Proceedings of the OSACT 2022 Workshop @LREC2022, pages 146–153
Marseille, 20 June 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

Ahmed Wasfey, Eman Elrefai , Marwa Muhammad, Haq Nawaz
Tactful AI , BambooGeeks , Freelancer , PUCIT
{ahmedsleemce, eman.lotfy.elrefai, marwa.mohammad.matar, haqnawaz99}@gmail.com

Abstract

The Holy Qur'an is the most sacred book for more than 1.9 billion Muslims worldwide, and it provides a guide for their behaviours and daily interactions. Its miraculous eloquence and the divine essence of its verses (Khorami, 2014)(Elhindi, 2017) make it far more difficult for non-scholars to answer their questions from the Qur'an. Here comes the significant role of technology in assisting all Muslims in answering their Qur'anic questions with state-of-the-art advancements in natural language processing (NLP) and information retrieval (IR). The task of constructing the finest automatic extractive Question Answering system from the Holy Qur'an with the use of the recently available Qur'anic Reading Comprehension Dataset(QRCD) was announced for LREC 2022 (Malhas et al., 2022) which opened up this new area for researchers around the world. In this paper, we propose a novel Qur'an Question Answering dataset with over 700 samples to aid future Qur'an research projects and three different approaches where we utilised self-attention based deep learning models (transformers) for building reliable intelligent question-answering systems for the Holy Qur'an that achieved a partial Reciprocal Rank (pRR) best score of 52% on the released QRCD test set.

Keywords: Qur'an, Extractive Question Ansewring , Deep Learning, NLP , Transformers

1. Introduction

Reading Comprehension, which is the skill of reading a text and then answering questions about it, is a difficult task for machines that continues to pique the interest of many academics and will continue to do so for many years to come (Rajpurkar et al., 2016). Unlike open domain question answering systems (Mishra and Jain, 2016) in extractive question answering, a passage or a context is provided so that the model can refer to it and predict where the answer is inside the passage as shown in Figure 4, and it is still a very challenging as it requires machines to have both natural language understanding and knowledge of the world or the domain in the case of domain-specific applications.

There are six official languages of the United Nations and the only Semitic language among these six languages is Arabic (Tahani et al., 2021). Mainly Arabic can be divided into three categories. Classical Arabic (CA) is the language of the Qur'an, Hadith and classical Islamic literature. Modern Standard Arabic (MSA) is the language used in the news, articles or modern Arabic era. Colloquial or dialectal Arabic is the variety of different dialects in different regions of Arabic speaking countries. Additionally, The grammatical structure of the Arabic language is exceedingly rich and complex(Khaled et al., 2018) as compared to the other languages. Arabic QA is addressed in a few studies since 2004(Bakari et al., 2016), and the existence of considerable, realistic datasets have always been crucial for propelling fields ahead, famous examples include SQUAD for English reading comprehension (Rajpurkar et al., 2016) and the Penn Treebank for syntactic parsing (Marcus et al., 1994). To assist in meeting the need to improve the Qur'anic reading comprehension dataset in order to boost the use of state-of-the-art data intensive models, we propose a new dataset with more than 700 Qur'an questions extracted from the publicly available Annotated Corpus of Arabic Al-Qur'an Question and Answer(AQQAC) dataset (Alqahtani and Atwell, 2018) and reformatted sample by sample to be usable for the extractive Ques-

Passage : "إن الله اصطفى آدم ونوحا وآل إبراهيم وآل عمران على العالمين. ذرية بعضها من بعض والله سميع عليم. إذ قالت امرأت عمران رب إني نذرت لك ما في بطني محررا فتقبل مني إنك أنت السميع العليم. فلما وضعتها قالت رب إني وضعتها أنثى والله أعلم بما وضعت وليس الذكر كالأنثى وإني سميتها مريم وإني أعيذها بك وذريتها من الشيطان الرجيم. فتقبلها ربها بقبول حسن وأنبتها نباتا حسنا وكفلها **زكريا** كلما دخل عليها زكريا المحراب وجد عندها رزقا قال يا مريم أنى لك هذا قالت هو من عند الله إن الله يرزق من يشاء بغير حساب. هنالك دعا زكريا ربه قال رب هب لي من لدنك ذرية طيبة إنك سميع الدعاء. فنادته الملائكة وهو قائم يصلي في المحراب أن الله يبشرك ب**يحيى** مصدقا بكلمة من الله وسيدا وحصورا ونبيا من الصالحين. قال رب أنى يكون لي غلام وقد بلغني الكبر وامرأتي عاقر قال كذلك الله يفعل ما يشاء. قال رب اجعل لي آية قال آيتك ألا تكلم الناس **ثلاثة أيام** إلا رمزا واذكر ربك كثيرا وسبح بالعشي والإبكار."

