Utilisation des termes complexes dans un système de recherche d'information en langue arabe

Siham Boulaknadel[1,2], Béatrice Daille[2], Driss Aboutajdine[1]

[1]GSCM-LRIT, Université Mohamed V, BP 1014 Agdal-Rabat, Maroc

[2] LINA FRE CNRS 2729 - Université de Nantes

2 rue de la Houssinière, BP 92 208 44 322 Nantes cedex 03, France

{siham.boulaknadel, beatrice.daille}@univ-nantes.fr

aboutaj@fsr.ac.ma

Abstract

To improve the performance of information retrieval systems, it seems important to identify key phrases which constitute a better representation of text semantic content than single word terms. In this paper, we adopt the standard method for multi-word term extraction for Arabic. We define the linguistic specifications and develop a term extraction tool. We experiment the term extraction program for document retrieval in a specific domain, evaluate two kinds of multi-word term weighting functions, considering either the corpus or the document, and demonstrate the efficiency of multi-word term indexing for both, weighing up to 5.8% of average precision.

Résumé

Pour améliorer les performances d'un système de recherche d'information, il semble intéressant d'identifier les termes complexes qui constituent une meilleure représentation du texte qu'au terme simple. Dans ce papier, nous adaptons la méthode standard pour l'extraction des termes complexes en langue arabe. Nous définissons les caractéristiques linguistiques et nous développons un outil d'extraction des termes. Nous évaluons l'outil d'extraction en recherche d'information dans un domaine spécifique et nous montrons l'impact de l'utilisation des termes complexes sur la précision d'un système de recherche d'information en langue arabe.

Mots-clés : langue arabe, terme complexe, indexation automatique, domaine spécifique.

1. Introduction

L'évolution très rapide d'Internet a permis l'émergence d'un savoir planétaire partagé mais a également généré plusieurs défis. En recherche d'information (RI), ceux-ci sont de trois types, à savoir, la gestion d'un volume d'informations, la présence de multiples supports et, finalement, le caractère plurilingue de la Toile. Dans ce dernier cas, l'importance grandissante d'autres langues que l'anglais a suscité le développement d'outils et de techniques automatiques afin de permettre leur traitement informatique. En comparaison à l'anglais et plus généralement aux langues sémitiques, la langue arabe présente des traits distinctifs, à savoir l'agglutination et la vocalisation.

Face à ces défis et sous l'impulsion des campagnes d'évaluation TREC-2001 (Gey et al., 2001), diverses approches (Larkey et al., 2002) (Aljlayl et al., 2002) ont été développées pour étudier l'apport des connaissances morphologiques, telle que la racinisation, sur des corpus

journalistiques[1] n'affectant pas un domaine spécifique. Dans le cadre de cet article, nous souhaitons étudier l'apport des connaissances syntaxiques à savoir les termes complexes sur la performance des SRI en langue arabe sur un corpus spécialisé d'environnement et en empruntant le protocole d'évaluation de TREC-2001. Pour ce faire, nous avons développé un système d'identification de termes complexes sur corpus qui produit des résultats de bonne qualité en terme de précision, en s'appuyant sur une approche mixte (Cabré et al., 2001) qui combine modèle statistique et données linguistiques pour obtenir des termes complexes pouvant servir à représenter des documents.

2. Spécifications linguistiques des termes complexes

Les termes se répartissent dans deux catégories, celle des termes simples qui sont constitués d'un seul « mot plein »[2] et qui ne peuvent être reconnus qu'en fonction de leur contexte (Jacquemin, 1995). Ils sont le plus souvent polysémiques, même à l'intérieur d'un domaine (Jacques, 2003), et celle des termes complexes qui contiennent au moins deux mots pleins, éventuellement reliés par des « mots grammaticaux »[3].

