Evaluation of Topic Identification Methods on Arabic Corpora

M. Abbas[1], K. Smaili[2], D. Berkani[3]
[1]Speech Processing Lab
crstdla
1 rue Djamel Eddine Alafghani 16011
Algeria
m_abbas04@yahoo.fr

[2]Parole Team
Inria-Loria
France
kamel.smaili@loria.fr

[3]Signal and Communications Lab
National Polytechnic School
Algeria.
dberkani@hotmail.com

Journal of Digital Information Management

***ABSTRACT**: Topic Identification is one of the important keys for the success of many applications. Indeed, there are few works in this field concerning Arabic language because of lack of standard corpora. In this study, we will provide directly comparable results of six text categorization methods on a new Arabic corpus Alwatan-2004. Hence, Topic Unigram Language Model (TULM), Term Frequency/Inverse Document Frequency (TFIDF), Neural Network, SVM, M-SVM and TR have been experimented, and showed that TR-Classifier is the most efficient among the set of classifiers, nevertheless, only binary SVM outperformed it thanks to its characteristics. Moreover, we should note that the size of Alwatan-2004 corpus used to achieve our experiments is considered the most important compared to any other Arabic corpus which had been used for topic identification experiments until now. In addition, we aim through using small sizes of vocabularies to reduce the time of computation. This is important for adaptive language modeling, particularly Topic Adaptation, which is required in real time applications such as speech recognition and machine translation systems. Our experiments indicate that the results are better than other works dealing with Arabic text categorization.*

Categories and Subject Descriptors
H.3.1 [Content Analysis and Indexing]: H.3.5 [Online Information Services]: Web-based services; **I.2.7 [Natural Language Processing]:** Text analysis

General Terms: Natural Language Processing, Text Analysis, Arabic Text Processing, Machine Translation, Text Categorization

Keywords: Topic Identification, Arabic Language, TULM, SVM, TR, Neural Network, Alwatan-2004 corpus

Received: 11 March 2011, Revised 15 June 2011, Accepted 21 June 2011

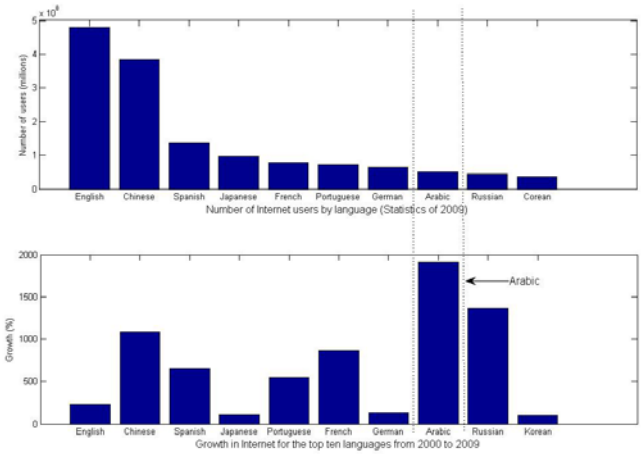
1. Introduction

The online language population is estimated as 1,733,993,741 persons in September 2009 by Internet World Stats. Growth in Internet between 2000 and 2009 reached 380.3 %, and gives us an idea about the amount data in the Web (http://www.internetworldstats.com/). Arabic is one of the top ten languages used in the web (2.9 %) and is the first language growing quickly (1,907.9 % from 2000 to 2009), -Figure 1-. For Arabic language, this has contributed to make texts more available in the Web, which is considered as an opportunity for researchers to prepare new corpora for language processing tasks.

For English, many corpora are considered as main referring Text Categorization collection, as Reuters versions[1] [1,2]. However, because of the lack of such referring corpora in Arabic, we built our own corpus, which size approaches 9000 documents. More details about this corpus are given in section 2.

Through this paper, we are dealing with Topic Identification, which has a strong impact on the enhancement of different applications including Machine Translation (MT), Automatic Speech Recognition (ASR) and Search Web Engines.

Indeed, the performance of ASR systems degrades when the field of application or the topic of the speech sequence is very specialized. It is necessary to use the appropriate language model based on this topic in order to get best performance [1,3,4,5,6,7,8].



Figuge 1. Statistics of Internet users by language

[1]The Reuters-21578 collection Apté split includes 12902 documents.

Hence, the purpose of this study is to make an evaluation, -by using an arabic corpus-, of a set of methods, in this case: TFIDF, Neural Networks, TULM, Support Vector Machines (SVM), Multi-Category SVM (M-SVM) and the TR-Classifier (TRiggers-based Classifier) which is a novel method presented in [9,10,11].

The five methods TFIDF, TULM, Neural networks, SVM and M-SVM are used as baselines. In fact, they have been evaluated for Topic Identification using large vocabularies [12] and they have well performed. However, we will show in this paper, the comparison of these well-known methods to the TR-Classifier by using small sizes of vocabularies.