Question : "كم ليلة امر الله زكريا الا يكلم الناس؟"
Answers : "**ثلاثة أيام**"

Question : "من هو ابن زكريا؟"
Answers : "**يحيى**"

Question : "من كفل السيدة مريم؟"
Answers : "**زكريا**"

Figure 1: : Question-answer pairs for a sample passage in the QRCD dataset. Each of the answers is a segment extracted from the passage.

tion Answering task. We also conducted many experiments with transformers like araBERT(Antoun et al., 2020a) some experiments by using the vanilla model for direct Question Answering fine tuning with different configurations and combinations of data and in other experiments we created araBERT based Qur'an masked language model by fine tuning the model on bare Qur'an verses first then using it for building the Question Answering model further more we also tried ensemble of different transformers to strengthen each other for better results and acheived 52% best pRR score on the test set and 62% score on the development dataset.

The rest of the paper is organized as follows: Section 2 summaries some of the work done previously in this field. Section 3 discusses the data and its challenges. Section 4 presents the basic ideas that were used as building blocks for the final three systems used in the final submissions which are defined in section 5. We evaluate the three systems performance and conclude the paper in section 6 and section 7respectively.

2. Related Work

In this section, we shed light on the previous work for both the Qur'anic datasets has questions and answers from the Holy Qur'an and Arabic Question Answering (QA) Systems.

2.1. Qur'anic datasets:

Several studies have been made to understand the Qur'anic text and extract knowledge from it. AyaTEC (Malhas and Elsayed, 2020) is the dataset for Arabic QA on the Holy Qur'an. All of the Qur'anic verses that directly answer the questions were exhaustively extracted and annotated. The answers are divided into a single answer, multiple answers and no answers with evaluation for each type. it proposed several evaluation measures to integrate the concept of partial matching. Also, there is a dataset have a partial matching for the required structure of the Qur'an QA task called AQQAC (Alqahtani and Atwell, 2018). It was annotated corpus of Arabic Al-Qur'an QA using machine learning.

2.2. Arabic QA Systems:

(Alsubhi et al., 2021) evaluated the performance of three existing Arabic pre-trained models(AraBERTv2-base, AraELECTRA, AraBERTv0.2-large) on Arabic QA. Al-Bayan(Abdelnasser et al., 2014) proposed a novel Question Answering system for the Qur'an, that extracted the answer from the retrieved verses accompanied by their Tafseer. (Mozannar et al., 2019) proposed an approach for open domain Arabic QA and introduced the Arabic Reading Comprehension Dataset (ARCD) and ArabicSQuAD and consisted of a document retriever using hierarchical TF-IDF and a document reader using BERT.

3. Data

3.1. Existing Datasets

We begin by investigating existing Qur'anic Reading Comprehension with Question Answering (QA) datasets. We highlight their structure and the challenges for each dataset. QRCD (Qur'anic Reading Comprehension Dataset) (Malhas and Elsayed, 2020) (Malhas et al., 2022) is the dataset which was provided by the organizers of the Qur'an QA competition with 1,337 question-passage-answer. AQQAC (Alqahtani and Atwell, 2018) is annotated corpus of Arabic Al-Qur'an Question and Answer with 1,225 question-answer. The last existing dataset we used is Arabic SQuAD(Mozannar et al., 2019). It's a machine translation of the Stanford Question Answering.

3.1.1. QRCD

QRCD (Qur'anic Reading Comprehension Dataset) (Malhas and Elsayed, 2020) (Malhas et al., 2022)is the main dataset we used. It helped us for understanding the task and creating a similar dataset following the same structure. The passage is extracted from some specific verses from the Qur'an. The passage is about one page from Qur'an or less than that. QRCD may have multiple same questions for different passages from the Qur'an. For the same question, there are a single answer or multiple answers with ranking. The first answer is the gold answer and the other answers after that are partially exact answers. The answer must be included in the passage. It is composed of 1,093 tuples of question-passage pairs that are coupled with their extracted answers to constitute 1,337 question-passage-answer triplets.It is divided to training, validation and testing dataset. For the training dataset, it is 710 samples, the development dataset is 109 samples and the testing is 352 samples. QRCD consists of a question-passage pair and the answers retrieved from the accompanying text. The dataset is formatted as shown in Figure 1.

