World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 2, No. 6, 193-196, 2012

Arabic Content Classification System Using statistical Bayes classifier With Words Detection and Correction

Abdullah Mamoun Hattab
Department of Computer science
Middle East University
Amman, Jordan

Abdulameer Khalaf Hussein
Department of Computer science
Middle East University
Amman, Jordan

Abstract— Automatic Arabic content classification is an important text mining task especially with the rapid growth of the number of online Arabic documents. This system is an enhancement of the implemented machine learning classification algorithm by applying detection and correction algorithm of Non-Words in Arabic text. This detection and correction algorithm is built on morphological knowledge in form of consistent root pattern relationships, and some morpho-syntactical knowledge based on affixation and morph-graphic rules to specify the word recognition and non-word correction process. Many researchers had been focused on Arabic content classification from only morphological view such as word's root and stemming techniques (prefixes and suffixes) which showed variant results. In this work, consider classification from a very different way which is the syntactical approach. This paper presents the results of experiments on document classification achieved on ten different Arabic domains (Economy, History, Family studies, Islamic, Sport, Health, Law, Stories, astronomy and Food articles) using statistical methodology. The performance of this classification system showed encouraging results compared with other existing systems.

Keywords- text mining; classification; Arabic text classification; Arabic language processing.

I. INTRODUCTION

Text categorization (TC – also known as text classification, or topic spotting) is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set. This task falls at the crossroads of information retrieval (IR) and machine learning (ML). TC has witnessed a booming interest in the last ten years from researchers and developers [1].

In the last ten years, content-based document management tasks had gained a prominent status in the information system field, due to the increased availability of documents in digital form and the ensuring need to access them in flexible ways [2].

The goal of text categorization is the classification of documents into a fixed number of predefined categories. Each document can be in multiple, exactly one, or no category at all [3].

The tremendous growth of available Arabic text documents on the Web and databases had posed a major challenge for researchers to find better ways to deal with such huge amount of information in order to enable search engines and information retrieval systems to provide relevant information accurately, which had become a crucial task to satisfy the needs of different end users. [4].

There are still few academic papers treating the problem of spell checking and correction in the Arabic computational community. Interest in commercial systems had been focused on the Arabic version of MS-Word. Most of the Arabic spell checkers are concerned with isolated word correction techniques, and are based, in particular on simple morphological analysis considering the keyboard effect for correcting single-error misspellings [5].

II. RELATED WORK

There are many researchers who had been concentrating on the field of Arabic language processing and diacritization.

[4] Addressed the issue of automatic classification or classification of Arabic text documents. They applied text classification to Arabic language text documents using stemming as part of the preprocessing steps. The results of their research had showed that applying text classification without using stemming; the support vector machine (SVM) classifier had achieved the highest classification accuracy using the two test modes with 87.79% and 88.54%. On the other hand, stemming had negatively affected the accuracy, where the SVM accuracy using the two test modes dropped down to 84.49% and 86.35%.

TABLE I: STATISTICS FOR MANHATTAN MEASURE

Category	Recall	Precision
Sports	0.882353	0.6
Economy	0.409091	0.93103448
Technology	0.45	0.209302
Weather	0.5	0.916667

TABLE II: STATISTICS FOR DICE'S MEASURE

Category	Recall	Precision
Sports	0.980392	0.78125
Economy	0.893939	0.951613
Technology	0.45	0.818182
Weather	0.1	1

[6] presented the results of classifying Arabic text documents using the N-gram frequency statistics technique employing a dissimilarity measure called the "Manhattan distance" and Dice's measure of similarity. The Dice measure was used for comparison purposes. Results showed that N-gram text classification using the Dice measure outperforms classification using the Manhattan measure. The results for the tri-gram method using the Dice measure exceed those for the Manhattan measure, reaching its highest recall value of 1 for the weather category, followed by 0.98 for the sports category, and 0.89 for the economy category as illustrated in table I and table II.

