Indexation à base des syntagmes nominaux

Amine Amri Maroua Mbarek Chedi Bechikh
Chiraz Latiri Hatem Haddad
Equipe de recherche URPAH, Faculté des Sciences Tunis El Manar
esc.amriamine@gmail.com, maroua.mbarek@yahoo.fr, chedi.bechikh@gmail.com
chiraz.latiri@gnet.tn, haddad.hatem@gmail.com

RÉSUMÉ

Cet article présente la participation de l'équipe URPAH à DEFT 2012. Notre approche exploite les syntagmes nominaux dans le cadre d'identification automatique des mots-clés indexant le contenu d'articles scientifiques ayant paru en revues de Sciences Humaines et Sociales, avec l'aide de la terminologie des mots clés (piste1) et sans terminologie (piste2).

ABSTRACT

This paper presents the URPAH team's participation in DEFT 2012.Our approach uses noun phrases in the automatic identification of keywords indexing the content of scientific papers published in a review of Human and Social Sciences, with assistance from the terminology of keywords (piste1) and without terminology (piste2)

MOTS-CLÉS : syntagmes nominaux, patrons syntaxiques, recherche d'information.

KEYWORDS: noun phrases, syntactic patterns, information retrieval.

JEP-TALN-RECITAL 2012, Atelier DEFT 2012: DÉfi Fouille de Textes, pages 33–39,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Cet article décrit la participation de l'équipe URPAH à ce défi. La tâche proposée dans le cadre du Défi Fouille de Texte (DEFT) en 2012 porte sur l'identification automatique des mots-clés indexant le contenu d'articles scientifiques. Nous nous sommes focalisé sur les mots-clés complexes et principalement les syntagmes nominaux (SNs).

2 Mots-clés syntagmatiques

2.1 Indexation à base des mots simple vs indexation à base des syntagmes nominaux

Un système de recherche d'information se pose le problème de reconnaître, au sein d'une collection de documents, les documents significatifs d'un ensemble d'informations. D'une part, l'information disséminée dans un texte n'est pas structurée et donc difficilement accessible et identifiable. D'autre part, les Systèmes de Recherche d'Information (SRI) doivent également offrir une interface d'aide à la formulation des requêtes, pour qu'elle soit une transcription valide du besoin d'information de l'utilisateur.

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'ambiguïté des termes isolés de leur contexte. Une solution à ce problème consiste à utiliser des termes complexes à la place de termes simples isolés (Boulaknadel, 2006). Cette approche se fonde sur l'hypothèse qu'un terme complexe est moins ambigu qu'un terme simple isolé.

2.2 Importance des syntagmes nominaux pour la recherche d'information

Les SRIs actuels se basent toujours sur l'hypothèse initiale qu'un document doit partager les termes d'une requête pour être identifié comme pertinent. Le problème de la RI semble alors se résumer à un simple calcul de correspondance entre un ensemble de mots clés de la requête de l'utilisateur avec l'ensemble des mots clés représentant le document.

Les systèmes de RI présentent des limites associées aux méthodes exploitées pour représenter les contenus textuels. Le passage du document ou de la requête en texte intégral à une représentation en " sac de mots ", telle qu'elle est présentée par la plupart des modèles de RI, implique des pertes d'informations conséquentes. Cette représentation souffre d'un sérieux inconvénient qui est le fait que les termes simples sont souvent ambigus et peuvent se référer à des concepts différents : 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'ambiguïté. Donc les mots simples ne peuvent pas être considérés comme un langage de représentation expressif et précis du contenu sémantique.

En fonction de ces difficultés associées à la complexité du langage naturel, une solution souvent évoquée est d'employer des unités lexicales complexes qui sont plus précises que les unités lexicales simples pour représenter les documents et requêtes afin d'améliorer les performances des SRIs (Haddad, 2003).

2.3 Termes complexes en RI

L'utilisation d'une représentation complexe revient à laisser les mots dans le contexte dans lequel les auteurs les ont écrits, en opposition à l'utilisation de mots simple, où les mots sont détachés de leurs contextes. L'hypothèse est que les termes complexes sont plus aptes à désigner des entités sémantiques (concepts) que les mots simples et constituent alors une meilleure représentation du contenu sémantique des documents (Mitra *et al.*, 1997).

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. Une étude comparative des résultats des approches d'extraction et d'indexation avec des termes complexes (statistique, linguistique et hybride) n'a pas abouti à des conclusions claires en ce qui concerne leur utilité en RI (Mitra *et al.*, 1997).

