AHJL at Qur'an QA 2023 Shared Task: Enhancing Passage Retrieval using Sentence Transformer and Translation

Hessa A. Alawwad
Imam Mohammad Ibn Saud Islamic
University (IMSIU)
haalawwad@imamu.edu.sa

Lujain A. Alawwad
Saudi Electronic University
L.alawwad@seu.edu.sa

Jamilah Alharbi
King Abdulaziz University
eng.jamilah@gmail.com

Abdullah I. Alharbi
King Abdulaziz University
aamalharbe@kau.edu.sa

Abstract

The Holy Qur'an is central to Islam, influencing around two billion Muslims globally, and is known for its linguistic richness and complexity. This article discusses our involvement in the PR task (Task A) of the Qur'an QA 2023 Shared Task. We used two models: one employing the Sentence Transformer and the other using OpenAI's embeddings for document retrieval. Both models, equipped with a translation feature, help interpret and understand Arabic language queries by translating them, executing the search, and then reverting the results to Arabic. Our results show that incorporating translation functionalities improves the performance in Arabic Question-Answering systems. The model with translation enhancement performed notably better in all metrics compared to the non-translation model.

1 Introduction

The Holy Qur'an holds significant relevance as it serves as the central holy book in Islam, guiding the beliefs and practices of over 1.9 billion Muslims worldwide. It provides essential spiritual guidance, imparts moral values, and establishes rules for living, exerting a profound influence on the lives of Muslims and their communities. Comprising 114 chapters (Suras) and 6236 verses (Ayas) of varying lengths, totaling approximately 80,000 Arabic words, the Qur'an, revealed over 1,400 years ago, is written in classical Arabic (Atwell et al., 2011). is considered to be linguistically complex because it uses a rich vocabulary, intricate sentence structures, and rhetorical devices like metaphors and allegories. Its verses can have multiple meanings depending on the context, allowing for various interpretations (Alasmari, 2020). Various studies have explored the Holy Qur'an for different NLP tasks, such as creating datasets, question answering (QA), retrieving related information, and and identifying topics (Adeleke et al., 2019; Mohd et al., 2021; Mohamed and Shokry, 2022; Malhas and Elsayed, 2022).

One recent study on applying NLP to the Qur'an relies on the Qur'an QA shared task (2022) (Malhas et al., 2022). They propose a task defined as giving a group of verses from a particular part of the Holy Qur'an and a question about those verses; a system needs to find the answer to the provided question. The organizers continued to provide this shared task, Qur'an QA 2023 Shared Task. However, they added a new task called the Qur'anic passage retrieval (PR) task. PR is defined as participants will be given a question in Modern Standard Arabic and a set of Qur'anic passages that cover the entire Holy Qur'an. The system is required to return a list of these passages, ranked in order of how likely they are to contain the answer to the provided question. The question may vary in complexity, ranging from simple and direct to more intricate and nuanced. However, some questions might not have an answer in the Holy Qur'an to make the task more realistic and challenging. In such cases, an adequate system should recognize that there is no answer. Otherwise, it should return a list of the top ten passages likely to contain the answer.

This paper describes our participation in the PR task (Task A) provided by the Qur'an QA 2023 Shared Task. Our proposed method is to translate the Arabic Questions into English and incorporate a paraphrasing module to enhance the retrieving process. The rest of the paper is organized as follows: Section 2 discusses the related work. Section 3 explains the data we used for our tests. Section 4 provides details of our experiments. Section 5

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 702–707
December 7, 2023 ©2023 Association for Computational Linguistics

presents the results tied to our research queries. Finally, we discuss potential next steps and conclude the paper.

2 Related Work

In the domains of Natural Language Processing (NLP) and Information Retrieval (IR), the task of Question Answering (QA) involves finding accurate answers to questions within a body of text. QA combines these two fields by requiring an understanding of language, as in NLP, and the ability to find the proper documents, as in IR (Alami et al., 2023). A typical QA system consists of several steps, including understanding questions, finding relevant text passages, and extracting answers to deliver precise responses from extensive textual sources (Alwaneen et al., 2022).

