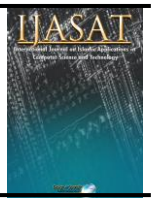

Challenges in the Islamic Question Answering Corpora

Sarah Alnefaie [1], Eric Atwell [2] and Mohammad Ammar Alsalka [3]

[1,2,3] University of Leeds, Leeds, UK
[2]King Abdulaziz University, Jeddah, Saudi Arabia
[1]scsaln@leeds.ac.uk, [2]e.s.atwell@leeds.ac.uk, [3]m.a.alsalka@leeds.ac.uk

Abstract
In the past, researchers in Islamic question-answering systems created their datasets to evaluate their systems due to the lack of the dataset, making it difficult to compare the performance of the systems. In the last three years, several studies have provided different datasets of Islamic questions and answers to the research community that can be used as a gold dataset to evaluate systems, a knowledge base for the system, and a training dataset to train pre-trained models. In this research, we review and explore the Islamic questions and answers datasets, study the percentage of its coverage of the Quran or Hadith, evaluate them using thirteen criteria, and identify their weaknesses which could serve as a basis for future research. We concluded that there is a limited number of Quran questions, and their answers are available in Arabic only. In addition, as far as we know, there is no Hadith Shareef questions and answers dataset.
***Keywords*:** Quran Question Answering Dataset, Islamic Corpus, Islamic Knowledge Base.

1. Introduction

The internet contains vast amounts of data written by various software, using assorted templates and formats. Gathering this data in one corpus is essential because it is the basic block in many computer science systems, such as classification and question-answering (QA). Corpus characteristics such as size and quality affect system performance. When the amount of data increases, the performance of the systems improves. For example, the performance of classification algorithms can be improved by increasing the data size more than enhancing the algorithms. On the contrary, when the data size is small in the QA system, the possibility of finding the answer decreases, so the system's performance decreases. Also, the most significant influencer affecting the classification's training phase is the data quality (Sapp et al., 2008).

A small number of studies created a corpus in the Islamic field. Recently, Mohammed et al. (2022) developed an English Islamic Articles dataset (EIAD) which contains ten thousand articles collected from websites. The target user of this dataset is new Muslims or people who want to convert to Islam. It is categorized into fifteen categories, each containing related folders, and each topic contains relevant articles. Nevertheless, it is a collection of articles from a limited number of sources, their topics are related only to new Muslims or who are willing to become Muslim, and it is unavailable.

Some current research tends to build an Islamic questions and answers (Q&A) corpus, which can be used for several purposes, such as a gold dataset in evaluating the QA systems, a knowledge base for the QA systems, or training the deep learning QA models. The availability of these datasets contributes to standardizing the QA systems' evaluation process, which facilitates comparing their performance. Contrary to what was happening before it was unavailable, each research had to build its assessment dataset. This paper reviews these

datasets, studying the extent to which it covers the Quran or Hadith and finds their flaws to open the door for development in the future. This paper only focuses on the existing datasets without considering other Islamic knowledge sources, such as websites and books, because they are not in a form you can use directly. The structure of the paper is as follows: The Islamic questions and answers corpora will be reviewed in the second section. At the same time, the methodology to assess these datasets will be presented in the third section. The conclusion will be discussed in the last section.

2. Literature Review

Several Q&A corpora have been developed for Islamic QA systems, which can be classified into question-answer pairs and question-passage-answer triplets as the following:

2.1 Question-Answer Pairs Datasets

This dataset consists of Q&A pairs used for systems that retrieve the answer to the question from the whole Islamic source, such as the Quran. Several datasets will present as follows:

Saeedi et al. (2014) developed the Quranjooy corpus, a collection of Quran Q&A pairs in Persian. The Q&A dataset was crawled from many collected websites. First, A list of found sites has been made. Then, the Quranic experts selected reliable sources. The total number of the collected Q&A was about 115000 that were put in the XML format. After that, the data cleaning process was applied, and the questions with grammatical or editing problems were excluded. The final corpus contained 6000 questions that cover many topics such as characteristics, place, personality, numerical, creature, behaviour, event, and date. Nevertheless, it does not cover all areas of the Quran and is written in Persian.