3.1.2. AQQAC

After searching for any open-source dataset that can work fine for Qur'an QA, The only related dataset we found is The Al-Qur'an Question and Answer Compilation (AQQAC) (Alqahtani and Atwell, 2018) is a collection of 1224 questions and answers regarding the Al-Qur'an. Combining more datasets, they can increase the dataset and get higher scores. AQQAC dataset consists of The question ID, question word (particles), chapter number, verse number, question topic, question type, Al-Qur'an ontology concepts (Alqahtani Atwell, 2018), and question source are all marked on each question and response. The goal of this corpus is to give a Question-Answering taxonomy for Al-Qur'an-related inquiries. This corpus might also be utilised as a data set for testing and evaluating Islamic IR systems. We aimed to use this dataset with the QRCD (Malhas and Elsayed, 2020) (Malhas et al., 2022) dataset

but we couldn't use it directly. The problem with it is that most of the passages and the answers aren't extracted from the Qur'an. They are extracted from the Altabari Tafseer. Most of the questions are valid to get and follow the same structure as the QRCD dataset. We used about 500 questions from the AQQAC dataset and added them to our data. We added some small passages and answers that are extracted from Qur'an.

Question: "ما مصير المنافقين يوم القيامة ؟ اذكر الآية الكريمة."
Answer: "الدرك الأسفل من النار .إِنَّ الْمُنَافِقِينَ فِي الدَّرْكِ الْأَسْفَلِ مِنَ النَّارِ وَلَن تَجِدَ لَهُمْ نَصِيراً{145}النساء."

Question: "ماذا تعلم آدم عليه السلام من الله جل جلاله وكان هذا العلم ليس عند الملائكة ؟"
Answer: "علم الأسماء . والدليل : وَعَلَّمَ آدَمَ الْأَسْمَاء كُلَّهَا ثُمَّ عَرَضَهُمْ عَلَى الْمَلَائِكَةِ فَقَالَ أَنبِئُونِي بِأَسْمَاء هَؤُلاء إِن كُنتُمْ صَادِقِينَ{31} البقرة"

Question: "ما هي السبع المثاني ؟"
Answer: "سورة الفاتحة .بِسْمِ اللهِ الرَّحْمنِ الرَّحِيمِ{1} الْحَمْدُ للهِ رَبِّ الْعَالَمِينَ{2} الرَّحْمنِ الرَّحِيمِ{3} مَالِكِ يَوْمِ الدِّينِ{4} إِيَّاكَ نَعْبُدُ وإِيَّاكَ نَسْتَعِينُ{5} اهدِنَــا الصِّرَاطَ الْمُستَقِيمَ{6} صِرَاطَ الَّذِينَ أَنعَمتَ عَلَيهِمْ غَيرِ الْمَغضُوبِ عَلَيهِمْ وَلاَ الضَّالِّينَ{7}الفاتحة"

Question: "اشرح الآية الكريمة.يَا أَيُّهَا الَّذِينَ آمَنُواْ عَلَيْكُمْ أَنفُسَكُمْ لاَ يَضُرُّكُم مَّن ضَلَّ إِذَا اهْتَدَيْتُمْ إِلَى اللهِ مَرْجِعُكُمْ جَمِيعاً فَيُنَبِّئُكُم بِمَا كُنتُمْ تَعْمَلُونَ{105}المائدة"
Answer: "يا أيها الذين آمنوا عليكم أنفسكم أي احفظوها وقوموا بصلاحها , لا يضركم من ضل إذا اهتديتم قيل المراد لا يضركم من ضل من أهل الكتاب وقيل المراد غيرهم لحديث أبي ثعلبة الخشني: سألت عنها رسول الله صلى الله عليه وسلم فقال: ائتمروا بالمعروف وتناهوا عن المنكر حتى إذا رأيت شحا مطاعا وهوى متبعا ودنيا مؤثرة وإعجاب كل ذي رأي برأيه فعليك نفسك إلى الله مرجعكم جميعا فينبئكم بما كنتم تعملون فيجازيكم به"رواه الحاكم وغيره

Figure 2: : Question-answer pairs with no passage in the AQQAC dataset.

