Fouiller un corpus en structurant sa terminologie

Olivier Ferret, Christian Fluhr, Françoise Rousseau-Hans, Jean-Luc Simoni

Fouiller un corpus en structurant sa terminologie

Olivier FERRET, Christian FLUHR, Françoise ROUSSEAU-HANS, Jean-Luc SIMONI
olivier.ferret@cea.fr , christian.fluhr@cea.fr , francoise.rousseau@cea.fr , jlsimoni@netcourrier.com

CEA/DTI, 91191 Gif-sur-Yvette Cedex

Mots clés :
Text mining, traitement automatique des langues, extraction terminologique, structuration terminologique, arbre de concepts

Keywords:
Text mining, natural language processing, term extraction, term structuring, concept trees

Palabras claves :
Text mining, tratamiento lenguaje natural

Résumé

L'exploitation d'un grand ensemble de textes, dans le contexte d'une veille technologique par exemple, passe souvent par l'utilisation d'un logiciel de Text Mining. L'objet de cet article est de présenter un nouvel algorithme pour la construction de représentations simplifiées de contenus textuels. La simple notion de cooccurrence entre termes y est dépassée pour mettre en évidence des relations de nature hiérarchique, traduisant des différences de niveau de généralité au sein de la terminologie extraite d'un ensemble important de textes. Cet outil fait appel à un pré-traitement linguistique des textes permettant de dégager leur terminologie sous une forme normalisée, à la fois composée de termes simples et de multi-termes.

1 Introduction

Un des principaux objectifs de la fouille d'un ensemble de textes est d'en faire émerger les notions les plus significatives ainsi que les relations existant entre ces notions. La réalisation du premier aspect s'appuie à la base sur des extracteurs terminologiques robustes du type de LEXTER 0, ACABIT 0 ou ANA 0. Ces outils extraient des candidats termes de façon large. Leur application doit donc être suivie d'un traitement permettant à l'expert du domaine de ne pas être submergé par un volume manuellement ingérable de termes. Ce traitement peut prendre la forme d'un filtrage, souvent réalisé selon des critères de fréquence.

Une autre façon de faciliter l'accès à l'information est de structurer les termes issus de l'extraction en fonction de leurs relations. Ce second aspect de la fouille textuelle doit permettre d'établir des liens de nature diverse entre les termes extraits. Le type de lien le plus évident est la relation d'équivalence, caractérisée en pratique par un critère de proximité assez stricte : on rassemble les termes qui sont des variantes les uns des autres, variantes pouvant être de nature morphologique, syntaxique, voire sémantique[1] 0. Ces regroupements s'appuient sur des outils tels que FASTR 0 ou bien SynoTerm 0, capables de reconnaître les variantes d'un terme. Le module de structuration de LEXTER, liant les termes partageant une même tête ou une même expansion, s'inscrit dans la même logique.

Pousser plus avant la structuration des termes oblige à élargir la nature des liens de proximité sémantique considérés. Suivant les approches adoptées, la nature de cette proximité est bien identifiée – les termes sont par exemple liés par une relation d'hyperonymie – ou bien elle reste indéterminée. Deux grandes approches se dégagent concernant la structuration 0 d'un ensemble de termes issus d'un corpus, la première basée sur la reconnaissance d'une série de formes spécifiques et la seconde sur des critères statistiques.

Dans la première approche, le système repère des patrons de reconnaissance liés à un type de relation spécifique. Par exemple, dans la phrase « le chat est un félin », il est possible d'extraire la relation d'hyperonymie entre le félin et le chat grâce au patron « SN1 est un SN2 » (SN étant la contraction de « *syntagme nominal* »). Cette approche est donc caractérisée par une grande précision des relations trouvées. En revanche, elle est peu productive dans la mesure où les marques linguistiques sur lesquelles reposent les patrons de reconnaissance se manifestent généralement avec une fréquence peu élevée. Dans le cas de la relation d'hyperonymie, qui retient plus particulièrement notre attention ici, elle est représentée par des travaux comme ceux de Hearst 0, Jouis 0 ou Morin 0.

