Comparison of Topic Identification methods for Arabic Language

Mourad Abbas, Kamel Smaïli

Comparison of Topic Identification methods for Arabic language

M. Abbas [2] **and K. Smaili** [1]
INRIA-LORIA, Parole team
[1] B.P. 101 - 54602 Villers les Nancy, France
[2] Ecole polytechnique Alger- Algérie
Tel.: +33 (0)3 83 59 20 22 - Fax: +33 (0)3 83 59 19 27
e-mail: smaili@loria.fr

Abstract

In this paper we present two well-known methods for topic identification. The first one is a TFIDF classifier approach, and the second one is a based machine learning approach which is called Support Vector Machines (SVM). In our knowledge, we do not know several works on Arabic topic identification. So that we decide to investigate in this article. The corpus we used is extracted from the daily Arabic newspaper *Akhbar Al Khaleej'*, it includes 5120 news articles corresponding to 2.855.069 words covering four topics : sport, local news, international news and economy.
According to our experiments, the results are encouraging both for SVM and TFIDF classifier, however we have noticed the superiority of the SVM classifier and its high capability to distinguish topics.

1 Introduction

Topic identification has several applications : documents categorization, selecting documents for WEB engines, speech recognition systems, etc.
State-of-the-art continuous speech recognition systems suffer from various problems. In unrestricted speech recognition process the vocabulary has to be as large as possible. Increasing the vocabulary increases the search space and results in performance degradation. A language model is one of the knowledge source which is used by an automatic speech recognition system in order to find the best hypotheses respecting linguistic criteria. One way to improve the results of a speech recognition system is to adapt the language model in accordance to the concerned utterance context. The problem of topic adaptation has already been largely addressed. In (Martin et al., 1997), (M. Mahajan and Huang, 1999),(Yang, 1999), (Bigi et al., 2000), (Bigi et al., 2001), (Brun et al., 2002), topic information is exploited in different ways, resulting each time in a significant reduction of the perplexity of the baseline language model and in sometimes in an improvement of the word error. Hence, these studies highlight the importance of topic adaptation.

Topic Identification is a supervised learning task consisting in identifying the topic of a text among a set of predefined topics. There is no formal definition of this concept. In what follows, a topic is viewed as a subset of the language associated to particular events. A document will be considered to be on a particular topic whenever its content is connected to the associated event.
In this article we present the performance of two classifiers: TFIDF and SVM which are evaluated on Arabic corpus extracted from *Akhbar Al Khaleej*. To our knowledge, topic identification for Arabic is very little covered, that is why our purpose in this article is to highliht this subject. We begin by presenting the specificity of Arabic language, then we give details about the two methods used in this paper and we present the results.

1.1 An overview of Arabic morphology

Arabic is a semitic language which is written from right to left, unlike Latin languages. An Arabia word may be composed of a stem, prefix and suffix. The stem is composed of a root and pattern morphemes. The prefix can be composed of several sub-prefixes including inflectional markers for tense, gender, and/or numbers. The suffix include zero or several sub-suffixes as some prepositions, conjunctions, determiners, possessive pronouns and pronouns. Most Arabic morphems are defined by three consonants, to which various affixes can be attached to create a word. For example, from the tri-consonant "ktb"

كتب

, we can inflect several different words concerning the idea of writing as presented in Table 1

There are many, many other derivations from this stem. The following example gives an idea about the different morphological segments existing in the word, and shows their equivalent in English:

And by her relations

Arabic	English
كَتَبَ	wrote
كِتَاب	book
كُتُب	books
يَكتُبُ	he writes
سَيَكتُبُ	he will write
كَاتِب	author
...	...

Table 1: An example of an Arabic word

Arabic	English
وَ	and
بِ	by
علَاقَاتِ	relations
هَا	her

Table 2: An example of an Arabic word

→ وَبِعَلَاقَاتِهَا

The example cited above shows that an Arabic word may correspond to several English words. Because of the variability of prefixes and suffixes, the morphological analysis is an important step in Arabic text processing. This makes segmentation of Arabic textual data different and more difficult than Latin languages. In the following, we developed a tool which split a word into prefixes, stem and suffixes. Some prefixes and stems have been kept, the suffixes have been removed for topic identification. This is due to the fact that we need only the sub-words which are meaningful for this task. have to be represented.

