Extraction et filtrage de syntagmes nominaux pour la Recherche d'Information

Chedi Bechikh Ali*, Hatem Haddad*,**

*Faculté des Sciences de Tunis
Département Infromatique
Laboratoire LIPAH
Campus Universitaire El Manar, Tunis
Tunisie
chedi.bechikh@gmail.com,
**ESSTHS, Université de Sousse
H. Sousse, Tunisie
haddad.hatem@gmail.com

Résumé. Nous proposons dans cet article un Système de Recherche d'Information (SRI) qui se base sur des techniques d'indexation de textes en langue naturelle. Nous présentons une méthode d'indexation de documents qui repose sur une approche hybride pour la sélection de descripteurs textuels. Cette approche emploie des traitements du langage naturel pour l'extraction des syntagmes nominaux et sur un filtrage statistique basé sur l'information mutuelle pour sélectionner les syntagmes nominaux les plus informatifs pour le processus d'indexation. Nous effectuons des expérimentations en utilisant le corpus Le Monde 94 de la collection CLEF 2001 et sur le SRI Lemur pour évaluer l'approche proposée.

1 Introduction

La plupart des Systèmes de Recherche d'Information (SRI) utilisent des termes simples pour indexer et retrouver des documents. Cependant, cette représentation n'est pas assez précise pour représenter le contenu des documents et des requêtes, du fait de l'ambigüité des termes isolés de leur contexte : si l'on considère le mot composé « *pomme de terre* », les mots simples pomme et terre ne gardent pas leur propre sens que dans l'expression « *pomme de terre* » et si on les utilise séparément ils deviennent une source d'ambigüité.

Une solution à ce problème consiste à utiliser des termes complexes à la place des termes simples isolés (Boulaknadel, 2006). L'hypothèse est que les termes complexes sont plus aptes à désigner des entités sémantiques que les mots simples et constituent alors une meilleure représentation du contenu sémantique des documents (Mitra et al., 1997).

Notre objectif consiste à acquérir des termes complexes représentatifs du contenu informationnel du corpus. Les termes complexes extraits doivent représenter le contenu des textes sous une forme compréhensive par l'ordinateur et riche en information. Ces termes extraits

sont utilisés pour effectuer l'indexation de corpus textuels, les termes d'indexation sont alors plus complets et plus précis, ils permettent d'atteindre une meilleure performance pour le SRI.

Les termes complexes peuvent être sélectionnés statistiquement, linguistiquement ou en combinant les deux approches. Les techniques statistiques permettent de découvrir des séries de mots ou de combinaisons de mots qui occurrent fréquemment dans un corpus. Les techniques linguistiques visent à extraire les dépendances ou les relations entre les termes grâce aux phénomènes langagiers.

Dans (Haddad, 2002), l'auteur fait l'indexation des documents et des requêtes après l'analyse linguistique et l'extraction des syntagmes nominaux (SNs). Les résultats des expérimentations montrent que l'intégration des SNs dans l'indexation permet d'obtenir de meilleures performances par rapport à l'utilisation des unitermes.

Le et Chevalet (Diem et Chevallet, 2006) utilisent une méthode d'extraction de connaissances hybride qui fusionne l'association entre les paires de termes extraits statistiquement avec les relations sémantique extraites linguistiquement. Les SNs sont organisés en réseaux de dépendance syntaxique (tête et expansion/ modificateur) en ajoutant les associations statistiques et sémantiques. L'information sémantique est étudiée à travers les relations : synonymie, hyperonymie, causalité.

Les auteurs dans (Woods et al., 2000; Haddad, 2003) ont montré que l'indexation avec des SNs extraits linguistiquement affecte plus positivement les résultats d'un SRI que celle avec des groupes de mots extraits statistiquement.

2 Méthodologie suivie

L'approche hybride d'extraction de connaissances montre son efficacité dans l'augmentation de la performance des SRIs. Nous choisissons alors de combiner entre une approche linguistique basée sur l'extraction des syntagmes nominaux (SNs) et sur un filtrage statistique basé sur l'information mutuelle (IM) car cette mesure est adaptée aux termes rares (Daille, 1994; Thanopoulos et al., 2002) ce qui est le cas des SNs, pour la représentation du contenu textuel du corpus.

