HAQA and QUQA: Constructing two Arabic Question-Answering Corpora for the Quran and Hadith

Sarah Alnefaie
King Abdulaziz University,
Jeddah, Saudi Arabia
University of Leeds, Leeds, UK
scsaln
@leeds.ac.uk

Eric Atwell
University of Leeds
Leeds, UK
e.s.atwell
@leeds.ac.uk

Mohammad Ammar Alsalka
University of Leeds
Leeds, UK
m.a.alsalka
@leeds.ac.uk

Abstract

It is neither possible nor fair to compare the performance of question-answering systems for the Holy Quran and Hadith Sharif in Arabic due to both the absence of a golden test dataset on the Hadith Sharif and the small size and easy questions of the newly created golden test dataset on the Holy Quran. This article presents two question–answer datasets: Hadith Question-Answer pairs (HAQA) and Quran Question–Answer pairs (QUQA). HAQA is the first Arabic Hadith question–answer dataset available to the research community, while the QUQA dataset is regarded as the more challenging and the most extensive collection of Arabic question–answer pairs on the Quran. HAQA was designed and its data collected from several expert sources, while QUQA went through several steps in the construction phase; that is, it was designed and then integrated with existing datasets in different formats, after which the datasets were enlarged with the addition of new data from books by experts. The HAQA corpus consists of 1598 question–answer pairs, and that of QUQA contains 3382. They may be useful as gold–standard datasets for the evaluation process, as training datasets for language models with question-answering tasks and for other uses in artificial intelligence.

1 Introduction

Natural language processing and artificial intelligence have been employed to computerize numerous tasks that require an expert in the field. One such task involves analyzing textual material to extract information that can be used to answer questions, including finding answers from Islamic religious texts such as Hadith sharif and the Quran.

The Holy Quran and the Hadith Sharif are the primary sources for millions of Muslims worldwide. Muslims draw from them for legislation, teachings, wisdom, knowledge, and a complete understanding of religion, making them important and fertile resources for answering their questions. Consisting of 30 parts, 114 suras and 6236 verses, the text of the Holy Quran is the word of God in classical Arabic (CA) (Atwell et al., 2010). The Quranic text has several characteristics, such as its series of verses of different lengths; one verse may cover various topics, and the same topic may be covered in many different verses. These characteristics lead to there being many challenges in processing and researching the Quranic text. Hadiths are the sayings and deeds of the Prophet Muhammad, may God bless him, that were handed down through a chain of narrators. They may consist of a short or long sentences about what the Prophet, may God bless him and grant him peace, said, his conversations with someone else or what he narrated to his companions about his actions regarding a particular matter. The significance of the Hadith lies in the Quran's directive for Muslims to follow the teachings of the Prophet Muhammad, since many of the topics that are touched upon in the Quran are mentioned in greater detail in the hadiths. For example, God commanded Muslims to pray according to the Holy Quran, but the method and mechanism for praying are mentioned in the Hadith Sharif. Processing hadiths faces the same challenges as processing the Quranic text. In addition, there are 33,359 hadiths in the Al-Sihah al-Sittah books (Altammami et al., 2020).

Much research effort has been devoted to developing a system that can respond to inquiries related to the Quran (Malhas et al., 2022), while only a few studies have focused on addressing questions from the hadiths. However, the primary difficulty of question-answering (QA) studies concerns the direct and easy questions and small size of the Quranic question–answer collections and the absence of a Hadith question–answer corpus. As a result, each study of building a QA system for hadiths has used its own dataset to evaluate the system, which has led to obstacles in comparing the results (NEAMAH and SAAD, 2017; Abdi et al.,

Proceedings of Recent Advances in Natural Language Processing, pages 90–97
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_010

2020; Maraoui et al., 2021). In addition, the small size of Quranic datasets in the training phase has affected the results of the language models in the Quran QA task (Malhas and Elsayed, 2022).

Therefore, we aim to enrich the Arabic Islamic language resources. The design objectives of our two question–answer datasets are as follows: (1) to use a variety of expert books, (2) to cover various types of questions and topics and different difficulty levels of the questions and (3) to collect as many questions as possible for use in training language models and systems evaluation.

