International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Evaluation Corpus for Restricted-Domain Question-Answering Systems for the Holy Quran

Bothaina Hamoud[1], Eric Atwell[2]

[1]College of Computer Science and Information Technology, Sudan University of Science and Technology, Khartoum, Sudan
[1]Preparatory Year, Umm Al-Qura University, Makkah, Saudi Arabia

[2]School of Computing, Faculty of Engineering, University of Leeds, Leeds LS2.9JT, England

Abstract: ***This paper presents the compilation of a corpus of question-answer pairs for the holy Quran. The corpus has been manually collected from a wide range of sources, and designed to represent the Quran Arabic-English Question and Answer Corpus (QAEQ&AC). QAEQ&AC is a written, bilingual corpus, which comprises Arabic and English text. First, question-answer pairs have been collected from several trusted expert sources. Then the data were merged and cleaned using Microsoft Excel. After that data were converted to the format that suitable for mining tools, where we have created a comma-separated value (CSV) file format. The corpus obtained consists of more than 1500 question-answer pairs which is nearly 50.000 words, divided over Arabic and English languages. It includes different question types such as what, when, why, etc., and different answer length. We anticipate that the current and subsequent versions of our corpus will be a valuable evaluation resource for computational linguists investigating Quran question and answer; it might be used as a gold standard in researches, that dealing with natural language processing, information retrieval, artificial intelligence. The corpus can be subjected to an annotation to derive linguistic information such as morphological, syntactic, semantic, and lexical information.***

Keywords: QAEQ&AC, Quran, corpus, data, question-answer pairs, dataset

1. Introduction

The creation of corpus is essential for implementing natural language processing solutions based on machine learning [1]. There is wider interest in evaluation of restricted-domain QA system, in contrast to open-domain QA system. A main characteristic of question answering in restricted domains is the integration of domain-specific information that is either developed for question answering or for other purposes. We chose the Quran domain to develop a question and answer corpus for the reason that: The Quran is a vital book as it is a core text which contributes to the lives of millions of people today. Quran texts are interesting and challenging for evaluation researchers and Language resources. It can also have a great effect on society, helping the general public to access and understand religious texts [2]. Nowadays, there is also a wide range of research projects focused on the Quranic domain. There is a need for more research on this book so as to make more sources available and accessible. Question and answer corpus for the holy Quran is a valuable resource for the Question Answering research communities and are probably attract more useful research. It can be used for the evaluation of question answering systems or any other application where questions and answers are needed. Research about the holy Quran has been receiving increasingly intensive attention and is an active and remarkable area.

There were no adequate existing resources specifically designed for Quran question and answer corpus. Existing Quran question and answer sources are scattered between different webpages, each of these sources has its own format and style. There is a need to create a unified dataset for Quran questions and answers, to be used in testing and evaluation in applications for Quran search and question-answering system.

1.1 Corpus

"Corpus is a collection of writings, conversations, speeches, etc., that people use to study and describe a language "(Merriam-Webster). The corpus can be in written language or spoken language or both. Text corpus is a structured set of a large text stored in many files that share the same parameters, usually, the language, the structure and the encoding. . Some popular corpora are British National Corpus (BNC), COBUILD/Birmingham Corpus, and IBM/Lancaster Spoken English Corpus. Corpus is used in many task such as hypothesis testing, statistical analysis, checking occurrences, and in the study of historical documents. Monolingual corpora represent in only one language; bilingual corpora represent in two languages while multilingual corpora represent in multiple languages. To make the corpora more useful they are often undergoing to an annotation operation, such as morphological, Statistical and semantics analysis. Corpora are used as essential knowledge base in corpus linguistics. Corpus may be open or closed. The analysis and processing of different types of corpora are the theme of a lot of work in computational linguistics, machine translation and speech recognition such as part-of-speech tagging to signal the lemma form of the word.

The rest of this paper is organized as follows: in Section 2 we introduce the current research related to corpus development, section 3 presents the data collection and resources employed to build the corpus, Section 4 outlines the preparation of the Quran questions and answers corpus. In Section 5 Some results are shown, and finally, we conclude the paper and give directions to future work in Section 6

2. Related Work

In recent years, large amount of question and answer corpora have been searched and developed:

In [3] a Yahoo-based Contrastive Corpus of Questions and Answers (YCCQA) was developed. It has been compiled using material downloaded from http://answers.yahoo.com/. The site offers an environment in which users post questions and answers. YCCQA contains about 90,000 question-answer pairs which make 29 million words of text. All the material collected has been posted by users between 2006 and 2009. Language material, consisting of questions and the accompanying answers, has been extracted for English, French, German, and Spanish. In [4] the KM database corpus has been developed, which is composed of question-answer pairs obtained from Knowledge Master (1999). Each of the pairs in KM represents a trivia question and its corresponding answer, such as the ones used in the trivia card game.