2.1. Typologie et composition des termes complexes

Nous présentons une étude linguistique sur les termes complexes du domaine de l'environnement. À notre connaissance, il n'existe pas d'étude sur la typologie générale des termes complexes en langue arabe. Il nous a semblé important de préciser quels types élémentaires de termes complexes sont effectivement présents dans ce domaine technique. Ensuite, nous examinons les différentes structures morphosyntaxiques des termes complexes arabes et nous établirons leur classement en fonction de leur structure morphosyntaxique.

2.1.1. Classification des structures élémentaires

La méthodologie que nous avons adoptée est la suivante : nous avons examiné le corpus et extrait des suites de mots susceptibles d'apparaître dans des positions syntaxiques variées et qui appartiennent à l'un des types élémentaires décrits par (Daille, 1994). Les termes complexes épousent des structures morphosyntaxiques exprimées en partie de discours. Pour vérifier le caractère terminologique du candidat terme extrait, deux solutions sont envisagées : la première consiste à utiliser la base terminologique AGROVOC[4] pour attester le candidat terme extrait. Si ce dernier n'existe pas dans la base terminologique, une deuxième solution est envisagée qui consiste à chercher la traduction de ces suites de mots en exploitant la propriété de compositionalité des sens des termes complexes. Ainsi, nous avons tiré profit de leur traduction française pour vérifier leur statut terminologique dans la banque terminologique Eurodicautom[5]. Quelques exemples rencontrés dans notre corpus sont présentés ci-dessous.

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001T55

[2] Mot dont le rôle essentiel est de porter un contenu (nom, verbe, adjectif).

[3] Mots dont le rôle essentiel est d'apporter la cohésion grammaticale de la phrase.

[4] www.fao.org/agrovoc/

[5] http://ec.europa.eu/eurodicautom/Controller

Patrons syntaxiques	Sous- patrons	Terme	Traduction
N ADJ		التلوث الكيميائي	Pollution chimique
N1 N2		تلوث الماء	Pollution de l'eau
	N1 ب N2	التلوث بالرصاص	Pollution au plomb
N1 PREP N2	N1 ل N2	التعرض للأمراض	Exposition aux maladies
	N1 من N2	التخلص من النفايات	Elimination des déchets

Tableau 1. Patrons syntaxiques

2.2. Variation des termes complexes

Pour étayer notre propos, nous présenterons en détail les variations que nous avons rencontrées pour les structures de type élémentaire de l'arabe. Nous avons suivi la typologie proposée par (Daille, 2005).

2.2.1. Variations graphiques

Nous notons comme des variations graphiques, le remplacement de certaine lettre comme ي en fin de terme complexe par la lettre ى. Par exemple, la structure élémentaire N1 N2 apparaît sous deux graphies différentes comme «التلوث الكيميائي » et «التلوث الكيميائى» « pollution chimique ».

2.2.2. Variations flexionnelles

Ces variations regroupent les différentes formes fléchies possibles pour un terme complexe. Les flexions concernent plus particulièrement la mise au pluriel du deuxième nom dans la structure N1 N2 et la définitude exprimée par le préfixe (ال).

Type	Terme complexe	Variante	Traduction
Nombre	تلوث المحيط	تلوث المحيطات	Pollution de l'océan/ Pollution des océan**s**
Définitude	التلوث الهوائي	تلوث هوائي	(**la**) Pollution atmosphérique

Tableau 2. Variantes flexionnelles

2.2.3. Variations morphosyntaxiques

Les variations morphosyntaxiques affectent la structure interne du terme de base, et les mots le composant subissent des modifications relevant de la morphologie dérivationnelle.