1.2 Documents representation

To process the documents, we have to build internal representations by transforming a document d to compact vector form. This operation is generally done after the tokenization of the corpus as explained in the previous section. The dimension of the vector corresponds to the number of distinct words or tokens in the training set. Each entry in the vector represents the weight of each term. For our purpose, after removing the non content words, we calculated both the frequency of each word, which is called Term Frequency, and the documents frequency of a word, that means the number of documents in which the word w occurs at least once. A general vocabulary is based on the word frequencies extracted from the Arabic newspaper corpus *Akhbar Al Khaleej* which contains 5120 news articles corresponding to more than 2.8 million of words. The first vocabulary contains 103706 distinct words, and finally the vocabulary used included all the words which appear more than 2. This leads to a vocabulary of 42877.

2 Topic Detection

Given a set of topics $T_1, T_2, \ldots, T_k$, the topic detection task consists in finding the topic(s) treated in a piece of text W (paragraph, article, ...).

Topic identification is based on topic training corpora, which represent the specificities of each topic. Given a text W, we want to identify the topic treated in this text. To do that, its specificities are compared with the ones of each topic.

3 The TFIDF classifier

The idea of this algorithm is to represent each document d as a vector $D = (d_1, d_2, \ldots, d_v)$ in a vector space. The vector elements are calculated as the combination of the term frequency $TF(w, d)$, which is the number of times the word w occurs in the document d, and the inverse document frequency $IDF(w)$ (Salton, 1991; Seymore and Rosenfeld, 1997).

$DF(w)$ is the number of documents in which the word w occurs at least once.
The value d_i is called the weight of word w_i in document d, and is given by the relation:
$d_i = TF(w, d) * IDF(w)$ with $IDF(w) = log(\frac{N}{DF(w)}$ N is the total number of documents.
To calculate the similarity between a document D_i and D_j we used the equation 1:

$$Sim(D_j, D_i) = \frac{\sum_{k=1}^{|V|} d_{jk} d_{ik}}{\sqrt{\sum_{k=1}^{|V|} (d_{jk})^2 \sum_{k=1}^{|V|} (d_{ik})^2}} \quad (1)$$

Topic	Training	Distinct words
International	755000	15078
Economy	578000	21108
Local	893000	17213
Sports	628000	13632

Table 3: Training corpora by topic

An article is assigned to the topic which gives the highest similarity.

4 The SVM method

The well known SVMs (Support Vector Machines) introduced by V. Vapnik (Vapnik, 1995) achieve biclass categorization. They have the advantage of being robust where it can handle a large number of features with good generalization performance. Another advantage of the SVM classifier is its capability to work with real and large-scale data. Basic SVM algorithm is able to recognize two different types of objects (vectors). The algorithm offers to do classification by building hyperplane in the R^N vector space and checking at which side found each vector. This operation may be described by a linear decision function: $f(x) = \sum_{i=1}^{n} w_i * x_i + b$ with w vector orthogonal to hyperplane and b distance from hyperplane to the origin. To decide to which class x belongs, one has to study the sign of the decision function $y = sgn(f(x))$. Since text categorization has been shown to be a linear problem (Joachims, 1998), and since exploratory research with other kernels did not yield performance improvements, we use only linear kernels. The SVM classification was performed with SVM^{light} (Joachims, 1998)

5 Experiments

In this section the TFIDF classifier and the SVM method are evaluated on real data extracted from an Arabic daily newspaper. We used 5120 articles, 90% of this corpus have been reserved for training and the rest for test. Table 5 summarizes the number of words for each topic and the number of words kept for a topic representation [1]

All the experiments presented in the next sections have been evaluated by the well-known measures : recall, precision and F1 given below.

[1] all the words occurred more than 3 times

	Recall	Precision	F1
International news	97.65	99.2	98.42
Local news	85.94	79.71	82.71
Economy	85.15	85.82	85.84
Sport	94.53	100	97.19

Table 4: The performance of the TFIDF classifier

$$\text{Recall} = \frac{\text{Nb texts correctly labelled}}{\text{Nb texts of topic}}$$

$$\text{Precision} = \frac{\text{Nb texts correctly labelled}}{\text{Nb texts labelled}}$$

$$F_1 = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall}+\text{Precision}}$$

5.1 The TFIDF classifier

We withdrawn the non content words. In addition, we removed the words occurring less than 3 times. Consequently, each document is represented by a vector of 42877 words. The table 4 presents the recall, precision and F1 measure values for the four topics :

The best result is obtained for the international news and followed by sport news.

