LowResContextQA at Qur'an QA 2023 Shared Task: Temporal and Sequential Representation Augmented Question Answering Span Detection in Arabic

Hariram Veeramani
Department of Electrical and Computer Engineering, UCLA, USA
hariram@ucla.edu

Surendrabikram Thapa
Department of Computer Science, Virginia Tech, Blacksburg, USA
sbt@vt.edu

Usman Naseem
College of Science and Engineering, James Cook University, Australia
usman.naseem@jcu.edu.au

Abstract

The Qur'an holds immense theological and historical significance, and developing a technology-driven solution for answering questions from this sacred text is of paramount importance. This paper presents our approach to task B of Qur'an QA 2023, part of EMNLP 2023, addressing this challenge by proposing a robust method for extracting answers from Qur'anic passages. Leveraging the Qur'anic Reading Comprehension Dataset (QRCD) v1.2, we employ innovative techniques and advanced models to improve the precision and contextuality of answers derived from Qur'anic passages. Our methodology encompasses the utilization of start and end logits, Long Short-Term Memory (LSTM) networks, and fusion mechanisms, contributing to the ongoing dialogue at the intersection of technology and spirituality.

1 Introduction

The Holy Qur'an considered the central religious text of Islam, is a source of profound wisdom, guidance, and spiritual insight for millions of people around the world (Touati-Hamad et al., 2020). Its rich and complex content spans a wide range of topics, encompassing historical narratives, moral teachings, legal principles, and metaphysical concepts (Ahmed and Atwell, 2016). For devout Muslims, seeking knowledge and understanding from the Qur'an is a fundamental aspect of their faith, and it serves as a cornerstone for theological, ethical, and philosophical discourse (Malhas et al., 2022).

In the age of information technology, the quest for a deeper comprehension of the Qur'an has extended beyond traditional exegesis, embracing digital tools and computational approaches (Bashir et al., 2023; Malhas and Elsayed, 2022; Ahmed and Atwell, 2016; Mohamed and El-Behaidy, 2021; Veeramani et al., 2023b,d,e). One such critical task in this domain is Qur'anic question-answering (QA), which bridges the sacred text with modern technology and linguistic analysis (Malhas et al., 2022). The goal of Qur'anic QA is to enable the retrieval of specific, contextually relevant answers (Malhas et al., 2022; Malhas and Elsayed, 2022, 2020) to a wide range of questions from the Qur'an's voluminous text.

This paper addresses the pressing need to develop and refine QA systems tailored for Qur'anic texts. In this paper, we provide a detailed description of our system for task B of the Qur'an QA 2023 shared task (Malhas et al., 2023). The task at hand involves providing accurate, contextually appropriate answers to questions posed in Modern Standard Arabic (MSA) regarding specific Qur'anic passages. These passages consist of consecutive verses from a particular Surah (chapter) of the Qur'an. The complexity of this task arises from the need to extract precise answers directly from the provided passage, ensuring that the responses are contextually relevant and adhere to the theological and linguistic nuances of the Qur'an.

Our model uses start and end logits, augmented by employing two model variants. Using two separate question-answering models enables us to explore different aspects of the task, capitalizing on the strengths of each model to ensure comprehensive coverage and accuracy in answer extraction. To further enhance the accuracy and relevance of our system, we pick the best start-end logits with Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) and multilayer perception. In this endeavor, we aim to advance both Natural Language Processing (NLP) and the accessibility of the Qur'an's profound wisdom. Our work not only provides a bridge between technology and spirituality but also strives to make the wealth of knowledge contained within the Qur'an more accessible to individuals seeking answers to a wide array of inquiries, whether they be of a religious, historical, or ethical nature. Addi-

