Identification de thèmes cas de l'Arabe standard

Mourad Abbas[1], Kamel Smaili[2], et Daoud Berkani[3]

[1] Speech Processing Laboratory, CRSTDLA,
1 rue Djamel Eddine Alfghani,
16000 Algiers, Algeria,
m_abbas04@yahoo.fr
[2] INRIA-LORIA, Parole Team, B.P 101,
54602 Villers les Nancy, France
smaili@loria.fr
[3] Signal and Communication Laboratory,
Ecole Nationale Polytechnique,
10 rue Hassan Badi, ELHarrach,
16200 Algiers, Algeria
dberkani@hotmail.com

Résumé Dans ce document, nous présentons les résultats de quelques expériences que nous avons menées concernant l'identification de thèmes en utilisant un corpus en langue Arabe. Nous avons choisi d'appliquer le classifieur TFIDF (Term Frequency/Inverse Document Frequency) qui est une référence dans ce domaine. Six thèmes ont été considérés pour réaliser cette tâche. En outre, l'évaluation a été réalisée en utilisant des données de différentes sources.

Mots clés Identification de thèmes, classifieur TFIDF, Arabe standard.

1 Introduction

L'identification de thèmes est un domaine de recherche qui a des applications dans plusieurs secteurs : reconnaissance de la parole pour adapter les modèles de langage, traduction automatique pour cibler la thématique de traduction, amélioration de la recherche dans le web, etc. Notre objectif est d'arriver à identifier le thème d'un document d'une manière efficace, ce qui aidera à améliorer le rendement des systèmes appliqués aux secteurs susmentionnés.
En effet, l'identification de thèmes est l'opération qui consiste à attribuer une étiquette ou plus à un flux de données textuelles. Les méthodes utilisées en catégorisation de textes sont celles adoptées pour l'identification de thèmes. Plusieurs travaux ont été menés au cours de la dernière décennie sur la catégorisation de textes : les classificateurs bayesiens [10,12,16], arbres de décision [8,10,12], réseaux de neurones [17,13], KNN [5,18]. D'autres approches adoptées utilisaient des systèmes experts comme le système "Construe" développé par Carnegie Group.

Plusieurs recherches ont été menées pour la classification des documents écrits en langue anglaise. D'autres langues ont eu le privilège d'être l'objet d'études, comme l'allemand, l'Espagnol, l'Italien [4], ainsi que les langues asiatiques comme le Japonais et le Chinois [14].

En ce qui concerne la langue Arabe, il y a peu de travaux dans ce domaine. En effet, dans [11], l'algorithme Naïve Bayes a conduit à des performances égales à 71.96% en terme de Rappel. Un autre système a été proposé dans [15], où ont été utilisées des méthodes statistiques de classification comme le Maximum d'entropie. Les résultats obtenus ont conduit à un taux de Rappel égal à 84.2% avec une valeur très basse de Précision, en l'occurrence 50%.
Concernant la langue française, et dans le but d'obtenir une identification thématique plus fiable, plusieurs efforts ont été déployés. En effet, Brun a proposé une méthode fondée sur la similarité mot-thème [3], qui a conduit à un taux de rappel égal à à 82.5%.
Dans nos expériences, nous avons utilisé le classifieur TFIDF pour l'identification de six thèmes. Le corpus utilisé est téléchargé à partir du site du journal arabophone Alwatan. Nous présenterons dans ce qui suit, quelques informations sur le corpus, ainsi que la définition du classifieur TFIDF. Ensuite nous exposerons les résultats de nos expériences.

2 Le corpus

La principale source de notre corpus est le journal arabophone Alwatan qui est un journal Omanais. Nous avons utilisé d'autres corpus de test issus des journaux "Akhbar Alkhaleej" de Bahrain, "Elkhabar" d'Algérie et "Alarabonline" [1]. Le corpus dont nous disposons est constitué de 9000 articles contenant environ 10 millions mots. Il est réparti sur six thèmes, en l'occurrence : Culture, Religion, Economie, Local, International et Sports. Nous présentons dans la table (1) le nombre de mots N constituant chacun des corpus thématiques extraits du journal "Alwatan", avant et après enlèvement des mots outils M.O.
La représentation de documents est une étape primordiale dans un système d'identification de thèmes. Un document ne peut être traité par un algorithme d'identification avant qu'il ne subisse une transformation particulière permettant sa représentation sous un format spécifique. Les mots sont adéquats pour la représentation des documents concernant les tâches de classification, cette méthode de représentation est appelée communément "Bag of words".
Pour la construction du vocabulaire, nous nous sommes basés sur une méthode connue : la fréquence de mots, bien qu'il existe d'autres méthodes comme la fréquence de documents ou la mesure d'information mutuelle.
Le vocabulaire ne doit pas contenir tous les mots du corpus d'apprentissage, car ceci pourra facilement fausser les résultats. En effet, d'après [6] et [9], les mots dont la fréquence ne dépasse pas un certain seuil n'apportent aucune information. La valeur de ce seuil reste empirique, toutefois des études optent pour la valeur 3, d'autres ont choisi la valeur 5 [2].

