Automatic Arabic Text Classification

S. Al-Harbi, A. Almuhareb, A. Al-Thubaity,
M. S. Khorsheed, A. Al-Rajeh

King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

Abstract

Automated document classification is an important text mining task especially with the rapid growth of the number of online documents present in Arabic language. Text classification aims to automatically assign the text to a predefined category based on linguistic features. Such a process has different useful applications including, but not restricted to, e-mail spam detection, web page content filtering, and automatic message routing. This paper presents the results of experiments on document classification achieved on seven different Arabic corpora using statistical methodology. The performance of two popular classification algorithms in classifying the aforementioned corpora has been evaluated.

Keywords: text mining, classification, feature selection, Arabic text classification.

1. Introduction

Most previous studies of data mining have focused on structured data. However, in reality substantial portions of the available information are stored in text databases that consist of large collections of documents from various sources such as books, articles, research papers, e-mail messages and web pages. With the existence of a tremendous number of these documents, it is tedious yet essential to be able to automatically organize the documents into classes to facilitate document retrieval and subsequent analysis. Automatic text classification is the process of assigning a text document to one or more predefined categories based on its content.

There are several research projects investigating and exploring the techniques in classifying English documents (Aas and Eikvil 1999). In addition to English language there are many studies in European languages such as French, German, Spanish (Ciravegna et al. 2000) and in Asian languages such as Chinese and Japanese (Peng et al. 2003). However, in Arabic language there is little ongoing research in automatic Arabic document classification.

The three main consecutive phases in building a classification system are as follows:

1. Compile the text documents in corpora and label them.
2. Select a set of features to represent defined classes.
3. Chosen classification algorithms must be trained and tested using the compiled corpora in the first stage.

This paper attempts to attain a better understanding and elaboration of Arabic text classification techniques by using the aforementioned stages. The remainder of the paper is organized as follows. Section 2 discusses previous work in Arabic text classification. In section 3, we present the corpora design and compilation. In section 4, we present our feature extraction tool, weighting techniques and stop lists. Section 5 presents an analytical and

experimental evaluation of the well-known classifications algorithms and section 6 concludes with a summary and some direction for future research.

2. Related work in Arabic text classification

Arabic is the mother language of more than 300 million people (El-Kourdi et al. 2004). Unlike Latin-based alphabets, the orientation of writing in Arabic is from right to left; the Arabic alphabet consists of 28 letters. Arabic words have two genders, feminine and masculine; three numbers, singular, dual, and plural; and three grammatical cases, nominative, accusative, and genitive. A noun has the nominative case when it is subject; accusative when it is the object of a verb; and the genitive when it is the object of a preposition. Words are classified into three main parts of speech, nouns (including adjectives and adverbs), verbs, and particles.

Much of the work in text classification treats documents as a bag-of-words with the text represented as a vector of a weighted frequency for each of the distinct words or tokens. Such a simplified representation of text has been shown to be quite effective for a number of applications (Diederich et al. 2003; Sebastiani 2002). There are several attempts to enhance text representation using concepts or multi-word terms (Mesleh 2007).

El-Kourdi et al., used Naïve Bayes algorithm to automatically classify Arabic documents. The average accuracy reported was about 68.78% (El-Kourdi et al. 2004). Sawaf et al. 2001 used statistical classification methods such as maximum entropy to classify and cluster news articles. The best classification accuracy they reported was 62.7%. In addition, El-Halees (2006) described a method based on association rules to classify Arabic documents. The classification accuracy reported was 74.41%. Al-Fedaghi and Al-Anzi's algorithm tries to find the root of the word by matching the word with all possible patterns with all possible affixes attached to it (Duwairi 2005).

Al-Shalabi et al's (1998) morphology system uses different algorithms to find the roots and pattern. This algorithm removes the longest possible prefix, and then extracts the root by checking the first five letters of the word. This algorithm is based on an assumption that the root must appear in the first five letters of the word. Khoja has developed an algorithm that removes prefixes and suffixes, all the time checking that it's not removing part of the root and then matches the remaining word against the patterns of the same length to extract the root (El-Kourdi et al. 2004; Larkey and Connell 2001).

3. Corpora

A representative corpus of labeled texts must be assembled. Each text in the data set must be assigned to one of the defined classes. Different training data sets are available for text classification in English. Reuters-21450 and Reuters-810000 collections of news stories are popular and typical examples. The Linguistic Data Consortium (LDC) provides two Arabic corpora, the Arabic NEWSWIRE and Arabic Gigaword corpus. Both corpora contain newswire stories. One of the aims of this paper is to scientifically compile representative training datasets for Arabic text classification that cover different text genres which can be used in the future as a benchmark. Therefore, seven different datasets were compiled covering different genres and subject domains.

The assembled corpus comprised 17,658 texts with more than 11,500,000 words. Such diversity in genre, class and text length will provide a clear insight into the classification

algorithms' performance and their ability to classify Arabic texts. The internet was used to compile the dataset. Saudi Press Agency (SPA) provides its newswires in six classifications which were used to label the texts collected from the SPA official web site.

