Topic Identification of Noisy Texts: Statistical Approaches

K. Abainia

Faculty of Electronics and Informatics,
University Houari-Boumediene of Sciences and Technology

***Abstract*—** This paper deals with the problem of automatic theme identification of noisy Arabic texts. Actually, there exist several works in this field based on statistical and machine learning approaches for different text categories. Unfortunately, most of the proposed approaches are suitable in clean and long texts. In this investigation, we carried out a comparative study between two different statistical approaches based on tf-idf. Hence, different configurations were used in both approaches to provide a large comparison. Furthermore, an in-house corpus called ANTSIX was created to evaluate the proposed approaches, which contains discussion forum texts related to 6 different topics.
Experimental results show that the two statistical approaches are suitable for topic identification of noisy Arabic texts, but each technique has advantages and drawbacks.

Keywords—Natural Language Processing; Text categorization; Automatic Topic Identification; Nearest Prototype; High Frequency.

I. Introduction

The amount of shared human knowledge is continuously increased over different numerical mediums, and that motivated many researchers to investigate in designing algorithms capable to retrieving and categorizing information (i.e. Text Mining field).
Nowadays, there exist a lot of works in text mining in different areas and different languages (Section 2 presents related works). However, there exist few works in the Arabic language, which is a difficult language having a complex morphology. Also, the Arabic language has some particularities, such as the use of diacritics (*Tashkil* in Arabic) and the *Shadda* character, which replaces the letter repetition twice.

Hence, the modification of one diacritic can change totally the word meaning. For instance, the Arabic word (خَرَجَ) with the *Fatha* diacritic at the middle (means "he went out" in English), it becomes (خَرِجَ) with the *Kasra* diacritic at the middle (means "he blended" in English).

In this investigation, we deal with the problem of topic identification of noisy Arabic texts, for which a new in-house corpus (ANTSIX) was constructed, which contains discussion forum texts related to 6 different topics (Section 3).

A comparison between two statistical approaches based on tf-idf was carried out to study the performance of each one in the case of noisy Arabic texts (Section 4). Then, the first one is called High Frequency approach, in which different n-gram features and different numbers of keywords were used and compared. On the other hand, the second approach is the Nearest Prototype approach, which is based on computing the similarity between documents. Hence, 4 different distance measures are used and compared.

Finally, the experimental results (described in Section 5) figured out that each approach had advantages and inconvenient, however, the two approaches are suitable for topic identification of noisy Arabic texts.

II. Related works

During the last decades, several researchers were interested in the topic identification field of written texts, because recognizing text topics is a primary step of other text mining fields.

For instance, a module for the topic identification was proposed in [Skorkovská, 2011], where it is embedded in a complex system containing a large dataset. The system is based on a tree of keywords and attributes one or more keywords to the text. The approach was tested on newspaper articles, and unfortunately the results were not suitable. Another work presented in [Massey, 2011], which is different from the existing techniques and is inspired from how the human brain processes the information. The algorithm requires an external dictionary to represent the knowledge base, and the accuracy was about 36%.

A comparative study of topic identification was performed in [Bigi, 2001], in which the authors tested five statistical methods of topic identification (i.e. unigrams, tf-idf, cache model, topic perplexity and weighted model) on newspaper and e-mail corpus. The max accuracy reached was about 97.1% on newspaper corpus, and it differs from corpus to another and from method to another. Another comparative study was performed in [McDonough, 1994], in which the authors divided the topic identification problem into 3 components: event generation, keyword event selection and topic modeling. Even, they tested different approaches and compared the relative effectiveness of each one. The accuracies depend on the number of keywords and the approach used in each component.

A graph approach was used previously in [Coursey, 2009] and [Coursey, 2009-bis], which was based on graph centrality and tested on Wikipedia topics. The accuracy was not high, but the results figured out that the use of external knowledge dictionary improves the baseline performances.