Dans l'approche statistique, chaque terme est caractérisé par ses contextes d'occurrence et le regroupement des termes s'effectue sur la base des proximités entre ces contextes. Selon les moyens d'analyse utilisés, la construction de tels contextes se fait sur la base de simples cooccurrences entre termes ou bien s'appuie sur des relations syntaxiques[2]. Les caractéristiques de cette approche sont complémentaires des caractéristiques de la première présentée : le type des relations mises à jour est plus indéterminé mais la productivité des méthodes utilisées est supérieure, à condition toutefois que le corpus considéré soit de taille suffisante. Cette seconde approche s'incarne au travers de systèmes tels que SEXTANT 0, LEXICLASS 0 et ZELLIG 0 ou dans un système tel que SAMPLER 0 dans le domaine plus spécifique de la veille technologique.

Le travail que nous présentons dans cet article relève de cette seconde approche. Plus précisément, il prend comme point de départ la notion de cooccurrence entre termes pour la dépasser et mettre en évidence des relations de nature hiérarchique, traduisant des différences de niveau de généralité, au sein de la terminologie extraite d'un ensemble important de textes. Cet outil fait appel à un pré-traitement linguistique des textes permettant de dégager leur terminologie sous une forme normalisée, à la fois composée de termes simples et de multi-termes.

2 Principes

La méthode de hiérarchisation décrite ici repose sur une hypothèse distributionnaliste du sens des termes : le sens d'un terme dans un corpus peut être caractérisé par l'ensemble des contextes

[1] On peut également être confronté à des variantes mixtes.

[2] On peut choisir par exemple de ne retenir que les mots faisant partie du même syntagme nominal ou bien ne sélectionner que les cooccurrences entre un verbe et son sujet ou son complément d'objet.

d'occurrence de ce terme dans le corpus. Dans le cas présent, le contexte d'occurrence d'un terme est constitué de l'ensemble des termes qui cooccurrent avec lui dans le cadre d'un paragraphe, le paragraphe étant considéré comme une unité textuelle significative du point de vue de la représentation du sens d'un terme.
L'ensemble des contextes d'occurrence d'un même terme enregistrés sur un corpus sont cumulés afin de former l'Ensemble Sémantique (ES) associé à ce terme. Du point de vue de l'hypothèse énoncée précédemment, l'ES d'un terme caractérise donc son sens. Le processus de hiérarchisation, qui, au-delà des termes, vise à s'appuyer sur leur sens, transpose la différence de généralité entre deux termes en une relation d'inclusion entre leur ES. Le principe général de la méthode de hiérarchisation présentée peut donc s'énoncer ainsi : un terme T1 est le générique d'un terme T2 si l'ES de T2 est inclus dans celui de T1.
La concrétisation de ce principe, développée initialement dans 0, puis dans 0, est réalisée au travers de trois grandes étapes : la première vise à mettre en évidence les notions importantes du corpus considéré ; la deuxième assure la construction de l'ensemble sémantique caractérisant, du point de vue de ce corpus, le sens de chaque notion retenue ; enfin, la troisième réalise la hiérarchisation proprement dite de ces notions. Cette dernière étape inclut la prise en compte du problème de la polysémie des termes, traité dans le même cadre que la hiérarchisation grâce à la partition des ensembles sémantiques des termes.

3 Extraction des notions importantes d'un corpus

La mise en évidence des notions importantes d'un corpus s'effectue dans le cas présent par une extraction de termes assez large, incluant une normalisation, suivie d'une sélection. Pour les unitermes, cette extraction repose sur l'étiqueteur morpho-syntaxique et le lemmatiseur associés au logiciel de recherche d'information SPIRIT 0. Pour les multi-termes, elle s'effectue par l'intermédiaire d'un extracteur de terminologie conjuguant classiquement un chunker et un ensemble de patrons lexico-syntaxiques 0. Le chunker a pour rôle de découper les phrases en grands groupes, appelés chaînes. On distingue plus précisément les chaînes nominales et les chaînes verbales. Les patrons lexico-syntaxiques[3] sont ensuite utilisés pour extraire les termes au sein des chaînes. L'extracteur opère lui-même à partir des résultats de l'analyse réalisée pour les unitermes.
La sélection des termes les plus significatifs se fonde à la fois sur leur type – seuls les termes nominaux sont retenus – et sur des critères statistiques : un terme n'est sélectionné que si sa fréquence dans le corpus est suffisamment élevée et s'il apparaît dans un nombre suffisamment important de documents. Ces deux critères sont en pratique représentés par des seuils fixés de façon expérimentale et constituent donc des paramètres.