In the field of information retrieval, using language models to rank documents based on their relevance to a query has been a popular method (Ponte and Croft, 2017). Earlier methods used count-based language models for each document to determine its likelihood of being relevant to a query (Zhai and Lafferty, 2004). Sentence similarity involves assessing the likeness between two texts, where each sentence pair is judged based on the notion that they have identical meanings (Achananuparp et al., 2008). Models for sentence similarity transform input texts into embeddings that capture the overall meaning and then compute their proximity according to some specific measure, such as cosine-similarity or dot product. In the Al-Bayan system by Abdelnasser et al. (2014), the researchers utilized the Holy Qur'an and Tafseer to identify verses with similar meanings using semantic analysis. They developed a semantic interpreter with machine learning to transform text into vectors representing Qur'anic concepts. These vectors, built from terms in the relevant documents, are weighted using the TF-IDF method. The system calculates the similarity between the vectors of a given question and terms in the Qur'an, and then highlights the most relevant terms to that question. These methods had challenges, like dealing with limited data.

Using commercial search engines as external sources for paragraph retrieval is one of the methods used in the literature. The EWAQ system, introduced by AL-Khawaldeh (2015), presents a novel passage retrieval (PR) method. This method fetches passages from search engines and calculates their relevance to a query based on "entailment similarity", employing cosine directional similarity as a metric. A similar method for passage retrieval was suggested by Bakari and Neji (2022). Initially, passages related to the query are fetched from Google using the question's keywords. These passages are then refined, standardized, and divided. Next, the questions and passages are examined linguistically, including identifying named entities, analyzing syntax, and assessing morphology. In the end, the main ideas of the question and the passage are presented logically.

More modern techniques use advanced language models like BERT to determine query relevance. Such methods have an advantage over older sparse retrieval techniques because they recognize word-based and more profound meaning similarities rather than just looking for exact keyword matches (Nogueira dos Santos et al., 2020). Karpukhin et al. (2020) aimed to develop an effective dense embedding model by merging the BERT pre-trained model with a dual-encoder setup. This model transforms text into a specific vector format and then indexes every passage for retrieval. They found that their model surpassed several other models in question-answer tests on various datasets like SQuAD (Rajpurkar et al., 2016) and TriviaQA (Joshi et al., 2017).

3 Methodology

3.1 Dataset Description

The dataset used consists of three main components: the Qur'anic Passage Collection (QPC), questions from the AyaTEC dataset, and relevance assessments for these questions against the QPC passages. The QPC was created by organizing the 114 Qur'anic chapters into topic-based segments using the Thematic Holy Qur'an (Swar, 2007), resulting in 1,266 distinct passages. The dataset was split into 70% training with 174 questions including 25 no-answer question, 10% development with 25 question including 4 no-answer question, and 20% testing sets with 52 question. 15% of the total questions are designed to have no corresponding answer in the Qur'an, termed as zero-answer questions to raise the challenge of the model's understanding. The Query Relevance Judgements (QRels) dataset includes 1,132 'gold standard' passage IDs from the Qur'an, each associated with a specific question from the AyaTEC dataset (Malhas and Elsayed, 2020) (Malhas, 2023). For questions that have no

Dataset	Structure	Preprocessing Applied
QPC	<passage-id> <passage-text>	None
Training, Dev, and Test	<question-id> <question-text>	None
QRels Gold	<question-id> Q0 <passage-id> <relevance>	None
Qur'an English Translation	<sura-id> <aya-id> <translation>	Cleaning
Questions (Post-Augmentation)	<question-id> <question-text> <question-en> <question-versions>	Translation and Paraphrasing

Table 1: Dataset Formatting and Structure

answer in the Qur'an, a placeholder value of "-1" is assigned as the passage ID.

Our datasets employ tab-delimited formatting and undergo different types of preprocessing. The architecture of these datasets is described in Table 1. We applied two primary components in our system for question preprocessing: translation and paraphrasing. The resulting structure of the question file post-augmentation is also outlined in Table 1.

We also used the English translation of the meanings of the Qur'an dataset from the Rowwad Translation Center (qur, 2023). It has a total of 6236 records, which represent the translation of every verse in the Quraan. The Ruwwad Centre for Translation has carefully examined each Arabic verse, consulting multiple sources of Arabic Tafseer and grammar. They have opted for modern phrasing and strived to maintain an arrangement that mirrors the original Arabic sequence as closely as possible.

