Topic Identification of Noisy Arabic Texts Using Graph Approaches

Kheireddine Abainia
USTHB University
Algiers
abainia@hotmail.fr

Siham Ouamour
USTHB University
Algiers
siham.ouamour@uni.de

Halim Sayoud
USTHB University
Algiers
halim.sayoud@uni.de

***Abstract*— This paper deals with the problem of automatic topic 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 methods are effective in clean and long texts. In this research work, we use an in-house dataset of noisy Arabic texts, which are collected from several Arabic discussion forums related to 6 topics.**

In this investigation, we propose a graph approach called LIGA for topic identification task. This approach was firstly introduced for language identification field. Moreover, we propose two other extensions in order to enhance LIGA performances. The experiments undergone on the Arabic dataset have shown quite interesting performances, reaching about 98% of accuracy.

Keywords—Topic Identification; Graph approach; Text Categorization; Natural Language Processing; TFIDF; Text mining.

I. INTRODUCTION

The growing amount of shared numerical information has attracted several researchers to work in automatic text categorization and knowledge extraction, Text Mining in general.

Nowadays, there exist a lot of works in text mining in different areas and different languages. However, there exist few works in the Arabic language. This language has a complex morphology and 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 work, we deal with the problem of topic identification applied to noisy Arabic texts. Thus, a new Arabic corpus (called ANTSIX) is constructed containing discussion forums related to 6 different topics. The dataset is unbalanced in term of text lengths (i.e. number of words).

We implemented three graph approaches. The first one (called LIGA) was introduced firstly for language identification by E. Tromp [1]. The second and third ones are modified versions of the LIGA approach based on tf-idf weights.

This paper is organized as follows. In section 2, we present some related works in topic identification. Section 3 defines the Arabic dataset used to evaluate the proposed approaches. In section 4 and section 5, we describe the preprocessing and the proposed approaches, respectively. Section 6 exposes and discusses the obtained results. A conclusion and some perspectives are given in section 7.

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 [2], 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 [3], 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 [4], 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 corpuses. 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 [5], 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 [6] and [7], 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 [8]. 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 [9], 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 [10], 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 [11], 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. DATASET

We manually constructed an Arabic dataset which we called ANTSIX or Arabic Noisy Texts in 6 Topics namely: health, economy, religion, sports, informatics and hunting. The texts are collected from different Arabic discussion forums, and they contain different kind of noises like URLs, tags, abbreviations, citations in other languages, typing errors…etc. Moreover, one forum text 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. The dataset contains 50 texts (encoded with UTF-8) per topic; thus, in overall there are 300 texts corresponding to 6 different topics.

We give in figure 1, an example of an Arabic forum text related to economics topic.

السلام عليكم ورحمة الله وبركاته

اعتقد انه حتى نتمكن من فهم النظرية الاقتصادية الكلية والجزئية من منظور إسلامي لابد من استيعاب وفهم هذه النظرية بمفهومها التقليدي. حيث ان عمر تلك النظرية ليس بالقصير فهي تمثل نتاج حضاري لابد من الاستفادة منه - سواء أتفقنا مع طروحاتهم ام اختلفنا معها .

ونحن بصدد الحديث عن نظرية اقتصادية إسلامية لانهدف الي ايجاد قوانين ونظريات مناقضة تماما لما هو مطروح في الفكر التقليدي, وإنما هي اتجاهات وأبعاد جديدة للمفاهيم المطروحة برؤية إسلامية.

Figure 1. EXAMPLE OF AN ARABIC FORUM TEXT BELONGING TO ANTSIX DATASET

IV. PREPROCESSING

The identification process is preceded by the preprocessing and stemming steps. Hence, the preprocessing step consists of removing insignificant characters like numbers, arithmetic operators, punctuation characters and non-Arabic characters (i.e. French and English characters)...etc. Furthermore, in case of Arabic language, an additional step of preprocessing is planned; this step concerns the removal of diacritics. Finally, stop words are removed after extracting all the words of the text.

The stemming is the process having the goal of reducing data dimensionality, and it consists of replacing the words by their roots or their stems. Thus, we can find the number of occurrences of the word and all its derivatives existing in the text.

