***I. J. Education and Management Engineering*, 2023, 5, 34-42**
Published Online on October 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijeme.2023.05.03

Introducing Arabic-SQuADv2.0 for Effective Arabic Machine Reading Comprehension

Zeyad Ahmed
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
Email: zeyadfahem1@gmail.com
ORCID iD: https://orcid.org/0009-0009-9562-5583

Mariam Zeyada
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
Email: mariamzeyada73@gmail.com
ORCID iD: https://orcid.org/0009-0005-3880-1668

Youssef Amin
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
Email: yusuf-mahmoud@hotmail.com
ORCID iD: https://orcid.org/0009-0007-6332-604X

Donia Gamal
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
Email: donia.gamaleldin@cis.asu.edu.eg
ORCID iD: https://orcid.org/0000-0002-0740-3086

Hanan Hindy*
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
Email: hanan.hindy@cis.asu.edu.eg
ORCID iD: https://orcid.org/0000-0002-5195-8193
*Corresponding Author

Received: 25 February, 2023; Revised: 12 May, 2023; Accepted: 23 June, 2023; Published: 08 October, 2023

Abstract: Machine Reading Comprehension (MRC), known as the ability of computers to read and understand unstructured text and then answer questions, is still an open research field. MRC is considered one of the most research-demanding sub-tasks in Natural Language Processing (NLP) and Natural Language Understanding (NLU). MRC introduces multiple research challenges. One of these challenges is that the models should be trained to answer all questions and abstain from answering when the answer is not covered in the given context. Another challenge lies in dataset availability. These challenges are amplified for non-Latin-based languages; Arabic as an example. Currently, available Arabic MCR datasets are either small-sized high-quality collections or large-sized low-quality datasets. Additionally, they do not include unanswerable questions. This lack of resources depicts the model as incapable of real-world deployments. To tackle these challenges, this paper proposes a novel large-size high-quality Arabic MRC dataset that includes unanswerable questions, named “Arabic-SQuAD v2.0"”. The dataset consists of 96051 triplets {question, context, answer} in an attempt to help enrich the field of Arabic-MRC. Furthermore, a Machine Learning (ML)-based model is introduced that is capable of effectively solving Arabic MRC-with-unanswerable questions. The results of the proposed model are satisfactory and comparable with Latin-based language models. Furthermore, the results show a significant improvement of the current state-of-the-art Arabic MRC. To be exact, the model scores 71.49 F1-score and 65.12 Exact Match (EM). This proposed dataset and implementation pave the way to further Arabic MRC; aiming to reach a state when MRC models could mimic human text reasoning.

Index Terms: SQuAD, AQAD, Machine Reading Comprehension, Question Answering.

1. Introduction

Machine Reading Comprehension (MRC) is a task introduced to test the degree to which a machine can understand natural languages [1]. This is achieved by asking the machine to answer questions based on a given context. A well-performing MRC model can substantially improve the way in which humans and machines interact with each other. This would impact different research domains including Human-Computer Interaction (HCI) and search engines [2].

However, currently available MRC models do not support Arabic language, and this is because Arabic MRC is an under-researched field compared to English and Latin-based languages [3]. The research and development of the English MRC field and the highly-performing English MRC models do not serve the field of Arabic MRC, because Arabic is a highly inflectional language, so its nature and structure are not similar to that of English language [4].

Given the limited research in the Arabic-MRC alongside the existence of AQAD dataset [5], a recently introduced dataset for Arabic-MRC that is derived from the popular and powerful SQuAD v2.0 dataset, there is a need to develop models that can understand the Arabic language and perform Arabic-MRC-with-unanswerable-questions task effectively. This paper aims to build an improved model for better Arabic language understanding with a competitive performance with respect to standard evaluation metrics (i.e., EM, F1 score).

The main contributions of this paper are as follows:

- Introducing “Arabic-SQuADv2.0” which is a new large benchmark Arabic MRC dataset.
- Introducing MRC model that can effectively solve the problem of Arabic-MRC with unanswerable questions.
- Achieving state-of-the-art performance with the introduced ML Arabic-MRC model.

The rest of this paper is organized as follows; Section II presents the previous literature related to Arabic MRC. Section III discusses the newly introduced dataset “Arabic-SQuADV2.0”. Section IV outlines the research methodology and the different conducted experiments. Section V explains the experimental setup and reports the results. Section VI discusses the analysis of the experiments using different Arabic MRC datasets. Finally, Section VII concludes the work and discusses the future directions.