Our contribution is threefold: (1) We present Quran Question–Answer pairs (QUQA), the most extensive reusable Quran question–answer collection, by integrating the existing available datasets and enlarging them using different resources and challenging questions. This dataset covers a large number of questions and more verses, with the questions being in modern standard Arabic (MSA) and the answers from Quran verses in CA. (2) We introduce Hadith Question–Answer pairs (HAQA), the first reusable Arabic hadith question–answer corpus, by collecting the data from different expert resources. (3) We make these two datasets available [1] to the research community, which will reflect positively on research on Islamic QA. They can be used as a golden test collection or as training and testing data in language model research.

In the following section, we discuss the existing related collections. We then outline the methodology for designing, collecting, and building our two datasets. After that, we show the resulting datasets. Finally, we present our conclusion.

2 Related Work

Most of the existing studies of building QA systems for the Holy Quran have involved creating test sets to evaluate their systems, but these datasets are unavailable. For example, datasets containing 263 question–answer pairs have been developed, and a small part of the questions were collected from websites, with the vast majority generated manually from Quranic text. These questions are solely about the 'Al-Baqarah' and 'Al-Fatiha' chapters (Hamdelsayed and Atwell, 2016; Adany et al., 2017). In addition, Hamoud and Atwell (2017) collected 1500 questions and answers from websites.

Alqahtani (2019) constructed the first available corpus of 1224 question–answer pairs called the Annotated Corpus of Arabic Al-Quran Question and Answer (AQQAC) that were gathered from the Islam – Quran & Tafseer website [2]. Studies have not used this dataset to evaluate QA systems for several reasons, such as (1) many of the answers only consist of interpretations and not evidence from the Quran, and (2) some of the questions include complete verses from the Quran written in CA, and the exact required answers are written in MSA by the scholar. Therefore, this dataset cannot be used directly since the exact answers do not use Quranic words (Sleem et al., 2022). Nevertheless, after cleaning this dataset and excluding answers that only contain interpretations, we found that only 1232 verses are used to answer the questions, covering only 19% of the Holy Quran. In addition, the number of questions (611) is small, and they are simple and taken straight from the text.

Malhas and Elsayed (2020) developed a dataset called AyaTEC, and the process of building this collection went through many stages. They began by collecting questions from different sources, then freelancers found the answers to these questions from the Quran. Finally, specialist religious scholars reviewed the datasets. In addition, they developed an extended version of AyaTEC called QRCD, which was intended to be an intensive machine reading comprehension (MRC) task. It has been used in several recent studies to train and test different language models to obtain a system for answering questions that performs well (Malhas et al., 2022).

Nevertheless, the size of this dataset is relatively small, and the number of questions is very limited. After excluding indirect answers, there are 169 questions and 1166 records, since one question may have more than one answer. Only 1247 verses are used to answer the questions, which means that this corpus only covers around 20% of the Quran. Not all correct answers are included (Alnefaie et al., 2022). The use of this collection in measuring the system's performance does not measure the strength of the actual answering system, since the nature of the most questions is direct.

Based on the above, to address the shortage of datasets, we design and create the QUQA by cleaning and integrating the existing datasets, enlarging them with more challenge questions from various

[1] http://github.com/scsaln/HAQA-and-QUQA

[2] http://http://islamqt.com/

sources and covering a more significant number of verses.

On the other hand, other researchers have been interested in finding answers to questions from hadiths and tested their systems by building different test collections. NEAMAH and SAAD (2017) collected hadiths and then asked university students to create questions from them, with the collection size reaching 12 questions. Abdi et al. (2020) built a collection of 3825 question–answer pairs by reading the hadiths and extracting questions manually. Maraoui et al. (2021) constructed a corpus of 33 questions from native Arabic speakers and online forums. Nevertheless, all these datasets are unavailable, and to the best of our knowledge, no Hadith Sharif question–answer datasets are publicly available. Therefore, we introduce the HAQA dataset to the research community to fill this gap.

3 Building QUQA and HAQA

The methodology for creating QUQA and HAQA went through several stages, consisting of designing the two datasets, identifying the data sources, and collecting and cleaning the data. Figure 1 shows the development methodology of these two datasets, which we now go on to discuss in detail.

