Vocabulary of Quranic Concepts: A semi-automatically created terminology of Holy Quran

Conference Paper · December 2012

DOI: 10.1109/INMIC.2012.6511467

CITATIONS
11

READS
2,308

3 authors:

Tsabbit Mukhtar
Institut Teknologi Sepuluh Nopember
2 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Hammad Afzal
National University of Sciences and Technology
62 PUBLICATIONS **272** CITATIONS
SEE PROFILE

Awais Majeed
Bahria University
13 PUBLICATIONS **52** CITATIONS
SEE PROFILE

Vocabulary of Quranic Concepts: *A semi-automatically created Terminology of Holy Quran*

Tayyeba Mukhtar, Hammad Afzal, Awais Majeed
Department of Computer Software Engineering
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
faisaltayyeba@gmail.com, hammad.afzal@gmail.com, awais.majeed@mcs.edu.pk

Abstract— **The identification and organization of terminology is the foremost step while organizing the domain knowledge for any domain as it is the terms and their inter-relationships that define the conceptual knowledge base. Quran, comprising the divine words of wisdom has been considered and used as prime source of knowledge and guidance for Muslims throughout the world for fourteen centuries. The concepts/topics discussed in Quran have been organized/indexed by many scholars which are used by Muslims who use them to search for guidance regarding various issues of daily life. In current era of information technology, various search services for Quranic topics are available online. They mostly use the terminologies (concepts hierarchy) manually built by scholars. In our work, we have used a semi-automatic approach to identify important concepts/topics from six English translations of Quran, and organized them into a hierarchical structure, named as Vocabulary of Quranic Concepts (VQC). C-NC Value method of term recognition is used to identify significant concepts, which are then manually analyzed by domain expert, and are then organized into a hierarchy using the term-head principle. Due to extreme sensitivity of this work, complete automation of system is avoided and outcomes at all steps are manually analyzed. Currently, we have developed a vocabulary from translation of only second chapter of Quran (Al-Bakara). VQC is available at:**

https://sites.google.com/a/mcs.edu.pk/codteem/projects/qwn

Keywords- Vocabulary of Quranic Concepts; Natural Language Processing

I. INTRODUCTION

Quran, the divine words of wisdom, is regarded as the ultimate source of knowledge. Muslims throughout the world not only read it but also seek guidance about various issues ranging from the basic beliefs such as *Tau'heed* – the oneness of God, *Akhirat* – the life after death) to complex daily life issues such as laws about inheritance, laws about wedding, human rights etc. A huge number of religious books and commentaries have been written by scholars since the revelation of Quran in seventh century A.D, explaining the various concepts mentioned in Quran. In last few years, the progress in the fields of information sciences, especially the techniques developed for information retrieval and information extraction have also inspired various search applications (Web applications, desktop tools etc) that are developed to aid a common man to search and seek guidance from Quran about any given topic/concept.

The terminology (terms and their relationships) plays vital role in organizing the knowledge about any domain. Creation of terminology is usually the first step towards organizing the information about any literature resource. Similarly, for Quranic studies, there are many projects which attempt to organize the information in Quran under certain topics/concepts. These terms, however, are mostly manually collected by experts. In our project, instead of collecting the terms manually, we have utilized the automatic term recognition technique to collect the significant terms from eight English translations of Quran automatically. We used Termine[1], a web service that uses C-NC value technique [1] of term recognition that has been successfully used in various other studies as well [2;3;8]. We then used head-modifier principle [4] to organize the terms under different categories, thus creating a concept hierarchy. The resulting terminology has a total of 592 terms. Our automatically created terminology is then manually analyzed by domain expert to filter out the incorrect terms (terms that are incorrectly recognized by Termine). We have named this terminology as Vocabulary of Quranic Concepts (VQC) and made is available online as XML file.

The rest of paper contains the related work that briefly introduces the C-NC Value technique and certain projects that created similar Quranic Terminologies. Section 3 describes the methodology of creating VQC. Section 4 concludes the paper.

II. BACKGROUND STUDY AND RELATED WORK

There has been an entire range of studies related to organizing the concepts/topics in Quran, spread across the fourteen hundred centuries since Quran revelation. However, for sake of brevity, we are describing here only those recent studies which have been used as basis for online search applications.

