Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study

Laila Khreisat
Dept. of Computer Science, Math and Physics
Fairleigh Dickinson University
285 Madison Ave, Madison NJ 07940
Khreisat@fdu.edu

Abstract- This paper presents 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 show that N-gram text classification using the Dice measure outperforms classification using the Manhattan measure.

Keywords: N-gram, classification, categorization, Arabic.

I. INTRODUCTION

The rapid growth of the Internet has increased the number of online documents available. This has led to the development of automated text and document classification systems that are capable of automatically organizing and classifying documents. Text classification (or categorization) is the process of structuring a set of documents according to a group structure that is known in advance. There are several different methods for text classification, including statistical-based algorithms, Bayesian classification, distance-based algorithms, k-nearest neighbors, decision tree-based methods [4] to name a few .

Text classification techniques are used in many applications, including e-mail filtering, mail routing, spam filtering, news monitoring, sorting through digitized paper archives, automated indexing of scientific articles, classification of news stories and searching for interesting information on the WWW.

The majority of these systems are designed to handle documents written in the English language, and therefore are not applicable to documents written in the Arabic language.

Developing text classification systems for Arabic documents is a challenging task due to the complex and rich nature of the Arabic language. The Arabic language consists of 28 letters. The language is written from right to left. It has very complex morphology, and the majority of words have a tri-letter root. The rest have either a quad-letter root, penta-letter root or hexa-letter root.

Previous work on Arabic text classification has used distance-based algorithms [5], Learning algorithms [10], and Bayesian classification methods [6] in developing automated text classification systems. Specifically, [8] used N-grams for searching Arabic text documents. They investigated di-grams and tri-grams. No stemming was performed. They concluded that the N-gram technique is not an efficient approach to corpus-based Arabic word conflation. [9] used tri-grams for indexing Arabic documents without any prior stemming. The work of [11] uses N-grams with and without stemming for text searching. Their results indicate that the use of tri-grams combined with stemming improved the performance of search retrieval, however, it was not statistically significant.

In this paper the behavior of the N-Gram Frequency Statistics technique for classifying Arabic text documents is studied. The technique employs a dissimilarity measure called the "Manhattan distance", and Dice's measure of similarity, for the purposes of classification. The Dice measure was used for comparison purposes. Results show that N-gram text classification using the Dice measure gives better classification results compared to the Manhattan measure.

A corpus of Arabic text documents was collected from online Arabic newspapers. 40% of the corpus was used as training classes and the remaining 60% of the corpus was used for classification. All documents, whether training documents or documents to be classified went through a preprocessing phase removing punctuation marks, stop words, diacritics, and non letters. For the training documents, the N-gram (N=3) frequency profile was generated for each document and saved in text files. Then for each document to be classified, the N-gram frequency profile was generated and compared against the N-gram frequency profiles of all the training classes. The Manhattan and Dice measures were computed.

Using the Manhattan measure, the category to which a document belongs is the one with the smallest Manhattan distance, and using the Dice measure, the category is the one with the largest Dice measure. The classification results using these two measures were compared in terms of recall and precision.

The rest of the paper is organized as follows: in section 2 the concept of N-grams is presented, section 3 describes the text preprocessing phase and section 4 gives detailed description of the classification procedure. Section 5 presents the classification results.

II. N-GRAMS

An N-gram [3] is an N-character slice of a string. The N-gram method is language independent and works well in the case of noisy-text (text that contains typographical errors). We used tri-grams for text classification. The tri-grams of a string or token is a set of continuous 3-letter slices of the string. For example, the tri-grams for the word المودعين are: عين, دعي ,ودع , مود لمو, ا لم . In general, a word of length w has w-2 tri-grams. According to **Zipf's law** [12] :

> *"The nth most common word in a human language text occurs with a frequency inversely proportional to n"*

This has the implication that documents belonging to the same class or category will have similar N-gram frequency distributions.

Figure 1 shows the Tri-gram frequency distribution for a text document belonging to the sports category from our corpus. It clearly shows that the frequencies of the most common Tri-grams are inversely proportional to their rank.