The paper is structured as follows: Section 2 describes the corpus we will use in this paper. Section 3, 4, 5, 6 and 7 introduce a definition of each classifier. Section 8 analyzes the results found by the experiments we conducted. Finally, our conclusions are summarized in Section 9.

2. Corpus description

Until now, there are no Arabic corpora made for text categorization purposes. However some researches are actually trying to scientifically compile representative training datasets for Arabic text classification that cover different text genres which can be used in the future as a benchmark [13].

Hence, few works dealing with Topic Identification or Text Categorization for Arabic language have been carried out by using non representative and small corpora [14,15,16,17]. Consequently, this could lead to erroneous results.

The only work in which the size of the training corpus is relatively representative is the one presented in [18]. However, the number of the used categories 3 is not sufficient to lead to reliable scores.

Our experiments are based on a corpus which size is relatively important, and composed of nearly 9000 articles "documents" that correspond to 10 million words[2]. We downloaded these articles from an online Arabic newspaper. A data-preprocessing step is necessary to prepare these texts for the next stage of treatment. Hence, each article is preprocessed to remove punctuation marks and digits, eliminate Stop-List words and achieve other treatments as light stemming [18]. Moreover, data need to be transformed to a form that is suitable for all text categorization algorithms used in this work. Therefore, we computed many parameters, as Term Frequencies, Document Frequencies, Average Mutual Information, etc. In order to achieve our experiments, we chose six topics –categories-: culture, religion, economy, local news, international news and Sports. The corpus' size of each topic is shown in Table 1.

Topics	N. words before	N. words after
Culture	1.359.210	1.013.703
Religion	3.122.565	2.133.577
Int. news	855.945	630.700
Economy	1.460.462	1.111.246
Loc. news	1.555.635	1.182.299
Sports	1.423.549	1.067.281
Total	9.813.366	7.139.486

Table 1. The size of the corpus before and after removing insignificant words

[2]This corpus is divided into six categories. It is released and available on the web: http://sourceforge.net/arabiccorpus

The vocabulary construction is based on the Term Frequency method which is quite simple and conducts to good performances [19,20]. Mutual information [3] and Document frequency are also good methods of terms' selection and lead to satisfactory results. First, we built a vocabulary corresponding to each topic (Topic Vocabulary) in order to be used by the TR-Classifier. Next, the remained methods use a general vocabulary that we built by making a concatenation of the six resulted Topic Vocabularies.

In the Text Categorization field, documents are usually represented by using the well-known Bag of Words method. The type of the representation depends both on the target task and the used method. In order to realize Topic Identification for Language Modeling as shown in [20], or classifying texts according to some specificities like historical period as presented in [21], the Bag of Words method seems to be very suitable. However, in the case of using the Perplexity method, we should consider a succession of words (n-grams) [20].

Each word of the document is weighted by an adequate value. The weights are those commonly used in text categorization [22,23,24], particularly for the TFIDF classifier.

In Tables 2 and 3, we present the most frequent Arabic words for each topic and their frequencies. Moreover, equivalent words in English are addressed.

According to tables 2 and 3, we can see that words are representative of the six topics. Bag of Words is undoubtedly a reliable method of representation for text categorization. However, it would be suitable to take into consideration some bigrams and trigrams. Indeed, if we take the example of Sports' topic that includes the two terms: كرة "ball" and قدم "foot" which frequencies are respectively 1148 and 898, we will clearly see that they are usually found as bigram كرة قدم "foot ball". Other examples can be mentioned as trigram: ولايات متحدة أمريكية "United States of America" which belongs to the Topic "International News" -Table 3-.

3. TFIDF Principle

The basic idea of the well-known TFIDF is to build a prototype vector per topic using a training set of documents. Vector components or weights are obtained by the product of the Term Frequency by the Inverse Document Frequency [22,23]. To identify a topic-unknown document, a similarity measure between this document and the prototype vector of a topic is computed.

Usually the similarity is calculated by using the cosine distance given by equation 1:

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

4. Neural Networks

Neural networks have been used in many artificial intelligence applications [25]. In text categorization area, some experiments have been carried out in order to evaluate this approach [26,27,28,1]. These systems consist to learn a non-linear mapping from vectors of a document space to a category. Training Neural Networks can be achieved either on all the categories or by using a separate network per category, which is considered a costly approach. In our case, we selected the last approach since the number of topics to be studied is small.

