# LARSA22 at Qur'an QA 2022: Text-to-Text Transformer for Finding Answers to Questions from Qur'an

*Youssef Mellah, ⁎Ibtissam Touahri, *Zakaria Kaddari, *Zakaria Haja, *Jamal Berrich, *Toumi Bouchentouf

*LARSA Laboratory of National School of Applied Sciences, 2Department of computer science, Superior School of Technology, Meknes
*Oujda, Morocco, ⁎Meknes, Morocco
y.mellah@ump.ac.ma, i.touahri@umi.ac.ma, kadzaki@gmail.com, zakaria.haja@ump.ac.ma, jberrich@gmail.com, t.bouchentouf@ump.ac.ma

## Abstract

Question Answering (QA) is one of the main focuses of Natural Language Processing (NLP) research. However, Arabic Question Answering is still not within reach. The challenges of the Arabic language and the lack of resources have made it difficult to provide powerful Arabic QA systems with high accuracy. While low accuracy may be accepted for general purpose systems, it is critical in some fields such as religious affairs. Therefore, there is a need for specialized accurate systems that target these critical fields. In this paper, we propose a Transformer-based QA system using the mT5 Language Model (LM). We finetuned the model on the Qur'anic Reading Comprehension Dataset (QRCD) which was provided in the context of the Qur'an QA 2022 shared task. The QRCD dataset consists of question-passage pairs as input, and the corresponding adequate answers provided by expert annotators as output. Evaluation results on the same DataSet show that our best model can achieve 0.98 (F1 Score) on the Dev Set and 0.40 on the Test Set. We discuss those results and challenges, then propose potential solutions for possible improvements. The source code is available on our repository[1].

**Keywords:** Question Answering (QA), Natural Language Processing (NLP), Transformer, mT5, Language Model (LM), QRCD

## 1. Introduction

The Holy Quran is the primary reference for about 1.6 billion Muslims around the world. The Qur'anic classical Arabic text is either instructive or narrative and it is composed of 114 chapters, 6,236 verses and 80k words (Malhas and Elsayed, 2020), the entire text is joined into one whole related concept which underlies a deep connection between the ideas and meaning that overlap whithin chapters and verses. Different researches have been conducted on the holy Quran, either to construct datasets, perform automatic text classification, Question answering (QA), semantic search based on ontology, topic assignment or optical text recognition tasks (Adeleke et al., 2019) (Alqahtani and Atwell, 2016) (Mohamed and Shokry, 2020) (Mohd et al., 2021). There is a need for automatic Qur'anic QA (Alsubhi et al., 2021) that helps to facilitate and improve the time of Qur'an knowledge acquirement since the holy Qur'an is the trustful legislated text used for teaching purposes and to respond to Muslim requests.The search engines are an automatic way to get information as they provide documents rich of relevant information to the user request. However, it doesn't pinpoint the answer. Hence, there is still a lack of methods that answer directly the user demands which makes answer search troubleshooting. In this context, reliable Question answering systems turn important to deal with this in different languages, while there are few attempts for Arabic QA investigation. QA is an interactive automatic process that belongs to the natural language processing field. QA systems provide the user with a solution that helps them to achieve the exact answer to a natural language question. The performance of a QA system is influenced by the size of training data, its quality as well as the used techniques. The Arabic language is still of scarce resources in comparison to English (Alsubhi et al., 2021) (Alanazi et al., 2021) (Maraoui et al., 2021) and yet the QA task needs

more preprocessing and deep linguistic knowledge when extracting answers from the holy Qur'an.

In the following, we describe our participation in Qur'an QA 2022 Shared Task that aims to answer questions in the holly Qur'an. The task organizers provide participating systems with a consecutive passage from the Holy Qur'an based on which they answer questions raised in Modern Standard Arabic by selecting one or many ranked text spans from the passage. For this, they provided the Qur'anic Reading Comprehension Dataset (QRCD ) composed of 1,337 question-passage-answer triplets.

Our paper proposed a model based on sequence to sequence (Seq2Seq) transformers using the mT5 Language Model that is fed by Arabic question and its corresponding passage (context), then extract the most relevant answers. The task starts by preprocessing and matching the question to the passage that contains the response, then we fine-tune the mT5 LM to get the adequate answer.

The paper is organized as follows, we present the related works, and afterward, we define the task steps and dataset as described by the organizers, we describe afterward the proposed approach besides analyzing the obtained results on both dev and test sets, then we give a conclusion with further work.

