SUST Journal of Engineering and Computer Science (JECS), Vol. 17, No. 2, 2016

Islamic Applications of Automatic Question-Answering

Mohamed Adany Hamdelsayed, Eric Atwell

Computer Science and Information Technology Faculty, Sudan University of Science and Technology (SUST), Khartoum, Sudan
School of Computing, Faculty of Engineering, University of Leeds, Leeds LS2 9JT, England
mohdn111@sustech.edu, E.S.Atwell@leeds.ac.uk

Received: 02.08.2015
Accepted: 10.11.2015

Abstract — A search engine aims to retrieve full documents whereas a question answering system aims to extract the exact answer. A question answering system involves the process of accepting a NL (Natural Language) question, analyzing, and processing to match against a knowledge base to generate the right answer from documents for users. For the Holy Quran this involves accepting the NL question and processing it to retrieve the right verse or verses from our Quran knowledge base. Question answering systems can use two types of algorithms: rule based techniques and/or AI (Artificial Intelligence) based techniques. Question Answering systems have three main components: question classification, information retrieval and answer extraction. We present a rule-based system for the Holy Quran that retrieves the right verse(s) from the Holy Quran instead of generating NL answers. We use a java program to extract the answer from a MS-Access database which contains our knowledge base for our Quran question answering system. We find that the system gives better results for the question after improving the system by removing stop words.

Keywords: *Corpus, Quran, Arabic, diacritics, stop words.*

المستخلص— يهدف محرك البحث لاسترجاع الوثائق الكاملة بينما يهدف نظام سؤال وجواب للاجابة عن السؤال المطلوب فقط. نظام الرد على السؤال يشمل عمليات إدخال السؤال(باللغة الطبيعية)، والتحليل، والمعالجة للحصول على الإجابة الصحيحة من قاعدة المعرفة التي تقابل السؤال من الوثائق الموجودة للمستخدمين. وبالنسبة للقرآن الكريم يتم إدخال السؤال للحصول على الآية الصحيحة المطابقة للسؤال. ويتكون نظام سؤال وجواب من ثلاث مكونات رئيسة : تصنيف السؤال، استرجاع المعلومات ، واستخلاص الجواب. قدمنا نظام يعتمد على القواعد لاسترجاع اللآيات الصحيحة من للقرآن الكريم بدلاً من توليد إجابات باللغة الطبيعية. تم استخدام برنامج بلغة جافا للحصول على الإجابات من قاعدة بيانات مايكروسوفت أكسس والتي تحتوي على قاعدة معرفة نظام الرد على السؤال. وجدنا أن النظام يعطي نتائج أفضل بعد تحسين النظام بواسطة إزالة الكلمات المتكررة المستبعدة.

INTRODUCTION

A question answering system involves sequences of processes: entering a question in natural language; analysis of this question, by filtering it and finding the key words for search, by tokenizing the question words and removing the unnecessary words from the question; to find the related document(s) that contains the related answer(s) if there is any answer.

Question answering systems research interacts with other areas of Computing such as natural language processing (NLP), information retrieval (IR) and information extraction (IE). Example research on applications of question answering in different areas were contributed in:

- Education: [1] presented a fully automatic question-answering system for intelligent search in e-learning documents.
- Biomedicine: [2] presented Biomedical question answering: a survey, and [3] proposed a usability survey of biomedical question answering systems.
- Linguistics: [4] proposed an online approach to English and Punjabi question answering; and a prototype Bangla question answering system using translation based on transliteration and table look-up as an interface for the medical domain was proposed by [5].

Question and answering systems (Q/A) components and architecture:

In general, all question and answering systems contain three components: question

classification, information retrieval, and answer extraction. Each one of these three components may be one main component or more subcomponents.