[1] http://www.nactem.ac.uk/software/termine/

978-1-4673-2252-2/12/$31.00 ©2012 IEEE

The *Qurany* Qur'an Search Tool[2] involves two modules: a *Keyword Search module* and a *Search for Concept module*. In order to implement Search for Concept they have embedded into their system an index of concepts retrieved from *Mushaf Al Tajweed*, compiled by Dr. Mohamed Habash, Director of the Islamic Studies Centre in Damascus, published by Dar Al-Maarifah in Syria and authenticated by the Al-Azhar Islamic Research Academy in Egypt [5].

A similar topic index has been created by a project Islamicity [6] which is a large Islamic resource that provides databases of translations of Qur'an as well as Hadeeth[3] . It provides four kinds of search: Word Search, Arabic (Phonetic) search, Topic Search and a pre-defined Topic Index search. However, the topics, in this study are also organized manually by experts.

C-NC Value Method: The automatic term recognition technique that we have used in our research is known as C-NC Value Method. The technique has been implemented in a web service Termine. The C-NC value method is a rule based statistical approach that identifies the lexical units using rules based on linguistic properties and then uses frequency of occurrences of those units to associate weight to those lexical units (thus identified as terms).

III. METHODOLOGY

The overall methodology is depicted in Fig. 1. The process starts with collection of English translations of Holy Quran. For this purpose, we have used English Translations of the Holy Qur'an by eight different authors. These authors are Sahih International, Muhammad Sarwar, Qaribullah & Darwish, Muhammad Asad, Tahir-ul-Qadri, Hilali & Khan, Marmaduke Pickthall and Ahmed Raza. At initial stage of research, we have focused on translation of only Second Chapter, Al-Bakara.

A. Identification of Significant Terms

The English translation is processed to label significant terms using *Termine*. All eight translations are put together into a single text file which is then processed. A total of 996 terms were obtained automatically. These terms were then validated by a domain expert to filter out the incorrectly identified terms. 578 out of 996 terms were marked as correct and valid terms. One major reason of relatively lower accuracy in term recognition using *Termine* is the presence of old style English words, e.g. *efface thou*, *doth*, *dost* etc. in translations we used. Such terms were identified as valid terms by *Termine*, which is not true in our case, and that is the reason we have made our methodology semi-automatic through involvement of domain expert due to high sensitivity of task at hand.

[2] http://quranykeywords.appspot.com

[3] In Islamic terminology, the term *Hadeeth* refers to reports of statements or actions of Prophet of Islam, Muhammad (peace be upon him), or of his tacit approval or criticism of something said or done in his presence. (*Encyclopedia of Islam*).

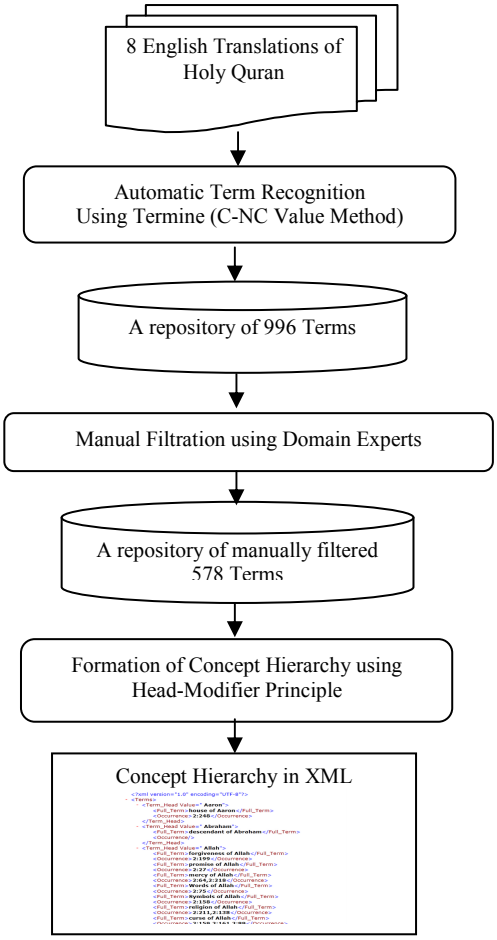

Figure 1. Overall methodology towards creating a concept hierarchy for Holy Quran

B. Creating the Concept Hierarchy

Majority of terms identified as result of applying *Termine* are multi word terms. The constituent words in multi-word terms have certain relationships with each other; one of them is identified as *Head-Modifier* relation. According to *Head-Modifier* principle, one element of the multi-word term acts as naming the general or semantic category to which the whole word belongs [4]. This element is generally termed as *head* and it also generally defines the semantic category as well that the multi-word term belongs to. Other elements of the multi-word terms distinguish the particular member from other members belonging to the same semantic category.

