Proceedings of the OSACT 2022 Workshop @LREC2022, pages 104–111
Marseille, 20 June 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

Aly Mostafa, Omar Mohamed
Department Of Computer Science, Faculty of Computers and Artificial Intelligence, Helwan University
Helwan, Egypt
{alymostafa, omar_20170353 }@fci.helwan.edu.eg

Abstract

Recently, significant advancements were achieved in Question Answering (QA) systems in several languages. However, QA systems in the Arabic language require further research and improvement because of several challenges and limitations, such as a lack of resources. Especially for QA systems in the Holy Qur'an since it is in classical Arabic and most recent publications are in Modern Standard Arabic. In this research, we report our submission to the Qur'an QA 2022 Shared task-organized with the 5^{th} Workshop on Open-Source Arabic Corpora and Processing Tools Arabic (OSACT5). We propose a method for dealing with QA issues in the Holy Qur'an using Deep Learning models. Furthermore, we address the issue of the proposed dataset's limited sample size by fine-tuning the model several times on several large datasets before fine-tuning it on the proposed dataset achieving 66.9% pRR 54.59% pRR on the development and test sets, respectively.
Keywords: Holy Qur'an, Question Answering System, Information retrieval,

1. Introduction

Extractive Question Answering is essential for extracting a span of text from a given context paragraph as the answer to a specified question. It is a subset of Natural Language Processing and Information Retrieval. Question Answering has made significant development in recent years with applications in search engines. This is because large pre-trained language models and self-supervised learning, such as BERT(Devlin et al., 2018), T5 (Raffel et al., 2019), MT5 (Xue et al., 2020), ALBERT(Lan et al., 2019), and BART(Lewis et al., 2019), which use the Transformer architecture (Vaswani et al., 2017) to develop robust language models for a range of NLP tasks specified by benchmarks, such as GLUE (Wang et al., 2018). Furthermore, new datasets, such as SQuAD (Rajpurkar et al., 2016), have brought more complicated questions with inference-based context to the question answering task. Arabic is a Semitic language spoken by about 250 million people in the Middle East and North Africa. It is the official language of 26 countries and one of the six official languages of the United Nations. Classical Arabic (CA), Modern Standard Arabic (MSA), and Colloquial Arabic are the three main variants of Arabic (Arabic,). CA is the style that illustrates the Holy Qur'an. The Holy Qur'an came to fruition in the sixth century CE, and Arabic has evolved over the centuries, but not significantly. CA is the basis of the mediaeval languages of Arab tribes. The phrase structure is the same as in MSA's current form. Several minor variations exist between the CA and MSA, like grammar and phrase punctuation. More than 1.8 billion Muslims around the world revere the Qur'an. It is the primary source of Islamic knowledge. The Holy Qur'an consists of 114 chapters and 6,236 verses of varying lengths, totalling around 80k Arabic words. An Ayah-meaning verse in the Qur'an- might refer to one or more topics, and several Ayahs might address the same topic in similar contexts. Because of the Arabic word structure, many regards a morphological study of the Arabic language as an involute endeavour; Arabic words consist of a maximum of four letters root verb. Also, each word consists of a root and other affixes (prefix, infix, suffix) and can have many interpretations depending on the diacritics marks that make the word two-dimensional. Finally, while the Arabic language is one of the most widespread globally, it is a low-resource language by many standers since it suffers from a scarcity of resources, such as lacking large pre-trained models and corpora. These problems make Arabic NLP, NLU, and IR research more challenging than in other languages like English or Chinese. Most of the attention in the Arabic literature research is towards Modern Standard Arabic due to the availability of its resources compared with CA, especially the QA systems on the Holy Qur'an (Abdelnasser et al., 2014), (Adany and others, 2017), (Hamdelsayed et al., 2017), (Hakkoum and Raghay, 2016), (Hamdelsayed and Atwell, 2016b), where a question is likely to be posed in MSA since it is the most common Arabic version used in Arabic-speaking countries today. While significant efforts have been made to offer dependable QA systems for other applications in the Arabic language (Brini et al., 2009), (Mohammed et al., 1993), (Nabil et al., 2017), (Abdelmegied et al., 2017), there have been few attempts to research QA for the Holy Qur'an. In this work, we present a method for dealing with QA in the Holy Qur'an. We analyzed the existing pre-trained models, namely MARBERTv2 and AraBERTv2, for QA in Arabic and found that the AraELECTRA model produces cutting-edge outcomes in a variety of Arabic