Table 1. Taille du corpus "journal Alwatan" avant et après enlèvement des mots outils

Thèmes	N. de mots avant	N. de mots après
Culture	1.359.210	1.013.703
Religion	3.122.565	2.133.577
International	855.945	630.700
Economie	1.460.462	1.111.246
Local	1.555.635	1.182.299
Sports	1.423.549	1.067.281
Total	9.813.366	7.139.486

3 Principe du classifieur TFIDF

L'idée de base du classifieur TFIDF est la représention des documents sous forme de vecteurs. Ainsi un document doc_i constitué de n mots : $doc_i = \{w_1, w_2, \ldots, w_n\}$, est transformé en un vecteur $D_i = \{d_{i1}, d_{i2}, \ldots, d_{i|V|}\}$. où $\mid V \mid$est la taille du vocabulaire. Chaque composante d_{ik} du vecteur représente une pondération du mot w_k. Elle est obtenue en effectuant le produit des deux grandeurs (valeurs statistiques) $TF(w, d)$ et $IDF(w)$. La fréquence de mots ou Term Frequency $TF(w, d)$ exprime le nombre de fois où le terme w apparaît dans le document d. Tandis que la fréquence de documents ou Document Frequency $DF(w)$ est le nombre de documents dans lesquels apparaît le terme w une fois au minimum. La valeur d_{ik} est obtenue en utilisant l'équation (1).

$$d_{ik} = TF(w_{ik}, d).IDF(w_{ik}) \tag{1}$$

L'inverse de la fréquence de documents (Inverse Documents Frequency) $IDF(w)$ est donné par la relation (2).

$$IDF(w) = log\left(\frac{\mid D \mid}{DF(w)}\right) \tag{2}$$

$\mid D \mid$ est le nombre total des documents. Le classifieur TFIDF représente aussi chacun des thèmes ou "classes" par un vecteur en se basant sur le corpus d'apprentissage concernant ce thème. Ainsi le thème T_j est représenté par le vecteur $D_j = \{d_{j1}, d_{j2}, \ldots, d_{j|V|}\}$. La similarité $sim(D_j, D_i)$ entre le thème T_j (représenté par le vecteur D_j) et le document doc_i (représenté par le vecteur D_i) est calculée en utilisant l'équation (3).

$$sim(D_j, D_i) = \frac{\sum_{k=1}^{|V|} d_{jk}.d_{ik}}{\sqrt{\sum_{k=1}^{|V|}(d_{jk})^2.\sum_{k=1}^{|V|}(d_{ik})^2}} \tag{3}$$

4 Expériences et résultats

Dans cette expérience, le corpus extrait du journal Alwatan est utilisé dans l'étape d'apprentissage. Comme nous l'avons mentionné, la taille de ce corpus avoisine 10 millions mots. Pour l'évaluation, nous avons pris trois corpus de trois journaux arabophones originaires de différents pays. Il s'agit du journal Algérien "Alkhabar", "Akhbar Alkhaleej" de Bahrain, et "Alarabonline". La taille de chacun de ces corpus de test correspond à 10 % de celle du corpus d'apprentissage. Nous avons utilisé un vocabulaire général dont la taille est 40000 mots.

Table 2. Evaluation par des corpus de test issus de trois journaux arabophones

Thème	Journal	Rappel (%)	Précision (%)	F_1(%)
Culture	Akhbar Alkhaleej	83.00	78.50	80.68
	AlKhabar	81.00	75.66	78.23
	Alarabonline	81.50	75.00	78.11
Religion	Akhbar Alkhaleej	92.75	90.00	91.35
	AlKhabar	90.00	92.50	91.23
	Alarabonline	91.00	91.00	91.00
Economie	Akhbar Alkhaleej	86.75	91.50	89.06
	AlKhabar	84.50	88.50	86.45
	Alarabonline	85.33	90.33	87.75
Local	Akhbar Alkhaleej	88.50	83.00	85.66
	AlKhabar	88.00	81.50	84.62
	Alarabonline	90.50	79.66	84.73
Politique	Akhbar Alkhaleej	92.66	87.50	90.00
	AlKhabar	91.75	86.33	88.95
	Alarabonline	93..55	88.50	90.95
Sports	Akhbar Alkhaleej	96.66	95.75	96.20
	AlKhabar	95.75	96.33	96.03
	Alarabonline	97.00	97.33	97.16
Moyenne	Akhbar Alkhaleej	90.05	87.70	88.85
	AlKhabar	88.50	86.80	87.64
	Alarabonline	89.81	86.97	88.36

Pour ce qui concerne l'évaluation des performances, nous avons utilisé trois mesures, en l'occurrence : le Rappel, la Précision et F_1. Le Rappel est obtenu en faisant un rapport du nombre des documents correctement étiquetés par le classificateur automatique et du nombre total des documents ayant cette même étiquette. Le calcul de la Précision se fait en divisant le nombre des documents correctement étiquetés par le nombre total des documents étiquetés par le classificateur. La combinaison de ces deux mesures donne la mesure F_1, qui est définie par la formule (4).