Culture	Religion	Economy
Freq / Arabic word / Eng. word	Freq / Arabic word / Eng. word	Freq / Arabic word / Eng. word
1867 / عربية / Arabic	4654 / صلى / Pray	2274 / دول / Countries
1308 / عربي / Arabic	4504 / وسلم / Peace	2117 / عمل / Work
1192 / عالم / World	3413 / رسول / Messenger	1747 / قطاع / Sector
1179 / عمل / Work	3045 / ناس / People	1646 / نفط / Oil
1103 / فيلم / Movie	2600 / قرآن / Koran	1555 / مليون / Million
1063 / كتاب / Book	2211 / إسلام / Islam	1487 / سوق / Market
1056 / فنية / Artistic	1844 / نبي / Prophet	1456 / شركة / Company
1053 / فن / Art	1675 / دين / Religion	1415 / ريال / Rial (money)
1042 / ثقافة / Culture	1622 / صلاة / Prayer	1317 / عماني / Omani
1030 / معرض / Exposition	1603 / حج / Pilgrimage	1243 / عربية / Arabic
948 / فنان / Artist	1593 / مسلم / Muslim	1103 / اقتصادية / Economic

Table 2. Words and their frequencies for the topics: Culture, Religion and Economy

Local News	International news	Sports
Freq / Arabic word / Eng. word	Freq / Arabic word / Eng. word	Freq / Arabic word / Eng. word
1875 / سلطنة / Sultanate	2193 / رئيس / President	2020 / مباراة / Match
1755 / عمل / Work	2109 / عراق / Iraq	1795 / فريق / Team
1524 / وزارة / Ministry	1710 / متحدة / United	1792 / مركز / Center
1438 / مهرجان / Festival	1494 / قوات / Forces	1680 / بطولة / Championship
1209 / فعاليات / Activities	1277 / حكومة / Government	1676 / منتخب / Team
1089 / صحية / Healthy	1246 / أمريكية / American	1431 / أولمبية / Olympic
1043 / تعليم / Teaching	1107 / ولايات / States	1195 / دورة / Tournament
1036 / مركز / Center	1084 / مجلس / Council	1148 / كرة / Ball
1000 / منطقة / Region	1070 / انتخابات / Elections	1046 / اتحاد / Union
998 / سعادة / Excellency	853 / وزير / Minister	948 / سباق / Race
948 / ولاية / State	842 / حرب / War	898 / قدم / Foot

Table 3. Words and their frequencies for the topics: Local News, International News and Sports

In addition, we used a multi-layer perceptron to realize the categorization task.

5. TR-Classifier Description

The main idea of the TR-Classifier is based on computing the Average Mutual Information (*AMI*) of each couple of words which belong to the vocabulary V_i [29,30,31,32]. "Triggers" of a word w_k are a list of words that have a high degree of correlation with it. The triggers that are considered important for a topic identification task are those which have the highest *AMI* values. Each topic is then endowed with a number of selected triggers M, calculated using training corpora T_i. We address in Table 4 an example showing the equivalence in English of the best five triggers of the topic Sports. More detailed description about this classifier can be found in [9,10].

The best five triggers
Team→National
Score→Competition
Olympic→Athens
Tournament→Athens
Final→Quarter

Table 4. Best five triggers of the topic Sports

On the contrary of TFIDF, SVM, M-SVM and TULM which use a general vocabulary, a vocabulary per topic is built for the TR-Classifier. Topic vocabularies are composed of the most frequent words which are ranked according to their frequency, from the maximum to the minimum. The size of each topic vocabulary used in our experiments is small, in this case, 300 words. The TR-Classifier performance is satisfactory, as it will be shown in the forthcoming sections.

6. Using SVM and M-SVM in Topic Identification

6.1 SVM: the Binary Case

Support Vector Machines have been widely used to handle the binary classification problem. Indeed, SVMs have a nice geometrical interpretation of the discrimination of one class from another, by a hyperplane with the maximum margin [33]. Consequently, the most of applications dealing with classification using SVM, conducted to very satisfactory performance.

The problem of separating a set of training vectors belonging to two different categories (topics) $S=\{(x_1,y_1),\ldots,(x_m,y_m)\}$, where $x_m \in R^n$ is a feature vector and $y_m \in \{-1,+1\}$ is a class label, can be realized by using a separating hyperplane of equation $\omega.x + b = 0$.

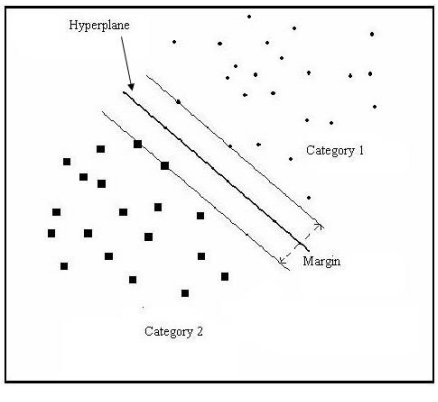


Figure 2. Binary separation by a linear SVM. Elements from the two classes "topics" are represented by ■ and • respectively

The boundary of the hyperplane which is determined by ω and b and maximizes the margin given by $2/\|\omega\|_2$ will generalize better than other possible hyperplanes. –Figure 2-.