In [5] a large training corpus consisting of question-answer pairs of a broad lexical coverage has been built. One million question-answer pairs from FAQ pages has been collected, this corpus is used to train various statistical models employed by their QA system in query analysis and answer extraction modules. Their system was intended to be applied to non-factoid questions. In [6] a collection of approximately 30,000 question-answer pairs were developed from the internet, they have been obtained from more than 270 Frequently Asked Question (FAQ) files on various subjects. The obtained FAQ were used by their project FAQ Finder.

In [7] a corpus of English question-answer pairs has been developed. The corpus consists of more than 70,000 samples. Each of these samples contains information that relates a question with its answer in four different contexts: exact match, sentence, paragraph and document. The developers claimed that their corpus suited to train on every stage of machine learning based QA system: question classification, information retrieval, answer extraction and answer validation. In [8] the Insurance QA Corpus has been created, it is a question answering corpus in insurance domain. It contains questions and answers collected from the website Insurance Library. The content of this corpus consists of questions from real world users, the answers were composed by professionals with deep domain knowledge.

3. Data Collection and Corpus Sources

Data collection is the systematic approach to gathering information from a variety of sources to get a complete and accurate data of an area of interest. Accurate data collection is essential to maintaining the integrity of research, and ensuring quality assurance. The process of collecting data can be relatively simple according to the type of tools used to collect the data. There are several tools that can be used to collect data. The data collection tools should be good enough to collect useful data in order to have better evaluations for the research. Selecting specific tools depend on the nature of the task, as well as the type of the required data.

The first step of building a corpus is to think about the resources of data. Question-answering research in the religious domain involves ethical concerns. Answering questions about Islamic beliefs requires great care to give an accurate answer for a given question, which can be accepted religiously and universally. For example, the answers should be stated as mentioned in the Quran as well as in Hadith books. Collecting question and answer from authoritative and credible sources is an important issue while low accuracies or wrong answer is not acceptable in the religious field and especially in the Quran domain.

In this paper we used four sources to collect Quran questions and answers corpus. Since the Internet is rich with data and easily accessible, the first and main source for our corpus collection is a web-based tool created by a group of scholars in the Islamic field. The second source is the Quran book. The third source is from some Muslims who came to the Holy Mosque in Makkah and got answers for their questions from a group of scholar in the Holy Mosque. And the fourth source is from a survey of previous research on Quran question-answering.

Frequently Asked Questions (FAQs) from the above-mentioned sources has been used to collect questions-answers pairs. We merge different data subsets from different sources to comprise the Quran question and answer dataset. We started by searching for web resources, and selected the following:

- The website turntoislam [9]: a popular website to learn about Islam, containing a huge library which consists of many topics about Islam in many languages. It includes questions along with their answers.
- Islamic Knowledge/Come towards Islam [10]: it contains monthly archives covering many topics concerned with Islam such as questions and answers about Quran, understanding Islam, Islamic facts, discussion of Quran chapters, teachings of the prophet Muhammad, haram (forbidden) food and drinks, Ramadan, women in Islam and many more. Its archives cover from March 2011 till February 2015
- The All-Quran [11] web site aims to have the holy Quran available to everyone by providing an easy way of audio streaming for a variety of Quran reciters and audio translations. It contains a tab for Islamic FAQ, in addition to annotated text of the Quran. It is a commercial-free website.
- The Siasat daily web site [12] also provides questions and answers about the holy Quran; it is written in three languages, English, Urdu, and Hindi.
- SULTAN ISLAMIC LINKS [13] Discover Islam, Muslim People, Holy Quran and Islamic Religions. It has the resource “you ask and the Quran answers”
- Islamic question and answer [14] which contains many question about Islam written in 13 languages.
- A set of questions along with their answer were gathered from some forum such as the official forum for Sheikh Dr. Mohammad Al Arifi [15], and hussaballa forum [16].