Set	Percentage	#Question-Passage Pairs	#Question-Passage-Answer Triplets
Training	65%	710	861
Development	10%	109	128
Test	25%	274	348

Table 1: QRCD Distribution

QA datasets (Antoun et al., 2020)(Abdul-Mageed et al., 2021). Our approach consists of three stages. First, the chosen model is fine-tuned using the Ar-TyDi QA dataset (Clark et al., 2020a). Second, we improved its efficiency by fine-tuning it on a larger corpora (Arabic-SQuAD and ARCD) (Mozannar et al., 2019). Finally, we fine-tuned the same model using QRCD (Qur'anic Reading Comprehension Dataset), achieving 66.9%, 54.59% partial Reciprocal Rank (pRR) (Malhas and Elsayed, 2020) on the development and test set, respectively. The rest of the paper is organized as follows. section 2 provides a review of previous Question Answering systems in the Holy Qur'an literature. section 3 describes the proposed dataset. section 4 proposes the model of the QA. section 5 discusses the results and performance evaluation. Finally, we conclude in section 6.

2. Related Work

This section discusses previous research and applications addressing Question-Answering challenges in the Holy Qur'an, as well as the methodologies, datasets, strengths employed, and drawbacks.
(Shmeisani et al., 2014) Their model is composed of three layers that apply a semantic approach of Arabic ontology for Qur'anic content and analyze user inquiries expressed in Arabic. A Question processing layer, based on POS tagging, a semantic layer, and a query builder, enhancing query keywords. Their method was able to retrieve answers even when the user's precise words were not found in the Holy Qur'an.

(Abdelnasser et al., 2014) Their model is a Question-Answering System that comprises two modules. The first module accepts the input in the form of a question and retrieves the appropriate verses based on their semantics. The second module returns the extracted answer from the retrieved verses alongside their Tafseer. Their method achieved 85% accuracy in the top three rankings.

(Adany and others, 2017) The authors proposed six distinct ways of dealing with the problem of question-answering in the Holy Qur'an. Approaches were eliminating stopwords, diacritics, and special symbols, using Lucene indexing patterns, exaggeration formulas patterns, and single, dual, and plural patterns. They test their model on a corpus of only two chapters (AlBagrah and Alfatihah chapters)

(Hakkoum and Raghay, 2016) Introduce a Holy Qur'an Q&A system based on the development of an ontology that comprises a set of concepts and semantic relations between distinct inquiries and responses. The model has a precision of 95% and a recall of 73%.

(Hamed and Aziz, 2016) The authors propose a QAS based on identifying verses based on their content by utilizing the Neural Network (NN) approach. Based on N-gram and similarity scores, the model retrieves the ranking verses. The dataset used was Abdullah Yusuf Ali's translation of the English version of the Holy Qur'an (Ali, 2000).

(Hamoud and Atwell, 2016) They presented a Question Answering System for the Holy Qur'an that retrieves answers based on the contextual meaning of the keywords in the user's query, with the used corpus consisting of queries and their answers. They also used question paraphrasing as a data augmentation method to improve Question Answering (QA) performance. For the Arabic version, they attained a precision of 79% and a recall of 76%.
(Hamdelsayed and Atwell, 2016a) Their method improved the retrieved responses by adding additional semantic meaning to the words. Thye achieved this by manipulating several aspects of the Arabic language such as numbering, single, dual, and plural. They tested their method to the test using a corpus of questions and answers from the Holy Qur'an.