## 2. Related Work

Many studies have been interested in providing reliable linguistic resources for the Holy Qur'an automatic annotation, processing and interpretation. In this context, (Osman et al., 2015) provided a Quranic Dataset that may be exploited by researches aiming to explore the holy Quran. (Dukes et al., 2013) presented the Quranic Arabic Corpus that includes morphological segmentation, part-of-speech, and syntactic analysis annotations. (Al-Salhi and Abdullah, 2022) constructed the ontology of Quranic stories which construction depends on the MappingMaster

---

domain-specific language technology. (Malhas and Elsayed, 2020) introduced AyaTEC, a collection of verse-based questions answering on the Holy Qur'an. The dataset covers 11 topic categories of the Holy Qur'an and includes 207 questions and 1,762 answers covering that target the information needs of both curious and skeptical users.

Question answering(QA) aims to search for an answer to given questions is one widely investigated task among the most challenging ones in Natural Language Processing (NLP) (Alsubhi et al., 2021). A large volume of QA studies was devoted to the English language. However, it is not the case when it comes to the Arabic language due to its scarce resources. (Alanazi et al., 2021) provided a systematic review of QA research. They have examined English relevant studies and techniques. (Maraoui et al., 2021) have used standardized hadith, narrators, and Tafsir to develop a QA system. The system reached an accuracy of 92%. Current studies Investigated advanced techniques achieving state-of-the-art results. The language pre-trained transformers models helped to achieve significant progress in many NLP tasks. (Xue et al., 2021) used pre-trained multilingual T5 (mT5) as a generative model to generate multilingual question and answer pairs. (Alsubhi et al., 2021) have evaluated the fine-tuned AraBERTv2-base (Antoun et al., 2021a), AraBERTv0.2-large, and AraELECTRA (Antoun et al., 2021b) models using Arabic-SQuAD (Mozannar et al., 2019), ARCD (Mozannar et al., 2019), AQAD (Atef et al., 2020), and TyDiQA-GoldP (Clark et al., 2020) datasets. (Alsaleh et al., 2021) applied AraBERT language model with its two versions to binary classify the QurSim dataset pairs of verses. The best result was AraBERTv0.2 with 92% accuracy.

Imbalanced classification techniques have been applied widely in the field of data mining. It is used to classify the imbalanced classes that are not equal in the number of samples. The problem with imbalanced classes is that the classification performance tends to the class with more samples while the class with few samples will obtain poor performance. This problem can be occurred in the Qur'anic classification due to the difference in the number of verses (B. S. Arkok and Zeki, 2021) (B. Arkok and Zeki, 2021). (B. S. Arkok and Zeki, 2021) applied methods for under-sample (RUS), random oversample (ROS), synthetic minority oversampling technique (SMOTE), and random methods to classify the Qur'anic topics that are imbalanced. Many metrics were used in this research to evaluate the experimental results. (B. Arkok and Zeki, 2021) aimed to address balanced classification. They performed the Qur'anic text classification using SVM, Naïve Bayes, KNN, and J48. (Abdelnasser et al., 2014) provided an Arabic QA system for the Holy Quran that retrieves the most relevant verses corresponding to a question, then extracts the passage that contains the answer from the Quran and its interpretation books (Tafseer). Their system reached an accuracy of 85%. (Samy et al., 2019) provided a review of the Arabic Question Answering Systems, their approaches and challenges.

## 3. Shared Task and DataSet

In this section, we introduce the context of our paper which is the "Qur'an QA 2022 Shared Task", and we describe the given dataset in the task.

### 3.1 Qur'an QA 2022 Shared Task

Modern Standard Arabic (MSA) is the most understandable and common language in the Middle East and North Africa (by more than 400M people), and it is the official language used in media such as TV shows and newspapers. For this reason, it is encouraging for the scientific community to exert more effort to build systems specialized in solving situations related to the Arabic language. "Qur'an QA 2022 Shared Task" (Malhas et al, 2022) is shown in the efforts of developing the field of Arabic NLP and it is an attempt to enrich it, especially QA in the Qur'an script context.