Most of the content of the web based corpus consists of questions from real world users, the answers were given by professionals with deep domain knowledge.

We did not rely exclusively on web sources. A set of questions and answers were gathered from some Muslims who came to the Holy Mosque in Mecca, and posted their questions to a group in the Holy Mosque who are leaders in the field of Islamic studies, this group gives their expert answers to questions posted by Muslims who come to the Holy Mosque. Beside that we included small test sets of questions and answers from previous research [17], [18], [19], [20].

There are some questions along with their answers stated in Quran book, such as the questions that were directed to Prophet from different denominations - Muslims, Jews and Christians. These questions were taken directly along with their answers from the holy Quran book. For example :) يسألونك عن الأهلة قل هي مواقيت للناس والحج...) (يسألونك عن الخمر والميسر قل فيهما إثم كبير ومنافع للناس وإثمهما أكبر من نفعهما ...). وهنالك عدد من الأسئلة الاخرى مثل السؤال عن اصحاب الكهف وذي القرنين والروح...الخ such questions and their answers were subjected to reformulation, table 1.shows some examples. Furthermore we extracted a set of questions and answers from some chapters of the Quran book, table 2 shows some examples. To reformulate these questions and their answers optimally and add some explanation to it we used some interpretation reliable books such as Almokhtsar in the interpretation of the Quran [21], Tafsir Center for Quranic Studies.

Table 1: Example, question and answer directed to Prophet

The location of the question and answer as stated in the Quran book	*Question after reformulation*	*Answer with some explanation from Tafsir books*
وَيَسْأَلُونَكَ عَنِ الرُّوحِ قُلِ الرُّوحُ مِنْ أَمْرِ رَبِّي وَمَا أُوتِيتُم مِّن الْعِلْمِ إِلاَّ قَلِيلا (الإسراء:85)	ما هي حقيقة الروح؟	لا يعلم حقيقة الروح إلا الله فهي سرٌّ من أسرار الله لم يطلع عليها عباده
وَيَسْأَلُونَكَ عَن ذِي الْقَرْنَيْنِ قُلْ سَأَتْلُو عَلَيْكُم مِّنْهُ ذِكْرًا إِنَّا مَكَّنَّا لَهُ فِي الأَرْضِ وَآتَيْنَاهُ مِن كُلِّ شَيْءٍ سَبَبًا (الكهف 83, 84)	من هو ذو القرنين؟	هو رجل طاف في الارض كلها وهو مثال للملك الصالح الذي آتاه الله ملكا فسخره في الدعوة لله تعالى وتعبيد الناس لهذا الدين
يَسْأَلُونَكَ عَنِ السَّاعَةِ أَيَّانَ مُرْسَاهَا ۖ قُلْ إِنَّمَا عِلْمُهَا عِنْدَ رَبِّي ۖ لَا يُجَلِّيهَا لِوَقْتِهَا إِلَّا هُوَ ۚ ثَقُلَتْ فِي السَّمَاوَاتِ وَالْأَرْضِ ۚ لَا تَأْتِيكُمْ إِلَّا بَغْتَةً ۗ يَسْأَلُونَكَ كَأَنَّكَ حَفِيٌّ عَنْهَا ۖ قُلْ إِنَّمَا عِلْمُهَا عِنْدَ اللَّهِ وَلَٰكِنَّ أَكْثَرَ النَّاسِ لَا يَعْلَمُونَ (الأعراف:187)	متى تقوم الساعة	علمها عند الله .وحده, لا يظهرها عند وقتها المقدر لها الا هو, خفى امر ظهورها على اهل السموات و اهل الارض, و لا تأتي الا فجأة

Table 2: Example, question and answer extracted from Quran book

The location in the Quran book, where the question and answer were extracted	*Extracted Question*	*Extracted answer*
وَإِذْ قَالَ إِبْرَاهِيمُ لِأَبِيهِ آزَرَ ..... (الانعام:74)	ما اسم والد سيدنا إبراهيم عليه السلام	آزر
وَيَسْأَلُونَكَ عَن ذِي الْقَرْنَيْنِ قُلْ سَأَتْلُو عَلَيْكُم مِّنْهُ ذِكْرًا إِنَّا مَكَّنَّا لَهُ فِي الأَرْضِ وَآتَيْنَاهُ مِن كُلِّ شَيْءٍ سَبَبًا (الكهف 83, 84)	ما هي عقوبة السارق والسارقة	يأمر الله تعالى بقطع يد السارق والسارقة
شَهْرُ رَمَضَانَ الَّذِيَ أُنزِلَ فِيهِ الْقُرْآنُ..... (البقرة:185)	في اي شهر انزل القرآن الكريم	انزل القرآن في شهر رمضان

