TR-Classifier and kNN Evaluation for Topic Identification tasks

Mourad Abbas, Kamel Smaïli, Daoud Berkani

TR-Classifier and kNN Evaluation for Topic Identification tasks

Mourad Abbas, Kamel Smaili, and Daoud Berkani

Abstract—**This paper focuses on studying topic identification for Arabic language by using two methods. The first method is the well-known kNN (k Nearest Neighbors) which is used as baseline. The second one is the TR-Classifier, mainly based on computing triggers. The experiments show that TR-Classifier has the advantage to give best performances compared to kNN, by using much reduced sizes of Topic Vocabularies. TR-Classifier performance is enhanced by increasing jointly the number of triggers and the size of topic vocabularies. It should be noted that topic vocabularies are used by the TR-Classifier. Whereas, a general vocabulary is needed for kNN, and it is obtained by the concatenation of those used by the TR-Classifier. In addition to the standard measures Recall and Precision used for the evaluation step, we have drawn ROC curves for some topics to illustrate more clearly the difference in performance between the two classifiers. The corpus used in our experiments is downloaded from an online Arabic newspaper. Its size is about 10 millions words, distributed over six selected topics, in this case: culture, religion, economy, local news, international news and sports.**

Index Terms—**TR-classifier, k Nearest Neighbors, Arabic corpus, topic vocabulary.**

I. Introduction

TOPIC identification has been sufficiently studied for Indo-European languages. Generally, the methods used are those of text categorization: Bayesian classifiers [1, 2, 3], decision tree [2, 3, 4], neural networks [5, 6], kNN "k Nearest Neighbors" [7, 8], etc. Nevertheless, for Modern Standard Arabic, few works have been carried out [9, 10, 11, 12, 13].

The aim of this study is to evaluate two text categorization methods applied to Arabic documents. The first method is TR-classifer [14, 15, 16], a new method based on triggers, and the second one is the famous k Nearest Neighbors.

The concept of triggers has been largely used in statistical language modeling. Indeed, Triggers had been used to improve and generalize the Cache model [17]. The latter enhances the probability of a word w_i when it occurs in its left context. A trigger model goes further and enhances the probability of a list of words which are correlated to w_i [18]. This means that for all the words which occur in the right context of a word, their probability will be increased. [19].

We take advantage of triggers and propose a new clustering method. The motivation behind the TR-classifier design is that the information found in the longer-distance history is significant [20]. Indeed, for a topic identification task, the presence of the term "guitar" could trigger another list of terms: "music", "dance", etc. Hence, the main idea of TR-classifier is to represent each topic by triggers and triggered words which characterize each topic, and then facilitating the identification.

The second method that we have used, is the well-known kNN. As this method is considered in [21] as one of the top-performing classifiers, we selected it to compare its performance with that of TR-classifier.

The Arabic corpus used in our experiments is downloaded from the website of the Omani newspaper Alwatan; it is composed of more than 9000 articles.

In section II, we give some details about both TR-classifier and kNN method. We talk, in Section III about some of Arabic language specificities. We present in IV and V respectively text preprocessing and evaluation measures. Section VI deals with corpus description, documents representation and vocabulary construction. Finally, experiments and results are exposed in section VII.

II. Description of the Evaluated Methods

A. An overview on the TR-Classifier

Let define the concept of Triggers before using it in TR-classifier. The triggers of a word w_k are the set of words that have a high degree of correlation with it [16, 20, 22, 23].

Given the following text:

A large number of Taliban fighters may have crossed the border into Pakistan after pressure from U.S. and Afghan forces, Afghan officials said on Monday…

In this example, the word *Taliban* may trigger *Pakistan, Afghan* and *U.S.* These words are triggered by *Taliban.* Consequently, if we are interested by retrieving articles

Manuscript received October 30, 2009.

Mourad Abbas is with Phonetics and Speech Processing Laboratory (CRSTDLA), 1 rue Djamel Edine Alafghani, Bouzareah, 16011, Algiers (phone: 213-219-41088; fax: 213-217-93719; e-mail: m_abbas04@yahoo.fr

Kamel Smaili is with Parole team, INRIA-LORIA 54602 Villers les Nancy, France (phone: 333-835-92022; fax: 333-835-91927; e-mail: smaili@loria.fr).

Daoud Berkani is with Signal and Communications Laboratory, Electrical and Computer Engineering Department of the National Polytechnic School, 10, rue Hassen Badi, El Harrach, 16200 Algiers (phone: 213-215-21494; e-mail: dberkani@hotmail.com).

concerning *Taliban*, the occurrence of the corresponding triggered words may concern an article about Taliban.

The main idea of the TR-classifier is based on computing the average mutual information of each couple of words which belong to the vocabulary V_i. Triggers are couple of words that have a high value of (*AMI*) [18, 24]. They are considered important for a topic identification task. Each topic is then characterized by a number of selected triggers *M*, extracted from the topic training corpus T_i.

Identifying topics by using TR-method consists in:

- Training step:
  - Associating triggers to each word $v_k \in V_i$, where V_i is the topic vocabulary of T_i.
  - Selecting the best *M* triggers which characterize the topic T_i.
- Test step:
  - Identifying triggers of each word w_k of the test document.
  - Computing Q_i values by using the TR-distance given by (1):

$$Q_i = \frac{\sum_{i,k} AMI(w_k, w_k^i)}{\sum_{l=0}^{n-1}(n-l)} \quad (1)$$

Where *i* stands for the i^{th} topic. *n* is the size of the test document. The denominator is a normalization coefficient which represents the number of times $AMI(w_k, w_k^i)$ is computed.

w_k^i are triggers extracted from the test document *d*, and characterizing the topic T_i.

