The Islamic University Journal (Series of Natural Studies and Engineering)
Vol. 15,No.1,pp 157-167, **2007,** ISSN 1726-6807, http//www.iugzaza.edu.ps/ara/research/

Arabic Text Classification Using Maximum Entropy

Alaa M. El-Halees
Department of Computer Science, The Islamic University of Gaza
,P.O. Box 108, Gaza, Palestine,
Email: alhalees@ iugaza.edu.ps

Abstract: In organizations, a large amount of information exists in text documents. Therefore, it is important to use text mining to discover knowledge from these unstructured data. Automatic text classification considered as one of important applications in text mining. It is the process of assigning a text document to one or more predefined categories based on their content. This paper focus on classifying Arabic text documents. Arabic language is highly inflectional and derivational language which makes text mining a complex task. In our approach, we first preprocessed data using natural language processing techniques such as *tokenizing, stemming* and *part-of-speech.* Then, we used maximum entropy method to classify Arabic documents. We experimented our approach using real data, then we compared the results with other existing systems.

ملخص: هناك الكثير من المعلومات موجودة بشكل نصي في المؤسسات. لذلك هناك أهمية كبيرة للتنقيب عن البيانات من هذا النوع. التصنيف التلقائي يعتبر أحد التطبيقات المهمة في مجال التنقيب النصي عن البيانات. التصنيف التلقائي هو عملية تصنيف الوثيقة إلى أحد التصنيفات المعرفة مسبقا بناء على محتوى الوثيقة. هذا البحث يركز على التصنيف التلقائي للمستندات المكتوبة باللغة العربية. هناك الكثير من التشكيل والتصريف في اللغة العربية مما يزيد صعوبة التنقيب النصي عن البيانات في المستندات. في هذا البحث نعمل أولا على المعالجة المسبقة للبيانات باستخدام تقنيات معالجة لغات البشر مثل الترميز (*tokenizing*) والتجزير (*Stemming*) والتصريف (*Part-of-Speech*) . ثم استخدم تقنية الانتروبي القصوى (maximum entropy) لتصنيف المستندات. وكذلك تم إجراء تجارب على هذه الطريقة باستخدام مستندات حقيقة وتم مقارنة النتائج مع أنظمة أخرى في تصنيف المستندات العربية.

Keywords: Text Data Mining, Automatic Documents Classification, Maximum Entropy, Arabic Natural Language Processing.

Introduction

Text mining is of growing importance as the volume of unstructured text in web pages, digital libraries and community wide intranets continue to increase. Robb [1] estimated that text documents account about 85% of organizations' knowledge stores. Therefore, more and more organizations need for text classification as important task in text mining that helps in organizing text documents. Automatic text classification (which also known as text categorization or topic spotting) is the process of assigning of a text document to one or more predefined categories based on their content. Automatic text classifications have been used in many applications such as real time sorting of files into folder hierarchies, topic identifications, dynamic task-based interests, automatic meta-data organization, text filtering and documents' organization for databases and web pages [2][3][4]. Many methods have been used for text classification such as naïve bayes [5], support vector machines [6], k-nearest neighbor [7] and maximum entropy [8]. From experiments no single technique consistently outperforms the others [8].

However in Arabic language there is a very limited work in automatic classification. Classifying Arabic text is different than classifying English language because Arabic is highly inflectional and derivational language which makes monophonical analysis a very complex task [9]. Also, in Arabic scripts some of the vowels are represented by diacritics that usually left out in the text which create ambiguity in that text. In addition, Arabic scripts do not use capitalization for proper nouns which are necessary in classifying documents classification [9].

This paper uses maximum entropy to classify Arabic documents. It shows that this approach outperforms the other existing systems.

The rest of the paper is organized as follows: Section 2 summaries related works in documents classifications. Section 3 gives a general description on maximum entropy as a method used to classify Arabic documents. Section 4 proposes system that implements our approach. Section 5 reports our experiments of the proposed method and compare it with other existing systems. Finally we close this paper with a summary and an outlook for future work.

