Recherche d'information en langue arabe : influence des paramètres linguistiques et de pondération de LSA

Siham Boulaknadel (1,2), Fadoua Ataa-Allah (2)

(1) LINA FRE CNRS 2729 – Université de Nantes
2 rue la Houssinière BP 92208 44322 Nantes cedex 03, France
siham.boulaknadel@univ-nantes.fr
(2) GSCM – Université Mohammed V
BP 1014 Agdal Rabat-Maroc
fadoua_01@yahoo.fr

Mots-clefs – Keywords

Recherche d'information, Analyse de la sémantique latente, Langue arabe, Racinisation

Information retrieval, Latent semantic analyses, Arabic language, Stemming

Résumé – Abstract

Nous nous intéressons à la recherche d'information en langue arabe en utilisant le modèle de l'analyse sémantique latente (LSA). Nous proposons dans cet article de montrer que le traitement linguistique et la pondération des unités lexicales influent sur la performance de la LSA pour quatre cas d'études : le premier avec un simple prétraitement des corpus; le deuxième en utilisant un anti-dictionnaire; le troisième avec un racineur de l'arabe ; le quatrième où nous avons combiné l'anti-dictionnaire et le racineur. Globalement les résultats de nos expérimentations montrent que les traitements linguistiques ainsi que la pondération des unités lexicales utilisés améliorent la performance de LSA.

We are interested in information retrieval in Arabic language by using latent semantic analysis method (LSA). We propose in this article to show that the linguistic treatment and weighting of lexemes influence the performance of LSA. Four cases are studied: the first with a simple pretreatment of the corpora; the second by using a stopword list; the third with arabic stemmer; the fourth where we combined stopword list and arabic stemmer. Broadly the results of our experiments show that the linguistic treatments as well as weighting of lexemes used improve the performance of LSA.

1 Introduction

En recherche d'information, le problème d'accès au texte est essentiellement dû à l'écart entre les termes utilisés dans les requêtes et les documents. L'appariement entre requête et document se fait donc par l'intermédiaire de leur représentation respective. Le modèle de recherche le plus souvent utilisé est le modèle vectoriel (Salton, 1983). Un des problèmes de ce modèle réside dans l'hypothèse d'indépendance faite sur les termes d'indexation : chaque terme d'indexation constitue une dimension de l'espace vectoriel, sans considération d'éventuelles relations entre termes.

En l'absence d'une connaissance approfondie de la collection de documents, la requête peut être formulé en des termes proches mais non identiques à ceux employés dans un document. Un certain nombre de chercheurs se sont intéressés à ce problème, soit par l'utilisation de réseaux sémantiques qui consiste à recourir à une base de connaissances linguistiques regroupant les mots sémantiquement proches et structuré selon des relations hyperonymiques et/ou synonymiques (Grefenstette, 1994), soit par l'extension de requêtes, opération par laquelle un certain nombre de termes issus de documents de la collection sont ajoutés à une requête.

La troisième possibilité que nous avons choisie, consiste à se servir des relations sémantiques implicites induites par les cooccurrences entre termes dans les documents. Ainsi le modèle de l'analyse sémantique latente (LSA) (Deerwester et al., 1990) consiste à réduire le nombre de dimensions de l'espace vectoriel en s'appuyant sur le fait que les documents traitant des mêmes sujets ont des vocabulaires proches et sont donc proches dans l'espace vectoriel.

Dans notre travail, nous avons sélectionné les schémas de pondération qui améliorent la performance de la méthode LSA pour le calcul de similarité, dans le cas de cinq corpus de petites tailles en langue arabe, tout en évaluant l'importance de différents paramètres linguistiques utilisés.

2 Présentation de LSA

L'analyse sémantique latente (LSA) consiste à réduire le nombre de dimensions de l'espace vectoriel par le biais d'une décomposition en valeurs singulières (SVD), de la matrice A en un produit de trois autres matrices : $\mathbf{A = U\ S\ V^T}$

Où U est une matrice orthogonale de taille (m x n) de description d'unité lexicale, V est une matrice orthogonale de taille (n x n) de description d'unité textuelle et S une matrice diagonale de taille (n x n).

