Filtrage thématique d'un réseau de collocations

Olivier Ferret

CEA – LIST/LIC2M
92265 Fontenay-aux-Roses Cedex
olivier.ferret@cea.fr

Résumé – Abstract

Les réseaux lexicaux de type WordNet présentent une absence de relations de nature thématique, relations pourtant très utiles dans des tâches telles que le résumé automatique ou l'extraction d'information. Dans cet article, nous proposons une méthode visant à construire automatiquement à partir d'un large corpus un réseau lexical dont les relations sont préférentiellement thématiques. En l'absence d'utilisation de ressources de type dictionnaire, cette méthode se fonde sur un principe d'auto-amorçage : un réseau de collocations est d'abord construit à partir d'un corpus puis filtré sur la base des mots du corpus que le réseau initial a permis de sélectionner. Nous montrons au travers d'une évaluation portant sur la segmentation thématique que le réseau final, bien que de taille bien inférieure au réseau initial, permet d'obtenir les mêmes performances que celui-ci pour cette tâche.

Lexical networks such as WordNet are known to have a lack of topical relations although these ones are very useful for tasks such as text summarization or information extraction. In this article, we present a method for automatically building from a large corpus a lexical network whose relations are preferably topical ones. As it does not rely on resources such as dictionaries, this method is based on self-bootstrapping: a collocation network is first built from a corpus and then, is filtered by using the words of the corpus that are selected by the initial network. We report an evaluation about topic segmentation showing that the results got with the filtered network are the same as the results got with the initial network although the first one is signicantly smaller than the second one.

Mots Clés– Keywords

Collocations, cooccurrences lexicales, réseaux lexicaux thématiques, analyse thématique
Collocations, lexical cooccurrences, topical lexical networks, topic analysis

1 Introduction

Depuis l'avènement de WordNet (Miller, 1995), un vif intérêt s'est développé autour des réseaux lexico-sémantiques. Ainsi que l'ont souligné certains auteurs (Harabagiu *et al.*, 1999), les réseaux de type WordNet présentent néanmoins un certain nombre d'insuffisances. L'une d'entre elles est l'absence en leur sein de relations thématiques. Ces relations caractérisent l'appartenance de deux mots à un même thème ou à une même situation comme c'est le cas

dans les couples *docteur-hôpital* ou *cambrioleur-policier*. Ces connaissances thématiques s'avèrent pourtant très utiles dans des domaines tels que l'extraction d'information ou le résumé automatique (Harabagiu, Maiorano, 2002) afin de déterminer ce qui est caractéristique d'une situation ou d'un thème. La difficulté d'en dresser un inventaire manuel du fait leur très vaste étendue a débouché sur un certain nombre de travaux concernant leur acquisition automatique, soit sous la forme de relations thématiques constituant un réseau d'associations lexicales ou bien de configurations de mots relatifs à un même thème.

La première approche est illustrée par le travail de Harabagiu et Moldovan (Harabagiu, Maiorano, 2002) concernant l'extraction de relations thématiques de WordNet par l'exploitation des définitions associées aux synsets. La seconde approche est typiquement représentée par le travail décrit dans (Lin, Hovy, 2000) portant sur l'apprentissage supervisé de signatures thématiques, représentations de thèmes obtenues par sélection et pondération du vocabulaire d'un ensemble de textes relatifs à ces thèmes. (Ferret, Grau, 1998) ont montré comment des représentations du même type peuvent être apprises de façon non supervisée. À mi-chemin entre ces deux approches, (Agirre *et al.*, 2001) construit des signatures thématiques centrées sur des synsets de WordNet. Les connaissances associées aux synsets sont utilisées pour sélectionner des textes en relation avec eux, textes dont on extrait ensuite le vocabulaire constitutif des signatures thématiques.

En dépit de leur intérêt, les travaux s'appuyant sur WordNet ont intrinsèquement un impact limité par le fait que WordNet est une ressource pour le moment unique quant à son degré d'élaboration et que ces travaux en exploitent tous les aspects, en particulier ceux absents des réseaux de même type (par exemple les définitions). Nous avons fait au contraire le choix d'une approche ne nécessitant pas de pré-requis importants, ce qui permet de l'appliquer potentiellement à un grand nombre de langues en dehors de l'anglais. La constitution d'un réseau de collocations, i.e. un ensemble de collocations liées par l'intermédiaire de leurs mots, apparaît à cet égard comme une solution intéressante. Mais compte tenu de l'hétérogénéité des relations sous-jacentes aux collocations, elle nécessite l'adjonction d'un processus de filtrage destiné à retenir préférentiellement les collocations de nature thématique. Pour réaliser ce filtrage, nous avons élaboré une méthode fondée sur l'amorçage. Le réseau de collocations à filtrer est utilisé par un système d'analyse thématique afin de mettre en évidence dans le corpus de construction de ce réseau des segments thématiquement homogènes et au sein de ces derniers, les mots caractérisant leur thème. Un réseau de collocations plus spécifiquement thématiques peut alors être construit en enregistrant les cooccurrences au sein de ces segments. Ce réseau est finalement utilisé pour éliminer du réseau initial celles de ses collocations qui sont les moins thématiques.