Some machine learning approaches were used in topic identification. For instance, the unsupervised neural network based on self-organizing map algorithm was presented in

[Lagus, 2002-bis]. The authors used a corpus of dialogue texts to evaluate their approach, and they compared different methods for model parameter estimation. The accuracy was about 87.7%, which was better for the longer dialogue segments. Another work based on neural networks performed in [Özmutlu, 2004], in which the authors used excite web search engine data logs. The results figured out the neural networks can achieve good accuracy, which may be compared to the results given by the human expert. A novel neural network different from the others was introduced in [Jo, 2009], and it was called neural text categorizer (NTC). Hence the input vector is a string vector and not a numerical vector like the others. The authors evaluated the NTC on newspaper corpus, where the accuracy was slightly better than the back propagation one (about 80% of accuracy).
A hierarchical approach based on the ontology was presented in [Tiun, 2001], in which the authors decomposed the classification into three steps: sentences extraction, keywords mapping on the ontology concepts and finally the ontology tree optimization. The experiments performed on Yahoo topics, and they reported 69.8% of best accuracy.

III. Corpus

A new in-house corpus ANTSIX (Arabic Noisy Texts in SIX topics) was manually constructed and containing 6 Topics namely: health, economy, religion, sports, informatics and hunting. Hence, the texts are collected over Arabic discussion forums, and they contain any kind of noises like URLs, tags, abbreviations, citations in other languages, typing errors…etc. On the other hand, we noticed that one text forum corresponding to a specific topic may contain words related to another topic.
The dataset is unbalanced in term of number of words, where the length ranges between 32 and 318 words. Hence, it contains 50 texts (encoded with UTF-8) per topic; thus, in overall there are 300 texts corresponding to 6 topics.

IV. Topic identification

The general scheme of topic identification followed in this work is based on four main steps illustrated by Figure 1. We have proposed two statistical approaches namely: High Frequency approach and Nearest Prototype approach, respectively.

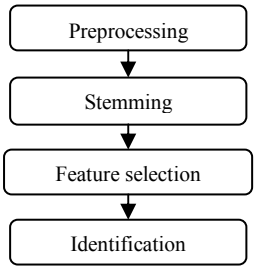

Figure 1. Global scheme of topic identification.

The preprocessing step consists of removing insignificant characters, removing non Arabic characters (e.g. English and French characters), removing Arabic diacritics (*Tashkil* in Arabic), and finally, removing the stop words.

The stemming is the process having the goal of reducing the dimensionality of the data and regrouping the relative words, where it consists of transforming the words to their roots or their stems.
Feature selection is the selection of a set of pertinent features, which bring good accuracies. Hence, the feature selection passes by two stages, the first one is attributing weights to the features using tf-idf techniques (equation 1). Whereas, the second stage consists in selecting a subset features by their frequency ranks.

$$tfidf\ (t\ ,\ d) = tf_{(t\ ,d)} * idf_t \qquad (1)$$

where $tf_{(t\ ,d)}$ is the term frequency t in the document d, and idf_t is the inverse document frequency of the term t and given by the following equation:

$$idf_t = \log(N/n) \qquad (2)$$

where N is the total number of documents, and n is the number of documents containing the term t.

A. High Frequency Approach

The High Frequency approach is based on selecting F most frequently features, which represent topic keywords. Next, the identification is based on computing the sum of keyword frequencies for each topic using the following equation:

$$proba_t = \sum_{i=1}^{F} freq_{t\,i} \qquad (3)$$

where $freq_{t\,i}$ is the frequency of the ith keyword of the topic t in the text. If the keyword does not exist in the text, then its frequency will be set to 0.

Finally, the promising topic is the one having the highest sum of frequencies.

B. Nearest Prototype Approach

The nearest Prototype approach is based on representing reference documents and unlabeled documents as profiles containing all features. Next, the similarity between reference profiles and input profiles is computed using distance measures. Finally, the promising topic is one having the minimal distance.
We have used 4 distance measure are used as follows: Euclidean, Cosine, Bray Curtis and Histogram Intersection distances.

V. Experiments and results

Our experiments were carried out on the ANTSIX corpus, where 60% of the corpus was used to creating reference profiles, whereas 40% was reserved for the test purpose. In the High Frequency approach, different n-gram features were used and compared including uni-gram words, bi-gram characters, tri-gram characters and tetra-gram characters. Moreover, different numbers of keywords were tested to study their impact on the topic identification of short and noisy texts (from 5 to 200 keywords). However, in the Nearest Prototype approach, only uni-gram words were used.

In the training process, all the texts of each topic were concatenated together in the same text after applying different preprocessing steps. Finally, the topic profiles were created using the previous approaches (described above).

