Comparing TR-Classifier and KNN by using Reduced Sizes of Vocabularies

Mourad Abbas, Kamel Smaïli, D Berkani

Comparing TR-Classifier and KNN by using Reduced Sizes of Vocabularies

M. Abbas* , K. Smaili** and D. Berkani***

* CRSTDLA /Speech Processing Laboratory, Algiers, Algeria. Email: m_abbas04@yahoo.fr
** INRIA-LORIA/Parole team, Villers les Nancy, France. Email: kamel.smaili@loria.fr
*** NPS/ Signal and Communication laboratory, Algiers, Algeria. Email: dberkani@hotmail.com

Abstract—**The aim of this study is topic identification by using two methods, in this case, a new one that we have proposed: TR-classifier which is based on computing triggers, and the well-known k Nearest Neighbors. Performances are acceptable, particularly for TR-classifier, though we have used reduced sizes of vocabularies. For the TR-Classifier, each topic is represented by a vocabulary which has been built using the corresponding training corpus. Whereas, the kNN method uses a general vocabulary, obtained by the concatenation of those used by the TR-Classifier. For the evaluation task, six topics have been selected to be identified: Culture, religion, economy, local news, international news and sports. An Arabic corpus has been used to achieve experiments.**

Keywords — **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].

The aim of this study is to evaluate two text categorization methods on Arabic corpora. The first one is TR-classifer [13], a new method based on computing triggers, and the second one is the k Nearest Neighbors.

The motivation behind the conception of the TR-classifier is that the information present in the longer-distance history is significant [14]. Indeed, for a topic identification task, the presence of the term "guitar" could trigger another list of terms: "music", "dance", etc. So, the main idea is to represent each topic by a number of triggers which allow characterizing 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 [15] as one of the top-performing classifiers, we selected it to evaluate TR-classifier by comparing performances of the two methods.

We should note that small sizes of topic vocabularies are chosen for TR-classifier. Whereas, a general vocabulary is needed for kNN, consequently it has been constructed by the concatenation of the topic ones.

The Arabic corpus used in our experiments is downloaded from the website of the Omani newspaper Alwatan[1]; it is composed of more than 9000 articles.

In section II, we give some details both about TR-classifier and kNN method. We talk, in Section III, about the corpus, the representation of documents and the vocabulary construction. Experiments and results are exposed in section IV.

II. METHODS DESCRIPTION

A. TR-Classifier

We start by giving a definition of Triggers as TR-classifier is based on computing them. So, triggers of a word w_k are the ensemble of words that have a high degree of correlation with it [13, 14, 16]. 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. Couples of words or "triggers" that are considered important for a topic identification task, are those which have the highest average mutual information (*AMI*) values [17, 18]. Each topic is then endowed with a number of selected triggers M, calculated using the training corpus of the topic T_i. Identifying topics by using TR-method consists in:
• Giving corresponding triggers for each word $w_k \in V_i$, where V_i is the vocabulary of the topic T_i.
  • Selecting the best M triggers which characterize the topic Ti.
  • In the test step, we extract for each word w_k of the test document, its corresponding triggers.
  • Computing Q_i values by using the TR-distance given by (1):

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

Where i stands for the ith topic. The denominator presents a normalization of *AMI* computation.
w_k^i are triggers included in 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 arg max Q_i.

---

[1] http://www.alwatan.com

B. k Nearest Neighbors

kNN has been applied to text categorization before two decades [19, 8, 20, 15]. Indeed, Yang compared it to a set of text categorization methods using the benchmark Reuters corpus (the 21450 version, Apte set) [15]. It has been found that KNN is one of the top-performing methods after SVM [15]. Many other researches have found that the kNN method achieves very good performances by using different data sets [15, 21, 22].

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 (2).

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

where D_j is the jth training document and D_i is the ith test document. $|V|$ is the size of the general 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 [15].

III. 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.

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 dress in table 1 the size of the entire corpus, before and after removing insignificant words.

