Islamic Application of Question Answering Systems: Comparative Study

Mohamed Adany Hamdelsayed\textsuperscript{1,3}, Ebtihal Mustafa Elamin Mohamed\textsuperscript{1}, MohamedAlmoayed TajAlsir Mohamed Saeed\textsuperscript{2}, Abakr Musa Ai Mhmoud\textsuperscript{1}, Edress Babiker Edress Mohamed\textsuperscript{1}, Maha Ali Mahmoud Ahmed Shamat\textsuperscript{1}, Eric Atwell\textsuperscript{4}

\textsuperscript{1}Computer Science and Information Technology Faculty, Sudan University of Science and Technology-SUST
\textsuperscript{2}Blue Nile University
\textsuperscript{3}Gabrah Academic College
\textsuperscript{4}School of Computing, Faculty of Engineering, University of Leeds, Leeds LS2 9JT, England

a_mohdn111@sustech.edu, mohdn111@gmail.com

ISSN: 2231-8852

ABSTRACT

A question answering system is an information retrieval system that retrieves relevant short answers that match the question, instead of retrieving relevant full documents in a standard information retrieval system. In this study, we use three prototypes uses different resources for answers: MS Access database in prototype1, text files in prototype2 and prototype3, for storing results, each of which is one verse (Ayah) from two chapters of the Holy Quran: AlBagarah and AlFati'hah. Prototype1 uses an MS Access database and rule-based system for retrieving matching answers after removing stop words and diacritics from the input Arabic question. Prototype2 and Prototype3 use indexed text files and the Lucene search engine and its techniques adapted to deal with our Arabic text corpus. These systems accept a question in natural language form, then find the relevant answers from the text files. We evaluated the Prototypes with thirty questions about the Quran, chosen randomly. Quran scholars from Gabrah educational college assessed the answers. There are many variations in results depending on many factors such as stemming, stop words use. We conclude by comparing each system’s benefits and drawbacks.

Keywords: Corpus, Quran, Arabic, diacritics, stop words
1. Introduction

A question answering system is defined as: "an automated approach to retrieve correct responses to the questions asked by a human in natural language" (Dwivedi & Singh, 2013). Many factors and dimensions control the question answering systems (Hirschman & Gaizauskas, 2002): applications, users, question types, answer types, evaluation, presentation.

It is consist of three main modules (A. Allam & Haggag, 2012): Query Processing Module, Document Processing Module and Answer Processing Module. The first module is Query Processing Module with much processing: accepting the question in natural language, then many preprocessing sequences such as: defining the type of question, removing diacritics, stopping words and some symbols, tokenizing the question words, stemming words. Also, we can use some techniques to add new words for search: semantically by using dictionaries, corpus, thesaurus and syntactically by adding many patterns for the root word to find the singular from plural and broken plural and vice versa. The second module uses the keywords extracted from the query to find relevant documents that containing these words, and ranking these documents. The third modules are used to extract answers from that retrieved documents if there is an answer found.

Question answering system also has two main domains (A. M. N. Allam & Haggag, 2012): open and closed domains. Also (Adany & Atwell, 2015) state that question answering system are applied in many areas such as education, biomedicine and linguistics.

2. Related work

Many studies were done in Arabic language and Holy Quran, these studies vary from one area to another. In the following part, we review some of these studies.

2.1 Arabic Language Studies

One of the important studies in Arabic stemmer is Khoja Stemmer designed by Khoja (Khoja). Depending on this stemmer many studies were done such as (Al-Kabi, 2013) and (Fareed, Mousa, & Elsisi, 2013). The first one identified that Khoja stemmer has some flaws came from missing some pattern that leads to some errors, these patterns added to the stemmer which result in 5% improvement in the accuracy to Stemmer. Also, the second study was done to use Arabic WordNet AWN and Arabic stemmer to enhance semantic question answering system; the system uses two levels of expansion by AWN and Stemmer, they found that when it uses stemmer only increase the accuracy, but when it uses the stemmer and AWN the accuracy is decreased.

