Utilisation des syntagmes nominaux dans un système de recherche d'information en langue arabe

Siham Boulaknadel

LINA FRE CNRS 2729- Université de Nantes
2 rue de la Houssinière, BP 92208 44322 Nantes cedex 03, France
siham.boulaknadel @univ-nantes.fr

RÉSUMÉ. *Dans un contexte riche, un système de recherche d'information doit être capable de trouver les meilleurs résultats possibles. Dans ce but, notre étude s'intéresse aux connaissances qui peuvent être extraites du contenu textuel des documents en associant la finesse d'analyse d'une approche linguistique à la capacité d'une approche statistique traitant des corpus de grandes tailles. L'approche statistique se base sur la fouille de données textuelles et principalement sur la technique d'analyse sémantique latente tandis que l'approche linguistique se base sur les syntagmes nominaux que nous considérons comme des entités textuelles plus susceptibles de représenter l'information contenue dans le texte que les termes simples. Par une expérimentation, sur une collection de documents arabes spécialisés dans le domaine de l'environnement nous montrons l'impact de l'utilisation des syntagmes nominaux sur la précision d'un système de recherche d'information.*

ABSTRACT. *In a rich information context, an information retrieval system must be able to ensure the best results. For this, the aim of our study consists in extracting the knowledge based on document textual contents by associating the analysis smoothness of a linguistic approach to the statistical approach capacity treating large corpus. The statistical approach is based on text mining mainly on the latent semantic analysis technique, while the linguistic approach is based on the noun phrases which are more susceptible to be used like textual entities in representing the text information than the simple terms. By experimentation in Arabic documents, specialized in the environment field, we show the use of noun phrase impact on the information retrieval system precision.*

MOTS-CLÉS : *Recherche d'information, Langue Arabe, Syntagmes nominaux, Pseudo-racinisation, Analyse Sémantique Latente.*

KEYWORDS : *Information retrieval, Arabic Language, Noun Phrase, Stemming, Latent Semantic Analysis.*

1. Introduction

L'objectif d'un système de recherche d'information (SRI) est de retrouver les documents pertinents parmi les premiers résultats car l'utilisateur cherche plutôt la précision dans les réponses et préfère un nombre restreint de documents répondant à son besoin qu'un grand nombre contenant la réponse mais noyée dans un ensemble de documents non pertinents (Mitra et al., 1997).

Une des solutions pour atteindre la précision maximale dans un SRI, est l'utilisation des techniques de racinisation consistant à rechercher les racines lexicales, introduites par les travaux de (Al kharashi, 1991) pour l'amélioration d'un SRI en arabe.

Une autre solution, que nous proposons, est d'associer une approche statistique avec une autre linguistique ; où la première consiste principalement à se servir des relations sémantiques implicites induites par les occurrences des termes dans les documents en utilisant l'analyse de la sémantique latente (Deerwester, 1991), et la deuxième se base sur les syntagmes nominaux que nous considérons comme des entités textuelles plus susceptibles à représenter l'information contenue dans le texte (Amar, 2000).

Dans la suite de notre article, nous exposons les caractéristiques de la langue arabe dans la partie 2, et nous présentons ensuite les principes de la LSA dans la partie 3. Dans la partie 4, nous détaillons la mise en œuvre de notre système de recherche d'information, ainsi que les résultats des expérimentations avant de conclure.

2. Particularité de la langue arabe

La langue arabe s'écrit et se lit de droite à gauche, son alphabet compte 28 consonnes changent de forme de présentation selon leur position. Certaines de ces caractéristiques peuvent être source d'ambiguïté.

Dans ce qui suit nous allons nous limiter aux ambiguïtés qui ont une incidence directe sur la recherche d'information.

2.1. *Voyellation*

Les signes de voyellation, sont notés sous la forme de signes diacritiques placés au dessus ou au dessous des lettres. On constate l'étendue du rôle que jouent les voyelles en arabe, non seulement parce qu'elles enlèvent l'ambiguïté mais aussi elles donnent l'étiquette grammaticale d'un mot indépendamment de sa position dans la phrase. Cependant, elles ne sont utilisées que pour des textes didactiques. Les textes courants rencontrés dans les journaux et les livres n′en comportent habituellement pas.

2.2. *Agglutination*

Contrairement aux langues latines, l'arabe est une langue agglutinante ; les articles, les prépositions et les pronoms collent aux adjectifs, noms, verbes et particules auxquels ils se rapportent ; ce qui engendre une ambiguïté morphologique au cours de l'analyse des mots.

