LKAU23 at Qur'an QA 2023: Using Transformer Models for Retrieving Passages and Finding Answers to Questions from the Qur'an

Sarah Alnefaie[1][2], **Abdullah N. Alsaleh**[1][2],
Eric Atwell[2], **Mohammed Ammar Alsalka**[2], **Abdulrahman Altahhan**[2]
King Abdulaziz University[1]
University of Leeds[2]
{scsaln, scanaa, e.s.atwell, m.a.alsalka, a.altahhan}@leeds.ac.uk

Abstract

The Qur'an QA 2023 shared task has two sub tasks: Passage Retrieval (PR) task and Machine Reading Comprehension (MRC) task. Our participation in the PR task was to further train several Arabic pre-trained models using a Sentence-Transformers architecture and to ensemble the best performing models. The results of the test set did not reflect the results of the development set. CL-AraBERT achieved the best results, with a 0.124 MAP. We also participate in the MRC task by further fine-tuning the base and large variants of AraBERT using Classical Arabic and Modern Standard Arabic datasets. Base AraBERT achieved the best result with the development set with a partial average precision (pAP) of 0.49, while it achieved 0.5 with the test set. In addition, we applied the ensemble approach of best performing models and post-processing steps to the final results. Our experiments with the development set showed that our proposed model achieved a 0.537 pAP. On the test set, our system obtained a pAP score of 0.49.

1 Introduction

The Arabic language poses many challenges in Naturual Language Processing (NLP), including in the areas of Machine Reading Comprehension (MRC) and Passage Retrieval (PR). One of the most prominent recent NLP techniques applied to MRC and PR tasks in the Arabic language is pre-trained transformer-based models, which can achieve state-of-the-art performance (Alsubhi et al., 2021, 2022).

There are PR studies that use a dense approach based on pre-trained models (Karpukhin et al., 2020). This approach has outperformed traditional information retrieval, such as TF-IDF (Sammut and Webb, 2010) with Modern Standard Arabic (MSA; Alsubhi et al., 2022). To our knowledge, the dense approach has not been researched with Classical Arabic (CA). Therefore, our proposed system for Task A of the Qur'an QA 2023 shared tasks uses the dense approach by fine-tuning the Arabic pre-trained models and then ensemble the best scores. The idea of Task A is to build a system to return a list of Qur'anic passages that contain answers to a posed question/query (Malhas et al., 2023). However, the challenging aspect of this task is that there are some questions that do not have an answer in the Qur'an. The first research question **RQ1**: Does using the Arabic pre-trained models in PR for CA outperform the traditional approach such as BM25?

Most recent studies on the Qur'an MRC task have tended to use transformers-based models along with Qur'anic Reading Comprehension Dataset (QRCD) (Malhas et al., 2022). We noticed that they improved the performance of the systems using three approaches: (1) using an additional MSA and/or CA datasets in fine-tuning (Mostafa and Mohamed, 2022; Aftab and Malik, 2022), (2) constructing an ensemble of different BERT models (3) applying appropriate post-processing steps on the result of the final ranked list (ElKomy and Sarhan, 2022). To the best of our knowledge, no studies have combined these three approaches. Therefore, we applied the combination of those approaches for Task B of the Qur'an QA 2023 shared task. The goal of Task B was to build a model that took a Qur'anic passage and MSA question as input and extracted a ranked list of up to 10 answer spans to that question from the passage as output (Malhas et al., 2023). The new challenge in the second version of this task was that there were no answers to some questions. The second research questions **RQ2** in this paper is: Does the combination of fine-tuning the models with a large CA dataset and/or MSA dataset, ensembling these models and then applying post-processing steps improve the results?

The paper's structure is as follows: In Section 2, related work is presented. Section 3 describes the datasets. This is followed by Section 4, which

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 720–727

explains the proposed models. In Section 5, the results are discussed. Finally, the paper provides a conclusion.