The head-modifier principle allows identifying the multi-word terms and grouping them into set of terms related to each other through the hyponymy relationship. The head of the word is the hypernym of all those terms. For instance, in the Qur'anic domain, the terms *blood money* and *bridal money* are used. *Blood money* represents the money given as retribution to the family of murdered according to the Law of Equality in Punishment, or Qisas. The *bridal money* (commonly referred as *Meher*) is given to the bride as part of the marriage contract. Both these terms belong to the semantic category of *Money* (or hyponyms of Money). Our method would group all hyponyms of *Money* that appear in translations of Quran and, therefore,

will enable the user to find out all the categories of money that need to be dispensed according to the Islamic law.

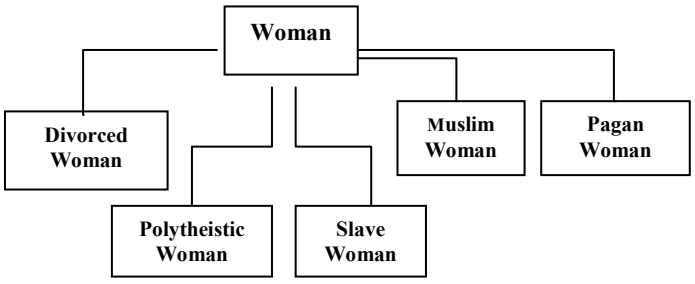

Fig. 2a

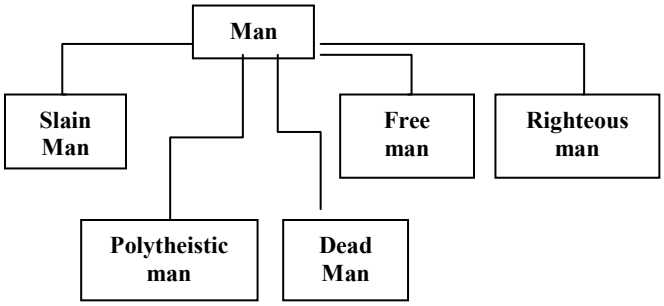

Fig. 2b

Figure 2. 2a and 2b represent the terms *Woman* and *Man* respectively along with their hyponyms as they appear in VQC. *Dead Man, Polytheistic Man, Free Man* etc have been arranged under the main concept of *Man*; *Slave Woman*, *Muslim Woman*, *Divorced Woman* are arranged under *Woman*

In the concept hierarchy, the head term is illustrated as parent node (or head node) whereas all its hyponyms (the multiword terms with that head) are illustrated as its children nodes. This relationship is further explained using simple examples of head terms *Man* and *Woman* along with their hyponyms (child terms) that appear in Quranic translations in Fig 2.

Upon further analysis of translation text by domain experts, it was found that there were some terms that were deemed important to be included into VQC but were missed by *Termine*. These terms were analyzed by domain experts to devise some rules to identify such missing terms. It was found that most of such terms were with patterns such as "*Noun1 of Noun2*". For example, *fear of Allah*, *inviolable house of worship*, *Jesus son of Mary*, *day of recompense*, *day of judgment* etc. Some of these terms were partially identified by *Termine* such as *inviolable house*, but the corresponding actual Quranic term is *inviolable house of worship* which was missed. Similarly, the terms *Jesus son* and *most acceptor* were recognized by *Termine*, but their complete terms *most acceptor of repentance* (an attribute of God), and *Jesus son of Mary* in the Qur'anic context were missed. A set of special rules were devised to identify such missing terms; these rules are given in Table 1.

TABLE I. ADDITIONAL RULES TO EXTRACT TERMS FROM QURANIC TRANSLATIONS

Rule to Extract	Example of Extracted Term
<Noun1>"of "<Noun2>	month of Ramadan
<Adjective><Noun1>of<Noun2>	most acceptor of repentance
<Noun1> <Noun2>of <Noun3>	jesus son of mary
<Noun1>of the<Noun2>	abode of the hereafter
<Noun1>of the<Noun2>and the <Noun3>	originator of the heavens and the earth

The terms identified using set of rules described in Table 1 were also included in VQC after manual filtration. In these terms, the term head is the substring appearing before the preposition "of" as the head-modifier principle requires the head to be set to the substring appearing before "of" in order to conform to the head of the corresponding structurally variant terms. For example, in *Day of Judgment*, *Day* is the head term; in *Month of Ramadan*, *month* is head term.