3.1 QUQA and HAQA Design

As a starting point for building the two datasets, we must define the structure of the collection, its metadata, and the format in which it will be available. We designed the Quran dataset based on the AyaTEC and the AQQAC designs, and the common metadata between the two corpora were adopted. Similar metadata were selected for HAQA to suit the nature of hadiths. Comma- separated values (CSVs) with UTF-8 encoding format were used because many systems can easily use them following their conversion into XML format. Every record in the QUQA CSV file includes the information listed in Table 1. The information in the HAQA records is similar.

3.2 Identifying Data Sources

To create the corpora, we used two sources, namely books and available datasets. Many books include questions and answers about the Quran and the Hadith Sharif, but they did not meet our requirements. For example, the answers in some sources are solely in the words of an expert and do not contain evidence from the Quran or the hadiths. Additionally, we did not have permission to publish the data of some sources in our datasets. The available datasets and books matching our requirements that were used to build QUQA are as follows:

- AQQAC: This was the first Islamic dataset made available to the research community and contains answers from Quranic verses, interpretations of the verses and explanations of them in the words of an expert. This dataset file is available in XLXS and XML formats. Among the topics covered by this dataset are stories of the prophets and previous nations, Islamic legal rulings and knowledge of unseen matters (Alqahtani, 2019).

- AyaTEC: This is a specialized dataset with answers from the Holy Quran. It consists of three XML files that must be linked together. The questions relate to 11 topics, including battles, humans, stories of the prophets and faith in God (Malhas and Elsayed, 2020).

- 900 Questions and Answers in Managing the Verses of the Book: This is a set of questions and answers from the Quran developed by the writer due to his belief that formulating material with questions and answers increases a person's understanding of the subject (ALmuselli, 2020).

- 100 Quranic Questions and Answers: This is a set of questions and answers from the Quran developed by the writer to answer people's questions and make them more aware of their religion (Alakeel, 2018).

The books that were used to create QUQA and HAQA are as follows:

- The Doctrine of Every Muslim in a Question-and-Answer Book and the Abridged Version of the Islamic Belief from the Quran and Sunnah: This is a series of publications by Sheikh Zeno that answers the most important questions in the Muslim faith (Zeno, 2004, 2007).

- Inference on Children's Treasure: This contains a set of questions covering the following topics: the most important matters of religion, the foundations of faith, belief, the principles of jurisprudence, etiquette, dealings between people, the Prophet's biography etc. This