2 Related Work

2.1 Task A: Passage Retrieval

Karpukhin et al. (2020) proposed their dense passage retrieval (DPR) system using BERT base and uncased models. Their system applies dual encoders for the passages to be transformed into dimensional real-valued vectors and then applies an index for all passages for retrieval. The input query is then encoded and mapped into the dimensional vector space and passages are retrieved that are near the query vector. Their approach outperformed other multiple open-domain QA techniques on several QA datasets such as TriviaQA and SQuAD. Sachan et al. (2022) proposed the unsupervised passage re-ranker (UPR), in which the system utilizes zero-shot question generation for re-ranking passages in order to improve passage retrieval. It then computes the relevance scores over the generated question and sort the results. Their approach outperformed DPR (Karpukhin et al., 2020) on several datasets, such as SQuAD and TriviaQA. Finally, Alsubhi et al. (2022) proposed a multilingual DPR model that was fine-tuned on Arabic datasets. Their model outperformed TF-IDF on Arabic datasets, which were ARCD (Mozannar et al., 2019) and TyDiQA-GoldP (Clark et al., 2020).

2.2 Task B: Machine Reading Comprehension

Recently, several researchers have built an MRC system to answer questions about the Qur'an. All these studies used QRCD_v1.1 in the fine-tuning and evaluation phases (Malhas et al., 2022; Malhas and Elsayed, 2022). Some studies have proposed further fine-tuning the model using MSA datasets (Mostafa and Mohamed, 2022; Malhas and Elsayed, 2022). Mostafa and Mohamed (2022) developed the Arabic Qur'an MRC model by fine-tuning the AraELECTRA model using three MSA datasets: Ar-TyDi, Arabic-SQuAD and Arabic Reading Comprehension Dataset (ARCD). Their model achieved a 0.54 pRR, 0.52 F1@1 and 0.23 EM. Other studies have proposed fine-tuning the model using the CA dataset. Sleem et al. (2022) fine-tuned AraBERTv02 using the Arabic Al-Qur'an Question and Answer Corpus (AQQAC) (Alqahtani, 2019). This model achieved scores of 0.52 pRR, 0.5 F1@1 and 0.25 EM.

ElKomy and Sarhan (2022) recommends using the training and development sets of QRCD_v1.1 to fine-tune five different Arabic BERT models. They then used these five models individually to find the answers for the QRCD test set. To obtain good results, they implemented an ensemble approach for the results of these models. Finally, post-processing was applied to enhance the results. The results showed a pRR of 56.6, an EM of 26.8 and F1@1 of 0.50.

To the best of our knowledge, no study has been conducted on the impact of the combination of the following three factors in building the Arabic Qur'an MRC model: First, Arabic pre-trained models are fine-tuned using CA and MSA datasets. Second, the ensembling approach was applied to the results using the majority vote. Finally, the final list was refined through several post-processing steps.

3 Datasets

3.1 Task A: Passage Retrieval

The data were comprised of the Qur'anic passage collection (QPC) and questions from AyaTEC (Malhas and Elsayed, 2020). The QPC was developed by segmenting the Qur'an passages into topics, which resulted in 1,266 passages. There were 199 questions that were derived from the AyaTEC dataset. The Query Relevance Judgements (QRels) dataset contained 1,132 gold (answer-bearing) Qur'anic passages that answered the questions; these data were used in training and development sets. Finally, the distribution of the dataset was 70%, 10% and 20% for training, development and testing sets respectively.

3.2 Task B: Machine Reading Comprehension

In this study, we used three different datasets, as follows:

QRCD: QRCD_v1.2 consists of 1,399 question–passage–answer triplets in the training and development splits, as shown in Table 6. It was split 70%, 10%, and 20% for the training, development and test sets respectively (Malhas and Elsayed, 2022, 2020).

ARCD: It consists of 1,395 question–passage–answer triplets for Wikipedia passages (Mozannar et al., 2019).