À partir d'un certain nombre k<n, nous nous apercevons de l'existence de valeurs singulières très faibles et qui peuvent être négligées dans la matrice.

De ce fait, il est démontré qu'il y a une meilleure approximation Ak de A qui est donnée par :

$$\mathbf{A_k = U_k\ S_k\ V_k^T}$$

Cette réduction va permettre de ne garder que les unités lexicales les plus significatives. À noter que k est déterminé de façon empirique en fonction du corpus utilisé et du degré de performance voulu.

Pour évaluer la performance de la LSA on utilise les deux mesures traditionnelles de précision et de taux de rappel (Salton, 1989).

3 Paramètres de pondération

La pondération des unités lexicales consiste à transformer l'occurrence d'une unité lexicale dans l'unité textuelle par une combinaison de pondérations locales L(i,j), indiquant l'importance de l'unité lexicale i dans l'unité textuelle j et pondérations globales G(i), indiquant l'importance de l'unité lexicale i dans l'ensemble des unités textuelles de la collection.

Avec f_{ij} la fréquence de l'unité lexicale i dans l'unité textuelle j, df_i le nombre d'unités textuelles auxquelles l'unité lexicale i appartient, gf_i le nombre total de fois où l'unité lexicale i apparaît dans la collection, N est le nombre d'unités textuelles, M le nombre des termes dans le corpus et p_{ij} est le rapport de f_{ij} par gf_i.

Pondération globale

Nom du schéma	Formule	Intérêt
Entropie	$1 - \sum_{j}^{N} \frac{p_{ij} \log p_{ij}}{\log(N)}$	Elle tient compte de la distribution des unités lexicales dans les unités textuelles et permet d'attribuer un poids minimum aux termes qui sont distribués de la même façon dans toutes les unités textuelles et un poids maximum aux termes qui sont concentrés dans quelques unités textuelles
Normal	$\frac{1}{\sum_{j}^{N} f_{ij}^{2}}$	Elle a pour effet de donner un poids élevé aux termes peu fréquents et elle ne dépend que de la somme des fréquences au carré et pas de la distribution de ces fréquences.
Gfldf	$\frac{gf_i}{df_i}$	Elles pondèrent tous deux les termes par le nombre des unités textuelles différentes dans lesquelles ils apparaissent. La différence entre les deux c'est que GfIdf augmente le poids des mots fréquents.
Idf	$\log_2\left(\frac{N}{df_i}\right)$	

Figure 1 : Paramètres de pondération utilisés

4 Traitements linguistiques

L'arabe est une langue sémitique s'écrivant de droite à gauche elle comporte 28 consonnes et 6 voyelles standard (3 longues : ا و ي et 3 courtes : ُ / ِ / َ). Le traitement automatique de l'arabe est difficile vu ses variations orthographiques et sa structure morphologique complexe.

Deux approches sont utilisées dans l'analyse morphologique de l'arabe, la première que nous avons choisie (Darwish, 2002) est une analyse morphologique assouplie ou racinisation qui consiste à essayer de déceler si des suffixes ou préfixes ont été ajoutés à l'unité lexicale : par exemple pour le duel (ان) dans (معلمان, deux professeurs), le pluriel des noms masculins (ين , ون) dans (معلمون, des professeurs) et féminins (ات) dans (مسلمات, musulmanes) ; la forme possessive (نا, كم, هم) dans (كتابهم, ses livres) et les préfixes dans les articles définis (ال, وال, بال, كال, فال).

La deuxième est une lemmatisation qui consiste à réduire les formes déclinées à une représentation canonique.

5 Expérimentations

Notre objectif est de sélectionner les schémas de pondération qui améliorent la performance de la méthode LSA pour le calcul de similarité, dans le cas des corpus de petite taille, tout en évaluant l'importance de l'utilisation d'un anti-dictionnaire et d'un racineur.