- *Question Classification:*

Questions have four types [3]: definition, factoid, list, and Boolean. In addition, a question can classified as flat or hierarchical. Depending on this classification, processing starts when we enter the question:

- The input question must be analyzed to extract the important information inside the text to move to next stage.
- Classifying the question depends on two types of taxonomy: hierarchical and flat taxonomies. [6] proposed a NSIR (pronounced answer) which used a flat taxonomy with seventeen classes; whereas [7] and [8] proposed a hierarchical taxonomy: the first classified the question types into nine classes (what, who, how, where, when, which, name, why, whom), and each one with its subclasses, while the second one proposed another taxonomy which had six main classes and each one with its subclasses.
- Two approaches were used for classifying documents: machine learning and rule based classifier.

- *Information retrieval:*

This stage is aimed to fetch and return the documents that are relevant to the query. The search engine marks the document relevant depending on the key words in the documents and then ranks it depending on other features such as: first searches for all of your keywords as a phrase, measures the adjacency between your keywords, measures the number of times your keywords appear on the page and etc. then ranking these documents in terms of relevance to the question text..

- *Answer Extraction:*

This stage is used to extract one or more candidate answer(s) from relevant documents, in which it chooses the passage(s) from the document(s) and displays them after ranking depending on specified criteria. In an advanced question and answering system there may be some concatenation from many passages to find the right answer.

Holy Quran and Arabic language:

Islam is now the second most popular religion in the world and more than a billion and a half people believe it. The main source of legislation in Islam is the Holy Quran and Sunnah, so these texts require special attention. Also many people need to know the instructions, commands and facts from the Quran and Sunnah. The Quran is written and cited in Arabic; so, for this computational question answering from the Quran depends on Arabic language processing. Here is a brief introduction to explain the importance of the Arabic language and Quran and its computational processing.

The Arabic language is a Semitic language, it is spoken and written by over 300 million people in the world. It is one of the official languages of the United Nations, also it is one of the main six language used in the United Nations, and it is ranked as the third most important international language after English and French.

[9] and [10] presented some Arabic challenges in NLP such as: morphological characteristics in which one token can take many meanings in the Arabic language. [9] listed several challenges of Arabic natural language processing: "… the orthographic convention of leaving out short vowels and other distinguishing marks in written text, derivation of Arabic language (by adding affixes (prefix, infix, or suffix)), inflection of Arabic language, no capital letters in Arabic language, and lack of digital resources in Arabic language such as: lexicons, Arabic corpora and dictionaries. Also [11] presented further challenges: written from right to left, diacritics, ambiguity, and data sparseness in Arabic texts. Also we can add the problem of existence of the Arabic phonemes or sounds like "ض", which is the Arabic character for a sound that does not exist in many other languages such as English.

RELATED WORK:

There has been a lack of question and answering systems for the Arabic language and Holy Quran. Until now Arabic question and answering systems research did not reach maturity for many reasons mentioned above, and attention to its importance only started recently.

The first Arabic question answering system was AQAS (Arabic Question Answering System) [12], which accepts a question and extracts answers only from structured data and not from raw text; it was designed for the restricted domain of radiation and its effects, and its knowledge-base used structured data only in a database; also no experimental evaluation results were reported.

[13] presented QARAB (Question answering for Arabic), which is an Arabic QA system that uses both natural language processing techniques and information retrieval. Its data was extracted from the Alraya newspaper published in Qatar. It retrieves only short passages that contain an answer, not necessarily the exact answer.

[14] presented ArabiQA which is another Arabic QA system that uses an Arabic Named Entity Recognition (NER) with the Arabic-JIRS (Java Information Retrieval System) to extract passages from Arabic documents. ArabiQA is an Arabic Q/A prototype based on the JIRS Passage Retrieval (PR) system and a Named Entities Recognition (NER) module. It works with factoid questions. In order to implement this module the authors developed an Arabic NER system and a set of patterns manually built for each type of question.