2 Construction du réseau de collocations initial

Le corpus sur lequel nous nous sommes appuyés pour ce travail est constitué de 24 mois du journal *Le Monde* sélectionnés entre 1990 et 1994, ce qui représente environ 39 millions de mots. Afin de construire le réseau de collocations initial, ce corpus a d'abord été pré-traité afin de caractériser les textes par leurs mots les plus thématiquement significatifs, en l'occurrence les noms, les verbes et les adjectifs, donnés sous forme lemmatisée. Les ambiguïtés de lemmatisation ont été levées grâce à un étiqueteur morpho-syntaxique. Les collocations ont ensuite été extraites en utilisant une fenêtre glissante selon la méthode décrite dans (Church, Hanks, 1990). Les paramètres de cette extraction ont été fixés afin de favoriser la capture de relations thématiques : une fenêtre assez large (20 mots), respectant la fin des tex-

tes et ne conservant pas l'ordre des collocations. Nous avons comme Church et Hanks adopté une évaluation de l'information mutuelle en tant que mesure de cohésion des collocations, mesure normalisée dans notre cas par l'information mutuelle maximale relative au corpus. Après filtrage des collocations les moins significatives (cohésion < 0,1 et moins de 10 occurrences), nous avons obtenu un réseau de 22.749 mots et 2.572.589 collocations.

3 Méthode de filtrage thématique

3.1 Construction des Unités Thématiques

La première étape du filtrage thématique d'un réseau de collocations consiste à définir, à partir du corpus ayant servi à la construction de ce réseau, un ensemble d'unités textuelles présentant les deux caractéristiques suivantes : chaque unité correspond à un segment de texte renvoyant à un seul thème et les mots qu'elle contient sont les mots de ce segment représentatifs de ce thème. Ces unités, déjà introduites dans (Ferret, Grau, 1998), sont appelées Unités Thématiques (UTs). Leur construction s'appuie sur l'utilisation conjointe d'un outil de segmentation thématique et du réseau de collocations à filtrer. L'outil de segmentation thématique permet de délimiter des segments thématiquement homogènes tandis que le réseau de collocations sert de support à la sélection de mots caractérisant les thèmes des segments. Dans le cas présent, nous faisons appel pour la construction des UTs à TOPICOLL (Ferret, 2002), un outil d'analyse thématique qui réalise simultanément ces deux tâches en exploitant les mots qu'il sélectionne dans un réseau de collocations pour segmenter les textes.

Dans le réseau de collocations initial, beaucoup de collocations sont composées de mots qui, de par leur généralité, se retrouvent dans de nombreux contextes différents et ne sont donc pas thématiquement discriminants. Le mécanisme de sélection de TOPICOLL repose sur l'hypothèse que dans un segment de texte renvoyant à un contexte thématique spécifique, il existe dans le réseau de collocations davantage de liens, directs ou par l'intermédiaire d'un autre mot du réseau, entre les mots de ce segment lorsque ceux-ci sont représentatifs du thème du segment que lorsque ce sont des mots assez généraux ou représentatifs d'un autre thème. Les résultats obtenus par TOPICOLL en matière de segmentation thématique (Ferret, 2002) tendent à prouver que cette hypothèse est justifiée.

Cette hypothèse se concrétise dans TOPICOLL au travers du processus suivant : une fenêtre délimitant l'espace de focalisation de l'analyse est déplacée sur le texte considéré. Cette fenêtre contient sous forme lemmatisée les mots pleins du texte issus de son pré-traitement, identique à celui appliqué aux textes lors de la construction du réseau de collocations initial. À chaque position de la fenêtre, on sélectionne les mots du réseau de collocations qui sont liés à au moins trois mots de la fenêtre. Il est à noter que sont ainsi sélectionnés à la fois des mots de l'espace du texte délimité par la fenêtre de focalisation et des mots du réseau n'en faisant pas partie. Ces derniers sont appelés *mots inférés*. Pour limiter l'impact du bruit présent dans le réseau initial, une élimination préalable des collocations dont la valeur de cohésion est inférieure à 0,12 est réalisée. Ce seuil a été déterminé expérimentalement sur la base des performances de TOPICOLL en matière de segmentation thématique.