A decision for labeling the test document with topic T_i is obtained by choosing *i* which maximizes Q_i.

The Average Mutual Information for a couple of words *a* and *b* is given by equation (2).

$$AMI(a,b) = P_{ab} + P_{\bar{a}b} + P_{a\bar{b}} + P_{\bar{a}\bar{b}} \quad (2)$$

with:

$$P_{ab} = p(a,b)\log\frac{p(a,b)}{p(a)p(b)}$$

$$P_{\bar{a}b} = p(\bar{a},b)\log\frac{p(\bar{a},b)}{p(\bar{a})p(b)}$$

$$P_{\bar{a}\bar{b}} = p(\bar{a},\bar{b})\log\frac{p(\bar{a},\bar{b})}{p(\bar{a})p(\bar{b})}$$

$$P_{a\bar{b}} = p(a,\bar{b})\log\frac{p(a,\bar{b})}{p(a)p(\bar{b})}$$

Where $P(a,b)$ is the probability that the words *a* and *b* are present in the same document.

$P(a,\bar{b})$ is the probability that the word *a* occurs with words different from b.

$P(\bar{a},b)$ is the probability that the word *b* occurs with words different from *a*.

$P(\bar{a},\bar{b})$ is the probability that the words *a* and *b* are not present in the same document.

$p(a)$ is the probability of the word *a*.

The AMI measures well the correlation between words since its computation relies on different probabilities values which are presented by equation (2). The usefulness of a trigger pair is not determined only by a high correlation but by its frequency too. In fact, in order to understand the importance of the AMI, we present here the example provided in [20]. If we consider the trigger (Brest - Litovsk) consisting of two rare words but highly correlated, and compare it to a much more common one (Stock – Bond) which is less correlated, then it may be preferable to select the latter, if we are constrained to incorporate only one of the two trigger pairs. We can see also in Section VII some examples of computed triggers.

B. K Nearest Neighbors

kNN has been applied to text categorization before two decades [25, 8, 26, 21]. Indeed, Yang compared it to a set of text categorization methods using the benchmark Reuters corpus (the 21450 version, Apte set) [21]. It has been found that KNN is one of the top-performing methods after SVM [21]. Many other researches have found that the kNN method achieves a high performance by using different data sets [21, 27, 28].

The TR-Classifier is an original method which leads to acceptable scores even with small vocabularies. The idea of comparing it to KNN aims to highlight whether its performance is better or worse than that of KNN, particularly when using small vocabularies. Let's note that, we are carrying out experiments in order to compare the TR-Classifier to other methods, such as SVM, M-SVM, genetic algorithms, etc.

The strategy of the kNN algorithm is quite simple, so that, to identify a topic-unknown document *d*, kNN ranks the neighbors of *d* among the training document vectors, and uses the topics of the k Nearest Neighbors to predict the topic of the test document *d*. The topics of neighbors are weighted using the similarity of each neighbor to *d*. In order to measure this similarity, the cosine distance is used, although other measures are possible, as the Euclidean distance. The cosine similarity is defined by (3).

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

where D_j is the j^{th} training document and D_i is the i^{th} test document. $|V|$ is the size of the global vocabulary. d_j and d_i

represent the weights of the words belonging respectively to D_j and D_i. To assign the test document d to the correct topic, a cutoff threshold is needed [21].

III. Some Specificities of Arabic

Arabic is a Semitic language which is written from right to left. Phonologically, it is composed of 28 consonantal phonemes and six vowels (long and short vowels). Short vowels "or diacritics" are omitted intentionally in most of texts.

Arabic is very inflectional compared to Indo-European languages. Indeed, nouns can be nominative مرفوع, accusative منصوب or genitive مجرور. They can be singular, dual and plural. The dual is formed by adding "ان" to the stem in the nominative case and "ين" in the genitive and accusative one. The plural is formed by the addition of "ون" or "ين" to the stem according with the inflection case. In the feminine case, the suffix "ات" must be added. However, there are many other plural forms: حصان → أحصنة, أرض → أراضي, جندب → جنادب .

In addition, in unvocalised arabic texts, we can find many words that can be written in the same form. For example درست could be equivalent to the following:

Darastu → I have studied.
Darasta → You have studied
Darasti → You have studied
Darasat → She has studied
Darrastu → I have taught
Darrasta → You have taught
Darrasti → You have taught
Darrasat → She has taught

Moreover, *Arabic uses an agglutinative strategy to form surface tokens* [29]. If we take the example ودرستهم "wadarrasathum", it is clear that the word stemming is more complex than its equivalent in English "and she has taught them". In fact, *the sparseness of Arabic data decreases the efficiency of the training significantly* [29]. That is why tokenization is necessary to tackle this problem.

Another particularity of Arabic language is the absence of capital letters in the orthography which is considered as one of the most widespread problems faced to NER[1] systems [29,30].

IV. Text Preprocessing

Text preprocessing is the basic stage needed for text categorization tasks. Its main objective is, in one hand to remove all the unnecessary particles and mistyping words and in another hand to transform document contents to a suitable form which can be used easily by different algorithms. In the following subsections, we present the most important operations.