The task is defined to find the correct answer(s) to a question in each passage (Figure 1), the questions are written in an MSA language, while the passage consists of verses from the Holy Qur'an written in Classical Arabic.



```
{
  "pq_id": "38:41-44_105",
  "passage": "واذكر عبدنا أيوب إذ نادى ربه أني مسني الشيطان بنصب وعذاب
              اركض برجلك هذا مغتسل بارد وشراب
              ووهبنا له أهله ومثلهم معهم رحمة منا وذكرى لأولي الألباب
              وخذ بيدك ضغثا فاضرب به ولا تحنث إنا وجدناه صابرا نعم العبد إنه أواب",
  "surah": 38,
  "verses": "41-44",
  "question": "من هو النبي المعروف بالصبر؟",
  "answers": [
    {
      "text": "أيوب",
      "start_char": 12
    }
  ]
}
{
  "pq_id": "74:32-48_330",
  "passage": "كلا والقمر. والليل إذ أدبر. والصبح إذا أسفر. إنها لإحدى الكبر
              نذيرا للبشر. لمن شاء منكم أن يتقدم أو يتأخر. كل نفس بما كسبت رهينة
              إلا أصحاب اليمين. في جنات يتساءلون. عن المجرمين. ما سلككم في سقر
              قالوا لم نك من المصلين. ولم نك نطعم المسكين
              وكنا نخوض مع الخائضين. وكنا نكذب بيوم الدين. حتى أتانا اليقين. فما تنفعهم شفاعة الشافعين",
  "surah": 74,
  "verses": "32-48",
  "question": "ما هي الدلائل التي تشير بأن الانسان مخير؟",
  "answers": [
    {
      "text": "لمن شاء منكم أن يتقدم أو يتأخر",
      "start_char": 76
    },
    {
      "text": "كل نفس بما كسبت رهينة",
      "start_char": 108
    }
  ]
}
```

Figure 1: Structure of the QRCD Dataset

From the accompanying passage for the given question, the model must return up to 5 possible responses, sorted from 1 (best) to 5 (lowest). Therefore, the partial Reciprocal Rank (pRR) will be the primary criterion for evaluating the competing models. Exact Match (EM) and F1@1, which are evaluation metrics applied only to the top predicted answer, will also be reported. The EM metric is a binary metric that only rewards a system if the top predicted answer exactly matches one of the gold answers. The F1@1 measurement, on the other hand, measures the token overlap between the best matching gold answer and the top predicted answer.

### 3.2 Dataset

The AyaTEC dataset was built with the help of three Islamic specialists in holy Qur'an interpretation (tafsir), and the questions were collected from different sources, questions were prepared by experts or asked by ordinary people on Islamic websites (Malhas and Elsayed, 2020), in the given dataset we have two types of samples, the single answer question samples, and the multi-answer question samples the first type is the questions that have only one answer for the question in the passage, the multi-answer

type is those which have multiple answers for the question in the same passage. The dataset is divided into training, development, and test set. Figure *2* shows that the question-passage-answer triplets represent the greatest count compared to the question-passage pairs, which is also greater than the count of passages. We note that the questions take the least part. Our model is trained on the question passage pair as input, and it is expected to provide the answer(s) to the question extracted from the given passage.
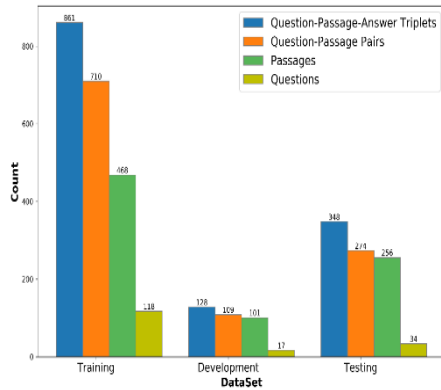


Figure 2: Task dataset parts distribution

## 4. Approach

Our approach consists of a Sequence-to-Sequence (Seq2Seq) model (Sutskever et al., 2014), based on the mT5 Language Model (Xue et al., 2021). We leverage the task as a Text-to-Text problem, which accepts an Arabic question and the passage as input, then extract the most relevant answers. Firstly, we preprocess and concatenate the question and the passage, then we fine-tune the mT5 LM to get an adequate answer. Figure *3* shows the general architecture of our approach.
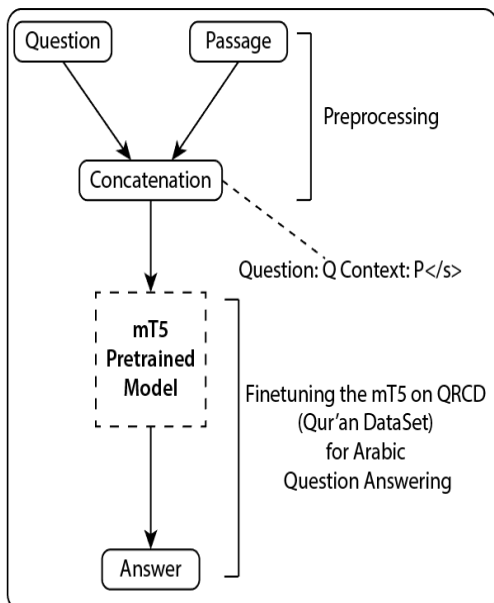


