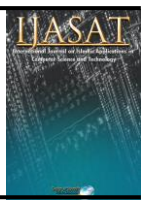

Automatic Knowledge Base Constructor for Al-Quran Retrieval System

Mohamad Fauzan Noordin[1,a], Sharyar Wani[2, b], Tengku Mohd T. Sembok[3,c], Roslina Othman[4,d]

[1, 2, 4] International Islamic University Malaysia, Malaysia
[3] National Defence University Malaysia
afauzan@iium.edu.my, bsharyarwani@gmail.com, ctmts@upnm.edu.my, droslina@iium.edu.my

ABSTRACT

Web 2.0 has changed the strategy of the world. The virtual world has a large impact on the society. There is enormous data on the web but the knowledge behind the data has not been utilized even to the slightest in comparison to its size. Web 3.0 aims at knowledge extraction from the data, there is need to develop means and ways to extract the knowledge behind the data. In this area of research, Muslim researchers have directed their works towards the availability of digital resources for Al-Quran and books of Hadith since they form the foundations of Islam. However, the research done so far has not gone deep into the area of knowledge representation of Al-Quran and Hadith. The current work looks into development of knowledge representation formalism for Al-Quran using the logical base as it is expressive in nature and has proven successful previously even in complex situations. The logical base needs indexing in order for efficient retrieval as well. It would be extremely difficult to maintain the consistency of the logical base if done manually. Hence this work primarly focuses on development of a automatic knowledge base constructor. The current work has a large significance, as it will ease the process of information access to the Muslim community by using the knowledge base for retrieval mechanisms. Not only that the work will be beneficial for Non-Muslims to know more about Al-Quran easily and thus gaining more and more information about Islam.

Keywords: Al-Quran, Web 3.0, Semantic Web, Logico-Linguistic, Knowledge Base, Automatic Constructor

1. Introduction:

The decentralized design of the Web has been a boon for the user contribution towards the information via the Internet. The World Wide Web is the information resource hub of the century and shows promising future directions. The current design has enabled the users to access the information with ease. The decentralized concept aiming at user contribution is the current design i.e. Web 2.0 (Dhingra & Bhatia, 2011).

The enormous potential of the Web in regards to information resourcing has not been utilized fully since machines lack analytical capabilities - to organize process and interpret the information in a desired and meaning-full manner. Search engines have become the most used entities on the web. However most of these search engines are still generic in nature (Ohshima, Jatowt, Oyama, Nakamura, & Tanaka, 2009).

Although the current search engines seem to have improved a lot in terms of speed, yet majority of the results obtained are extraneous to search entity. The current search engine result listing only adds to the confusion of the user most of time (Kumar, Rana, & Singh, 2012).

The solution to the problem is that computers be made capable of processing and interpreting the data. The machines need to have high analytical skills for this knowledge based information representation. Since machines do not have cognitive capabilities, therefore knowledge behind the information needs to be represented according to a proper and consistent formalism for the analytical skills to operate process and extract information from the data.

In artificial intelligence, it is necessary for knowledge to be represented. Therefore, models and methods for the same have a very important role. In the past, various knowledge models have been presented such as predicate logic, semantic nets, frames, fuzzy logic etc. Some of them have proven quite useful for designing intelligent systems and solving complex problems.

Van Do (Van Do, 2009) presents a computational network for real-time applications such as studying knowledge and solving problems in analytical geometry. The authors claim that they are easy to use and produce human readable solutions.

Bonino and Corno (Bonino & Corno, 2008) propose a semantic search based system using pruning algorithm which employs document self-similarity preserving the most significant components of the document conceptual vector. The proposed mechanism has been tested yielding better results. However there is no change in the conventional approach of search engines relying on keywords and their frequencies. (Tengku M. T. Sembok, 2013)

Information retrieval forms a key functionality of Semantic Web. Ontologies are standard for Semantic IR systems as they help in interaction between users and software's by using concepts, relations and semantic inference. The concepts and relations help to achieve the desired expressiveness. However, there are a few issues associated with them. Firstly, they return the documents relating to the queries made. The search engines on the other hand list almost all possible related texts. This does not satisfy the user needs who need a specific answer to the question/search queries they input. (Bekhti, Rehman, Al-Harbi, & Saba, 2011).