5.1 Données

Afin de bien évaluer nos résultats sur les corpus de petite taille, nous avons choisi sur Internet une version arabe des contes partiellement voyellés de 1800 mots : « Le paysan énergique »[1], de H.Darwish « Fleurs du miel »[2] , «Musique de la nature »[3] et « Sous les branches »[4] de K.Abid et «Les chaussures en bois»[5] de J.alhamad. Nous avons appliqué la transcription de Buckwalter qui consiste à transcrire l'alphabet arabe en alphabet latin (Buckwalter, 2002). Nous avons décidé de construire un anti-dictionnaire général qui contient l'ensemble des unités lexicales grammaticales extraites du dictionnaire arabe[6] ensuite nous avons choisi de formuler les requêtes avec aussi peu de variations que possible par rapport à la formulation d'origine. L'ensemble des requêtes que nous avons établi est de l'ordre de 71 requêtes.

5.2 Calculs des courbes

Nous avons segmenté par la suite chacun de ces contes en paragraphes, ce qui nous a permis de construire cinq corpus dont le nombre d'unités textuelles (paragraphes) varie entre 8 et 24. Après avoir transformer ces contes et requêtes textuelles en mode vectoriel, nous avons calculé la précision moyenne sur l'ensemble des requêtes de chaque corpus, en faisant varier k de 2 à r (le rang de la matrice correspondant à chaque corpus). Les paragraphes retournés étaient ceux dont le vecteur faisait un angle de cosinus supérieur à un seuil de 0.9 avec les vecteurs requêtes.

5.3 Résultats

Nous avons effectué des tests pour vingt trois schémas de pondération, plus un autre test où nous avons utilisé la méthode LSA sans appliquer aucune pondération à la matrice originale. Effectivement, d'après les tests appliqués sur le corpus « Musique de la nature », nous avons remarqué que la performance de la LSA sans pondéré la matrice originale est relativement

---

[1] http://www.awu-dam.org/book/99/child99/5-a-d/book99-ch008.htm
[2] http://www.awu-dam.org/book/99/child99/11-h-a/book99-ch001.htm
[3] http://www.awu-dam.org/book/99/child99/11-h-a/book99-ch003.htm
[4] http://www.awu-dam.org/book/99/child99/11-h-a/book99-ch012.htm
[5] http://www.comp.leeds.ac.uk/latifa/research.htm
[6] http://www.almeshkat.net/books/archive/books/muajm arabia.zip

inférieure de 6% par rapport à celle où nous appliquons le schéma de pondération 'Pondération Local Logarithmique * Entropie de Dumais' ; et les schémas 'Pondération Local Logarithmique*Entropie Globale' et 'Pondération Local Logarithmique*GFIDF'.

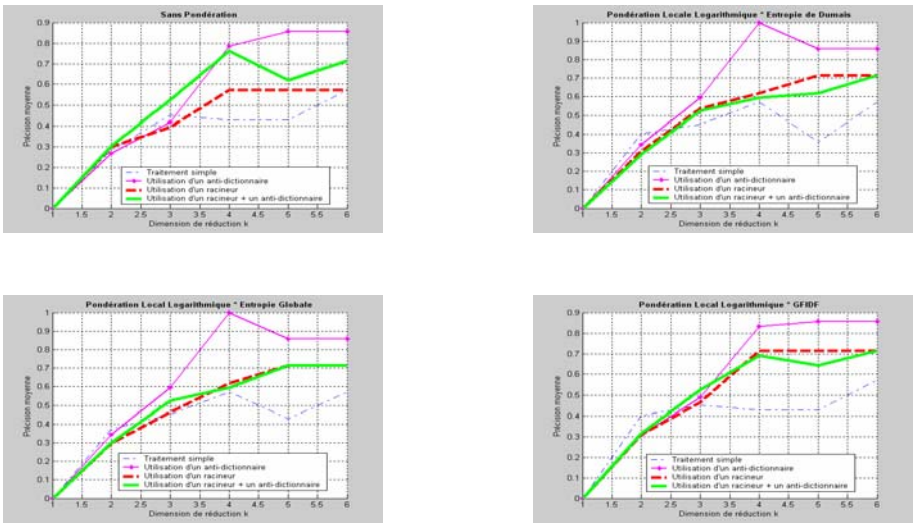
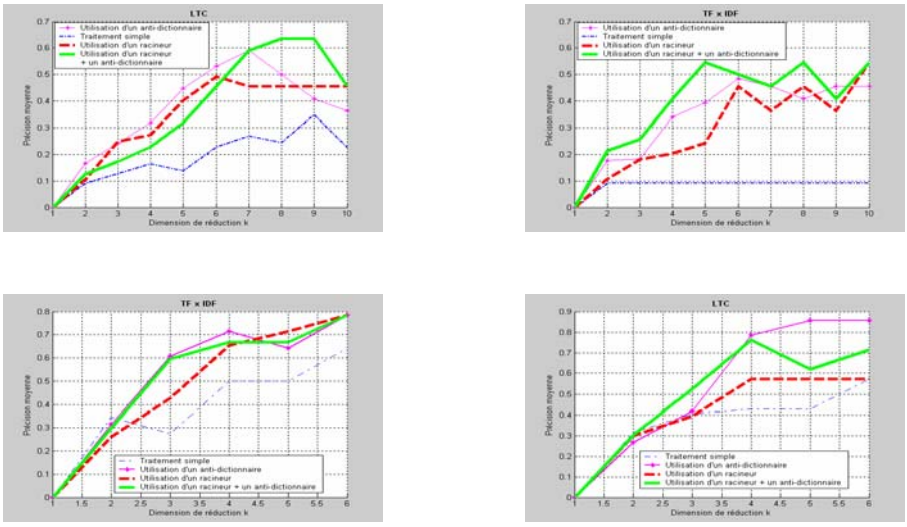