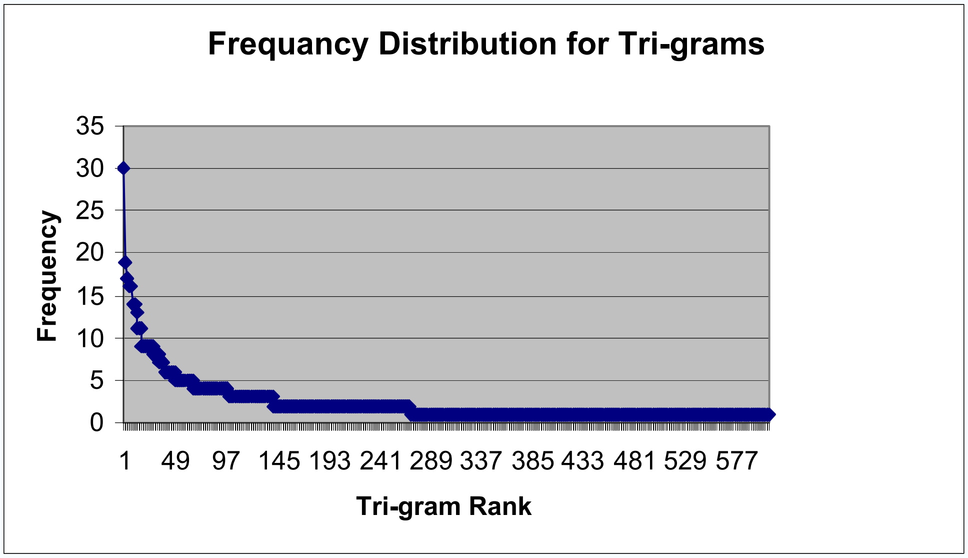

Figure 1. Frequency Distribution of Tri-grams

III. TEXT PREPROCESSING

All text documents went through a preprocessing stage. This was necessary due to the variations in the way text can be represented in Arabic. The preprocessing was performed for the documents to be classified and the training classes themselves. Preprocessing consisted 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 [7].
3) Replace initial أ, إ آ with ا .
4) Replace final ى followed by ء with ئ.

IV. N-GRAM BASED TEXT CLASSIFICATION

A corpus of Arabic text documents was built using Arabic news articles collected from online websites of several Arabic newspapers. The corpus consisted of text documents covering 4 categories: sports, economy, technology and weather. The technology and weather documents were very small in size ranging from 1 KB to 4 KB. Sports and economy documents were much larger ranging from 2 KB to 15 KB for sports documents and 2 KB to 18 KB for economy documents. The smaller documents constituted about 2% of the total number of documents in the sports and economy category.

All these documents went through the text preprocessing step outlined above in section 3. 40% of the corpus was selected as training classes, and the remaining 60% was used for testing the classification procedure. The documents used for training went through the same procedures as did the documents to be classified. Specifically, each document selected to be part of the training classes, was preprocessed as outlined above in section 3. Then the N-gram profile was generated. Generating the N-gram profile consisted of the following steps:

1) Split the text into tokens consisting only of letters. All digits are removed.
2) Compute all possible N-grams , for N=3 (Tri-grams)
3) Compute the frequency of occurrence of each N-gram.
4) Sort 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.

Each document to be classified, went through the text preprocessing phase, then the N-gram profile was generated as described above. The N-gram profile of each text document (document profile) was compared against the profiles of all documents in the training classes (class profile) in terms of similarity. Specifically, two measures were used. The first measure is a distance or dissimilarity measure, called the “Manhattan distance” [1]. 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 for the N-gram in the class profile and calculate the difference between their positions. For N-grams not found in the class profile, a maximum value is assigned. After all N-grams in the document profile have been exhausted, the sum of the distance measures is computed.

$$\text{Manhattan } (P_i, P_j) = \sum_{h=1}^{k} \left|(P_{ih} - P_{jh})\right|$$

where P_i, P_j represent two N-gram profiles

The class that has the smallest Manhattan distance is chosen as the class for the document being classified.