A. Feature extraction

Features or types of information are extracted in order to be used by classifiers to find categories. For that, many techniques are usually used such as removing stop-list words and stemming.

For Text Categorization purposes, the bag of words method is the most used. In fact, in the earliest work related to documents representation, Luhn [31,32] proposed the following ideas:

Luhn observed that authors emphasize an aspect of a subject by repeating certain words. They also usually use the same meaning of a word through a text. He also observed that only a limited number of words are used to express a particular idea. All these reasons conduct the community to reduce the space of representation of the original corpus that leads to good performance [33, 34].

For that, we proposed a stop-list of words that will be discarded from the document representation. This stop-list contains all the function words as: في، فيها على، عليهم، من، منه etc. Furthermore, the definite articles in all their forms such as: بالـ، فالـ، كالـ، الـ are deleted. This pre-processing is easier for Indo-European languages for many reasons, among them we can mention the fact that definite articles and function words are isolated from the other words, and the plural form in Arabic has more variants than in French or English. For example, the plural forms of the two arabic words تلميذ and مدرسة are respectively تلاميذ and مدارس. However, for English, the equivalent is obtained simply by ending the two words by "s": student → students, school → schools. Nevertheless, on the contrary to Arabic, there are few irregularities for English, for example: man → men**,** woman → women, child → children.

B. Feature selection

Even though the stop-word removed and word stemming done, the dimensionality still high. That is why, it is necessary to select a subset of relevant features which allow to represent documents adequately.

Feature selection can be carried out by using various methods as Term Frequency, Document Frequency [35, 36], Information Gain [35, 37, 38, 39], Mutual Information [39, 40] and Transition Point Technique [41]. In our experiments, we used Term Frequency.

V. Evaluation Measures

In order to evaluate the classifiers performance, at least three standard measures are used, in this case: Recall, Precision and F1 measure. Recall is defined to be the ratio of correct assignments by the classifier divided by the total number of correct assignments. Precision is the ratio of correct assignments by the classifier divided by the total number of the system's assignments [31]. The combination of the two measures with an equal weight gives the F1 measure, presented by the following expression:

$$F1 = \frac{2.\mathrm{Re}\,call.\mathrm{Pr}\,ecision}{\mathrm{Re}\,call + \mathrm{Pr}\,ecision}$$

[1] NER stands for Named Entity Recognition.

Other performance metrics which are used in many fields, could measure the classifiers' performance. False Acceptation Rate (*FAR*) which is defined to be the probability that the classifier incorrectly accepts the wrong document, and the False Rejection Rate (*FRR*) which is defined to be the probability that the classifier incorrectly rejects the desired document. Hence for a topic T_i , *FAR* and *FRR* are given by the following expressions:

$$FAR = \frac{Number\ of\ documents\ in correctly\ labelled\ T_i}{Number\ of\ documents \notin T_i}$$

$$FRR = \frac{Number\ of\ documents\ in correctly\ rejected\ T_i}{Number\ of\ documents \in T_i}$$

The two measures take their values from 0 to 1. In fact, if *FAR* equals 0, then all documents are correctly labeled. On the contrary, the value 1 indicates that all documents are incorrectly labeled.

However, we can decide that a classifier has a good performance if both *FAR* and *FRR* have jointly minimal values.

The plot of *FAR* against *FRR* is called the ROC curve (Receiver Operating Characteristic). The ROC curve of a perfect classifier would go through the point (0,0), it can be used to quickly visualize the quality of the classifier.

In the Experiments and Results section, we will evaluate the TR-Classifier and kNN by using, in addition to Recall and Precision, ROC curves for some topics in order to highlight the difference in performance between the two classifiers.

The idea behind using these metrics is to have a look at the evaluation from many different angles.

VI. Corpus Representation

We started by downloading Arabic texts from the archives of the Omani newspaper Alwatan of the year 2004. The size of the extracted corpus is about 10 millions terms which correspond to 9000 articles, distributed over six topics, in this case: Culture, religion, economy, local news, international news and sports. 90 % of these articles are reserved to training and the rest to the evaluation.

TABLE I.
Number of Terms Before and After Eliminating Insignificant Words

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

We should note that we have realized some elementary operations for topic identification, as eliminating insignificant words that do not bring any information, as function words, and also words whose frequencies are less than a definite threshold. We address in Table I the size of the entire corpus, before and after removing insignificant words.

The construction of the vocabulary has been made by using the term frequency method which gives good results though its simplicity [35]. Other terms selection methods as Mutual Information [40] and Document Frequency lead also to a satisfactory performance.

The kNN method uses a general vocabulary, whereas the TR-classifier uses a vocabulary per topic, i.e., six topic vocabularies are built, in our case.

We should note that these vocabularies are very small; indeed the size of each topic vocabulary is 300 terms. Nevertheless, they are composed of terms ranked from the maximum, to the minimum according to their frequencies. The reason behind the vocabularies size reduction is to make the topic identification process faster.

Documents need to be transformed to a compact vector form, and the dimension of the vector corresponds to the size of the vocabulary. Each word of the document is weighted by a definite value. The weights or vector components are those commonly used in text categorization, particularly for the TFIDF classifier [42].

Hence, after removing insignificant words, we calculated both the frequency of each word "Term Frequency", and the Document Frequency of a word *w*, that means the number of documents in which the word *w* occurs at least once. The weight of each term results then from the product of Term Frequency and Inverse Document Frequency [42, 43, 44].