2. Related Work

Many efforts have been made to solve the problem of Arabic MRC, from creating Arabic MRC datasets and developing models able to understand and answer Arabic questions given a certain context. Hence this section discusses prior research related to the work presented in two subsections; (1) Existing Dataset, and (2) Existing Models.

A. Existing Dataset

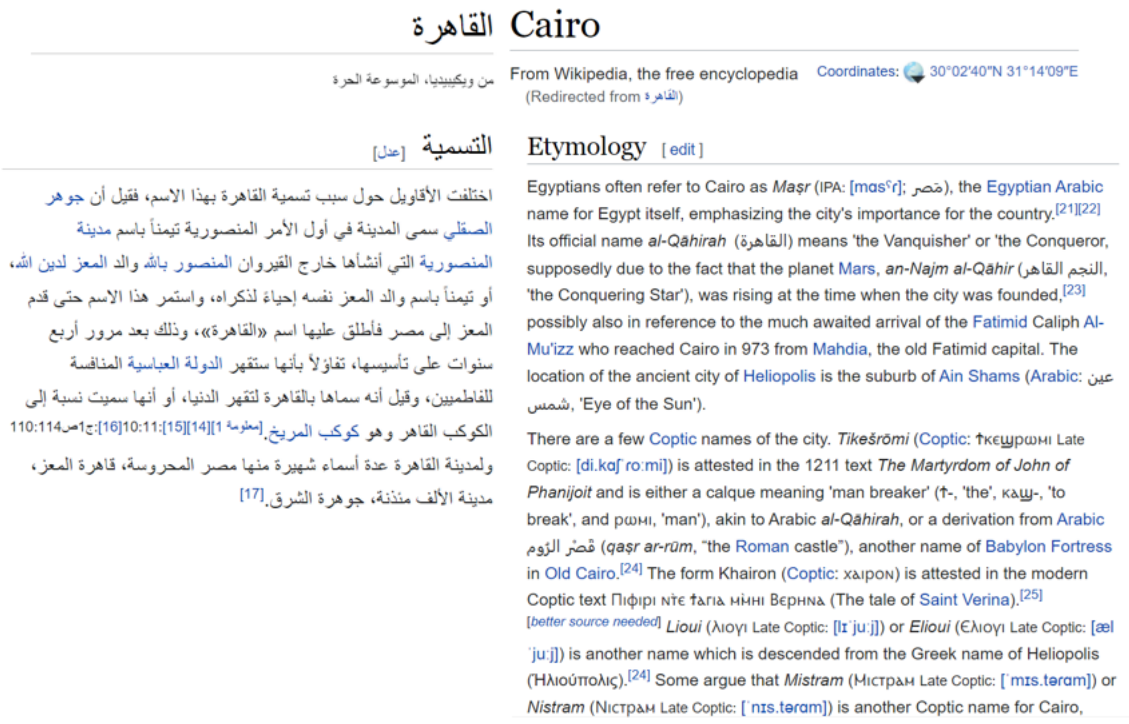
One of the most used English MRC datasets are the MS MARCO[6] and the SQuAD dataset with its different versions [7, 8]. The existing Arabic datasets can be categorized into two categories based on the collection method; (1) Crowd-Workers collectors; the same method purposed by the SQuAD authors [7], and (2) Automated Neural Machine Translation (A-NMT) as of the SQuAD dataset.

An Arabic version of SQuAD 1.1 was introduced using A-NMT for the first 231 articles [9] resulting in a large A-MRC dataset. However, the dataset is represented in low quality due to the poor translation compared to the new NMT models which are based on transformers. Moreover, the dataset contains mislabeled answers due to the corruption of the answer’s span.

In [10], the authors proposed the ARCD dataset based on crowd workers using only 5 articles taken from Arabic Wikipedia. This resulted in a small high-quality data, however, ML models could not benefit from data of this size.

In [5], the AQAD dataset was introduced to tackle two problems; having high-quality large data that can be enough to train an MRC model, and also adding the unanswerable questions. The authors’ method of collecting the dataset was based on finding the Arabic equivalent articles to the ones on SQuAD 2.0 using a similarity model and doing NMT on QAs only without the contexts so the translation is as minimal as possible.

The AQAD method was promising but had faulty assumptions. These include assuming the existence of the exact contexts and answers across different languages. For example, it is not guaranteed that the Wikipedia article for a topic T in Arabic would have the same content as the English one. As a result, the questions asked in the English context could not be answered for the Arabic equivalent article, resulting in a mislabeled dataset. This assumption could be noticed by proofreading the AQAD dataset samples. Figure 1 shows an example of this false assumption. The figure shows a snippet of the Wikipedia article about Cairo in both Arabic and English languages. It is observed that the content varies in size as well as there is a difference in the presented facts about Cairo.