According to the results of this survey, we identified many flaws. To begin, several of the approaches for Q&A in the Holy Qur'an operate poorly. In addition, numerous studies aim to develop a static Qur'an ontology or hierarchical tree based on Qur'anic ontology. There is a scarcity of dynamic tools that prompt users to query using an abstract statement or a question. While there is a lot of research on developing Qur'an ontologies, there is little research on experimenting with state-of-the-art Deep Learning models for Q&A. In addition, there is a scarcity of structured, reusable, and publicly accessible gold standard test datasets for the Q&A in the Holy Qur'an. As a result, there are limitations in comparing the performance of different Q&A systems for the Holy Qur'an. The goal of this study is to solve previous

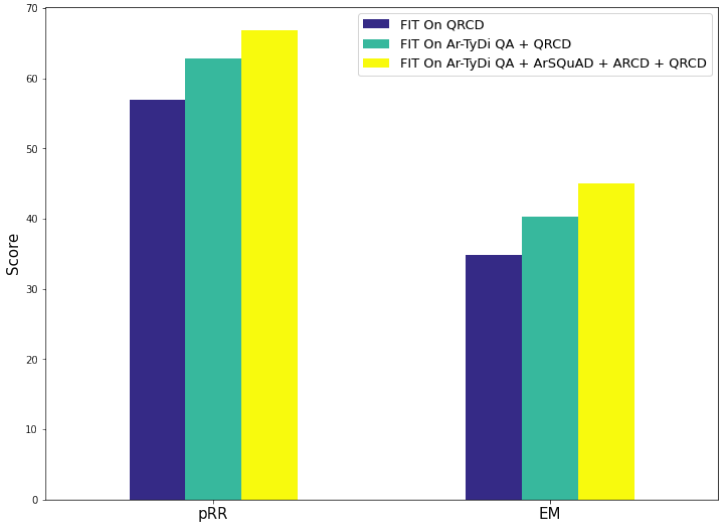

Figure 1: ARAELECTRA Fine-tuned (FIT) on different datasets. X-axis shows the pRR and EM of the different models. Y-axis shows the Score of each model

limitations by applying and analyzing state-of-the-art models and training those models on QRCD (Qur'anic Reading Comprehension Dataset) (Malhas and Elsayed, 2020)(Malhas and Elsayed, 2022)(Malhas et al., 2022).

3. Dataset

QRCD (Qur'anic Reading Comprehension Dataset) (Malhas and Elsayed, 2020)(Malhas and Elsayed, 2022) is a publically accessible dataset for extractive Question-Answering tasks (Machine Reading Comprehension) for the Holy Qur'an. There are two types of question passages in the dataset: question-passage pairs (1,093 records) and question-passage-answer triplets (1,337 records). Since the same question in the QRCD can be presented in multiple Qur'anic chapters, each passage may have multiple occurrences. Furthermore, the same passage may accompany different questions. The source of the Qur'anic text in QRCD is the Tanzil project download page, which provides validated versions of the Holy Qur'an in a variety of programming techniques. The simple-clean text style was chosen for convenience of usage. The proposed dataset is divided into three sets: training, development, and test; the dataset distribution is illustrated in Table 1

4. Methodology

In this section, we discuss the essential components of the proposed method, starting with an overview of the pre-trained models used and going over the fine-tuning process on the various datasets to overcome the dataset's small sample size problem. Finally, we present the fine-tuning of the QRCD.

4.1. ELECTRA

The Electra pre-training (Clark et al., 2020b) method involves training two neural networks, a generator G and a discriminator D. Each one essentially consists of a bi-directional encoder (e.g., Small BERT). The generator has been conditioned to conduct masked language modelling (MLM). MLM initially chooses a random set of positions (integers between 1 and n) to mask out m given an input. The generator's learning goal is to anticipate the original identities of the masked-out tokens. The discriminator is trained to discriminate between tokens in the data and tokens substituted by generator samples. While the Electra pre-training approach appears to be similar to GAN (Goodfellow et al., 2014), there are key variations. First, it was observed that changing the token status from "fake" to "real" after generating the correct token improves the results on downstream tasks. Second, GANs have trained adversarially to deceive the discriminator, whereas the Electra model is trained with maximum probability. Finally, The Electra pre-training's loss is a combination of MLM loss and discriminator loss as follows:

$$\min_{\theta_G,\theta_D} \sum_{\in\mathcal{X}} \mathcal{L}_{\text{MLM}}(,\theta_G) + \lambda\mathcal{L}_{\text{Disc}}(,\theta_D)$$