BUILDING A DATA SET FOR HOLY QURAN QUESTIONS AND ANSWERS:

Our first task was extracting an example set of questions and answers from the Holy Quran (Surat Al-Fatiha and Al-Baqarah Chapters) to be the core of a question and answers corpus to be used as a gold standard corpus. This work was done by reading the holy Quran and devising suitable questions relating to each verse and its answer, which is the verse number in this chapter or any other related verses number(s) in other chapter(s). These questions and answers were written in an Excel spreadsheet, see Table 1.

Table 1: Quran QA spreadsheet header

الرقم NO	السؤال Question	السورة Chapter	الآية Verse	رقم السورة Chapter No

Our example dataset contains 215 questions and more answers because several verses can be answers to one question. Also the last column is for comments to explain some verses when it needs more explanation to be clear. These questions were then validated and revised in Gabrah College by Islamic and Arabic scholars. Table 2 has some examples from the validated spread sheet.

Table 2: Quran QA sample after expert validation

الرقم	السؤال	السورة	الآية
1	لمن الحمد؟	1	2
2	ما هي صفات الله التي ذكرت في سورة الفاتحة؟	1	2-3-4
99	أين يؤتى النساء؟	2	222
104	هل الرجال والنساء على درجة واحدة ولِم ؟	2	228
210	هل يحاسبنا الله على ما نبديه أو نخفيه في أنفسنا؟	2	284

A second example Quran QA dataset containing 47 questions was extracted from an Islamic web site. In this source there are some problems in presenting data, as it used some English letters like r and U. We extracted the questions and related answers from the website as illustrated in Table 3.

Table 3: Quran QA sample from WWW

الرقم	السؤال	السورة	الآية
1	ما هي السبع المثاني ؟	1	1-2-3-4-5-6-7
7	من هن الأزواج المطهرة ؟	2	25
16	ماذا كان شرط قوم موسى u حتى يؤمنوا ؟	2	55
21	ماذا كانت وصية إبراهيم u لأولاده وكذلك وصية يعقوب u لأولاده ؟	2	132

Finally we combined the two sources in the third spreadsheet containing 263 Quran questions and their answers. This spreadsheet was sorted by verse numbers. It was easy to check there is no duplication here because we only covered two chapters of the Quran; but if we continue there may be duplications and so we may need another column as a cross-check key to show the chapter number and the verse or verses. Each answer that includes more than one verse was marked in the end of the rows. With the combined spreadsheet the data was processed as follows:

To differentiate the two different sources we changed the font type and size as in Table 4.

Table 4: Differentiate between sources

لمن الحمد؟	1	2
ما هو الكتاب الوحيد الذي لا يوجد أي ريب أو شك فيه ؟	**2**	**2**

If the question and its answers are found in both the two sources: colored yellow as in Table 5.

Table 5: differentiate between two questions from different sources

15	كم عدد السماوات؟	2	29	مكرر
16	ما عدد السماوات ؟	2	29	مكرر

If there are two questions repeated and followed by another two repeated questions we colored first yellow and second blue as in Table 6.

If there are three questions repeated from the two sources but the second and the third interleaved in spreadsheet sorting we colored them as yellow and blue and red as shown in Table 7.

Finally we add a new column for verse English translation by Abdullah Yusuf Ali, who "was a British-Indian Islamic scholar who translated the

Qur'an into English. His translation of the Qur'an is one of the most widely known and used in the English-speaking world. He was also one of the trustees of the East London Mosque." Our final version of the Quran QA corpus is illustrated in Table 8.

Table 6: double repeated question from different sources

15	كم عدد السماوات؟	2	29	مكرر
16	ما عدد السموات ؟	2	29	مكرر
17	ماذا علم الله نبيه آدم؟	2	31	مكرر
18	ماذا تعلم آدم عليه السلام من الله جل جلاله وكان هذا العلم ليس عند الملائكة ؟	2	31	مكرر

Table 7: Three repeated question from different resources

20	من هو الذي لم يسجد آدم ؟	2	34	مكرر
21	**من الذي رفض أمر الله تعالى بالسجود لآدم عليه السلام؟**	2	34	مكرر
22	أين أمر الله آدم وزوجه أن يسكنا؟	2	35	مكرر
23	**أمر الله تعالى آدم وزوجه أن لا يقربا شيئاً ما هو ؟**	**2**	**35**	مكرر
24	عندما خلق الله تعالى آدم عليه السلام وحواء ماذا كان سكنهما ؟	2	35	مكرر
25	**ما الشيء الذي منعه الله تعالى عن آدم وحواء في الجنة ؟**	**2**	**35**	مكرر

Table 8: the final format of the Quran QA corpus.

