Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 5 (3): 580-583, 2005
ISSN 1812-5654

Statistical Classifier of the Holy Quran Verses (Fatiha and Yaseen Chapters)

Mohammed Naji Al-Kabi, Ghassan Kanaan, [1]Riyad Al-Shalabi,
[1]Kahlid M.O. Nahar and Basel Mohammed Bani-Ismail
Department of Computer Information Systems,
[1]Department of Computer Science, Yarmouk University, P.O. Box 566, 21163 Irbid, Jordan

Abstract: Automatic Text Categorization (ATC) refer to the process of building software tools capable of assigning unseen documents to predefined categories or subjects. This study aims to automatically classify the verses (Ayat, sentences) of the Fatiha and Yaseen Surahs (Chapters) in the Quran according to the classifications of Islamic scholars. Our automatic text categorization is based on the traditional linear classification function (score function). A system (classifier) has been designed and implemented to categorize the different verses in each Sura (Chapter). This system fully normalizes the verses in the first stage and then the verses are categorized to classes for which they have highest score. The categorization process in this paper depends heavily on a specialized corpus of the Fatiha and Yaseen Surahs built by the authors. The corpus of the whole Quran has not been built before. To build a comprehensive corpus of the Holy Quran requires much time and efforts. Hence a corpus of the Fatiha and Yaseen Surahs was only built and this corpus will be extended to include all the Quran in later future work. This limitation of the corpus leads to a limitation of the categorization of the system to the Fatiha and Yaseen Surahs only. The accuracy of the system can be improved if a more powerful stemmer and a corpus is used. This study lays the foundation stone of building a full corpus of the Holy Quran and a classifier of different verses, which can be used to prove the unity of the subject and different verse similarities.

Key words: Arabic Text Categorization, Quran text classifier, arabic text classification, data mining, classification, statistical analysis

INTRODUCTION

Automated Text Categorization (ATC) is the task of building software tools capable of classifying text (or hypertext) documents under predefined categories or subject codes. ATC has witnessed a booming interest in recent times, due to the availability of ever larger numbers of text documents in digital form and to the ensuing need to organize them for easier use. The dominant approach is nowadays one of building text classifiers automatically by learning the characteristics of the categories from a training set of pre-classified documents (http://mason.gmu.edu/~kersch/JIIS/Special/Issues/TextCategory.html).

Data mining, is a new technology aims to find patterns in data. Similarly, text mining aims to find patterns in text. Some authors defined it as the analysis of text in order to extract useful information for different applications. Mostly, text is unstructured, formless and relatively difficult to deal with in comparison to the data stored in databases.

Natural language corpora are primary sources of information about language use. They represent a huge linguistic knowledge bank that can be tapped through the use of various data analysis tools to discover trends, patterns or other linguistic phenomena which may be incorporated into other language processing tasks. For example, corpora can support detailed studies of how particular words are used, by providing extensive examples of natural language sentences in context. Information about word frequency, co-occurrence, collocations etc. can be derived from corpora and used to build statistical language models, for word sense disambiguation or speech recognition[1].

In Holy Quran we have what we call a unity of subject. Holy Quran divided into 114 chapters (Surahs) and each chapter consists of a number of verses (Ayat). This study aims to classify any verse to a predefined subjects, since the Quran as book is not classified on subjects.

ALGORITHM

This algorithm is fully implemented using Microsoft Visual Basic. Visual Basic was used because it adopts Unicode which leads to the support of the Arabic

Corresponding Author: Dr. Mohammaed Naji Al-Kabi, Department of Computer Information Systems, Yarmouk University, P.O. Box 566, 21163 Irbid, Jordan Tel: 00 962 795 189 632 E-mail: mohammednaji@yahoo.com

language. In this case we will not need to use Arabization software.