Hence, we are dealing with an optimizing problem which solution consists to minimize $\|\omega\|_2^2$ under the constraints:

$$y_i((\omega.x_i)+b) \geq 1 \quad \forall i \in \{1,.....,m\} \tag{3}$$

If the training set is not separable, SVM try to minimize $\|\omega\|_2^2$ and at the same time separate the set with a minimum number of errors. This can be done by minimizing (4):

$$\frac{1}{2}\|\omega\|_2^2 + C\sum_{i=1}^{m}\xi_i \tag{4}$$

where $\xi_i \geq 0$ satisfies the constraints:

$$\begin{cases}(\omega.x_i)+b \geq +1-\xi_i & if\ y_i = +1\\ (\omega.x_i)+b \leq -1+\xi_i & if\ y_i = -1\end{cases} \tag{5}$$

Slack variables ξ_i are introduced in order to limit the number of vectors that pass into the margin or into the other side of the hyperplane. The parameter C is defined by the user; It balances contributions from the first and second terms of expression (4). An example is presented in section (6.2) in which M-SVM is used to identify some topics, or in other words to make a separation between vectors[3].

6.2 M-SVM: the Multi-Category Case

The multi-class classification is recently the object of several researches. It is used when the number of categories is superior than 2. Let us consider the following set:

$S_N = \{(x_1, C(x_1),...,(x_i, C(x_i),.....,(x_N, C(x_N)\}$, where x_i is the i^{th} training vector, and $C(x_i) \in \{1,....., Q\}$ is its category "topic". On the contrary to the binary classification, the multi-class approach consists to find more than one hyperplane, according to the number of the considered categories. In order to compute the parameters of these hyperplanes, a set of functions from R^d into R^Q represented by equation (6) is used:

$$h_k(x) = \omega_k^T x + b_k \tag{6}$$

with $[\omega_k] \in R^{Qd}$ and $[b_k] \in R^Q$. $h_k(x)$ represents the solution corresponding to the category C_k of a given point x. Thus, the objective function given by equation (7) will be minimized [33,34]:

$$J(\omega) = \frac{1}{2}\sum_{k=1}^{Q-1}\sum_{l=k+1}^{Q}\|\omega_k - \omega_l\|_2^2 \tag{7}$$

under the constraints :

$$\{\quad \forall i \in \{1,.....,N\}, \forall k \in \{1,.....,Q\} / \quad C_k \neq C(x_i)\ \}$$
$$(\omega_{C(x_i)} - \omega_k)^T x_i + b_{C(x_i)} - b_k \geq 1$$

ω_k and ω_l stand for the parameters of the hyperplanes equations. For more understanding, we present in Figure 3 an example that shows how a given point x is attributed to one class by using a separation with more than one hyperplane.

In fact, separation between classes A, B and C is realized by using many hyperplanes. In fact, the nearest hyperplanes of the point x, which belongs to the zone A, are determined by the difference between the parameters (ω_A, ω_B) and (ω_C, ω_C). Figure 3 represents hyperplanes which determine the zone A in which x belongs.

[3]Here, vectors represent a set of documents which belong to different topics.

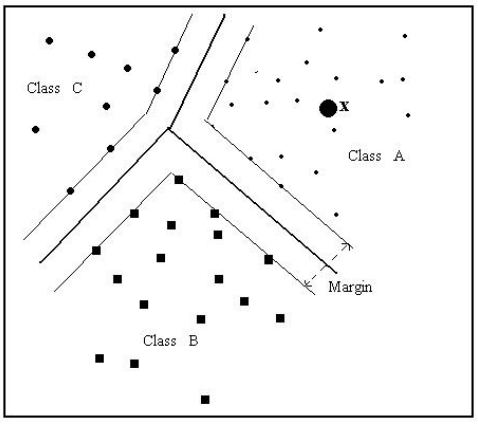


Figure 3. Multi-class separation

For the non-separable case the non-negative variables ξ_i are introduced, $1 \leq i \leq N$, $1 \leq k \leq Q$.

The objective function (8) is modified to become:

$$J(\omega,\xi) = \frac{1}{2}\sum_{k=1}^{Q-1}\sum_{l=k+1}^{Q}\|\omega_k - \omega_l\|_2^2 + C\sum_{i=1}^{N}\sum_{k=1}^{Q}\xi_{ik} \tag{9}$$

under the constraints:

$$\{\quad \forall i \in \{1,.....,N\}, \forall k \in \{1,.....,Q\} / \quad C_k \neq C(x_i)\ \}$$
$$(\omega_{C(x_i)} - \omega_k)^T x_i + b_{C(x_i)} - b_k \geq 1 - \xi_{ik}$$

More detailed theoretical definitions about M-SVM can be found in [34,35,36,37]. As in the binary case, the C parameter can be modified in order to tune the M-SVM algorithm. The following experiment shows the effect of varying the parameter C on results. In fact, the six aforementioned topics had to be identified. Then, we have selected 1400 documents for training and 10 % of the corpus had been reserved for test. The size of the vocabulary is 8000 words. Table 5 summarizes M-SVM performances variations according to the chosen value of C.