4.2. ARAELECTRA

The ARAELECTRA model (Antoun et al., 2020) is an ELECTRA-based Arabic language representation model. The goal of developing this model for the Arabic language is to improve current Arabic reading comprehension. The model is a bidirectional transformer encoder model with 136M parameters that includes 12 encoder layers, 12 attention heads, 768 hidden sizes, and 512 maximum input sequence lengths. As stated in the previous section, the ARAELECTRA pre-training objective replaced token detection (RTD). The pre-training dataset is identical to that of ARABERTV0.2 (Antoun et al.,). The dataset used is a collection of Arabic corpora

Model	Dataset	Loss Function	pRR
MARBERTv2	QRCD	Cross-Entropy	50.46
ARAELECTRA	QRCD	Cross-Entropy	56.07
		Focal-Loss + Label Smoothing	57.60
		Focal-Loss	**59.84**
ARAELECTRA	Ar-TyDi QA + QRCD	Cross-Entropy	62.88
ARAELECTRA	**Ar-TyDi QA + ArSQuAD + ARCD + QRCD**	**Cross-Entropy**	**66.9**
		Focal-Loss	61.2
		Dice-Loss	61.4

Table 2: The Results Of Applying Different Models on Development Set

totalling 8.8 billion words, the majority of which are news articles.

4.3. Fine-Tuning On Ar-TyDi QA

TyDi QA (Clark et al., 2020a) is a question-answer dataset with 204K question-answer pairs that covers 11 typologically diverse languages. TyDi QA's typology includes a wide range of languages. TyDi QA has a significant benefit in that data is collected directly in each language without the usage of neural machine translation. Furthermore, to give a realistic information-seeking activity, the questions are written by individuals who want to know the answer but do not yet know the answer, as opposed to SQuAD. We trained our model in the Arabic subset of the TyDi QA, the training examples were 15,364 and testing 941 with a total of 16305 samples. The model achieved an F1 score of 85.7% and an exact match of 72.9%.

4.4. Fine-Tuning On Arabic SQuAD and ARCD

Arabic-SQuAD: (Mozannar et al., 2019) The Arabic-SQuAD was obtained by a neural machine translation from the English version; the translation was carried out using the Google Translate neural machine translation (NMT) API; the authors chose to translate SQuAD version 1.1 because it was the most popular benchmark for Machine Reading Comprehension (MRC). Out of the 536 articles in the SQuAD training set, they only translated the first 231. The final distribution of the dataset is 48,344 questions on 10,364 paragraphs.

Arabic Reading Comprehension Dataset (ARCD): (Mozannar et al., 2019) ARCD's questions was written by proficient Arabic speakers. They retrieved the top 1000 viewed articles on Wikipedia in 2018 and then randomly sampled 155 articles. They tried to make the articles' topics as diverse as possible, including religious and historical figures, sports celebrities, countries, and companies. Finally, they requested a worker to create three question-answer pairs for each paragraph in unambiguous Modern Standard Arabic, with the answer to each question being an exact span of text from the article's paragraph. ARCD is composed of 1,395 questions along with their passages and answers.

Model Training: To train the proposed model, we merged the previous datasets, Arabic SQuAD and ARCD. The model was initially fine-tuned on Ar-TyDiQA, then the same model was used to fine-tune on the combined two datasets, yielding 49,739 questions, with the training set including 39,791 questions and the test set containing 9,948 questions, together with their passages and answers. The model obtained an F1 score of 70.05% and an exact match score of 36.47%.

4.5. Fine-Tuning On QRCD

Following the previous phases of fine-tuning the model on the Ar-TyDi QA and Arabic SQuAD + ARCD, we acquired the model's weights and fine-tuned it again on QRCD to improve performance. On the Development and Test sets, the model achieved 66.9% pRR and 54.59% pRR, respectively. We follow this approach due to the QRCD's small sample size because it is known that Deep Learning models require a large sample size even if it is pre-trained on a large corpus, and the transferred knowledge from previous dataset training helped the model retrieve better answers because the MSA(the style in which Ar-TyDi QA, Arabic SQuAD, and ARCD are written) is similar in many characteristics to CA (the style in which the Holy Qu'ran is written). This method is used in a variety of fields (Thrun and Pratt, 2012), (Menegola et al., 2017), (Jang et al., 2019), (Silver et al., 2013), not just in the Machine Reading Comprehension task. Even if the model suffers from catastrophic forgetting (McCloskey and Cohen, 1989) during the different phases of our training pipeline, which is split into three phases as detailed in the previous section, it may benefit from the transferred knowledge.