3.2 Model Setup

The proposed cross-lingual model architecture is depicted in Figure 1, and its components are explained in detail. The general components of the model are the English translation module, paraphrasing module, and information retrieval module which is based on the sentence-transformer model.

For the translation and paraphrasing, we used OpenAI ChatCompletion API, gpt-3.5-turbo model, and the prompts: "You will be provided with a sentence in Arabic, and your task is to translate it into English." And "You will be provided with an English question, and your task is to paraphrase it." Respectively for each task. The temperature of the model is 0.9, with 150 maximum tokens. The translation process was proposed to enhance the quality of the processing of the used models, as they performed poorly in Arabic directly. The paraphrasing was proposed to enhance further the accuracy of the answer retrieved.

The retrieved documents of different paraphrases are aggregated and sorted according to their similarity scores, eliminating duplicate documents in case the same document is retrieved from multiple paraphrases. The model handles no-answer questions by setting a threshold value of similarity score in an attempt to eliminate irrelevant documents. such that a document is accepted as an answer if its score exceeds the threshold value. The threshold value was determined according to the analysis conducted during the model experimentation.

The information retrieval model was built using a semantic search (Reimers, 2022). It is also known as dense retrieval, which transforms the search query into a vector representation and identifies document embeddings that are proximate in the vector space. The lexical search seeks exact word-for-word matches of the query terms within the set of documents, failing to account for synonyms and acronyms. Semantic search, on the other hand, converts the search query into a vector format and fetches document embeddings that are close to that vector space.

The initial retrieval system could fetch documents that may not be highly relevant to the search query. To address this, a second-layer re-ranker is employed, which uses a cross-encoder to evaluate and score the relevance of all candidate documents in relation to the specified search query as shown in Figure 1.

In our study, we employ two distinct models to assess the efficacy of document retrieval in a question-answering context. Model A which is a Semantic Search that employs 'msmarco-distilbert-base-tas-b'[1] sentence Transformer model as the bi-encoder, 'cross-encoder/ms-marco-MiniLM-L-6-v2'[2] model as the cross encoder. Model B which is a semantic search that employs OpenAI's best embeddings 'text-embedding-ada-002' engine as the bi-encoder and OpenAI's 'text-davinci-003" engine[3] as the cross encoder. The two models serve

[1] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-tas-b

[2] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2

[3] https://platform.openai.com/docs/models/overview

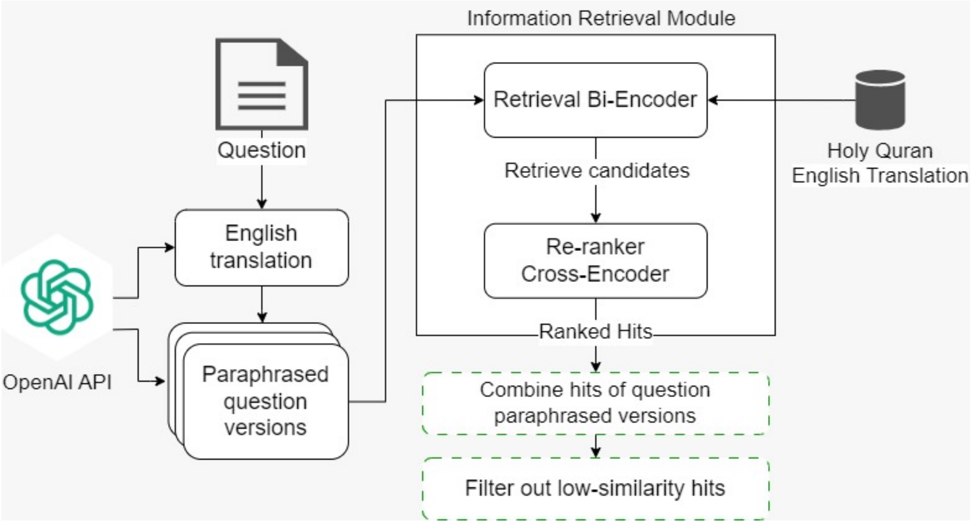
Information Retrieval Module
Question
English translation
OpenAI API
Paraphrased question versions
Retrieval Bi-Encoder
Retrieve candidates
Re-ranker Cross-Encoder
Ranked Hits
Holy Quran English Translation
Combine hits of question paraphrased versions
Filter out low-similarity hits