2.1 Extraction des syntagmes nominaux

Nous effectuons, d'abord l'analyse linguistique avec un étiqueteur, qui génère une collection étiquetée. Ensuite, on utilise cette collection étiquetée et on en extrait un ensemble de SNs. Les syntagmes nominaux candidats sont extraits par repérage de patrons syntaxiques. Nous adoptons la définition des patrons syntaxiques dans (Haddad, 2002), où un patron syntaxique est une règle sur l'ordre d'enchainement des catégories grammaticales qui forment un SN :

- V : le vocabulaire extrait du corpus
- C : un ensemble de catégories lexicales
- L : le lexique $\subset V \times C$

Un patron syntaxique est une règle de la forme :

$$X := Y_1 Y_2 Y_k ... Y_{k+1} ... Y_n$$

Avec Yi $\in$ C et X un syntagme nominal, exemples : SUBC ADJQ : « échelle planétaire », « relation diplomatique », etc. Nous nous basons dans nos travaux sur les 10 patrons syntaxiques

les plus susceptibles de contenir le maximum d'informations (Haddad, 2002), alors on ne va étudier que les SNs composés de deux ou de trois termes.

2.2 Sélection des meilleurs descripteurs

Bien que les SNs soient généralement pertinents, il peut être nécessaire de n'en sélectionner que les « meilleurs » du corpus, pour cela nous utilisons un filtrage statistique qui consiste à employer une mesure statistique afin de leurs donner un score de qualité. Pour notre cas, ce filtrage statistique est effectué en calculant l'information mutuelle (IM) entre les composants de chaque SN et en fixant différents seuils de cette mesure pour ne conserver que les SNs plus pertinent pour l'indexation. La mesure de l'information mutuelle consiste à comparer la probabilité d'apparition des cooccurrences de mots (m1, m2) à la probabilité d'apparition de ces mots séparément. Cette mesure est donnée par (Church et Hanks, 1990) :

$$IM(m1, m2) = \log_2(\frac{P(m1, m2)}{P(m1) * P(m2)}) \quad (1)$$

Où P est la probabilté.

Dans notre cas l'IM est utilisée pour détecter les syntagmes nominaux les plus pertinents pour les utiliser ensuite dans notre processus de RI. Donc pour les syntagmes composés de deux mots $X := Y_1Y_2$ Avec Yi $\in$ C et X un SN, l'IM va être calculée de la façon suivante :

$$IM(Y_1, Y_2) = \log_2(\frac{P(Y_1, Y_2)}{P(Y_1) * P(Y_2)}) \quad (2)$$

Pour les SNs composés de trois mots $X := Y_1Y_2Y_3$, l'IM va être calculée de la façon suivante :

Si Y_2 est une préposition, alors :

$$IM(Y_1, Y_3) = \log_2(\frac{P(Y_1, Y_3)}{P(Y_1) * P(Y_3)}) \quad (3)$$

Sinon

$$IM(Y_1, (Y_2, Y_3)) = \log_2(\frac{P(Y_1, (Y_2, Y_3))}{P(Y_1) * P(Y_2, Y_3)}) \quad (4)$$

Où P(Y_i) est une estimation de la probabilité d'apparition du mot Y_i qui est calculée à partir de la fréquence d'apparition du mot Y_i dans le document où il apparait, normalisée par N le nombre de mots contenu dans le même document.

P(Y_i, Y_j) est une estimation de la probabilité que les deux mots apparaissent ensemble dans le même document. Cette probabilité est estimée par la fréquence d'apparition du couple (Y_i, Y_j) divisé par N.

3 Expérimentations et résultats

Pour tester notre approche hybride d'extraction de SNs, nous avons utilisé le SRI Lemur [1], le modèle vectoriel et la mesure de pondération *tf.idf* (Salton et Yang, 1973).

Le corpus utilisé est le corpus « *Le Monde* » fournit lors la compagne d'évaluation CLEF 2001 [2] de taille 157 MB et composé de 44 013 documents, un ensemble de 49 requêtes est