So, we summarize the different preprocessing steps by the following points:

- remove insignificant characters
- remove French and English characters
- remove Arabic diacritics
- separate contracted words
- strip multiple word separators
- remove stop words
- Stem the rest of words.

V. PROPOSED APPROACHES

We propose 3 graph approaches for topic identification: the first one is called (LIGA), and it was firstly introduced for language identification [1]; the second and third ones are modified versions of the LIGA approach.

A. Graph Based Topic Identification (LIGA)

This approach was introduced by E. Tromp for language identification field based on character n-grams. In this work, we propose to use this method with uni-gram words instead of character n-grams.

In the LIGA approach, the training documents of each topic are used to create the topic graphs, in which each node contains a word string and its number of occurrences in the training documents. Furthermore, each edge represents the link between two consecutive words in the training documents, and the weight of the edge represents the number of occurrences of these consecutive words.

Let us assume that we have a set of topics T, and each topic $t_i \in T$ is represented by the extended graph model $G_i = (V_i, E_i, \mathcal{L}_i, W_{vi}, W_{ei})$, defined as follows:

- V_i and E_i are respectively a set of nodes and a set of edges;
- a labeling function $\mathcal{L}_i : V_i \rightarrow T$ used to assign vertices to the graph;
- a function $W_{vi}: V_i \times T \rightarrow \mathbb{N}$, is assigned for each vertex $v \in V_i$ a weight;
- a function $W_{ei}: E_i \times T \rightarrow \mathbb{N}$, is assigned for each edge $e \in E_i$ a weight.

1) Model construction and learning:

Each topic $t_i \in T$ has a set of labeled documents $D = \{d_j \in D \ / \ 1 \leq j \leq m\}$ reserved for the training and constructing the models.

For each document $d_j \in D$, a list of ordered words W_j is extracted. Subsequently, for every word $w \in W_j$, a new vertex v is assigned to the graph by $\mathcal{L}_i(v) = w$ with a weight equal to 1, only if $v \notin V_i$, else $(v \in V_i)$ the vertex weight is incremented by one.

$$W_{vi}(v, t_i) = \begin{cases} W_{vi}(v, t_i) + 1 & if\ v \in V_i \\ 1 & else \end{cases} \quad (1)$$

A new edge $e \in \{(u, v) \in V_i \times V_i : (\mathcal{L}_i(v) = w_j \wedge \mathcal{L}_i(u) = w_{j+1}) \Rightarrow (w_j, w_{j+1} \in W_{d_i})\}$ is created between the two vertices (v,u) but only if $e \notin E_i$. Else, the edge weight is incremented by one.

$$W_{ei}(e, t_i) = \begin{cases} W_{ei}(e, t_i) + 1 & if\ e \in E_i \\ 1 & else \end{cases} \quad (2)$$

The training documents are passed one by one to construct the graphs, and finally, we obtain *n* graphs at the end of the training (i.e. each graph corresponds to one topic). Figure2 illustrates the graph construction using the following Arabic text (Arabic sentence):

"للمحافظة على الصّحة يجب زيارة طبيب الصّحة أحيانا"

which means in English "to preserve the health it is recommended to often visit the health doctor".

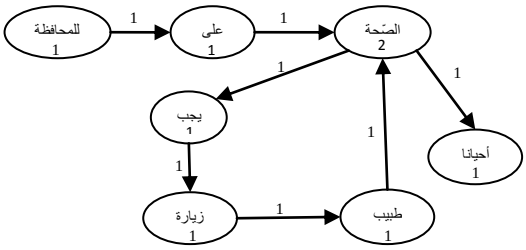

Figure 2. GRAPH MODEL CONSTRUCTED FROM AN ARABIC TEXT

2) Classifying a document:

The unlabeled text is represented as a path of words $\pi = (V, E, \mathcal{L}, v_{start})$, where v_{start} is the starting node. The path structure is similar to the graph structure without node weights, and further, a vertex $v \in V_i$may occur more times in the path.