Figure 1: Model Architecture of the Passage Retrieval.

as a comprehensive setup incorporating a translation component to facilitate multilingual query processing. Built on an advanced neural network architecture, those models with Translation are capable of understanding and interpreting queries in the Arabic language. The translation feature allows it to translate the queries into a common language, perform the search, and then translate the results back into the original language, if necessary. In this work, we elaborate on the three setups used in our experiments: model A with translation and paraphrasing. model A with translation and no paraphrasing, and model B with translation and no paraphrasing.

3.3 Experiments Setup

All the pre-trained models were used in a zero-shot manner. With no fine-tuning on the dataset explained in Dataset Description. The primary metric for evaluation is the Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR). All the experiments were carried out on a single T4 GPU and implemented using Google Collaboratory. We will use our built model with the translated question and no paraphrasing as the baseline for comparison.

4 Result and Discussion

In our research, we initially focused on assessing the capabilities of the semantic search integrated with a translation component. The metrics used for performance evaluation included MRR and MAP. The results of the three proposed models are shown in Table 2. According to the scores on the dev set, the SBERT semantic search without paraphrasing was the best-performing model, with a MAP score of 0.343 and an MRR score of 0.413. When it comes to the test set, the SBERT semantic search without paraphrasing had the highest MAP score of 0.132, while the OpenAI semantic search without paraphrasing had the highest MRR score of 0.389.

Model	Metric	Dev	Test
SBERT with paraphrasing	MAP MRR	0.284 0.408	0.120 0.291
SBERT without paraphrasing	MAP MRR	0.343 0.413	0.132 0.302
OpenAI without paraphrasing	MAP MRR	0.221 0.369	0.199 0.389

Table 2: Performance results for the three proposed models: SBERT semantic search with paraphrasing, SBERT semantic search without paraphrasing and OpenAI semantic search without paraphrasing.

Our findings indicate that the translation-augmented version exhibited significant improvements across all metrics when compared to the model without translation. For instance, the MAP score witnessed an increase from 0.003 using an Arabic sentence transformer model 'medmediani/Arabic-KW-Mdel'[4] to 0.343 using the English sentence transformer model 'msmarco-distilbert-base-tas-b' on the dev set, suggesting that the translation component greatly enhanced the model's ability to retrieve more relevant documents. Overall, integrating translation into the system substantially improved its performance, validating our

[4] https://huggingface.co/medmediani/Arabic-KW-Mdel

hypothesis that translation is a crucial element for improving retrieval quality in the Arabic question-answering (QA) environment.

Building on this, we also introduced a second model that involved multiple paraphrased versions of the input question for even more precise retrieval. The results of the versions of the question produced by the paraphrasing component were sorted, and the duplication in the retrieved answers was deleted. The result of both T-test (Semenick, 1990) and Mann-Whitney U test (McKnight and Najab, 2010) shows no significant difference in MAP and MRR scores with adding the paraphrasing component to the base model.

In the case of questions with no answer, the test set contained 7 questions with no answers, the best model was able to correctly say 'No answer' to four questions, 0.57 of the questions. The threshold value for eliminating irrelevant documents is set to -5, where documents with a score of -6 and below are considered irrelevant.

The test set has in total 7 questions that did not have corresponding answers (no-answer questions). Interestingly, out of these 7 questions, our best-performing model accurately identified 'No answer' for 4 of them, giving us a 57% accuracy rate in this specific context. In order to filter out irrelevant documents, we established a threshold value of -5, which means that any documents scoring -6 or lower were considered irrelevant.

Some questions in the test set are not direct and cannot be solved with similarity measures but rather require some inference methodology to infer the question from the given context.

Certain questions within the test set are indirect and present challenges when addressed through similarity measures. To effectively tackle these questions, a more nuanced approach, specifically an inference methodology, is necessary in order to ascertain the intention of the question from the given context.