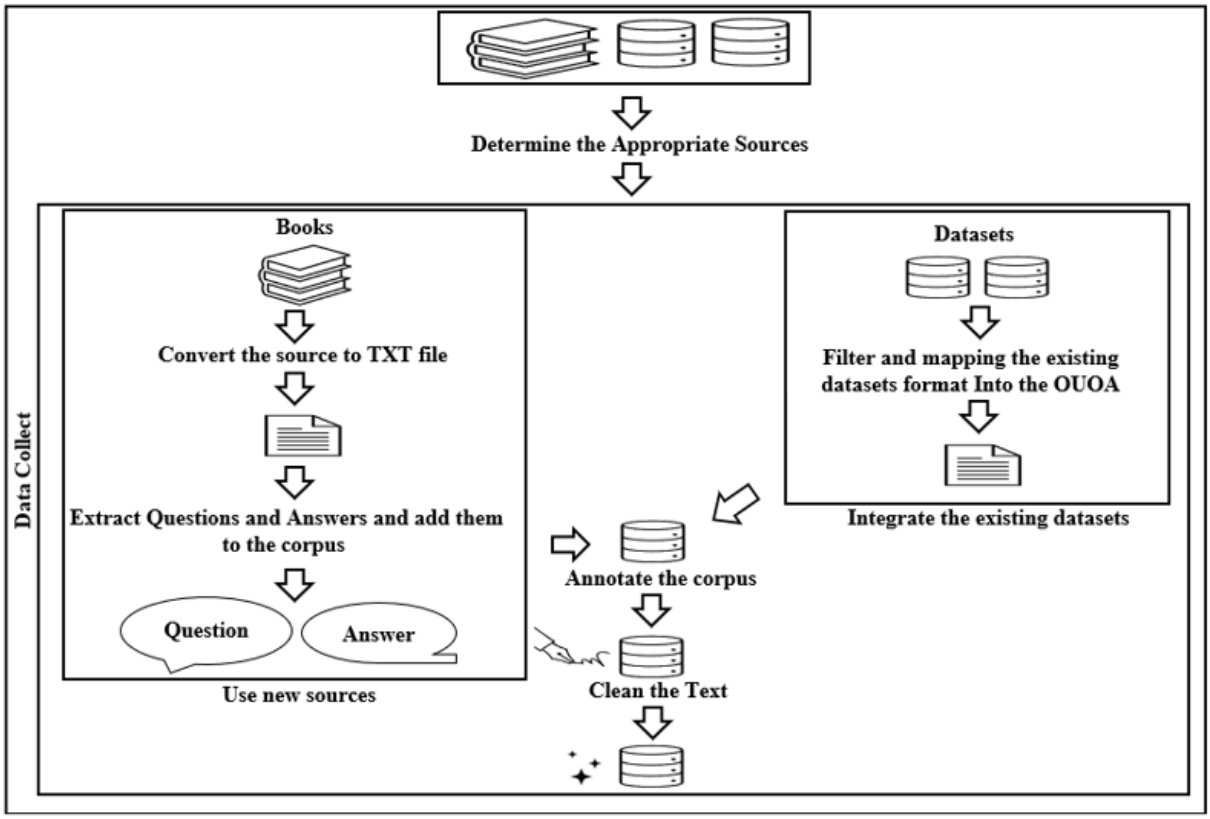

Figure 1: The development methodology of QUQA and HAQA.

book's questions are related to the basics of religion, the Prop het's life, faith, matters related to the afterlife etc. (Al-Wadi, 2016).

- Prayer (1770) Question and Answer: This contains people's questions related to the topic of prayer, with answers taken from the Quran and the hadiths (Al Alami, 2022). This selection achieved the design goals, since the most significant questions, in terms of their type, topic and source, were included in these corpora.

3.3 Data Collection

This stage consisted of two steps, the first being to integrate the existing datasets and the second to use new sources. As mentioned earlier, there are two corpora in the Quran domain, AQQAC and AyaTEC, while the hadiths have no dataset. We wrote a Python program to convert the existing two datasets into the structure and format of our dataset. In the second step, we used the new sources to enlarge QUQA and create HAQA. The sources of the two collections consist of six books, some of which are available in text format and some not. Therefore, we wrote a Python program using OCR to convert some of the books into text format, which we reviewed manually. After that, we wrote a program that extracted questions and answers from text files and put them in our files. As a final step in this stage, we filled in the metadata using Python, or manually in some cases.

3.4 Cleaning the Data

Cleaning data is the process of detecting and fixing errors and incorrect information. Such errors include misspellings, missing information, unwanted items, and noisy and duplicate data. This cleaning process improves the quality of the resulting data, which reflects positively on the purpose of collecting it. There are two methods for cleaning data: one that is manual and the other automated. We used the manual method to discover spelling errors, missing information and duplicate information, although this approach usually takes time and effort. In addition, we used the automated approach by applying some regular expressions to remove extra spaces and non-Arabic characters. An example of a QUQA final record is shown in Table 2 and Table 3 while a HAQA example is shown in Table 4 and Table 5. We combined the answers of duplicate questions.

Annotation	Description
Record Id	The unique record number.
Question Id	A question number may appear many times in this dataset due to the following: 1.The question has many different answers. 2.The question has one answer, but it is mentioned in many different verses in the Quran.
Question Text	The question text
Quetion Type	The type of question can be a factoid (F), a confirmation (C) or a description (D).
Question Start Word	The question keyword.
Answer ID	The number of unique answers to the same question: 1.If the question has only one answer in a sense that comes totally or partially from different verses with different syntax, the numbering appears as 1.1, 1.2, 1.3 etc. 2.If the question has different answers in the same or different verses, the numbering appears as 1, 2, 3 etc.
Full Answer	The whole answer consists of expert commentary, the Quranic verse and the hadith.
Expert Commentary	An answer uses an expert's words alone.
Answer Instances	The exact part of a verse that answers the question. The verse may contain more than one answer, and each answer considers an answer instance.
Quran Full Verse Answer	A complete verse that considers or contains the answer.
Chapter Name	The chapter name.
Chapter Number	The chapter number.
Verses Number Start	The number of the first verse.
Verses Number End	The number of the last verse.
Source Name	The name of the source.
Source Link	The link to the source.
Credibility	Yes, if an Islamic expert has reviewed the answers; no, if they have not done so.
Question ID in the Original Dataset	The question ID in the original dataset.