1. http ://www.lemurproject.org
2. http ://www.clef-initiative.eu

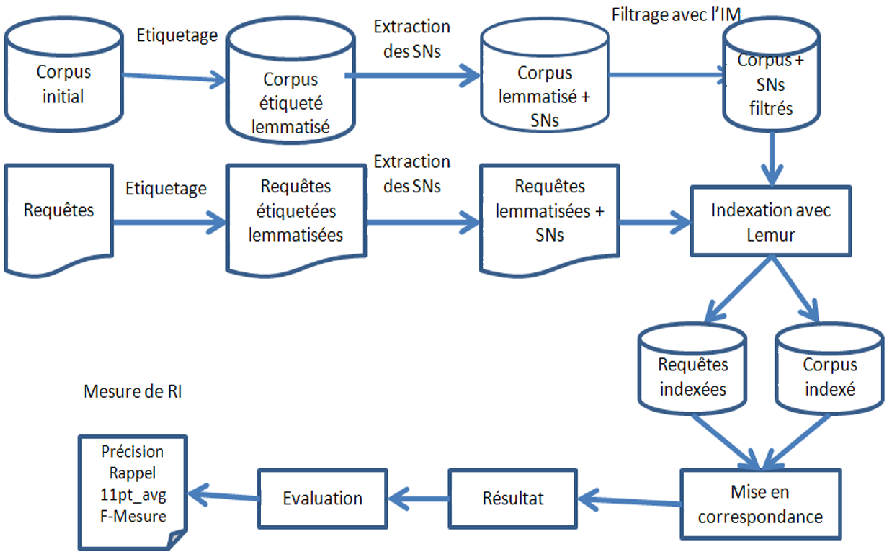

FIG. 1 – *Processus expérimental.*

associées à ce corpus numérotées de 41 à 90.

La Figure 1 décrit le processus expérimental suivi. Nous lemmatisons les documents et les requêtes en utilisant un étiqueteur, ensuite nous passons à l'extraction des SNs en utilisant un programme que nous avons développé.

Le filtrage des SNs s'effectue comme décrit dans la section précédente selon le score d'information mutuelle obtenu par chaque SN, nous fixons alors un seuil minimal et/ou un seuil maximal de l'information mutuelle, au delà duquel les SNs sont retenus. Nous avons défini alors deux variables *S_Min* et *S_Max* qui correspondent respectivement au seuil minimal et au seuil maximal que prendra l'IM.

Nous enlevons par la suite les mots vides lors de l'indexation et cela avec le SRI Lemur. Pour évaluer notre approche d'indexation, on se focalise sur la précision à faible taux de rappel et cela en étudiant la précision à 3, 5 et 10 documents (P_3, P_5, P_10), la précision à 11 point de rappel (11 pt_avg) et la F-mesure sont aussi étudiés.

Pour ces expérimentations le corpus initial sera noté *Corpus-I*, le corpus qui contient les SNs sera noté *Corpus-SN* et le corpus qui contient des SNs après filtrage sera noté *Corpus-Fi*

3.1 Résultats du filtrage avec l'IM

Selon les expériences on a déterminé que le score de l'information mutuelle pour les SNs varie entre -2 et 12. Nous varions alors *S_Min* et *S_Max* entre cet intervalle pour déterminer les meilleurs paramètres pour le filtrage statistique.

Après analyse du Tableau 3.1, nous remarquons que l'approche de filtrage selon l'IM a permis l'amélioration de la P_3 lorsque *S_Max*= 7, 8 et 9 avec une amélioration de 1.4% par rapport au *Corpus-I*. Nous remarquons aussi l'amélioration de la P_5 pour *S_Max*= 4, 7, 8 et

	S_Min	S_Max	P_3	P_5	P_10	11 pt_avg	F-mesure
Corpus-I	-	-	0.9467	0.9240	0.9040	0.4660	0.1860
Corpus-SN	-	-	0.9400	0.9320	0.8920	0.4472	0.1838
Corpus-Fi	-	4	0.9467	0.9280	0.9060	0.4663	0.1861
	-	6	0.9400	0.9240	**0.9080**	0.4764	0.1874
	-	7	**0.9600**	0.9280	0.9060	0.4850	0.1890
	-	8	**0.9600**	**0.9440**	0.9020	**0.4873**	**0.1897**
	-	10	**0.9600**	0.9360	0.9040	0.4656	0.1859
	2	-	0.9400	0.9320	0.8920	0.4472	0.1838
	6	-	0.9333	0.9320	0.8900	0.4399	0.1825
	8	-	0.9133	0.9080	0.8540	0.4101	0.1771
	10	-	0.9200	0.9000	0.8620	0.4214	0.1812

TAB. 1 – *Résultats avant et après filtrage.*

10, et pour *S_Min*= 2 et 6, la meilleure amélioration est de 2.1% pour *S_Max*= 8 par rapport au *Corpus-Fi*. Pour la P_10, trois stratégies ont permis l'amélioration de cette mesure avec *S_Max*= 4, 6 et 7, la stratégie *S_Max*= 6 a permis une amélioration de 0.2% par rapport au *Corpus-I*.