In special cases, we made exception in the rule of Head-Modifier principle. There are a significant number of occurrences of attributes of Allah in Quran which appear with the pattern *Attribute of Allah* e.g. *Promise of Allah*, *Mercy of Allah*. In these cases, we grouped all such attributes (terms) with *Allah* (shown and explained later in Fig. 4).

C. *Serialization of VQC in XML form*

We have made VQC available in XML form. The format of XML file is given in Fig. 3.

```
<Terms>
   <Term_Head Value = "term_head_value">
      <Full_Term> Value of Full Term<Full_Term>
      <Occurrence>X:Y</Occurrence>
  </Term_Head>
</Terms>
```

Figure 3. Vocabulary of Quranic Concepts in XML Format

We have devised a format of XML that presents the value of term, the head of term and also the location of the occurrence of that term in Quran in the format X:Y where X denotes the (Surah) Chapter number and Y denotes the Ayat (Line) number. A snapshot of VQC presenting some of the attributes of Allah is given in Fig. 4.

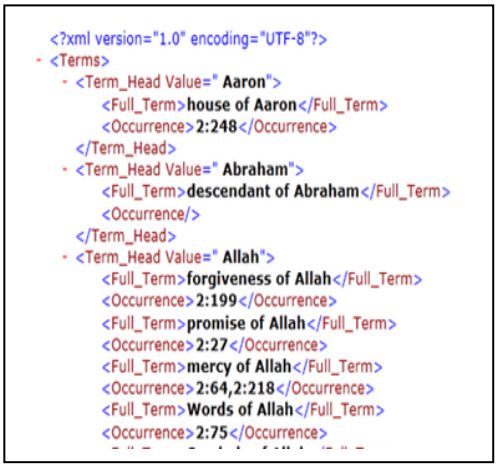

Figure 4. A snapshot of XML file containing Vocabulary of Quranic Concepts

IV. CONCLUSION

We have presented a methodology to identify significant terms that describe important concepts appearing in Quran. The concepts are organized into a vocabulary that is presented in hierarchical form to group the similar concepts (to be specific, the terms that have hypernym-hyponym relationship). There are several other such indexes available as well, however, to the best of our knowledge, they are all manually created. We want to clearly state that all such efforts of creating the indexes of concept of Quran have been done very carefully, most of them have been built after years of efforts of highly knowledgeable religious scholars. Our system, by no means, is comparable to those; however, we have tried to incorporate a methodology to automatically identify significant terms; this methodology can easily be scaled to involve other religious texts (e.g. Hadeeth).

The core of our methodology is automatic term recognition, however, due to sensitivity of the issue and lack of room of ambiguity or error, the automatic term recognition process is followed by a manual filtration by domain expert as this task demands 100% accuracy that Natural Language Processing (NLP) systems cannot guarantee.

The vocabulary built is available in XML form. We have formatted it in the way that it can be easily embedded into other Information Processing systems (such as for Searching applications). We have also made available the location of each term in Quranic text so that any person searching for any concept can also find the related Ayat (line of text) in Quran. A search tool is also developed as proof of concept [9], however its details are beyond the scope of this paper.

REFERENCES

[1] K. Frantzi, S. Ananiadou, H. Mima. “Automatic recognition of multiword terms: the C-value/NC-value method”. In International Journal on Digital Libraries, vol. 3, no.2, pp.117-132, 2000.

[2] H. Afzal, R. Stevens, G. Nenadic: “Towards Semantic Annotation of Bioinformatics Services: Building a Controlled Vocabulary”, Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008): pp. 5-12, 2008.

[3] M. Krauthammer, G.Nenadic: “Term Identification in biomedical literature”. Journal of Biomedical Semantics. Vol 37(6), pp 512-526, 2004.

[4] A Hippisley, , D Cheng. and K Ahmad: “The head-modifier principle and multilingual term extraction”, In Journal of Natural Language Engineering, vol 11(2), 129-157, 2005.

[5] H. Mohamed, “Mushaf Al Tajweed”, Dar-Al-Maarifah, Syria, 2001.

[6] Islamicity, “Quran Search”; http://www.islamicity.com/QuranSearch.

[7] T. Mukhtar, H. Afzal. “Quranic WordNet Database”. In Press.

[8] C Brewster, S Jupp, J Luciano, D Shotton, R D Stevens and Z Zhang: “Issues in learning an ontology from text”. BMC Bioinformatics, Vol 10(5). 2009.

[9] Tayyeba Mukhtar: “Creating an Infrastructure for semantic search of the Holy Quran using English WordNet”. MS Thesis. National University of Sciences and Technology, Islamabad, Pakistan. 2012.