VII. Experiments and Results

A. TR-Classifier Performance

As we mentioned in section III, TR-classifier uses a vocabulary per topic, and the words of each vocabulary are ranked according to their frequencies. In these experiments we used much reduced sizes of the six topic vocabularies, in this case: 100, 200 and 300 terms.

Hereafter, in Tables II and III, we present the best triggers which characterize two different topics, in this case sports and culture.

TABLE II.
THE TEN FIRST TRIGGERS CHARACTERIZING THE TOPIC CULTURE AND THEIR CORRESPONDINGS IN ENGLISH

culture	
Arabic	English
ملتقى → ثقافة	Culture → Meeting
قصيدة → شاعر	Poet → Poem
رواية → قصة	Novel → Story
مسلسل → شخصية	Personage → Serial
أفلام → جمهور	Public → Movies
تشكيلي → معرض	Exposition → Plastic
مسلسل → فنان	Artist → Serial
لوحة → تشكيلي	Plastic → Painting
فرقة → مسرح	Theater → Group
أفلام → سينما	Cinema → Movies

TABLE III.
THE TEN FIRST TRIGGERS CHARACTERIZING THE TOPIC SPORTS AND THEIR CORRESPONDINGS IN ENGLISH

sports	
Arabic	English
منتخب → وطني	Team → National
رصيد → منافسة	Score → Competition
أثينا → اولمبية	Olympic → Athens
أثينا → دورة	Tournament → Athens
ربع → نهائي	Final → quarter
حارس → لاعب	Player → Goalkeeper
موسم → مباريات	Match → Season
حارس → ملعب	Stadium → Goalkeeper
حارس → شوط	Half-time → Goalkeeper
متر → مسابقة	Race → Meter

For example, the first important couple of words for the topic Culture is that for which the Average Mutual Information is higher, in this case: Culture → Meeting (ثقافة → ملتقى), followed by a list of triggers.

Otherwise, we notice that the obtained triggers as Team → National (وطني → منتخب), Score → Competition (منافسة → رصيد), Olympic → Athens (اولمبية → أثينا) describe well the topic Sports.

The evaluation of the TR-classifier has been made by varying both topic vocabularies sizes and triggers number *N*.

The choice of N = 20, with a topic vocabulary size 100, lead to an average recall rate equal to 71.55 %. For some topics, we achieved good results; nevertheless the performance is degraded for other ones. The fact that these topics had unsatisfactory Recall values is due to their variety. Indeed, splitting them to subtopics turns out to be convenient, even necessary to have a best performance. Otherwise, the three remaining topics are relatively easily identified, particularly topic "Sports" which had a Recall rate equal to 93.33 %.

In order to enhance the performance, we have conducted other experiments, in which we increased *N*. Indeed, when taking N = 40, average recall rate attains 79.44 %, which represents an improvement of 8 %. Values of *N* = 60 and *N* = 80 conducted respectively to a Recall of 82.33 % and 83.11 %. In this case, the improvement of the performance is very slight. Hence, we decided to use larger vocabularies: 200 and 300. The performance in terms of Recall by using a size of vocabulary 100 is presented in Figure 1.

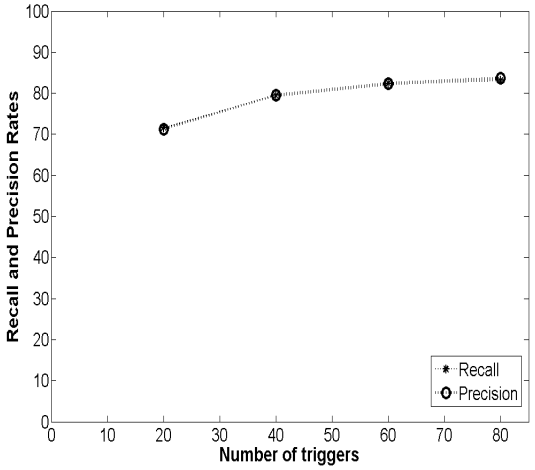

Fig. 1. TR-Classifier performances using a vocabulary size 100 in terms of Recall and Precision

Using a size of vocabulary 200 conducted to a slight improvement of Recall by nearly 1%. Nevertheless, this improvement necessitated to take *N* = 160 to reach 84 % in terms of Recall (See Figure 2).

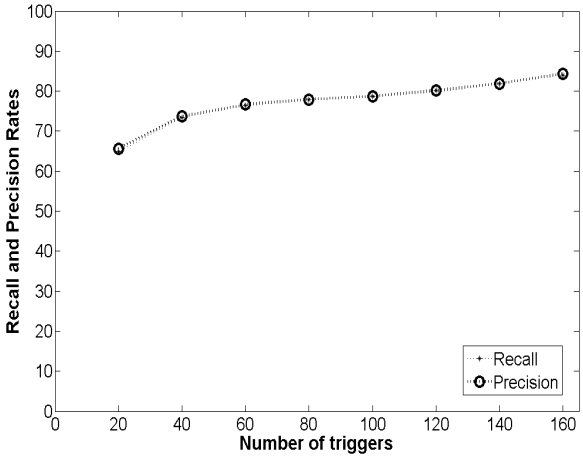

Fig. 2. TR-Classifier performances using a vocabulary size 200 in terms of Recall and Precision

Therefore, since the choice of vocabulary size 200 has not brought significant improvement compared with previous experiments, we continued with changing vocabulary size, and varying triggers number. We obtained better Recall values when using a vocabulary size of 300. Indeed, for N = 250, the best Recall rate achieved is 89.69 %. To see more clearly we show performance of the method in Figure 3.