Figure 3: General architecture for our approach

We used mT5, which was pre-trained on a new Common Crawl-based DataSet covering 101 languages, Arabic language included. This model achieves top results on many NLP benchmarks while being flexible enough to be fine-tuned for a variety of important downstream tasks (Jawahar et al., 2021) (Agarwal et al., 2020) (Farahani et al., 2021) (Rothe et al., 2021). Another reason for choosing mT5 is that it is based on an encoder-decoder (Seq2Seq), which makes it appropriate for the Qur'an shared task. AraT5 (Nagoudi et al., 2021), is also a powerful Seq2Seq model which has better or comparable performance to mT5, but the source code is still not yet ready to use, at the time of writing this paper. In this sense, we fine tune mT5 on QRCD considering the input sequence (question + passage) as text and the output answer as text also.

### 4.1 Preprocessing

Given a question Q and a passage P from the Qur'an text, we form the input sequence as follows:

*Question : Q Context : P</s>*

Where </s> tag denotes the end of the input sequence. For all text in the dataset, we removed some stop words and since the original format of QRCD DataSet files is JSON, we converted the input and the output sequences to Tabulate-Separated Values in TSV files, the favorite format for fine-tuning mT5.

### 4.2 Fine-Tuning of mT5

In recent years, Transfer Learning (TL) has led to a new wave of cutting-edge results in Natural Language Processing (NLP). The power of TL comes from pretraining a model on abundantly available unlabeled text data with a self-supervised task. After that, the model can be refined on smaller labeled data sets, which often results in (much) better performance than training on the labeled data alone. The recent success of transfer learning was sparked in 2018 by ULMFiT (Howard and Ruder, 2018), ELMo and BERT (Devlin et al., 2019). The 2019 year saw the development of a wide variety of new methods like GPT (Radford et al., 2019), XLNet (Yang et al., 2019), RoBERTa, ALBERT (Lan et al., 2019), Reformer and MT-DNN (Liu et al., 2019). The pace of progress on the ground has made it difficult to assess the most significant improvements and their effectiveness when combined. After a preliminary study that we did, we did not use these models because most of them do not deal with the Arabic language. Or, for for multilingual models (which also process Arabic), their architectures do not help to use them for seq2seq tasks (text generation or extraction task, the case of the task being processed), which led us to fine-tune the mT5 pretrained model.
Our implementation details are explained in the next section.

## 5. Experimentation and Results

Using the original mT5 recipe, we consider three model sizes: Base (580M), Large (1.2B) and XL (3.7B). The increase in parameters comes from the larger vocabulary used in mT5 (covering 101 languages).
The following setup is used across all our experiments. We used a global batch size of 64 with a max input length of 354 and a max target length of 150. AdamW was used as

the optimizer with a constant learning rate set to 0.003. Models were trained for 15k steps and we saved the best checkpoint every 5k steps. We kept the other default hyperparameters suggested by T5 (Raffel et al., 2019) and mT5 papers. Finally, we selected the best checkpoint based on the Dev Set validation results. The training was performed on a TPU v3, using Google Colab Pro, with Google Cloud Platform (GCP) for storage.

We finetuned our models on the Train Set of QRCD Dataset, then we evaluated them on the Development Set. Finally, we generated predictions on the Test Set and submitted our runs for evaluation by the community of the Qur'an QA 2022 shared task using Codalab platform. Table *1* shows the results of our experimentations over the Development Set and Table *2* shows our best run result on the Test Set.

| Model | pRR | Exact Match | F1 Score |
|---|---|---|---|
| mT5-Base | 0.79 | 0.65 | 0.79 |
| mT5-Large | 0.91 | 0.71 | 0.91 |
| mT5-XL | 0.98 | 0.97 | 0.98 |

Table 1: Results of evaluation on the Dev Set

| Model | pRR | Exact Match | F1 Score |
|---|---|---|---|
| mT5-XL | 0.43 | 0.20 | 0.40 |

Table 2: Best result of evaluation on the Test Set

As shown in the results tables, our best model achieves 0.98, 0.97 and 0.98 (pRR, EM and F1 scores respectively) on the Development Set, which are very good results, but it decreases on the Test Set, mainly on the EM metric.