L'équipe XEROX durant TREC-5 (Hull *et al.*, 1997) a testé l'impact de la reconnaissance de la dépendance syntaxique des mots pour éliminer le bruit dans les couples de mots extraits statistiquement et réduire le silence par la reconnaissance de paires de termes reliés syntaxiquement. Les résultats de ces expérimentations montrent que l'indexation avec des termes complexes extraits syntaxiquement affecte plus positivement les résultats d'un SRI que les groupes de mots extraits statistiquement dans les cas où les requêtes sont longues.

Dans (Haddad, 2002), l'auteur fait l'indexation des documents et des requêtes après l'analyse linguistique et l'extraction des syntagmes nominaux (SNs). Outre l'indexation classique avec des unitermes, il a testé l'indexation des unitermes et des SNs ensemble dans un même vecteur et il a testé aussi l'indexation des unitermes et des SNs séparément. 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 et en particulier, la séparation des unitermes et des SNs dans deux sous-vecteurs différents donne les meilleurs résultats.

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. L'extraction des SNs est faite en utilisant les patrons syntaxiques selon des règles basés sur les catégories grammaticales. Dans ce cas, les SNs sont organisés en réseaux de dépendance syntaxique (tête et expansion/modificateur) en ajoutant les associations statistiques et sémantiques. La mesure de la qualité sémantique permet d'évaluer l'importance d'un terme et sa contribution à la représentation du contenu du corpus. Cette approche combine l'information statistique basée sur le calcul de fréquence de termes et l'information syntaxique sur la structure des SNs dans un réseau de dépendance. L'information sémantique est étudiée à travers les relations : synonymie, hyperonymie, causalité.

La plupart des travaux montrent que l'utilisation des syntagmes offre un avantage pour un SRI (Woods *et al.*, 2000), les auteurs dans (Hull *et al.*, 1997; Mitra *et al.*, 1997; 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.

2.4 Les syntagmes nominaux

Plusieurs travaux menés par des linguistiques ont montré le lien entre SNs et thèmes (ce dont on parle ou ce dont il est question) d'une part, et d'autre part entre syntagmes verbaux et thèmes (ce qu'on en dit ou le propos) (Amar, 2000). Plus précisément, ils s'accordent sur le fait que seuls les groupes nominaux peuvent être des référents (Amar, 2000). C'est pourquoi dans le domaine de la RI, les SNs ont eu plus d'attention puisque c'est le thème qui est intéressant plus que le rhème. Donc dans notre travail, on a choisi les SNs comme représentant des thèmes et comme descripteur au lieu d'utiliser les mots isolés.

Il reste néanmoins très difficile de placer les SNs réellement à un niveau sémantique. De manière pratique, c'est la structure syntaxique qui sert de passerelle vers le niveau sémantique. En effet, les auteurs dans (Carballo et Strzalkowski, 2000) indiquent qu'un traitement linguistique, pour une représentation des documents avec des termes complexes, peut couvrir, contrairement à une représentation avec des mots simples, certains aspects sémantiques du contenu des documents. Nous nous intéressons alors aux SNs au niveau syntagmatique de l'analyse linguistique sans prendre en considération les niveaux sémantique et paradigmatique. Une analyse de surface avec des patrons syntaxiques semble suffisante comme l'atteste les travaux de Debili (Debili, 1982).

Dans le cadre de l'analyse syntaxique d'une phrase, on parle de segmentation en unités fonctionnelles appelées syntagmes. Les syntagmes peuvent avoir la même fonction qu'un mot seul et ils peuvent également inclure un ou plusieurs autres syntagmes. Linguistiquement, un SN peut être caractérisé d'une part par les catégories grammaticales de ces composantes et d'autre part par les règles syntaxiques de l'agencement de ces composantes. Les catégories grammaticales des éléments d'un syntagme nominal sont : substantif, préposition, conjonction, article, adjectif, verbe à l'infinitif, participe passé et adverbe. L'ordre d'enchaînement de ces catégories dans un SN respecte des règles linguistiques qui permettent d'avoir des SNs corrects. A partir de ces deux caractéristiques des SNs, des patrons syntaxiques peuvent être construits (Debili, 1982). Ces patrons décrivent les catégories grammaticales et l'ordre dans le quel les éléments d'un syntagme nominal doivent apparaitre.

2.5 Extraction des syntagmes nominaux

Nous avons opté pour une approche linguistique pour l'extraction de syntagmes nominaux à partir du contenu des articles scientifiques de DEFT 2012. Notre approche vise à extraire les dépendances ou les relations entre termes grâce aux phénomènes langagiers. Nous effectuons d'abord l'analyse linguistique avec un étiqueteur, qui génère une collection étiquetée. Chaque mot est alors associé à une « étiquette » syntaxique. Cette étiquette correspond à la catégorie

syntaxique du mot. 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 ∈ C et X un syntagme nominal.
Exemples :
ADJQ SUBC : « premier ministre", "petite échelle », etc.
SUBC PREP SUBC : « job d'été », « programmes de prévention », etc.

