GYM at Qur'an QA 2023 Shared Task: Multi-Task Transfer Learning for Quranic Passage Retrieval and Question Answering with Large Language Models

Ghazaleh Mahmoudi [*1], **Yeganeh Morshedzadeh**[*2], **Sauleh Eetemadi**[1]
[1] School of Computer Engineering, Iran University of Science and Technology, Iran
[2] School of Engineering, The University of British Columbia, Canada
gh_mahmoodi@comp.iust.ac.ir, yeganeh.morshedzadeh@ubc.ca, sauleh@iust.ac.ir

Abstract

This work addresses the challenges of question answering for vintage texts like the Quran. It introduces two tasks: passage retrieval and reading comprehension. For passage retrieval, it employs unsupervised fine-tuning sentence encoders and supervised multi-task learning. In reading comprehension, it fine-tunes an Electra-based model, demonstrating significant improvements over baseline models. Our best AraElectra model achieves 46.1% partial Average Precision (pAP) on the unseen test set, outperforming the baseline by 23%.

1 Introduction

Question Answering (QA) for vintage, religious texts like the Quran presents unique challenges for natural language understanding systems. Understanding the concepts and connections in the Quran requires deep semantic reasoning to map questions to relevant passages and surface correct answers. To advance research in this domain, the Qur'anQA 2023 Shared Task [1] proposes two sub-tasks focused on machine comprehension of the Quran (Malhas et al., 2023).

Task A on passage retrieval requires matching Modern Standard Arabic (MSA) free-text questions to Quran verses potentially containing the answer. This tests semantic similarity between questions and passages. We propose using sentence encoders (Reimers and Gurevych, 2019) to derive dense vector representations for questions and passages. These vectors can be indexed and searched efficiently to find relevant matches.

Task B on reading comprehension focuses on extracting span answers from a given passage. This is framed as a machine reading comprehension task. However, given its literary Arabic and frequent need for theological reasoning, it is especially difficult for the Quran. We formulate the task as extractive QA and experiment with span prediction models like AraElectra (Antoun et al., 2021).

The Qur'anQA 2023 shared task includes two sub-tasks that form an end-to-end QA pipeline. Task A retrieves candidate passages potentially containing answers. This narrows the search space from the entire Quran to a small set of relevant verses. Task B then extracts answer spans from these candidates. The tasks work sequentially: passage retrieval provides context to reading comprehension, which verifies answers. Together they comprise an end-to-end QA system over the Quran.

Our key contribution to this work is utilizing transfer learning and model adaptation techniques to develop customized QA models for the limited Qur'anQA 2023 shared task dataset. After experimenting with several Arabic and multi-lingual language models (LMs) we choose AraElectra and AraBERT (Antoun et al.) as strong candidates. These models provide contextual representations of Arabic text learned from broad domains. In this work, we aim to address these research questions:

- How can we adapt LMs for Qur'anQA with limited task data?
- Which methods (e.g., transfer learning, data augmentation, unsupervised pretraining, etc.) improve the accuracy of the Quranic domain?

Through experiments, we analyze different strategies for unsupervised sentence embeddings and supervised task-specific fine-tuning. Despite the scarce training data, this allows the model to learn specialized embeddings for Quranic comprehension. Our work provides insights into adapting pre-trained language models to new domains with limited labeled data. By combining broad pre-trained knowledge with targeted fine-tuning, we develop customized QA models capable of reasoning about the Quran's abstract concepts and archaic language. The source code is available at GitHub[2].

[*] These authors contributed equally.
[1] https://sites.google.com/view/quran-qa-2023/home
[2] github.com/ghazaleh-mahmoodi/Quran-QA_2023_Shared-Task

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 714–719
December 7, 2023

2 Task A: Passage Retrieval

For a free-text question in MSA, the system must retrieve and rank Quranic passages that potentially contain answers to the question from a corpus covering the entire Quran.

3 Data

For this work, we utilize the training and development datasets provided by the Qur'anQA 2023 organizers (Malhas and Elsayed, 2020; Swar, 2007; Malhas, 2023), a summary of which is provided in Table 1. Across both train and development splits, there are 30 zero-answer questions, meaning that they have no answers in the Quran passages.