Figure 2 : L'influence des schémas de pondération sur la performance de la LSA

Pour évaluer l'importance de l'utilisation d'un anti-dictionnaire et d'un racineur, nous avons extrait la précision maximale de l'ensemble des précisions moyennes résultantes des tests réalisés. Nous avons présenté l'évolution de la précision moyenne de la méthode LSA, en appliquant les deux schémas de pondération « Tf x IDF » et « LTC », pour le corpus « Sous les branches » sur la figure 3-(a) et sur la figure 3-(b) pour le corpus « Musique de la nature ».



(a)

(b)

Figure 3 : Evolution de la précision en fonction du nombre de dimensions de l'espace pour un seuil de 0.9

Globalement les courbes calculées montrent que la performance de LSA s'améliore en utilisant soit un anti-dictionnaire soit un racineur, soit les deux. Néanmoins pour certaines requêtes les traitements linguistiques n'améliorent pas la performance de la LSA. Ceci est dû au racineur qui échoue à traiter certains pluriels et verbes. Par exemple pour les pluriels irréguliers des noms comme « طفل, enfant » « أطفال, enfants » qui ne sont pas une combinaison des formes singulières. Pour les verbes irréguliers comportant des consonnes particulières dites faibles (و, ا, ي) qui sont soit conservée, soit remplacée ou éliminée lors de leur déclinaison, exemple «قال , il a dit » «يقول , il dit » . Vu aussi la petite taille de nos corpus, par conséquent on peut dire que la LSA reste sensible dans le cas où on a peu de données à traiter.

6 Conclusion

Nous avons proposé une approche pour améliorer la méthode de l'analyse sémantique latente (LSA) en intégrant les paramètres linguistiques et de pondération. L'évaluation a montré l'intérêt d'appliquer conjointement le traitement linguistique et la pondération des unités lexicales pour pouvoir améliorer la performance de la LSA. Dans la suite de nos travaux, nous envisageons d'étendre cette étude à l'utilisation de la lemmatisation.

Références

Buckwalter T.(2002), Buckwalter Arabic Morphological Analyzer Version 1.0, http://www.ldc.upenn.edu/Catalog/CatologEntry.jsp?catologId=LDC2002L49.

Darwish K. (2002), Building a Shallow Arabic Morphological Analyzer in One Day, Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02) , pp. 47-54.

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

Grefenstette G. (1994), Explorations in automatic thesaurus discovery, New York, Kluwer Academic Publishers.

Salton G. (1989), Automatic text processing the transformation analysis and retrieval of information by computer, New York, Addison-Wesley.

Salton G. (1983), An Introduction to Modern Information Retrieval, New York, McGraw-Hill.