Nous nous basons dans nos travaux sur les 10 patrons syntaxiques les plus susceptibles de contenir le maximum d'information (Haddad, 2002).

3 Description des corpus et résultats

3.1 Corpus d'apprentissage

L'ensemble du corpus d'apprentissage constitue de 281 documents. La tâche se subdivise en deux pistes. La première piste contient 140 documents avec une liste de la terminologie des mots clés. Une deuxième piste contient 141 articles sans terminologie.

3.2 Corpus de test

L'ensemble du corpus de test est constitué de 187 documents. La tâche se subdivise en deux pistes. L'une contient 93 documents avec une liste de la terminologie des mots clés. Une deuxième piste contient 94 articles sans terminologie.

3.3 Résultats et discussion

Le tableau ci-dessous présente la précision, le rappel et la F-mesure pour chaque tâche. Les runs soumis au corpus de test piste 1 avec terminologie (tâche 1) et au piste 2 sans terminologie (tâche 2) obtiennent respectivement des précisions de 0.16 et 0.12.

Notre participation dans l'atelier DEFT 2012 est basée sur une approche automatique d'extraction de mots-clés à base de syntagmes nominaux. L'objectif est d'identifier les mots clés, tels qu'ils ont

	Precision	Recall	F-mesure
Tâche 1	0.1694 [91/537]	0.1694 [91/537]	0.16
Tâche 2	0.1203 [58/482]	0.1198 [58/484]	0.12

TABLE 1 – Résultats

été choisis par les auteurs, pour indexer des articles scientifiques. Dans un premier temps, nous utilisons un analyseur syntaxique pour analyser les documents et étiqueter les mots. Le système utilise la collection étiquetée pour extraire l'ensemble des syntagmes nominaux.

Étant donné que chaque document doit être représenté par un nombre fixe de mots-clés, le système procède alors à un filtrage automatique des SNs. Notre approche de filtrage est basée sur le nombre d'occurrences des SNs dans chaque document. Les SNs les plus fréquents sont alors considérés par le système comme étant les mots-clés. Ce processus automatique de filtrage justifie les résultats de notre approche. En effet, vu la taille relativement petite des documents, les SNs extraits sont d'une part peu fréquents mais aussi d'une autre part possèdent tous le même nombre d'occurrences dans un document. Si le nombre de mots-clés requis pour un documents est n, notre système sélectionne alors les n premiers SNs rencontrés dans le document.

4 Conclusion et perspectives

Dans cette contribution, nous avons présenté une approche d'extraction automatique de mots-clés à base de syntagmes nominaux. Notre approche se base sur un traitement linguistique pour l'extraction des SNs à base de patrons syntaxiques. Les SNs extraits sont alors automatiquement filtrés pour ne garder que le nombre requis de mots-clés pour chaque document. C'est ce processus de filtrage qui sera remis en question dans nos perspectives.

Références

AMAR, M. (2000). Les fondements théoriques de l'indexation : une approche linguistique. ADBS éditions.

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

CARBALLO, J. P. et STRZALKOWSKI, T. (2000). Natural language information retrieval : progress report. *Inf. Process. Manage.*, 36(1):155–178.

DEBILI, F. (1982). Analyse syntaxico-semantique fondee sur une acquisition automatique de relations lexicales-semantiques. habilitation à diriger des recherches.

DIEM, L. T. H. et CHEVALLET, J.-P. (2006). Extraction et structuration des relations multi-types à partir de texte. *In RIVF'06*, pages 53–58, Ho Chi Minh Ville, Viêt-Nam.

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*, pages 277–286, Manaus, Brazil.

HULL, D. A., GREFENSTETTE, G., SCHULZE, B. M., GAUSSIER, E., SCHÜTZE, H. et PEDERSEN, J. O. (1997). Xerox trec-5 site report : Routing, filtering, nlp, and spanish tracks. *In Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, pages 167–180.

MITRA, M., BUCKLEY, C., SINGHAL, A. et CARDIE, C. (1997). An analysis of statistical and syntactic phrases. *In* DEVROYE, L. et CHRISMENT, C., éditeurs : *RIAO*, pages 200–217.

WOODS, W. A., BOOKMAN, L. A., HOUSTON, A., KUHNS, R. J., MARTIN, P. et GREEN, S. (2000). Linguistic knowledge can improve information retrieval. *In Proceedings of the sixth conference on Applied natural language processing*, ANLC '00, pages 262–267, Stroudsburg, PA, USA. Association for Computational Linguistics.