C	Recall
25	86.66
35	85
50	88.33
200	84.16

Table 5. Tuning M-SVM by varying the C parameter

7. An overview on TULM

Topic Unigram Language Model (TULM) represents a topic by a vector $D = \lfloor d_1,...,d_{|V|} \rfloor$, where d_k is the number of occurrences of the word w_k in the document w_1^N and $|V|$ is the size of the vocabulary V. The Topic Unigram Language Model is based on counting the occurrences' number of each word for each topic. The posterior probability for a topic j is given by equation (10):

$$P(T_j / W_1^N) = \frac{P(T_j)P(W_1^N / T)_j}{\sum_{k=1}^{j} P(T_k)P(W_1^N / T)_k} \tag{10}$$

$P(T_j)$is the a priori probability of the topic T_j, and $P(W_1^N / T_j$ is the likelihood of sequence W_1^N given a topic T_j [38].

8. Methods Performances

In this experiment we aim to compare the TR-classifier to the aforementioned set of methods by using small sizes of vocabularies. The TR-classifier uses a vocabulary per topic V_i, $(i = 1, 2,\ldots,6)$, which is not the case for the other methods. That is why we constructed a global vocabulary by concatenating the six topic vocabularies V_i which are composed of words of highest frequencies.

The size of the resulted global vocabulary after concatenation is 800 words. Experiments showed that the used methods are ranked as following: SVM, TR, TFIDF, TULM, Neural Networks and M-SVM. We should note that SVM is not appropriate in the case where the number of categories is superior to two. Certainly, we have seen that results achieved by SVM are satisfactory; however it realizes only binary categorization. Table 6 summarizes SVM performances per topic.

Compared to the previous experiments related to SVM and achieved in [12,39] in which the size of the global vocabulary is 40000 words, SVM performance did not decrease when we used a smaller global vocabulary (800 words). This is due firstly to the robustness of this method and its capacity to give best scores of classification, and secondly to ranking words of the vocabulary according to their frequencies (from max to min). In fact, the advantage of selecting the highest frequencies words is to have a good representation by avoiding, as much as possible, the overlap caused by the different topic vocabularies. Anyway, according to our experiments, results obtained by using SVM are better than those achieved in other works [40,41].

Topic	Recall (%)	Precision (%)	F1(%)
Culture	97.33	95.51	96.41
Religion	96.93	99.32	96.11
Economy	96.26	96.57	96.42
Local	96.13	96.55	96.34
International	98.26	96.88	97.56
Sports	99.20	99.59	99.40
average	97.35	97.40	97.37

Table 6. SVM Performance per topic

M-SVM is the extension of SVM to the multi-category case, in which all topics are considered at once. This method has been a subject of many researches, but dealing only with small data sets. M-SVM has a strong theoretical background; consequently, results should be better than SVM (in condition to use a larger vocabulary [39]). Average M-SVM performance, in terms of Recall, is about 84%. Table 7 summarizes M-SVM performances per topic.

The TFIDF classifier outperformed a little bit the M-SVM by nearly 2% in terms of Recall. While the TR-classifier conducted to a Recall rate of 89.67% which is viewed as an encouraging result, since the maximal size of each topic vocabulary didn't exceed 300 words. It is shown in [10] that increasing concomitantly the number of triggers and the size of vocabularies enhance TR-Classifier performance. Therefore, we have selected a number of triggers which equals to 250. Detailed results of TFIDF, TULM and TR-Classifier are exposed respectively in tables (8), (9), (10) and (11).

Topic	Recall (%)	Precision (%)	F1(%)
Culture	75	78	76.47
Religion	95	96	95.50
Economy	83.5	75	79.02
Local	74	64	68.64
International	86.75	83	84.83
Sports	90	89.5	89.75
average	84.04	80.91	82.44

Table 7. M-SVM Performance per topic

TULM performance averages 85.61 % (in terms of Recall), which is almost the same result for the TFIDF classifier (85.88 %). The Neural Network categorization system is ranked after the TULM and TFIDF. As shown in previous tables, results given by each method indicate that there are some topics for which performance is lower than the other ones, in this case:

Topic	Recall (%)	Precision (%)	F1(%)
Culture	71.33	88.43	78.96
Religion	93.33	86.95	90.03
Economy	83.33	80.64	81.96
Local	80	76.92	78.43
International	93.33	84.33	88.60
Sports	94	100	96.91
average	85.88	86.21	86.04

Table 8. TFIDF performance per topic

Topic	Recall (%)	Precision (%)	F1(%)
Culture	70.55	89.5	78.90
Religion	94	86.33	90.00
Economy	82.66	82.33	82.50
Local	78.33	80	79.15
International	94.50	85.66	89.86
Sports	93.66	98.33	95.94
average	85.61	87.02	86.31

