Journal of Theoretical and Applied Information Technology
10th April 2013. Vol. 50 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EXPERIMENTS WITH ARABIC TOPIC DETECTION

[1]RIM KOULALI, [2]ABDELOUAFI MEZIANE

Department of Mathematics and Informatics, LARI Laboratory, Sciences College, Mohammed I University
MOROCCO
E-mail: [1] rim.koulali@gmail.com , [2] abdelouafi_meziane@yahoo.fr

ABSTRACT

The continuous growth of information on the Internet and the availability of a large mass of electronic documents in Arabic language make Natural Language processing (NLP) tasks play an important role to enhance and facilitate the access and the exploitation of information. Among available NLP tasks, we are interested in Arabic Topic Detection. Our objective is to realize an indexing system capable of identifying the general topics discussed in Arabic unvowelized documents. The proposed topic detection system of Arabic texts is based on Mutual Information for Topic Oriented Vocabulary (TOV) and classification according to Jaccard and adapted TF-IDF indicators. The experimental results are presented in terms of precision, recall and F1 measure evaluating the influence of factors such as: vocabulary length and morphological analysis on Arabic Topic Detection.

Keywords: *Natural language processing (NLP), Topic Detection (TD), Topic Oriented Vocabulary (TOV), Mutual Information (MI), Jaccard Indicator, TF-IDF.*

1 INTRODUCTION

Automatic processing of Arabic documents is an active research field that received increasing attention to confront this language to the technological challenges. Arabic Language Processing calls for innovative techniques that take into consideration, particularities and complex morphological composition of the language.

Topic detection is an important task in Automatic Natural Languages Processing (Martin et al., 1997), (Allan et al., 1998), (Yang et al., 2002) due to its various application fields: Summarization, Speech Recognition, Question-Answer system… Topic Detection enables the automatic identification of semantic content and assignment of a topic label to a given document. Topic Detection is based on supervised machine learning using a training corpus to represent each topic with a specific model obtained using a wide range of text processing approaches, text representation methods, and distance measures to estimate similarities between topics and documents vectors. Thus, the topic of a given document is detected whenever their contents are similar

Our work is part of the semantic processing of unvowelized Arabic documents and aims to develop a Topic Detection system for Arabic texts based Topic Oriented Vocabularies (TOV), using Mutual Information, Jaccard indicator and an adaptation of the TF-IDF indicator to topic detection. Through experiments, we study the influence of the vocabulary length and using a morphological analysis on the performances of the Topic Detection system,

This paper is organized as follows: In section 2 we present related works. In section 3; our approach is explained. Section 4 presents the data set and details the conducted experiments and obtained results. The last section concludes the paper.

2 RELATED WORKS

Works on topic detection in Arabic documents are very few due to the complex morphological specificities of the Arabic language.

(Sawaf et al., 2001) realized a classification and a clustering document system based on the extraction of key phrases using statistical methods such as Mutual Information and maximum entropy modelling.

(El-Kourdi, 2004) used Naïve bayes algorithm to classify unvowelized Arabic web documents and achieved 68.78% in cross validation and 62% in evaluation set experiments.

A comparative study between two methods of classification: TF-IDF and SVM by (Abbas and Smaili, 2005) shows that both of the two methods give good result for Arabic documents: TF-IDF achieves results of 90.95 % in term of F-measure.

SVM classifier outperforms the results obtained by TF-IDF by more than 7.5% in terms of F-measure.

The experiments of (Abbas and Berkani, 2006) included the study of three statistical methods: the TF-IDF classifier, the SVM method and the Topic Unigram Language Model, and showed the superiority of the SVM classifier and its high capability to distinguish topics.

(Abbas et al, 2010) proposed a study of topic identification for Arabic language by using two methods: the k Nearest Neighbors which is used as a baseline and the TR-Classifier which gives the best performances using reduced size of topic vocabularies with 90% in terms of recall.

3 TOPIC DETECTION SYSTEM

Our developed topic detection system relies on topic oriented vocabularies to classify Arabic documents. Each topic is described by a vector of words (vocabulary) that outline specifically and accurately the topic (Brun et al., 2003). Various methods for vocabulary generation were proposed in the literature. Although the frequency of words in the corpus is the most intuitive approach, other methods based on probabilistic measurements are more efficient such as: Mutual Information, Gain Information, and Unigram Model....