4 Constitution des Ensembles Sémantiques des termes

En accord avec la définition de la notion d'Ensemble Sémantique (ES) donnée au §2, la première phase de leur constitution consiste à enregistrer pour chaque terme sélectionné l'ensemble des termes sélectionnés qui cooccurrent avec lui dans l'espace d'un paragraphe et ce, pour tous les paragraphes du corpus où ce terme apparaît. Les ES ainsi obtenus sont ensuite filtrés afin d'en éliminer les mots supposés les moins liés avec le terme de référence auquel ils sont associés. Le sens d'un terme étant représenté par ses différentes occurrences, l'évaluation de la force de cette liaison repose sur l'importance du nombre de fois où le terme de référence et le terme considéré cooccurrent. On fait plus précisément appel à l'indice d'inclusion 0 :

$$I(Ti \rightarrow Tj) = \frac{coocc(Ti, Tj)}{occ(Ti)}$$

où $I(Ti \rightarrow Tj)$ est l'indice d'inclusion du terme Ti dans le terme Tj, i.e. la proportion des occurrences de Ti se caractérisant par une cooccurrence avec une occurrence de Tj. Pour être retenu, le terme T_{ES} de l'ES du terme T_{ref} doit remplir la double condition suivante : $I(T_{ES} \rightarrow T_{ref}) > S1$ et $I(T_{ref} \rightarrow T_{ES}) > S2$, où

[3] Un patron lexico-syntaxique est simplement une suite de catégories morpho-syntaxiques et/ou de mots. Exemple : ARTICLE NOM ADJECTIF.

S1 et S2 sont là encore des seuils fixés expérimentalement. Ces deux conditions permettent de s'assurer ce qui lie les deux termes est suffisamment représentatif de leur sens.

5 Hiérarchisation des termes

5.1 Méthode

La construction proprement dite de la hiérarchie de termes[4] s'effectue de façon itérative et descendante, donc à partir des termes les plus généraux. Son principe général est le suivant : à chaque itération de l'algorithme, on définit un ensemble de termes racines parmi les termes restant à classer et l'on effectue ensuite le rattachement (relation de type générique/spécifique) de ces racines aux termes déjà présents dans la hiérarchie. Les termes racines représentent les termes pour lesquels il n'est pas possible de trouver un père parmi les termes restant à classer. Les termes racines de plus haut niveau, i.e. les racines de la hiérarchie, sont soit fournis *a priori* (directement par l'utilisateur ou en faisant appel à un autre critère d'intérêt), soit issus d'une première application de l'algorithme de détermination des termes racines. Globalement, les itérations se poursuivent tant que l'ensemble des termes restant à classer n'est pas vide. La décroissance de cet ensemble intervient à l'issue de chaque itération, lorsque les termes nouvellement rattachés à la hiérarchie en sont supprimés.
Chaque itération de l'algorithme de construction de la hiérarchie de termes se décompose plus précisément en quatre étapes :

1. choix, parmi les termes restant à classer, des termes racines que l'on va rattacher à la hiérarchie ;
2. détermination des génériques potentiels pour chacun des termes racines. Ces génériques représentent ceux des termes ayant été précédemment rattachés à la hiérarchie susceptibles d'être le père (ou encore le générique) du terme racine considéré ;
3. choix du générique de chacun des termes racines (cf. 5.2) ;
4. rattachement des termes racines à la hiérarchie de termes (cf. 5.3).