The classification provided for each article or news story in Saudi newspapers was used to label the texts in Saudi Newspapers (SNP) corpus. The classifications used by Arabic internet portals were used to classify WEB and Discussion Forums datasets. For the Writers corpus, the writer's name was used to label his/her writings. Islamic Topics and Arabic Poems were collected from well organized web and trusted web sites which provide a classification for each text. Table 1 illustrates the genres, number of texts and classes for each dataset.

Genre	No. of Text	Classes
Saudi Press Agency (SPA)	1,526	Cultural News, Sports News, Social News, Economic News, Political News, General News
Saudi News Papers (SNP)	4,842	Cultural News, Sports News, Social News, Economic News, Political News, General News, IT News
WEB Sites	2,170	IT, Economics, Religion, NEWS Medical, Cultural, Scientific
Writers	821	Ten writers
Discussion Forums	4,107	IT, Economics, Religion, NEWS Medical, Cultural, Scientific
Islamic Topics	2,243	Hadeeth, Aqeedah, Lughah, Tafseer, Feqah
Arabic Poems	1,949	Hekmah, Retha, Ghazal, Madeh, Heja, Wasf
Total	17,658	

Table 1: The number of text in each dataset.

As part of building the compiled corpora, we implemented a tool for Arabic Text Classification (ATC Tool) in order to accomplish feature extraction and selection. This tool is able to perform the following main functions:

1. Automatically divide the dataset into two partitions - training and testing - according to the user input of training and testing size.
2. Extract the lexical features (single word) and generate the feature frequency profile for both the training set and testing set with options to explore the profile for each class and each file.
3. Calculate the importance of each feature locally (for each class) based on Chi Square.
4. Generate training and testing matrices.

Also, the user has the option to exclude stop words from the texts or to generate the frequency profile of a certain list of words. Stop words frequently occur in all corpora with no any added value. Table 2 shows examples of stop words.

Stop word (Arabic)	English
في	in
على	on
أين	where

Table2: The Example of stop words.

4. Feature selection and weighting

Classification algorithms cannot deal directly with texts. Instead, these texts are represented as a vector with *m* elements where *m* denotes the number of features which are mostly the text words. This kind of text representation typically leads to high dimension input space which normally affects the efficiency of classification algorithms. Several methods are used to reduce the input space by choosing a subset of features that may lead to better classification. Chi-Squared statistics (X^2) (Alexandrov et al. 2001; Dunham 2003) for instance is one of the most used metrics for feature selection (Forman 2003). The Chi-Squared can be used as a (i) goodness-of-fit test between a set of data and a particular statistical distribution, or (ii) test for independence or for association between two factors or variables. The Chi-Squared statistics formula is related to information-theoretic feature selection functions which 'try to capture the intuition that the best terms for the class *c* are the ones distributed most differently in the sets of positive and negative examples of *c*' (Mesleh 2007; Sebastiani 2002).

The Chi-Squared statistic is given by the following formula:

$$X^2 = \sum \frac{(O-E)^2}{E} \quad (1)$$

Where O equals Observed value and E equals expected value and can be calculated using observed values.

The Chi-Squared statistic (X^2) computes the dependency between two factors the term *t* and the class *c* and considered as test with one degree of freedom. The higher the value of Chi-Squared, the higher the dependency or association between the term and the class. For feature selection problems in text classification, the Chi-Squared statistic is only used to rank features according to their usefulness and is not used to judge the statistical dependence of the term *t* and class *c*.

To calculate X^2 for a term *t* and a particular class *c*, the contingency table (see Table 3) of a term *t* and the class *c* can be used to illustrate the idea.

	c	Not *c*	Total
t	A	B	A+B
Not *t*	C	D	C+D
Total	A+C	B+D	N

Table 3: The contingency table of t and c

The expected value for each cell can be calculated using the observed values given in the contingency table (Table 2) where E = (row total X column total)/grand total. Equation 2 illustrates Chi-Squared statistics using Table 2.

$$X^2(t,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (2)$$

Where N=Total number of documents in the corpus.

A= Number of documents in class *c* that contain the term *t.*

B= Number of documents that contain the term *t* in other classes.

C= Number of documents in class *c* that does not contain the term *t*.

D= Number of documents that does not contain the term *t* in other classes.

5. Experiments and results

The goal of this experiment is to evaluate the performance of two popular classification algorithms (SVM and C5.0) on classifying Arabic text using the seven Arabic corpora described in Section 3. Two data mining software were used here: RapidMiner[1] and Clementine[2]. The RapidMiner open source software was used to provide an implementation for the SVM algorithm and Clementine for the C5.0 decision tree algorithm.

The classification method used in this experiment is based on preliminary experiments conducted earlier and has shown satisfactory results. The method involves applying the Chi-Squared statistic (as described in Section 4) to select the top 30 terms of each class in the training dataset. Chi-Squared is applied on document frequency instead of term frequency. The training and testing matrices are formatted using a Boolean representation that is to set the value 1 for terms that exist in the document and the value 0 for terms that do not. The dataset split size is 70% for training and 30% for testing in each corpus.