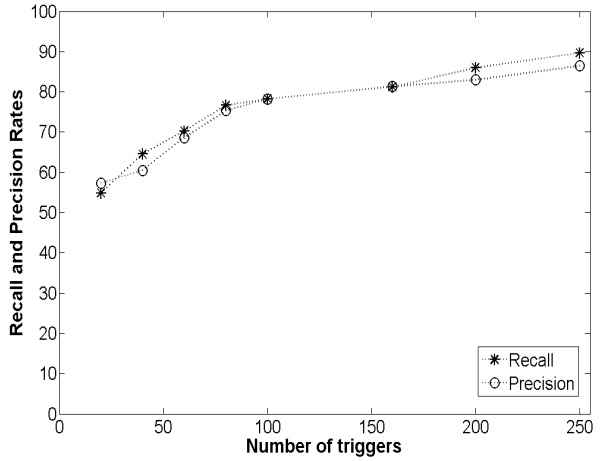

Fig. 3. TR-Classifier performances using a vocabulary size 300 in terms of Recall and Precision

B. kNN Evaluation

The main computation made by kNN is the sorting of training documents in order to find the k Nearest Neighbors for the test document. The value of k is usually optimized by several trials on the training and validation sets. In fact, if k is too large, small classes are overwhelmed by big ones. In practice, k is usually optimized by many trials on the training and validation data sets.

A general vocabulary is used by the kNN method. Thus we constructed it by concatenating the six topic vocabularies used in the previous experiment. The resulted size of this general vocabulary is 800 words.

Thus, we selected empirically different values of k. Recall and Precision rates related to each value of k are presented respectively in Tables IV, V, VI and VII.

TABLE IV.
PERFORMANCES OF THE KNN METHOD BY TAKING K=12

Topics	Recall (%)	Precision (%)
Culture	80	40.81
Religion	76	91.93
Economy	64.66	89.81
Local	54	70.43
International	70	81.39
Sports	82	94.61
Average	71.11	78.16

TABLE V.
PERFORMANCES OF THE KNN METHOD BY TAKING K=8

Topics	Recall (%)	Precision (%)
Culture	78.66	44.19
Religion	76	89.76
Economy	64.66	89.81
Local	60	69.23
International	72.66	81.34
Sports	83.33	93.28
Average	72.55	77.93

TABLE VI.
PERFORMANCES OF THE KNN METHOD BY TAKING K=4

Topics	Recall (%)	Precision (%)
Culture	76	49.78
Religion	75.33	94.95
Economy	68.66	81.74
Local	69.33	70.27
International	80	85.11
Sports	84.66	92.70
Average	75.66	70.09

Experiments showed that the performance on average is enhanced by using small values of k. Indeed, for k=12, 8, 4, 3, average Recall rates are respectively: 71.11%, 72.55%, 75.66% and 75.88%, - See Figure 4 -. Nevertheless, the performance's evolution is different from one topic to another. Indeed, the performance for all topics is enhanced when k decreases excepting the topic Culture whose performances diminish. This is due on a big part, to its variety.

Overall, we should note that the kNN' performance is lower than those of TR-Classifier by nearly 14 % which is considered as an important difference between the two methods.

TABLE VII.
PERFORMANCES OF THE KNN METHOD BY TAKING K=3

Topics	Recall (%)	Precision (%)
Culture	76	49.35
Religion	78	95.90
Economy	72	83.72
Local	68	69.86
International	76.66	84.55
Sports	84.66	93.38
Average	75.88	79.46

We can see clearly in Figure 5, the performance in terms of Recall of the two classifiers, for the six studied topics, in the case of using the size 300 for topic vocabularies and 800 for the general vocabulary.

In order to show the performance of the two classifiers, we selected some topics for which we drew ROC curves.

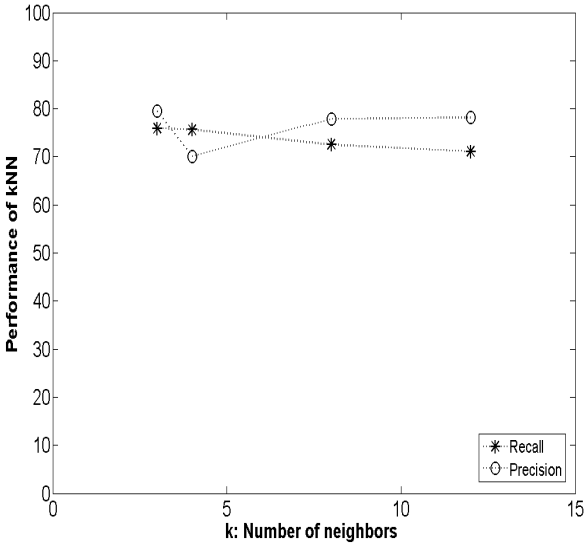

Fig. 4. kNN performances versus number of neighbors

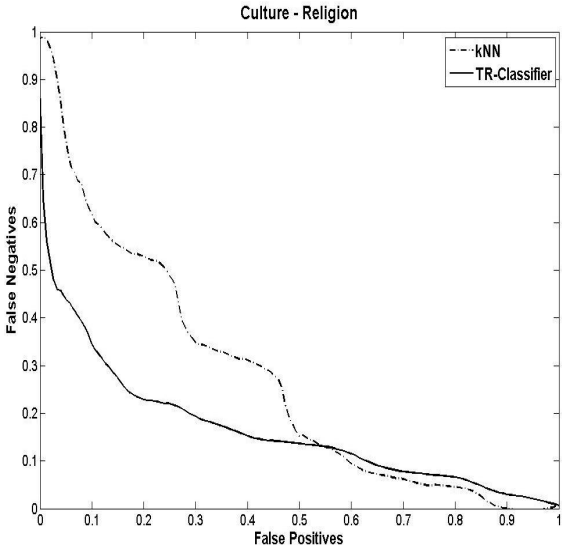

Fig. 6. ROC curves for the topics Culture/Religion.