Quran Question–Answer pairs (QUQA): It consists of 3,382 question–passage–answer triplets regarding the Arabic Holy Qur'an. This dataset was built using the available Qur'an AQQAC dataset

Model	Encoder	MAP	MRR
BM25 (Robertson and Zaragoza, 2009)	-	0.17	0.313
ArabicBERT (Safaya et al., 2020)	bi-encoder	**0.511**	**0.687**
	cross-encoder	0.292	0.452
CL-AraBERT (Malhas and Elsayed, 2022)	bi-encoder	0.489	0.7
	cross-encoder	0.318	0.481
AraBERT (Antoun et al.)	bi-encoder	0.461	0.662
	cross-encoder	0.351	0.54
CAMeL-BERT (Inoue et al., 2021)	bi-encoder	0.455	0.606
	cross-encoder	0.351	0.505
Ensemble ArabicBERT & CL-AraBERT	bi-encoder	**0.534**	**0.73**
Ensemble ArabicBERT & CL-AraBERT & CAMeL-BERT	bi-encoder	0.487	0.688
Ensemble ArabicBERT & CL-AraBERT & AraBERT	bi-encoder	0.485	0.682

Table 1: The results of the development set by BM25, individual Arabic pre-trained models and the ensemble method. MAP is the official evaluation metric. The cross-encoder is used for re-ranking the list of answers output by the bi-encoder.

(Alqahtani, 2019) and five available books. It is available in the Github repository. [1]

4 Proposed Models

4.1 Task A: Passage Retrieval

Sentence transformers, also known as Sentence-BERT (SBERT), introduced a bi-encoder that transforms a pair of sentences independently and maps them to a dense vector for efficient comparison when performing an information retrieval task (Thakur et al., 2021). Our proposed system uses a bi-encoder method for a semantic search task by further training Arabic pre-trained models with the QRCD_v1.1 (Malhas et al., 2022). We also used the cross-encoder "mmarco-mMiniLMv2-L12-H384-v1" [2] for re-ranking; however, it did not improve the performance of the individual models.

Training the Models: We trained a set of four models using the SBERT architecture with Arabic pre-trained models: ArabicBERT (Safaya et al., 2020), CAMeL-BERT (Inoue et al., 2021), AraBERT (Antoun et al.) and CL-AraBERT (Malhas and Elsayed, 2022). Two datasets were used for training the models: the training set of Task A and the QRCD_v1.1. Since most of the data were duplicated between the QRCD_v1.1 and the training set of PR task, we used the NoDuplicatesDataLoader function to remove any copies prior to training. We used the MultipleNegativesRankingLoss (MNRL) loss function, as it allowed for two similar or positive sentences without labels to be computed. Finally, the QPC dataset were encoded for each model. All the models used the following parameters: 5 epochs, a learning rate of 2e-5, max length 512 and batch size of 16.

Ensemble Approach: The ensemble method used for this task was to retrieve the top 20 answers from the Arabic pre-trained models. If the answer was found in all outputs, we then summed up the scores and divided by the number of models to obtain the average score. These answers were then put at a top-ranked list by descending order of averaged score. If there were remaining places in the ranked list, we added answers that had the highest scores out of all the models. Finally, we capped the ranked list at 10 answers [3].

4.2 Task B: Machine Reading Comprehension

The pre-trained transformer-based models were the basis of our methodology. As a first step, we fine-tuned all available Arabic pre-trained models with the QRCD_v1.2 training set. There were nine Arabic pre-trained models: AraBERT base, AraBERT large, CAMeL-BERT, ArabicBERT, CL-AraBERT, AraELECTRA (Antoun et al., 2021), MARBERT, ARBERT (Abdul-Mageed et al., 2021) and QARiB (Abdelali et al., 2021). When we conducted our experiments, we set the batch size to 8 for AraBERT large and 16 for the rest of the models, the number of epochs to 4, and the learning rate to 1e-4. We

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

[2] https://huggingface.co/nreimers/mmarco-mMiniLMv2-L12-H384-v1

[3] The code can be accessed here https://github.com/AlsalehAbdullah/Quran_PR_Task

