

Apport des ontologies dans un système de recherche d'informations arabe

Med El Amine ABDERRAHIM
Université Abou Bekr Belkaid Tlemcen

Résumé

Dans le cadre de la recherche d'information pour les textes en langue arabe, la récupération de mots clé est jugée insuffisante, car les termes utilisés dans la requête peuvent présenter par rapport aux documents de la base, des différences sur plusieurs plans, par exemple:

- des variations morphologiques comme dans « مدرسة » et « مدرستان », « خيل » et « خيول »;
- des variations lexicales (on utilise pour le même sens des mots différents) comme dans le cas dans « فرس » et « خيل »;
- des variations sémantiques comme dans le cas de « الحجر: مرادف الصخر » et « الحجر: أنثى الخيل ».

L'utilisation des ontologies pour l'enrichissement (expansion) de la requête utilisateur peut constituer une solution (parmi d'autres) pour résoudre le problème des variations sémantiques, en effet, les ontologies offrent des ressources sous la forme de relations sémantiques, ils permettent d'étendre le champ de recherche d'une requête, ce qui a pour conséquence l'amélioration des résultats de la recherche.

Dans ce modeste travail, nous proposons la réalisation d'une interface utilisateur qui emploie un analyseur morphologique pour récupérer les formes de base des mots arabe présents dans la requête de l'utilisateur, pour ensuite faire appel à WordNet Arabe pour enclencher le processus d'expansion. La requête ainsi étendue est envoyée à Google.

Mots Clés : TALN Arabe, Recherche d'Information Arabe, expansion de la requête, Wordnet Arabe.

1. Introduction

L'utilisation des ontologies dans un Système de Recherche d'Information (SRI) peut être envisagée à trois niveaux :

- L'indexation des documents et requêtes se fait en utilisant les concepts de l'ontologie et non pas les mots clés.
- Avant d'être envoyée, la requête de l'utilisateur pourra être enrichie par les concepts jugés proches dans l'ontologie et ceci, par le biais de l'utilisation des relations comme la généralisation/spécialisation, la synonymie...
- Le filtrage des documents selon un domaine particulier pour des profils d'utilisateur. [7], [8], [15], [13], [5], [19].

La qualité des réponses obtenues par un SRI ne dépend pas seulement de la qualité du processus l'appariement requête/documents mais aussi de la requête formulée par l'utilisateur, d'où l'intérêt de la reformulation de la requête.

Dans cette optique, les travaux de recherche distinguent deux approches pour la reformulation d'une requête:

a) La reformulation directe : elle consiste à ajouter de nouveaux termes à la requête initiale en s'appuyant sur les liens de co-occurrences entre les termes ou sur des ressources externes telles que les ontologies et les thesaurus. On parle dans ce dernier cas de reformulation de requêtes basée sur les concepts «Concept-based Query Reformulation».

b) La reformulation indirecte : elle consiste à modifier la requête de l'utilisateur en tenant compte d'une liste de documents déjà jugés sélectionnés. Ce processus est appelé réinjection de la pertinence «relevance feed-back» si le processus est supervisé et de pseudo réinjection de pertinence

« blind relevance feedback » si le processus est automatique. Il est clair que les termes rajoutés à la requête, dans le cadre de cette approche, proviennent seulement des documents de la collection.

Les travaux sur les SRI pour les textes arabes ne sont pas nombreux à notre connaissance. Parmi ces travaux on trouve :

- le système de Bessou [10]. Ce dernier adopte la notion de schème comme base pour lemmatiser les mots et les substituer par leurs lemmes dans les opérations d'indexation et de recherche. Ce système n'a pas été évalué par ses auteurs.
- Le système de Farag [12]. Il propose d'assister l'utilisateur dans la formulation de sa requête par l'utilisation des modèles de n-gram. Pour les opérations d'indexation et de recherche, Farag utilise les services du moteur de recherche Google.
- Le système de Abderrahim [4]. Ce système utilise un analyseur morphologique pour récupérer les formes de base des mots arabe présents dans la requête de l'utilisateur, pour ensuite faire appel à WordNet Arabe pour enclencher le processus d'expansion. Dans cette étude l'évaluation de l'apport réel de la reformulation de la requête n'a pas été effectuée pour des raisons de complexité de la tâche.