السؤال	نص الآية	Translation	رقم الآية	رقم السورة	اسم السورة
بم افتتحت سورة الفاتحة	بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ	In the name of Allah, Most Gracious, Most Merciful.	1	1	الفاتحة
لمن الحمد	الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ	Praise be to Allah, the Cherisher and Sustainer of the worlds;	2	1	الفاتحة
ماهي صفات الله التي وردت في سورة الفاتحة	الرَّحْمَنِ الرَّحِيمِ	Most Gracious, Most Merciful	3	1	الفاتحة

Building Question Answering Prototype for the Holy Quran:

We built our program using java standard edition version 8. The system involves the following steps:

1. The input string (user query) is tokenized into a string for each word.

The code for splitting and counting the numbers of words in the input string :

```
String[] tokens = splitString.split(delims);
 int tokenCount = tokens.length;
```

2. Use tokenized strings as keywords for search.

```
String[] words;
   words = new String[tokenCount];
```

3. The tokenized words are stored in an array of string

```
for (int j = 0; j < tokenCount; j++)
        {
          words[j] = tokens[j];
         }
```

4. Then the tokenized words are compared with the database (after making the connection and storing their fields in variables), each one with the all words in the record; if any word matches then display the record.

The code for this operation is:

```
for (int i = 0; i < tokenCount1; i++)
{
  for (int j = 0;j < tokenCount; j++)
   if ((Num1[i]).equals(words[j]))
  flag = true;
```

The system works as follows:

1. Accept a question as in figure 1.

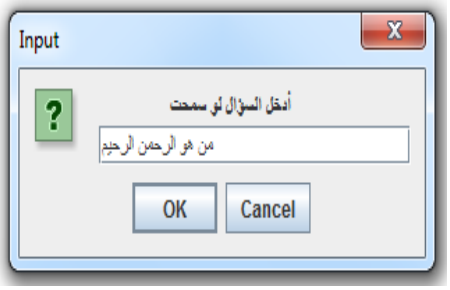

Figure 1: the question input message

Then the above processing steps must be done as follows:

2. The string is tokenized into words.
3. Using the tokenized keywords for search.
4. Then the tokenized words are compared with the database, each one with the all words in each record (verse). If any word matches then display the record. This may lead to unnecessary results as in the following figure 2 (only part of result).
5. Also at this stage if we use Modern Standard Arabic in our question which is the standard language in use today we may not find direct

matches to Quran verses (we can use the English translation and the question as an evidence) because verses use diacritics. If we search about: الحي القيوم there is no result as in Figure 3 and 4.

Figure 2: results without removing diacritics and stop words

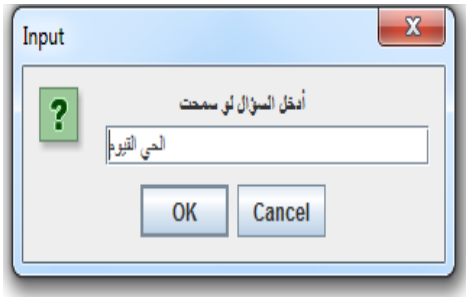

Figure 3: The question input

Figure 4: The relevant answer to the above question

6. To solve these two problems we remove some diacritics such as: (ّ، ْ، ٌ، ٍ، ً، ُ، ِ، َ) and some stop words such as : (ما ،من ،كيف، متى); this enhance the system results such as in Figure 5 and 6.