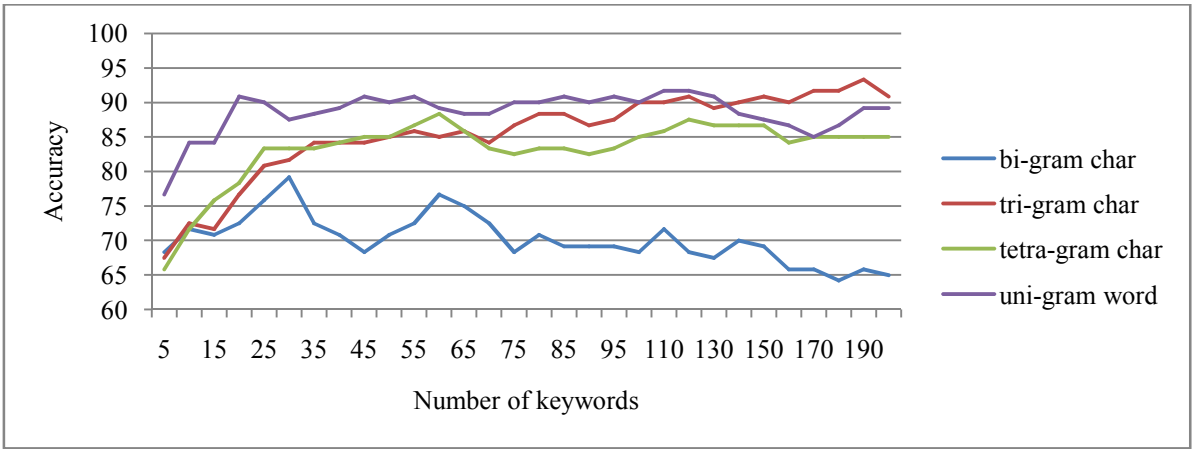

Figure 2. Comparative accuracies obtained by the High Frequency approach.

Figure 2 exposes different accuracies obtained by the High Frequency approach with uni-gram words, bi-gram characters, tri-gram characters and tetra-gram characters using different numbers of keywords. As depicted in the figure, in overall, the uni-gram words are more accurate than the other features. On the other hand, tri-gram characters reach the best accuracy (about 93.33% of accuracy). Contrariwise, the worst feature is bi-gram characters.

We notice that the increase of the number of keywords, leads to continuously increasing the tri-gram accuracy to reach 93.33% as a max accuracy, on the contrary of the bi-gram accuracy, which is decreased over the number of keywords.

Therefore, in term of optimizing the computation time by using a small number of keywords it is preferable to use uni-gram words as feature. Contrariwise, it is recommended to use tri-gram characters to reach best performances.

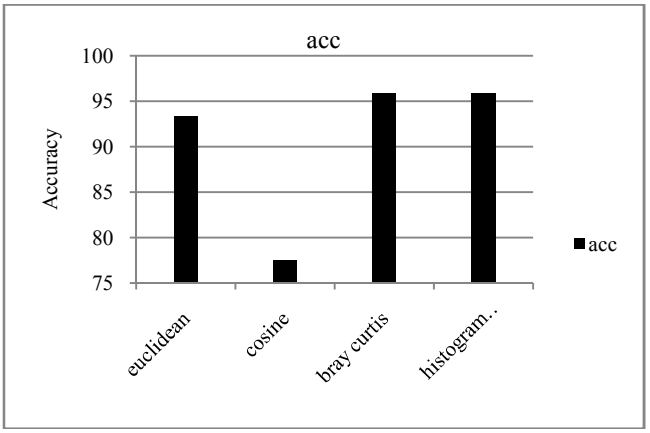

Figure 3. Comparative accuracies obtained by of the Nearest Prototype approach.

Figure 3 exposes the accuracy obtained by the Nearest Prototype approach using the Euclidean, Cosine, Bray Curtis and Histogram Intersection distance.

As depicted in Figure 3, the Bray Curtis and Histogram Intersection distances are more accurate (95.83% of accuracy) than Euclidean and Cosine distances, which are commonly used in the literature and proved good performances compared to machine learning techniques [Bigi, 2001] [Abbas , 2011].

We notice from the two figures that the tf-idf Nearest Neighbor approach using Cosine distance is not suitable for topic identification of noisy Arabic texts, and the Euclidean distance brings the same accuracy as High Frequency using tri-gram characters (93.33%). However, the Nearest Prototype approach is more reliable than the Highest Frequency approach in terms of performances. Contrariwise, in terms of optimizing computation time the High Frequency approach is cheaper than Nearest Prototype approach.