Structure	Terme complexe	Traduction
N1 N2 ⇔ N1 ADJ	صناعة بترول ⇔ صناعة بترولية	Industrie pétrole/ Industrie pétrolière
N1 ADJ ⇔ N1 PREP N2	بئر نفطي ⇔ بئر من النفط	Puit pétrolier/ Puit de pétrole

Tableau 3. Variantes morphosyntaxiques

2.2.4. Variations syntaxiques

Les variantes syntaxiques modifient la structure interne de la structure du type élémentaire sans affecter les catégories grammaticales des mots pleins qui restent identiques. Nous distinguons :

Type	Sous-type	Terme complexe	Variante
Modification	Insertion	التكوين للتربة	التكوين المستمر للتربة
		Composition du sol	Composition permanente du sol
	Postposition	درجة الحرارة	درجة الحرارة العالية
		Degré de température	Degré de température élevé
Coordination	Expansion	تلوث التربة	تلوث المياه و التربة
		Pollution du sol	Pollution du sol et des eaux
	Tête	المخاطر من التلوث	المخاطر والوقاية من التلوث
		Risques de la pollution	Risques et prévention de la pollution

Tableau 4. Variantes syntaxiques

3. Méthode d'extraction des termes complexes

L'objectif de notre travail consiste à définir une méthode d'acquisition de termes complexes à partir de corpus pouvant servir à représenter les documents en recherche d'informations. Nous partons de l'idée qu'un texte n'est pas seulement un sac de mots, mais c'est un ensemble fortement structuré de termes qui permettent de communiquer des informations d'une grande précision. Les mots simples ne peuvent pas être considérés comme un langage de représentation expressif et précis du contenu sémantique. Le but que nous nous fixons est d'extraire les termes complexes :

- En adoptant la même approche que dans (Daille, 2005). Nous recherchons les termes qui épousent l'une des structures des termes complexes de l'arabe, ainsi des patrons syntaxiques sont utilisés pour identifier des candidats termes (termes complexes) dans les textes annotés par l'outil de Diab (Diab et al., 2003).
- En utilisant les mesures statistiques (LLR (Dunning, 1994), FLR (Nakagawa et al., 2003), T-score (Church et al., 1991)) pour distinguer parmi ces candidats termes (CTs) lesquels sont effectivement des termes complexes.

4. Expérimentation et Evaluation

4.1. Corpus

Pour démontrer l'intérêt de représenter le contenu textuel par des unités lexicales complexes dans un processus de recherche d'information, nous devons disposer d'un corpus de langue de spécialité riche en terme de variation de genres. A notre connaissance, il n'existe pas un corpus répondant à ces critères. Ainsi, nous avons décidé de construire un corpus à partir du web dans le domaine de l'environnement, restreint aux thématiques suivantes : la pollution, la purification de l'eau, la dégradation du sol, la préservation de la forêt, les catastrophes

naturelles. Elles font l'objet d'une importante production langagière en arabe, comme l'atteste la présence de nombreux sites sur le web. Pour la récolte de documents, nous avons effectué :

- Une recherche sur le web à l'aide du moteur http://www.google.com/intl/ar/ pour l'arabe.
- Une recherche interne sur des portails, notamment « Al-Khat Alakhdar »[6] et « Akhbar Albiae »[7], en utilisant le cas échéant le moteur de recherche propre au site.

Le tableau présente quelques indications des caractéristiques de la collection.

Caractéristiques	Corpus
Nombre de documents	1 062
Nombre total de mots	475 148
Nombre total de mots différents	54 705

Tableau 5. Caractéristiques du corpus

4.2. Evaluation des mesures statistiques

Nous avons mesuré la performance des mesures statistiques décrites précédemment en terme de précision. Le tableau ci-dessous contient les données obtenues à l'aide de ces mesures. La précision est mesurée en fonction du nombre de termes identifiés corrects par rapport à l'ensemble de candidats termes extraits. Nous avons mesuré la performance du système sur les 100 premiers éléments de la liste, classés par chaque mesure statistique.

Mesure	Précision
LLR	85%
T-score	57%
FLR	60%

Tableau 6. Performance des mesures statistiques

Ainsi, dans l'ensemble des documents, un tri à l'aide de la LLR permet d'obtenir une bonne concentration des termes en tête de la liste de candidats termes. Les performances obtenues à l'aide de la LLR nous conduisent à conclure qu'il s'agit d'une mesure permettant de bien cerner le potentiel terminologique de certains des candidats termes recensés.