A named entity recognition NER is one of the natural language processing done to enhance performance in some applications such as information retrieval, question answering system and machine translation. The study of (Benajiba & Rosso, 2008) uses an Arabic Named Entity Recognition system (ANERSys) to enhance results. The study uses some additional information such: Base Phrase Chunks, Part-Of-Speech tags and changing the probabilistic model. The results improved by 10 points.
QARAB is presented by (Hammo, Abuleil, Lytinen, & Evens, 2004), which is question answering system that accepts question in natural language to find the relevant passage as an answer. The answers extracted from Al-Raya newspaper published in Qatar, the system uses two strategies word index and root index, the system tested by native Arabic user entered 112 questions the result evaluated as following: recall 79.6% and precision 100%.

Another important study in the Arabic language is an ontology-based formalism for the Arabic language using verbs and their derivatives model proposed by Belkredim & Sebai (2009) to build some patterns from its root nouns and verbs that can be used in Question Answering Systems, Information Retrieval, Machine Translation and Natural Language Generation. Also, they found that derivation from verbs is very easily from derivation from verbal nouns. Some difficulties appeared such as lack of logical structure, their complexity and different variations.

2.2 Holy Quran Studies

There is a shortage in the computation of Holy Quran studies, but there are some researches done in this area. The following paragraphs describe some of these studies.

Leeds University designed Quranic Arabic Dependency Treebank (QADT) which is Quranic Arabic Corpus part available online source for Arabic and Holy Quran for researchers and collaborative users to develop to enhance its accuracy (Dukes, Atwell, & Sharaf, 2010). The system displays the tree bank and illustrates the choice of syntactic representation, and highlights key parts of the annotation guidelines. By adopting i’rab they can enrich corpus. The main problem of QADT is dealing with semantic, but this is another component of this project.

One of the Quranic studies is done by (Adany & Atwell, 2015) for building a corpus for Holy Quran (AlBagarah and AFatihah chapters) question answering system; the corpus builds In database and text files to use by question answering system applications. Also, they build question answering system application using rule-based techniques. When the system was applied, they noticed that increase in matching when the diacritics were removed from questions and results.

Another Quranic study in semantic (Hashem A. J. Shmaisani, 2014) proposed approaches that covered some parts of this area in Arabic language for organising textual entities and their relationships. They used the Holy Quran to test this method. Also they used ontology based on the relation extractor model. In this study, there are two contributions Arabic relation extractor and Arabic ontology for the Holy Quran. The main drawbacks the system designed for only Holy Quran.

QurSim proposed by (Sharaf & Atwell, 2012) which is a corpus of Quranic text that similar and relatedness in verses, this corpus is a useful resource for evaluating relatedness and similarity and has two key characteristics: high quality of relatedness and has more than 7600 pairs of related verses. Only 33% has shared the same roots. Also, the authors explain that the verses vary in their text length which leads to two main problems: the long verses may contain many topics, and hence, the short verses may have only one topic.
3. Building prototype for question answering system

In this part, we will build two programs and then make comparisons between the two prototypes and discuss the results. Also, we will try to enhance one of these two programs and make another comparison and discuss these results.

3.1. First prototype:

Adany & Atwell (2015) explained the prototype very well, but in general, the system accepts the question and make many preprocessing such as: tokenizing the question, removing stopping words, some symbols and diacritics from the query and the verses to enhance the result. In general, the system has advantages: give more accurate answers than the other prototypes because some words repeated in many verses in the Holy Quran such as الله، ما، في، على. On the other hand, the system does not use any semantic methods such as stemmer to increase results.

3.2. The second prototype:

The second program depends on indexing, which was used by Lucene search engine. The system searches through our corpus which contains the following files (Holy Quran: Al-Baqarah and Al-Fatihah chapters, verses of each chapter).

Searches using Lucene search engine packets, which is used reverse indexing which indexes words in files and gives each word index to tag the words and files that contain these words the following table (Table 1) provide examples of the indexing process in the second prototype.