To compute the similarity between the path and graphs we use the so-called *path matching scores (PM)*. The *PM* function is defined as $PM : T \rightarrow N$ assigning an integer number to any topic $t_i \in T$. Hence, initially, all *PM* scores of the all topics are initialized to zero $PM(t_i) = 0$ for every $t_i \in T$.

To compute the *PM* score, the path is traversed node by node, and we look for the existence of each path node in every graph G_i. If the path node v exists in the graph G_i $(v \in G_i)$, we then add the node weight $W_{vi}(v, t_i)$ of the graph G_i to the corresponding $PM(t_i)$ score using equation (3).

$$PM(t_i) = \begin{cases} PM(t_i) + W_{vi}(v, t_i) & if\ v \in G_i \\ PM(t_i) & else \end{cases} \quad (3)$$

Even, we look for the existence of the path edges in the graph G_i, if the edge e exists in the graph $G_i (e \in G_i)$, we then add the edge weight $W_{ei}(e, t_i)$ of the graph $G_i$to the $PM(t_i)$ score using equation (4).

$$PM(t_i) = \begin{cases} PM(t_i) + W_{ei}(e, t_i) & if\ e \in G_i \\ PM(t_i) & else \end{cases} \quad (4)$$

Finally, after matching the entire path onto the graphs, the text topic is identified by the one having the highest *PM* score.

$$topic = argmax_{t_i \in T}(PM(t_i)) \quad (5)$$

We take an example of the following Arabic text " الطبيب هو من يعطي الدواء", which means in English "only the doctor can prescribe medicaments". After creating the path and matching it in the previous graph, we obtain 1 as a *PM score* (*PM* = 1), because it exists only one path node in the graph (الطبيب).

Figure 3. PATH CONSTRUCTED FROM AN ARABIC TEXT

B. Topic Identification Graph Approach (TIGA1)

We proposed a modified version of the LIGA approach, which we called TIGA1 (Topic Identification Graph Approach). TIGA1 is based on tf-idf weights instead of standard frequencies; hence, it consists in computing tf-idf weights using the previous node weights computed by LIGA. Thus, the weights of the nodes in the LIGA graph representing words frequencies are replaced by the new tf-idf weights, whereas the edges weights are the same as those used in the LIGA approach.

$$tfidf(v, t_i) = W_{vi}(v, t_i) * idf_v \quad (6)$$

Where $W_{vi}(v, t_i)$ is the weight of the node v in the graph G_i, and idf_v is the inverse graph frequency (equivalent to the inverse document frequency) of the node $v$and given by the following equation:

$$idf_v = \log(n / M_v) \quad (7)$$

Where n is the number of topics, and M_v is the number of the graphs containing the node v.

C. Topic Identification Graph Approach (TIGA2)

The second extension of the LIGA approach consists in replacing the nodes and edges weights of the LIGA graph (corresponding to the number of occurrences) by the tf-idf weights. These tf-idf weights are computed using the previous node and edge weights (obtained by the LIGA approach) by the equations 6 and 8, respectively. Thus, we replace the old node and edge weights with their corresponding tf-idf weights.

$$tfidf(e, t_i) = W_{ei}(e, t_i) * idf_e \quad (8)$$

Where $W_{ei}(e, t_i)$ is the weight of the edge e in the graph G_i, and idf_e is the inverse graph frequency (equivalent to the inverse document frequency) of the edge $e$and given by the following equation:

$$idf_e = \log(n / M_e) \quad (9)$$

Where n is the number of topics, and M_e is the number of the graphs containing the edge e.

VI. EXPERIMENTS AND RESULTS

For the evaluation task, we conducted several experiments on the ANTSIX dataset, where 60% of the overall dataset is used in the training step, and 40% is reserved for the test.

In these experiments, we have tested different *PM* functions to identify the topic of the input text (test text) ; *PM1* is based only on the node weights (equation 3), *PM2* is based only on the edge weights (equation 4) and *PM3* is based on the node weights and the edge weights (equations 3 and 4).