Secondly, the current Semantic IR systems require complex queries while the others provide very simple query language. The results produced are not expressive in nature. On the other hand current engines suffer from the problem of having no user interaction with the retrieved resources and even queries are not provided with any competences (Sy et al., 2012).

The efficiency of a question answer system is directly dependent on a consistent knowledge base, which in turn depends on a sound knowledge base. Artificial intelligence will require this since it needs that everything related to the problem domain should be a part of it. The fundamental principle is that knowledge organization is the key principal behind the degree of efficiency of the result .So it becomes a necessity to understand that knowledge cannot be represented randomly and a better system would require a better knowledge base. A sound & consistent formalism to represent data is mandatory for an efficient knowledge base (Tanwar, Prasad, & Aswal, 2010).

The logical formalism developed needs to apply to the normal text in order to produce a sound knowledge base. Every sentence would have a different grammatical structure, which

implies there will be an entirely different logico-linguistic rule being applied to it. Nevertheless, it will be highly cumbersome to do a manual design and population of the knowledge base. As such, there is a need of automatic translation for design of knowledge bases.

Al-Quran and Hadith form the basis of Islam i.e.; they are the foundation for the teaching of Islam. In the recent years, a lot of work has been done in regards to the availability of digital resources for Al-Quran, Hadith and other works by Muslim scholars to be digitally available. However not many works have been directed towards the area of knowledge representation in this area. The existing keyword based search retrieves much irrelevant information. It is therefore a challenge to represent this knowledge so that it becomes easy for the users to access it. The current effort is to have an efficient digital knowledge representation for Al-Quran. The current work looks into the knowledge representation of Al-Quran. Extraction of knowledge from religious texts has been a focus of many researches in the past few years. Attempts have been made to develop applications for representation and retrieval of such knowledge. In this block of research, researchers have looked into retrieving knowledge from Al-Quran and Hadith, etc. in order to ease the process of learning (Baqai, Basharat, Khalid, Hassan, & Zafar, 2009).

Baqai et al. (Baqai et al., 2009) have looked into leveraging Semantic Web technologies for such knowledge discovery involving data integration and semantic annotation. They propose to modify their system to produce agent based intelligent modeling and retrieval framework for Al-Quran.

In the context of logical formalism for knowledge representation, (Sembok, Zaman, & Kadir, 2008) present the approach of unified linguistic-logical representation which has shown improvements over the benchmark numbers in question answering. The unification is considered in terms of function such as user profiling, declaring world knowledge etc.
The researchers further present an efficient mechanism for answer extraction by introducing ground term as expanded notion of the answer generated literally with existential quantifying within the theorem-proving paradigm. The researchers have designed their knowledge base using logical linguistic model similar to their previous works whereby they have implemented the enhanced answer extraction. (Rabiah, Sembok, & Halimah, 2009).

However, no major works have been done in the area pertaining to representing the knowledge of Al-Quran in a consistent and formalistic manner. Such formalism would eventually lead to development of a question answer system that would retrieve knowledge with high levels of efficiency. However it is necessary to have an automated tool for the developing the consistent knowledge base for Al-Quran.

The objectives of this research can be summed up as:

i. To auto construct knowledge representation for Quranic verses using an automated tool.
ii. To index the designed knowledge base automatically for further access e.g., retrieval, etc.

2. Automatic Knowledge Base Constructor

2.1 Implementation

This work presents the automatic translator tool to construct a consistent knowledge base for Al-Quran based on rules reducing the design span to minimum at the same time. The automatic translator uses the logical formalism developed to populate the knowledge base. The automatic translator will solve the crisis of translating, ordering, populating the knowledge base with enormous data as well as avoiding the hassle of manual translation. Not only is it difficult but impossible owing to the amount of phrases involved from 6236 Verses.

The automatic translator presented in this paper will accept the text as input, process it using parsers, apply the designated rules of application, and finally export a new set of data into the knowledge base. The new set of data is the input data filed accordingly as per the logical rules developed.

The automatic translator accepts the Quranic verses stored in .txt file and loads it in the core. It sends the original text to the database and receives an index number in exchange, which gets stored. This later helps us for successful retrieval and other logical representations of the input text.