VI. Conclusion

In this investigation, several experiments of topic identification, in noisy Arabic texts, have been conducted and commented. An in-house corpus ANTSIX containing noisy Arabic texts have been manually constructed and used to evaluate and compare the proposed methods.

In this research work, a comparative study was carried out between two statistical approaches with different configurations. The first one is the High Frequency approach, which consists in computing the frequency of the keywords, where different numbers of keywords and different features were used and compared: uni-gram words, bi-gram character, tri-gram characters and tetra-gram characters. On the other side, the second approach was the Nearest Prototype approach based on TF-IDF, in which 4 distance measures were used and compared: Euclidean, Cosine, Bray Curtis and Histogram Intersection distances.

Experimental results figured out that the use of uni-gram words in the High Frequency approach optimized the computation time, whereas the use of tri-gram characters enhanced

the performances. Contrariwise, the use of bi-gram and tetra-gram characters was not suitable at all.
It was also noticed that the Bray Curtis and Histogram Intersection distances in the Nearest Prototype approach were more accurate than Euclidean and Cosine distances by reaching the highest accuracy (95.83%).
The second approach was better than the first one in term of accuracy, whereas the first approach was optimized in term of computation time.
Therefore, the standard statistical approaches can be considered to deal with topic identification of noisy Arabic texts, and they bring acceptable accuracies. As future work, we suggest trying other techniques and methods based on n-gram characters and words to enhance the identification performances.

References

[Abbas , 2011] M. Abbas and K. Smaili and D. Berkani, "Evaluation of Topic Identification Methods on Arabic Corpora", Jouranl of Digital Information Management, Vol.9 (2011), pp. 185-192.

[Bigi, 2001] B. Bigi, A. Brun, J. Haton, K. Smaili and L. Zitouni, A Comparative Study of Topic Identification on Newspaper and Email, Proceedings of the 8th International Symposium on String Processing and Information Retrieval, SPIRE'2001, Laguna de San Rafael, Chile, November 13-15, 2001, pp. 238-241.

[Coursey, 2009] K. Coursey and R. Mihalcea, Topic Identification Using Wikipedia Graph Centrality, Proceedings of NAACL HLT'2009, Boulder, Colorado, June 1-3, 2009, pp. 117-120.

[Coursey, 2009-bis] K. Coursey, R. Mihalcea and W. Moen, Using Encyclopedic Knowledge for Automatic Topic Identification Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), Boulder, Colorado, June 2009, pp. 210-218.

[Jo, 2009] T. Jo, Neural Text Categorizer for Exclusive Text Categorization, Proceedings of the First International Conference on Networked Digital Technologies, NDT'09, Ostrava, July 28-31, 2009, pp. 26-31.

[Lagus, 2002-bis] K. Lagus and J. Kuusisto, Topic Identification in Natural Language Dialogues Using Neural Networks, Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue, Philadelphia, July11-12, 2002, pp. 95-102.

[McDonough, 1994] J. McDonough, K. Ng, P. Jeanrenaud, H. Gish, and J. R. Rohlicek, Approaches to topic identification on the SWITCHBOARD corpus, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'94, Adelaide, SA, April 19-22, 1994, pp. 385-388.

[Massey, 2011] L. Massey, Autonomous and Adaptive Identification of Topics in Unstructured Text, Proceedings of the 15th International Conference, KES'2011, Kaiserslautern, Germany, September 12-14, 2011, pp. 1-10.

[Özmutlu, 2004] H. Özmutlu, F. Çavdur, S. Özmutlu and A. Spink, Neural network applications for automatic new topic identification on excite web search engine data logs, Proceedings of the American Society for Information Science and Technology, Vol.41, No.1, 2004, pp. 310-316.

[Skorkovská, 2011] L. Skorkovská, P. Ircing, A. Pražák and J. Lehečka, Automatic Topic Identification for Large Scale Language Modeling Data Filtering, Proceedings of the 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011, pp. 64-71.

[Tiun, 2001] S. Tiun, R. Abdullah and T. Kong, Automatic Topic Identification Using Ontology Hierarchy, Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing, CICLing'01, Mexico City, Mexico , February 18–24 , 2001, pp. 444-453.