La présence de l'étape 2 est motivée par des raisons de complexité algorithmique : les critères utilisés à l'étape 3 pour choisir le générique d'un terme racine sont en effet trop coûteux pour être appliqués à l'ensemble des termes déjà hiérarchisés. La détermination des génériques potentiels d'un terme racine permet d'écarter rapidement les termes qui en sont visiblement trop éloignés. Elle consiste à énumérer l'ensemble des termes déjà hiérarchisés et à ne retenir que ceux remplissant les deux conditions suivantes :

- la taille de l'ES du terme déjà hiérarchisé doit être supérieure à celle de l'ES du terme racine ;
- le terme déjà hiérarchisé doit faire partie de l'ES du terme racine.

Cette recherche des génériques potentiels est utilisée non seulement comme filtre dans le choix du générique d'un terme racine mais également pour la sélection des termes racines réalisée à l'étape 1. En pratique, un terme racine est en effet défini comme un terme n'ayant pas de générique potentiel parmi les termes déjà hiérarchisés.

5.2 Choix du générique d'un terme racine

Ainsi que nous l'avons évoqué à propos des principes, la recherche du terme *Tg* susceptible d'être le père d'un autre terme *Ts* dans la hiérarchie est fondée sur des critères de recouvrement entre l'ES de *Tg* et celui de *Ts* : l'ES de *Tg* doit recouvrir celui de *Ts* de façon significative mais cette intersection ne doit pas être trop petite par rapport à l'ES de *Tg*. Le cas contraire serait en effet le signe d'un écart de niveau de généralité trop important entre *Tg* et *Ts* pour qu'ils soient liés directement.
Plus formellement, ces critères s'expriment en faisant appel au coefficient de recouvrement des ES. Ainsi, le coefficient de recouvrement de l'ES du terme *Ti* par celui du terme *Tj* est donné par :

$$C(Ti \rightarrow Tj) = \frac{card(ES(Ti) \cap ES(Tj))}{card(ES(Ti))}$$

[4] Nous parlerons dans ce qui suit de hiérarchie de termes pour désigner le résultat de notre algorithme de structuration. Plus formellement, il s'agit d'une forêt d'arbres puisque ses racines peuvent être multiples.

Le premier critère énoncé ci-dessus conduit à ne retenir que les termes *Tg* tels que $C(Ts \rightarrow Tg) > S3$. Le second sélectionne les termes *Tg* tels que $C(Ts \rightarrow Tg) > S4$. Comme les seuils précédents, S3 et S4 sont des paramètres déterminés de manière expérimentale.
Si l'ensemble des génériques potentiels ainsi filtré comporte plus d'un seul élément, on choisit le générique dont le coefficient de recouvrement mutuel avec le terme racine considéré est maximal. Pour deux termes *Ti* et *Tj*, ce coefficient est donné par le produit suivant :

$$C(Ti, Tj) = C(Ti \rightarrow Tj) \times C(Tj \rightarrow Ti)$$

À l'issue de cette troisième étape, chaque terme racine possède un point de rattachement à la hiérarchie de termes.

5.3 Rattachement des termes à la hiérarchie

Le rattachement d'un terme *T* à un terme père s'accompagne d'une adaptation de son ES caractérisant le fait que *T* se trouve dans le contexte de son terme père. Cette adaptation se justifie par la nature intrinsèque des ES. L'ensemble des termes qui les constituent recèle une certaine hétérogénéité. Deux facteurs y contribuent fortement :