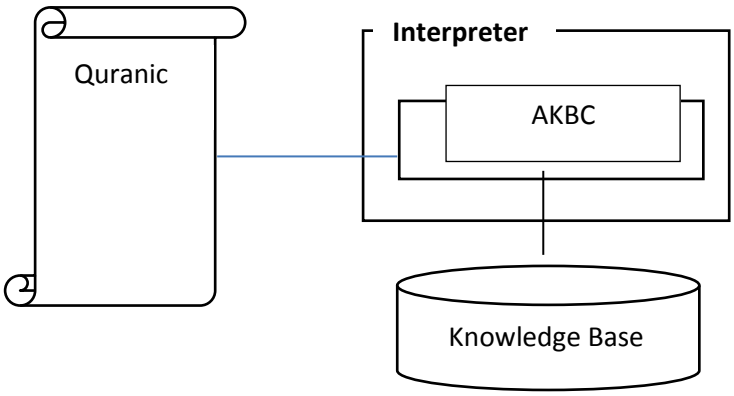

Figure 1: Implementation of Automatic Knowledge Base Constructor for Al-Quran

The automatic translator then stipulates a call to its internal function to activate the parser. The parser libraries are accessed and initialized for parsing. The text is then passed on for parsing. As soon as the parser gets into activity, it asks the end user the type of parser and parser library that should be implemented. Based on the selection, the text is parsed accordingly and two outputs are presented – one is the syntactic tree of the parsed text that primarily focuses on the phrase structure of the sentences. Second is the output with the phrases and a special index number of each phrase. It should be noted that indexing is a very crucial phenomenon that has to be affirmed at every stage so that the retrieval process later can be easy, fast and accurate.

The parsed text with the indexes is sent into another section of the knowledge base and an additional index is automatically assigned based on the designed indexing mechanism. This is the second set of data with which the knowledge base is populated with.

As the phrase structure is available, the formalism developed is brought into operation. The formalism application module of the automatic translator is activated primarily. The

phrase input serves as a flag for activation for the module. The automatic translator applies the formalism on the parsed text and a new arrangement of phrases is obtained as a result. The result is sent to the knowledge base as another set of data. As a requisite of storing any kind of data, the third set of data is only granted index numbers for retrieval and accuracy. This marks the end of population of the knowledge base process using the automatic logical translator tool. It is important to note that all sets of data in the knowledge base are already indexed which implies that the data can be accessed easily and accurately.

2.1.1 Example

Input: Chapter 3, Verse 35

"When the wife of Imran said, "My Lord, indeed I have pledged to You what is in my womb, consecrated [for Your service], so accept this from me. Indeed, You are the Hearing, the Knowing."

Output after translation

The representation of the verses was obtained using the formalism developed. In this example the obtained result is:

Sentence 1:

said(wife_of_Imran,my_Lord,pledged(i,to_you),what_is(in_my_womb,for_your_service),accept(this,from_me)).

Sentence 2:

are(you,the_Hearing,the_Knowing).

3. Results & Discussion:

The automatic translator was tested with different sets of date. 5 sets of different data were incorporated into the automatic translator to verify the accuracy and speed of the translator. The speed can be checked at the run-time itself and the accuracy was checked by testing the database tables manually.

Table 1: Translator speed test

Test Data Set	**Time [secs)]**
Set 1	6
Set 2	6.5
Set 3	6.5
Set 4	7
Set 5	6

Set 1 – Set 5 are verses from Al-Quran that were provided as input to the translator to populate the knowledge base based on the logical formalism. The following are the sets 1-5 given as inputs to the translator:

Set 1: "But when she delivered her, she said, "My Lord, I have delivered a female." And Allah was most knowing of what she delivered, and the male is not like the female. "And I

have named her Mary, and I seek refuge for her in You and [for] her descendants from Satan, the expelled."

Set 2: "So her Lord accepted her with good acceptance and caused her to grow in a good manner and put her in the care of Zechariah. Every time Zechariah entered upon her in the prayer chamber, he found with her provision. He said, "O Mary, from where is this to you?" She said, "It is from Allah. Indeed, Allah provides for whom He wills without account."

Set 3: "So the angels called him while he was standing in prayer in the chamber, "Indeed, Allah gives you good tidings of John, confirming a word104 from Allah and honorable, abstaining , and a prophet from among the righteous."

Set 4: "When the wife of Imran said, "My Lord, indeed I have pledged to You what is in my womb, consecrated [for Your service], so accept this from me. Indeed, You are the Hearing, the Knowing."

Set 5: "At that, Zechariah called upon his Lord, saying, "My Lord, grant me from Yourself a good offspring. Indeed, You are the Hearer of supplication."