Model	QRCD	QRCD +QUQA	QRCD +ARCD	QRCD +QUQA +ARCD
AraBERT Large	0.165	**0.482**	0.162	-
AraBERT Base	0.402	**0.458**	**0.433**	**0.49**
MARBERT	**0.326**	0.089	0.291	-
ARBERT	0.357	**0.38**	0.343	-
QARiB	**0.307**	0.301	0.278	-
CAMeL-BERT	0.401	**0.406**	0.362	-
ArabicBERT	**0.332**	0.330	0.313	-
AraELECTRA	**0.332**	0.248	0.218	-
CL-AraBERT	0.373	**0.383**	0.358	-

Table 2: The pAP@10 result of fine-tuned different Arabic pre-trained models by using different combinations of the datasets.

attempted to improve the performance using the following three optimisation approaches [4]:

Transfer Learning: We conducted three experiments using this approach. We further fine-tuned the models using different datasets. In the first experiment, we used the CA dataset QUQA. Second, the MSA ARCD was used. Finally, a combination of the QUQA dataset and ARCD was used only for the models that showed an improvement in performance when using one of these two datasets individually.

Ensemble Approach: We used majority voting among the models to produce the final ranked-list results. We took the top 20 answers with their scores for each question from each model. We then computed the total score for each answer. The total score was the sum of the scores obtained from the answers from all models. After that, we sorted the answers for each question based on the total score. Finally, we adopted the top 10 answers as the final ranked list.

Post-Processing: There were two issues when producing the ranked list: uninformative answers (as shown in Figure 1) and overlapping answers (as shown in Figure 2). The first issue was solved by removing these answers from the list. The second was overcome by applying a redundancy elimination algorithm (ElKomy and Sarhan, 2022).

5 Results and Discussion

5.1 Task A: Passage Retrieval

The official evaluation metric used for this task was mean average precision (MAP), but the mean reciprocal rank (MRR) was also reported.

[4]The code can be accessed here https://github.com/scsaln/RC

Model	pAP@10
Ensemble **Vanilla** (All)	0.466
Ensemble **Vanilla** (Best)	0.517
Ensemble **POST** (Best)	**0.537**

Table 3: The results of the ensemble approach. Ensemble **Vanilla** (All) refers to applying the ensemble approach to all models. Ensemble **Vanilla** (Best) represents applying the ensemble approach to the best two performed models (the bert-large-arabertv02 and the bert-base-arabertv02). Ensemble **POST** (Best) refers to the **Vanilla** (Best) after applying the postprocessing step.

Validation: As for the validation results, the BM25 scored the lowest, with a 0.17 MAP. As for the pre-trained models, ArabicBERT performed the best of the individual models using a bi-encoder with a 0.511 MAP, while the ensemble of ArabicBERT and CL-AraBERT performed the best with the validation set with 0.534. Therefore, to address **RQ1**, the Arabic pre-trained models outperformed BM25 (See Table 1).

Testing: For the test set, we chose three methods based on their performances with the validation set. They were: ArabicBERT, CL-AraBERT and the ensemble of ArabicBERT and CL-AraBERT. The test set results did not reflect the performances on the validation set, as it can be seen in Table 4. CL-AraBERT performed the best with a 0.124 MAP while the performance of the ensemble method was a close second with a 0.117 MAP. The ensemble method and CL-AraBERT have successfully answered two questions with a perfect score of 1 MAP while 21 questions scored a 0 MAP. Some

of these happened to be a no-answer, which the models have failed to identify.

5.2 Task B: Machine Reading Comprehension

The evaluation metric for Task B was partial average precision (pAP) (Malhas and Elsayed, 2022, 2020).

Validation: Column QRCD in Table 2 presents the results of the models when they were fine-tuned using only the QRCD dataset. The AraBET base model outperformed the other models with a 0.402 pAP@10.