- la cooccurrence de deux termes dans une même unité sémantique, ici le paragraphe, n'est que partiellement indicatrice de la nature du lien qui les unit : ce lien peut être de nature sémantique, ce qui est avant tout recherché, mais il peut être également de nature syntaxique, voire être le résultat d'une simple coïncidence, même si le filtrage du vocabulaire significatif du corpus ainsi que le filtrage des ensembles sémantiques contribuent à réduire sensiblement ce cas de figure ;
- un terme est soumis au phénomène de la polysémie (plusieurs sens pour un terme)[5]. Ce constat est déjà réalisable en langue générale. Mais la polysémie prend plus d'importance encore dans le cadre d'une définition du sens des termes par rapport à un corpus, comme c'est le cas ici. La pluralité des sens a en effet tendance à s'accroître car les contextes sont plus précis.
  Or, les cooccurrences recueillies sont des cooccurrences entre termes et non entre sens de termes. Même si un terme possède un sens majoritaire dans un corpus, il est très rare qu'il soit unique. Les cooccurrences concernant ce sens majoritaire se mêlent donc aux cooccurrences recouvrant ses sens minoritaires, ce qui contribue à l'hétérogénéité des ES. Le problème est encore plus important lorsque cohabitent dans un même corpus plusieurs sens d'un même terme dans des proportions relativement équilibrées. Ce problème sera plus spécifiquement abordé au §6.

L'algorithme de hiérarchisation des termes participe néanmoins à la réduction de l'hétérogénéité des ES. Lorsqu'un terme est rattaché à la hiérarchie de termes, son ES est mis à jour en fonction du générique qui lui a été choisi. L'ES résultant est appelé *Ensemble Sémantique en Contexte* (ESC) du terme[6].
Pour être plus précis, l'ESC d'un terme est le résultat de l'intersection de l'ES initial de ce terme et de l'ESC du générique auquel il se trouve rattaché. Cette opération apporte dans une certaine mesure une réponse face aux deux sources d'hétérogénéité des ES mentionnées ci-dessus :

- les mots dont la présence dans l'ES initial d'un terme peut être considérée comme contingente ont par définition peu de chances de se retrouver parmi les mots de l'ESC du générique auquel ce terme est rattaché. Ils sont donc éliminés par l'intersection des deux ensembles sémantiques ;
- cette opération permet également de filtrer les mots dont la présence dans l'ES initial d'un terme correspond à des cooccurrences impliquant des sens de ce terme minoritaires dans le corpus et différents du sens incarné par le générique de rattachement de ce terme. Ces mots ont en effet moins de chances de se trouver dans l'ESC du générique que dans l'ES initial du terme.

De par leur définition, les ESC sont de plus en plus précis à mesure que l'on s'enfonce dans les niveaux de la hiérarchie de termes. Après qu'un terme a été rattaché à la hiérarchie, l'ES que l'on prend en compte pour représenter ce terme au niveau de l'algorithme de hiérarchisation n'est plus en

[5] Le problème de l'homonymie (un mot renvoyant à des sens sans relation sémantique) se pose également pour les unitermes de la même façon.
[6] La notion de contexte fait référence ici au générique auquel le terme a été rattaché.

effet son ES initial mais son ESC. Au niveau le plus haut, c'est-à-dire le niveau des racines de la hiérarchie de termes, il y a par construction identité entre les ESC et les ES initiaux.

6 Traitement de la polysémie

Même dans un corpus assez homogène, un terme peut admettre des sens suffisamment différents pour qu'il soit intéressant de les faire apparaître, en particulier dans la perspective de la fouille d'un corpus. La détection et le traitement de ce phénomène sont réalisés dans le prolongement de la méthode de hiérarchisation présentée. À la suite du rattachement d'un terme *T* à un terme père, on teste si la partie de l'ES initial de *T* restant après différence entre l'ES initial de *T* et son ES en contexte peut être rattachée à un terme de la hiérarchie. Dans l'affirmative, on vérifie également que l'ES du terme père et la partie de l'ES de *T* déjà recouverte par les termes pères des sens précédemment distingués pour *T* ne sont pas trop proches. Une telle proximité signifierait en effet que le nouveau sens que l'on cherche à différencier est en fait très proche des sens déjà distingués et que donc, une telle différenciation ne se justifie pas. En pratique, cette condition est vérifiée en comparant à un seuil, S5, le coefficient de recouvrement mutuel de l'ES du générique trouvé et de la partie de l'ES de *T* déjà recouverte (cf. 5.2). Si ce coefficient est supérieur au seuil fixé, un nouveau sens est créé, avec comme ESC le résultat de l'intersection de la partie restante de l'ES de *T* et de l'ESC du générique trouvé. La création de nouveaux sens pour le terme *T* se poursuit ainsi itérativement jusqu'à ce que les conditions de rattachement de la partie résiduelle de l'ES initial de *T* à un terme de la hiérarchie ne puissent plus être remplies. Le seuil S5 contrôle ainsi directement le taux de polysémie que l'on accepte et donc, la finesse des distinctions de sens opérées au niveau de la hiérarchie de termes construite.