Table 9. TULM performance per topic

Topic	Recall (%)	Precision (%)	F1(%)
Culture	75	86.66	80.40
Religion	92	87.33	89.60
Economy	81.33	85.33	83.30
Local	75.25	90.50	82.20
International	92.55	83.33	87.70
Sports	94.50	90.66	92.54
average	85.10	87.30	86.20

Table 10. Neural Networks performance

Topic	Recall (%)	Precision (%)	F1(%)
Culture	82.66	80.55	81.60
Religion	96.33	83.56	89.50
Economy	83.50	84.05	83.77
Local	86.25	82.53	84.35
International	93.33	90.66	91.97
Sports	96	97.33	96.66
average	89.67	86.44	88.02

Table 11. TR-Classifier performance per topic

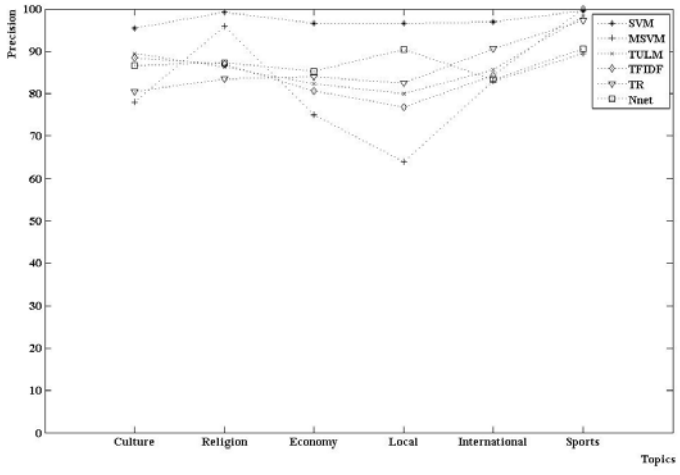


Figure 4. Recall values for the evaluated methods

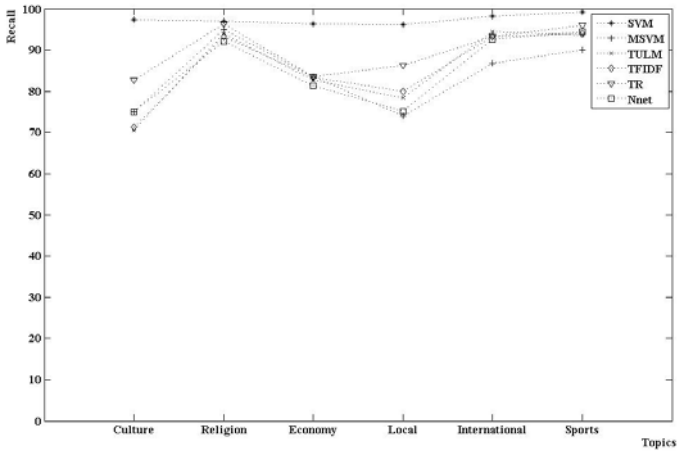


Figure 5. Precision values for the evaluated methods

Culture, Local and Economy, which are usually miscellaneous. Consequently, it is necessary to subdivide them to subtopics. For more clearness and visibility, figures 4 and 5 illustrate methods performance in terms of Recall and Precision. According to figure 4, all methods excepting SVM show that topic corpora has a great importance to determine classifiers' performance; this is deduced by the rapprochement of Recall values, given by all the used methods, concerning each topic. For example, the Religion topic corpus seems to be more representative than the corpus related to the Economy topic.

Moreover, we should note that results concerning Topic Identification for Arabic language are very promising, at least for the aforementioned methods that have been tested for French as well and led to lower performance [20].

Finally, as an illustrative example, we present in table (12) the performance of some of the aforementioned methods (SVM and TULM to be specific), that have been experimented in [14,40,41] for Arabic text categorization. When compared their performances (in terms of Recall) to our experiments, we clearly see that our results are better. The determining factor for enhancing the performance is the size of training corpora.

9. Conclusions

In this paper, some of the methods of Topic Identification have been evaluated using an Arabic corpus. It should be noted that this is for the first time a relatively important size of Arabic corpora had been used for a topic identification task.

Experiments	Our experiments	Experiment 1 [40]	Experiment 2 [41]	Experiment 3 [14]
Data set (number of documents)	9000	5121	1145	1500
SVM	97.35	77.80	84.90	-
TULM (NB)	85.60	74	-	62

Table 12. Performance of SVM and TULM

TFIDF, TULM, Neural Networks, SVM and M-SVM are used as baselines for the TR-Classifier. In fact, the main characteristic of this one is its capacity to capitalize on triggers in order to represent each topic as faithful as possible.

On the contrary of the other methods in which a general vocabulary is used, we have constructed one vocabulary per topic for the TR-Classifier. We should point out that the advantage of TR-classifier is achieving a good performance by using reduced sizes of topic vocabularies. In addition, the triggers number is an important factor for results improvement. Indeed, as shown in [9,10] for a given size of topic vocabularies, performance is enhanced when increasing the triggers number. That is why we selected 250 triggers for each topic.