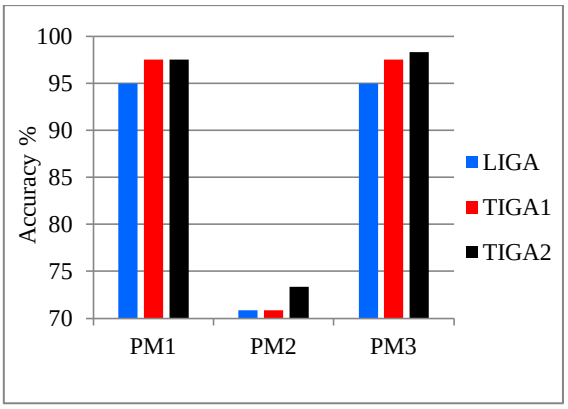

Figure 4. ACCURACIES OBTAINED BY LIGA, TIGA1 AND TIGA2 ON THE ANTSIX DATASET

The accuracy is defined by the following formula:

$$\text{Accuracy} = \frac{number\ of documents\ well\ recognized}{total\ number\ of\ documents} \quad (10)$$

Figure 4 exposes the accuracies of the three approaches LIGA, TIGA1 and TIGA2 using different path-matching functions: *PM*1, *PM*2 and *PM*3. From the figure, we notice that the best accuracy is reported by TIGA2 algorithm, whereas the worst one is that given by LIGA algorithm. This result proves that the two proposed extensions (TIGA1 and TIGA2) have improved the LIGA scores.

Concerning the *path matching score*, the *PM*2 gives the worst scores, comparing to those obtained by *PM*1 and *PM*3, with a difference of 20%. Thus, the *PM*2 seems to be not suitable for topic identification.

TABLE 1. ACCURACIES OF LIGA, TIGA1 AND TIGA2 APPROACHES USING DIFFERENT *PM* FUNCTIONS

	LIGA	**TIGA1**	**TIGA2**	**average**
***PM*1**	95	97,5	97,5	96,67
***PM*2**	70,83	70,83	73,33	71,66
***PM*3**	95	97,5	**98,33**	96,94

The use of *PM*1 brings good results like the original function *PM*3, except in the case of the TIGA2 approach, where the *PM*3 accuracy is slightly greater than the *PM*1 one (about 0.83% of difference). Therefore, in term of optimizing the computation time, it is preferable to use *PM*1 instead of *PM*3.

The use of tf-idf weights (used by TIGA1 and TIGA2 approaches) increases the topic identification accuracy, because topic keywords have high weights, contrariwise the other terms (i.e. not keyword terms).

We also noticed, from the results, that TIGA2 brings better results than the other approaches using *PM2* and *PM3*, because the edge weights are biased by the tf-idf algorithm. Hence, successive words (i.e. edges) that exist in all topics have small weights, contrariwise those belonging to specific topics. In this aspect, we notice that in the Arabic language writing style, there exist a set of successive words commonly used in any text.

VII. CONCLUSION

In this investigation, several topic identification experiments have been conducted and commented in noisy Arabic texts. For that purpose, we constructed an Arabic dataset which we called ANTSIX. This dataset contains a collection of discussion forum texts corresponding to 6 different topics.

We implemented 3 graph approaches: the first one is the LIGA approach; the second and third ones called TIGA1 and

TIGA2 represent modified versions of the LIGA. LIGA uses the word frequencies as weights of the graph nodes and edges, whereas TIGA1 and TIGA2 use the tf-idf as weights instead of the standard word frequencies.

Results show that the proposed approaches: TIGA1 and TIGA2 are more accurate than LIGA for topic identification of noisy Arabic texts (forum texts). Also, the best accuracy is given by the TIGA2 approach (i.e. about 98.33% of accuracy).

As perspectives, we are interested in testing statistical approaches and machine learning classifiers associated to n-grams (as features). On the other hand, we plan to build some fusion techniques between all the developed classifiers to enhance the performances.

REFERENCES

[1] E. Tromp and M. Pechenizkiy, Graph-Based N-gram Language Identification on Short Texts, Proceedings of Benelearn 2011, The Hague, Netherlands, 2011, pp. 27-35.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.