L'analyse de la 11 pt_avg, montre que cinq stratégies du *Corpus-Fi* ont permis l'amélioration de cette mesure lorsque *S_Max*= 2, 4, 6, 7, 8 et 10, la meilleure amélioration est obtenue pour *S_Max*= 8, ce qui a permis d'augmenter cette mesure de 4.6% par rapport au *Corpus-Fi*.

Pour la F-mesure le meilleur score est obtenu pour le *Corpus-Fi* lorsque *S_Max*= 8 avec une augmentation de 2% par rapport au *Corpus-I* et de 3.2% par rapport au *Corpus-SN*.

Nous remarquons que pour la 11 pt_avg la meilleure performance est obtenue par le *Corpus-Fi* pour *S_Max*=8, c'est à dire que les SNs les plus pertinents ont un score d'IM inférieur à 8. Nous remarquons aussi que les résultats obtenus par le *Corpus-SN* sont inférieurs à ceux obtenus par le *Corpus-I*, cela peut être expliquer par le fait que les SNs sélectionnés ne sont pas tous de bonne qualité.

4 Conclusion

Cet article présente notre méthode d'indexation basée sur la sélection et le filtrage des SNs. Nous avons opté pour une méthode hybride d'extraction des connaissances, qui combine à la fois l'analyse linguistique fondée sur l'extraction des SNs et l'analyse statistique. Les SNs candidats sont extraits d'un corpus étiqueté par repérage de patrons syntaxique. Nous procédons par la suite à un filtrage statistique basé sur l'IM pour ne sélectionner que les SNs les plus pertinents.

Les résultats ont montré que cette méthode permet d'améliorer les performances de notre SRI et que la meilleure performance est obtenue pour le cas où *S_Max*= 8, malgré que les SNs étaient de mauvaise qualité car la stratégie *Corpus-SN* qui utilise les SNs a dégradé les performances du SRI.

Pour les travaux futurs, c'est le processus de filtrage qui sera mis en question, que ce soit pour le filtrage linguistique qui permettra d'extraire des SNs de meilleurs qualités, que ce soit pour le filtrage statistique.

Références

Boulaknadel, S. (2006). Utilisation des syntagmes nominaux dans un système de recherche d'information en langue arabe. In *CORIA*, pp. 341–346.

Church, K. W. et P. Hanks (March 1990). Word association norms, mutual information, and lexicography. *Comput. Linguist. 16*(1), 22–29.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Thèse de doctorat, Université Paris 7.

Diem, L. T. H. et J.-P. Chevallet (2006). Extraction et structuration des relations à partir de texte. In *RIVF'06*, Viêt-Nam, pp. 53–58.

Haddad, H. (2002). *Extraction et Impact des connaissances sur les performances des Systèmes de Recherche d'Information*. Thèse de doctorat, Université Joseph Fourier.

Haddad, H. (2003). French noun phrase indexing and mining for an information retrieval system. In *String Processing and Information Retrieval, 10th International Symposium*, Manaus, Brazil, pp. 277–286.

Mitra, M., C. Buckley, A. Singhal, et C. Cardie (1997). An analysis of statistical and syntactic phrases. In *RIAO*, pp. 200–217.

Salton, G. et C. S. Yang (1973). On the specification of term values in automatic indexing. *Journal of Documentation. 29*(4).

Thanopoulos, A., N. Fakotakis, et G. Kokkinakis (2002). Comparative evaluation of collocation extraction metrics. In *In Proceedings of the 3rd Language Resources Evaluation Conference*, pp. 620–625.

Woods, W. A., L. A. Bookman, A. Houston, R. J. Kuhns, P. Martin, et S. Green (2000). Linguistic knowledge can improve information retrieval. In *Proceedings of the sixth conference on Applied natural language processing*, ANLC '00, pp. 262–267.

Summary

In this paper, we propose an information retrieval system based on natural language processing and statistical indexing techniques. We present a method of indexing documents based on a hybrid approach for selecting text descriptors. These techniques use natural language processing to extract noun phrases and statistical filtering based on mutual information to select the most representative noun phrases for the indexing process. We evaluate our system using the corpus Le monde 94 of the collection CLEF 2001 and the IRS Lemur.