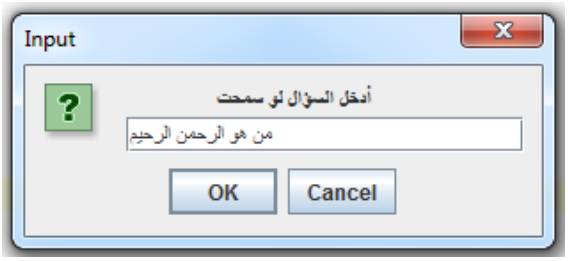

Figure 5: Another input question

7. We notice that the number of results becomes 5 instead of 151. Also if we removes punctuations from verses as well we find results instead of no results as in Figure 7 and 8.

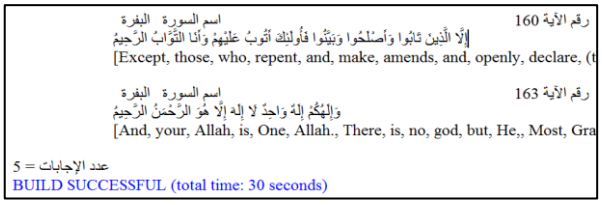

Figure 6: The answer after removing stop words and diacritics

Figure 7: the last question

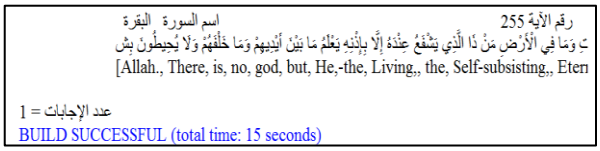

Figure 8 : the result after removing diacritics and stop words and punctuation.

A number of observations can be summarized as follows:

- Tokenization is a very important stage because we need to make comparisons between each word in the question and each word in the verses.
- Removing stop words and diacritics increases the efficacy and efficiency of the system.
- The link between the answers and any other fields in the database table can decrease answers; for this we remove the question as evidence in search, but we can add another field to increase the efficiency such as أسباب النزول.
- The system needs a lot of memory because it uses a lot of strings and arrays of strings; the system stores the entire Quran QA database in memory.
- Finally, we summarize the results in table 8 and figure 9 which show the increase of the matching answers after removing diacritics (little) and stop words (more).

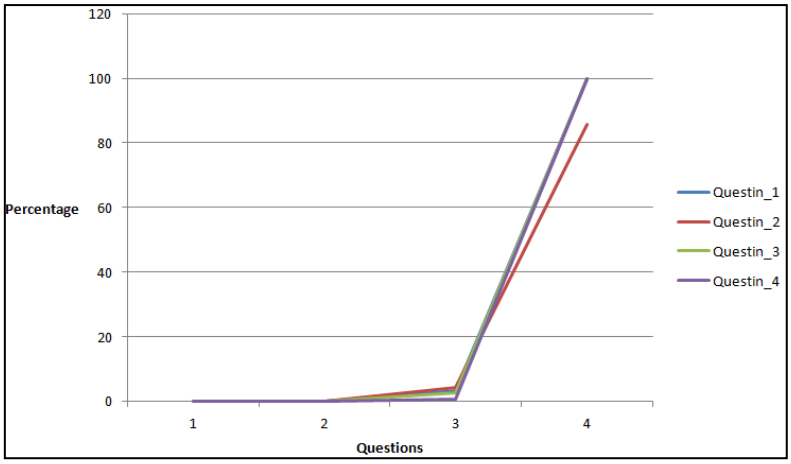

Figure 9: Increase in matching answers

Table 8: comparison of results after removing stop words and diacritics

question	Not removing anything	removing stop word only	removing diacritics	removing stop words and diacritics
	Number of matches results and its percentages			
من هو الرحمن الرحيم	0 = 0%	0 = 0%	151 =3.3%	5 =100%
من هم بني اسرائيل	0 = 0%	0 = 0%	163=4.2%	7=85.7%
ما هو الصيام	0 = 0%	0 = 0%	70=2.85	2=100%
من هو الحي القيوم	0 = 0%	0 = 0%	147=0.68%	1=100%

CONCLUSIONS

Our Question Answering system for Holy Quran applied for only two chapters: Al-Bagrah and Al-Fatihah (البقرة والفاتحة). We noticed that the system enhances results after removing stop words and diacritics. It can applied for whole Holy Quran as a basis for a Question Answering system.