Figure 1 shows a diagram of the major components of our system:

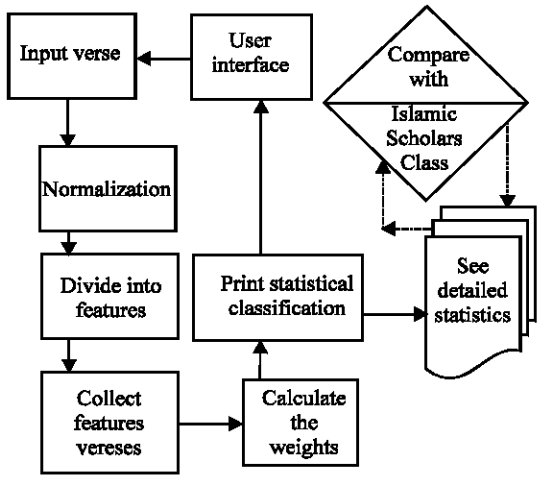

Fig. 1: Major components of the system

Figure 2 describes the algorithm used in this study:

```
Algorithm Classify (Input Sura: String, Input verse: String, Output Statistical Classification of verse)
Begin
1. Choose a Sura;
2. Select a Specific verse From The Chosen Sura;
3. Normalize (verse, Features_set)
4. For Each Feature in Features_set Do
   Begin
      Find all verses in Holy Quran In Which  It Occurs and insert it in verses_Set.
   4.1 For Each verse In verses_set Do
      Begin
      Find all themes(subjects) related to verse
      End
      Themes_frequency= frequency of all Themes for Each feature

    4.2 For each theme in themes_set  Do
      ; Note: Number of  Themes  in General = 15
       Begin
      Weight (feature,theme)= frequency of theme (subject)/Themes_frequency
       End
   End
5. For theme =1 To Number Of themes Do
   Begin
     For feature =1 To Number Of features Do
     Begin
       Sum (theme)=Sum (theme)+ Weight (feature, theme)
     End
   End
6. Max=sum (1)
   For theme=2 to Number Of themes DO
   Begin
      If sum (theme)>Max then Max=sum(theme)
   End
7. Return StatistcalClass=Subject (Max)
End
```

Fig. 2: Algorithm of the system

METHODOLOGY

Present methodology can be summarized by the following steps:

1. Select the desired Sura (Chapter) of the Holy Quran.
2. Select the verse you want to classify.
3. Subdivide the verse into features (keywords).
4. Try to find the recurrences of the keywords (features) in other Surahs (Chapters).
5. For each verse extracted from previous step try to know what is the subject this verse is talking about.
6. Collection of such information needs a holy Quran corpus that contains words a long with the verse and sura it was mentioned.
7. Step 6 was built manually and we depends on: http://www.alnoor-world.com/
8. The system aims to build the following Table 1:
9. The previous table shows that we have to take the maximum summation of subject (S1) which indicates the subject or class of the verse.
10. General subjects that we found are relevant to Muslim scholars classifications are:

- Islam Basics (Islam Pillars) (أركان الإسلام)
- Faith (الإيمان)
- General and Political Relations (العلاقات السياسية والعامة)
- Science and Art (العلوم و الفنون)
- Holy Quran (القرآن الكريم)
- Organizing Financial Relations (تنظيم العلاقات مالية)
- Human and Social Relations (العلاقات الاجتماعية والإنسانية)
- Al-Jehad (الجهاد)
- Religions (الديانات)
- Judicial Relations (العلاقات القضائية)
- Working (العمل)
- Stories and History (القصص و التاريخ)
- Human and Ethical Relations (الإنسان و العلاقات الأخلاقية)
- Trade, Agriculture and Industry (التجارة والصناعة والزراعة)
- Call for Allah (Dawa) (الدعوة إلى الله)

Table 1: Simple view of gained table

	Subject (Theme)			
Word	S_1	S_2	S_3	S_4
W_1	75%	12%	10%	3%
W_2	5%	45%	25%	25%

IMPLEMENTATION

After selecting sura (chapter) and the verse by the user, the system starts normalizing the verse by removing, diacritical marks, punctuations and stop words. In addition to parsing the verse into different tokens.