Les mots sélectionnés pour chaque position de la fenêtre sont conservés et finalement, ne sont retenus pour constituer l'UT associée à un segment que les mots ayant été sélectionnés pour au moins 75% des positions du segment, condition reprise de (Ferret, Grau, 1998) et visant de

nouveau à réduire le nombre de mots sélectionnés sur la base de collocations non thématiques.

3.2 Filtrage des Unités Thématiques

Les UTs ainsi construites font l'objet d'un double filtrage. Le premier vise à écarter les UTs peu significatives sur le plan thématique. Certains textes entremêlent si étroitement plusieurs thèmes qu'il est impossible de les différencier grâce à un découpage linéaire du texte. Le résultat des méthodes de segmentation existantes n'est alors pas significatif et les UTs construites n'ont pas l'homogénéité thématique qui devrait les caractériser. Nous considérons qu'un tel cas de figure est détecté lorsque aucun mot d'un segment ne figure parmi les mots de l'UT qui lui est associée. Au-delà de cette première contrainte, nous avons choisi de ne retenir que les UTs contenant au moins deux mots de leur segment originel. Un thème étant une configuration d'unités sémantiques, nous considérons en effet qu'il doit être représenté au minimum par deux de ses mots pour être attesté.

Le second filtrage est interne à chaque UT et s'opère parmi ceux de ses mots ne faisant pas partie de son segment originel, c'est-à-dire les mots inférés. Ne conserver que les mots du segment originel serait un peu restrictif pour la construction d'un réseau de collocations thématique dans la mesure où leur nombre est faible (rarement plus de trois). Néanmoins, dans la perspective du filtrage du réseau de collocations initial, il est nécessaire de restreindre les mots inférés à ceux supposés les plus thématiquement proches des mots du segment sélectionnés. Nous avons donc appliqué le même principe de sélection que précédemment : un mot inféré est retenu s'il est lié, par l'intermédiaire du réseau de collocations, à au moins trois mots sélectionnés du segment. Par ailleurs, nous avons aussi imposé un seuillage assez restrictif en fréquence et en cohésion concernant les collocations supportant ces liens (cohésion $\geq 0{,}15$ et fréquence ≥ 15).

3.3 Construction d'un réseau de collocations thématiques

À l'issue de l'étape précédente, on dispose d'un ensemble d'UTs rassemblant chacune des mots que l'on suppose assez fortement liés sur le plan thématique. En enregistrant les cooccurrences entre ces mots sur l'ensemble des UTs produites à partir du corpus ayant permis la construction du réseau initial, on obtient un ensemble de collocations qui sont préférentiellement de nature thématique, même si elles ne sont pas exclusivement de ce type compte tenu du caractère non supervisé de leur sélection. La fréquence d'une collocation est dans ce cas directement donnée par le nombre d'UTs dans lesquelles les deux mots qui la constituent sont présents simultanément. Aucune différenciation n'est réalisée en fonction de l'origine des mots constituant les UTs.

3.4 Filtrage du réseau initial

Le réseau de collocations obtenu à l'issue de l'étape précédente regroupe un sous-ensemble des collocations du réseau initial mais également des collocations n'y figurant pas. Parmi elles, certaines sont pertinentes sur le plan thématique mais d'autres ne le sont pas. Comme il est difficile de juger globalement du rapport entre les collocations intéressantes et celles qui ne le sont pas, il paraît plus prudent dans un premier temps de laisser de côté ces nouvelles collocations et de ne retenir que celles qui sont communes avec le réseau initial.

Le réseau de collocations thématiques n'est de ce fait utilisé que comme filtre du réseau de collocations initial. Préalablement, il est lui-même filtré afin de supprimer les collocations dont la fréquence est trop faible, ce qui est signe de leur absence de représentativité. En pratique, seules sont conservées les collocations thématiques de fréquence supérieure à 5. Cette valeur a été déterminée de manière expérimentale sur la base de l'utilisation du réseau filtré final (cf. section 4). Le filtrage du réseau initial se limite à ne retenir que le sous-ensemble de ses collocations également présentes dans le réseau thématique. Les valeurs de fréquence et de cohésion des collocations ainsi retenues sont celles du réseau initial, conformément à la perspective adoptée de filtrage du réseau initial.

4 Résultats et évaluation

Nous avons appliqué la procédure de filtrage présentée au réseau de collocations de la section 2. La première étape a conduit à un ensemble de 382.208 UTs dont seulement 59% ont été conservées après filtrage. Le réseau construit à partir de ces UTs est constitué de 11.674 mots et de 2.864.473 collocations. Parmi celles-ci, environ 70% étaient nouvelles par rapport au réseau initial et n'ont donc pas été retenues. Finalement, le réseau filtré comporte 7.223 mots et 400.963 collocations.