## 6.   Analysis and discussion

We have achieved good results in the Developpement set. However, there was a degradation when applying the model to the test set. This may be due to many reasons among which the ones presented in Table *3*, in which we give some examples of our model prediction and its interpretation based on our estimated human response since the gold answers aren't released yet. The model has a high performance when it comes to extractive answers such as in the first example. Moreover, it may generate singular terms from plurals such as in the second example. In the third example, our model generates close answers. However, the system may generate distant responses since sometimes the names of Prophets and messengers may be considered angels, which raises the need for named entities recognition. Our model may generate the same verse as the given passage if the response is difficult to be extracted as in the fifth example. Moreover, sometimes our model is not only extractive but it may become generative as in example six. The model can generate the best answer but with a low rank which has been fixed to five as the organizers of the competition consider the top five ranked answers. At this level, the right answer may be not reached yet. Increasing the number of estimated answers causes an increase in execution time and computational requirements.

| N | Passage | Question | Response | Interpretation |
|---|---|---|---|---|
| 1 | والذين يتوفون منكم ويذرون أزواجا يتربصن بأنفسهن أربعة أشهر وعشرا فإذا بلغن أجلهن فلا جناح عليكم فيما فعلن في أنفسهن بالمعروف والله بما تعملون خبير. ولا جناح عليكم فيما عرضتم به من خطبة النساء أو أكننتم في أنفسكم علم الله أنكم ستذكرونهن ولكن لا تواعدوهن سرا إلا أن تقولوا قولا معروفا ولا تعزموا عقدة النكاح حتى يبلغ الكتاب أجله واعلموا أن الله يعلم ما في أنفسكم فاحذروه واعلموا أن الله غفور حليم. لا جناح عليكم إن طلقتم النساء ما لم تمسوهن أو تفرضوا لهن فريضة ومتعوهن على الموسع قدره وعلى المقتر قدره متاعا بالمعروف حقا على المحسنين. وإن طلقتموهن من قبل أن تمسوهن وقد فرضتم لهن فريضة فنصف ما فرضتم إلا أن يعفون أو يعفو الذي بيده عقدة النكاح وأن تعفوا أقرب للتقوى ولا تنسوا الفضل بينكم إن الله بما تعملون بصير". As for those among you who die and leave widows behind, their widows shall wait by themselves for four months and ten days. When they have reached their term, there is no blame on you regarding what they might honorably do with themselves. God is fully acquainted with what you do. You commit no error by announcing your engagement to women, or by keeping it to yourselves. God knows that you will be thinking about them. But do not meet them secretly, unless you have something proper to say. And do not confirm the marriage tie until the writing is fulfilled. And know that God knows what is in your souls, so beware of Him. And know that God is Forgiving and Forbearing. You commit no error by divorcing women before having touched them, or before having set the dowry for them. And compensate them— the wealthy according to his means, and the poor according to his means—with a fair compensation, a duty upon the doers of good. If you divorce them before you have touched them, but after you had set the dowry | كم مدة عدة الأرملة ؟ /kam mudat eidat al'armala/ How long is a widow's waiting period? | أربعة أشهر وعشرا / 'arbaeat 'ashhur waeashran/ four months and ten | Our model extracts the correct answer from a complicated paragraph |