Cet article constitue une extension des travaux présentés dans Abderrahim [4]. Il entre donc dans le cadre de l'assistance de l'utilisateur par l'amélioration de sa requête (reformulation directe) en utilisant un analyseur de la langue arabe et une ressource lexicale externe (Wordnet Arabe). Dans ce qui suit nous allons décrire l'architecture, le mode de fonctionnement et l'expérimentation de notre module pour l'enrichissement de la requête Arabe.

2. Architecture et fonctionnement du module d'enrichissement

Notre système d'enrichissement de la requête utilisateur (voir la figure 1) se compose de deux modules importants: l'analyseur morphologique et le module de recherche des concepts à partir de Wordnet Arabe.

En effet, Wordnet Arabe est une base de données lexicale librement disponible pour l'arabe standard. Cette base de données suit la conception et la méthodologie du Princeton Wordnet pour l'anglais et d'EuroWordnet pour les langues européennes. Sa structure est celle d'un thésaurus, il est organisé autour de la structure des synsets, c'est-à-dire des ensembles de synonymes et de pointeurs décrivant des relations vers d'autres synsets. Chaque mot peut appartenir à un ou plusieurs synsets, et à une ou plusieurs catégories du discours. Ces catégories sont au nombre de quatre: nom, verbe, adjectif et adverbe. Wordnet est donc un réseau lexical dont les synsets sont les nœuds et les relations entre synsets sont les arcs. Il faut noter toutefois que Wordnet Arabe est une des rares ressources pour la langue générale arabe disponible en ligne. Il compte actuellement¹ 11269 synsets et 23481 mots. [14], [16], [11], [17], [20], [21].

L'entrée de l'analyseur morphologique est une requête écrite en arabe. Un premier traitement consiste à segmenter la requête en formes. Le séparateur blanc étant la marque des frontières des formes, ce traitement ne devra donc poser aucun problème. Le deuxième traitement permet de supprimer les mots outils (mots vides) de la requête en consultant un lexique prédéfini des mots outils. Le troisième et le dernier traitement réalise les

¹Décembre 2009 ; voir la page <http://www.lsi.upc.edu/~mbertran/>

taches de segmentation et d'analyse de chaque forme de la requête. L'opération de segmentation de la forme procède par l'accès aux différentes tables (clitiques et affixes) pour détecter la présence de proclitique, enclitique, préfixe et suffixe dans la forme. Le résultat étant un ensemble de cinq segments (proclitique, préfixe, radical, suffixe, enclitique). Par ailleurs, l'opération d'analyse effectue un accès au dictionnaire des formes simples pour vérifier l'existence du radical. S'il existe, l'analyseur lui associe l'ensemble de ses informations linguistiques (la base, racine...) [2]. A l'issue de l'analyse morphologique de la requête, l'analyseur produit un ensemble d'informations (base, racine, catégorie grammaticale, ensemble de traits syntaxiques...) qui représente la solution morphologique hors-contexte calculée dans le modèle linguistique utilisé. Toutefois, ce qui nous intéresse dans le cadre de notre étude est évidemment l'ensemble des formes de base composant la requête initiale.

Le module de recherche des concepts à partir de Wordnet permet de construire la liste des concepts les plus proches pour chaque forme de base de la requête. Par exemple, à partir de la forme «هجرة» on construit la liste suivante «انتقال ،

«تنقل ، سفر رحلة ، مسيرة»

Il faut de toute façon faire remarquer que notre système utilise deux ressources de données:

- une base de données linguistique (contenant les différents lexiques ainsi que l'ensemble des clitiques et affixes propres à la langue arabe) utilisée par l'analyseur morphologique ; son contenu est détaillé dans [1], [2],[3].

- Wordnet Arabe.