The translator processed the texts at an average speed of 6.4 secs. The data in the knowledge base was compared to the original data and the logical formalism and was found to be matching. The phrase arrangement after the logical formalism is applied is the most important set to be verified in the knowledge base and it was found that the knowledge base was populated according to the formalism. In addition, every phrase was indexed for later process of retrieval or access for other operations.

4. Conclusion:

This research aimed at the development of an automatic translator of data to design a knowledge base for Al-Quran. It is important that the translation process be carried systematically and consistently. The speed of translation could be another factor to look upon later in the experimentation and research process. The current translator solves the problem of translating the text and design of knowledge base automatically based on the designed set of rules.

The evaluation of the current work shows that the automatic translator can construct the knowledge base for Al-Quran in a consistent manner with proper indices. This would eventually help in easy acess of the knowledge contained in the base which is necessary for efficient outcomes in information retrieval or more specifically question answer systems.

The automatic translator easens the process of knowledge base formation for Al-Quran. It has a larrge significance, as it will to help to form a sound knowledge base for A-Quran retrieval system, thus making information access to the Muslim community. Also this will help Non-Muslims to know more about Al-Quran easily and thus helping in inviting them to Islam.

5. References

Baqai, S., Basharat, A., Khalid, H., Hassan, A., & Zafar, S. (2009). *Leveraging semantic web technologies for standardized knowledge modeling and retrieval from the Holy Qur'an and religious texts*. Paper presented at the Proceedings of the 7th International Conference on Frontiers of Information Technology, Abbottabad, Pakistan.

Bekhti, S., Rehman, A., Al-Harbi, M., & Saba, T. (2011). Aquasys: An Arabic Question-Answering System Based On Extensive Question Analysis And Answer Relevance Scoring. *International Journal of Academic Research, 3*(4), 45-54.

Bonino, D., & Corno, F. (2008, 1-5 Sept. 2008). *Self-Similarity Metric for Index Pruning in Conceptual Vector Space Models.* Paper presented at the Database and Expert Systems Application, 2008. DEXA '08. 19th International Workshop on.

Dhingra, V., & Bhatia, K. K. (2011). Towards Intelligent Information Retrieval on Web. *International Journal on Computer Science & Engineering, 3*(4), 1721-1726.

Kumar, S., Rana, R. K., & Singh, P. (2012). A Semantic Query Transformation Approach Based on Ontology for Search Engine. *International Journal on Computer Science & Engineering, 4*(5), 688-693.

Ohshima, H., Jatowt, A., Oyama, S., Nakamura, S., & Tanaka, K. (2009). Towards Improving Web Search: A Large-Scale Exploratory Study of Selected Aspects of User Search Behavior. In G. Vossen, D. E. Long & J. Yu (Eds.), *Web Information Systems Engineering - WISE 2009* (Vol. 5802, pp. 379-386): Springer Berlin Heidelberg.

Rabiah, A. K., Sembok, T. M. T., & Halimah, B. Z. (2009). *Broadened answer extraction of QA system capability using expanded notion of logical-linguistic approach*. Paper presented at the Proceedings of the international conference on Computational and information science 2009, Houston, USA.

Sembok, T. M. T., Zaman, H. B., & Kadir, R. A. (2008). *IRQAS: information retrieval and question answering system based on a unified logical-linguistic model*. Paper presented at the Proceedings of the 7th WSEAS International Conference on Artificial intelligence, knowledge engineering and data bases, Cambridge, UK.

Sy, M.-F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., & Ranwez, V. (2012). User centered and ontology based information retrieval system for life sciences. *BMC Bioinformatics, 13*(1), 1-12. doi: 10.1186/1471-2105-13-S1-S4

Tanwar, P., Prasad, T. V., & Aswal, S. (2010). Comparative Study of Three Declarative Knowledge Representation Techniques. *International Journal on Computer Science & Engineering*, 2274-2281.

Tengku M. T. Sembok, R. A. K. (2013). A Unified Logical-Linguistic Indexing for Search Engines And Question Answering. *International Journal Of Mathematical Models And Methods In Applied Sciences, 7*(1), 8.

Van Do, N. (2009). Computational Networks for Knowledge Representation. *Proceeding of World Academy of Science, Engineering and Technology (ICCSISE 2009), 56*.