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*, School of Computer Engineering, Iran University of Science and Technology, Iran
Yeganeh Morshedzadeh*, School of Engineering, The University of British Columbia, Canada
Sauleh Eetemadi, School of Computer Engineering, Iran University of Science and Technology, Iran

Introduction

Question answering (QA) over religious texts like the Quran is challenging. Understanding the concepts requires mapping questions to relevant passages and extracting correct answers. This involves deep semantic reasoning with the complex, archaic language.

This work aims to advance question answering (QA) capabilities for limited old text of Quran. It focuses on developing models to address two key QA tasks for the Quran: passage retrieval and reading comprehension.

The passage retrieval task aims to match free-text queries written in Modern Standard Arabic to relevant verses in the Quran containing potential answers. The reading comprehension task then extracts answer spans from retrieved passages by reading and understanding the verses given a question.

The proposed approach leverages techniques like transfer learning, unsupervised and supervised fine-tuning, and ensemble modeling with large, pre-trained Arabic language models to specifically tailor them for Quranic semantics and reasoning.

We utilize transfer learning to adapt large pre-trained language models like AraElectra and AraBERT for these tasks. Despite limited Quran-specific data, unsupervised and supervised fine-tuning techniques allow the models to learn specialized embeddings that can be used for Quranic comprehension.

Task A: Passage Retrieval

Given a free-text question in Modern Standard Arabic (MSA), the task requires retrieving and ranking relevant Quran verses that potentially contain the answer. The system searches over the entire Quran corpus to find the most similar passages.

We first derive passage embeddings using unsupervised strategies like contrastive self-supervised learning (SimCSE) and reconstruction denoising autoencoders (TSDAE). A bi-encoder model is then trained to compare question and passage vectors, fine-tuned on external QA data. The models rank up to 1000 retrieved passages per question.

- **Data:** The data contains 30 zero-answer questions across splits. Augmented with Mr. TyDi Arabic QA dataset.

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

- **Methodology:** The system employs the Sentence-Transformers framework with the AraBERT contextual language model as a base. Additional unsupervised pretraining techniques like TSDAE and SimCSE are used to further tune the semantic sentence encodings on in-domain Quranic text. A bi-encoder model is then trained in a supervised manner using both the Quranic question-passage datasets as well as complementary out-of-domain Arabic question-passage data from the MR TyDi dataset. This multi-task learning leverages contrastive, triplet, and other losses to optimize question-passage similarity predictions.

		AraBERT-TSDAE-Contrastive	AraBERT-SimCSE-Contrastive	AraBERT-SimCSE-Triplet
Sentence Embedding	TSDAE	✓		
	SimCSE		✓	✓
Training Loss	Denoising Auto-Encoders	✓		
	Contrastive	✓	✓	
	Triplet			✓
	Multiple Negative	✓	✓	✓
Dataset	Quran Question-Passage	✓	✓	✓
	Mr TyDi	✓	✓	✓

Task A Models Summery

- **Results:**
  - The AraBERT–SimCSE-Contrastive model performs the best, achieving a Mean Average Precision (MAP) at 10 of 0.6522 on the training set and a Mean Reciprocal Rank (MRR) at 10 of 0.2613 on the development set. Top model for MRR metric:
  - The AraBERT–TSDAE-Contrastive configuration obtains the highest MAP@10 of 0.0545 and MRR@10 of 0.1581 on the official test set.

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

Task A MAP@10 and MRR@10 Results

Task B: Reading Comprehension

The reading comprehension task focuses on extracting span answers from retrieved Quran passages. Given a question and associated passage, the model must identify all text snippets that contain answers stated in the passage.

We formulate this as an extractive question answering problem and experiment with the AraElectra model architecture. AraElectra is first pre-trained on large general domain Arabic text before fine-tuning on the task.

- **Data**

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

- **Methodology:** The system fine-tunes variations of the AraElectra contextual language model, which has been pre-trained on broad Arabic textual data. Specifically, models adapted using the Arabic SQuAD v2 and TyDiQA machine reading comprehension datasets are employed. This allows the model to learn strong linguistic knowledge and adapt it to Quran QA. Ensembling predictions from both models is also evaluated.

	Dataset			Model and Environment Setting			
	SQuADv2	TyDiQA	QRCD v1.2	Epoch	Batch Size	Max sequence Length	Document Stride
AraElectra-SQuADv2	✓		✓	30	4	256	64
AraElectra-TyDiQA		✓	✓	1	8	256	64

Task B train setting

- **Result:**

The AraElectra model finetuned on Arabic SQuAD v2 performs the best, achieving 46.1% partial Average Precision (pAP) at 10 on the unseen official test set. This represents a 13.5% absolute improvement over the baseline. On the dev set, this model obtains 48.5% pAP@10.

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

Task B pAP@10 result

The results thus demonstrate promising multi-span extraction capabilities, gained from pre-training on SQuAD data.

Further improvements can be made in the post-processing to handle cases of overlapping predicted answers. Additionally, optimizing the model's confidence estimates for start and end tokens of spans could improve discerning valid versus invalid answers. But overall, transfer learning is shown to be an effective strategy for adapting to this domain.

Conclusion

This work demonstrates strategies for adapting large language models to question answering over the Quran given limited labeled data. Key techniques that are shown to be effective include unsupervised fine-tuning, negative sample mining, multi-task learning, and transfer learning.

For passage retrieval, unsupervised pretraining approaches like TSDAE and SimCSE allow models to learn better representations and rankings compared to training from scratch. For reading comprehension, transfer learning from Arabic SQuAD enables the AraElectra model to excel at answer span prediction, despite scarce Quran-specific annotations.

By leveraging external datasets and maximizing limited Quranic annotations, the models learn to map questions expressed in modern Standard Arabic to archaic Quranic passages and extract answers spanning obsolete terminology and abstract concepts. Overall, leveraging additional datasets is beneficial for improving model performance on the target Quran QA domain. The techniques presented provide insights into specializing state-of-the-art NLP models to comprehend Quranic semantics given modest labeled data.