3.1.3. Arabic SQuAD

Arabic SQuAD (Mozannar et al., 2019) is a dataset with 48,344 questions on 10,364 paragraphs. The paragraphs are extracted from Wikipedia articles so this dataset isn't related to Qur'an QA. We aimed to make the model deal with the Arabic question answer in general. After that, we expected the model can perform better when we apply the Qur'an QA.

When we trained the transformer model using QRCD

passage: 'S XOFF ، المزود بقارئ الشريط الورقي الأوتوماتيكي كونترول ASR عندما تلقى جهاز تيليتيبي 33 الإرسال على تسبب في استئناف قارئ Q XON ، اختصار للإرسال ، تسبب في توقف قارئ الشريط تلقي كونترول الشريط . أصبحت هذه التقنية معتمدة من قبل العديد من أنظمة تشغيل الكمبيوتر في وقت مبكر باعتبارها إشارة المصافحة تحذير المرسل لوقف انتقال بسبب تجاوز الوشيكة يستمر حتى يومنا هذا في العديد من الأنظمة كتقنية التحكم ثانية S بـ كونترول Q بمعناها ولكن يتم استبدال كونترول S اليدوي في الإخراج . في بعض الأنظمة تحتفظ كونترول لبدء وإيقاف لكمة الشريط T DC4 و كونترول R DC2 لتعيين كونترول ASR 33 لاستئناف الإخراج . يمكن أيضا تكوين في بعض الوحدات المجهزة بهذه الوظيفة ، كان الحرف المقابل لحروف التحكم على كيكاب الموجود أعلى الحرف هو على التوالي TAPE و TAPE. '

question: 'ما انتهى قارئ الشريط الورقي التلقائي للتوقف؟'
answers: 'تحذير'

Figure 3: : Question-answer pairs for a sample passage in the Arabic SQuAD dataset. Each of the answers is a segment extracted from the passage.

(Malhas and Elsayed, 2020) (Malhas et al., 2022) only, the pRR scores were 44% or less than that. We did many experiments without adding any external data. The model performed well with adding more data and increased in pRR score by 10% when adding our data so our team decided to use Arabic SQuAD for two approaches.

One is to train the model using Arabic SQuAD only then using this pre-trained model for our task. This approach didn't perform better as expected. Another approach is to combine the QRCD dataset with the Arabic SQuAD dataset and our collected data. The reason to apply this is that modern Arabic isn't different so much from the Qur'an. We can increase more data with modern Arabic that enabling the model to train to extract more answers from different questions. This approach performed better sometimes than normal. We did random portions with different sizes and used them for the last advanced approaches.

3.2. Challenges

Previously available work combines many ways to produce their own data. However, there are certain obstacles to each strategy. We highlight the hurdles that must be solved in order to create a large-scale, high-quality dataset.

3.2.1. Limited question-answer for QRCD

QRCD (Malhas and Elsayed, 2020) (Malhas et al., 2022) is a small data which has about 700 questions. Most of them are repeated questions for different passages from Qur'an verses so the model is limited by a few questions to train and extract the answer. That changes our direction towards creating a similar data with following the same structure to add it to QRCD and all training data become doubled in size which QRCD is 710 samples and our data is about 730. We focus on adding more different types of questions.

3.2.2. Unavailability of open-source datasets

Qur'an QA is a hard task and collecting data can consume time with inaccurate answers. We are interested in doing a deep search to get any open-source data that can help us with this task. However, we found only annotated data with its passage and answer from tafsir and few of them from Qur'an AQQAC (Alqahtani and Atwell, 2018) as mentioned previously.

3.2.3. Automated vs Manual dataset

With the previous 2 challenges. We decided to collect our own data to combine it with QRCD (Malhas and Elsayed, 2020) (Malhas et al., 2022) and with the help of AQQAC (Alqahtani and Atwell, 2018). The first plan is to generate an automated dataset that enables us to create a large-scale dataset. Creating it manually means consuming time with generating a small-scale dataset. The challenges with AQQAC are that most of the answers aren't extracted from their specific passages, the passage and answers are mixed with Qur'an verses and tafsir, There are answers with the same meaning but with different words that aren't found in the Qur'an, and there are types of questions like" ما معني كلمة ", " ما الدليل من القرآن " aren't related to Qur'an QA tasks. We couldn't solve all these issues so we decided to do it manually by using the questions of AQQAC.