7 Résultats et discussion

7.1 Résultats

L'algorithme que nous avons décrit précédemment a été implémenté dans le cadre d'un prototype développé par Jean-Luc Simoni dans le cadre de sa thèse. Celui-ci a été testé sur un corpus composé de 19 numéros de la revue en sciences de l'éducation SPIRALE, chaque numéro représentant environ 200 pages de texte. Ce corpus est issu des travaux de l'Action de Recherche Concertée (ARC) A3 du réseau francophone de l'ingénierie de la langue Francil, action dédiée à la construction automatique de terminologie et de relations sémantiques entre termes à partir de corpus. Les figures 1 et 2 donnent un aperçu de la hiérarchie de termes construite à partir de ce corpus. Les valeurs utilisées pour les différents paramètres évoqués précédemment sont les suivantes :

Filtrage des ES : S1 = 0,01 et S2 = 0,01 (cf. §4)
Choix du générique d'un terme : S3 = 0,01 et S4 = 0,01 (cf. §5.2)
Gestion de la polysémie : S5 = 0,05 (cf. §6)

```
élève
  texte
          mot
              lettre
              passage
              phrase
              mémoire
              retour
              son
          langue
              unité
              grammaire
          expression
          récit
              personnage
              roman
              suite
          proposition
              liste
          document
```

Figure 1. Extrait de la hiérarchie construite à partir de la racine de plus haut niveau « élève »

Les graphes obtenus ne doivent pas être considérés comme des représentations d'une réalité ontologique, mais bien comme des représentations du contenu d'un corpus. Le fait par exemple que « arithmétique » a comme père indirect « entier » est ainsi spécifique du corpus considéré et le rapport pourrait être inversé pour un autre corpus. La relation peut même être plus spécifique encore comme c'est le cas pour la relation « élève » → « texte ». Cette particularité est néanmoins un atout pour une compréhension globale rapide d'un corpus. On remarquera par ailleurs que les relations mises en évidence, d'un point de vue général, sont non seulement de type générique/spécifique mais également fréquemment de nature métonymique (relation entre « texte » et « mot » ou entre « récit » et « personnage »).

```
élève -> problème -> mathématique ->
énoncé

calcul
    entier
        fraction
            arithmétique
            nombre rationnel
        exercice application
            activité préparatoire
                fonction didactique
        nombre décimal
```

Figure 2. Extrait d'une sous-hiérarchie plus profonde [7]

L'évaluation de ce type d'outils demeure pour une large part un problème ouvert dans la mesure où la définition d'une hiérarchie de termes servant de référence aussi bien que la mise en œuvre d'une procédure d'évaluation manuelle apparaissent très difficiles. Le fait que les hiérarchies construites sont représentatives d'un corpus spécifique rend en particulier très aléatoire la comparaison avec des thésaurus construits manuellement. Dans la perspective de l'assistance à la fouille de textes, nous orientons plutôt notre travail en direction de l'étude de l'influence des différents paramètres de

[7] Le chemin séparant la tête de cet extrait (« calcul ») de la racine de plus haut niveau (« élève ») est donné en préambule.

l'algorithme présenté sur la forme de la hiérarchie produite en nous appuyant pour ce faire sur les travaux relatifs aux mesures de similarité entre arbres et aux tests statistiques permettant la comparaison de matrices de distance (par exemple le test de Mantel ou une adaptation de la mesure Kappa) 0. Une évaluation plus indirecte de l'intérêt des hiérarchies construites est également envisageable dans le cadre de la recherche d'information au travers de leur utilisation pour la reformulation automatique de requêtes (cf. paragraphe suivant).