[7] Applied the Support Vector Machines (SVM) model in classifying Arabic text documents. The results compared with the other traditional classifiers Bayes classifier, K-Nearest Neighbor classifier and Rocchio classifier. Two experiments were used to test the different classifiers. The first experiment used the training set as the test set, and the second experiment used Leave one testing method. Experimental results performed on a set of 1132 documents, showing that Rocchio classifier gave better results when the size of feature set is small while SVM outperform the other classifiers when the size of the feature set was large enough. Classification rate exceeds 90% when using more than 4000 features. Leave one method led to more realistic results over the use of training set as a test set. Classification accuracy results are illustrated in figure 1.

$$p(c|d_j) = \frac{e^{z_{jc}}.p(c)}{e^{z_{jc}}.p(c) + p(\overline{c})}$$

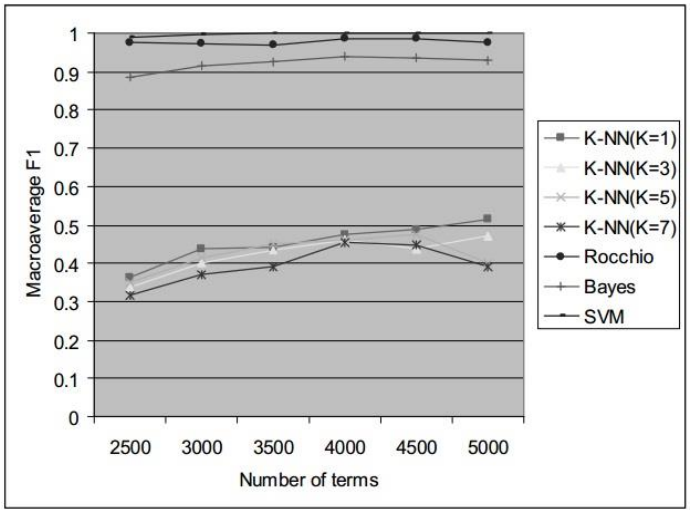

Figure 1: Classifiers performance using Leave One testing

III. CLASSIFICATION METHOD

A naive Bayes classifier is a well-known and highly practical probabilistic classifier, and had been employed in many applications. It assumes that all attributes of the examples are independent of each other given the context of the class, that is, an independent assumption. [8]. Bayesian classification and decision making is based on probability theory and the principle of choosing the most probable or the lowest cost option. In the context of text classification, the probability of class *c* given a document d_j is calculated by Bayes' theorem as follows:

Equation 1:

$$P(c|d_j) = \frac{P(d_j|c)p(c)}{P(j)} = \frac{p(d_j|c)p(c)}{p(d_j|c)p(c) + p(d_j|\bar{c})p(\bar{c})} = \frac{\frac{p(d_j|c)}{p(d_j|\overline{c})}.p(c)}{\frac{p(d_j|c)}{p(d_j|\overline{c})}.p(c) + p(\overline{c})}$$

Equation 2:

$$z_{jc} = \log\frac{p(d_j|c)}{p(d_j|\overline{c})}$$

Equation 3:

Using Equation (3), we can get the posterior probability $p(c|d_j)$ by obtaining z_{jc}, which is a form of log ratio similar to the BIM retrieval model[9]. The log ratio means that the linked independence assumption [10], which explains that the strong independent assumption can be relaxed in the BIM model, is sufficient for the use of naïve Bayes text classification model. With this framework, two representative

naïve Bayes text classification approaches are well introduced in [11] which designated the pure naive Bayes as multivariate Bernoulli model and the unigram language model classifier as multinomial model.