3. Présentation du LSA

La méthode LSA est fondée sur le fait que des mots qui apparaissent dans le même contexte sont sémantiquement proches. Le corpus est représenté sous forme matricielle. Les lignes représentent les mots et les colonnes représentent les différents contextes choisis (un document, un paragraphe, une phrase, etc.). Chaque cellule de la matrice représente le nombre d'occurrences des mots dans chacun des contextes du corpus. Deux mots proches au niveau sémantique sont représentés par des vecteurs proches. La mesure de similarité est généralement définie par le cosinus de l'angle entre les deux vecteurs. L'étude effectuée par (Dumais, 1992) a montré de bons résultats de la méthode LSA par rapport au modèle vectoriel, vu que la LSA est distinguée par une décomposition en valeurs singulières.

4. Mise en œuvre d'un SRI

Les travaux réalisés, au cours de cette étude, ont pour objectif d'évaluer l'impact des connaissances syntagmatiques sur l'analyse sémantique latente pour la recherche d'information dans un corpus en langue arabe.

4.1. *Construction du corpus*

Les textes que nous avons choisis sont extraits d'un ensemble d'articles des sites Web « Al-Khat Alakhdar »[1] et « Akhbar Albiah »[2], dont les sujets couvrent plusieurs thématiques environnementales.
Dans une première étape, notre corpus est constitué de 550 documents et de 15 requêtes de types mot-clé, dont la pertinence est évaluée manuellement. Notre corpus contient 324 338 mots dont 44 325 sont distincts.

4.2. *Protocole expérimental*

Le protocole expérimental comprend une transcription dite de Buckwalter[3], une extraction des syntagmes nominaux, « tokenisation », et une pseudo-racinisation.

[1] http://www.greenline.com.kw
[2] http://www.4eco.com
[3] http://www.ldc.upenn.edu./catalog/catalogEntry.jsp?catalogId=LDC2002L49

4.2.1. *Prétraitement du corpus*

Le prétraitement du corpus permet de formater les données textuelles et de les rendre directement exploitables pour les traitements ultérieurs. La mise en œuvre de notre système de recherche d'information (SRI) est représentée par la figure 1.

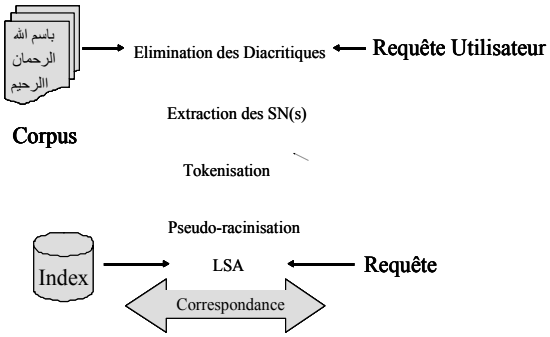

Figure 1. *Schéma d'un SRI pour la langue Arabe*

4.2.1.1. Extraction des SN(s)

Nous nous intéressons aux syntagmes nominaux SN(s) au niveau syntagmatique de l'analyse linguistique sans prendre en considération les niveaux sémantiques et pragmatiques. Pour cela, nous avons utilisé l'analyseur morpho-syntaxique (Diab et al., 2004) qui est un analyseur de surface basé sur un apprentissage supervisé et donne en sortie une collection de textes étiquetés où tous les mots ont été catégorisés.
Une fois que la collection est étiquetée le système extrait l'ensemble des syntagmes nominaux et l'utilise pour l'indexation.

4.2.1.2. Pseudo-racinisation

L'approche que nous avons choisie (Larkey *et* al., 2002) est une analyse morphologique assouplie qui consiste à essayer de déceler les préfixes et les suffixes ajoutés à l'unité lexicale. Notre choix est motivé par le fait que l'indexation par des pseudo-racines donne de meilleurs résultats pour le modèle vectoriel, ce qui est approuvé dans les travaux de (Aljlayl *et al.,* 2002).

4.3. Evaluation

Le protocole va nous permettre d'évaluer l'éventuelle amélioration de la précision du SRI par l'utilisation des SN(s). Nous avons effectué deux expérimentations :

– Indexation avec des SN(s).

– Indexation avec des unitermes.

Pour chaque document ou requête, un nouvel index est crée où les SN(s) extraits sont ajoutés. Ces SN(s) sont alors indexés indépendamment des unitermes en utilisant une pondération Okapi BM-25 (Darwish, 2003).