Further, Sheker et al. (2016) proposed a corpus of 1094 Q&A for the prayer Fatwas domain, which was gathered from the Ibn-Othaimeen Fatawas Book (Ibn Baz et al., 2003). The corpus specializes in prayer questions only and is small in size. Moreover, the answer is scholarly opinion and belief without any evidence from Quran or Hadith.

Hamdelsayed and Atwell (2016) and Adany (2017) recommend a Q&A dataset about Surat Al-Fatiha and Al-Baqarah Chapters in the Holy Quran. This corpus was used for two purposes in these QA prototypes: a knowledge base for the system and a gold standard evaluation dataset. Building this corpus involved several steps: they collected the Q&A form websites, extracted the questions by reading Quran text, and integrated these two sets into one file with different formats such as text, Access, and Microsoft Excel. They elicited 47 questions from Islamic websites, but the gathered text suffered from many problems, such as diacritics and English characters. Therefore, the text needed some cleaning process. They also extracted the appropriate questions from the verses. One question may have many answers. The number of questions generated using this approach is 215. These questions have been reviewed and validated by a scientist of Islamic scholars from Gabrah College. The combined dataset contained 263 questions. The data in the file was arranged according to the verse number to detect duplicate questions. The final file contains five columns: the question, the answer, the verse number, the sura name, and a column for Abdullah Yusuf Ali's English translation of the verse has been added to the file. However, the corpus covers only sura Al-Fatihah and sura Al-Baqarah, and the number of questions is very few.

Moreover, Hamoud and Atwell (2017) created a dataset of the Q&A about the Quran by going through the following steps: they manually gathered 1500 Q&A from Islamic websites, previous research, and experts in Mecca Holy Mosque and then merged them into one

knowledge base. Next, the corpus was cleaned and put in an appropriate format. Finally,it was used as knowledge base for the QA search tool. The eight websites used in this knowledge are turntoislam1, Islamic Knowledge/Come towards Islam2, All-Quran3, The Siasat4, SULTAN ISLAMIC LINKS5, Islamic question and answer6, Sheikh Dr. Mohammad Al Arifi official forum7, and the Sheikh Hussaballa forum8. They used some examples mentioned in the Gusmita et al. (2014), Abdelnasser et al. (2014), Shmeisani et al. (2014), and Hamdelsayed and Atwell (2016) studies. One of the services available in the Holy Mosque is the possibility to ask questions to Islamic experts. They communicated with Muslims who returned from Mecca to collect the questions that they asked the experts and the answers they got. However, this corpus has only 1500 Q&A pairs. No evidence from the Quran exists in a lot of the answers. Q&A in this corpus without reference.

NEAMAH and SAAD (2017) built a corpus of twelve questions to evaluate their Hadith QA system. They asked fifteen students from UKM universities to create these questions.
In addition, Abdi et al. (2020) constructed an Arabic corpus of 3825 Q&A pairs about Hadith from the Sahih al-Bukhari collection. Two human experts built the corpus manually by (1) reading all Hadiths in order, (2) generating questions for each Hadith, (3) removing the duplicated questions, and (4) linking each question with the correct Hadith. There are different types of questions, but the most significant percentage was for WH questions.
Maraoui et al. (2021) collected a dataset of 100 Q&A from online forums and native Arabic speakers. The distribution of this dataset is as follows: 13 questions about the Tafsir topic, 33 questions about the Hadith topic, and 54 questions about the Hadith narrator profile. Nevertheless, it is an unreliable and small corpus.