القاهرة

من ويكيبيديا، الموسوعة الحرة

التسمية [عدل]

اختلفت الأقاويل حول سبب تسمية القاهرة بهذا الاسم، فقيل أن جوهر الصقلي سمى المدينة في أول الأمر المنصورية تيمناً باسم مدينة المنصورية التي أنشأها خارج القيروان المنصور بالله والد المعز لدين الله، أو تيمناً باسم والد المعز نفسه إحياءً لذكراه، واستمر هذا الاسم حتى قدم المعز إلى مصر فأطلق عليها اسم «القاهرة»، وذلك بعد مرور أربع سنوات على تأسيسها، تفاؤلاً بأنها ستقهر الدولة العباسية المنافسة للفاطميين، وقيل أنه سماها بالقاهرة لتقهر الدنيا، أو أنها سميت نسبة إلى الكوكب القاهر وهو كوكب المريخ.[معلومة 1][14][15]:10:11[16]:ج1ص110:114 ولمدينة القاهرة عدة أسماء شهيرة منها مصر المحروسة، قاهرة المعز، مدينة الألف مئذنة، جوهرة الشرق.[17]

Cairo

From Wikipedia, the free encyclopedia Coordinates: 30°02′40″N 31°14′09″E
(Redirected from القاهرة)

Etymology [edit]

Egyptians often refer to Cairo as *Maṣr* (IPA: [mɑsˤr]; مَصر), the Egyptian Arabic name for Egypt itself, emphasizing the city's importance for the country.[21][22] Its official name *al-Qāhirah* (القاهرة) means 'the Vanquisher' or 'the Conqueror, supposedly due to the fact that the planet Mars, *an-Najm al-Qāhir* (النجم القاهر, 'the Conquering Star'), was rising at the time when the city was founded,[23] possibly also in reference to the much awaited arrival of the Fatimid Caliph Al-Mu'izz who reached Cairo in 973 from Mahdia, the old Fatimid capital. The location of the ancient city of Heliopolis is the suburb of Ain Shams (Arabic: عين شمس, 'Eye of the Sun').

There are a few Coptic names of the city. *Tikešrōmi* (Coptic: ϯⲕⲉϣⲣⲱⲙⲓ Late Coptic: [di.kɑʃˈro:mi]) is attested in the 1211 text *The Martyrdom of John of Phanijoit* and is either a calque meaning 'man breaker' (ϯ-, 'the', ⲕⲁϣ-, 'to break', and ⲣⲱⲙⲓ, 'man'), akin to Arabic *al-Qāhirah*, or a derivation from Arabic قَصْر الرّوم (*qaṣr ar-rūm*, "the Roman castle"), another name of Babylon Fortress in Old Cairo.[24] The form Khairon (Coptic: ⲭⲁⲓⲣⲟⲛ) is attested in the modern Coptic text Ⲡⲓⲫⲓⲣⲓ ⲛ̀ⲧⲉ ϯⲁⲅⲓⲁ ⲙⲏⲛⲏ Ⲃⲉⲣⲏⲛⲁ (The tale of Saint Verina).[25][*better source needed*] *Lioui* (ⲗⲓⲟⲩⲓ Late Coptic: [lɪˈju:j]) or *Elioui* (Ⲉⲗⲓⲟⲩⲓ Late Coptic: [æl ˈju:j]) is another name which is descended from the Greek name of Heliopolis (Ἡλιούπολις).[24] Some argue that *Mistram* (Ⲙⲓⲥⲧⲣⲁⲙ Late Coptic: [ˈmɪs.tərɑm]) or *Nistram* (Ⲛⲓⲥⲧⲣⲁⲙ Late Coptic: [ˈnɪs.tərɑm]) is another Coptic name for Cairo,

Fig. 1. AQAD English-Arabic Context Content Variation; Left: Arabic, Right: English. Source: Wikipedia

B. *Existing Systems and Models*

This section discusses the different systems trained on A-MRC without handling unanswerable questions. In [10], SOQAL system proposed an open domain question answering model consisting of two parts; document retrieval from Arabic Wikipedia article and MRC to answer questions [11, 12]. The system achieved 61.31 F1-score using BERT transformer [13].

In [14], AraElectra is fine-tuned on the ARCD dataset achieving 71 F1-score. In [5], the authors proposed Multi-Lingual BERT trained on AQAD dataset achieving 37 F1-score. There are not many contributions regarding the models in Arabic MRC due to the lack of good-quality Arabic datasets which urges the need of introducing a new dataset for Arabic MRC.