5.2 The SVM method

The Joachims tool SVM^{light} is used in our experiments for biclass discrimination. We used 1152 articles from each topic for training and 128 articles for test. Training consists of presenting positive and negative data. The negative data in our experiments consists of any other topic different from the one we want to learn. In all the experiments, we kept the same number of articles for positive and negative data. The table 5.2 shows respectively the values of recall, precision and F1 measure. This table shows that the SVM gives good results for Arabic topic identification. In fact, International news topic is well discriminated. It is never confused with Economy and Sport and reciprocally. In less than 1% of cases it is confused with local news topic. It is clear from this table that local news topic is the one which is slightly confused with all the other topics even with sport which could be considered as a very special. This topic has to be splited to more precise sub-topics. Table 6 shows the decrimination between a specific topic and a mixture of the the three other topics. This leads to the same conclusion, the Arabic topics are well discriminated.

To give an idea about the performances of both

Topic	International			Local			Economy			Sport		
	Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec	F_1
International	-	-	-	99.22	100	99.61	100	99.22	99.61	100	100	100
Local	99.22	100	99.61	-	-	-	89.06	92.68	90.83	97.66	99.21	98.43
Economy	100	99.22	99.61	89.06	92.68	90.83	-	-	-	97.66	100	98.81
Sport	100	100	100	97.66	99.21	98.43	97.66	100	98.81	-	-	-

Table 5: Recall, precision and F_1 for SVM biclass discrimination

	Recall	Precision	F1
International news	99.21	100	99.60
Local news	89.68	93.39	91.49
Economy	96.03	91.67	93.79
Sport	96.83	100	98.39

Table 6: SVM discrimination between a topic and topic mixtures

	Recall	Precision	F1 measure
TFIDF	90.82	91.18	90.95
SVM	97.26	98.52	97.88

Table 7: The mean values of recall, precision and F1

methods (SVM, TFIDF), we summarized the values of recall, precision and F1 measure, from the previous tables, in the table 7. We can conclude that SVM overcomes the results of TFIDF classifier for Arabic topic identification even if we showed in other works (Brun et al., 2002) that SVM is not the best method for classification. Neverthless, for Arabic language and with 4 topics the SVM performance are very interesting and important.

6 Conclusion

In this work we investigated topic identification for Arabic language, two well-known methods have been tested : TFIDF and SVM. The SVM methods achieves very high results 97.88 in terms of F_1. This method shows its capability to discriminate topics. Some of the studied topics are distinguished very easily. The SVM classifier outperforms the results obtained by TFIDF by more than 7.5% in terms of F_1 measure. As presented in (Yang, 1999), it would be interesting to study the methods performance according to the size of training data. This study is under work, we have now to increase the number of topics for Arabic and to compare the results obtained with those we achieved for French with other methods (Bigi et al., 2001). The idea is to try to understand if these methods are sensitive to the language.

References

B. Bigi, R. De Mori, M. El-Bèze, and T. Spriet. 2000. A fuzzy decision strategy for topic identification and dynamic selection of language models. *Special Issue on Fuzzy Logic in Signal Processing, Signal Processing Journal*, 80(6).

B. Bigi, A. Brun, J.P. Haton, K. Smaïli, and I. Zitouni. 2001. Dynamic topic identification: Towards combination of methods. In *Recent Advances in Natural Language Processing (RANLP)*, pages 255–257, Tzigov Chark, Bulgarie.

A. Brun, K. Smaïli, and J.P. Haton. 2002. Contribution to topic identification by using word similarity. In *International Conference on Spoken Language Processing (ICSLP2002)*.

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, pages 137–142.

D. Beeferman M. Mahajan and X. Huang. 1999. Improved topic-dependent language modeling using information retrieval techniques. In *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing.*

S. Martin, J. Liermann, and H. Ney. 1997. Adaptive topic-dependent language modelling using word-based varigrams. In *Proceedings 3rd European Conference on Speech Communication and Technolog.*

G. Salton. 1991. Developments in Automatic Text Retrieval. *Science*, 253:974–979.

K. Seymore and R. Rosenfeld. 1997. Using Story Topics for Language Model Adaptation. In *Proceeding of the European Conference on Speech Communication and Technology.*

V. Vapnik. 1995. *The Nature of Statistical Learning Theory.* Spinger, New York.

Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1-2):69–90.