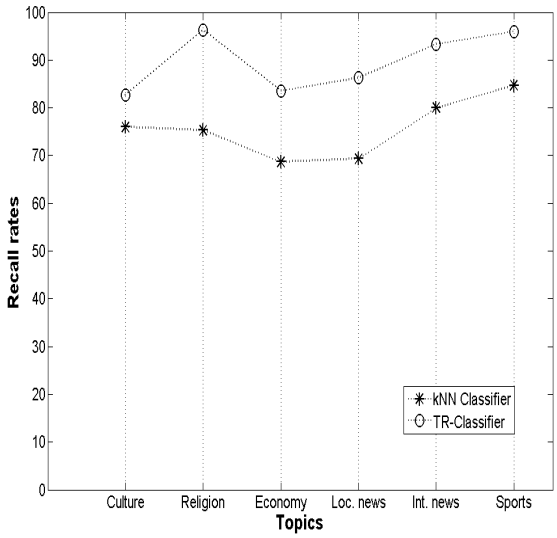

Fig. 5. TR-Classifier performances compared to kNN ones.

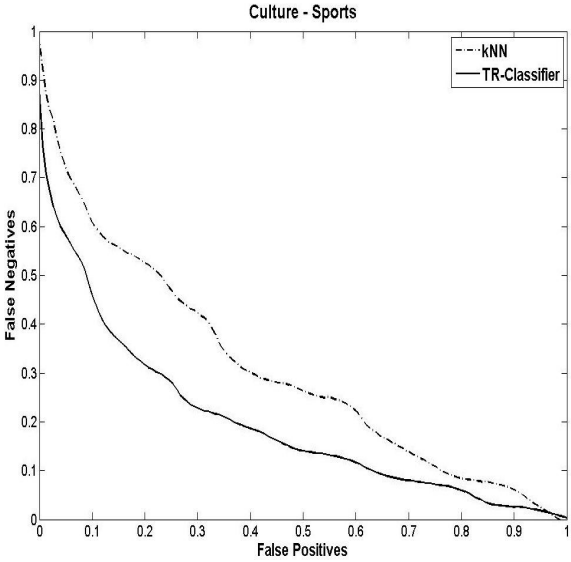

Fig. 7. ROC curves for the topics Culture/Sports.

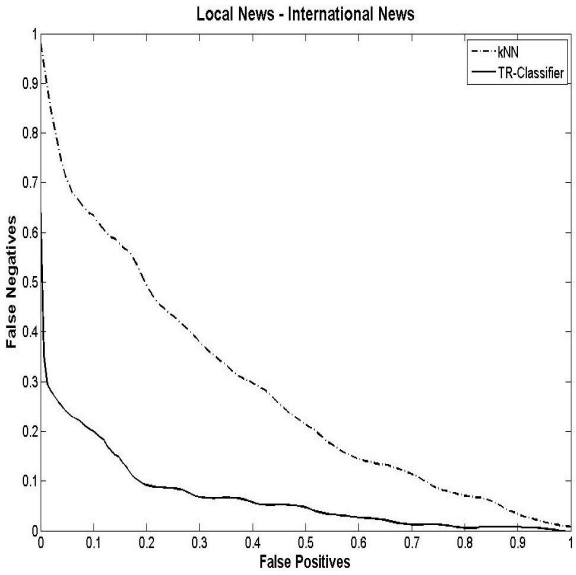

Fig. 8. ROC curves for the topics Local News/International News.

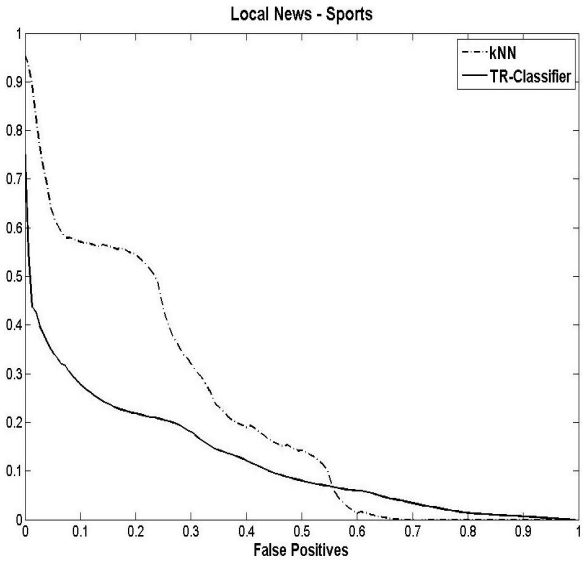

Fig. 9. ROC curves for the topics Local News/Sports.

Figures 6, 7, 8 and 9 show that both TR-Classifier and kNN go through the point (0,0) for the selected topics. However ROC curves related to TR-Classifier are closer to the point (0,0) than kNN.