3. Introducing “ARABIC-SQUADV2.0” Dataset

As mentioned in Section II, there is a lack of a high-quality large dataset for Arabic MCR. Moreover, the existing dataset does not tackle the problem of unanswerable questions except for the AQAD dataset [5]. The AQAD dataset collection methodology produces a large dataset but with a lot of mislabeling as discussed in the previous section. This lack increases the need for a new Arabic dataset for MRC with unanswerable questions following the generation of the well-known SQuADv2.0 [8].

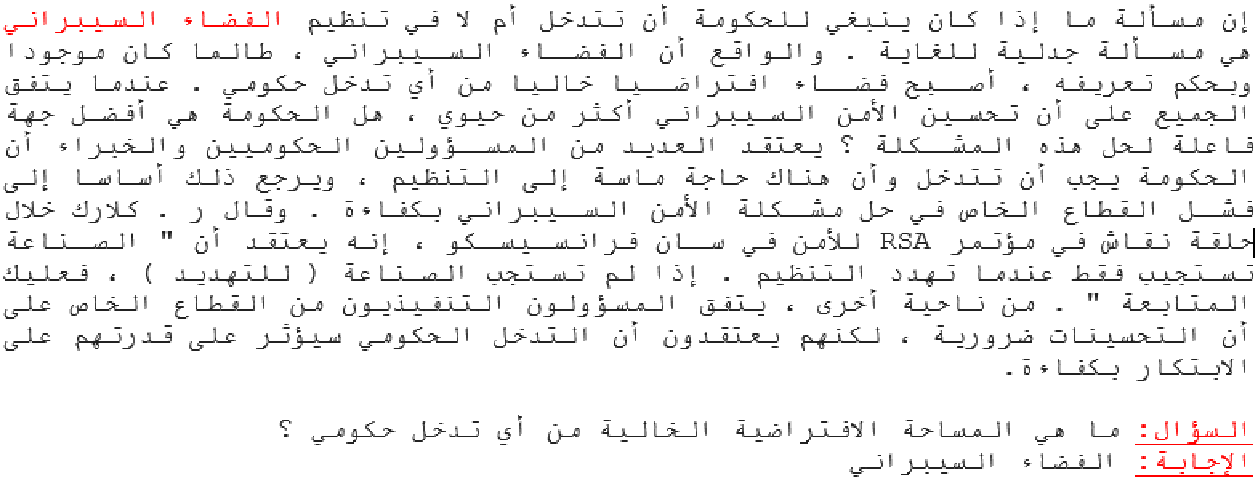
This paper introduces, proposes, and provides a new Arabic MCR dataset “Arabic-SQuADv2.0”. The dataset is created through automated machine translation using state-of-art models on MS Azure using the SQuADv2.0 train dataset. In the case of unanswerable questions, the dataset is not concerned with whether or not the plausible answer is in a contiguous span in the new context as the model should abstain from answering unanswered questions.

The proposed ***“Arabic-SQuADv2.0”*** dataset size is ***96051*** samples. Each sample is presented as a triplet {question,context, answer}. The dataset is split as follows: 80% for training, 10% for validation, and 10% for testing corresponding to the following sizes; 76840 triplets for training, 9605 triplets for validation, and 9606 triplets for testing.

The percentage of answerable questions and unanswerable questions in each dataset split is: 55% answerable questions and 45% unanswerable questions. It has been ensured that the same ratio of answered And unanswered exists in every split to ensure consistency of the data by maintaining the same distribution. Finally, the data follows the SQuADv2.0 [8] format with newly generated IDs for consistency.

The dataset is labeled and publicly available on the HuggingFace at
https://huggingface.co/datasets/ZeyadAhmed/Arabic-SQuADv2.0.

Figure 2 depicts a sample triplet from the newly generated dataset.



إن مسألة ما إذا كان ينبغي للحكومة أن تتدخل أم لا في تنظيم الفضاء السيبراني
هي مسألة جدلية للغاية . والواقع أن الفضاء السيبراني ، طالما كان موجودا
وبحكم تعريفه ، أصبح فضاء افتراضيا خاليا من أي تدخل حكومي . عندما يتفق
الجميع على أن تحسين الأمن السيبراني أكثر من حيوي ، هل الحكومة هي أفضل جهة
فاعلة لحل هذه المشكلة ؟ يعتقد العديد من المسؤولين الحكوميين والخبراء أن
الحكومة يجب أن تتدخل وأن هناك حاجة ماسة إلى التنظيم ، ويرجع ذلك أساسا إلى
فشل القطاع الخاص في حل مشكلة الأمن السيبراني بكفاءة . وقال ر . كلارك خلال
حلقة نقاش في مؤتمر RSA للأمن في سان فرانسيسكو ، إنه يعتقد أن " الصناعة
تستجيب فقط عندما تهدد التنظيم . إذا لم تستجب الصناعة (للتهديد) ، فعليك
المتابعة " . من ناحية أخرى ، يتفق المسؤولون التنفيذيون من القطاع الخاص على
أن التحسينات ضرورية ، لكنهم يعتقدون أن التدخل الحكومي سيؤثر على قدرتهم على
الابتكار بكفاءة.