3.2.4. Qur'an Experts to add accurate answers

The Holy Qur'an is a classical Arabic with Complex Word Structure. Before answering any questions from Qur'anic passage manually that requires knowing the tafsir and the meaning of the passage with related context events. The questions are in modern Arabic and understanding it easier. For that, we care about constructing it by Qur'an scholars despite it consuming time and cost but it was our only choice. We have in our team a Qur'an scholar who added the passage and answers for every question and created some similar questions. The constructed data by a Qur'an scholar in our team is about 730 questions with answers. It's called AQAQ (Arabic Question Answers from the Holy Qur'an dataset).

3.3. Dataset Collection

These previous challenges made us do it manually adding question by question. The source of questions in this dataset is Corpus of Arabic Al-Qur'an Question and Answer(AQQAC)(Alqahtani and Atwell, 2018) as mentioned previously. This data helped us to do authorized questions and answers. This data answers the question from the Qur'an and Tafseer. It consists of 1225 questions. We filtered about 500 questions and removed every answer from Tafsier because our task is to get the answers from the Qur'an only.

passage: 'يا أيها الناس إنا خلقناكم من ذكر وأنثى وجعلناكم شعوبا وقبائل لتعارفوا إن أكرمكم عند الله أتقاكم إن الله عليم خبير'

question: 'ما مقياس التفاضل عند الله تعالى بين الناس'

answers: 'إن أكرمكم عند الله أتقاكم'

Figure 4: : Question-answer pairs for a sample passage in the AQAQ dataset. Each of the answers is a segment extracted from the passage.

We structured it like the original dataset and added multiple same questions with other different passages. There are 2 versions. One was used for the first and second submissions with 625 questions. The second version has 732 questions and is used for the third submission. By this data, the scores are increased by 5% for the development evaluation for the first version. The second version is increased by 10% for the development evaluation.

4. Methodology

In this section, we illustrate the main components and ideas that we used for constructing our solution approaches explained in the upcoming section 5. We focus on the main four elements. First, **Masked Language Models** that were the basic building blocks for our approaches, we have not only made use of existing Arabic pre-trained MLM like BERT (Devlin et al., 2018) ,ELECTRA (Clark et al., 2020) and XLM-ROBERTA (Conneau et al., 2019), but also we created our Qur'anic MLM by fine-tuning araBERT (Antoun et al., 2020a) on the Holy Qur'an's verses only, the second step is **Preprocessing** where we put together the two sequence for tokenization and encode the tokens to be used for **fine tuning** the MLM for our task which is the third step, finally we tried using **ensemble** of many fine-tuned QA models and take a vote of the best answer to achieve better results, more details for every step in the following subsections.

4.1. Masked Language Models

Attention mechanisms have been proved to be extremely efficient for understanding context for natural languages (Hu, 2019). Consequently, Transformers (Vaswani et al., 2017) have shown qualitative superiority over previously used sequence models in most NLP tasks, so we tried to make the best use of them, especially with their availability and ease of fine-tuning.

4.1.1. Pre-trained Models

HuggingFace provides a wide variety of pretrained ready-for-use transformers for more than 175 languages For Arabic language araBERT (Antoun et al., 2020a) with 136 million parameters -for base version- is one of the best options, as it was trained on a combination of large Arabic corpora along with araELECTRA (Antoun et al., 2020b) and other publicly available Arabic question answering models on their hub, was the foundation for many experiments we did by directly fine-tuning them for extractive question answering with different configurations and combinations from datasets mentioned above in section 3.

4.1.2. Qur'anBERT

Because this task is very specific for only a set of verses we needed an MLM that really understands the Qur'an more than any corpus or Arabic text so We used a Qur'an dataset from Kaggle -The Holy Qur'an competition- that contained the Holy Qur'an verses with diacritics [1], we removed all diacritics and signs of stopping (like صلي ، مـ) and transformed it into a text file that contains all verses to be ready for training the model. We used this data to fine-tune araBERT as Qur'anic MLM with the help HugingFace Dataset utility we used DataCollatorForLanguageModeling with masked language probability of 0.1 and 0.15 for different trails to create our Qur'anBERT that was better at filling masks in Qur'an verses, for example, the original model was unable to predict the word "الدين" in surah 1 verse 4 " مالك يوم الدين " if we masked it with [MASK] spe-

[1] Kaggle the Holy Qur'an competition `https://www.kaggle.com/zusmani/the-holy-Qur'an`

cial token "مالك يوم [MASK]" while the Qur'anBERT predicted it easily, this indicated more understanding-better attention weights- for the Qur'anic verses and words. so we used this model also in our experiments along with previously mentioned models.