Table 1: QUQA annotation.

Record Id	Question Id	Question Text	Question Type	Question Start Word	Answer ID	Answer Instances
2350	1345	Perhaps one person had succeeded in saving society and saving a nation! There is a beautiful verse that indicates this meanin Mention it?	D	Mention	1	At length, when they came to a (lowly) valley of ants, one of the ants said, 'O ye ants, get into your habitations, lest Solomon and his hosts crush you (under foot) without knowing it.'

Table 2: Example of QUQA Record– Part 1.

Chapter Name	Chapter Number	Verses Number Start	Verses Number End	Source Name	Credibility	Question ID in the Orignal Dataset
An-Naml	27	18	18	900 Questions and Answers in Managing the Verses of the Book for ALmuselli.	yes	19.15

Table 3: Example of QUQA Record– Part 2.

Record Id	Question Id	Question Text	Question Type	Question Start Word	Answer ID	Full Answer
472	404	What is the name of the battle during which the Prophet, peace be upon him, took a wound to the head and had his front teeth damaged?	F	What	1	The Battle of Uhudl. It has been narrated on the authority of Anas that the Messenger of Allah (may peace be upon him) had his front teeth damaged on the day of the Battle of Uhudl and got a wound to his head. (Sahih Muslim, 1791).

Table 4: Example of HAQA Record– Part 1.

Expert Commentary	Hadith Full Answer	Answer Instances	Source Name		Question ID in the Orignal Dataset
The Battle of Uhudl	It has been narrated on the authority of Anas that the Messenger of Allah (may peace be upon him) had his front teeth damaged on the day of the Battle of Uhudl and got a wound to his head. (Sahih Muslim, 1791).	On the day of the Battle of Uhudl	Inference on Treasure	Children's	595

Table 5: Example of HAQA Record– Part 2.

4 Evaluation of the Corpora

QUQA is an Arabic question-and-answer dataset on the Holy Quran consisting of 3382 records and over 301,000 tokens. Since some questions may have more than one answer, there are 2189 questions. The answers in this corpus are extracted from 2930 verses of the Holy Quran. Accordingly, this dataset covers almost 47% of the Quran. We noticed that the questions in the new dataset are more diverse and challenging than those in the previous datasets, as shown in Table 2 and Table 3. In contrast to the two existing datasets, whose questions are considered to be direct and explicit because they include the words found in the answer, extracting the answer is easy. Table 6 shows the comparison results between our corpus and the two existing corpora. This dataset covers many topics, including worship, the most important matters of religion, the foundations of faith, belief, the principles of jurisprudence, etiquette, matters related to the afterlife, dealings between people, the life of the Prophet, battles, humans, and stories about prophets. There are 199 single-answer and 1990 multiple-answer questions. The single-answer questions are ones that have only one answer found in one or several verses in the Quran, with answers that are repeated in different places in the Quran being semantically and/or syntactically similar. The multiple-answer questions have several different answers to the question.

In addition, when we analyzed the Arabic HAQA dataset of Hadith sharif answers, we found that there are 1598 records and 1359 questions. The hadiths in this collection were taken from various sources of basic hadith books; for example, there are hadiths from Al-Bukhari, Muslim, Al-Tirmidhi, Al-Nasai, Ibn Majah, Imam Ahmad, Ibn Shaybah and others. The most important matters of religion, battles, biographies of men about the Prophet Muhammad, the foundations of faith, belief, the principles of jurisprudence, etiquette, dealings between people, the life of the Prophet, worship and others are the main topics covered by this dataset.

5 Conclusion and Future Work

This paper presents the process of building two Islamic religious corpora in Arabic. QUQA and HAQA are two datasets that contain questions and answers about the Holy Quran and the Hadith Sharif, respectively. Since these corpora include more than 4900 records, they are considered to be the largest Islamic corpora available [3] to the research community.

