiel de structuration bien plus important en considérant par exemple une forme dynamique des opérateurs structurant l'espace.

Nous avons tenté dans une étude empirique d'évaluer ce potentiel ce qui nous a conduit à des résultats assez significatifs qui mériteront d'être confirmés à plus large échelle mais qui motivent dores et déjà les recherches à venir sur les pistes évoquées pour l'ajustement des opérateurs dynamiques. Enfin, il conviendra de diffuser rapidement une première version logicielle de cette méthodologie d'extraction de connaissances, susceptible d'intervenir dans de nombreux processus de fouille de textes.

Références

Belmandt, Z. (1993). *Manuel de Prétopologie et ses Applications*. Hermes Sciences Publications.

Bhogal, J., A. Macfarlane, et P. Smith (2007). A review of ontology based query expansion. *Information Processing & Management 43*(4), 866–886.

Biemann, C. (2005). Ontology learning from text – a survey of methods. *LDV-Forum 20*(2), 75–93.

Brissaud, M. (1975). Les espaces prétopologiques. *Compte-rendu de l'Académie des Sciences 280*.

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.

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.

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

Harris, Z. (1968). *Mathematical structures of language*. Wiley.

Hotho, A., S. Staab, et G. Stumme (2003). Ontologies improve text document clustering. *Data Mining, IEEE International Conference on 0*, 541.

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 *EKAW '02 : 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 EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, et K. J. Miller (1990). Introduction to wordnet : An on-line lexical database. *Int J Lexicography 3*(4), 235–244.

Mylonas, P., D. Vallet, P. Castells, M. Fernandez, et Y. Avrithis (2008). Personalized information retrieval based on context and ontological knowledge. *Knowledge Engineering Review 23*(1), 73–100.

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.

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.

Summary

Lexical-Semantic structures play an essential role in text mining processes. By coding the semantic relations between concepts of discourse, they can enrich the reasoning capabilities. The development of such structures being largely limited by the efforts required for their construction, we propose a new formalism for the automatic acquisition of a terminological ontology from texts. We introduce a pretopological formalization of the term space and define a generic structuring model. We present a first empirical study that reveals the potential of the proposed approach in term of knowledge extraction.