Furthermore, Munshi et al. (2022) created an Arabic Fatwa dataset with an 850K record. Each record contains questions, answers, Fatwa topic, and publication date. It aims to focus on social media questions and answers channels, Unlike the existing dataset focusing on the Quran and Hadith text. Usually, the users post the question, and a highly qualified expert answers the question in his opinion without evidence. They gathered questions from various countries, backgrounds, and accents. The resource can be classified as Government authentic sources such as Al-ifta-SA9and Dar-al-ifta-EG10 and untrusted websites such as fatawapedia 11, Islamweb12, Islamway13, binothaimeen 14, AlFawzan15, Islamqa16, and binbaz17. Some websites only contain articles, so they treat the title as a question and the article as an answer. Nevertheless, it is Fatwas without evidence from Quran or Hadith.

[1] http://turntoislam.com/community/threads/100-questions-on-quran.10052, time of access 30/5/2015
[2] https://islamicknowledge2all.wordpress.com/2011/10/30/question-and-answers-about-quran-3/
[3] http://www.allquran.com/islamic_material/frequently_asked_questions.html.
[4] http://www.siasat.com/english/news/questionsanswers-about-holy-quran?page=0%2C0 ,30/5/2015
[5] http://www.sultan.org/.
[6] http://islamqa.info/en/
[7] http://www.3refe.com/vb/
[8] http://hassabala.yoo7.com/t714-topic?highlight=500+%D3%C4%C7%E1.
[9] https://www.alifta.gov.sa
[10] https://www.dar-alifta.org/ar/Default.aspx?sec=fatwa&1&Home=1
[11] http://fatawapedia.com/
[12] https://www.islamweb.net/ar/
[13] https://ar.islamway.net/fatawa/source/
[14] https://binothaimeen.net/site
[15] https://www.alfawzan.af.org.sa
[16] https://islamqa.info/
[17] https://binbaz.org.sa/fatwas/kind/1

All previous datasets are unavailable to evaluate the Islamic QA system. To the best of our knowledge, only two datasets were recently developed for the Arabic Quran and are available to the public. These studies will be discussed as follows:

Alqahtani (2019) built a corpus of 2224 questions and their answers, which is Quran verse, to evaluate his QA system, but the available version only has 1224 Q&A pairs. It is called Annotated Corpus of Arabic Al-Quran Question and Answer (AQQAC). The construction of this collection went through four stages. First, the data has been gathered from the Islamic expert Ashur's book (2002) and the Islam – the Quran and Tafseer18 website. Second, he removed unwanted elements, such as non-letter characters, to clean the data. Third, he wrote the Q&A in a separate Excel column. Finally, the Q&A was annotated with descriptive information. This information is related to the questions, such as the ID, type, topic, and ontology concepts, while the other is about the answer location in the Quran (chapter and verse numbers) and the Q&A references. However, there are deficiencies in this dataset that affect its use for evaluating QA systems: (1) many of the answers are an interpretation of the verses only, and they do not contain any Quranic verses. (2) Other answers contain verses mixed with an interpretation. (3) Also, some answers are written in MSA with supported verses, but the problem is that these answers have the same meaning as the Quran but with different terms. After filtering these cases, the dataset will contain only 473 out of 1224 Q&A (Aftab and Malik, 2022). (4) the dataset is not diverse because it was built from only two sources. (5)When we analyzed this dataset, we found that 2601 verses appeared as an answer to one question and 47 verses appeared as an answer to two questions out of 6236 Quran verses, which means that it covered only 41% of the Quran with one question.

Further, Malhas and Elsayed (2020) constructed a gold stander Arabic corpus of 207 Q&A about the Quran to unify the QA systems' assessment process called AyaTEC. This corpus covers 11 topics: provisions of Islam, stories of prophets, former nations, the unseen, universe & God's creations, worshipping, jihad, battles & wars, faith in God & believers, prophet Mohamad, humans/Mankind, linguistics of Quran and others. However, Almost half of the questions are about the first two topics. They gathered the questions from users directly or from multiple sources such as Books, YouTube videos, and previous studies Abdelnasser et al. (2014). Two freelancers answered the questions from UpWork8 19. Subsequently, three religious scholars reviewed the answers to verify their validity. However, the data size is too small. Based on our analysis, the questions are only about 1573 verses, constituting 25% of the Quran. These verses answer one or more questions based on the statistics shown in Table 1. There is no variety in the questions, as half of the questions are on only two topics.