7.2 Discussion

L'originalité de la démarche présentée repose sur le couplage entre un traitement linguistique approfondi et un algorithme original de structuration hiérarchique de l'information. Les résultats obtenus fournissent un aperçu global du contenu d'un corpus de textes.
De nombreuses autres solutions logicielles visent de la même façon à faciliter le traitement de grandes masses d'informations. Ces logiciels de Text Mining sont pour la plupart fondés sur la construction de cartographies textuelles, émanant soit d'une approche linguistique (TROPES 0, …), soit d'une approche statistique (UMAP 0, …). Depuis quelques années cependant, le traitement linguistique réalisé en amont des traitements mathématiques est de plus en plus élaboré. Des logiciels intégrant des algorithmes de classification automatique (WORDMAPPER 0, …) ou de clustering (SAMPLER, LEXIMINE 0, …) sont ainsi précédés d'un traitement linguistique plus ou moins poussé (extraction des expressions, reconnaissance des mots vides et/ou des pluriels, …).
Dans cette étude, l'utilisation de l'étiqueteur morpho-syntaxique et du lemmatiseur du logiciel SPIRIT fournit une normalisation complète des formes rencontrées dans les textes. Cela permet de gommer certaines différences de forme pouvant s'avérer dommageable pour le traitement statistique utilisé lors de la construction des liens hiérarchiques. De plus, les différents processus mis en œuvre pour l'extraction des termes garantissent une grande qualité des multitermes, avec peu de bruit.
Par ailleurs, l'algorithme de structuration apporte lui-même des possibilités nouvelles par rapport à d'autres algorithmes plus classiques pour l'établissement de relations entre termes, tels que par exemple la méthode des mots associés. Les liens qu'il met en évidence sont en effet dirigés et créent une hiérarchie entre les termes, donnant ainsi des informations supplémentaires sur l'organisation des termes dans le corpus.

Plusieurs types d'applications sont envisagés pour les travaux présentés.

- **Aide à la navigation ; accès facilité à l'information**

Lors d'une recherche d'information dans une base documentaire, une hiérarchie de termes représentative de cette base peut être utilisée comme espace conceptuel pour naviguer de manière superficielle dans l'information afin de cerner rapidement les domaines d'intérêt existants dans les textes disponibles.

- **Aide à la reformulation**

Les hiérarchies sont alors considérées comme des échantillons représentatifs du vocabulaire des textes. La reformulation est une fonction qui permet d'élargir, de manière automatique ou contrôlée, le vocabulaire d'une question documentaire d'un utilisateur afin d'améliorer le taux de rappel de documents pertinents.
Dans le contexte présent, la reformulation en mode automatique peut prendre la forme d'un enrichissement des mots de la question par ses spécifiques. En mode contrôlé, la visualisation des parties pertinentes des hiérarchies de termes permet à l'utilisateur de choisir les termes pertinents à utiliser pour la reformulation.

- **Aide à l'élaboration de thésaurus**

La mise en évidence de liens typés hiérarchiques permet d'envisager l'utilisation de ces hiérarchies comme une aide automatique à l'élaboration de thésaurus. Cette potentialité est sans nul doute assujettie à la représentativité du contenu de la base de documents initiatrice de ces hiérarchies.

- **Aide à la détection de signaux faibles**

Certains signaux faibles peuvent donner lieu à la génération de hiérarchies particulières pouvant servir d'alertes lors de l'interprétation par un analyste.

8 Conclusion

Dans le cadre de cet article, nous avons présenté une méthode permettant d'organiser les termes extraits d'un corpus selon une relation de type générique/spécifique. Cette méthode, de nature statistique, fait l'hypothèse que le sens d'un terme dans un corpus est caractérisé par l'ensemble de ses contextes d'occurrence dans ce corpus. Elle permet par ailleurs de rendre compte de la polysémie des termes de façon concomittante à la construction des hiérarchies de termes.
Cette méthode a été validée au travers d'une implémentation et d'un test sur plusieurs corpus de natures différentes (revue scientifique, brevets, base réglementaire). Cette implémentation fait actuellement l'objet d'une redéfinition et d'une réécriture, en particulier afin de permettre le test d'un ensemble plus large d'hypothèses et l'insertion dans des contextes applicatifs allant de l'aide à la navigation jusqu'à la veille technologique.