For M-SVM, only a linear kernel is used in the framework of this work. Thus, in our next works, we aim to experiment more kernels as polynomial and gaussian ones.

Compared to the rest of the experimented methods, TR is the best classifier after SVM, nevertheless, we have not exploited vocabularies of bigger sizes to get maximum performance.

Moreover, even we used an Arabic corpus which has relatively an important size, we aim to build a more representative corpus, in order to be used in the fu2ture by researchers, for Topic Identification and Text Categorization tasks.

References

[1] Yang, Y (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval Journal* 1(1/2) 67-88.

[2] Sebastiani, F (2002). Machine learning in automated text categorization. *ACM Computing Surveys* 34 (1) 1–47.

[3] Seymore, K., Chen, S., Rosenfeld, R (1998). Nonlinear interpolation of topic models for language model adaptation. *In: Proc. of the International Conference on Spoken Language Processing,* pages 2503-2506, Sydney, Australia.

[4] Clifton, C., Cooley, R., Rennie, J (2004) TopCat: Data Mining for Topic Identification in a Text Corpus. *IEEE Transactions on Knowledge and Data Engineering* 16(8) 949-964.

[5] Mahajan, M., Beeferman, D., Huang, X (1999) Improved Topic-dependent Language Modeling using information Retrieval Techniques. *In: Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing*, p. 541-544, Phoenix, USA.

[6] Martin, S., Liermann, J., Ney, H (1997) Adaptive topicdependent language modelling using word based varigrams. *In: Proc. Of the 3rd European Conference on Speech Communication and Technolog*, p.1447-1450, Rhodes, Greece.

[7] Bigi, B., De Mori, R., El-Bèze, M, Spriet, T (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) 1085-1097.

[8] Bigi, B., Brun, A., Haton, J.P., Smaili, K., Zitouni, I (2001). Dynamic topic identification: Towards combination of methods, *In: Proc. of the Recent Advances in Natural Language Processing (RANLP'01)*, p. 255-257, Tzigov Chark, Bulgary.

[9] Abbas, M (2008). Topic Identification for Automatic Speech Recognition, Phd thesis, Electrical and Computer Engineering Department, National Polytechnic School, Algiers.

[10] Abbas, M., Smaili, K., Berkani, D (2009). A Trigger-based Classifier, *In: The 2nd international conference on Arabic Language Resources and Tools*, 22-23 April, Egypt.

[11] Abbas, M., Smaili, K., Berkani, D (2009). Comparing TRClassifier and kNN by using Reduced Sizes of Vocabularies. *In: Proc. of the 3rd International Conference on Arabic Language Processing*, p. 1-4. Rabat, Morocco.

[12] Abbas, M., Smaili, K., Berkani, D (2009). Topic Identification for Arabic Texts. *Journal of Computer Science and Engineering in Arabic* 2(3) 9-21.

[13] Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S., Al-Rajeh, A (2008). Automatic Arabic Text Classification. *9es Journees Internationales d'Analyse Statistique des Donnees Textuelles JADT*, p. 77-83, Lyon, France.

[14] El-Kourdi, M., Bensaid, A., Rachidi, T (2004). Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. *In: Proc. Of the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages,* pages 51-58, Geneva, Switzerland.

[15] Sawaf, H., Zaplo, J., Ney, H (2001). Statistical Classification Methods for Arabic News Articles, *In: Arabic Natural Language Processing in ACL 2001,* Toulouse, France.

[16] El-Halees, A (2006). Mining Arabic Association Rules for Text Classification, *In: Proc. of the 1st international conference on Mathematical Sciences,* p. 15-17, Al-Azhar University of Gaza, Palestine.

[17] Syiam, M. M., Fayed, Z. T., Habib, M. B (2006). An Intelligent System for Arabic Text Categorization. *IJICIS* 6 (1) 1-19.

[18] Duwairi, R., Al-Refai, M., Khasawneh, N (2007). Stemming Versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization, *In: 4th International Conference on Innovations in Information Technology*, p. 446-450, Dubai, UAE.

[19] Yang, Y., Pedersen, J. O (1997). A comparative study on feature selection in text categorization. *In: Proc. of the 14th International Conference on Machine Learning*, pages 412-420, San Francisco, US.

[20] Brun, A. 2003. Topic Detection and Language Model adaptation for automatic Speech Recognition. PhD thesis, Henri Poincare University, Nancy1, France.

[21] HaCohen-Kerner, Y., Mughaz, D., Beck, H., Yehuda, E (2008). Words as Classifiers of Documents according to their Historical Period and the Ethnic Origin of their Authors. *Cybernetics and Systems* 39 (3) 213-228.

[22] Joachims, T (1996). A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, School of Computer Science Carnegie Mellon University Pittsburgh.

[23] Salton, G. 1991. Developments in automatic text retrieval, Science 253. 974–979.