5. Impact des termes complexes dans un SRI

5.1. Architecture envisagée d'un SRI en langue arabe

L'architecture proposée peut être synthétisée en trois étapes : les documents et requêtes passent dans un premier temps par un module de prétraitement, qui consiste en :

- Elimination des diacritiques.
- Normalisation : La normalisation transforme une copie du document original dans un format standard plus facilement manipulable. Cette étape est considérée nécessaire à cause des variations qui peuvent exister lors de l'écriture d'un même mot arabe. Le document est normalisé comme suit :

[6] http://www.greenline.com.kw

[7] http://www.4eco.com

- Suppression des caractères spéciaux ;
- Remplacement de إ, آ et أ avec ا ;
- Remplacement de la lettre finale ي avec ى ;
- Remplacement de la lettre finale ة avec ه.

- Tokenisation et assignation des étiquettes grammaticales.

Nous avons divisé le corpus en phrases et les mots sont ainsi analysés morphologiquement en utilisant l'étiqueteur de diab (Diab, 2003).

- Indexation.

Nous utilisons une indexation libre en utilisant l'extracteur de termes complexes présenté en section 3.

5.2. Données

Nous utilisons la collection de documents présentés dans la section 4.1. Pour nos expérimentations, un jeu de 30 requêtes a été construit en s'inspirant des campagnes d'évaluation TREC. Elles comportent quatre champs : un titre nommant le thème, une description énonçant complètement l'objet de la recherche, un développement explicitant des critères de validité des rapprochements, des mots-clés fournissant le contexte terminologique et les concepts concernés. Cette forme apporte une information aussi complète et détaillée que possible, y compris des connaissances avancées sur le domaine grâce aux mots-clés.

<titre> حماية الغابات </titre>
<desc> البحث على النصوص التي تتحدث عن حماية الغابات.</desc>
<narr> النصوص ذات صلة بملاحقة قاطعي الأشجار، طرق النهوض بالتشجير ونشر المساحات الخضراء. </narr>
<titre> Préservation de la forêt </titre>
<desc> Trouver les documents qui parlent de la préservation de la forêt. </desc>
<narr> Sont pertinents les documents qui évoquent les manières de favoriser le reboisement, la diffusion des espaces verts et la poursuite des destructeurs des forêts </narr>

Tableau 7. Exemple de requêtes du corpus

5.3. Résultats

Les évaluations que nous présentons ont été effectuées sur les données de notre collection. Outre l'indexation classique avec des unitermes (UT) où il s'agit de trouver les meilleurs résultats suivant les différents paramètres de pondération, nous avons testé les stratégies d'indexation suivantes :

La stratégie (appelée Sm) qui permet d'indexer les unitermes et les termes complexes séparément. Pour chaque document ou requête un nouvel index est crée où les termes complexes extraits sont ajoutés. Ces termes complexes sont indexés indépendamment des unitermes. Cela crée deux sous-vecteurs : le premier correspond aux unitermes et le deuxième aux termes complexes.

La stratégie (appelée Smp) qui emploie un index de termes complexes en pondérant par Okapi BM-25 les termes de la requête pour représenter la présence et l'importance du terme de la requête dans le document donné.

En comparant les taux de rappel et précision de la collection, nous pouvons constater qu'en intégrant des termes complexes dans l'indexation, nous obtenons de meilleures performances par rapport aux meilleurs résultats obtenus par l'utilisation des unitermes (les performances sont exprimées en terme de précision moyenne (Préc.moy) en 11 points de rappel et en pourcentage de variation par rapport aux performances de l'indexation avec des unitermes).