5 Conclusion

In this study, we explored the linguistic complexity of the Holy Qur'an, which holds profound influence over approximately two billion Muslims worldwide. Our engagement in the Qur'an QA 2023 Shared Task's PR task (Task A) led us to employ two distinct models: the Sentence Transformer and OpenAI's embeddings, both aimed at effective document retrieval. A significant feature of our approach was the integration of a translation mechanism to facilitate the interpretation of Arabic queries. Upon evaluation, the translation-enhanced model showcased superior performance across all metrics in comparison to its non-translation counterpart.

References

2023. Rowwad translation center, the noble qur'an encyclopedia.

Heba Abdelnasser, Maha Ragab, Reham Mohamed, Alaa Mohamed, Bassant Farouk, Nagwa M El-Makky, and Marwan Torki. 2014. Al-bayan: an arabic question answering system for the holy quran. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 57–64.

Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen. 2008. The evaluation of sentence similarity measures. In *Data Warehousing and Knowledge Discovery: 10th International Conference, DaWaK 2008 Turin, Italy, September 2-5, 2008 Proceedings 10*, pages 305–316. Springer.

A Adeleke, NA Samsudin, ZA Othman, and SK Ahmad Khalid. 2019. A two-step feature selection method for quranic text classification. *Indonesian Journal of Electrical Engineering and Computer Science*, 16(2):730–736.

Fatima T AL-Khawaldeh. 2015. Answer extraction for why arabic questions answering systems: Ewaq. *World of Computer Science & Information Technology Journal*, 5(5).

Hamza Alami, Abdelkader Mahdaouy, Abdessamad Benlahbib, Noureddine En-Nahnahi, Ismail Berrada, and Said El Alaoui Ouatik. 2023. Daqas: Deep arabic question answering system based on duplicate question detection and machine reading comprehension. *Journal of King Saud University-Computer and Information Sciences*, page 101709.

Jawharah Saeed N Alasmari. 2020. *A Comparative Analysis of The Arabic and English Verb Systems Using the Qur'an Arabic Corpus [A corpus-based study]*. Ph.D. thesis, University of Leeds.

Tahani H Alwaneen, Aqil M Azmi, Hatim A Aboalsamh, Erik Cambria, and Amir Hussain. 2022. Arabic question answering system: a survey. *Artificial Intelligence Review*, pages 1–47.

Eric Atwell, Claire Brierley, Kais Dukes, Majdi Sawalha, and Abdul-Baquee Sharaf. 2011. An artificial intelligence approach to arabic and islamic content on the internet. In *Proceedings of NITS 3rd National Information Technology Symposium*, pages 1–8. Leeds.

Wided Bakari and Mahmoud Neji. 2022. A novel semantic and logical-based approach integrating rte technique in the arabic question–answering. *International Journal of Speech Technology*, 25(1):1–17.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Rana Malhas. 2023. *Arabic Question Answering on the Holy Qur'an*. Doctoral dissertation.

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.

Patrick E McKnight and Julius Najab. 2010. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.

Ensaf Hussein Mohamed and Eyad Mohamed Shokry. 2022. Qsst: A quranic semantic search tool based on word embedding. *Journal of King Saud University-Computer and Information Sciences*, 34(3):934–945.

Masnizah Mohd, Faizan Qamar, Idris Al-Sheikh, and Ramzi Salah. 2021. Quranic optical text recognition using deep learning models. *IEEE Access*, 9:38318–38330.

Cicero Nogueira dos Santos, Xiaofei Ma, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Beyond [CLS] through ranking by generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1722–1727, Online. Association for Computational Linguistics.

Jay M Ponte and W Bruce Croft. 2017. A language modeling approach to information retrieval. In *ACM SIGIR Forum*, volume 51, pages 202–208. ACM New York, NY, USA.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers. 2022. Sbert semantic searchs.

Doug Semenick. 1990. Tests and measurements: The t-test. *Strength & Conditioning Journal*, 12(1):36–37.

M. N. Swar. 2007. *Mushaf Al-Tafseel Al-Mawdoo'ee*. Dar Al-Fajr Al-Islami.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.