The contents of our specialized corpus are from alnoor-world web site. The construction of our specialized corpus depends on a freeware program from this site, facilitating the process of searching of any word in the Quran[2].

The system use an executable code written by Hilat E. to remove stop words in the normalization process[3].

This system depends on holy Quran corpus and since such corpus has not been built yet, we build a specialized corpus for Fatiha and Yaseen Surahs (Chapters). Figure 3 shows a Database tables represent a entity relationship of the system:

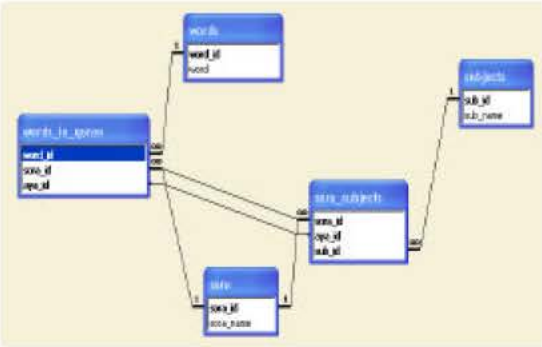

Fig. 3: Entity relationship of the system

This corpus is used to collect all the verses of holy Quran that contains a tokenized word (feature) from a verse under classification, in addition to the subject(s) classification for each verse according to Islamic scholars. Afterward, assume that:

- The number of words remained in the verse after normalization equal to N.
- The number of subjects these N words talking about is M.
- The total number of the predefined subjects of the holy Quran is S.
- P(Si) represents the percentage of a specific subject relative to the other resulting subjects and is computed by the following formula:

$$p(S_i) = \frac{\text{frequency of theme } S_i \text{ for each word}}{\text{frequency of all themes S for each word}}$$

The system also used a two dimensional table of the verse words (features) after normalization and subject percentage of each word. This table correlate the main features (keywords) of any verse with other verses in the same chapter (Sura) and within other chapters (surahs).

In order to determine the class of the verse, the following computation has to be done depending on the Table 1:

1. Summation of subject percentages to each word in the verse as shown in the following equation:

$$\text{Sum}(S_i) = \sum_{j=1}^{N} P(S_i), i = 1,2,\ldots,M$$

We will get 15 values (number of subjects *M*) from the above equation.

2. The prediction of the class of the verse will depends on the following equation:

$$\text{Class} = \text{Subject}(\text{MAX}(\text{Sum}(S_i)))$$

Next, we show the execution of the system based on our algorithm. Figure 4 shows the Interface of the system. The interface enable the user to choose the sura (chapter) and the verse.

Fig. 4: User interface

Figure 5 shows the 2D table, besides the statistics related to the verse.

Fig. 5: Classification details

CONCLUSIONS

In this study we have described the design and successful implementation of a new text classifier suitable for classifying different verses of the Holy Quran. The text classifier has been implemented using Microsoft Visual Basic 6.0.

This work needs a full corpus for the Holy Quran in order to get more precise results. The Yaseen Sura was selected due to its size and the variety of subjects it discusses.

The system has been tested on the Fatiha and Yaseen Surahs (Chapters) and showed 91% accuracy in classifying different verses. The results of the system are compared with the classifications of Islamic scholars to all verses of the Quran.

The accuracy of this system can be improved substantially if a full corpus is built and a better stemmer is used.

REFERENCES

1. Wanjiku, N., 2003. Semantic analysis of kiswahili words using the self organizing map. Nordic J. African Studies, pp: 407-425.
2. Naik and Zakir, 2004. The quran and modern science, compatible or incompatible? -Islamic Research Foundation -PDF Courtesy of www.Ahya.org
3. Al-Shalabi, R., G. Kanaan, J.M. Jaam, A. Hasnah and E. Hilat, 2004. Stop-word removal algorithm for arabic language. Proceedings of 1st International Conference on Information and Communication Technologies: from Theory to Applications, ICTTA'04, (Damascus, Syria, April 2004). IEEE-France, pp: 545-550.