[24] Seymore, K., Rosenfeld, R (1997). Using story topics for language model adaptation. *In: Proc. of the European Conference on Speech Communication and Technology*, pages 2-5, Rhodes, Greece.

[25] Mitchell, T (1996). Machine Learning. McGraw Hill.

[26] Wiener, E., Pedersen, J.O., Weigend, A.S (1995). A neural network approach to topic spotting. *In: Proc. of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95)*, p. 317-332, 1995.

[27] Ng, H.T., Goh, W.B., Low, K.L (1997). Feature selection, perceptron learning, and a usability case study for text categorization. *In: Proc. Of the 20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, p. 67-73, 1997.

[28] Harrag, F., Al-Qawasmah, E. (2010). Improving Arabic Text Categorization Using Neural Network with SVD. Journal of Digital Information Management 8 (4) 233-239.

[29] Rosenfeld, R (1994). Adaptive Statistical Language Modeling: A Maximum Entropy Approach, PhD thesis, Computer Science Department, Carnegie Mellon University.

[30] Haton, J.P., Cerisara, C., Fohr, D., Laprie, Y., Smaili, K (2006). Speech Recognition from signal to its interpretation, Dunod, France. 392.

[31] GuoDong, Z., KimTeng, L (1999). Interpolation of n-gram and mutual information based trigger pair language models for Mandarin speech recognition. *Computer Speech and Language* 13. 125-141.

[32] Tillman, C., Ney, H (1996). Selection criteria for word trigger pairs in language modeling, *In:* Laurent Miclet and Colin de la Higuera, editors, Grammatical inference: Learning syntax from sentences. *Lecture Notes in Artificial Intelligence*, 1147 95-106.

[33] Vapnik, V. N (1998). Statistical Learning Theory. John Wiley and Sons, Inc., N.Y. 736.

[34] Lee, Y., Lin, Y.,Wahba, G. (2004). Multicategory Support Vector Machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance Data, *Journal of the American Statistical Association* 99 (465) 67-81.

[35] Guermeur, Y., Elisseeff, A., Paugam-Moisy, H (2000). A new multi-class SVM based on a uniform convergence result. *IJCNN'00*, 4 183-188.

[36] Bredensteiner, E. J., Bennet, K. P (1999). Multicategory Classification by Support Vector Machines *Computational Optimizations and Applications.* 12. 53-79.

[37] Guermeur, Y., Pollastri, G., Elisseeff, A., Zelus, D., Paugam-Moisy, H., Baldi, P (2004). Combining Protein Secondary Structure Prediction Models with Ensemble Methods of Optimal Complexity *Neurocomputing* 56. 305-327.

[38] Bigi, B., Brun, A., Haton, J.P., Smaili, K., Zitouni, I (2001). A Comparative Study of Topic Identification on Newspaper and E-mail, *In: Proc. of the IEEE International Conference on String Processing and Information Retrieval*, p. 238-241, Chile.

[39] Abbas, M., Smaili, K., Berkani, D (2009). Multi-Category Support Vector Machines for Identifying Arabic Topics. *Research in Computing Science: Advances in Computational Linguistics* 41. 217-226.

[40] Alsaleem, S (2011) Automated Arabic Text Categorization using SVM and NB, *International Arab Journal of e-Technology* 2(2) 124-128.

[41] MESLEH, A (2007) Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System, *Journal of Computer Science* 3(6) 430-435.

Authors Biographies

Mourad Abbas was born in Algiers in 1971. He obtained an engineer diploma in Electronics form USTHB (Algiers) in 1997, then the Magister in 2002. He obtained a Phd in Electronics from the National Polytechnic School in 2008. He is senior fellow and head of Speech Processing Laboratory (crstdla). He published his research in more than 20 papers. His research interests include speech processing, speech recognition, speech synthesis, text categorization, machine translation, etc. Dr. Abbas became a senior member of Iacsit (International Association of Computer Science and Information Technology) in 2009.

Kamel Smaïli was born in Algeria in 1963; he obtained an engineer diploma in computer science from (USTHB-Algiers). He then obtained a research master and a PhD. in 1991 from Nancy1 University (France). In 2001, he obtained an HDR. He is professor at university Nancy2. His research interest concerns statistical language modeling for speech recognition and speech-to-speech translation. Pr. Smaïl advised more than 10 Phd students. He took part to several program committees: Eurospeech, ICASSP, and reviewed papers for several journals: Computer speech and language, Speech communication. He published his research in more than 55 international conferences and journals.

Daoud Berkani received the engineer diploma and Master degree with Red Award from Polytechnic Institute of Kiev in 1977, then the Magister and Phd degrees from the National Polytechnic School (NPS). He is full Professor teaching signal processing and information theory in NPS. In 1992, he joined the Electrical Engineering Department of Sherbrooke University (Canada) where he taught signal processing and was a member of the Speech Coding team. His current research interests include signal and communications, information theory concepts and clustering. He is author of more than 150 papers.