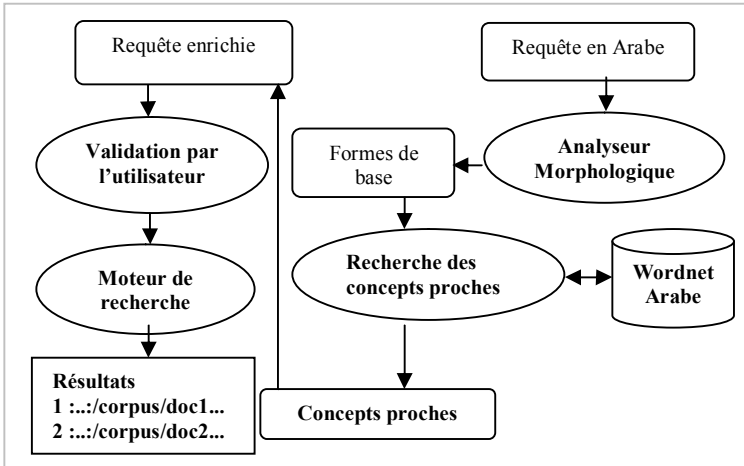


Figure 1 : Architecture de l'interface de recherche

Après la saisie de la requête, le texte de cette dernière est envoyé à l'analyseur morphologique pour produire une liste de formes de base qui va servir à générer une liste des concepts proches en utilisant wordnet arabe. La liste ainsi générée avec le texte de la requête initiale forment le texte de la requête enrichi. Cette dernière est envoyée au moteur de recherche après sa validation par l'utilisateur. En ce qui concerne le moteur de recherche, nous avons utilisé une API libre (disponible gratuitement en ligne) fourni par yahoo pour l'interrogation de notre base de données documentaire et la récupération des résultats.

3. Expérimentation et discussion

Pour tester l'apport de l'enrichissement de la requête, nous avons construit une base de données documentaire (corpus) de 89,2 Mo de taille et comprenant 10200 documents (fichier de

texte en Arabe) issus de différents domaines (histoire, science, politique, sport, religion,...). Nous avons ensuite construit 20 requêtes pour les soumettre au moteur de recherche (voir tableau n°1 pour des exemples de requêtes). Ces requêtes sont traitées par l'analyseur morphologique et par le module de recherche des concepts proches pour produire les requêtes enrichies (voir tableau n°1 pour des exemples de requêtes enrichies).

N° requête	Requête	Requête enrichie
3	إِنْتِهَاك	إِنْتِهَاك, مُخَالَفَة, تَعَدٍّ, تَجَاوُز, خَرَق
5	إِهْدَار	إِهْدَار, إِسْرَاف, تَبْذِير
12	إِبْدَاع	إِبْدَاع, إِبْتِكَار, تَكْوِين, خَلْق
15	تَجْهِيْز	تَجْهِيْز, تَوْفِيْر, تَرْوِيْد

Tableau 1: Exemple de requêtes

Dans un premier temps nous avons indexé l'ensemble des documents de notre corpus. Le résultat de cette opération nous a permis d'obtenir un index d'une taille de 18,6 Mo. Cet index sera utilisé pour faire la recherche des informations en réponse aux 20 requêtes.

Dans le cadre de cette expérimentation et après la soumission des requêtes initiales (requête avant) et enrichies (requête après) au moteur de recherche nous avons enregistré dans le tableau suivant (voir le tableau N°2) le nombre de documents trouvés avant et après enrichissement ainsi que le nombre de documents pertinents avant et après enrichissement.

N° requête	Nb de doc trouvés Avant (1)	Nb de doc pertinents Avant (2)	Nb de doc trouvés Après (3)	Nb de doc pertinents Après (4)
1	2207	1067	3913	1078
2	5066	225	5407	279
3	745	13	810	16
4	1322	201	2839	312
5	169	11	951	43
6	2057	114	4124	239
7	2121	984	2383	1038
8	1850	582	7120	1297
9	4531	953	5925	998
10	4661	435	5480	616
11	2672	593	8070	1075
12	920	367	3013	828
13	5751	831	7371	1140
14	6791	1370	7745	1441
15	920	459	7636	1440
16	2017	710	4618	1132
17	1696	29	7253	83
18	2208	100	2352	124
19	367	12	1992	93
20	146	54	1536	56

Tableau 2: Résultats de l'expérimentation

Avec une simple comparaison entre la colonne (1) et la colonne (3) du tableau 2, nous pouvons déduire que l'enrichissement de la requête améliore le nombre de documents retournés. La comparaison de la colonne (2) et (4) prouve aussi qu'il y a une

amélioration au niveau des documents pertinents trouvés après enrichissement. Ce qui démontre que la reformulation de la requête, par utilisation de l'ontologie Wordnet Arabe, est très bénéfique, elle permet donc d'améliorer les performances d'un SRI Arabe en diminuant le silence. Cependant, nous remarquons qu'il y a une augmentation du nombre de documents retrouvés mais non pertinents, ce qui implique que l'enrichissement de la requête produit aussi du bruit dans le système. Par exemple la reformulation de la requête N°20 a produit 1480 (1536-56) documents retournés non pertinent contre seulement 92 (146-54) documents retournés non pertinents avant enrichissement.