2.2 Question-Passage-Answer Triplets Datasets

This dataset type is classified as a reading comprehension dataset, usually composed of a question, a passage, and an answer extracted from that passage. It is commonly used to measure a person's understanding of a text by asking questions about it. Four studies built this type of dataset for the Arabic Quran, which can be divided into three approaches. First, some researchers changed the structure of the available question-answer pairs datasets to fit the triplets structure (Question, Passage, Answer) and added new questions similar to the existing ones. Second, Some other researchers only modified the available dataset from (question, answer) pairs to (question, passage, answer) triplets. The last approach's concept was based on enlarging the question-passage-answer triplets dataset by reformulating the questions.

[18] http://islamqt.com/
[19] https://www.upwork.com/

Table 1: Statistics of the number of questions about the verses in the AyaTec dataset.

Number of verses	1022	337	145	40	18	18	5	4	2
Number of questions about these verses	1	2	3	4	5	6	7	8	9

2.1.1 Restructure the Existing Datasets and Added New Questions Approach

The Qur'anic Reading Comprehension Dataset (QRCD), a publicly accessible20 dataset proposed by Malhas et al. (2022), is restructured and expanded version of the AyaTEC. It is composed of 1,093 pairs of question-passage written in JSON file format, but because the questions can have more than one answer, the triplets become 1,337 question-passage-answer, as shown in Table **2**. The passage may appear more than once, but with different questions, and at the same time, the question may appear more than once but with a different passage. The triplets are classified into training, development, and test set.

Table 2:The statistics of the QRCD

	The percentage	Number of the Pairs	Number of the Triplets	Passages	Questions
Training	65%	710	861	468	118
Development	10%	109	128	101	17
Test	25%	274	348	256	34
All	100%	1,093	1,337	825	169

However, the number of non-repetitive questions is very few, only 169. In addition, the types of questions are not diverse. The dataset contains different questions type such as Why (لماذا), Who (من), How (كيف), Where (أين), When(متى), What (ما - ماذا), Do (هل), How much (كم), and In any (بأي). Nevertheless, about 80% of the question types are Why, Who, What, and Do (Keleg and Magdy, 2022). Moreover, not all the correct answers were added to the dataset. For example, the gold answer to the question "متى يحل الأسلام دم الشخص", which translates to "When does Islam allow the blood of a person?" is "قاتلو في سبيل الله الذين يقاتلونكم" meaning "Fight in the way of Allah those who fight you." At the same time, there is another correct answer in the passage, which is "فمن اعتدى عليكم فاعتدوا عليه بمثل ما اعتدى عليكم" "So whoever has assaulted you, then assault him in the same way that he has assaulted you." The meaning of the other correct answer is the same as the gold answer, according to Al-Tabari (1954), As shown in Figure 1. Furthermore, only 3325 verses are mentioned in this dataset which covers only 53% of the Quran. 1685 were mentioned once as an answer to one question only, while the statistics of the rest verses appear in Table 3.