4.2. Preprocessing

In preprocessing, the questions and passages are tokenized and converted into a suitable format that can be used to fine-tune transformers. The input to the tokenizer is the question and the passage of that question. The tokenizer output a dictionary containing the followings(pritesh1, 2022).

- input-ids: Stores the tokens ids of question and passage including the [CLS] and [SEP] tokens which indicate the beginning of the question, and SEP between question and the passage.
- token-type-ids: Stores 0s and 1s to differentiate between the passage and question borders. It has 0 for question tokens and 1s for passages tokens.
- attention-Mask: for every token indicates which tokens must not have attention like padding, and proceeding tokens.

Then, the start and the end position of the answer - index of the first and last tokens - have been added to the tokenizer output dictionary, after calculating it by comparing given start character with tokens' offset mapping returned from the tokenizer

4.3. Training

Fine-tuning large models like transformers on a small dataset like we have was challenging so we avoided using large batch sizes and kept our trails in the range [2, 4] and tried increasing the learning rate for faster convergence {2e-5, 1e-4, 2e-4}. in All experiments, we used the trainer API [2].

4.4. Model Ensemble

Ensemble models are a machine learning technique for combining multiple models in the prediction process. These models are known as base estimators or weak learners, and it is a solution to the many challenges of developing a single estimator like low accuracy and high sensitivity.

5. Solution Approaches for Our Submitted Results

As you know in the methodology section, our team developed different techniques to improve the system results. Let's discover their effect on this task.

5.1. Qur'anBERT

As mentioned earlier in section 4.1.2 we fine-tuned our masked language model for Qur'an that we named Qur'anBERT the first submitted results -stars_run01- were generated from the ensemble of many trained models most of them were Qur'anBERT based models, different combinations of data were used for training the weak learners, some were trained on QRCD (Malhas and Elsayed, 2020) (Malhas et al., 2022) data only other were trained on QRCD augmented by 625 sample unfinished AQAQ dataset or some other models with random portions of Arabic Squad dataset (Mozannar et al., 2019), also with variety of training configurations (epochs, batch sizes, learning rate, or weight decay) as explained with code in our repo [3]

5.2. Ensembling with K-Folds-Like Data Splitting

For the second submission -stars_run05- We have divided the competition dataset merged with our extracted AQAQ dataset into distinct 8-Folds instead of using normal bootstrapping as we might get more dissimilar weak learners for better final ensemble model. Then, the training of 8 transformers has been accomplished with distinct configurations per split of data. In the following table, all model names and configurations are shown. (Note, the list values are registered in the same order of running folds) We have 8 folds with different configurations as mentioned above. Each fold model has its own behavior on the development dataset (Figure 2). As noticed from the curves below that the model 2 has the best behavior across all metrics pRR, exact match, f1.

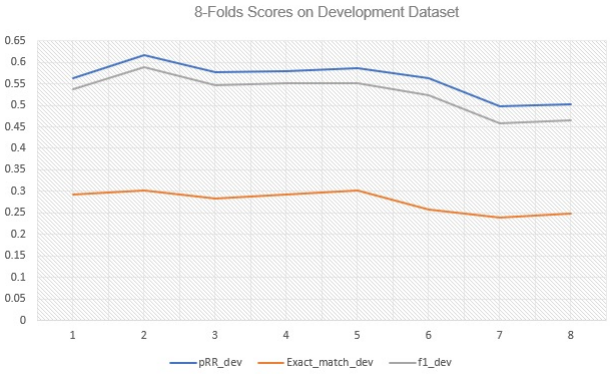

Figure 5: k-fold development results for the configurations mentioned in table 1

5.3. Expanding AQAQ dataset

For this approach, We focused on measuring the effect of the size of the data on the performance. We

[2] https://huggingface.co/docs/transformers/main_classes/trainer

[3] https://github.com/EmanElrefai/Qur'an_QA/tree/main/Qur'anBERT

Tokenization	token-model-name = ”bert-base-arabertv02” max-length= 384 truncation=”only-second” return_offsets_mapping=True padding=”max_length”
Training	QA_model_name = [’bert-base-arabertv02’, bert-base-arabertv02’, ’AraElectra-base-finetuned-ARCD’, ’AraElectra-base-finetuned-ARCD’ ’AraElectra-base-finetuned-ARCD’, ’AraElectra-base-finetuned-ARCD’, ’arap_qa_bert_v2’, ’arap_qa_bert_v2’] learning_rate=2e-5, per_device_train_batch_size= [2,3,2,2,2,2,2,2] per_device_eval_batch_size= 2 num _train_epochs= [5,5,5,4,3,2,7,10]
Inference	Extracting the highest 5 start and end scores from the all predicted scores returned from the model