السؤال: ما هي المساحة الافتراضية الخالية من أي تدخل حكومي ؟
الإجابة: الفضاء السيبراني

Fig. 2. Sample Triplet from the Proposed Arabic-SQuADv2.0 dataset

4. Research Methodology

This research went through multiple experiments to reach the best-optimized model. The process started with experimenting the use of the existing dataset (AQAD dataset) on the baseline model AraBERT [15] with question-answering head on top of the implementation presented by the original BERT authors [13]. Then, another baseline model was used which consumed less training time; named AraElectra [14]. During this phase, different parameters, ranging from weighted losses, data augmentation, coupling, and decoupling methods, were experimented and evaluated on both the AQAD dataset and the proposed Arabic-SQuADv2.0 dataset.

A. *Pre-trained AraBERT\AraELECTRA Baseline Model*

BERT is designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning both left and right context in all layers. As a result, there is no need for a huge architectural change to fine-tune BERT on a specific task. Adding an extra layer would be enough to achieve state-of-the-art [13].

ELECTRA is the same as BERT but instead of hiding input tokens, in ELECTRA, they are subverted by replacing them with plausible alternatives sampled from a small generator network [16]. Then, a discriminative model is trained that determines whether each token in the input is replaced by a generator sample or not rather than a model that predicts the original identities of the tokens. This is an improvement over MLM-based pre-training because the model learns from all input tokens rather than just the ones that were excluded. This yields a better language understanding and by experimentation, it achieves less training time. The training of these Arabic-Transformers is discussed in detail in [14, 15].

B. *Modules on Top of AraBERT\AraELECTRA*

In this section, an overview of the used modules on top of AraBERT\AraELECTRA is presented that will be later used for models discussed in Section C. An explanation of each type of head is discussed followed by the chosen hyper-parameters. These hyper-parameters are recommended by Hugging-Face authors in [17]

Question Answering Head In this experiment, a span classification head is added on top of BERT [13], as proposed by its authors, which acts as extractive question answering. It is formed of a linear layer on top of the hidden state's output with two vectors S"start span" E"end Span" $S, E \in R^H$ Where H is hidden vector size.

Classification Head Sequence Classification head is added on top of BERT [13] for prediction of whether or not the question can be answered with a given context. It is a linear layer on top of the hidden state's output with size 768 followed by a drop-out layer with probability 0.1 and a linear layer at the end of the size of 2 for the two classes "Unanswerable-Answerable"

Coupling Head This module uses the same head as the Question Answering head but with a small extension sigmoid unit for predicting whether or not a question is answerable.

Decoupling Head This one uses two separate AraELECTRA baseline models. One of the models with Question Answering Head and the other with Classification Head. The Classification model is trained using the whole data, while the Span model is trained using answerable questions data only. This is because the classification model's role is to predict if the given question is an answerable one or not.

C. *Proposed Models*

This paper discusses six different proposed models to reach the best performing one. Each model ran through multiple experiments and fine-tuning of its hyper-parameters based on [18, 19]. The optimal hyper-parameters for each model are the only ones mentioned below.

1. **AraBERT with Question Answering Head (AQAD)** This model uses AraBERT as its baseline model with the Question Answering head on top. It is trained on the AQAD dataset for 4 epochs, freezing the first 4 layers and twice weighted loss for answerable questions.
2. **AraELECTRA With Question Answering Head (AQAD**) The same experiment as the previous one but with AraELECTRA baseline instead.
3. **AraELECTRA with Question Answering Head (Data Augmented)** The same experiment as the previous but it is trained on the AQAD-ARCD-Arabic-SQuADv1 to get the benefit of larger dataset without the weighted loss.
4. **AraELECTRA with Coupling Head (AQAD)** This model uses AraELECTRA as its baseline model with the coupling head on top and trained on the AQAD dataset for 4 epochs, freezing the first 4 layers.
5. **AraELECTRA with Decoupling (AQAD)** This model uses AraELECTRA as its baseline model with the Decoupling head on top and trained on the AQAD dataset for 2 epochs, freezing the first 4 layers for both classification and span model.
6. **AraELECTRA with Decoupling (Arabic-SQuADV2.0)** This model uses AraELECTRA as its baseline model with the Decoupling head on top and trained on the Arabic-SQuADv2.0 dataset freezing the first 6 layers for both models, classification model trained for 8 epochs with the twice weighted loss for answerable questions and the span model trained for 4 epochs.