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

The construction of the vocabulary has been made by using the term frequency method which gives good results though its simplicity [23]. Other terms selection methods as Mutual Information [24] and Document Frequency lead also to satisfactory performances.

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 [25]. Hence, after removing insignificant words, we calculated both the frequency of each word, which is called 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 [25, 26, 27].

IV. EXPERIMENTS AND RESULTS

A. TR-Classifier Performances

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 300 terms.

The evaluation of the TR-classifier has been made by varying both topic vocabularies sizes and triggers number N. So, we present performances on average in figure 1.
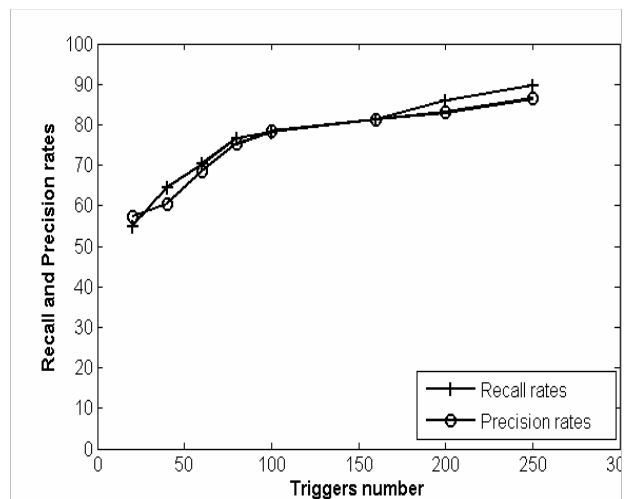

Figure 1. TR-Classifier performances using a vocabulary size 300 in terms of Recall and Precision

The number of triggers is important to achieve good performances. According to figure 1, we notice that average Recall (R) and Precision (P) rates improve when increasing the triggers number. Indeed, for N=20, R and P are respectively equal to 54.88 % and 57.30 %. These values continue to be enhanced when N is augmented to 250, in this case R=89.67 % and P=86.44 %.

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.

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

Performances are lower than those of TR-Classifier by 14 % which is considered as an important difference between the two methods. We present in figure 2, performances in terms of Recall of the two classifiers, for the six studied topics.

TABLE II.
PERFORMANCES OF THE KNN METHOD

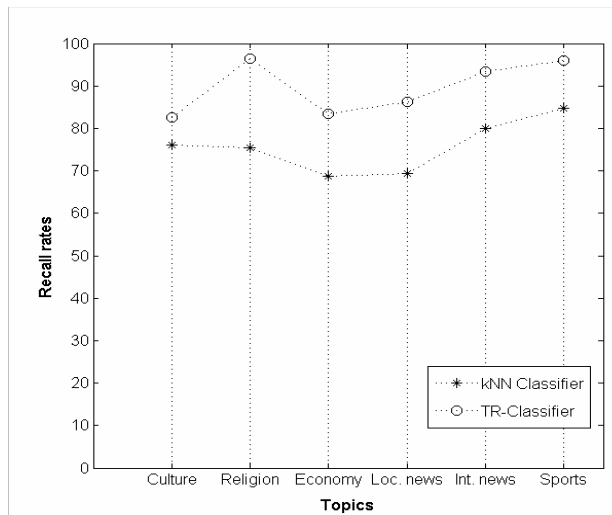| Topics	Recall (%)	Precision (%)
Culture	76	49.78
Religion	75.33	94.95
Economy	68.66	81.74
Local	69.33	70.27
International	80	85.11
Sports	84.66	92.70
Average	75.66	70.09

Figure 2.   TR-Classifier performances compared to kNN ones

CONCLUSION

In this paper, two methods of topic identification have been presented. Their performances have 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 performances by using reduced sizes of topic vocabularies, compared to kNN. The reason behind that, 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 best performances; nevertheless in the case of small vocabularies, as shown in the aforementioned experiments, its performances didn't exceed 76 % in terms or Recall.

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

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. September 2006.

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

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

[15] 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.

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

[17] 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.

[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] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

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

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

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