5. Results and Discussion

5.1. Performance Metrics

Because this is considered a ranking task, the QA system is needed to return up to 5 potential answers. The pRR

Question من هم الملائكة المذكورون في القرآن؟
واتبعوا ما تتلو الشياطين على ملك سليمان وما كفر سليمان ولكن الشياطين كفروا يعلمون الناس السحر وما أنزل على الملكين ببابل **هاروت وماروت** وما يعلمان من أحد حتى يقولا إنما نحن فتنة فلا تكفر فيتعلمون منهما ما يفرقون به بين المرء وزوجه وما هم بضارين به من أحد إلا بإذن الله ويتعلمون ما يضرهم ولا ينفعهم ولقد علموا لمن اشتراه ما له في الآخرة من خلاق ولبئس ما شروا به أنفسهم لو كانوا يعلمون. ولو أنهم آمنوا واتقوا لمثوبة من عند الله خير لو كانوا يعلمون
أرأيت الذي ينهى. عبدا إذا صلى. أرأيت إن كان على الهدى. أو أمر بالتقوى. أرأيت إن كذب وتولى. ألم يعلم بأن الله يرى. كلا لئن لم ينته لنسفعا بالناصية. ناصية كاذبة خاطئة. فليدع ناديه. سندع **الزبانية**. كلا لا تطعه واسجد واقترب
Question ما هي منزلة من يقتل في سبيل الله؟
فليقاتل في سبيل الله الذين يشرون الحياة الدنيا بالآخرة **ومن يقاتل في سبيل الله فيقتل أو يغلب فسوف نؤتيه أجرا عظيما**. وما لكم لا تقاتلون في سبيل الله والمستضعفين من الرجال والنساء والولدان الذين يقولون ربنا أخرجنا من هذه القرية الظالم أهلها واجعل لنا من لدنك وليا واجعل لنا من لدنك نصيرا. الذين آمنوا يقاتلون في سبيل الله والذين كفروا يقاتلون في سبيل الطاغوت فقاتلوا أولياء الشيطان إن كيد الشيطان كان ضعيفا
Question هل يجب ذكر اسم الله على المأكل والمشرب؟
فكلوا مما ذكر اسم الله عليه إن كنتم بآياته مؤمنين. وما لكم ألا تأكلوا مما ذكر اسم الله عليه وقد فصل لكم ما حرم عليكم إلا ما اضطررتم إليه وإن كثيرا ليضلون بأهوائهم بغير علم إن ربك هو أعلم بالمعتدين. وذروا ظاهر الإثم وباطنه إن الذين يكسبون الإثم سيجزون بما كانوا يقترفون. ولا تأكلوا مما لم يذكر اسم الله عليه وإنه لفسق وإن الشياطين ليوحون إلى أوليائهم ليجادلوكم وإن أطعتموهم إنكم لمشركون
إن الذين كفروا ويصدون عن سبيل الله والمسجد الحرام الذي جعلناه للناس سواء العاكف فيه والباد ومن يرد فيه بإلحاد بظلم نذقه من عذاب أليم. وإذ بوأنا لإبراهيم مكان البيت أن لا تشرك بي شيئا وطهر بيتي للطائفين والقائمين والركع السجود. وأذن في الناس بالحج يأتوك رجالا وعلى كل ضامر يأتين من كل فج عميق. ليشهدوا منافع لهم **ويذكروا اسم الله في أيام معلومات على ما رزقهم من بهيمة الأنعام** فكلوا منها وأطعموا البائس الفقير. ثم ليقضوا تفثهم وليوفوا نذورهم وليطوفوا بالبيت العتيق.