PQ_ID: 2:190-194_400
Passage:
وقاتلوا في سبيل الله الذين يقاتلونكم ولا تعتدوا إن الله لا يحب المعتدين. واقتلوهم حيث ثقفتموهم وأخرجوهم من حيث أخرجوكم والفتنة أشد من القتل ولا تقاتلوهم عند المسجد الحرام حتى يقاتلوكم فيه فإن قاتلوكم فاقتلوهم كذلك جزاء الكافرين. فإن انتهوا فإن الله غفور رحيم. وقاتلوهم حتى لا تكون فتنة ويكون الدين لله فإن انتهوا فلا عدوان إلا على الظالمين. الشهر الحرام بالشهر الحرام والحرمات قصاص **فمن اعتدى عليكم فاعتدوا عليه بمثل ما اعتدى عليكم** واتقوا الله واعلموا أن الله مع المتقين.
Question: متى يحل الإسلام دم الشخص؟
Gold Answer:
1- **قاتلوا في سبيل الله الذين يقاتلونكم**
Predicted Answer:
1- ' واقتلوهم حيث ثقفتموهم وأخرجوهم من حيث أخرجوكم والفتنة أشد من القتل ولا تقاتلوهم عند المسجد الحرام حتى يقاتلوكم فيه فإن قاتلوكم فاقتلوهم كذلك جزاء الكافرين. فإن انتهوا فإن الله غفور رحيم. وقاتلوهم حتى لا تكون فتنة ويكون الدين لله فإن انتهوا فلا عدوان إلا على الظالمين. الشهر الحرام بالشهر الحرام والحرمات قصاص **فمن اعتدى عليكم فاعتدوا عليه بمثل ما اعتدى عليكم**'
2- ' وقاتلوهم حتى لا تكون فتنة ويكون الدين لله فإن انتهوا فلا عدوان إلا على الظالمين. الشهر الحرام بالشهر الحرام والحرمات قصاص **فمن اعتدى عليكم فاعتدوا عليه بمثل ما اعتدى عليكم**'
3- '**فمن اعتدى عليكم فاعتدوا عليه بمثل ما اعتدى عليكم**'
4- 'الشهر الحرام بالشهر الحرام والحرمات قصاص **فمن اعتدى عليكم فاعتدوا عليه بمثل ما اعتدى عليكم**'

Figure 1. An example of another correct answer can be added to the QRCD dataset

[20] https://gitlab.com/bigirqu/quranqa

Table 3:Statistics of the number of questions about the verses in the QRCD dataset.

Number of verses	1685	1046	397	116	57	57	12	9	3
Number of questions about these verses	1	2	3	4	5	6	7	8	9

Wasfey et al. (2022) constructed the Arabic Question Answers from the Holy Qur'an dataset (AQAQ). The construction process involved several stages :First, they selected 500 questions from AQQAC, kept the verses only as answers, and deleted any other answers. Next, an expert answered all these questions and added the passages. Finally, similar questions were added to bring the size of these datasets to 732 question-passage-answer triplets. Nevertheless, It is available[21] , but it cannot be opened.

2.1.2 Restructure the Existing Datasets Approach

Aftab and Malik (2022) extracted 473 Q&A pairs from the AQQAC to augment the QRCD dataset and enhance their QA system. The AQQAC structure is different from QRCD. The record in AQQAC contains the question and the answer, which is a verse or sequential verses from the Quran, while the record in QRCD contains the question, passage, which is a verse or sequential verses, and the answer that is a partial sentence or word from the passage. Therefore, by adding a context column, they restructured AQQAC to be the same as the QRCD structure. For each question, they filled this column through the following steps: determining the location of the verses that appear in the answer using Tanzil Quran text. Then extract the verse preceding the answer's verse, the answer's verse, and the verse after it. Finally, they add these verses to the context column. Nevertheless, this data set is unavailable and did not add any additional questions or answers to the AQQAC.

2.1.3 Reformulating the Existing Questions Approach

Ahmed et al. (2022) paraphrased the question to augment the training and development set of the QRCD by reordering the word and replacing the words with synonyms. This available22 augmentation data was used to fine-tune the model so that the model could find the correct answer to the different question formulas in the testing phase. Its size is around 466 question-passage-answer triplets. The number of verses mentioned once to answer one question is 273, and the rest of the statistic appears in Table 4.