The generated topic oriented vocabularies describe a semantic relationship between words in documents of the same topic. The topic detection System matches documents against the generated TOV to determine the general topic of each document. The implementation of our system is based on several phases to identify and classify the texts according to their general topic.

3.1 Documents pre-processing

This phase is crucial in order to extract relevant information. It consists of the following steps:

- Unify the documents encoding: the unification of the encoding is used to represent all the documents in the same repository. We adopted the UTF-8 which supports the Arabic language;
- Document normalization: Suppression of: vowels, Latin words, symbols, numbers, markers, special characters...
- Stop words elimination: Suppression of noisy words by comparing each word with the elements of a hand crafted list containing over 600 stop words including: prepositions, demonstrative pronouns, identifiers, logical connectors...
- Root and stem extraction: Although several articles on classification estimate that working with words roots favors the obtaining of efficient results due to the reduction of the noise and best qualification of words, we conduct a comparative study between roots and stems to evaluate which is the most effective for the Arabic language. To achieve that, we used the morphological analyzer: Alkhalil(A. Boudlal et al., 2011). We adapted Alkhalil to recover for each document two lists: one for stems and the other for roots. Alkhalil realizes morphological analysis for each word in the corpus and returns among other morphological information all possibly related stems and roots to the considered word. So, we used a Viterbi algorithm to keep only the stems and roots that are relevant to the context.

3.2 vocabulary oriented topic generation

Our approach is based on the generation of a specific vocabulary for each topic.

The vocabulary is composed by words which define specifically the topic. Various methods are used to create these vocabularies, we used the Mutual Information (MI) method which measures the influence of a word W on each topic T and attribute the word to the appropriate topic (the one that has the greatest MI value) based on the formula:

$$MI\ (W, T)=\log\ (P\ (W \mid T)) - \log\ (P(W)) \qquad (1)$$

We calculate the MI for each word in the train corpus with each topic. The word will be affected to the topic which gives the maximal value (Yang and Pedersen, 1997).

Finally, we generate six vocabularies specific to each topic of the training corpus by considering for each one the affected words ordered by decreasing MI. We considered vocabularies of various lengths: 50, 100, 150, 200, 250 and 1000.

3.3 topic detection

During this phase, we tested our system on the test corpus with 2035 articles belonging to six topics.

Each document is represented with a vector of words composing it and each topic is represented by its vocabulary vector.

The topic detection of a new document consists in calculating the similarity between the vector representing the document and those representing the topics. We based our similarity calculus on two approaches: the first one based on the Jaccard indicator and the second one using TF-IDF as term weighting method and Cosine similarity,

3.3.1 Jaccard Indicator

The Jaccard indicator (Real and Vargas, 1996) measures the degree of similarity between two documents. The indicator is expressed as:

$$sj = \frac{mc}{md + md' - mc} \qquad (2)$$

WHERE:

- m_d : Total number of words of the first document.
- $m_{d'}$: Total number of words of the second document.
- m_c : Number of common words between the two documents.

The choice of Jaccard indicator is motivated by the fact that it employs words in their brut state and thus semantic information is accounted for.

3.3.2 TF-IDF

We adapted the classic TF-IDF (Salton, 1991), (Seymore and Rosenfeld, 1997) to assign a wight for each word of each topic vocabulary.

The weight of the k^{th} vocabulary word of topic j is expressed ad fellow:

$$t_{jk} = \mathrm{nf}_k^j * \mathrm{idf}_k \qquad (3)$$

Where nf_k^j is the frequency of the word k in documents of the training corpus relative to topic j.

Let df_k be the number of documents not relative to topic j in which the word k appears at least once and N the total number of corpus documents. idf_k , the inverse document frequency, is given by:

$$\mathrm{idf}_k = \log\left(\frac{N}{df_k}\right) \qquad (4)$$

Test documents are also represented by vectors containing weights of their words. To judge the similarity between a topic t and a document d, we used the Cosine similarity:

$$\cos(\theta) = \frac{\sum_{k=1}^{n} d_{ik} t_{jk}}{\sum_{k=1}^{n} d_{ik}^2 \sum_{k=1}^{n} t_{jk}^2} \qquad (5)$$

The smaller the θ is the bigger is the similarity between a test document d and topic t. The topic of highest similarity will be assigned to the test document.