4. Data Preparation

Since there were no existing resources specifically designed for Quran Q&A we merged different data subsets for the purpose of a Quran question and answer corpus task. Data preparation covers all tasks to build the final dataset from the initial raw data. Tasks include table, record, and attribute selection; merging; formatting; cleaning; and transformation for modeling tools. These tasks can be performed multiple times, and not in a specific order [22].

4.1 Creating a data file

The obtained data must be processed or organized for analysis. For instance, these involve placing data into rows and columns in a table format for further analysis, such as within a spreadsheet. While our collected data sets have different format and style, a data file was created using Microsoft Excel 2010, and the collected data were merged into standard spreadsheet worksheet format, after that their style and format were unified.

4.2 Data cleaning

Data cleaning is the process of detecting and correcting corrupt or inaccurate data. Data quality problems exist in single data collections, such as files and databases, due to misspelled during data entry, missing information or invalid data. As for the data that is integrated from multiple sources, the need for data cleaning is largely increase due to the presence of heterogeneous data sources. Data cleaning is a phase in which noise and irrelevant data are removed from the collection; it refers to identifying incorrect, incomplete, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this noisy data. Unclean data can contain mistakes such as spelling or punctuation errors, incorrect data associated with a field, incomplete or outdated data, or even data that has been duplicated in the database. Data cleaning may also involve activities like harmonization and standardization of data. The goal of the data cleaning process is to maintain a meaningful data by removing elements that may hinder the analysis which affects the quality of the results, as the use of incorrect or inconsistent data may inevitably lead to false results.

When we import or paste the data from the Internet into Excel's worksheet, this may end up with unclean data, which will cause errors in the next phases. At this phase the original data were copied to another worksheet, and then cleaned by removing the inconsistent and the unwanted data from the dataset. To clean data we used two methods: the manual method, and the automatic method. Manual method was used to remove incorrect or incomplete data while the automatic method may be not useful in such a situation. In automatic method XLTools was used to remove extra spaces, line breaks, delete non-printable characters etc., while the manual method may take painstaking hours or may not guarantee to detect and remove all the mistakes. We also handled the blank cells as they can create problems if not treated in advance. The duplicated data were removed as well as spell checker were applied on the data set to correct spelling errors, where it is nothing that reduces the credibility of the work more than spelling error. While the search and replace in data cleaning is indispensable, we searched for inappropriate words and then replace them with more proportional. The text can also be changed to the uppercase, lowercase or other common capitalization, as well as the transformation of cell formats can be applied to change numbers to text and vice versa. Figure 4.1 bellow shows data cleaning using Excel tools (xlTools).

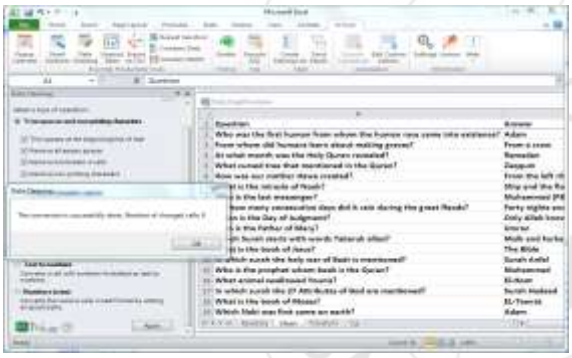

Figure 1: Data cleaning using Excel tools (xlTools)

4.3 Data transformation

Data transformation is the consolidation and transformation of data into forms suitable for mining. Data transformation allows the mapping of the data from its given format into the format expected by the modeling tool. Excel is a great tool to use when we need to take the data in a specific format, then processed to be converted into another format, then push the results to another tool to further processing. For example if we want to export Excel file to other applications that support the comma-separated value format (CSV) format, we can convert the worksheet first to CSV format and then export the .csv file to those programs. Excel works as a transfer tool for the data to be transferred from one system to another, where it supports many file formats.