115

| # | Verse | Question | Answer | Note |
|---|---|---|---|---|
| | for them, give them half of what you specified—unless they forego the right, or the one in whose hand is the marriage contract foregoes it. But to forego is nearer to piety. And do not forget generosity between one another. God is seeing of everything you do[2]. | | | |
| 2 | ولقد ذرأنا لجهنم كثيرا من الجن والإنس لهم قلوب لا يفقهون بها ولهم أعين لا يبصرون بها ولهم آذان لا يسمعون بها أولئك كالأنعام بل هم أضل أولئك هم الغافلون". <br><br> We have destined for Hell multitudes of jinn and humans. They have hearts with which they do not understand. They have eyes with which they do not see. They have ears with which they do not hear. These are like cattle. In fact, they are further astray. These are the heedless. | هل أشار القرآن إلى العضو الذي يعقل به الإنسان؟ / hal 'ashar alquran 'iilaa aleudw aladhi yueqil bih al'iinsan / Did the Qur'an refer to the organ with which a person can reason? | الأذن / al'udhun/ Ear | Generate Singular terms |
| 3 | يا أيها النبي لم تحرم ما أحل الله لك تبتغي مرضات أزواجك والله غفور رحيم. قد فرض الله لكم تحلة أيمانكم والله مولاكم وهو العليم الحكيم. وإذ أسر النبي إلى بعض أزواجه حديثا فلما نبأت به وأظهره الله عليه عرف بعضه وأعرض عن بعض فلما نبأها به قالت من أنبأك هذا قال نبأني العليم الخبير. إن تتوبا إلى الله فقد صغت قلوبكما وإن تظاهرا عليه فإن الله هو مولاه وجبريل وصالح المؤمنين والملائكة بعد ذلك ظهير. عسى ربه إن طلقكن أن يبدله أزواجا خيرا منكن مسلمات مؤمنات قانتات تائبات عابدات سائحات ثيبات وأبكارا", <br><br> O prophet! Why do you prohibit what God has permitted for you, seeking to please your wives? God is Forgiving and Merciful. God has decreed for you the dissolution of your oaths. God is your Master. He is the All-Knowing, the Most Wise. The Prophet told something in confidence to one of his wives. But when she disclosed it, and God made it known to him; he communicated part of it, and he avoided another part. Then, when he informed her of it, she said, "Who informed you of this?" He said, "The All-Knowing, the All-Informed, informed me." If you repent to God, then your hearts have listened. But if you band together against him, then God is his Ally, as is Gabriel, and the righteous believers. In addition, the angels will assist him. Perhaps, if he divorces you, his Lord will give him in exchange wives better than you: submissive, believing, obedient, penitent, devout, fasting—previously married, or virgins. | من هم الملائكة المذكورون في القرآن /man humm almalayikat almadhkurun fi alquran/ Who are the angels mentioned in the Qur'an? | جبريل وصالح المؤمنين جبريل وصالح / المؤمنين / Gabriel and the righteous believers | The need for named entities recognition |
| 4 | وكذلك أنزلناه آيات بينات وأن الله يهدي من يريد. إن الذين آمنوا والذين هادوا والصابئين والنصارى والمجوس والذين أشركوا إن الله يفصل بينهم يوم القيامة إن الله على كل شيء شهيد. ألم تر أن الله يسجد له من في السماوات ومن في الأرض والشمس والقمر والنجوم والجبال والشجر والدواب وكثير من الناس وكثير حق عليه العذاب ومن يهن الله فما له من مكرم إن الله يفعل ما يشاء. <br><br> Thus We revealed it as clarifying signs, and God guides whomever He wills. Those who believe, and those who are Jewish, and the Sabeans, and the Christians, and the Zoroastrians, and the Polytheists—God will judge between them on the Day of Resurrection. God is witness to all things. Do you not realize that to God prostrates everyone in the heavens and everyone on earth, and the sun, and the moon, and the stars, and the mountains, and the trees, and the animals, and many of the people? But many are justly deserving of punishment. Whomever God shames, there is none to honor him. God does whatever He wills. | ما الدلائل على أن القرآن ليس من تأليف سيدنا محمد (ص)؟ / ma aldalayil ealaa 'ana alquran lays min talif sayidina muhamad (s)?/ What is the evidence that the Qur'an was not written by our master Muhammad (PBUH)? | من يهن الله فما له من مكرم / man yahin allah fama lah min makram / Whomever God shames, there is none to honor him. | The wrong answer is repeated for all ranks. |