First, we addressed **RQ2**, which was related to whether the combination of the three factors enhanced the performance of the Qur'an MRC models. The first factor further fine-tuned the models using the CA dataset QUQA and/or MSA ARCD. The results are shown in columns 'QRCD + QUQA', 'QRCD + ARCD' and 'QRCD + QUQA + ARCD' in Table 2. There are three interesting observations in the results. First, using the QUQA dataset led to improvements in more than half of the models. The best score was the pAP@10 of 0.482, obtained by AraBERT large. Second, when we trained the model using the ARCD dataset it enhanced the performance of the AraBERT base model only with 0.433 pAP@10. Third, using QUQA and ARCD at the same time to train the AraBERT base improved results with 0.49 pAP@10 compared to using QUQA and ARCD separately. For the second factor, we used the ensemble method for all the models; however, this approach did not yield the best performance with a result of 0.466 pAP@10. We then ensembled two of the best performing individual models, which were AraBERT base and AraBERT large. The results outperformed the other models with 0.517. For the third factor, we note that the post-processing step improved the results based on the Ensemble '**POST** (Best)' row shown in Table 3.

Testing: For the test set, we chose two methods based on the performance of the development set. They were (1) the ensemble of AraBERT base and AraBERT large with post-processing and (2) the AraBERT base model. The ensemble with the post-processing approach achieved a 0.498 pAP@10, while the AraBERT base model achieved the best performance with a 0.5 pAP@10, as it can be seen in Table 5.

When we analysed the model answers to questions from the development set, we identified the following: The model worked as a simple match model. When part of the passage contained words from the question, it retrieved this part as an answer to the question, even though the meaning of this part did not answer the question (see Figure 3). Therefore, the system failed to predict the correct answer when the answer has semantically similar words to the question (see Figure 4).

Model	MAP	MRR
Ensemble	0.117	0.36
ArabicBERT	0.07	0.20
CL-AraBERT	**0.124**	0.375

Table 4: Test set results of Task A.

Model	pAP@10
Ensemble **POST** (Best)	0.498
AraBERT Base	0.5

Table 5: Test set results of Task B.

6 Conclusion

This paper presented our contributions to Task A: PR and Task B: MRC of the Qur'an QA 2023 shared task. Our proposed PR method was to train several Arabic pre-trained models with QRCD dataset using SBERT architecture and then ensemble the combination of these models. The ensemble method did not yield the best performance with the test set, although it had the best score with the development set. CL-AraBERT achieved the best results with a 0.124 MAP. Our proposed MRC system is based on combining the transfer learning and ensemble approaches for the best-performing models. Initially, we fine-tuned nine different Arabic pre-trained models using different data collections. We then applied the ensemble approach to the two best-performing models. Finally, we implemented appropriate post-processing steps. The combination of the base and large variants of AraBERT achieved the best results on the development set, with a 0.537 pAP@10. The second-highest score was achieved by base AraBERT with a 0.49 pAP@10. The results of applying these two models to the test set showed that the base AraBERT model was the best with a score of 0.5 pAP@10, while the ensemble model achieved a score of 0.49 pAP@10.

Limitations

One of the most important factors affecting the performance of pretraining models is the size of the dataset. The size of the dataset used in the training in this study is miniscule compared to the size of the data available in the English language. Therefore, we noticed weak performance of the models in Arabic. There is an urgent need to build large data collections in Arabic.

Acknowledgement

The authors would like to express their deepest gratitude to King Abdulaziz University for the support. We thank the reviewers for their constructive comments.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Esha Aftab and Muhammad Kamran Malik. 2022. erock at qur'an qa 2022: Contemporary deep neural networks for qur'an based reading comprehension question answers. 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 96–103.

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

Kholoud Alsubhi, Amani Jamal, and Areej Alhothali. 2021. Pre-trained transformer-based approach for arabic question answering: A comparative study. *arXiv preprint arXiv:2111.05671*.

Kholoud Alsubhi, Amani Jamal, and Areej M. Alhothali. 2022. Deep learning-based approach for arabic open domain question answering. *PeerJ Computer Science*, 8.

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*, 8:454–470.

Mohammed ElKomy and Amany M Sarhan. 2022. Tce at qur'an qa 2022: Arabic language question answering over holy qur'an using a post-processed ensemble of bert-based models. *arXiv preprint arXiv:2206.01550*.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). 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, Online. Association for Computational Linguistics.

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 *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.