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 708–713

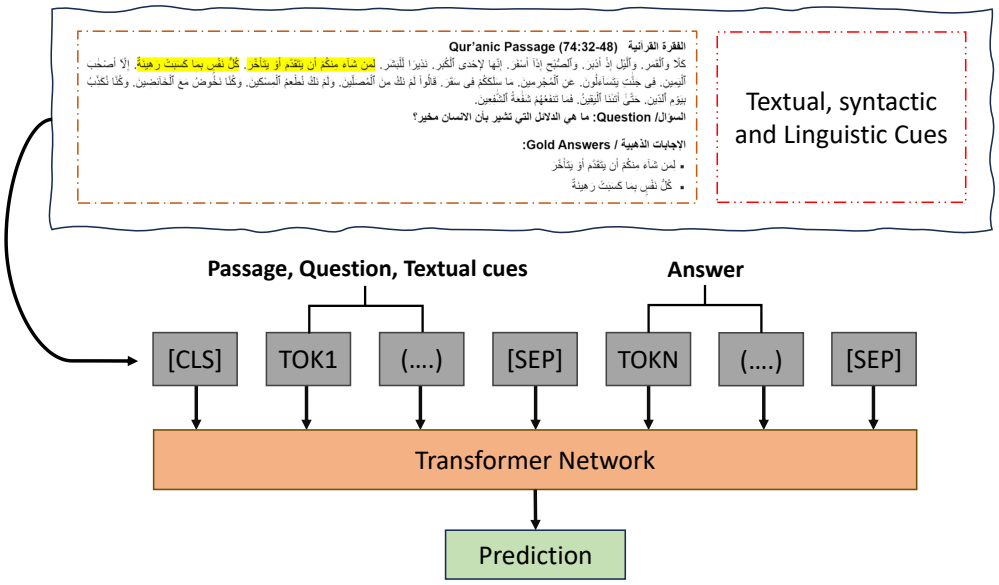

Figure 1: A brief overview of a question-answering model for extracting answers from Qur'anic passages.

tionally, it offers a valuable resource for scholars, educators, and researchers engaged in Qur'anic studies, empowering them to navigate the text efficiently and extract pertinent information.

2 Task Description

We only participated in Task B of Qur'an QA 2023. For task B, given a specific passage from the Qur'an consisting of consecutive verses within a particular Surah, along with a free-text question posed in MSA regarding that passage, the system's objective is to extract all answers to the question that are explicitly stated within the provided passage. The answers extracted must be in the form of text spans directly sourced from the given passage. In order to enhance the task's realism and difficulty level, some questions may not have a corresponding answer within the provided passage. In such instances, the ideal system should return no answers. Conversely, when there are answers present in the passage, the system should return a ranked list of up to 10 answer spans that are relevant to the question.

The evaluation measure utilized for this task is partial Average Precision (pAP) (Malhas and Elsayed, 2022). This metric plays a central role in assessing the performance of Question-Answering (QA) systems by incorporating partial matching. It acknowledges and rewards QA systems that retrieve answers that may not necessarily occupy the top rank and may only partially match one of the gold-standard answers. Additionally, pAP is particularly well-suited for evaluating questions that may have one or more valid answers within the accompanying passage. For questions where no answer exists within the provided passage, the evaluation approach is straightforward. A "no-answer" system output is granted full credit, while any other response is assigned a score of zero. To arrive at an overall evaluation score, the pAP measure is calculated and averaged across all questions, providing a comprehensive assessment of the QA system's performance. This metric is designed to capture the system's effectiveness in terms of accuracy and ranking relevance, offering a holistic view of its capabilities in the context of Qur'anic text-based question-answering.

3 Dataset

Task B utilizes the QRCD (Qur'anic Reading Comprehension Dataset) v1.2. This dataset (Malhas and Elsayed, 2020, 2022) currently consists of 1,155 question-passage pairs, forming 1,399 question-passage-answer triplets. The data split for training, development, and test sets is targeted at 70%, 10%, and 20%, respectively. A unique aspect of this dataset is the inclusion of "zero-answer questions", which make up 15% of the questions and are questions without answers in the Holy Qur'an. This addition aims to provide a more realistic and challenging reading comprehension task.

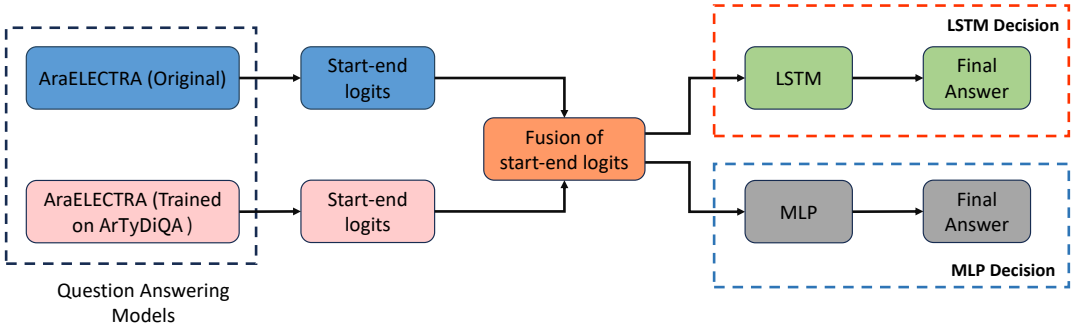

Figure 2: Framework of our methodology for extracting answers from Qur'anic passages.

4 System Description

In this section, we describe two different question-answering models we use along with our methodology. Figure 2 represents our approach for the Qur'an question-answering task.

4.1 Question Answering Models

QA models are a subset of NLP models designed to answer questions in human language automatically. They employ machine learning and deep learning techniques to understand questions and relevant text to extract suitable answers. In the context of Qur'an QA, Figure 1 provides an overview of our approach to extracting answers from Qur'anic passages.