Figure 3 outlines the final pipeline for all the experiments. The experiments passed through different phases including; data preprocessing, QA model training for both classification and span heads until reaching the model evaluation. During the experiments, the performance of each model is evaluated on the validation set. The results are discussed in detail in Section D.

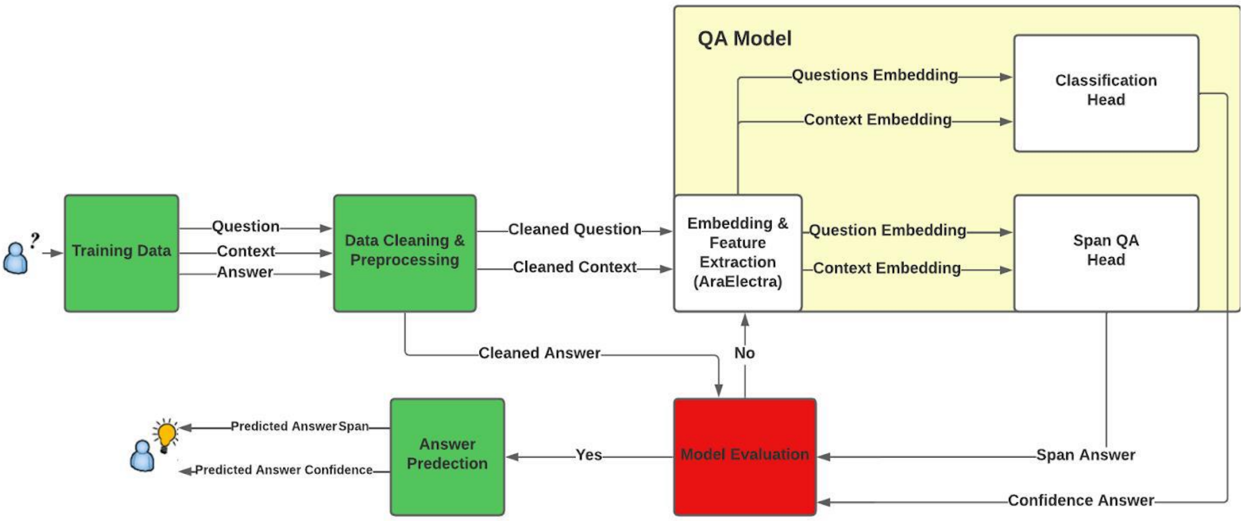


Fig. 3. Proposed model pipeline

5. Experimental Results and Details

A. *Dataset*

As discussed before, the experiments evaluation started with using the AQAD dataset as it was the available benchmark dataset and it looked promising. However, after evaluation and finding that the results are not satisfactory, the dataset was reconsidered. The authors started proofreading the data and found out that it is mislabeled. This is one of the motivations behind the generation of the proposed Arabic-SQuADV2.0 dataset.

B. *Evaluation Metrics*

Two metrics are used to measure the performance of the proposed models; Exact Match (EM) score and F1-score using the official SQuADv2.0 evaluation script [8]. In the case of unanswerable question, the model should abstain from answering and both EM, F1-scores should be 1, otherwise, it should be 0. The evaluation script also reports answer\no-answer EM and F1-scores.

C. Compute and Hyper-parameters

The experiments are evaluated on MS Azure compute with 1xTesla K80 GPU and using PyTorch and HuggingFace. The optimizer is AdamW learning rate= $3e-5$ as the HuggingFace recommendation, with dynamic padding to reduce training time by around 50%.

D. Experimental Results

The proposed models are evaluated on the validation set, and the results are summarized in Table 1 and Table 2.
In total, there are six models, as discussed earlier in Section C. The AraELECTRA-Decoupling-Arabic-SQuADv2.0 achieves the best results with F1-score of 71.49 and EM 65.12 score. This score becomes the new state-of-the-art for Arabic-MRC extractive question answering task for both "Answerable and Unanswerable Questions" as initiated in the SQuAD2.0.

Table 1. F1-Score and EM scores for different proposed models.