Ali Mostafa and Omar Mohamed. 2022. Gof at qur'an qa 2022: Towards an efficient question answering

for the holy qu'ran in the arabic language using deep learning-based approach. 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 104–111.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. 2019. Neural arabic question answering. *arXiv preprint arXiv:1906.05394*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Claude Sammut and Geoffrey I. Webb, editors. 2010. *TF–IDF*, pages 986–987. Springer US, Boston, MA.

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.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

A QRCD Dataset Distribution

In this appendix, Table 6 presents the distribution of the dataset.

B The problems of the list of answers

In this appendix, Figure 1 and Figure 2 present the problems we encountered in the list of answers.

Dataset	**# Q**	**# Q-P Pairs**	**# Q-P-A Triplets**
Training	174	992	1179
Development	25	163	220

Table 6: QRCD distribution. # Q shows the number of the questions, # Q-P Pairs show the number of the questions–passage pairs and # Q-P-A Triplets show number of questions–passage–answers triplets.

C The Analysis and Discussion of Task B

In this appendix, Figure 3 and Figure 4 present the discussion of Task B Machine Reading Comprehension.

```
"pq_id": "13:18-24_360",

"passage": " للذين استجابوا لربهم الحسنى والذين لم يستجيبوا له لو أن لهم ما في الأرض
جميعا ومثله معه لافتدوا به أولئك لهم سوء الحساب ومأواهم جهنم وبئس المهاد. أفمن يعلم
أنما أنزل إليك من ربك الحق كمن هو أعمى إنما يتذكر أولو الألباب . الذين يوفون بعهد الله
ولا ينقضون الميثاق . والذين يصلون ما أمر الله به أن يوصل ويخشون ربهم ويخافون سوء الحساب
. والذين صبروا ابتغاء وجه ربهم وأقاموا الصلاة وأنفقوا مما رزقناهم سرا وعلانية ويدرؤون
بالحسنة السيئة أولئك لهم عقبى الدار . جنات عدن يدخلونها ومن صلح من آبائهم وأزواجهم
وذرياتهم والملائكة يدخلون عليهم من كل باب . سلام عليكم بما صبرتم فنعم عقبى الدار ."

"question": "هل سيجمع الله بين المؤمنين وأبنائهم وأهلهم في الجنة ؟",

{"answer":"في",

"rank":1, "score":0.1957549469953647, "strt_token_indx":12, "end_token_indx":12}
```


Figure 1: Example of an uninformative answer.


```
"pq_id":"2:177-177_419":

"question": "هل احترم الإسلام الأنبياء ؟"

[{"answer":"من آمن بالله واليوم الآخر والملائكة والكتاب والنبيين",

"rank":1,"score":0.9420806664550877,"strt_token_indx":10,"end_token_indx":17},

{"answer":"البر من آمن بالله واليوم الآخر والملائكة والكتاب والنبيين",

"rank":2,"score":0.042539567979458445,"strt_token_indx":9,"end_token_indx":17},

{"answer":"آمن بالله واليوم الآخر والملائكة والكتاب والنبيين",

"rank":3,"score":0.01242817290648292,"strt_token_indx":11,"end_token_indx":17}
```


Figure 2: Example of repeated answers.


```
"pq_id":"28:85-88_322":

"question":"هل تدبر القرآن فرض ؟"

"Gold answer":"[]"

"Model answer":"إن الذي فرض عليك القرآن لرادك إلى معاد"
```


Figure 3: Example 1 of an incorrect answer.


```
"pq_id":"11:50-60_337":

"question": "ما هي الإشارات للدماغ أو لأجزاء من الدماغ في القرآن ؟"

"Gold answer":"ناصيتها ",

"Model answer 1":"يرسل السماء عليكم مدرارا ويزدكم قوة إلى قوتكم"

"Model answer 2":" ولا تتولوا مجرمين . قالوا يا هود ما جئتنا ببينة وما نحن بتاركي
آلهتنا عن قولك وما نحن لك بمؤمنين"

"Model answer 3":"إن نقول إلا اعتراك بعض آلهتنا بسوء قال إني أشهد الله واشهدوا أني"

"Model answer 4":"استغفروا ربكم ثم توبوا إليه"
```


Figure 4: Example 2 of an incorrect answer.