<table>
<thead>
<tr>
<th>Word</th>
<th>Documents contain the words</th>
</tr>
</thead>
<tbody>
<tr>
<td>الرحمن</td>
<td>1-1.txt, 1-3.txt, 2-163.txt</td>
</tr>
<tr>
<td>الرحمٍ</td>
<td>1-1.txt, 1-3.txt, 2-37.txt, 2-128.txt, 2-160.txt, 2-163.txt, 2-182.txt</td>
</tr>
<tr>
<td>الحج</td>
<td>2-150.txt, 2-189.txt, 2-196.txt, 2-197.txt</td>
</tr>
<tr>
<td>الشيطان</td>
<td>2-36.txt, 2-168.txt, 2-208.txt, 2-268.txt, 2-275.txt</td>
</tr>
</tbody>
</table>

Also, the following figure (Fig1) shows part of the indexing result of verses files.

Figure 1: The indexing of Holy Quran verses (Al-Baqarah and Al-Fatihah)
Secondly, the user inserts its query through the search program as shown in figure (Fig 2).

![Inserting query in the second prototype](image)

Figure 2: Inserting query in the second prototype

Then the search program makes many processes for the query, these processes are:

1. Removing stopping words such as (كيف، هو، من).
2. Stemming the original words that the query uses to search.
3. Displaying these words after stemming, as in figure (Fig 3):

![The main words for the query after stemming some letters](image)

Figure 3: The main words for the query after stemming some letters

4. Then the system search index to find these words, if it found, then the system reads the file that corresponds to these words and display its contents and as in figures (Fig 4):

![The result of the query in figure 11](image)

Figure 4: The result of the query in figure 11

3.3. The third prototype

It is the same The third one builds by adding new patterns for irregular plural (broken plural in Arabic language) depending on the Lucene search engine (second prototype) applying in Quran verses corpus, in this part we use these patterns to increase the precision of results, because the system return each broken plural to its singular and use all the words that
produced as a new word for searching. We find only one word in AlBagarah and Elfatihah chapters, which is تصرف. When we use the word تصرف which is singular and appeared in AlBagarah chapter verses 164. We use the query: هل للرياح تصرف in all prototypes we find the following results as in Fig 5:

![Figure 5: entering the query "هل للرياح تصرف"](image)

1. The first prototype result as in Fig 6.

![Figure 6: The first prototype result for the query "هل للرياح تصرف"](image)

2. The second prototype result as in Fig 7, but in the third, there is a technical problem:

![Figure 7: The second prototype result for the query "هل للرياح تصرف"](image)

However, when we use the plural تصاريف in the query we find the following results:

3. The first prototype result as in figure Fig8:
Figure 8: The second prototype result for the query "هل للرياح تصرف".

4. The second prototype result as in figure Fig 9:

The second prototype result as in figure Fig 9:

5. The third prototype result:

When we enter the question the system gives two screens, the first for the query after stemming and the second for the unique patterns.

6. The system first stems the main word in the query as in figure Fig 10:

Figure 10: The third prototype result for the query "هل للرياح تصرف".

7. Then the system checks each word and matching with the pattern that added to the prototypes and displays the new words as in figure Fig 11:

Figure 11: The third prototype result for the query "هل للرياح تصرف" after adding new words.

Figure 12: The second prototype result for the query "هل للرياح تصرف".
Comparison between prototypes (one, two and three):

When we are comparing the first prototype with the second and third, we noticed that:

8. First, the first prototype may give the exact matching, and it may be a correct answer as in the following question (من هم المغضوب عليهم), figures Fig 13 represent this, and the result of query represented as in figure Fig14 and Fig15:

![Input](image)

Figure 13: Question that the first prototype gives the best result

Your Question is:

من هم المغضوب عليهم

[The, way, of, those, on, whom, Thou, hast, besto

عدد الإجابات = 1

BUILD SUCCESSFUL (total time: 8 seconds)

Figure 14: the matching result of the above question for the first prototype

![Figure 13](image)

![Figure 14](image)

![Figure 15](image)