Table 1: k-folds experiment details

worked on creating more examples during last moments of the competition’s deadline. We reached about 732 samples instead of 625 samples. This combined with QRCD (Malhas and Elsayed, 2020) (Malhas et al., 2022) dataset is about 1,593 samples. We did shuffle for all data and used araBertv0.2 transformer base version. We clearly noticed during experiments that increasing AQAQ to be 732 samples worked better than AQAQ with 625 samples for the same model. This is for the third submission stars_run06. In Table 2 below are the configurations that are used for the third submission.

6. Results

In this section, we show the results of the three different approaches on both development dataset and test dataset, metrics used for evaluation are partial Reciprocal Rank (pRR) which is a variant of the traditional Reciprocal Rank evaluation metric that considers partial matching (Malhas and Elsayed, 2020) for 5 predicted answers , Exact Match (EM) and F1@1 the top predicted answer only. All three systems perform better than fine tuning araBERT base model with QRCD data which gives 44% pRR 24% exact match and 42% F1@1 on the development dataset. Throughout most of the experiments and as shown in table 3 and table 4, we noticed that ensemble and Qur’anBERT were not

Tokenization	token-model-name = ”bert-base-arabertv02” max-length= 512 truncation=”only-second” return_offsets_mapping=True padding=”max_length”
Training	QA_model_name = ’aubmindlab/bert-base-arabertv02’ learning_rate=2e-5, per_device_train_batch_size= 2 per_device_eval_batch_size= 2 num _train_epochs= 4

Table 2: Expanding AQAQ dataset experiment details

as effective as increasing the size of AQAQ data with fresh samples, which might be attributed to the following reasons.

We have only 1500 Qur’an training examples in the best case scenario -QRCD training set with AQAQ data - which makes the splits coming out of bootstrapping to be very similar, consequently, the weak learners of the ensemble will be almost identical, especially given that, transformers are known to be extremely data intensive.

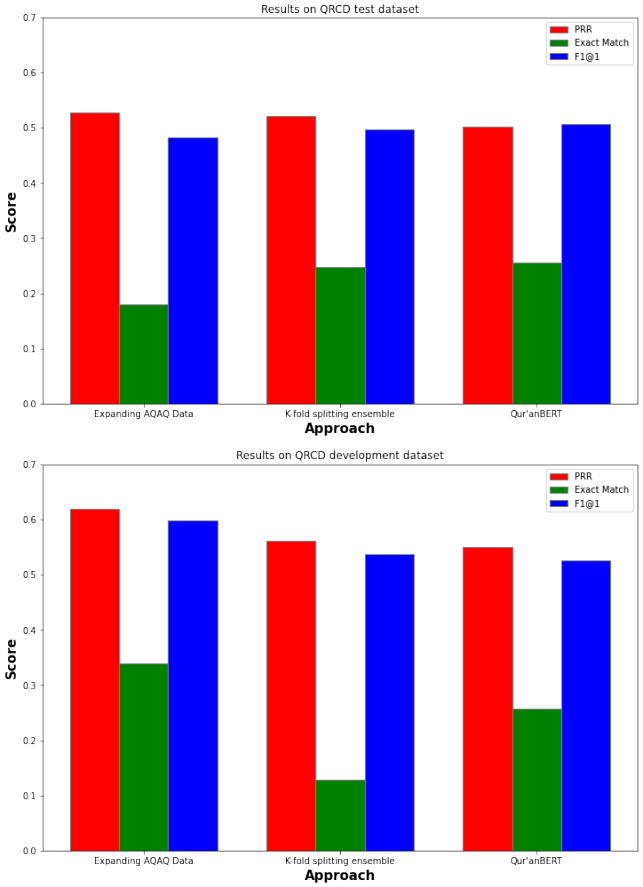

Figure 6: Results of the three approaches on QRCD development and test dataset.

On the other hand, Masked Language Models are known to be trained on a large corpus of the text so the 6236 verses of the Qur'an are tiny in size compared to that enormous amount of data that araBERT is trained on (Antoun et al., 2020a) for example. Additionally, in the Question Answering task model needs to learn the attention between the two different sequences of the passage -which is part of Qur'an- and the question -which is a *non-Qur'anic* sentence- unlike the single Qur'anic sequence in the Masked Language task.