Figure 2: Samples Of Questions Combined With Context and Predicted Answers(in Green).

is best for measuring the retrieving performance of the system. pRR is a Reciprocal Rank variation in which the system may retrieve answers that partially match the gold ground-truth answers at different rankings. The suggested approach gives credit to responses at any rank but adds a penalty as the rank of the answers increases (1 top/best to 5 lowest), as shown below.

$$pRR(R) = \frac{m_{r_k}}{k}; k = min\{k|m_{r_k} > 0\},$$

where k denotes the rank position (in our case k = {1 to 5}). m_r is computed as follows:

$$m_r = \max_{a \in A} f_m(r, a)$$

Where $f_m(r, a)$ is an answer-match function that matches a system answer r with the answer a in our case, we utilise the F1 measure applied across questions.

5.2. Experimental Results

In this section, we will present the results of experimenting with various models and architectures trained on different datasets, as well as the impact on performance. We analyzed multiple pre-trained models tailored for Arabic QA and found that the best performing models were MARBERTv2 and ARAELECTRA on multiple datasets. The results of our experiments showed that ARAELECTRA greatly outperformed MARBERTv2 on QRCD, which incentivised our choice to continue using ARAELECTRA in our pipeline.

MARBERTv2: (Abdul-Mageed et al., 2021) was trained using the same data as MARBERT and ARBERT, as well as the AraNews dataset (**?**), but with a longer sequence length of 512 tokens over 40 epochs, totalling 29B tokens. MARBERTv2 didn't obtain competitive results on QRCD, yielding 50.46, 32.11, and 50.5 pRR, exact match, and F1 on the development set, respectively .

Datasets: We fine-tuned ARAELECTRA on three datasets(Ar-TyDi QA, Arabic SQuAD, ARCD, and QRCD) as described in the previous sections. ARAELECTRA fine-tuned in the three datasets yielded the best performance of all the experiments. The comparison between the model's performance was fine-tuned on different datasets and tested on the development set illustrated in Figure 1.

Loss Functions: We experimented with several loss functions to see how they affected performance because the data imbalance issue is more severe for MRC tasks (Rajpurkar et al., 2016), (Bajaj et al., 2016), (Rajpurkar et al., 2018), with a negative-positive ratio of 200-50. Because the MRC task is thought to anticipate the start and end indices of the answer based on the query and context, only two tokens (start and end) are considered positive, while the others are considered negative. We tested Focal-loss (Lin et al., 2017) which is a dynami-

cally scaled cross-entropy loss by applying a modulating term to focus the learning process on low confidence examples (hard misclassified) and down-weight the contribution of the high confidence examples (easy classified). Also, we applied Dice-Loss (Sorensen, 1948), (Dice, 1945), which is an F1- oriented statistic used to gauge the similarity of two sets. We only evaluated those losses in the last phase (fine-tune on QRCD), because we didn't have enough time to experiment with it in the early phases. But, we believe applying those loss functions in the early phases may enhance overall performance (Li et al., 2019). The results of different experiments are shown in Table 2

5.3. Results Analysis

The test set has 238 samples. We analysed the results across all competition participants. For each sample in the test set, we have the minimum, median, and maximum pRR across all submitted runs from all teams. We obtained 73 samples equal to the maximum pRR, 124 samples larger than the median, 52 samples less than the median, 62 samples equal to the median, and 21 samples equal to the minimum pRR from all 238 samples. Figure 2 Samples of the test set questions and their context are illustrated, with the predicted answers highlighted in green.

6. Conclusion and Future Works

In this study, We proposed a method for dealing with QA in the Holy Qur'an. The ARAELECTRA model's efficiency and performance were improved by fine-tuning it on the Ar-TyDi QA, Arabic-SQuAD, and ARCD datasets before fine-tuning it on the competition dataset (QRCD). Furthermore, because the dataset is imbalanced, experimenting with different loss functions to observe how they affect the model performance resulted in a higher model pRR score using the Cross-Entropy loss, which achieved 66.9% on the development set and 54.59% on the test set. In future work, we aim to experiment with alternative loss functions in the early stages of our technique to see whether it improves model performance and efficiency. Moreover, Increasing the dataset size may improve the model's robustness.