Datasets	AQQAC	AyaTEC	QuQA
# Records	616	1166	3382
# Questions	611	169	2189
#Verses in the answers	1232	1247	2930
% of Quran coverage	19%	20%	47%

Table 6: Comparing QuQA, AQQAC and AyaTEC.

These two datasets enrich the resources of the Arabic language, which suffers from a shortage of datasets in comparison with English and other languages. They open the door to conduct much more research in the field of artificial intelligence and the task of studying the nature and understanding of classical Arabic texts.

In the future, we plan to enlarge these two corpora to cover a significant number of hadiths and Quranic verses, including a greater variety of question types and challenging questions that will improve the dataset's quality. Different languages, such as English, can be added to them. In addition, a question–answer corpus can be built for other Islamic books using the same methodology, enhancing the state-of-the-art of Islamic QA systems.

References

Asad Abdi, Shafaatunnur Hasan, Mohammad Arshi, Siti Mariyam Shamsuddin, and Norisma Idris. 2020. A question answering system in hadith using linguistic knowledge. *Computer Speech & Language*, 60:101023.

Mohamed Adany Hamdelsayed Adany et al. 2017. *An automatic question answering system for the Arabic Quran*. Ph.D. thesis, Sudan University of Science and Technology.

Faisal bin Misfer bin Moawad Al Alami. 2022. *Prayer (1770) Question and Answer*.

Faisal bin Misfer bin Moawad Al-Wadi. 2016. *Inference on children's treasure.* Dar Knoz Al-Islam.

Fouzia Alakeel. 2018. *Quranic questions and answer*.

Duraid ALmuselli. 2020. *900 questions and answers in managing the verses of the book*. Altafseer, Erbil.

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. 2022. Challenges in the islamic question

[3] http://https://github.com/scsaln/HAQA-and-QUQA

answering corpora. *International Journal on Islamic Applications in Computer Science And Technology*, 10(4):1–10.

Mohammad Mushabbab A Alqahtani. 2019. *Quranic Arabic semantic search model based on ontology of concepts*. Ph.D. thesis, University of Leeds.

Shatha Altammami, Eric Atwell, and Ammar Alsalka. 2020. Constructing a bilingual hadith corpus using a segmentation tool. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3390–3398. The European Language Resources Association (ELRA).

Eric Atwell, Nizar Habash, Bill Louw, Bayan Abu Shawar, Tony McEnery, Wajdi Zaghouani, and Mahmoud El-Haj. 2010. Understanding the quran: A new grand challenge for computer science and artificial intelligence. *ACM-BCS Visions of Computer Science 2010*.

Mohamed Adany Hamdelsayed and Eric Atwell. 2016. Islamic applications of automatic question-answering. *Journal of Engineering and Computer Science*, 17(2):51–57.

Bothaina Hamoud and Eric Atwell. 2017. Evaluation corpus for restricted-domain question-answering systems for the holy quran. *International Journal of Science and Research*, 6(8):1133–1138.

Rana Malhas and Tamer Elsayed. 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.

Rana Malhas and Tamer Elsayed. 2022. Arabic machine reading comprehension on the holy qur'an using cl-arabert. *Information Processing & Management*, 59(6):103068.

Rana Malhas, Watheq Mansour, and Tamer Elsayed. 2022. Qur'an qa 2022: Overview of the first shared task on question answering over the holy qur'an. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 79–87.

Hajer Maraoui, Kais Haddar, and Laurent Romary. 2021. Arabic factoid question-answering system for islamic sciences using normalized corpora. *Procedia Computer Science*, 192:69–79.

NABEEL NEAMAH and SAIDAH SAAD. 2017. Question answering system supporting vector machine method for hadith domain. *Journal of Theoretical & Applied Information Technology*, 95(7).

Ahmed Sleem, Eman Mohammed lotfy Elrefai, Marwa Mohammed Matar, and Haq Nawaz. 2022. 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. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 146–153.

Muhammad bin Jamil Zeno. 2004. *The abbreviation of the Islamic belief from the Qur'an and Sunnah*.

Muhammad bin Jamil Zeno. 2007. *The doctrine of every Muslim in a question and answer*.