2. Related Work

There are many research in classifying English documents (i.e. [10] have a survey). In addition to English language there are many research in European languages such as German , Italian and Spanish [11] and in Asian languages such as Chinese and Japanese [12]. However, in Arabic language there are only four researches which are: El-Kourdiet. al. [13] use naïve bayes algorithm to automatic Arabic document classification. The average accuracy reported was about 68.78%. Another system is called *Siraj* from *Sakhr*. The system is available at (siraj.sakhr.com) but it has no technical documentation to explain the method used in the system neither the accuracy of the system. The third system proposed by sawaf et. al. [14] who used statistical classification methods such as maximum entropy to classify and cluster News articles. The best classification accuracy they reported was 62.7% with precision of 50% which is a very low precision in this field. In addition, El-Halees [15] described a method based on association rules to classify Arabic documents. The classification accuracy reported was 74.41%.

3. Maximum Entropy Framework for Text Classification

The maximum entropy model estimates probabilities based on the principle of making as few assumptions as possible, other than the constrained imposed. The constraints are derived from training process which express a relationship between the binary features and the outcome [8] [16]. In text classification, maximum entropy is a model which assigns a class *c* of each word *w* based on its document *d* in the training data *D*. Conditional distributed *p(c|d)* is computed as follows [16]:

$$p(c \mid d) = \frac{1}{Z(d)} \exp(\sum_{i} \alpha_i f_i(d,c)) \qquad (1)$$

Where Z(d) in equation (1) is a normalization function which is computed as:

$$Z(d) = \sum_{c} \exp(\sum_{i} \alpha_i f_i(d,c)) \qquad (2)$$

And the parameter α_i in equation (2) must be learned by estimation. It can be estimated by a iterative way using algorithms such as Generalized Iterative Scaling (GIS) [17], Improved Iterative Scaling (IIS) [18], or L-BFGS Algorithm [19]. In our research we tested all of them and found that L-BFGS is the most accurate one.

In equation (2), $f_i\ (d,c)$ is a binary valued feature which makes a prediction about the outcome. In classification the feature presented by each instance to be classified. The type of feature could be Boolean that presents the word in the text, or integer which presents frequency of the word in the text. In this work integer type is used because it gives more information than Boolean. More precisely the feature is formulated as [8]:

$$f_{(w,c')}(d,c) = \begin{cases} 0 & \text{if } c \neq c' \\ \dfrac{N(d,w)}{N(d)} & \text{Otherwise} \end{cases} \qquad (3)$$

Where *N(d,w)* in equation (3) is the number of times word *w* occurs in document *d*, and *N(d)* is the number of words in *d*.

4. System Description

This work constructed a system called ArabCat to implement the proposed method that classifies Arabic documents. The structure of the system is depicted in Figure 4.1. The system consists of the following parts.