AraELECTRA (Original)–Model 1: AraELECTRA (Antoun et al., 2021) is an Arabic language representation model that is pre-trained using the replaced token detection (RTD) objective. This objective is similar to the masked language modeling (MLM) objective used by other pre-trained language models, such as BERT and RoBERTa (Liu et al., 2019; Gururangan et al., 2020; Veeramani et al., 2023c,a,f). However, instead of masking tokens in the input sequence, the RTD objective replaces some tokens with a special [MASK] token and then trains the model to distinguish the original tokens from the replaced tokens. It was pre-trained on a large corpus of Arabic text, including news articles, books, and social media posts. AraELECTRA has been shown to outperform previous Arabic language representation models on various natural language processing (NLP) tasks, including question answering, sentiment analysis, and named entity recognition. It is also smaller and faster than previous models, making it more suitable for deployment on resource-constrained platforms. In addition to the original model, we also test AraElectra-ARCD[1].

AraELECTRA-ArTyDiQA–Model 2: This version of AraELECTRA is trained on the extensive ArTyDiQA dataset, which offers several advantages for question answering. Firstly, its pre-training on ArTyDiQA, a substantial Arabic question-answering dataset, equips it with a strong grasp of the Arabic language's nuances and its usage in the context of question answering. This enhanced language understanding enables AraELECTRA-ArTyDiQA to comprehend the intent of questions better and effectively extract relevant information from the corpus. Additionally, as AraELECTRA is built upon the ELECTRA architecture (Clark et al., 2019), it benefits from rapid and effective learning facilitated by the ArTyDiQA dataset, which adeptly captures the intricacies of Arabic question answering.

4.2 Answer Span Start-End Logits

In QA models, start and end logits are critical components that facilitate the extraction of answers from a given passage. These logits are computed for each token in the passage when the model analyzes a question and a text. They represent the likelihood that a token serves as the start or end point of the answer. By comparing these logits, the model identifies potential answer spans by selecting tokens with the highest combined scores. The final answer span is determined by choosing a continuous sequence of tokens with the highest joint likelihood based on the start and end logits. In some cases, QA models may further enhance answer selection by scoring and ranking multiple possible spans, ultimately presenting the span with the high-

[1] https://huggingface.co/salti/AraElectra-base-finetuned-ARCD

est overall score, which usually includes contextual information beyond just the logits. This mechanism ensures the model provides accurate and contextually relevant answers to the posed questions. In our case, we take start and end logits from both the models we used.

We employ two distinct QA model settings in our approach. In the first setting, we utilize start and end logits independently. These logits are processed by passing them through a Multi-Layer Perceptron (MLP) layer. This configuration allows each model to make individual predictions based on its understanding of the input, ensuring a level of independence in their responses. In the second setting, we introduce a fusion process for the start and end logits obtained from two separate models. These fused logits are then fed into the MLP layer. This fusion mechanism enables the models to collaboratively refine their predictions, potentially benefiting from the diverse insights each model offers.

We also utilize Long Short-Term Memory (LSTM) networks and MLP in our experiment. The LSTM component enhances the models' ability to capture temporal dependencies across passages/answers along with the sequential representation of the input data. It promotes local context understanding and global features, further optimizing the models' performance in delivering accurate and contextually relevant answers to the posed questions.

4.3 Decision Mechanism

We adopt MLP (Rosenthal et al., 2017; Kanagasabai et al., 2023)and LSTM to extract finer features to reinforce the confidence in picking the right start-end logit pair from one of the abovementioned models. In both our Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) components, we apply specific mechanisms to refine the models' predictions. For the LSTM, we utilize a softmax function. Softmax is employed to transform the LSTM output into a probability distribution over possible answer spans. This ensures that the model assigns a probability score to each pair of start-end logits, indicating its likelihood of being part of the answer span.

On the other hand, in the MLP layer, we employ an argmax of the computed class probabilities/classes to identify the answer's starting and ending points with the highest probability score. This is done along with the logits processed through the MLP. The argmax function selects the start-end logit pair with the highest predicted probability as the start of the answer span and the token with the highest predicted probability as the end of the answer span.

5 Results

In Table 1 showcasing results for Quran question answering, models are evaluated based on their performance measured in partial Average Precision (pAP). Model 1 and Model 2 achieve pAP scores of 0.367 and 0.406, respectively. Model 2 achieves a pAP score of 0.474 on the test set. A configuration combining both models (fused logits) using the LSTM branch achieves a pAP score of 0.411 during evaluation. The model configuration involving fusing logits with the MLP layer excels with a pAP score of 0.435 during evaluation, and we expect even better performance on the test dataset. Because of the unavailability of the test dataset, we only report the best submission score. Similar to the previous model, using the MLP branch but this time with AraElectra-ARCD instead of AraELECTRA (original), we achieved the highest pAP score of 0.442 on the evaluation dataset. These scores reflect the efficacy of different model configurations in Quranic question answering, with fused logits using the MLP branch displaying the highest overall performance, particularly on the test dataset.