Bibliographie

[1.] ASSADI H., Construction d'ontologies à partir de textes techniques. Application aux systèmes documentaires, Thèse de l'Université Paris 6, 1998
[2.] BARBIÉRI B., Vers une construction automatique de concepts, Thèse de l'Ecole centrale de Paris, 1992
[3.] BOURIGAULT D., LEXTER, a Terminology Extraction Software for Knowledge Acquisition from Texts, 9ème Knowledge Acquisition for Knowledge Based System Workshop (KAW'95), Banff, Canada, 1995
[4.] BOURIGAULT D. et JACQUEMIN C., Term extraction + term clustering: An integrated platform for computer-aided terminology, 9ème conference of the European Chapter of the Association for Computational Linguistics (EACL'99), Bergen, Norvège, 1999
[5.] BOURIGAULT D. et JACQUEMIN C., Construction de ressources terminologiques, dans Ingénierie des langues, éditeur J-M. Pierrel, Hermès, Paris, 2000
[6.] DAILLE B., Extraction automatique de terminologie monolingue, conférence Informatique et Langues Naturelles (ILN'93), Nantes, 1993
[7.] DEBILI F., Analyse syntaxico-sémantique fondée sur une acquisition automatique de relations lexicales-sémantiques, Thèse de doctorat d'état de l'Université Paris XI, 1982
[8.] ENGHEHARD C., Acquisition de terminologie à partir de gros corpus, conférence Informatique et Langue Naturelle (ILN'93), Nantes, p 373-384, 1993
[9.] FERRET O., GRAU B. et JARDINO M., A cross-comparison of two clustering methods, Workshop on Evaluation for Language and Dialogue Systems, ACL, Toulouse, 2001
[10.] FLUHR C., *SPIRIT : un système d'exploration de données textuelles*, Le Traitement Informatique des Corpus Textuels, INALF, 1994
[11.] GHIGLIONE R., LANDRÉ A, BOMBERG M. et MOLETTE P., L'Analyse Automatique des Contenus, Dunod, 1998
[12.] GREFENSTETTE G., *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publisher, Boston, MA, 1994
[13.] GRIMMERSOFT, http://www.grimmersoft.com
[14.] HABERT B., NAULLEAU E. et NAZARENKO A., Symbolic word clustering for medium-size corpora, 16ème International Conference on Computational Linguistics (COLING'96), Copenhague, Danemark, 1996
[15.] HAMON T. et NAZARENKO A., A step towards the detection of semantic variants of terms in technical documents, 36ème Annual Meeting of the Association for Computational Linguistics et 17ème International Conference on Computational Linguistics (COLING-ACL'98), Montréal, Canada, 1998
[16.] HEARST M., Automatic acquisition of hyponyms from large text corpora, 14ème International Conference on Computational Linguistics (COLING'92), Nantes, 1992
[17.] JACQUEMIN C., *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*, Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes, 1997
[18.] JOUIS C., Constribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse de textes. Réalisation d'un prototype : le système SEEK, Thèse en informatique de l'Ecole des Hautes Etudes en Sciences Sociales, Paris, 1993

[19.] JOUVE O., Les outils d'analyse et de filtrage d'information : exemple du projet Sampler, IDT'98, Paris, 1998
[20.] LEXIQUEST, http://www.lexiquest.com
[21.] MICHELET B., *L'analyse des associations*, Thèse de doctorat de l'Université Paris VII, Paris, 1988.
[22.] MORIN E., Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes, TALN'98, Paris, 1998
[23.] SIMONI J-L, Accès à l'information à l'aide d'un graphe de termes construit automatiquement, Thèse en information scientifique et technique de l'Université Paris VII, 2000
[24.] TRIVIUM, http://www.trivium.fr