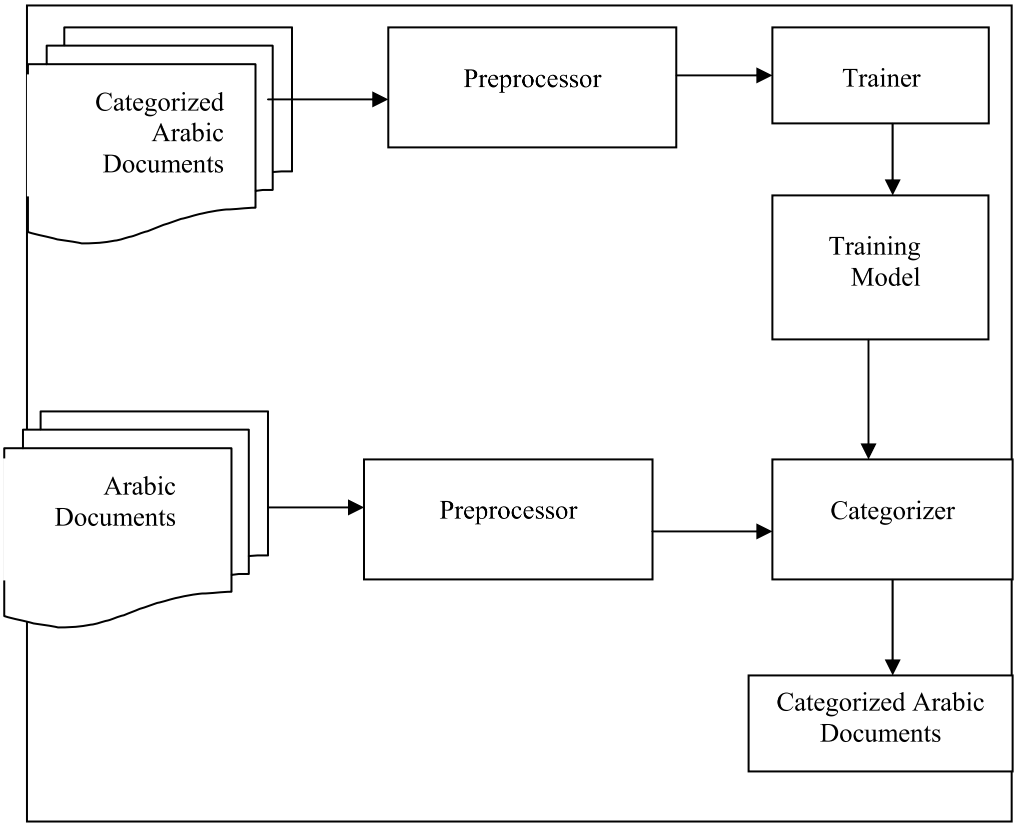

Figure 4.1: ArabCat System Structure

4.1 Corpus

Maximum entropy is a supervised learning technique. Therefore, it needs a training corpus. Our experiments trained the system using Arabic documents collected from the Internet. It mainly collected from Aljazeera Arabic news channel (www.aljazeera.net) which is the largest Arabic site [13]. The documents categorized into sex domains: *politics*, *sports*, *culture and arts*, *science and technology*, *economy* and *health*.

4.2 Preprocessor

Before applying maximum entropy algorithm, for both training and testing data, some preprocessing in the text have been performed. All the experiments are performed after normalizing the text. In *normalization*, the text is converted to UTF-8 encoded and punctuations and non-letters are removed. Also, some Arabic letters are normalized such as أ, إ, and آ are converted to ا, and ى replaced by ي and ة to ه. Then text is parsed by a *parser* and all stopwords are removed. Stopwords are terms that are too frequent in the text. These terms are insignificant. So, removing them reduces the space of the items significantly in the training and testing text. In addition, in Arabic script, some of the vowels are represented by diacritics which usually left out in the texts [9]. If there is any case of ambiguity, then a *tokenizer* is used to generate all possible forms.

Then, the roots are extracted from the text using a *stemmer*. The experiments are performed with and without stemming the training and testing text. Then the results are compared. In addition, some experiments are performed with only part of the text (mainly nouns and proper nouns). Therefore, *part-of-speech* is used to tag the text and nouns and proper nouns are chosen and the rest of the text is neglected.

4.3 Trainer

Trainer part takes the training documents, after the preprocessing, and applies maximum entropy algorithm [16] [20] to trains the data. Outline of the algorithm that implemented in the trainer is given in Figure 4.2. The output of the trainer is a training model which will be used to classify a new document.

4.4 *Categorizer*

The input of this part is an unlabeled Arabic documents after applying the same preprocessor used for training phase. The output is classifying each new document into one of the classes. In this paper we used the classes as *politics*, *sports*, *culture and arts*, *science and technology*, *economy* and *health*.