To augment the limited size of the Quran-specific data, we incorporate additional datasets during fine-tuning. For this passage retrieval task, we leverage the multi-lingual Mr. TyDi dataset, which contains monolingual question-passage pairs for information retrieval in 11 different languages (Zhang et al., 2021). We utilized the Arabic portion to fine-tune our proposed model.

Split	# Question	# Question-Passage Pairs
Training	174	972
Development	25	160
Test	52	-
All	251	1132

Table 1: Task A Dataset Distribution

3.1 System

Our implementation leverages the Sentence-Transformers framework (SBERT) (Reimers and Gurevych, 2019) to derive question and passage embeddings optimized for semantic similarity search. This provides an efficient method to match questions to relevant passages based on learned representations. SBERT provides a Siamese BERT network architecture optimized for semantic textual similarity. We used AraBERT, a BERT variant pre-trained on Arabic Wikipedia and news corpus.

To derive semantic vector representations of questions and Quran verses, the proposed passage retrieval approach trains a sentence embedding model, also known as a bi-encoder model. In order to achieve this, first, using **unsupervised** methods, AraBERT is fine-tuned on Quran passages to get sentence embedding. In the second step, the bi-encoder is trained on Mr. TyDi's Arabic dataset and Quran question-passage pairs using **supervised multi-task learning**.

3.1.1 Unsupervised Fine-Tuning: Learning Sentence Embedding

We experiment with TSDAE (Wang et al., 2021) and SimCSE (Gao et al., 2021) as the unsupervised training approach for encoding questions and passages.

TSDAE (Transformer-based Denoising Auto-Encoder) is a denoising Auto-Encoder trained to reconstruct corrupted passages, learning robust representations that capture semantic meaning.

SimCSE (Simple Contrastive Learning of Sentence Embeddings) is a contrastive self-supervised learning approach to derive passage embeddings. SimCSE is trained to predict a passage from itself, using only standard dropout as noise for data augmentation.

By transfer learning, these models learn robust passage representations that capture semantic meaning without the need for labeled data.

3.1.2 Supervised Fine-Tuning: Training Bi-Encoder using Question-Passage Pairs

After unsupervised fine-tuning convergence, a mean pooling and dense layer are added to the last layers of the bi-encoder. This bi-encoder is then fine-tuned end-to-end on Mr. TyDi and our question-passage pairs dataset. More specifically, the bi-encoder takes paired question and passage embeddings as input to predict relevance in a multi-task approach.

3.1.3 Model Specific Preparation

Models are trained with a combination of multiple negative ranking (Henderson et al., 2017), contrastive (Hadsell et al., 2006), and triplet (Dong and Shen, 2018) losses. As the models are trained in a multi-task manner, different loss functions are used for each dataset. A summary of the models is deprecated in Table 3. These three models were trained for 3 epochs with a batch size of 64, taking approximately 48 minutes in total on Nvidia GeForce RTX 3090 GPU.

As for the **Quranic question-passage** pairs, either a contrastive loss or triplet loss was incorporated:

- When using **contrastive loss**, we benefited from BM25 retrieval over the full corpus to mine negative passages for contrastive learning. More specifically, for each question in the training data, we first retrieve the top 1000 most relevant passages using BM25. We then label the ground truth

Model Name	Train Set		Development Set		Test Set	
	MAP	MRR	MAP	MRR	MAP	MRR
AraBERT-TSDAE-Contrastive	0.1502	0.3206	0.1365	**0.2613**	**0.0545**	**0.1581**
AraBERT-SimCSE-Contrastive	**0.6522**	**0.7646**	**0.1459**	0.2573	0.0315	0.1023
AraBERT-SimCSE-Triplet	0.5243	0.6580	0.1082	0.1693	0.0116	0.0356

Table 2: Task A MAP@10 and MRR@10 Results

passage associated with the question as positive examples (label 1). The BM25 retrieved passages that do not match any ground truth passages are used as hard negatives (label 0). Each <*question, positive passage, label=1*> and <*question, negative passage, label=0*> is added as a training example. By learning attempts to maximize similarity for positive pairs and minimize it for mined negatives.