VIII. Conclusion

In this paper, two methods of topic identification have been presented. Their performance has been tested on an Arabic corpus that we have constructed using many thousands of texts, downloaded from an online newspaper. One of these methods is the TR-Classifier: a new technique that we exposed in this paper and the second one is the well-known kNN.

The strong point of the TR-Classifier is its ability to realize better performance by using reduced sizes of topic vocabularies, compared to kNN. The reason behind this is the significance of the information present in the longer-distance history that TR-Classifier uses.

Undoubtedly, kNN is one of the best methods which give satisfactory performance; nevertheless in the case of small vocabularies, as shown in the aforementioned experiments, its performance didn't exceed 76 % in terms or Recall.

In perspectives, we aim to enhance the TR-Classifier performance by using greater sizes of vocabularies, though it outperforms kNN by 14 %, which is considered as a satisfactory result.

IX. References

[1] K. Tzeras, and S. Hartman, "Automatic Indexing Based on Bayesian Inference Networks," In: Proc. 16th Ann. Int. ACM SIGIR Conference on Research and development in Information Retrieval (SIGIR'93), 1993, pp. 22-34.

[2] DD. Lewis, and M. Ringuette, "Comparison of two Learning Algorithms for Text Categorization," In: Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), 1994.

[3] I. Moulinier, *Is Learning Bias an issue on Text Categorization Problem?,* Technical Report, LAFORIA-LIP6, University Paris VI, 1997.

[4] N. Fuhr, S. Hartman, G. Lustig, M. Schwantner, and K. Tzeras, "A rule-based Multistage Indexing Systems for Large Subject fields," in Proceedings of RIAO'91, 1991, pp.606-623.

[5] E. Wiener, J. O. Pedersen, and A.S. Weigend, "A neural network approach to topic spotting," in Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), University of Nevada, Las Vegas, 1995, pp. 317-332.

[6] 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 Development in Information Retrieval (SIGIR'97), 1997, pp. 67-73.

[7] 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, vol. 35, pp. 48-63, 1992.

[8] 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), 1994, pp. 13-22.

[9] M. El-Kourdi, A. Bensaid, and T. Rachidi, "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," 20th International Conference on Computational Linguistics, Geneva, 2004.

[10] H. Sawaf, J. Zaplo, and H. Ney, "Statistical Classification Methods for Arabic News Articles. Arabic Natural Language Processing," Workshop on the ACL'2001, Toulouse, 2001.

[11] M. Abbas, and K. Smaili, "Comparison of Topic Identification Methods for Arabic language," in proceedings of the International conference on Recent Advances in Natural Language Processing RANLP05, Bulgary, 2005, pp. 14-17.

[12] M. Abbas, and D. Berkani. "Topic Identification by Statistical Methods for Arabic language," Wseas Transactions on Computers", Athens, Issue 9, vol. 5. pp. 1908-1913., 2006.

[13] Rosso Y. Benajiba, M. Diab, P. Rosso, "Arabic Named Entity Recognition: An SVM-based Approach. In: IEEE Transactions on Audio, Speech and Language Processing, vol. 15, num. 5. Special Issue on Processing Morphologically Rich Languages, pp. 926-934, 2009.

[14] M. Abbas, K. Smaili, D. Berkani, "Comparing TR-Classifier and kNN by using Reduced Sizes of Vocabularies". The 3rd International Conference on Arabic Language Processing, CITALA 2009, Mohammadia School of Engineers, Rabat, Morroc, pp.1-4, 2009.

[15] M. Abbas, K. Smaili, D. Berkani. A Trigger-based Classifier. The 2nd International Conference on Arabic Language Resources and Tools (MEDAR 2009), 22-23 April 2009, Cairo, Egypt

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

[17] R. Kuhn and R. DeMori, "A cache-based natural language model for speech recognition," IEEE Trans. PAMI, vol. 12, no. 6, pp. 570–582, 1990.

[18] C. Tillman and H. Ney, *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, pp. 95-106, 1996.

[19] C. Tillmann and H. Ney, "Word trigger and the EM algorithm," in Proceedings of the Conference on Computational Natural Language Learning, Madrid, Spain, 1997, pp. 117–124.

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

[21] Y. Yang, and X. Liu, "A re-examination of text categorization methods," in Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'99,1999, pp 42-49.

[22] J.P. Haton, C. Cerisara, D. Fohr, Y. Laprie, and K. Smaili, *Speech Recognition from signal to its interpretation,* France: Dunod, 2006.

[23] C. Lavecchia, K. Smaïli, D. Langlois and J.P. Haton "Using inter-lingual triggers for machine translation", Tenth International Conference on Spoken Language Processing", Antwerp, Belgium, 2007

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

[25] B. Masand, G. Lino, and D.Waltz, "Classifying news stories using memory based reasoning," In 15th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92), 1992, pp. 59-64.

[26] M. Iwayama and T. Tokunaga, "Cluster-based text categorization: a comparison of category search strategies," in Proceedings of the 18th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95), 1995, pp, 273-281.

[27] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," In: Proceedings of the European Conference on Machine Learning, 1998.

[28] L. Baoli, C. Yuzhong, and Y. Shiwen, "A Comparative Study on Automatic Categorization Methods for Chinese Search Engine," in Proceedings of the Eighth Joint International Computer Conference. Hangzhou: Zhejiang University Press, 2002, pp. 117-120.