5.0 Experiments and Results

To experiment our method, we built ArabCat system that classifies Arabic documents. The system built using Java Programming Language and used AraMorph package, which is Arabic Morphology analysis package from http://www.qamus.org/morphology.htm, for the preprocessor part of the

```
Input: Training Data (e.g. set of  Arabic Documents)  $D$
Output: Training model
For each document $d$ in $D$
    For each class $c$
        Compute $U[c]$ as $\exp \sum_{i} \alpha_i f_i(d,c)$
    Next $c$
    Compute $Z$ as $\sum_{c} U[c]$
    For each class $c$
        For each $j$ such as $f_j(d,c) \neq 0$
            $O[j] += f_j \times U[c] / Z$
        Next $j$
    Next $c$
Next $d$
For each indicator $f_j$
    Re-estimate $\alpha_j$ using $O[j]$
Next $j$
```

Figure 4.2: An outline of maximum entropy algorithm that trains the ArabCat system

system. We tested the system using real data collected from many Arabic web sites such as www.elaph.net, www.palestine-info.info and www.islamonlone.net and others.

In our experiments, we computed recall (the percentage of the total documents for the given topic that are correctly classified) and precision (the percentage of predicted document for the given topic that are correctly classified) which are generally accepted ways of measuring system's performance in this field [21], and combined these measures into f-measure. F-measure is a standard statistical measure that is used to measure the performance of a classifier system [22]. The f-measure is an average parameter based on precision and recall [23].

We tested our system and compared its overall performance with others exiting systems. The results are recorded as in Table 5.1. For [15], [13] , [14] we get the number from their published work. For Sakhr's Categorizer we used the same test data we used for ArabCat system. From this table we can notice that ArabCat system has better performance than other systems.

However, Sawaf et. al.[14] have better recall but they had a very low precision. In addition, our system has the best Precision.

System	Recall	Precision	F-Measure
ArabCat System	80.48	80.34	80.41
El-Halees	74.48	74.34	74.41
El-Kourdi et. al.	71.96	67.88	67.83
Sakhr's Categorizer	73.78	47.35	57.68
Sawaf et. al	84.2	50	62.7

Table 5.1: Performance of ArabCat systems compared with other systems.

When we recorded the performance for each class of the six categories, we got the results as in Table 5.2. From the table we can see that the best performance is recorded in *sports* domain that because *sports* has limited space comparing to other domains. Also, it shows that *politics* has lowest performance may be that also because politics have a large space domain.

Category	Recall	Precision	F-Measure
Politics	72.22	72.22	72.22
Sports	100.00	100.00	100.00
Culture and arts	93.33	77.78	84.85
science and technology	70.59	80.00	75.00
Economy	68.18	83.33	75.00
Health	78.57	68.75	73.33

Table 5.2: The Performance of ArabCat in each of the domain category**.**

To see the effectiveness of the preprocess, we conducted experiments in each preprocess stage and we recorded the results in Table 5.3. We first tested the system without any preprocessing we got overall f-measure of 68.1%. When we did only normalization in the preprocessing, the performance increased to 70.25%. When we did normalization and tokenizing the performance increased to 71.20%. Finally, when we did the experiment only in nouns and proper nouns (by using part-of speech) and neglected the other words in the text, the performance increased to 80.41% which the largest increase.

Method	Recall	Precision	F-Measure
Without Preprocessing	67.89	68.46	68.13
With Normalization	70.49	70.00	70.25
With Normalization + tokenizer	71.24	71.16	71.20
With Normalization + tokenizer + POS	80.48	80.34	80.41

Table 5.3: The effectiveness of preprocessing stage in the performance of ArabCat system

Besides this increase of performance, Table 5.4 shows that using POS to choose nouns and proper nouns also reduces the training size by 65.5% and decrease the training time by about 50%.
From this we can conclude that the using nature language processing techniques increase significantly the performance of the system especially in Arabic language which needs special treatment.