$$F_1 = \frac{2RP}{R+P} \tag{4}$$

Les résultats exposés dans la table (2) sont presque similaires. De ce fait on peut tirer la conclusion que l'Arabe standard est une langue uniforme, du moins pour les pays où nous avons puisé les données de test textuelles. Les performances du classifieur TFIDF ne sont pas affectées sachant que nous avons utilisé un corpus d'apprentissage dont la source est le journal "Alwatan", et d'autres corpus de test émanant d'autres sources. En effet, l'évaluation effectuée par l'utilisation de ces corpus de test varie de 87.64% à 88.85% en terme de mesure F_1. Ce qui est très proche du résultat obtenu en utilisant les données de test issues du journal "Alwatan" qui est de 89.94%.

5 Conclusion

Le travail que nous avons présenté dans cet article concerne l'évaluation du classifieur TFIDF dans la catégorisation des textes arabes. Nous étions motivés d'expérimenter ce classifieur qui est bien connu dans ce domaine, toutefois non utilisé pour la langue Arabe. Nous considérons que ses performances qui varient entre 87.64% à 88.85% en terme de F_1 sont satisfaisantes, particulièrement dans le cas où on les compare avec celles du système ARABCAT, basé sur le maximum d'entropie et qui a conduit à une valeur de F_1 égale à 80.41% [7].
En perspective, nous projetons de réaliser une meilleure représentation des documents, en prenant en considération l'étape de lemmatisation qui permettra d'améliorer les performances.

Références

1. M. Abbas. *Identification de thèmes pour la reconnaissance automatique de la parole.* PhD thesis, Ecole Nationale Polytechnique, Algiers, 2008.
2. C. Apté, F. Damerau, and S.M. Weiss. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3) :233–251, 1994.
3. A. Brun. *Détection de thème et adaptation des modèles de langage pour la reconnaissance automatique de la parole.* PhD thesis, Henri Poincaré University, Nancy1, 2003.
4. F. Ciravegna, L. Gilardoni, A. Lavelli, M. Ferraro, N. Mana, S. Mazza, J. Matiasek, W. Black, and F. Rinaldi. Flexible text classification for financial applications : the facile system. In *PAIS-2000, Prestigious Applications of Intelligent Systems sub-conference of ECAI2000*, 2000.
5. R. H. Creecy, B. M. Masand, S. J. Smith, and D. L. Waltz. Trading mips and memory for knowledge engineering : Calssifying census returns on the connection machine. *Comm. ACM*, 35 :48–63, 1992.
6. I. Dagan, Y. Karov, and D. Toth. Mistake-driven learning in text categorization. In *2nd conference on Empirical Methods in Natural Language Processing, EMNLP-97*, pages 55–63, Providence, US, 1997.
7. A. El-Halees. Arabic text classification using maximum entropy. *The Islamic University Journal (Series of Natural Studies and Engineering)*, 15(1) :157–167, 2007.
8. N. Fuhr, S. Hartman, G. Lustig, M. Schwantner, and K. Tzeras. A rule-based multistage indexing systems for large subject fields. In *RIAO'91*, pages 606–623, 1991.
9. T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213, 1996.

10. DD. Lewis and M. Ringuette. Comparison of two learning algorithms for text categorization. In *the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)*, 1994.
11. T. Rachidi. M. El-Kourdi, A. Bensaid. Automatic arabic document categorization based on the naïve bayes algorithm. In *20th International Conference on Computational Linguistics*, Geneva, August 28th 2004.
12. I. Moulinier. Is learning bias an issue on text categorization problem ? Technical report, LAFORIA-LIP6, Université Paris VI, 1997.
13. H. T. Ng, W.B. Goh, and K.L. Low. Feature selection perceptron learning, and a usability case study for text categorization. In *20th Ann Int ACM SIGIR Conference on Research and De velopment in Information Retrieval (SIGIR'97)*, pages 67–73, 1997.
14. F. Peng, X. Huang, D. Schuurmans, and S. Wang. Text classification in asian languages without word segmentation. In *the Sixth International Workshop on Information Retrieval with Asian Languages (IRAL 2003)*, Sapporo, Japan, July 7 2003. Association for Computational Linguistics.
15. H. Sawaf, J. Zaplo, and H. Ney. Statistical classification methods for arabic news articles. In *the Workshop on the ACL'2001*, Toulouse, France, July 2001.
16. K. Tzeras and S. Hartman. Automatic indexing based on bayesian inference networks. In *16th Ann. Int. ACM SIGIR Conference on Research and development in Information Retrieval (SIGIR'93)*, pages 22–34, 1993.
17. E. Wiener, J. O. Pedersen, and A.S. Weigend. A neural network approach to topic spotting. In *the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95)*, pages 317–332, Nevada, Las Vegas, 1995. University of Nevada, Las Vegas.
18. Y. Yang. Expert network : Effective and efficient learning from human decisions in text categorization and retrieval. In *17th Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 13–22, 1994.