We recommended: building a full corpus for the full set of Holy Quran verses, and build another prototype to make comparison to find what is the best model. The prototype can be made available online for more users. In addition, we can use our dataset for AI techniques instead of rule based techniques to add more efficiency and/or accuracy to the system. Finally, we can also add more evidence such as Hadith and Asbab Alnizoul (أسباب النزول) for more documentation and precision in Answers to Questions.

REFERENCES

[1] P. Kumar, S. Kashyap, A. Mittal, and S. Gupta, "A Fully Automatic Question-Answering System for Intelligent Search in E-Learn...," , International Journal on E-Learning, 4(1), 149-166. Norfolk, VA: AACE. Retrieved June 25, 2014 , Int. J., vol. 4, pp. 149–166, 2005.

[2] S. J. Athenikos and H. Han, "Biomedical question answering: a survey.," Comput. Methods Programs Biomed., vol. 99, no. 1, pp. 1–24, Jul. 2010.

[3] M. a Bauer and D. Berleant, "Usability survey of biomedical question answering systems.," Hum. Genomics, vol. 6, p. 17, Jan. 2012.

[4] V. Gupta, "A Proposed Online Approach of English and Punjabi Question Answering," Int. J. Eng. Trends Technol., vol. 6, no. 5, pp. 292–295, 2013.

[5] Nafid Haque "A prototype framework for a Bangla question answering system using translation based on transliteration and table look-up as an interface for the medical domain April 2010 University of Malta," 2010.

[6] D. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal, "Probabilistic Question Answering on the Web," J. Am. Soc. Inf. Sci. Technol., vol. 56, pp. 571–583, 2005.

[7] A. Diekema, X. Liu, J. Chen, H. Wang, N. Mccracken, O. Yilmazel, and E. D. Liddy, "Question Answering : CNLP at the TREC-9 Question Answering Track," in Science And Technology, 2001.

[8] Li, X. & Roth, D., (2002). Learning question classifiers. In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), p.556-562.

[9] Kulick, S.,Ann Bies and Mohamed Maamouri, (2010). Consistent and Flexible Integration of Morphological Annotation in the Arabic Treebank. Proceedings of the Seventh International Conference

on Language Resources and Evaluation (LREC 2010), pp. 1–8.
[10] A. M. Ezzeldin, “The 13th International Arab Conference on Information Technology ACIT ’ 2012 Dec . 10-13 A Survey Of Arabic Question Answering : Challenges , Tasks , Approaches , Tools , And Future Trends,” pp. 409–414, 2012.
[11] Benajiba Y. and Rosso P, (2008). Arabic Named Entity Recognition using Conditional Random Fields. Proceedings of Workshop on HLT & NLP within the Arabic World, LREC’08, 2008 Al-Sabbagh, R. and Girju, R. (2012). YADAC: Yet Another Dialectal Arabic Corpus. LREC 2012.
[12] Mohammed F., Nasser K., Harb H., (1993). A knowledge based Arabic Question Answering system (AQAS), ACM SIGART Bulletin, pp. 21-33.
[13] B. Hammo and S. Lytinen, “QARAB: A Question Answering System to Support the Arabic Language,” ACL2002Computational Approaches to Semit. Lang., p. 11, 2002.
[14] Benajiba Y., Rosso P., Lyhyaoui A, (2007). Implementation of the ArabiQA Question Answering System's components, Proceedings of Workshop on Arabic Natural Language Processing, 2nd ICTIS Information Communication Technologies Internatioanl Symposium, Fez.