	All text	**Only nouns and Proper Nouns**
Performance	71.20%	80.41%
Training Size	609 KB	210 KB
Training Time	1.45 Sec	0.73 Sec

Table 5.4: The effectiveness of POS on Performance, Training Size and Training Time

In maximum entropy the parameter α_i must be learned by estimation. There are three algorithms in research which do estimates this parameters which are: Generalized Iterative Scaling (GIS) [17], Improved Iterative Scaling (IIS) [18], or L-BFGS Algorithm[19]. In our work we conduct our experiments in the three algorithms, and in each one we used different numbers of iterations. Figure 5.5 shows the performance of each algorithm with the number of iterations. From the figure we found that L-BFGS is the most accurate one and the performance stable beginning at 50 as number of iterations.

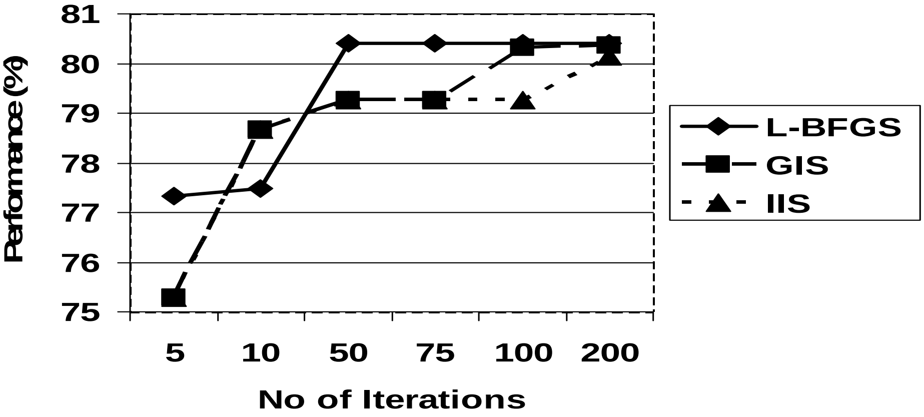

Figure 5.5: The estimation of parameter α_i in Maximum Entropy.

Finally, we tested the scalability of the system by training the system in different data sizes and recorded the performance in each size. From Figure 5.6 we found that when the training size increases it almost linearly increases the system performance.

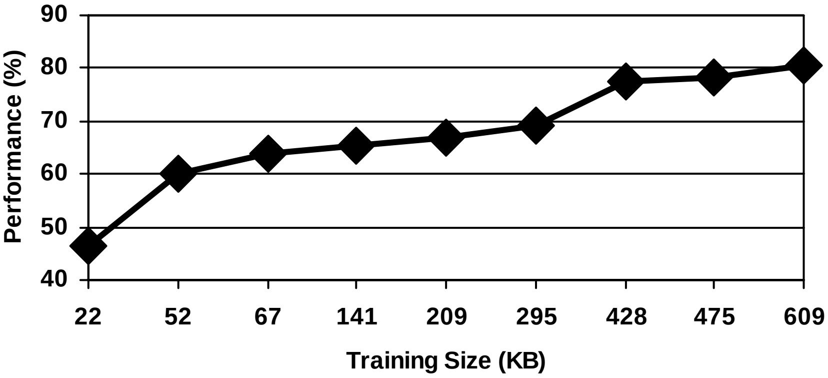

Figure 5.6: Scalability measures of the ArabCat System.

6.0 Summary

This work used maximum entropy method to classify Arabic documents. It proposed a method and system that focuses in natural language processing techniques to preprocess the documents before applying the method. The paper proposed a structure for the system which has five parts. Then it built the parts in a system called ArabCat. It used real data sets to test the systems. The experiments found that the system has a better performance than all the systems built for Arabic classification task. Also, it found that using the preprocessing techniques increases the f-measure from 68.13% to 80.41%. And it decreased the training data sets size hence deceases the training time. The experiments also tested the system in all maximum entropy's parameter α_i estimations algorithms and it found the best algorithm performance was L-BFGS algorithm with 50 as number of iteration. Finally, the scalability of the system was measured and it found that the system is almost scalable linearly with the size of training data. For future work, the system may test more domains and hierarchies of domain (e.g. subdomains). It may also train and test for huge number of documents. Also, more advanced techniques could be used to reduce the dimensionality of the text.