Évaluer la qualité thématique d'un réseau de collocations est, comme toute tâche d'évaluation d'une ressource linguistique, assez difficile. Le jugement humain direct étant peu praticable de par la taille des données à considérer et un réseau lexical thématique de référence n'existant pas pour le français, nous avons opté pour une évaluation indirecte. Celle-ci a été réalisée au travers de l'utilisation par TOPICOLL d'un réseau de collocations pour assurer une tâche de segmentation thématique. Cette tâche consiste à redécouvrir les frontières d'un ensemble de textes concaténés, textes assimilables à des paragraphes. L'évaluation a été réalisée sur un ensemble de 49 textes longs de 133 mots en moyenne, extraits du journal *Le Monde* (1995) et couvrant 11 thèmes. Les résultats du Tableau 1 sont des moyennes obtenues sur 10 ordonnancements différents de ces textes. La précision est définie par N_c / N_b et le rappel par N_c / D, où N_b est le nombre de bornes trouvées par TOPICOLL, N_c est le nombre de bornes trouvées correctes et D le nombre total de frontières de texte.

Systèmes	Rappel	Précision	F1-mesure
TOPICOLL_1 (réseau non filtré)	0,85	0,79	0,82
TOPICOLL_2 (filtrage thématique)	0,85	0,79	0,82
TOPICOLL_3 (filtrage fréquentiel)	0,83	0,71	0,77

Tableau 1 : Précision/rappel de la segmentation pour le corpus du *Monde*

TOPICOLL_1 correspond à une version de TOPICOLL avec le réseau de collocations initial, TOPICOLL_2, avec le réseau filtré thématiquement et enfin TOPICOLL_3, avec un réseau de taille la plus proche possible du réseau filtré thématiquement mais obtenu par un seuillage de la fréquence (à 14 occurrences) et de la cohésion (à 0,14) des collocations. Dans ce dernier cas, le réseau obtenu comporte 17.639 mots et 196.374 collocations. TOPICOLL_1 réalisant un seuillage en cohésion et fréquence (0,13 et 13), nous avons repris ce même seuillage dans TOPICOLL_2 afin d'obtenir des résultats comparables. Le réseau effectif de TOPICOLL_2 est ainsi constitué de 7.160 mots et 183.074 collocations tandis que celui de TOPICOLL_1 est formé de 18.958 mots et 341.549 collocations. Comme le montre clairement le Tableau 1, le filtrage thématique permet de réduire le réseau de collocations de 46% tout en conservant les mêmes performances. Comparativement, un seuillage non spécifiquement thématique conduit

à une dégradation significative des performances pour une réduction de volume un peu moins importante. Ce résultat tend donc à montrer que le caractère effectif de la méthode de filtrage thématique que nous avons proposée.

5 Conclusion et perspectives

Le travail que nous avons présenté avait pour objectif de construire automatiquement un réseau lexical de nature thématique en limitant autant que possible les outils et les ressources linguistiques utilisés. Nous avons choisi d'aborder ce problème par le biais du filtrage thématique d'un réseau de collocations et proposé une méthode pour ce faire fondée sur l'amorçage. Son évaluation indirecte au travers de l'utilisation de son résultat par une méthode de segmentation thématique a montré son intérêt. Cette évaluation doit toutefois être poussée plus avant, notamment en comparant les réseaux lexicaux ainsi produits avec des réseaux équivalents construits ou contrôlés manuellement. La mise à disposition dans le cadre de WordNet de relations thématiques extraites des définitions des synsets (Harabagiu *et al.*, 1999) devrait permettre une avancée intéressante dans cette direction.

Références

Agirre E., Ansa O., Martinez D., Hovy E. (2001), Enriching WordNet concepts with topic signatures, Actes de *Workshop on WordNet of the NAACL'01 Conference*.

Church K. W., Hanks P. (1990), Word Association Norms, Mutual Information, And Lexicography, *Computational Linguistics*, Vol. 16(1), pp. 177-210.

Ferret O. (2002), Using collocations for topic segmentation and link detection, Actes de *COLING 2002*, pp. 260-266.

Ferret O., Grau B. (1998), A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts, Actes de *ECAI'98*, pp. 155-159.

Harabagiu S, Maiorano S (2002), Multi-Document Summarization with GISTEXTER, Actes de *LREC 2002*.

Harabagiu S, Miller G.A., Moldovan D (1999), WordNet 2 - A Morphologically and Semantically Enhanced Resource, Actes de *SIGLEX'99*, pp. 1-8.

Lin C.Y., Hovy E. (2000), The Automated Acquisition of Topic Signatures for Text Summarization, Actes de *COLING 2000*.

Miller G.A. (1995), WordNet: A lexical Database, *Communications of the ACM*, Vol. 38(11).