• For **triplet loss**, similarly, BM25 is used to mine negatives but used in a different format and structure. Specifically, for each question, the top 100 BM25 retrieved passages are obtained. Then for each positive passage, negative passages are sampled to be used in forming of <*question, positive passage, negative passage*> triplets. Finally, for all of the question-passage pairs, multiple such triplets are created by pairing them with each possible negative passage from the BM25 results. Triplet loss optimizes the model to ensure the positive passage embedding is closer to the question than the negative passage.

For **Mr. TyDi**, the samples follow a format <*question, positive passage, negative passage*> and accordingly, multiple negative ranking loss function is used.

3.2 Results

To evaluate system performance, we report the official metrics of Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) on train, development, and test splits.

On the training set, our best-performing model is AraBERT-SimCSE-Contrastive, achieving a MAP@10 of 0.6522 and MRR@10 of 0.7646. Contrastive learning approaches generally outperform the triplet loss in our experiments. On the development set, AraBERT-SimCSE-Contrastive obtains the best MAP@10 of 0.1459 while AraBERT-TSDAE-Contrastive achieves the highest MRR@10 of 0.2613 score. Our top-performing model on the official test set is AraBERT-TSDAE-Contrastive, with a MAP@10 of 0.0545 and MRR@10 of 0.1581. Table 2 summarizes the full results on dataset distributions for top-10.

	ATC [3]	ASC [4]	AST [5]
TSDAE	✓	-	-
SimCSE	-	✓	✓
Denoising AE [6]	✓	-	-
Contrastive	✓	✓	-
Triplet	-	-	✓
Multiple Negative	✓	✓	✓
Quran Q-P [7]	✓	✓	✓
Mr. TyDi	✓	✓	✓

Table 3: Task A Models Summery

3.3 Discussion

Overall, our results demonstrate performance for passage retrieval on the Qur'anQA dataset. Observing the results of the development set indicates that the models are effective at retrieving all relevant passages containing name entities, which appeared in both the question and the passage. However, performance suffers for questions that are only relevant to a single obscure passage.

The unsupervised learning approaches of TSDAE and SimCSE both improve results compared to other methods we experimented with Arabic LMs. TSDAE in particular excels at ranking the relevant passages higher, leading to better MRR. This shows the value of its robust representations learned by reconstructing passages. The unsupervised fine-tuning allows the model to generalize better despite the limited size of the Quranic dataset.

4 Task B: Reading Comprehension

Given a Quranic passage that consists of consecutive verses in a specific Surah [8] of the Quran and a

[3] AraBERT-TSDAE-Contrastive (GYM_Run1)
[4] AraBERT-SimCSE-Contrastive (GYM_Run0)
[5] AraBERT-SimCSE-Triplet (GYM_Run2)
[6] Auto-Encoders
[7] Question-Passage
[8] A surah is a chapter in the Quran consisting of a set of verses revealed to the Islamic prophet Muhammad. There are 114 surahs in the Quran.

free-text question posed in MSA over that passage, a system is required to extract all answers to that question that is stated in the given passage.

4.1 Data

For Task B, we use Qur'anic Reading Comprehension Dataset (QRCD v1.2) (Malhas and Elsayed, 2022, 2020; Malhas et al., 2022) which consists of question-passage pairs combined with one or more annotated answers (15% of the questions have no answers). The dataset distribution is illustrated in Table 4.

Split	%	#Q	#Q-P	#Q-P-A
Training	70%	174	992	1179
Development	10%	25	163	220
Test	20%	51	431	-
All	100%	250	1586	1399

Table 4: Task B dataset distribution. #Q shows the number of questions. #Q-P shows the number of question-passage pairs. #Q-P-A shows the number of question-passage-answer triplets in the dataset.