References

[1] Robb, D., *Text Mining Tools Take on Unstructured Information*. Computerworld, 21 June (2004).

[2] Sauban, M. , Pfahringer, B, *Text Categorization Using Document Profiling*. Principles of Data Mining and Knowledge Discovery. (2003)

[3] Yang, Y., Slattery, S., Ghani, R., *A Study of approaches to hypertext Categorization.* Journal of Intelligent Information Systems Vol. 18 p. 219-214 (2002).

[4] Sebastiani, F.**,** *Machine learning in automated text categorization*, ACM Computing Surveys (CSUR) Vol. 34 , Issue 1. P:1 - 47 (2002)

[5] Lewis, D., *Naïve (Bayes) at forty: The Independent Assumption in Information Retrieval*. In Machine Learning: ECML-98, 10th European Conference on Machine Learning. p 4-15 (1998).

[6] Joachims, T., *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. In Proceedings of ECML-98, 10th European Conference on Machine Learning. Pages 137-142. (1998).

[7] Yang, Y., *An Evaluation of Statistical Approaches to Text Categorization* . Journal of Information Retrieval . (1999).

[8] Nigam, K., Lafferty, J., McCallum, A., *Using Maximum Entropy for Text Classification*. In *IJCAI-99* Workshop on Machine Learning for Information Filtering, pp. 61-67. (1999).

[9] Hammo, B., Abu-Salem, H., Lytinen, S., Evens, M., *QARAB: A Question Answering System to Support the Arabic Language*. Workshop on Computational Approaches to Semitic Languages. ACL 2002, Philadelphia, PA, July. p 55-65 (2002).

[10] Kjersti ,A. Eikvil, L., *Text categorization - A survey.* Report No. 941, June, 19 (1999).

[11] Ciravegna, F., Gilardoni, L., Lavelli, A., Ferraro, M., Mana , N., Mazza, S., Matiasek, J., Black, W., Rinaldi, F., *Flexible Text Classification for Financial Applications: the FACILE System*. In Proceedings of PAIS-2000, Prestigious Applications of Intelligent Systems sub-conference of ECAI2000. (2000)

[12] Peng, F., Huang, X., Schuurmans, D., Wang, S., *Text Classification in Asian Languages without Word Segmentation*. In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages (IRAL 2003), Association for Computational Linguistics, July 7, Sapporo, Japan. (2003).

[13] El-Kourdi, M., Bensaid, A., Rachidi, T., *Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm*. 20th International Conference on Computational Linguistics . August 28th. Geneva (2004).

[14] Sawaf, H., Zaplo, J., Ney, H., *Statistical Classification Methods for Arabic News Articles*. Arabic Natural Language Processing, Workshop on the ACL'2001. Toulouse, France, July (2001).

[15] El-Halees A., *Mining Arabic Association Rules for Text Classification* In the proceedings of the first international conference on Mathematical Sciences. Al-Azhar University of Gaza, Palestine, 15 -17 (2006). *To be appear.*

[16] Berger, A., Pietra, D., Pierta D,. *A Maximum Entropy Approach to Natural Language Processing*. Computational Linguistics, Vol., 22. p. 39-71 (1996).

[17] Darroch, J., Ratcliff, D., *Generalized Iterative Scaling for Long_Linear Model* . Annals of Mathematical Statistics, 43 (5): 1470 - 1480 (1972).

[18] Berger, A., *The Improved Iterative Scaling Algorithm: A Gentle Introduction*. Technical report**.** (1997)

[19] Malouf, R., *A comparison of algorithms for maximum entropy parameter estimation*. In Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002). P 49-55. (2002).

[22] Apte , Damerau, Weiss, *Towards Language Independent Automated Learning of Text Categorization Models* . Research and Development in Information Retrieval P 23-30. (1994)

[23] Chinchor, N., *Named Entity task definition*, In Proceedings of the Seventh Message Understanding Conference. (1998).