---

| | | | | |
|---|---|---|---|---|
| 5 | ولقد آتينا موسى تسع آيات بينات فاسأل بني إسرائيل إذ جاءهم فقال له فرعون إني لأظنك يا موسى مسحورا. قال لقد علمت ما أنزل هؤلاء إلا رب السماوات والأرض بصائر وإني لأظنك يا فرعون مثبورا. فأراد أن يستفزهم من الأرض فأغرقناه ومن معه جميعا. وقلنا من بعده لبني إسرائيل اسكنوا الأرض فإذا جاء وعد الآخرة جئنا بكم لفيفا".<br><br>We gave Moses nine clear signs—ask the Children of Israel. When he went to them, Pharaoh said to him, "I think that you, Moses, are bewitched." He said, "You know that none sent these down except the Lord of the heavens and the earth—eye openers; and I think that you, Pharaoh, are doomed." He resolved to scare them off the land, but We drowned him, and those with him, altogether. After him, We said to the Children of Israel, "Inhabit the land, and when the promise of the Hereafter arrives, We will bring you all together." | ما هي معجزات النبي موسى عليه السلام؟<br>/ ma hi muejizat alnabii musaa ealayh alsalam /<br>What are the miracles of Prophet Moses, peace be upon him? | - | The difficulty of response extraction |
| 6 | ويوم نحشرهم جميعا ثم نقول للذين أشركوا أين شركاؤكم الذين كنتم تزعمون. ثم لم تكن فتنتهم إلا أن قالوا والله ربنا ما كنا مشركين. انظر كيف كذبوا على أنفسهم وضل عنهم ما كانوا يفترون. ومنهم من يستمع إليك وجعلنا على قلوبهم أكنة أن يفقهوه وفي آذانهم وقرا وإن يروا كل آية لا يؤمنوا بها حتى إذا جاءوك يجادلونك يقول الذين كفروا إن هذا إلا أساطير الأولين. وهم ينهون عنه وينأون عنه وإن يهلكون إلا أنفسهم وما يشعرون.<br><br>On the Day when We gather them all together, then say to the idolaters, "Where are your idols, those you used to claim?" Then their only argument will be to say, "By God, our Lord, we were not idolaters." Look how they lied to themselves. And what they invented deserted them. Among them are those who listen to you; but We place covers over their hearts, to prevent them from understanding it, and heaviness in their ears. Even if they see every sign, they will not believe in it. Until, when they come to you, to argue with you, those who disbelieve will say, "These are nothing but myths of the ancients." They keep others from it, and avoid it themselves; but they ruin only their own souls, and they do not realize. | هل أشار القرآن إلى العضو الذي يعقل به الإنسان؟<br>/ hal 'ashar alquran 'iilaa aleudw aladhi yueqil bih al'iinsan /<br>Did the Qur'an refer to the organ with which a person can reason? | العضو من العضو<br>/ aleudw min aleudw/<br>Organ from organ<br><br>قلوبهم أكنة<br>/qulubuhum 'akuna /<br>Their hearts are veils | Generate new terms<br><br>The correct answer hasn't the first rank |

Table 3: Error analysis

We have performed an in depth analysis of the proposed model and the obtained results and we have concluded the following points. Our model can be improved by addressing stemming, lemmatization, or word root and using synonyms to match a large set of similar questions. Also, the analysis done on the false predictions may be affined according to the questions using association rules. Since QRCD is a dataset of limited size yet imbalanced, we can use external Arabic data mainly Qur'an data and Tafsir to pretrain T5 from scratch then do fine tuning and explore data augmentation to augment the number of entries in the dataset. The data augmentation will be mainly applied to questions for preserving the meaning of the Qur'an verses. We assume also that larger base models could provide even more improvement. In fact, we can use mT5-XXL (about ×4 larger than mT5-XL), but this requires more resources precisely in terms of TPU and RAM. We have tested this model on Google Colab Pro (with 32G RAM) but it gives Coda out of memory.

## 7. Conclusion

In this paper, we presented our approach based on the Language Model mT5. We have finetuned it on the QRCD DataSet proposed by the Qur'an QA 2022 shared task. We fine-tuned our model using the Train Set, then we evaluated our model on the Development Set. After that, we generated predictions on the Test Set and then get an official evaluation from the community of the shared task. The result are compared in the overview paper (Malhas et al, 2022). Our best model can achieve very good results on the Development Set but less on the Test Set.

Finally, we discussed obtained results and challenges, then proposed potential solutions for possible improvements, that we can leverage as future works.

## 8. Bibliographical References

Abdelnasser, H., Ragab, M., Mohamed, R., Mohamed, A., Farouk, B., El-Makky, N.M. and Torki, M., 2014, October. Al-Bayan: an Arabic question answering system for the Holy Quran. *In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 57-64).

Adeleke, A., Samsudin, N.A., Othman, Z.A. and Khalid, S.A., 2019. A two-step feature selection method for quranic text classification. Indones. *J. Electr. Eng. Comput. Sci*, 16(2), pp.730-736.

Agarwal, O., Kale, M., Ge, H., Shakeri, S. and Al-Rfou, R., 2020. Machine translation aided bilingual data-to-text generation and semantic parsing. *In Proceedings of the 3rd International Workshop on Natural Language*

*Generation from the Semantic Web (WebNLG+)* (pp. 125-130).

Alanazi, S.S., Elfadil, N., Jarajreh, M., Algarni, S., 2021. Question Answering Systems: A Systematic Literature Review. *Int. J. Adv. Comput. Sci. Appl. 12.* https://doi.org/10.14569/IJACSA.2021.0120359

Alqahtani, M. and Atwell, E., 2016, June. Arabic quranic search tool based on ontology. *In International Conference on Applications of Natural Language to Information Systems* (pp. 478-485). Springer, Cham.