Table 4: Statistics of the number of questions about the verses in the augmentation dataset.

Number of verses	273	145	23	21	0	0	4	0	0
Number of questions about these verses	1	2	3	4	5	6	7	8	9

The creation process of the question-passage-answer triplets dataset for the Quran is at an early stage, and it needs much work as the data size is very small, and the number of questions does not exceed hundreds. However, the users usually need one comprehensive answer to their question from the entire Quran without specifying a specific paragraph. The number of questions in QRCD is only 169, while the number of pairs is 1337, meaning that one question has many answers. Additionally, there is no possibility for questions in this dataset whose answers depend on the entire Quran. For example, the number and locations of the appearance of the Prophet Abraham, peace be upon him.

[21] https://github.com/EmanElrefai/Quran_QA/tree/main/datasets
[22] https://github.com/motazsaad/Quran-QA

3. Methodology

This study aims to survey and evaluate the existing corpus of Q&A about the Islamic text, which is more than 200 questions. This operation seeks to identify flaws in the current dataset to develop a more significant and better collection based on reusing the available dataset.

3.1 The Islamic Question Answering Corpora Evaluation Criteria

There are several criteria to evaluate the corpus, such as the scope, size, and purpose of constructing this corpus. Thirteen criteria were used in this study to assess the existing Q&A corpus. Some of these measures are adapted from Alrehaili and Atwell (2014) to evaluate Quran ontologies. The details of these criteria are listed in the following:

1. Source of the Question:
   Books (B) Websites (WE) Expert (E)
   Users (U) Previous studies (P) The writer extracted the questions from the verse (WR)
2. Source of the Answer:
   Books (B) Websites (WE) Expert (E)
   Previous studies (P) The writer who extracted the questions from the verse (WR)
3. Purpose of Creating this Corpus: The authors create this dataset for the question-answering systems to use as:
   Gold dataset for evaluation (G) A knowledge base (K)
4. Questions Type: The Questions type in this corpus
   Factoid (F) Other: Yes/No, Facts, List, Definition, Arguments, relation, etc. (O)
5. Answer Type: the answer natural in this corpus
   Plain Texts (PT) Verse (V) Verse details (verse number and sura name) (VD)
   Hadith (H) Tafsir's (T) Part of the verse (PV)
6. The Corpus Language:
   Arabic (A) English (E) Persian (P)
7. Corpus Size: The number of Q&A in this corpus.
8. Scope of the Corpus:
   Quran (Q) Hadith (H) Tafsirs (T) Fatwa (F)
9. The Data Coverage Area:
   The whole book (All) Part of the book (Part)
10. Topics Coverage:
11. Corpus Availability: is this corpus available to the users?
    Yes (Y) No (N)
12. Corpus Formats:
    Text (T) CSV (C) XML (X)
    Access Database (A) JSON Lines file (J)
13. Validation Approaches: used to validate the corpus:
    Reviewed by an Islamic Expert (R) Answered by an Islamic expert (A) None (N)

3.2 Comparing the Existing Corpora of the Islamic Question Answering

The previous evaluation criteria are used to compare the Islamic Q&A corpora in Table 5. Based on Table 5, In order for the dataset to be suitable, it is better to be characterized by the following: (1) it is from trusted sources, (2) it contains all kinds of questions, (3) its size is large, (4) it covers the Quran or Hadith completely, and (5) it is available to the user in several languages.