The second measure used is the Dice measure [1] of similarity $\text{Dice}(P_i, P_j) = \dfrac{2\left|P_i \wedge P_j\right|}{\left|P_i\right| + \left|P_j\right|}$

Where $|P_i|$ is the number of elements (N-grams) in profile P_i.

Using the Dice measure, the class with the largest measure is chosen as the class for the text document being classified.

The results obtained using the Manhattan measure and the Dice measure were compared in terms of precision and recall. Precision and recall are defined in [1] as follows:

$$\text{Precision} = \frac{CC}{TCF}$$

$$\text{Recall} = \frac{CC}{TC}$$

Where,

CC :number of correct categories(classes) found.
TCF : total number of categories found
TC : total number of correct categories

V. RESULTS

To compare the performance of the tri-gram technique using the Manhattan measure , and the Dice measure, the recall and precision values were computed. These values are shown in tables 1 and 2 respectively. The best result for the tri-gram method using the Manhattan measure was achieved for the sports category with a recall value of 0.88, and the worst result was for the economy category with a recall value of 0.409.

Table 1. Recall and precision using Manhattan measure

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 2. Recall and precision using Dice's measure

Statistics for Dice's measure		
category	**Recall**	**Precision**
Sports	0.980392	0.78125
Economy	0.893939	0.951613
Technology	0.45	0.818182
Weather	1	1

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.

The two measures produced equal low recall values for the technology category. The reason for this is attributed to the nature of the newspaper articles covering technological issues. They tend to be very diverse covering a vast range of topics. As a result, the training classes for the technology category did not provide full coverage of all the different topics in the category. Overall, classification using the Dice measure outperformed classification using the Manhattan measure. The Manhattan measure has provided good classification results for English text documents [2]. The poor performance of the measure for Arabic in this study, can be attributed to the nature of the Manhattan measure, and the complex morphological structure of Arabic, which is quite different than the structure for English. Stemming text documents before generating the N-grams may give us comparable results for the two measures.

VI. CONCLUSION

This paper 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.

REFERENCES

[1] R. Baeza-Yates, and B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999.

[2] W. B. Cavnar, and J. M. Trenkle, "N-Gram Based Text Categorization," Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[3] M. Damashek, "Gauging Similarity with n-grams: Language-Independent Categorization of Text," Science 267, pp 843 – 848, 10 February 1995.

[4] M. H. Dunham, Data Mining: Introductory and Advanced Topics. Prentice Hall 2003

[5] R. M. Duwairi, "A Distance-based Classifier for Arabic Text Categorization," In Proceedings of the 2005 International Conference on Data Mining, Las Vegas USA 2005.

[6] M. El-Kourd, A. Bensaid, and T. Rachidi, "Automatic Arabic document categorization based on the Naïve-Bayes Algorithm," Workshop on Computational Approaches to Arabic Script-based Languages, COLING-2004, University of Geneva, Geneva, Switzerland, August 2004.

[7] S. Khoja, Personal communication.

[8] H. S .Mustafa, and Q. Al-Radaideh "Using N-Grams for Arabic Text Searching," Journal of the American

Society for Information Science and Technology, 55(11), pp 1002-1007, 2004.

[9] J. Savoy, and Y. Rasolofo, Report on the TREC-11 Experiment: Arabic, Named Page and Topic Distillation Seraches, TREC-11 2002.

[10] H. Sawaf, J. Zaplo, and H. Ney, "Statistical classification methods for Arabic news articles," Arabic Natural Language Processing in ACL2001, Toulouse France July 2001.

[11] J. Xu, A. Fraser, and R. Weischedel, "Empirical Studies in Strategies for Arabic Retrieval,". SIGIR '02 Tampere Finland, 2002.

[12] G. K. Zipf, "Human Behavior and the Principle of Least Effort, an Introduction to Human Ecology," Addison-Wesley, Reading, Mass., 1949.