Alsaleh, A.N., Atwell, E. and Altahhan, A., 2021, April. Quranic Verses Semantic Relatedness Using AraBERT. *In Proceedings of the Sixth Arabic Natural Language Processing Workshop* (pp. 185-190). Leeds.

Al-Salhi, R.Y. and Abdullah, A.M., 2022. Building Quranic stories ontology using MappingMaster domain-specific language. *International Journal of Electrical & Computer Engineering* (2088-8708), 12(1).

Alsubhi, K., Jamal, A. and Alhothali, A., 2021. Pre-trained transformer-based approach for Arabic question answering: A comparative study. arXiv preprint arXiv:2111.05671.

Alwaneen, T.H., Azmi, A.M., Aboalsamh, H.A., Cambria, E. and Hussain, A., 2022. Arabic question answering system: a survey. *Artificial Intelligence Review*, 55(1), pp.207-253.

Antoun, W., Baly, F., Hajj, H., 2021a. AraBERT: Transformer-based Model for Arabic Language Understanding. ArXiv200300104 Cs.

Antoun, W., Baly, F., Hajj, H., 2021b. AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding. ArXiv201215516 Cs.

Arkok, B. and Zeki, A.M., 2021, June. Classification of Quranic Topics Using Ensemble Learning. In *2021 8th International Conference on Computer and Communication Engineering (ICCCE)* (pp. 244-248). IEEE.

Arkok, B.S. and Zeki, A.M., 2021. Classification of Quranic topics based on imbalanced classification. *Indones. J. Electr. Eng. Comput. Sci*, p.678.

Atef, A., Mattar, B., Sherif, S., Elrefai, E. and Torki, M., 2020, November. AQAD: 17,000+ Arabic Questions for Machine Comprehension of Text. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)* (pp. 1-6). IEEE.

Clark, J.H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V. and Palomaki, J., 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8, pp.454-470.

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*.

Dukes, K., Atwell, E. and Habash, N., 2013. Supervised collaboration for syntactic annotation of Quranic Arabic. *Language resources and evaluation*, 47(1), pp.33-62.

Farahani, M., Gharachorloo, M. and Manthouri, M., 2021, March. Leveraging ParsBERT and Pretrained mT5 for Persian Abstractive Text Summarization. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)* (pp. 1-6). IEEE.

Howard, J. and Ruder, S., 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Jawahar, G., Nagoudi, E.M.B., Abdul-Mageed, M. and Lakshmanan, L.V., 2021. Exploring text-to-text transformers for english to hinglish machine translation with synthetic code-mixing. *arXiv preprint arXiv:2105.08807*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R., 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Liu, X., He, P., Chen, W. and Gao, J., 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

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), pp.1-21.

Maraoui, H., Haddar, K. and Romary, L., 2021. Arabic factoid Question-Answering system for Islamic sciences using normalized corpora. *Procedia Computer Science*, 192, pp.69-79.

Mohamed, E.H. and Shokry, E.M., 2020. QSST: A Quranic Semantic Search Tool based on word embedding. *Journal of King Saud University-Computer and Information Sciences*.

Mohd, M., Qamar, F., Al-Sheikh, I. and Salah, R., 2021. Quranic optical text recognition using deep learning models. *IEEE Access*, 9, pp.38318-38330.

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

Nagoudi, E.M.B., Elmadany, A. and Abdul-Mageed, M., 2021. AraT5: Text-to-Text Transformers for Arabic Language Understanding and Generation. *arXiv preprint arXiv:2109.12068*.

Osman, M., Hilal, A., Alhawarat, M., 2015. Fine-Grained Quran Dataset. *Int. J. Adv. Comput. Sci. Appl. 6.* https://doi.org/10.14569/IJACSA.2015.061241

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), p.9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. https://doi.org/10.48550/ARXIV.1910.10683

Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P., 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. https://doi.org/10.48550/ARXIV.1606.05250

Rothe, S., Mallinson, J., Malmi, E., Krause, S., Severyn, A., 2021. A Simple Recipe for Multilingual Grammatical Error Correction. https://doi.org/10.48550/ARXIV.2106.03830

Samy, H., E., E., Shaalan, K., 2019. Arabic Question Answering: A Study on Challenges, Systems, and Techniques. *Int. J. Comput. Appl.* 181, 6–14. https://doi.org/10.5120/ijca2019918524

Sutskever, I., Vinyals, O. and Le, Q.V., 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C., 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *ArXiv201011934 Cs*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R. and Le, Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.