4.2 System

Our solution for Task B is using the AraElectra-based model (Antoun et al., 2021) that is pre-trained on general domain Arabic language data. We propose two strategies for fine-tuning this model on the QRCD v1.2 dataset in addition to other complimentary datasets. The description of each model's training settings is summarized in Table 5. The hardware used is a GPU.1080Ti.xlarge with 31.3GB RAM. In the following sections, we briefly explain how we train each model.

4.3 Models Specifications

We chose **AraElectra-SQuADv2** (Ahmed, 2023a) model which is fine-tuned using the Arabic-SQuADv2.0 (Ahmed, 2023b) dataset. Specifically, AraElectra-SQuADv2 is the AutoModelForQuestionAnswering model from the transformers library in HuggingFace initialized with AraElectra model (Antoun et al., 2021). This model was trained on question-answer pairs, including unanswerable questions targeting QA task. We further fine-tuned this model using the QRCD v1.2 dataset (submitted as *GYM_Run0*).

We select **AraElectra-TyDiQA** (Ahmed et al., 2022) which fine-tuned on TyDi QA (Clark et al., 2020) dataset. Similarly, we fine-tuned this model on the QRCD v1.2 (submitted as *GYM_Run1*).

We incorporated **ensemble modeling** which is a machine-learning technique for combining multiple models in the prediction process. More specifically, by finding the top 10 answers using both AraElectra-SQuADv2 and AraElectra-TyDiQA, we can aggregate the given scores for all specified spans that are common among these runs/models (submitted as *GYM_ensemble*). The aggregation process works as follows:

I. We consider the output results of both AraElectra-SQuADv2 and AraElectra-TyDiQA models for each given question.
   - If the answers are the same, we sum the model's output scores.
   - Otherwise, we keep the answer without changing the score.

II. Finally, based on the newly calculated scores, we sort the output results of the two models and consider the top 10 outputs as the final output of the ensemble model.

	AraElec-SQuADv2	AraElec-TyDiQA
SQuADv2	✓	-
TyDiQA	-	✓
QRCD v1.2	✓	✓
Epoch	30	1
Batch Size	4	8
Max Seq Len [9]	256	256
Doc Strid [10]	64	64

Table 5: Task B train setting

4.4 Results

Reading Comprehension is evaluated with partial Average Precision (pAP), which accounts for partial matches and multiple answers. Our best configuration, AraElectra-SQuADv2, beats the task's baseline by 23.0% and reaches 48.5% pAP@10 on the dev set and 13.5% while achieving 46.1% pAP@10 on the test set (Table 6). Our experiments indicate that in comparison with other models, including an AraBERT, the AraElectra model gives better results on the Qur'anQA Task. Also, the use of the Arabic-SQuADv2.0 dataset, which is similar to QRCD v1.2, significantly improves the result.

4.5 Discussion

The results demonstrate that transfer learning from large Arabic NLP datasets (TyDiQA and SQuADv2) is an effective strategy for adapting models to Qur'anQA despite limited task-specific

[9]The maximum length of a feature.

[10]The authorized overlap between two part of the context when splitting is needed.

Model	Dev	Test
AraElectra-SQuADv2	**0.485**	**0.461**
Ensemble	0.481	0.458
AraElectra-TyDiQA	0.431	0.430
Baseline	0.255	0.326

Table 6: Task B pAP@10 result

training data. Pre-training on broad domains equips models like AraElectra with useful linguistic and semantic knowledge of Arabic that transfers well to Qur'anQA. Fine-tuning on the small QRCD v1.2 dataset provides the final layer of adaptation to handle Quranic syntax, terminology, and reasoning requirements.

Our best approach leverages Arabic SQuADv2 and is able to effectively identify questions with multiple answers and specify the start and end tokens of each answer. Among the answers, there were cases where the predicted answers overlap; hence, having a mechanism to handle overlapping predictions could improve the results. Additionally, it would be beneficial to optimize the model's confidence scores for predicting start and end tokens, such that falling below a threshold indicates no answer.

Overall, our results demonstrate promising multi-span extraction capabilities gained via pre-training on SQuADv2 data. However, enhancements to prediction post-processing and confidence modeling could further improve the handling of overlap and no-answer cases. This would move towards more human-like discernment of when extracted snippets represent valid or invalid answers.