Dans le cadre de cet expérience, un problème est donc traité, il s'agit de palier le problème des variations lexicales, autrement dit, notre module interroge wordnet arabe pour récupérer des mots différent lexicalement, mais reliés à ceux de la requête initiale par des relations sémantiques, telles que la synonymie, la généralisation et la spécialisation. Un problème que nous n'avons pas traité dans le cadre du présent travail s'est posé au niveau du choix du sens (synset) à prendre dans le cas de la polysémie dans les mots de la requête.

L'évaluation de l'apport réel de l'enrichissement de la requête arabe est une tâche très délicate et demande par conséquent beaucoup d'investigations. Toutefois si nous allons se baser sur notre expérimentation nous pouvons dire que l'apport des ontologies dans le domaine de la RI se caractérise principalement par la réduction du silence dans les réponses aux requêtes utilisateurs et par la simplification de la

tâche d'expression de la requête (assistance dans la formulation de la requête).

Comme perspectives, il nous reste maintenant les tâches suivantes :

- fixer le nombre de concepts qui doivent être choisis pour chaque relation utilisée (*synonymie*, *hypernymie*, *hyponymie...*),
- fixer le nombre de termes formant un concept d'extension résultant d'une relation sémantique,
- fixer les poids à affecter aux mots des concepts résultant de l'extension,
- l'étude de l'influence de l'utilisation des concepts composés de l'ontologie,
- l'étude de l'apport de chaque relation sémantique utilisée dans le processus d'enrichissement.

4. Conclusion

L'idée de cet article est d'exploiter une ressource lexicale et un analyseur morphologique pour reformuler (par expansion) la requête de l'utilisateur afin d'améliorer les résultats de la recherche. Le mode de fonctionnement de cette approche se résume ainsi : on commence la phase d'expansion de la requête en analysant les mots de la requête à l'aide de l'analyseur morphologique. Wordnet arabe est ensuite interrogé pour récupérer une liste des termes reliés à la requête par des relations de synonymie, généralisation et spécialisation. L'ensemble des termes constitués par cette dernière liste ainsi que la liste des mots de la requête initiale forme ainsi la requête enrichie qui sera validée par l'utilisateur et envoyée au moteur de recherche.

Pour tester cette approche nous avons utilisé un moteur de recherche avec wordnet arabe. Les résultats d'expérimentation confirment le fait que les requêtes étendues permettent d'élargir le champ de recherche dans un SRI arabe. En revanche, il apparaît plus prudent d'étudier avec plus de détail le problème du bruit dans les résultats renvoyés par le SRI pour pouvoir juger objectivement l'apport réel de l'approche dans l'amélioration de la qualité d'un SRI arabe.

Beaucoup de tests et d'améliorations restent donc à faire et comme travail actuel, nous sommes en train d'augmenter le nombre de requêtes et de varier le corpus de texte arabe.

Bibliographie

1. Abderrahim, M. El A., & al.: Un modèle objet pour le traitement automatique de l'arabe voyellé ou non. JeTIC'2007. Bechar 21/22 avril (2007)
2. Abderrahim, M. A. : Un analyseur morphologique pour l'arabe voyellé ou non. SIIE'2008 : 1ère Conférence Internationale, Systèmes d'Information et Intelligence Economique, SIIE 2008 Hammamet – Tunisie, 14-16 Février 2008, Proceedings tome II, IHE éditions, ISBN 9978-9973-868-20-6, pp. 324--339 (2008)
3. Abderrahim, M. A. : Vers un dictionnaire unifié pour le TALN arabe. 2ème Colloque International de Traductologie et TAL, Oran les 7 et 8 juin (2008)
4. Abderrahim, M. A.: Vers la recherche d'information de contenus en arabe fondée sur l'enrichissement des requêtes. SIIE'2009 : 2ème Conférence Internationale, Systèmes d'Information et Intelligence Economique, SIIE 2009 Hammamet – Tunisie, 12-14 Février 2009, Proceedings IHE éditions, ISBN 9978-9973-868-21-3, pp. 598--607 (2009)
5. Abouenour, L., Bouzoubaa, K., Rosso, Paolo : Système de Question/Réponse dans le cadre d'une plateforme intégrée : cas de l'Arabe. Rencontre Nationale en Informatique : Outils et applications. RNIOA'08 Les 05, 06 et 07 Juin (2008)