En comparant les taux de rappel et précision (voir Figure 2), nous pouvons constater que l'utilisation des SN(s) dans l'indexation n'améliore pas les performances obtenues en utilisant les unitermes, ceci confirme les résultats obtenus pour la langue anglaise (Fagan, 1987).

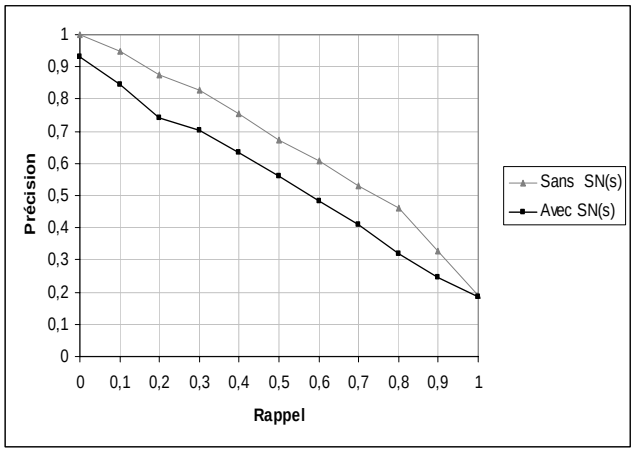

Figure 2. *Influence respective des SN(s) et des unitermes sur le SRI*

En examinant les résultats, nous pouvons expliquer que cette dégradation est due à la taille des SN(s) et aussi au manque de la normalisation de ces derniers ; par exemple «**كارثة تلوث الهواء**» «lit. catastrophe de la pollution de l'air » et « **تلوث الهواء** » «lit. pollution de l'air » qui devrait être normalisé sous le SN « **تلوث الهواء** » « pollution de l'air ». Cela peut être aussi expliqué par le fait que nous avons utilisé un extracteur des SN(s) basé sur l'apprentissage supervisé dépendant du corpus annoté et non pas sur un schémas structuré par les règles syntaxiques.

5.Conclusion et perspectives

Dans cette contribution, nous avons présenté une approche de traitement de données textuelles qui associe la finesse d'analyse d'une approche linguistique à la capacité d'une approche statistique traitant des corpus de grandes tailles.

L'utilisation des SN(s) dans le processus d'indexation n'a pas montré une amélioration des performances du SRI, cela est dû à l'utilisation d'un extracteur des SN(s) basé sur l'apprentissage supervisé, dépendant du corpus annoté.

Ces résultats nous laissent penser qu'une extraction plus fine respectant des schémas syntaxiques pourrait améliorer la performance de notre SRI.

6. Bibliographie

Aljlayl M., Frieder O., « On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach », *In 11 the International Conference on Information and Knowledge Management (CIKM)*, November 2002, Virginia (USA), p. 340-347.

Al Kharashi I., Microcomputer based Arabic Information Retrieval System, Comparing words, stems, and roots as index terms, Doctoral dissertation, Illinois Institute of Technology, University Microfilm Ann Arbor, 1991.

Amar M., Les Fondements théoriques de l'indexation une approche linguistique, ADBS éditions, Paris, 2000.

Darwish K.,Probabilistic Methods for Searching OCR-Degraded Arabic Text, Doctoral dissertation, University of Maryland, 2003.

Deerwester S, Dumais S.T., Furnas G.W., Landauer T.K., Hrashman R., « Indexing by latent semantic analysis », *Journal of the american society for information science*, Vol 41(6), 1990, p. 391-407.

Diab M., Hacioglu K., Jurafsky D., « Automatic tagging of arabic text : from raw text to base phrase chunks », *In HLT-NAACL*, May 2004, p. 149-152.

Dumais S.T., Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval, Technical Memorandum Tm-ARH-017527, 1992, Bellcore.

Fagan J., Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-syntactic methods, Doctoral dissertation, Cornell University, 1987.

Larkey L. S., Ballesteros L. and Connell M., « Improving Stemming for Arabic Information Retrieval : Light Stemming and Cooccurrence Analysis », *In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002)*, Tampere, Finland, August 2002, p. 275-282

Mitra M., Buckley C., Singhal A., Cardi C., «In analysis of statistical and syntactic phrases». *In 5ème Conférence de Recherche d'Information Assistée par Ordinateur (RIAO'1997)*, Montreal, Canada, Juin 1997, p. 200-214.