ID	Proposed Model Name	Dataset	F1-Score	EM
1	AraBERT with QA Head	AQAD	34.72	19.57
2	AraELECTRA with QA Head	AQAD	36.86	20.35
3	AraELECTRA with QA Head	AQAD-ARCD-ASQuADv1	24.44	3.69
4	AraELECTRA with Coupling Head	AQAD	N/A	N/A
5	AraELECTRA with Decoupling Head	AQAD	38.06	38.05
6	**AraELECTRA with Decoupling Head**	**Arabic-SQuADv2.0**	**71.49**	**65.12**

Table 2. Model 6 Full Evaluation.

Evaluation Metric	Result
Classification Model Accuracy	84.5
EM	**65.12**
F1 Score	**71.49**
Answerable questions (Has answer) EM	56.15
Answerable questions (Has answer) F1	67.80
Non Answerable questions (Has no answer) EM	75.80
Non Answerable questions (Has no answer) F1	75.80

6. Result Analysis

A. Model Performance Analysis

Starting the experiments with AraBERT with QA Head did not achieve good results as shown in Table 1 (row 1). By analysing the SQuAD leaderboard and MRC survey [2], it was found that the literature started to switch to ELECTRA for a lighter and better NLU model.

Switching to AraELECTRA did not achieve any significant result improvement from the previous model as shown in Table 1 (row 2). The last two experiments and the third model's results were not satisfactory which lead to more literature investigation. It was found that most of the results reported in the literature are neglecting the fact that there are unanswerable questions and use plausible answers as ground truth answers [5].

The fact of having unanswerable questions inspired the idea of the coupling head instead of span abstaining. However, by experimenting the model could not converge given the two tasks together; (a) predicting confidence of answerability, and (b) predicting the span of the answer. This is reported as N/A in Table 1 (row 4).

This further influenced the decoupling head which achieved around 38 F1-score. This score is close to the AQAD authors' results in [5]. However, this score can be misleading as it reflects the results of the classification head in the decoupling method which performed well (Table 1(row 5)). The span head in this case scored only 3 F1-score on answerable questions.

Achieving good results in predicting if the question is answerable or not raises questions about why the span model results were not matching the classification model performance. By proofreading the dataset, it was found that owing to the process with which the AQAD dataset [5] was collected, it was a mislabeled dataset as previously discussed in Section A.

The last key finding about the AQAD dataset presented the need for a high-large quality dataset for the Arabic language resulting in introducing The "Arabic-SQuADv2.0". Applying the experiment of the decoupling head on The "Arabic-SQuADv2.0" achieved an F1-score of 71.49 and EM 65.12 score. The classification head achieved 85.09% of accuracy. This becomes the new state-of-the-art for Arabic-MRC and proving that the decoupling method and the presented dataset were effective for the Arabic language MRC. It should be noted that a strong neural system that gets 86 F1-score on SQuAD 1.1 achieves only 66 F1-score on SQuAD 2.0 [8]

7. Conclusion and Future Work

MRC is one of the fields of machine natural language understanding in which machine understanding can be evaluated by answering questions about given paragraphs/contexts. The work on Arabic-MRC is little due to the lack of high-quality publicly available datasets for the Arabic language, and accordingly lack of models; unlike; English language and Latin-based languages.

This paper worked in the direction of building an Aabic-MRC model that is mature enough to tackle answerable and unanswerable questions. The paper started with the experiments using the AQAD benchmark dataset. However, by experimenting and proofreading, it was found that the AQAD had a faulty assumption regarding data contexts [5]. The authors figured out that the AQAD collectors based their dataset on the assumption of one-to-one mapping between English and Arabic contexts, which is not the case.

Therefore, this paper introduces the “Arabic-SQuADv2.0” which is a large-size good-quality dataset with unanswerable questions. Then, the paper introduces a model that proved to sufficiently understand the Arabic language and succeeded to solve MRC-with-unanswerable-questions. The model achieves new state-of-the-art scores for Arabic-MRC-with-unanswerable-questions.

To sum up, this work achieved the objective of enriching the field of Arabic-MRC through the following:

- Introducing a model that proved to sufficiently understand the Arabic language and succeeded to solve MRC-with-unanswerable-questions problem by scoring EM and F1 scores that have not been achieved before in the field of Arabic-MRC-with-unanswerable-questions, as shown in Table 2.
- Contributing to enriching the Arabic-MRC datasets, by introducing Arabic-SQuAD v2.0, and accordingly, paving the path for further development of the whole field of machine understanding of Arabic language.

This work could be continued by extending the “Arabic-SQuADv2.0” dataset with the help of crowd workers. Furthermore, other training methods could be experimented to improve the models’ performance.