The results of this experiment are shown in Table 4 using the accuracy measure. Accuracy is computed by dividing the number of the correctly classified document by the total number of documents in the testing dataset. The overall results for this experiment are promising compared to the reported work on Arabic text classification, even though the results are not directly comparable due to using different datasets, methods, and measures.

Classifier	Writers	SNP	Poems	Islamic	SPA	Web	Forums	Average
SVM	75.61%	72.73%	**36.42%**	**86.42%**	73.25%	68.67%	67.45%	**68.65%**
C5.0	86.43%	79.49%	**49.15%**	**92.12%**	79.81%	81.79%	80.13%	**78.42%**

Table 4: SVM and C5.0 classification accuracy using seven Arabic corpora.

The C5.0 algorithm outperformed the SVM algorithm by about 10%; the SVM average accuracy is 68.65%, while the average accuracy for the C5.0 is 78.42%. A drawback of the C5.0 algorithm is that it is a black box algorithm that is only commercially available.

[1] RapidMiner is freely downloaded at: http://rapid-i.com/ .

[2] Clementine is a data mining software from SPSS Inc.

Apart from the Arabic Poems corpus, the results associated with all the other six corpora range from good to excellent. Removing the Poems corpus from the set improves the average accuracy to 74.02% and 83.30% for SVM and C5.0 respectively. The poor performance of this corpus is probably to do with how well poems are normally written (at least in Arabic). A conceivable requirement for writing an excellent poem is to use a rich vocabulary that makes the poem distinctive from all other poems. This aspect of poem writing complicates the task of the classification algorithm in identifying plausible patterns.

The corpus with the most accurate result is the Islamic Topics corpus; accuracy of 92.12% using C5.0, and accuracy of 86.42% using SVM. This excellent result may indicate that the classes in this corpus are well defined and that there are a number of distinctive terms associated with each class in the corpus. Another factor that may lead to this high performance is the fact that this corpus is the one with the least number of classes (five classes only). In this experiment, accuracy does not correlate either with the number of classes nor with the size of the corpus in terms of the number of samples.

Another notable result that was also reported by others is that accuracy varies among classes. For example, in the SPA corpus, the *Sport* class has a neat classification accuracy of 96.88%, while the *General* class has a noticeably poor accuracy of 45.37%. Other classes have accuracies of between 73.33% and 82.19%.

6. Conclusions

This paper presented the results of classifying Arabic text documents on seven different Arabic corpora by using a recognized statistics technique. A tool was implemented for feature extraction and selection and the performance of two popular classification algorithms (SVM and C5.0) has been evaluated on classifying Arabic corpora. C5.0 classifier, in general, gives better accuracy. In our future work, we plan to introduce other classification algorithms in addition to the ones used here. Additionally, we plan to utilize other feature selection and weighting methods and compare them with the methods already used. Finally, we will continue to investigate the effect of each factor on the accuracy of the classification of Arabic text.

References

Aas K. and Eikvil L. (1999). *Text Categorisation: A survey*. Technical report, Norwegian Computing Center.

Alexandrov M., Gelbukh A. and Lozovo. (2001). *Chi-square Classifier for Document Categorization*. 2nd International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City.

Al-Shalabi R. and Evens M. (1998). *A Computational Morphology System for Arabic. Workshop on Semitic Language Processing.* COLING-ACL'98, University of Montreal, Montreal, PQ, Canada. pp. 66-72.

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

Diederich J., Kindermann J. L., Leopold E. and Paaß G. (2003). Authorship attribution with support vector machines. *Applied Intelligence*, 19(1/2): 109-123.

Dunham M. H. (2003). *Data Mining: Introductory and Advanced Topics*. Prentice Hall.

Duwairi R. M. (2005). A Distance-based Classifier for Arabic Text Categorization. In *Proceedings of the International Conference on Data Mining*, Las Vegas USA.

El-Halees A. (2006). Mining Arabic Association Rules for Text Classification. In *Proceedings of the first international conference on Mathematical Sciences*. Al-Azhar University of Gaza, Palestine.

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

Forman G. (2003). An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res*. 3, 1289-1305.

Larkey L. and Connell M. E. (2001). Arabic information retrieval at UMass in TREC-10. In *Proceedings of TREC*, Gaithersburg: NIST.

Mesleh A. A. (2007). Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System. *Journal of Computer Science* 3(6): 430-435.

Peng F., Huang X., Schuurmans D. and Wang S. (2003). 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, Sapporo, Japan.

Sawaf H., Zaplo J. and Ney H. (2001). *Statistical classification methods for Arabic news articles*. Natural Language Processing in ACL2001, Toulouse, France.

Sebastiani F. (2002). Machine learning in automated text categorization**.** *ACM Computing Surveys*, Vol. 34 number 1. pp.1-47.