IV. Detection And Correction Method

The Arabic spell checking algorithm was defined in [5]. Their algorithm defined three types of word misspelling and these types are typographic, cognitive and phonetic errors. Errors are categorized into single misspelling errors or multi-error misspellings, based on the analysis presented in [12] approximately 80% of all misspelling errors in Arabic refer to single error misspellings. Many cases for errors can be found in any language electronic documents that include Arabic language too, such cases as:

A. Substitution: According to [12] 41.5% of errors are belonging to substitution's case, i.e. (cut → cur) the replacement of (T-) to (R-). In some cases the substitution leads to non-real words and not only different meanings.
B. Deletion: From all single errors rate 23% are deletion errors [12], i.e. (busy → buy) ,the letter (S) had been missed which gave a misleading meaning word.
C. Insertion: An additional letter inserted by mistake to word, around 15% of errors belong to this case [12], i.e. (play→plaay), and the letter (A) had been duplicated and affected the word meaning.
D. Transposition: Swapping two letters, [12] approximately 4% are transposition, i.e. (read → raed), caused by swapping the character (A) with (E).

Grammatical and semantic errors can be regarded as a major source of producing real word errors. [5] Mentioned that errors cannot be detected without using syntactical, semantic, statistical knowledge or any combination between them to build a detection system.

V. Proposed model

The proposed system consists of three models. The first model is the preprocessing model that deals with corpus data and encoding, punctuation marks, white spaces and empty lines. The second model is the detection and correction model for the corpus, the result of second model is a clean corpus free of unwanted characters and contain lower rate of spelling mistakes. The third model is the classification model that calculates to which domain each text belongs.

A corpus of Arabic documents was built using Arabic news and magazines articles collected from several Arabic newspapers. The corpus consists of text documents covering many categories (Economy, History, Family studies, Islam, Sport, Health, Law, Stories, astronomy and Food articles). Documents sizes were within average from 3KB to 8 KB. All documents are subjected through the text preprocessing steps. This was necessary due to the variations in the way text can be represented in Arabic.

This part of the proposed system, the detection and correction model, classify words into a non-words or a misspelling if the morphological analysis, dictionary look up and the sub-sequential compositional process if it fails to find a model for that word within the defined knowledge base. This process includes extracting a valid root within a consistent root pattern relationship or a stem from a non-derivative word form.

The preprocessing model in the proposed system performed for documents to be cleaned from non-recognized characters and convert these documents to be UTF-8 encoding. Preprocessing consists of the following steps:

1) Convert text files to UTF-8 encoding.
2) Remove punctuation marks, diacritics, non-letters, stop words. The definitions of these were obtained from the Khoja stemmer [13].

Classification model the final part of the proposed system. Training corpus will go through the same procedures as documents to be classified. Each selected document to be a part of the training classes will be preprocessed as explained in the above procedures. Then the N-gram profile will be generated. Generating the N-gram profile consists of the following steps:

1) Splitting the text into tokens consisting only of letters. All digits are removed.
2) Computing all possible N-grams, for N=3 (Tri-grams).
3) Computing the frequency of occurrence of each N-gram.
4) Sorting the N-grams according to their frequencies from most frequent to least frequent. Discard the frequencies.
5) This gives us the N-gram profile for a document. For training class documents, the N-gram profiles were saved in text files.

The N-gram profile of each text document (document profile) is compared against the profiles of all documents in the training classes (class profile) in terms of similarity. Two measures are used. The first measure is a distance or dissimilarity measure, called the Manhattan distance. It calculates a rank-order statistic for two profiles by measuring the difference in the positions of an N-gram in two different profiles. For each N-gram in the document profile, search must be performed for the N-gram in the class profile and then calculating the difference between their positions. For N-grams that are not found in the class profile, a maximum value is assigned. After that all N-grams in the document profile have been exhausted. The second measure, after all N-grams in the document profile have been exhausted, the sum of the distance measures is computed.

VI. EXPERIMENTS AND RESULTS

The goal of this experiment is to evaluate the performance of the detection and correction method with a popular classification algorithm on classifying Arabic text using Arabic corpora that covers ten domains (Economy, History, Family studies, Islam, Sport, Health, Law, Stories, astronomy and Food articles). Two run applied for the proposed system, first run without the detection and correction method and the second is with misspelling detection and correction. The same data will be used in both experiments.

The results of these experiments are shown in Table III and IV using the accuracy measure. Accuracy is computed by dividing the number of the correctly classified document and the total number of documents in the testing dataset. The overall results for these experiments are very promising compared to the reported work on Arabic text classification.