Approach	pRR	EM	F1@1
Expanding AQAQ	0.6195	0.3394	0.5983
K-folds ensemble	0.562	0.128	0.537
Qur'anBERT	0.5495	0.2568	0.526

Table 3: Results of different approaches on development dataset

Approach	pRR	EM	F1@1
Expanding AQAQ	0.528	0.256	0.507
K-folds ensemble	0.521	0.247	0.4966
Qur'anBERT	0.502	0.18	0.483

Table 4: Results of different approaches on test dataset

7. Conclusion and Future Work

We proposed three different approaches with remarkable results. The best and the highest one was expanding AQAQ with 0.528 (pRR) in testing. Our data team did their best to add more to be 732 samples eventually. The second approach was ensembling the most three best models and it's called K-folds ensemble approach. Its pRR score was 0.521. The last one was training Bert model on Qur'an then we used this pretrained model for this specific task Qur'an QA. The pRR score of Qur'anBert was 0.502. The scores for the three approaches are similar. For all three approaches, we developed the system with transformer models. The task was difficult and need more time and effort to collect more data that made our scores couldn't excceed a half in pRR.

Also, we proposed a new Arabic Question Answers dataset from the Holy Qur'an called AQAQ about 625 question-answers. We used it for K-folds ensemble and Qur'anBert approach. Then we increased it to be 732 samples and it was used for expanding AQAQ dataset approach only. We achieved highest scores by expanding it.

In future work, The data team aims to increase the AQAQ dataset to be the biggest one related to this task. We plan to cover all kinds of questions with answers, add complex questions and increase the question with multiple answers. In order to improve the scores of the system, we plan to expand our experiments for more approaches in different directions as many different adaptations, tests, and experiments and achieve higher scores in pRR. We plan to make the system and AQAQ dataset publicly available to the research community.

8. Reproducibility

All code, data, and experiments for this paper are available at GitHub [4]

9. Bibliographical References

Abdelnasser, H., Ragab, M., Mohamed, R., Mohamed, A., Farouk, B., El-Makky, N. M., and Torki, M. (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.

Alqahtani, M. and Atwell, E. (2018). Annotated corpus of arabic al-quran question and answer.

Alsubhi, K., Jamal, A., and Alhothali, A. (2021). Pretrained transformer-based approach for arabic question answering: A comparative study. *arXiv preprint arXiv:2111.05671*.

Antoun, W., Baly, F., and Hajj, H. (2020a). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Antoun, W., Baly, F., and Hajj, H. (2020b). Araelectra: Pre-training text discriminators for arabic language understanding. *arXiv preprint arXiv:2012.15516*.

Bakari, W., Bellot, P., and Neji, M. (2016). Literature review of arabic question-answering: Modeling, generation, experimentation and performance analysis. In *Flexible Query Answering Systems 2015*, pages 321–334, cham. Springer International Publishing.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elhindi, Y. (2017). Metaphors in the quran: A thematic categorization. *QURANICA-International Journal of Quranic Research*, 9(1):1–20.

Hu, D. (2019). An introductory survey on attention mechanisms in nlp problems. In *Proceedings of SAI Intelligent Systems Conference*, pages 432–448. Springer.

[4] https://github.com/EmanElrefai/Qur'an_QA

Khaled, S., Siddiqui, S., Alkhatib, M., and Monem, A. A. (2018). Challenges in arabic natural language processing. *Computational Linguistics, Speech and Image Processing for Arabic Language*, pages 59—-83.

Khorami, M. (2014). Eloquence of repetition in quran and arabic old poetry. *Language Related Research*, 5(2):91–110.

Malhas, R. and Elsayed, T. (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.

Malhas, R., Mansour, W., and Elsayed, T. (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)*.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Mishra, A. and Jain, S. K. (2016). A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3):345–361.

Mozannar, H., Hajal, K. E., Maamary, E., and Hajj, H. (2019). Neural arabic question answering. *arXiv preprint arXiv:1906.05394*.

pritesh1. (2022). An explanatory guide to bert tokenizer.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Tahani, A., Azmi, A. M., Aboalsamh, H. A., Cambria, E., and Hussain, A. (2021). Arabic question answering system: A survey. *Artificial Intelligence Review*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.