Figure 15: The matching result of the above question for the second prototype
From these results, we noticed that the first system gave only one matching result because it has many stopping words that decreased the precision of results. The first prototype removes the words: \( \text{هم} \), \( \text{عليهم} \) from the query while the second prototype does not eliminate these words; which gave some wrong results (e.g. 37 and 38) and one matching. Secondly, the first prototypes use the database to store Quran verses where the second prototype uses indexing files, and each one contains one verse. This makes the second prototype faster than the first because the systems scan only the files that contain the relevant words, where the first scans the entire database to find the relevant documents. Also the second (Ebtihal Mustafa AlAmin, 2015) and the third prototypes (Mahmoud, 2015) (Ebtihal Mustafa AlAmin, 2015) are distinguished from the first prototypes by using some patterns that change the plural in broken plural to singular, also the third one have three patterns where the first one have only one pattern. These patterns help the system to check if the word in broken plural in the weight of تفاعيل or فعاعيل or فاعاعيل then it returns to its singular. However, in the two chapters, we used as corpus AlFatihah and AlBagarah chapters we find only one verse matching these rules as in the above figures.

4. Experiments, Results and Discussions

4.1. Experiments and results
In this part, we discuss our experiments, which it judged by Scholars from Jabrah college. We use 30 question from Corpus designed by (Adany & Atwell, 2015) applied in the three prototypes. The following tables (2,3,4) and figures (16,17), explain the results.

<table>
<thead>
<tr>
<th>Question Number</th>
<th>Prototype (1)</th>
<th>Prototype (2)</th>
<th>Prototype (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Right answer</td>
<td>Matching</td>
<td>Right answer</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>76</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>91</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>110</td>
<td>1</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>111</td>
<td>1</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>119</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>121</td>
<td>1</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>132</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>
We used table 1 to generate table 3 and the charts in Figures 16 and 17.

Table 3: Percentages of right and wrong answers

<table>
<thead>
<tr>
<th></th>
<th>Prototype (1)</th>
<th>Prototype (2)</th>
<th>Prototype (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Percentage</td>
<td>Number</td>
</tr>
<tr>
<td>Total of question</td>
<td>30</td>
<td>100%</td>
<td>30</td>
</tr>
<tr>
<td>Right answers</td>
<td>21</td>
<td>70%</td>
<td>24</td>
</tr>
<tr>
<td>Wrong answers</td>
<td>9</td>
<td>30%</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 16: displays percentages of right and wrong answer for prototype 1
We notice that prototypes 2 and 3 gave a correct answer because they used stemmer which increases the accuracy.

Table 4: displays percentages of right and wrong answer in matching

<table>
<thead>
<tr>
<th></th>
<th>Prototype (1)</th>
<th>Prototype (2)</th>
<th>Prototype (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Percentage</td>
<td>Number</td>
</tr>
<tr>
<td>Right answers</td>
<td>21</td>
<td>10.88%</td>
<td>24</td>
</tr>
<tr>
<td>Wrong matching</td>
<td>193</td>
<td>89.11%</td>
<td>2270</td>
</tr>
</tbody>
</table>

Also, we notice that in matching there an increasing in the precision results and vice versa the stemming decrease the precision of results.

4.2. Discussions

Depending on the results we find that results showed the following:
1. Using patterns increase the results rapidly (Recall).
2. Using patterns increase the right results (Precision).
3. In the Holy Quran using of special pattern broken plural (the pattern of تفاعيل) is rarely used.
4. Increase the number of query words, leads to increasing the number of matching words rapidly, many answers may be wrong.

5. Conclusion and recommendation

5.1. Conclusion

Quranic and Arabic language studies begin to take attention by scientist, especially in information retrieval and question answering system, for this we need to unify efforts to go to the maturity and filling gaps in the lack of researches in this area. Also, we need to use many technologies in question answering system to reach high accuracy and matching. Also, we the efforts of Arabic language and Islamic scholars to help scientist to go on.
5.2. Recommendation

We need to merge these prototypes in only one, also we need to build a general model for the Arabic question answering system to help scientists and linguistics, and last we need to create stemmer by choosing the suitable pattern to search.

REFERENCES