TABLE III: EXPERIMENTS RESULTS OF THE PROPOSED SYSTEM

Classifier	Economy	History	Family	Islam	Sport	Health
Only Bayes	66.2	76.4	84.1	86.4	71.9	64
Bayes With D&C	74.6	80.3	89.6	90.2	75.1	69.6

TABLE IV: EXPERIMENTS RESULTS OF THE PROPOSED SYSTEM

Classifier	Law	Stories	astronomy	Food	Average
Only Bayes	34.8	67	52.4	65.3	66.85
With D&C	45.8	72.9	53.9	66	71.77

The detection and correction algorithm outperformed the Bayes algorithm by about 10%, without checking misspelling errors accuracy is 68.85%, while the average accuracy for the classification system with misspellings detection and correction is 71.77%.

VII. CONCLUSION

The proposed system studied the problem of Arabic content classification and the techniques used to build full automated Arabic classifier using statistical base method and misspelling detection and correction. Text corpus is collected from the newspapers, websites, articles and books, which cover ten domains Economy, History, Family studies, Islam, Sport, Health, Law, Stories, Astronomy and Food. A tool was implemented to evaluate classifying performance on Arabic corpora using two approaches. First approach is Bayes method with misspelling detection and correction that shows 66.85% an average classification accuracy rate and the second approach is Bayes without misspelling detection and correction 71.77% classification rate. The classifier that uses detection and correction gives better accuracy in all domains, with a variety in increased percentage for each domain. After all, increasing of 4.92% can be more improved by using other classification algorithms that based on phrase based and machine learning in addition to the one used in the proposed system. Additionally, more enhancements can be done on the misspelling detection and correction algorithm by complementing its knowledge base rules and morphological analyzer.

REFERENCES

[1] Fabrizio Sebastiani. Text categorization. In Alessandro Zanasi (ed.), Text Mining and its Applications, WIT Press, Southampton, UK, 2005, pp. 109-129

[2] Fabrizio Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1-47, 2002.

[3] T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning (ECML), Springer, 1998.

[4] Wahbeh A., Al-Kabi M., Al-Radaideh Qasem A., Al-Shawakfa E., and AlSmadi I., (2011), The Effect of Stemming on Arabic Text Categorization: An Empirical Study, "International Journal of Information Retrieval Research (IJIRR)", IGI Publisher, 1(3): pp. 54-70.

[5] Bassam Haddad abd M.Yaseen,"Detection and Correction of Non-Words in Arabic: A Hybrid Approach, " International Journal of Computer Processing of Oriental Languages, IJCPOL,Vol. Vol. 20, , No.Number 4, World Scientific Publishing , New Jersey, London, Singapore, Beijing, Shanghai, Hong Kong, Taipei, Chennai, 2007

[6] Laila Khreisat: Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study. DMIN 2006: 78-82

[7] Tarek Fouad Gharib, Mena Badieh Habib, and Zaki Taha Fayed, "Arabic Text Classification Using Support Vector Machines", International Journal of Computers and Their Applications, VOLUME 16, NO. 4, December 2009.

[8] Sang-Bum Kim, Hee-Cheol Seo, Hae-Chang Rim: Poisson naive Bayes for text classification with feature weighting. IRAL 2003: 33-40, 2003.

[9] Karen Sparck Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments - part 1. Information Processing and Management, 36(6):779–808.

[10] William S. Cooper, Fredric C. Gey, and Daniel P. Dabney.1992. Probabilistic retrieval based on staged logistic regression. Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval, pages 198–210.

[11] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. Machine Learning, 39(2/3):103–134.

[12] Ben Hamadou, A.,"The phases of computational analysis of Arabic towards detecting and correcting of errors," Second Conference for Arabization of Computers, in Arabic, 1994.

[13] Khoja, S. and Garside, R. Stemming Arabic Text. Computing Department, Lancaster University, Lancaster, U.K. ttp://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps, 1999.