Code Availability: The code that was used to conduct the experiments in this work can be found at the following GitHub repository:`https://github.com/Alymostafa/GOF-Qur-an-QA-2022-Shared-Task-Code`

7. Bibliographical References

Abdelmegied, A., Ayman, Y., Eid, A., El-Makky, N., Fathy, A., Khairy, G., Nagi, K., Nabil, M., and Yousri, M. (2017). A modified version of alquans: An arabic language question answering system. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, pages 184–199. Springer.

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.

Abdul-Mageed, M., Elmadany, A., and Nagoudi, E. M. B. (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, August. Association for Computational Linguistics.

Adany, M. A. H. et al. (2017). *An automatic question answering system for the Arabic Quran*. Ph.D. thesis, Sudan University of Science and Technology.

Ali, A. Y. (2000). *The holy Qur'an*. Wordsworth Editions.

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

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

Arabic.). Arabic language.

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., et al. (2016). Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Brini, W., Ellouze, M., Mesfar, S., and Belguith, L. H. (2009). An arabic question-answering system for factoid questions. In *2009 International Conference on Natural Language Processing and Knowledge Engineering*, pages 1–7. IEEE.

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

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

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*.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Ben-

gio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hakkoum, A. and Raghay, S. (2016). Semantic q&a system on the qur'an. *Arabian Journal for Science and Engineering*, 41(12):5205–5214.

Hamdelsayed, M. A. and Atwell, E. (2016a). Using arabic numbers (singular, dual, and plurals) patterns to enhance question answering system results. In *IMAN'2016 4th International Conference on Islamic Applications in Computer Science and Technologies*. Leeds.

Hamdelsayed, M. A. and Atwell, E. (2016b). Islamic applications of automatic question-answering.

Hamdelsayed, M. A., Mohamed, E. M. E., Saeed, M. T. M., Ai, A. M., Mhmoud, E. B. E. M., Mahmoud, M. A., Shamat, A., and Atwell, E. (2017). Islamic application of question answering systems: Comparative study. *Journal of Advanced Computer Science and Technology Research*, 7(1):29–41.

Hamed, S. K. and Aziz, M. J. A. (2016). A question answering system on holy quran translation based on question expansion technique and neural network classification. *J. Comput. Sci.*, 12:169–177.

Hamoud, B. and Atwell, E. (2016). Using an islamic question and answer knowledge base to answer questions about the holy quran. *International Journal on Islamic Applications in Computer Science And Technology*, 4(4):20–29.

Jang, Y., Lee, H., Hwang, S. J., and Shin, J. (2019). Learning what and where to transfer. In *International Conference on Machine Learning*, pages 3030–3039. PMLR.

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*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., and Li, J. (2019). Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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. and Elsayed, T. (2022). Official repository of qur'an qa shared task. `https://gitlab.com/bigirqu/quranqa`.

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)*.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F. V., Avila, S., and Valle, E. (2017). Knowledge transfer for melanoma screening with deep learning. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pages 297–300. IEEE.

Mohammed, F., Nasser, K., and Harb, H. M. (1993). A knowledge based arabic question answering system (aqas). *ACM SIGART Bulletin*, 4(4):21–30.

Mozannar, H., Maamary, E., El Hajal, K., and Hajj, H. (2019). Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy, August. Association for Computational Linguistics.

Nabil, M., Abdelmegied, A., Ayman, Y., Fathy, A., Khairy, G., Yousri, M., El-Makky, N. M., and Nagi, K. (2017). Alquans-an arabic language question answering system. In *KDIR*, pages 144–154.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

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

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Shmeisani, H., Tartir, S., Al-Na'ssaan, A., and Naji, M. (2014). Semantically answering questions from the holy quran. In *International Conference on Islamic Applications in Computer Science And Technology*, pages 1–8.

Silver, D. L., Yang, Q., and Li, L. (2013). Lifelong machine learning systems: Beyond learning algorithms. In *2013 AAAI spring symposium series*.

Sorensen, T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skar.*, 5:1–34.

Thrun, S. and Pratt, L. (2012). *Learning to learn*. Springer Science & Business Media.

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.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.