Models	Eval	Test
Model 1	0.367	-
Model 2	0.406	0.474
Fused Logits (LSTM branch)	0.411	-
Fused Logits (MLP branch)	0.435	-
Fused Logits (AraElectra-ARCD + AraElectra) MLP Branch	**0.442**	-

Table 1: Results for various models with the dataset provided. All values are given with the pAP metric.

6 Conclusion

In summary, our paper contributes to developing precise Question-Answering (QA) systems for Qur'anic texts. By employing advanced techniques and models, we significantly improve answer accuracy and contextuality. Notably, certain model configurations, particularly those incorporating fused logits with the MLP branch, excel in achieving high partial Average Precision (pAP) scores across both evaluation and test datasets. This research not only

advances the field of Natural Language Processing (NLP) but also offers an invaluable resource for a diverse audience, ranging from scholars and educators to individuals seeking a deeper understanding of the Qur'an. It bridges technology and spirituality, promoting the harmonious integration of ancient wisdom with modern technology.

Limitations

This work exhibits several limitations. Firstly, the modest size of the QRCD dataset may restrict the models' full potential, warranting consideration for larger and more diverse Qur'anic text datasets. Furthermore, while our models aim for contextuality, capturing the intricate theological and linguistic nuances of the Qur'an remains an ongoing challenge. Addressing these limitations is essential to enhance the versatility and robustness of Question-answer models for Qur'anic texts and potentially expand their utility to broader NLP applications.

Ethics Statement

The Qur'anic text, being a sacred and religious source, is treated with the utmost respect and sensitivity. We have taken measures to ensure that our research and models align with cultural and religious considerations, and we do not engage in any activities that may cause harm or disrespect to any community or belief system. Additionally, we adhere to guidelines on data usage, compliance with applicable laws and regulations, and ethical conduct in research. We aim to contribute positively to the field of Natural Language Processing while promoting inclusivity, respect, and responsible use of technology.

References

Rasha Ahmed and ES Atwell. 2016. Developing an ontology of concepts in the qur'an. *International Journal on Islamic Applications in Computer Science and Technology*, 4(4):1–8.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Huzaifa Bashir, Aqil M Azmi, Haq Nawaz, Wajdi Zaghouani, Mona Diab, Ala Al-Fuqaha, and Junaid Qadir. 2023. Arabic natural language processing for qur'anic research: A systematic review. *Artificial Intelligence Review*, 56(7):6801–6854.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Rajaraman Kanagasabai, Saravanan Rajamanickam, Hariram Veeramani, Adam Westerski, and Kim Jung Jae. 2023. Semantists at ImageArg-2023: Exploring Cross-modal Contrastive and Ensemble Models for Multimodal Stance and Persuasiveness Classification. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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, Marseille, France. European Language Resources Association.

Rana Malhas, Watheq Mansour, and Tamer Elsayed. 2023. Qur'an QA 2023 Shared Task: Overview of Passage Retrieval and Reading Comprehension Tasks over the Holy Qur'an. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore.

Ensaf Hussein Mohamed and Wessam H El-Behaidy. 2021. An ensemble multi-label themes-based classification for holy qur'an verses using word2vec embedding. *Arabian Journal for Science and Engineering*, 46:3519–3529.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada. Association for Computational Linguistics.

Zineb Touati-Hamad, Mohamed Ridda Laouar, and Issam Bendib. 2020. Quran content representation in nlp. In *Proceedings of the 10th International Conference on Information Systems and Technologies*, pages 1–6.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023a. Automated Citation Function Classification and Context Extraction in Astrophysics: Leveraging Paraphrasing and Question Answering. In *Proceedings of the second Workshop on Information Extraction from Scientific Publications*, Online. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023b. DialectNLU at NADI 2023 Shared Task: Transformer Based Multitask Approach Jointly Integrating Dialect and Machine Translation Tasks in Arabic. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023c. Enhancing ESG Impact Type Identification through Early Fusion and Multilingual Models. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023d. KnowTellConvince at ArAIEval Shared Task: Disinformation and Persuasion Detection in Arabic using Similar and Contrastive Representation Alignment. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023e. LowResourceNLU at BLP-2023 Task 1 2: Enhancing Sentiment Classification and Violence Incitement Detection in Bangla Through Aggregated Language Models. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023f. Temporally Dynamic Session-Keyword Aware Sequential Recommendation system. In *2023 International Conference on Data Mining Workshops (ICDMW)*.