En particulier, la stratégie Smp donne de meilleurs résultats que dans le cas où aucune pondération n'est utilisée. Cette amélioration est perceptible, où la stratégie Sm a permis d'augmenter les performances de 3,6% alors que cette augmentation est de 5,8% dans le cas de la stratégie Smp. En examinant les résultats de la Stratégie Sm et de la Stratégie Smp, nous constatons que le nombre de documents pertinents retrouvés est presque identique pour les deux stratégies. Ce qui diffère est le classement des documents trouvés. En effet, la stratégie Smp permet de favoriser le classement de ces documents en les mettant en tête de la liste des documents trouvés. Ceci se reflète dans les résultats de la précision à faibles taux de rappel (précision à 5, 10 et 20 documents). Ces résultats, particulièrement la précision à 5 et 10 documents, confirme notre hypothèse que les termes complexes aident à augmenter la précision d'un SRI.

	Préc.moy	Diff
UT	26,1%	
Sm	29,7%	+3,6%
Smp	31,9%	+5,8%

Tableau 8. Précision moyenne à 11points de rappel

	5 doc		10 doc		20 doc	
	Préc.moy	Diff	Préc.moy	Diff	Préc.moy	Diff
UT	51,6%		43,9%		40,8%	
Sm	53,2%	+1,4%	45,1%	+1,2%	42,1%	+1,3%
Smp	58,3%	+6,7%	45,8%	+1,9%	42,2%	+1,4%

Tableau 9. Précision moyenne à 5, 10, 20 documents

L'intégration des termes complexes dans le processus d'indexation a montré une influence réelle sur les performances d'un SRI. Particulièrement, l'indexation des termes complexes en pondérant les requêtes, a révélé de meilleurs résultats. Ces constatations confirment notre hypothèse avancée que l'utilisation des termes complexes constitue une représentation plus précise du contenu des documents que les unitermes et renforce notre approche d'indexation en considérant les termes complexes comme support des termes d'indexation.

6. Conclusion

Dans ce papier, nous présentons l'apport des termes complexes dans un SRI en langue arabe. Ainsi, nous définissons les spécifications linguistiques des termes complexes en arabe et nous développons un outil d'extraction de ces termes. Les termes sélectionnés sont ensuite utilisés dans un système de recherche d'information en langue arabe. Nos expériences ont été effectuées sur un domaine spécialisé dans le domaine de l'environnement, nous avons comparé différents types d'indexation, et nous avons conclu qu'une indexation qui combine termes complexes et la pondération Okapi-BM25 a permis d'augmenter les performances du SRI en terme de précision/rappel.

Bibliographie

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

Cabré M. T., Bagot R. E. and Platresi J. V. (2001). Automatic term detection: A review of current systems. In Bourigault D., Jacquemin C., L'Homme M. C. (eds.), *Recent Advances in Computational Terminology*. Volume 2 of Natural Language Processing. John Benjamins, pages 53-88.

Church K., Gale W., Hanks P. and Hindle D. (1991). Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. U. Zernik, 115-164.

Daille B. (2005). Variations and application-oriented terminology engineering. *International journal of theoretical and applied issues in specialized communication*, Vol. (11): 181-197.

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

Diab M., Hacioglu K. and Jurafsky D. (2004). Automatic tagging of arabic text: From raw text to base phrase chunks. In *Proceedings of HLT-NAACL 2004*, Boston, pages 149-152.

Dunning T. (1994). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, Vol. (19): 61-74.

Gey F.C., Oard D.W. (2001). The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries. In *Proceedings of the 2001Text Retrieval Conference (TREC-2001)*. National Institute of Standards and Technology, pages 16-26.

Jacques M.P. (2003). *Approche en discours de la réduction des termes complexes dans les textes spécialisés*. PhD thesis, Université de Toulouse II, France.

Jacquemin C. (1995). Etat de l'art sur l'analyse des noms composés et des termes. *Technical Report* 89, Institut de Recherche en Informatique de Nantes, Nantes.

Larkey L.S., Ballesteros L. and Connell M. (2002). 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, pages 275-282.

Nakagawa H., Mori T. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, Vol.(9), pages 201-219.