References

[1] Danqi Chen. *Neural reading comprehension and beyond*. Stanford University, 2018. https://purl.stanford.edu/gd576xb1833.

[2] Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9:3698, 09 2019. doi: 10.3390/app9183698.

[3] Saeed Salah, Mohammad Nassar, Raid Zaghal, and Osama Hamed. Towards the automatic generation of arabic lexical recognition tests using orthographic and phonological similarity maps. *Journal of King Saud University-Computer and Information Sciences,* 2021.

[4] Mohamed Shaheen and Ahmed Magdy Ezzeldin. Arabic question answering: systems, resources, tools, and future trends. *Arabian Journal for Science and Engineering*, 39(6):4541–4564, 2014.

[5] Adel Atef, Bassam Mattar, Sandra Sherif, Eman Elrefai, and Marwan Torki. AQAD: 17,000+ Arabic questions for machine comprehension of text. pages 1–6, 112020. doi: 10.1109/AICCSA50499.2020.9316526.

[6] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. CoRR, abs/1611.09268, 2016.

[7] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264.

[8] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable questions for SQuAD. *CoRR*, abs/1806.03822, 2018.

[9] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s neural machine translation system: Bridging the gap between human and machine translation, 2016. doi: 10.48550/ARXIV.1609.08144.

[10] Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. Neural Arabic question answering. *In Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4612.

[11] Mariam Biltawi, Arafat Awajan, and Sara Tedmori. Towards building an open-domain corpus for arabic reading comprehension. *In Proceeding of the 35th Int. Bus. Inf. Manage. Assoc.(IBIMA)*, pages 1–27, 04 2020.

[12] Amit Mishra and Sanjay Jain. A survey on question answering systems with classification. *Journal of King Saud University - Computer and Information Sciences*, 28, 11 2015. doi: 10.1016/j.jksuci.2014.10.007.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[14] Wissam Antoun, Fady Baly, and Hazem Hajj. Ara-ELECTRA: Pre-training text discriminators for Arabic language understanding. *In Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics.

[15] Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer-based model for arabic language understanding. *In LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

[16] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555, 2020.

[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[18] Jaejun Lee, Raphael Tang, and Jimmy Lin. What would Elsa do? freezing layers during transformer fine-tuning. *arXiv*, 2019.

[19] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. *arXiv*, 2020.

Authors' Profiles

Zeyad Ahmed is a Software Engineer. Zeyad earned his bachelor's degree (2022) in Computer Science, majoring in Artificial Intelligence, from the Faculty of Computer and Information Sciences at Ain Shams University, Cairo, Egypt. His research interests include Natural Language Processing, Computer Vision, and AI Knowledge Retention Systems.

Mariam Zeyada is a Digital Transformation Consultant and self-employed IGCSE Computer Science Teacher. Mariam received her bachelor's degree (2022) in Computer Science, majoring in Artificial Intelligence, from the Faculty of Computer and Information Sciences at Ain Shams University, Cairo, Egypt. Her research interests include Machine and Deep Learning, Natural Language Processing, IOT, and Robotics.

Youssef Amin is a Computer Science graduate. He received his bachelor's degree (2022) in Computer Science, majoring in Artificial Intelligence, from the Faculty of Computer and Information Sciences at Ain Shams University, Cairo, Egypt. His research interests include Natural Language Processing and Artificial Intelligence.

Donia Gamal is an Assistant Lecturer at the Computer Science department at the Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt. Donia received her bachelor's degree with honors (2013) and master's (2018) degrees in Computer Science from the Faculty of Computer and Information Sciences at Ain Shams University, Cairo, Egypt. Her research interests include Natural Language Processing, Deep Learning, and Artificial Intelligent.

Hanan Hindy is a Lecturer at the Computer Science department at the Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt. Hanan did her Ph.D. at the Division of Cyber-Security at Abertay University, Scotland, UK. Hanan received her bachelor's degree with honors (2012) and master's (2016) degrees in Computer Science from the Faculty of Computer and Information Sciences at Ain Shams University, Cairo, Egypt. Her research interests include Machine and Deep Learning, Intrusion Detection Systems, and Cyber Security.

How to cite this paper: Zeyad Ahmed, Mariam Zeyada, Youssef Amin, Donia Gamal, Hanan Hindy, "Introducing Arabic-SQuADv2.0 for Effective Arabic Machine Reading Comprehension", International Journal of Education and Management Engineering (IJEME), Vol.13, No.5, pp. 34-42, 2023. DOI:10.5815/ijeme.2023.05.03