In this phase the cleaned data were copied to another excel workbook, because CSV format does not support workbook containing several worksheets. After that the workbook was saved as a CSV format. CSV files are a common data exchange format that stores tabular data in plain text format, which can be read using any standard text editor. CSV is supported by many applications and therefore a large amount of tabular data can be transferred between these applications [23] Since the CSV files are plain text, this makes it easy to be understand by normal user or even by beginner, as well as it allow users to diagnose data problems easily.

5. Result

The QAEQ&AC is a written, bilingual corpus, which comprises Arabic and English text. In its current form, QAEQ&AC contains about 1500 question-answer pairs, which makes about 42500 words of text. As shown in Figures 2 and 3 these are divided over the Arabic and English languages as follows: 1000 Arabic question-answer pairs, they makes about 20.000 words, and 500 English question-answer pairs, which is about 22500 words . Contrary to other contrastive corpora, this version of QAEQ&AC does not contain parallel translated texts. All texts are originals. As a result, the subcomponents of the corpus can be used independently as language-specific corpus. The QAEQ&AC is not a general corpus; it is specifically dedicated in Quran domain, it includes different question types such as what, when, why, etc. and different answer length, the answer for a question can be a short text, or longer text for questions that need more explanation. Fig.4.1 shows the types of question that we used. QAEQ&AC will be a valuable evaluation resource for computational linguists investigating Quran question and answer; it might be used as a gold standard in researches, that dealing with natural language processing, information retrieval, artificial intelligence. The corpus can be subjected to an annotation to derive more linguistic information such as morphological, syntactic, semantic, and lexical information.

Most of questions and answers of this dataset were created from scholars works found on the internet and hence one would assume they were accurate and verified. However, the work is going on, to be subjected to further manual validation by Quranic scholars before making it available on the internet and incorporating them into wider applications. As well as to create another set of questions to be semantically identical but syntactically different from the questions already collected from different sources. At the end our data set file will list the original questions and the set of variants for that original along with their answers.

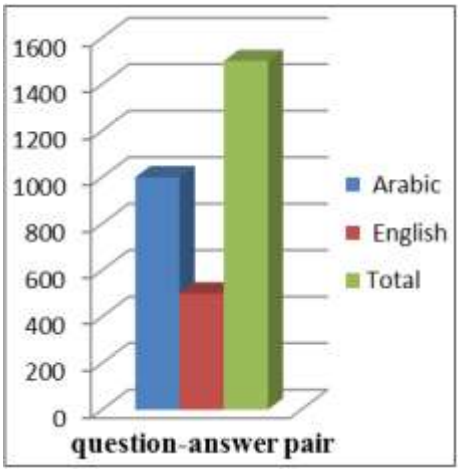

Figure 2: Questions and answers by language

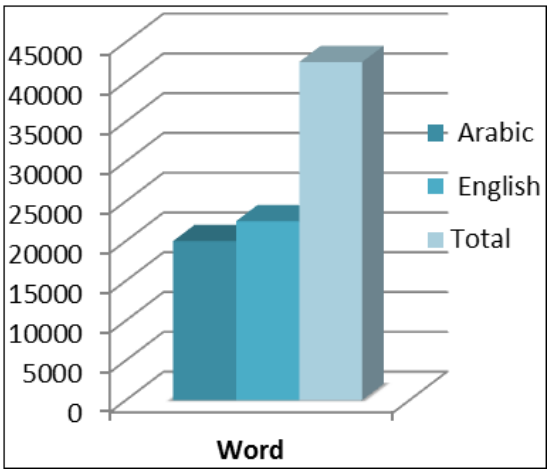

Figure 3: Word count by language

Table 3: question types

Domain identifier	*The meaning*
What	Indicates a question asking about things
Who	Indicates a question asking about persons
When	Indicates a question asking about time
Where	Indicates a question asking about a place.
How many	Indicates a question asking about countable things
Why	Indicates a question asking about cause
How	Indicates a question asking about the condition or situation
Others	All other questions

6. Conclusion and Future Work

Collecting data manually is a big challenge, as the automatic method has not always been successful in filtering inappropriate or unwanted data. We plan to develop this corpus and add more question-answer pairs. In conclusion, we believe that creating an integrated Quran question and answer corpus dataset is an important resource that we would like to apply in a task challenge, aimed at improving the state-of-the-art of online Quran question answering systems.

References

[1] E. Boldrini, S. Ferrández, R. Izquierdo, D. Tomás, & J. L. Vicedo, (2009, March). "A Parallel Corpus Labeled Using Open and Restricted Domain Ontologies". In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 346-356). Springer, Berlin, Heidelberg.