Table 5:Comparing the Islamic question-answer corpus

Criteria	(Saeedi et al., 2014)	(Sheker et al., 2016)	(Hamdelsayed and Atwell, 2016) and (Adany, 2017)	(B. I. Hamoud, 2017)	AQQAC (Alqahtani, 2019)	AyaTEC (Malhas and Elsayed, 2020)	QRCD (Malhas et al., 2022)	(Aftab and Malik, 2022)	(Wasfey et al., 2022)	(Abdi et al., 2020)	(Munshi et al., 2022)
1. Source of the Question:	WE	B	WR and WE	WE, E, and P	B and WE	B, WE, E, P, and U	B, WE, E, and P	P	E and P	E	WE
2. Source of the Answer:	WE	B	WR and WE	WE, E, and P	B and WE	E	E	P	E and P	B	WE
3. Purpose of Creating this Corpus	K	K	G and K	K	G	G	G and K	K	K	G	K
4. Questions Type	-	-	-	-	F, and O.	F, and O.	F	-	F	-	-
5. Answer Type	PT	PT	V and VD	PT, V, and VD	PT, V, T, and VD	V, VD, and PV	V, VD, and PV	V, VD, and PV	V, VD, and PV	H	PT
6. The Corpus Language	P	A	A and E	A and E	A	A	A	A	A	A	A
7. Corpus Size	6000 Q&A pairs	1094 Q&A pairs	263 questions with many verses answer	1500 Q&A pairs	2224 Q&A pairs	207 questions with many verses answers (1573 verses)	1,337 question-passage-answer triplets	473 question-passage-answer triplets	432 question-passage-answer triplets	3825 Q&A pairs	850K Q&A pairs
8. Scope of the Corpus	Q	F	Q	Q	Q and T	Q	Q	Q	Q	H	F
9. The Data Coverage Area:	-	-	Part	-	Part	Part	Part	-	Part	-	-
10. Topics Coverage	30 topics	Prayer	-	-	-	11 topics	-	-	-	-	-
11. Corpus Availability	N	N	N	N	Partly Y	Y	Y	N	Y	N	N
12. Corpus Formats	X	-	T, C, and A	C	C	T and X	J	J	J	-	-
13. Validation Approaches	A	A	R and A	A and N	A	R	R	A	R and A	A	A and N

4. Conclusion

This study surveys the existing Islamic questions and answers dataset and compares them using thirteen measures. The following points summarize what we found: (1) As far as we know, no current public questions and answers corpus specialized in Hadith or Tafsir exists. The existing corpora cover only Arabic Quran questions. (2) A thousand is the average corpus size which is considered very small. (3) The question types in the datasets are limited. Even if there is a slight diversity in the questions, the most number of questions of a particular type. (4) Some datasets specialise in only specific areas, or even if they specialize in several areas, the most significant proportion of the questions specialise in a limited number of areas. (5) Each corpus uses different file formats such as text, CSV, JSON, and XML. (6) Many answers in the available corpora do not have any verse or Hadith evidence. (7) Some datasets contain many answers to

one question that varies by passage, which usually contradicts the user's desire to find one comprehensive answer. At the same time, it does not contain all the correct answers. (8) no corpus covers all the verses of the Quran. (9) Most of these datasets cover any verse of the Quran with only one or two questions. As a result, there is a need to construct a reliable and big-size Islamic questions and answers corpus that covers the whole Quran or Hadith using various questions for each verse or Hadith. This corpus can be used in many systems such as the question-answering system.

References

Abdelnasser, H., Ragab, M., Mohamed, R., Mohamed, A., Farouk, B., El-Makky, N. M., and Torki, M. (2014). Al-Bayan: an Arabic question answering system for the Holy Quran. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 57–64.

Abdi, A., Hasan, S., Arshi, M., Shamsuddin, S. M., and Idris, N. (2020). A question answering system in hadith using linguistic knowledge. Computer Speech \& Language, 60, 101023.

Adany, M. A. (2017). An automatic question answering system for the Arabic Quran. Ph.D. thesis, Sudan University of Science and Technology.

Aftab, E., and Malik, M. K. (2022, Jun). eRock at Qur'an QA 2022: Contemporary Deep Neural Networks for Qur'an based Reading Comprehension Question Answers. Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022).

Ahmed, B. H., Saad, M. K., and Refaee, E. A. (2022, Jun). QQATeam at Qurán QA 2022: Fine-Tunning Arabic QA Models for Qurán QA Task. Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022).