[29] Y. Benajiba, M. Diab and P. Rosso, "Arabic Named Entity Recognition: An SVM-based approach," In: Proc. Int. Arab Conf. on Information Technology, ACIT-2008, Hammamet, Tunisia, December, 16-18.

[30] Y. Benajiba, P. Rosso and J.M. Benedí, "ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy, " In: Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007, Springer-Verlag, LNCS (4394), pp. 143-153.

[31] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," IBM Journal, pp. 309-17, 1957.

[32] H. P. Luhn, "The automatic creation of literature abstracts," IBM journal, pp. 159-65, 1958.

[33] C. Y. Lin, Robust automated topic identification, Phd thesis, Faculty of the Graduate School, University of Southern California, 1997.

[34] D. Lewis, An evaluation of phrasal and clustered representations on a text categorization task. In Croft et. al. (Ed.), Proceedings of SIGIR-95, 15th ACM International Conference on Research and Development in Information Retrieval, New York, ACM Press, 1995, pp. 37-50.

[35] Y. Yang, and J. O. Pedersen, "A comparative study on feature selection in text categorization," in 14th International Conference on Machine Learning, 1997, pp. 412-420, San Francisco, USA.

[36] D. J. Ittner, DD..Lewis, DD. Ahn, "Text categorization of low quality images," in 14th Annual Symposium on Document Analysis and Information Retrieval, SDAIR-95, 1995, pp. 301-315.

[37] F. Sebastiani, "Machine learning in automated text categorization", ACM Computing Surveys, vol. 34(1), pp. 1-47, 2002.

[38] D. Mladenic, *Machine learning on non-homogenous distributed text data*, Phd thesis, Computer and Information Science, Slovenia university, 1998.

[39] A. Brun, *Topic detection and statistical language model adaptation for speech recognition (In French)*, Computer Science, Nancy 2 University, 2003.

[40] K. Seymore, S. Chen, and R. Rosenfeld, "Nonlinear interpolation of topic models for language model adaptation," in Proceedings of the International Conference on Spoken Language Processing, 1998.

[41] D. Pinto, H. Jiménez, and P. Rosso, "Clustering abstracts of scientific texts using the transition point technique," in Proceedings of the 7th Int. Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2006, Springer-Verlag, LNCS, vol. 3878, pp. 536–546, 2006.

[42] T. Joachims, *A probabilistic analysis of the rocchio algorithm with tfidf for text categorization*. Technical report, School of Computer Science Carnegie MellonUniversity Pittsburgh, 1996.

[43] G. Salton, "Developments in Automatic Text Retrieval," Science, 253, pp. 974-979, 1991.

[44] K. Seymore, and. R. Rosenfeld, "Using Story Topics for Language Model Adaptation," in Proceeding of the European Conference on Speech Communication and Technology, 1997.

Mourad Abbas was born in Algiers in 1971. He obtained an engineer diploma in Electronics form University of Science and Technology (Algiers) in 1997, then the Magister in 2002. He obtained a Phd from the Electrical and Computer Engineering Department of the National Polytechnic School (Algiers) in 2008.

He joined c.r.s.t.d.l.a. in 2000 in where he was a responsible for studies then research associate, senior research associate and actually he is senior fellow and head of Phonetics and Speech Processing Laboratory. He taught Signal processing and Radio communications and advised engineers and actually he advises Master students. 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 and a member of ASTF (Arab Science and Technology Foundation) in the same year.

Kamel Smaïli was born in Algeria in 1963, he obtained an engineer diploma in computer science from University of Science and Technology (USTHB-Algiers). He then obtained a research master from university of Nancy 1 (France). In 1991, he obtained a PhD diploma from university of Nancy 1. His PhD concerned the development of a

continuous speech dictation machine based on statistical models. In 2001, he obtained an HDR (Habilitation à Diriger la Recherche) in the topic of language models for speech recognition and machine translation.

He spent two years at CNRS and since 2002, he is professor at university Nancy 2. His research interest since 20 years concerns statistical language modeling for speech recognition and since 2000 speech-to-speech translation. He proposed several original ideas: retrieving phrases based on class-phrases, purging statistical language models from impossible events, Cache-features language model, multilingual triggers,…

Pr. Smaïli participated to several European and French projects concerning speech recognition: COCOS, MULTIWORKS, COST, MIAMM, IVOMOB (RNRT project). He advised more than 10 Phd students and participated to 20 Phd committees through the world. He took part to several program committees: Eurospeech, ICSLP, ICASSP, SIIE, TAIMA, TAL, and reviewed papers for several journals: Computer speech and language, Speech communication, EURASIP. He published his research in more than 55 international conferences and journals and in more than 20 francophone 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 Sc. D. degrees from the National Polytechnic School (NPS), Algiers.

In 1979, he became Lecturer, Associated Professor then full Professor teaching signal processing and information theory in the Department of Electronics of NPS. During this period, his research activities involved the applications of signal processing and the source coding theory.

In 1992, he joined the Department of Electrical Engineering of University of Sherbrooke, Canada, where he taught signal processing. He was a member of the Speech Coding team of the University of Sherbrooke. He has been conducting research in the area of speech coding and speech processing in adverse conditions.

In 1995, he backs to the Department of Electrical and Computer Engineering of the National Polytechnic, Algiers. His current research interests include signal and communications, information theory concepts and clustering and adaptive algorithms applied to speech and image processing. He is author of more than 150 papers.