[2] E.S. Atwell, C. Brierley, and M. Sawalha, 2012. "Proceedings of LREC'2012 Workshop LRE-Rel: Language Resources and Evaluation for Religious Texts".

[3] H. De Smet, (2009). Yahoo-based Contrastive Corpus of Questions and Answers.

[4] D. Ravichandran, A. Ittycheriah, S. Roukos, "Automatic derivation of surface text patterns for a maximum entropy based question answering system". In: Proceedings of the HLT-NAACL Conference (2003)

[5] R. Soricut, & E. Brill, "Automatic question answering using the web: Beyond the factoid. Information Retrieval", 9(2), 191-206.

[6] R. D. Burke, K. J. Hammond, V. Kulyukin, S. L. Lytinen, N. Tomuro, & S. Schoenberg, (1997). "Question answering from frequently asked question files: Experiences with the FAQ finder system". AI magazine, 18(2), 57-66

[7] D. Tomás, J. L. Vicedo, E. Bisbal, & L. Moreno, (2009). "TrainQA: a Training Corpus for Corpus-Based Question Answering Systems". Polibits, (40), 5-11.

[8] M. Feng, B. Xiang, M. R. Glass, L. Wang, & B. Zhou,. Applying deep learning to answer selection: "A study and an open task. In Automatic Speech Recognition and Understanding (ASRU)", 2015 IEEE Workshop on (pp. 813-820). IEEE.

[9] Turn to Islam Community, "questions-on-Quran", http://turntoislam.com/community/threads/100-questions-on-quran.10052, time of access 30/5/2015

[10] Islamic Knowledge/Come towards Islam "questions and answers about Quran", https://islamicknowledge2all.wordpress.com/2011/10/30/question-and-answers-about-quran-3/

[11] All Quran "Islamic material/frequently asked questions(FAQ)", http://www.all-quran.com/islamic_material/frequently_asked_questions.html.

[12] The siasat daily "Questions and answers about the holy Quran", http://www.siasat.com/english/news/questions-answers-about-holy-quran?page=0%2C0 ,30/5/2015

[13] Sultan;" Discover Islam and Muslim Beliefs. Learn about The Real Islam. Correct your information about Islam Religion" http://www.sultan.org/.

[14] Islam Question and answer. "question and answer about Islam". http://islamqa.info/en/

[15] Official forum for Sheikh Dr. Mohammad Al Arifi http://www.3refe.com/vb/

[16] Hassabala forum, Encyclopedia for religious questions and their answers (500 questions) http://hassabala.yoo7.com/t714-topic?highlight=500+%D3%C4%C7%E1.

[17] R. H. Gusmita, Y. Durachman, S. Harun, A. F. Firmansyah, H. T. Sukmana, & A. Suhaimi, (2014). "A rule-based question answering system on relevant documents of Indonesian Quran Translation" Proceedings of International Conference on Cyber and IT Service Management (CITSM). pp. 104-107.

[18] H. Abdelnasser, R. Mohamed, M. Ragab, A. Mohamed, B. Farouk, N. El-Makky, & M. Torki, (2014). "Al-Bayan: an Arabic question answering system for the holy Quran". Proceedings of the EMNLP Workshop on Arabic Natural Langauge Processing (ANLP), pages 57–64, Doja, Qatar.

[19] M. A. Hamdelsayed, & E. S. Atwell, (2016). "Islamic Applications of Automatic Question-Answering". Journal of Science and Technology, 17(2).51-57

[20] H. Shmeisania, S. Tartirb, A. Al-Na'ssaanc, & M. Najid, "Semantically Answering Questions from the Holy Quran". 2nd International Conference on Islamic Applications in Computer Science And Technology, 12-13 Oct 2014, Amman, Jordan

[21] Almokhtsar in the interpretation of the Quran. (المختصر في تفسير القرآن الكريم/ جماعة من علماء التفسير) https://islamhouse.com/ar/books/2795156/

[22] Pete Chapman, Julian Clinton Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer and Rüdiger Wirth "CRISP-DM 1.0. Step-by-step data mining guide"
[23] B. Hamoud, & E. S. Atwell, (2016). "Quran question and answer corpus for data mining with WEKA". In the Conference of Basic Sciences and Engineering Studies (SGCAC). pp. 211-216. IEEE.