Conclusion

This work demonstrates adapting LMs to Qur'anQA with limited data. Key techniques include unsupervised fine-tuning, negative sample extracting with BM25, multi-tasking, and transfer learning. For passage retrieval, unsupervised strategies like TSDAE and SimCSE improve ranking over training from scratch. In reading comprehension, leveraging Arabic SQuAD allows AraElectra to excel at span prediction despite scarce Quran annotations. Overall, leveraging additional datasets benefited models in both sub-tasks. We provide insights into tailoring state-of-the-art NLP techniques to learn specialized behavior for machine comprehension of the Quran's semantics given modest labeled data.

Limitations

The main constraint we faced was the lack of labeled data. To overcome this, we used similar non-Quranic datasets. While this affected the model's quality during training, it improved its ability to perform well on unseen data.

An important aspect to consider in the context of this research is the wealth of Tafsirs[11] available for the Quran, authored by religious scholars spanning different time periods and languages. These Tafsirs provide invaluable insights into the interpretations and nuances of the Quranic text, shedding light on the historical, linguistic, and cultural contexts in which the verses were revealed. The Quran, being a deeply layered and intricate scripture, often carries layers of meaning that extend beyond the literal words and Tafsirs help unravel these layers. Incorporating Tafsirs into the model's training data could enable it to better capture these nuanced interpretations and subtle connections, potentially leading to more accurate and contextually informed question-answering for vintage texts like the Quran.

Another challenge in passage retrieval we encountered was when the input question had no corresponding answer in the Quranic passages. In these cases, the model's performance suffered because we had to apply a threshold to the output scores, which were not fine-tuned specifically for this task. Additionally, the difference between the questions in Modern Standard Arabic (MSA) and the diverse variations of Quranic texts presented another challenge. One additional challenge we faced in this task was the lack of negative passages. To address this, we had to generate a set of negative passages using the BM25 method, as previously explained in detail. However, the quality of these negative passages plays a crucial role in the model's training. One approach we considered was to treat all passages, except the positive ones, as negatives. However, due to the imbalance between positive and negative samples and GPU limitations, we decided not to pursue this approach. But this approach can be examined in future work.

Acknowledgements

We'd like to thank the organizers for introducing this task. We are glad that we had the opportunity to engage more in the challenges and progress in the information retrieval task.

[11]Tafsir is Quranic exegesis that explains, interprets, contextualizes, or comments on Quran verses.

References

Basem Ahmed, Motaz Saad, and Eshrag A. Refaee. 2022. QQATeam at qur'an QA 2022: Fine-tunning Arabic QA models for qur'an QA task. 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 130–135, Marseille, France. European Language Resources Association.

Zeyad Ahmed. 2023a. Arabic Machine Reading Comprehension: Effective Models and Introducing Arabic-SQuAD v2.0. https://github.com/zeyadahmed10/Arabic-MRC, https://huggingface.co/ZeyadAhmed/AraElectra-Arabic-SQuADv2-QA. Original-date: 2021-11-04T18:03:17Z.

Zeyad Ahmed. 2023b. Arabic SQuAD v2.0 Dataset based on the popular SQuADv2.0 with unanswered questions for more challenging task. https://huggingface.co/datasets/ZeyadAhmed/Arabic-SQuADv2.0. Original-date: 2022-06-29T18:03:17Z.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

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.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Xingping Dong and Jianbing Shen. 2018. Triplet loss in siamese network for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *ArXiv*, abs/1705.00652.

Rana Malhas and Tamer Elsayed. 2020. Ayatec: Building a reusable verse-based test collection for arabic question answering on the holy qur'an. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(6).

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 *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, pages 79–87.

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.

Rana R Malhas. 2023. *ARABIC QUESTION ANSWERING ON THE HOLY QUR'AN*. Ph.D. thesis.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marwan N. Swar. 2007. *Mushaf Al-Tafseel Al-Mawdoo'ee*. Dar Al-Fajr Al-Islami, Damascus.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.