4 EXPERIMENTS AND RESULTS

4.1 Data set

For the set up of our experiments, we used a corpus of over 20.291 articles, collected from the Arabic newspaper Wattan of the year 2004 (Abbas, Smaili, & Berkani, 2010). The corpus contains articles covering the six following topics: culture, economics, international, local, religion and sport. The repartition of documents is described in Table 1. The corpus was divided into two subsets of documents. Thus, 9/10 of the corpus was dedicated to training the feature selection system (Topic vocabulary construction), whereas 1/10 of the overall documents formed the evaluation corpus.

Table 1: Number of documents and words per topic.

Topics	Number of articles	Number of words
Culture	2782	1.359.210
Economy	3468	3.122.565
International	2035	855.945
Local	3596	1.460.462
Religion	3860	1.555.635
Sport	4550	9.813.366

4.2 Evaluation metrics

In order to evaluate the classifiers performances, three standard metrics are used: Recall, Precision and F1-measure formulated as follows:

$$Recall = \frac{Number\ of\ correctly\ labelled\ documents}{number\ of\ topic's\ documents} \quad (6)$$

$$Precision = \frac{Number\ of\ correctly\ labelled\ documents}{number\ of\ labelled\ documents} \quad (7)$$

$$F1-measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

4.3 Results and discussion

We conducted several experimentations on the test corpus to study and compare the influence of factors such as:

- The variation of the vocabulary size.
- The morphological nature of words: stem or root.
- The use of nouns only.

Figure 1 depicts the F-measure metric realized by our proposed topic identification system using Jaccard Indicator over two morphologically distinct corpora.

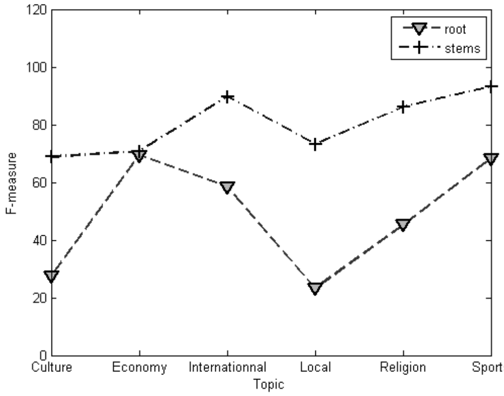

Figure 1: F-measure, Stems vs Roots vocabulary of 250 words.

The former was obtained by stemming the corpus documents; while the latter's documents contain only the roots of words. It shows that the developed system realizes higher performances when using the stemmed corpus. The best performance is obtained for the Local topic with 14.87% of enhancement.

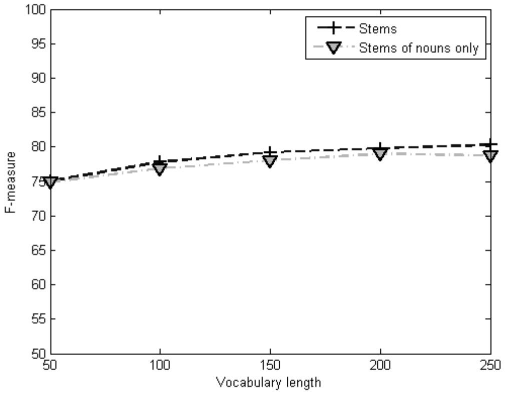

Figure 2: F-measure, Stems vs stems of nouns only.

To study the impact of removing non-noun words from the corpus on the system performances we used Stanford Part-Of-Speech Tagger for Arabic language to identify nouns contained in our corpus and stripped documents from other words. Figure 2 indicates that the system performances are not heavily affected by this operation. The performance decrease is around 1%. Thus, we conclude that nouns in the Arabic language are more useful to construct VOT than verbs, adjectives, adverbs …. As they hold the essential of semantic information.

Statistical classifiers such as TF-IDF are widely used in topic detection literature. We compare the performance of the developed system for TF-IDF and Jaccard indicator.

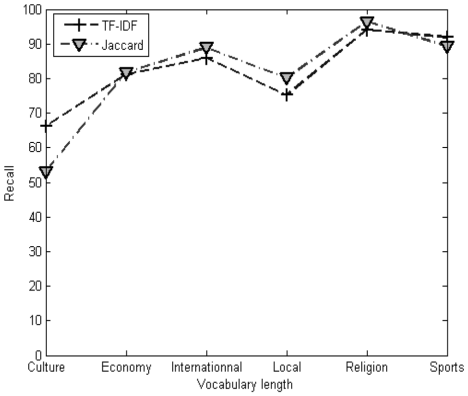

Figure 3: Recall, TF-IDF vs Jaccard for a vocabulary of 1000 stemmed words.