Al-Tabari, M. (1954). جامع البيان عن تأويل القرأن. Dar Al-Fikr.

Alqahtani, M. M. (2019). Quranic Arabic Semantic Search Model Based on Ontology of Concepts. Ph.D. thesis, The University of Leeds.

Alrehaili, S. M., and Atwell, E. (2014). Computational ontologies for semantic tagging of the Quran: A survey of past approaches. LREC 2014 Proceedings.

Ashur, Q. (2002). 1000 soual wa jawab fi al Quran. Dar Ibn Hazm.

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

Hamdelsayed, M. A., and Atwell, E. (2016). Islamic applications of automatic question-answering.

Hamoud, B., and Atwell, E. (2017). Evaluation corpus for restricted-domain question-answering systems for the holy Quran. International Journal of Science and Research, 6(8), 1133–1138.

Hamoud, B. I. (2017). A Question Answering System Design about the Holy Quran. Ph.D. thesis, Sudan University of Science and Technology.

Ibn Baz, Abdul Aziz and AlUthaymeen, Muhammad and Al-Madkhalee, R. (2003). Three Essays on the Obligation of Veiling. edited by Translated by Abu Maryam Ismaeel Alarcon Toronto: Al,. Ibaanah Book Publishing.

Keleg, A., and Magdy, W. (2022, Jun). SMASH at Qur'an QA 2022: Creating Better Faithful Data Splits for Low-resourced Question Answering Scenarios. Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022).

Malhas, R., and Elsayed, T. (2020). AyaTEC: building a reusable verse-based test collection

for Arabic question answering on the Holy Qur'an. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(6), 1–21.

Malhas, R., Mansour, W., and Elsayed, T. (2022, Jun). Qur'an QA 2022: Overview of the first shared task on question answering over the holy qur'an. Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022).

Maraoui, H., Haddar, K., and Romary, L. (2021). Arabic factoid Question-Answering system for Islamic sciences using normalized corpora. Procedia Computer Science, 192, 69–79.

Mohammed, M., Amin, S., and Aref, M. M. (2022, March). An English Islamic Articles Dataset (EIAD) for developing an IslamBot Question Answering Chatbot. 2022 5th International Conference on Computing and Informatics (ICCI), 303–309.

Munshi, A. A., AlSabban, W. H., Farag, A. T., Rakha, O. E., Al Sallab, A., and Alotaibi, M. (2022). Automated Islamic Jurisprudential Legal Opinions Generation Using Artificial Intelligence. Pertanika Journal of Science \& Technology, 30(2).

NEAMAH, N., and SAAD, S. (2017). QUESTION ANSWERING SYSTEM SUPPORTING VECTOR MACHINE METHOD FOR HADITH DOMAIN. Journal of Theoretical \& Applied Information Technology, 95(7).

Saeedi, P., Heidari, S., and Farhoodi, M. (2014). Creating quranic question taxonomy. 2014 22nd Iranian Conference on Electrical Engineering (ICEE), 1070–1074.

Sapp, B., Saxena, A., and Ng, A. Y. (2008). A Fast Data Collection and Augmentation Procedure for Object Recognition. AAAI, 1402–1408.

Sheker, M., Saad, S., Abood, R., and Shakir, M. (2016). Domain-specific ontology-based approach for Arabic question answering. Journal of Theoretical and Applied Information Technology, 83(1), 43.

Shmeisani, H., Tartir, S., Al-Na'ssaan, A., and Naji, M. (2014, October). Semantically answering questions from the Holy Quran. International Conference on Islamic Applications in Computer Science And Technology, 1–8.

Wasfey, A., Elrefai, E., Marwa, M., and Haq, N. (2022, Jun). Stars at Qur'an QA 2022: Building Automatic Extractive Question Answering Systems for the Holy Qur'an with Transformer Models and Releasing a New Dataset. Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022).