6. Attia, M. A.: Arabic Tokenization System. ACL-Workshop on Computational Approaches to Semitic Languages. Prague. (2007)
7. Baziz, Mustapha: Application des Ontologies pour l'Expansion de Requêtes dans un Système de Recherche d'Informations. Rapport de DEA Informatique de l'Image et du Langage (2IL). Université Paul Sabatier & Institut National Polytechnique de Toulouse, Année 2001/2002 (2002)
8. Baziz, Mustapha : Indexation conceptuelle guidée par ontologie pour la recherche d'information. Thèse de doctorat, université Paul Sabatier (2005)
9. Ben Othmane, C. Z.: De la synthèse lexicographique à la détection et la correction des graphies fautives arabes. Thèse université de Paris-Sud décembre (1998)
10. Bessou, Sadik, Saadi, Abdelahlim, Touahria, Mohamed : Vers une recherche d'information plus intelligente application à la langue arabe. SIIE'2008 : 1ère Conférence Internationale Systèmes d'Information et Intelligence Economique SIIE 2008 Hammamet, Tunisie, 14-16 Février (2008)
11. Christiane, Fellbaum, William, Black, Sabri, Elkateb, Antonia, Marti, Adam, Pease, Horacio, Rodriguez, Piek, Vossen : Constructing Arabic WordNet in Parallel with an Ontology. <http://www.globalwordnet.org/AWN/meetings/meet20050901/Fellbaum.ppt> (2005).
12. Farag, Ahmed, Andreas, Nürnberger: AraSearch: Improving Arabic text retrieval via detection of word form variations. SIIE'2008, 1ère Conférence Internationale Systèmes d'Information et Intelligence Economique, SIIE 2008 Hammamet – Tunisie, 14-16 Février (2008)
13. Haïfa, Zargayouna: Indexation sémantique de documents XML. Thèse de Doctorat de l'université Paris XI Orsay, Décembre (2005)
14. Horacio, Rodríguez, Sabri, Elkateb, William, Black, Piek, Vossen, Adam, Pease, Christiane, Fellbaum: Building a WordNet for Arabic, <http://www.adampease.org/Articulate/publications/LREC.pdf> (2006)
15. Mezaour, Amar-Djalil: Recherche ciblée de documents sur le web. Thèse de Doctorat de l'Université Paris_Sud, juin (2005)
16. Musa, Alkhalifa. : Arabic WordNet and Arabic NLP. JETALA 5-7 June, Rabat (2006)

17. Sabri, Elkateb, William, Black, Piek, Vossen, David, Farwell, Adam, Pease, Christiane, Fellbaum.: Arabic WordNet and the Challenges of Arabic. <http://www.mt-archive.info/BCS-2006-Elkateb.pdf> (2006).
18. Saidi, M. El F., Abderrahim, M. A. E. : El-WAFI : un méta moteur base sur l'enrichissement de la requête arabe. Mémoire d'ingénieur d'état en informatique, université de Tlemcen, Algérie, juin (2008)
19. Sais, Fatiha : Transformation d'informations structures en documents XML guidée par une ontologie. Mémoire de DEA Information-Interaction-Intelligence, université Paris-Sud, septembre (2004)
20. William, J., Black, Sabri, El-Kateb: A Prototype English-Arabic Dictionary Based on WordNet. <http://www.fi.muni.cz/gwc2004/proc/95.pdf> (2004)
21. William, BLACK, Sabri, ELKATEB, Horacio, RODRIGUEZ, Musa, ALKHALIFA, Piek, VOSSSEN, Adam, PEASE, Christiane, FELLBAUM: Introducing the Arabic WordNet Project .
<http://www.globalwordnet.org/AWN/meetings/GWApaper.pdf>