Figure 3 indicates that the recall of our system using Jaccard indicator on stemmed documents outperforms the results obtained by TF-IDF except for the culture topic. This can be explained by the fact that the culture topic contains numerous words that are shared with other topics and this fact results in classification errors. The F-measure results depicted in Figure 4 correlate with Recall.

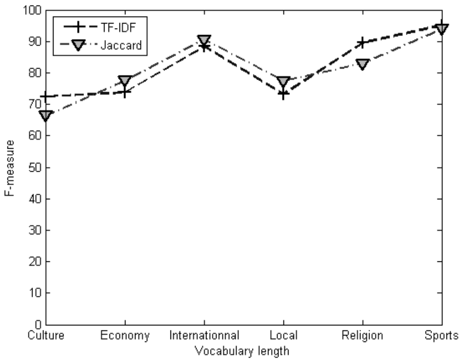

Figure 4: F-measure, TF-IDF vs Jaccard for a vocabulary of 1000 stemmed words.

5 CONCLUSIONS

We developed a topic detection system based on topic oriented vocabulary for the Arabic Language using Jaccard indicator. Our system manages to achieve 84% of correctly classified documents over Wattan corpus. We conducted several experiments to improve our system performances.

We showed that working with the stemmed corpus is more efficient than using roots. Also, we found that stripping Wattan corpus from non-noun words does not affect our system performances. The resulting performances decrease is only of 1%. The results obtained for Jaccard indicator were compared to the standard classifier TF-IDF. The results prove that the use of Jaccard indicator is judicious in Arabic Language topic detection.

REFRENCES:

[1] Martin, S., Liermann, J., Ney, H., 1997. Adaptive topic-dependent language modelling using word-based varigrams. *Fifth European Conference on Speech Communication and Technology*, volume 3, pp. 1447–1450.

[2] Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y., 1998. Topic detection and tracking pilot study: Final report. *Proceedings of the DARPA broadcast news transcription and understanding workshop*, pp. 194 to 218.

[3] Yang, Y., Zhang, J., Carbonell, J., Jin, C., 2002. Topic-conditioned novelty detection. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 688 to 693.

[4] Sawaf, H., Zaplo, J., Ney, H., 2001. Statistical Classification Methods for Arabic News Articles. *Natural Language Processing in ACL2001*, Toulouse, France.

[5] El-Kourdi, M., Bensaid , A., Rachidi,T. 2004. Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm, *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland, pp. 51 to 58.

[6] Abbas, M., Smaili, K., 2005. Comparison of Topic Identification Methods for Arabic Language, *Recent Advances in Natural Language Processing*, Bulgaria, pp. 14 to 17.

[7] Abbas, M., Berkani, D., 2006. Topic Identification by Statistical Methods for Arabic language. *Wseas Transactions on Computers*, 5(9), pp. 1908 to 1913.

[8] Abbas, M., Smaili, K., Berkani , D., 2010, TR-Classifier and KNN Evaluation for Topic Identification tasks, *Special Issue on Advances in Arabic Language Processing, the International Journal on Information and Communication Technologies (IJICT)*, 3(3), pp. 65 to 74.

[9] Brun, A., Smaili, K., Haton, JP., 2003.Nouvelle approche de la sélection de vocabulaire pour la détection de thème, *TALN 2003*, Batz-sur-Mer, France, pp. 45 to 54.

[10] A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane M. Ould Abdallahi Ould Bebah and M. Shoul., 2011: Alkhalil MorphoSys: A Morphosyntactic analysis system for non vocalized Arabic, *Seventh International Computing Conference in Arabic (ICCA 2011).*Riyadh.

[11] Yang, Y., Pedersen, J. 1997. A comparative study on feature selection in text categorization. *14th International Conference on Machine Learning, ICML-97*, San Francisco, pp. 412 to 420.

[12] Real, R., Vargas, J.M., 1996. The probabilistic basis of Jaccard's index of similarity, *Systematic biology*, Oxford University Press,45(3), pp. 380 to 385.

[13] Salton, G., 1991. Developments in Automatic Text Retrieval. *Science*, 253, pp.974 to 979.

[14] Seymore, K., Rosenfeld, R., 1997. Using Story Topics for Language Model Adaptation. *In Proceeing of the Eeuropean Conference on Speech Communication and Technology*