DTW at Qur'an QA 2022: Utilising Transfer Learning with Transformers for Question Answering in a Low-resource Domain

Damith Premasiri[1], Tharindu Ranasinghe[1], Wajdi Zaghouani[2], Ruslan Mitkov[1]

[1]University of Wolverhampton, UK
[2]Hamad Bin Khalifa University, Qatar
{damith.premasiri, tharindu.ranasinghe, r.mitkov}@wlv.ac.uk
wzaghouani@hbku.edu.qa

Abstract

The task of machine reading comprehension (MRC) is a useful benchmark to evaluate the natural language understanding of machines. It has gained popularity in the natural language processing (NLP) field mainly due to the large number of datasets released for many languages. However, the research in MRC has been understudied in several domains, including religious texts. The goal of the Qur'an QA 2022 shared task is to fill this gap by producing state-of-the-art question answering and reading comprehension research on Qur'an. This paper describes the DTW entry to the Quran QA 2022 shared task. Our methodology uses transfer learning to take advantage of available Arabic MRC data. We further improve the results using various ensemble learning strategies. Our approach provided a partial Reciprocal Rank (pRR) score of 0.49 on the test set, proving its strong performance on the task.

Keywords: Machine Reading Comprehension, Transformers, Transfer Learning, Ensemble Learning, Qur'an

1. Introduction

Machine Reading Comprehension (MRC) is a challenging Natural Language Processing (NLP) application (Baradaran et al., 2022). The concept of MRC is similar to how humans are evaluated in examinations where a person should understand the text and answer questions based on the text. Similarly, the goal of a typical MRC task requires a machine to read a set of text passages and then answer questions about the passages. MRC systems could be widely applied in many NLP systems such as search engines and dialogue systems. Therefore, the NLP community has shown a great interest in MRC tasks over recent years.

The most common way of dealing with MRC tasks is to train a machine learning model on an annotated dataset. Over the years, researchers have experimented with different machine learning approaches ranging from traditional algorithms such as support vector machines (Suzuki et al., 2002; Yen et al., 2013) to embedding based neural approaches such as transformers, with the latter providing state-of-the-art results in many datasets. We discuss them thoroughly in Section 2. However, an annotated dataset is an essential requirement for these machine learning models. Identifying this, the NLP community has developed several datasets in recent years. The most popular MRC dataset is the Stanford Question Answering Dataset (SQuAD), which contains more than 100,000 annotated examples (Rajpurkar et al., 2016). SQuAD dataset has been extended to several languages including Arabic (Mozannar et al., 2019), Dutch (Rouws et al., 2022), Persian (Abadani et al., 2021) and Sinhala (Jayakody et al., 2016). However, MRC datasets have been limited to common domains such as Wikipedia

and MRC in low-resource domains, including religious books, have not been explored widely by the community (Baradaran et al., 2022). Moreover, most researchers focus on a few popular MRC datasets, while most other MRC datasets are not widely known and studied by the community (Zeng et al., 2020).

Qur'an QA 2022 shared task(Malhas et al., 2022) has been organised to address these gaps in MRC research. The goal of the shared task is to trigger state-of-the-art question answering and reading comprehension research on a book that is sacredly held by more than 1.8 billion people across the world. The shared task relies on a recently released dataset of 1,337 question-passage-answer triplets extracted from the holy Qur'an (Malhas and Elsayed, 2020). Despite the novelty, the dataset poses several challenges. Firstly, since the dataset contains texts from Qur'an, modern embedding models would have problems encoding them. Therefore, we experimented with different pre-processing techniques to handle the texts from Qur'an. Secondly, the dataset is small compared to other MRC datasets such as SQuAD (Rajpurkar et al., 2016), and it would be difficult to fine-tune the state-of-the-art neural models. We experiment with different techniques such as transfer learning and ensemble learning to overcome this. We show that state-of-the-art neural models can be applied in smaller MRC datasets utilising the above methods.

We address two research questions in this paper:

RQ1: Do ensemble models provide better results compared to single models?

RQ2: Can other Arabic MRC resources such as SO-QAL (Mozannar et al., 2019) be used to improve the results for Qur'an MRC?

The code of the experiments has been released as an open-source Github project[1]. The project has been released as a Python package[2] and the pre-trained machine learning models are freely available to download in HuggingFace model hub[3]. Furthermore, we have created a docker image of the experiments adhering to the ACL reproducibility criteria[4].

The rest of the paper is structured as follows. Section 2 presents an overview of MRC datasets and machine learning models. Section 3 describes the data we used in the experiments. In Section 4 we explain the experiments carried out. Section 5 discusses the results answering the research questions. Finally, the paper outlines future works and provides conclusions.

2. Related Work

Machine reading comprehension is not newly proposed. The earliest known MRC system dates back to 1977 when (Lehnert, 1977) developed a question answering program called the QUALM. In 1999 (Hirschman et al., 1999) constructed a reading comprehension system exploiting a corpus of 60 development and 60 test stories of 3rd to 6th-grade material. Due to the lack of high-quality MRC datasets and the poor performance of MRC models, this research field was understudied until the early 2010s. However, with the creation of large MRC datasets and with the success of word embedding based neural models in the NLP field, research in MRC has been popular in recent years. We present the related work in MRC in two broad categories; datasets and models.

Datasets In 2013, (Richardson et al., 2013) created the MCTest dataset which contained 500 stories and 2000 questions. This dataset can be considered the first big MRC dataset. A breakthrough in MRC was achieved in 2015 when (Hermann et al., 2015) defined a new dataset generation method that provides large-scale supervised reading comprehension datasets. This was followed by the creation of large scale MRC datasets such as SQuAD(Rajpurkar et al., 2016). Later the SQuAD dataset has been expanded to many languages including Arabic (Mozannar et al., 2019), Dutch (Rouws et al., 2022), French (d'Hoffschmidt et al., 2020) and Russian (Efimov et al., 2020). Furthermore, SQuAD has been extended to low-resource languages such as Persian (Abadani et al.,

2021) and Sinhala (Jayakody et al., 2016) proving that SQuAD has been an important benchmark in MRC research. MRC datasets have been compiled on different domains such as news (Trischler et al., 2017), publications (Dasigi et al., 2021) and natural sciences (Welbl et al., 2017). As far as we know, Qur'an Reading Comprehension Dataset used in this shared task is the first dataset created on religious texts (Malhas and Elsayed, 2020).

Methods Most MRC systems in the early 2000s were rule-based or statistical models (Riloff and Thelen, 2000; Charniak et al., 2000). These models do not provide good results compared to the neural methods introduced in recent years (Baradaran et al., 2022). (Hermann et al., 2015) developed a class of attention based deep neural networks that learn to read real documents and answer complex questions with minimal prior knowledge of language structure. Since 2015, with the emergence of various large scale, supervised datasets, neural network models have shown state-of-the-art results in MRC tasks. The recently introduced transformer models such as BERT (Devlin et al., 2019) have already exceeded human performance over the related MRC benchmark datasets (Zeng et al., 2020). A critical contribution of the SQuAD benchmark is that it provides a system to submit the MRC models and a leaderboard to display the top results[5]. This has enabled the NLP community to keep track of the state-of-the-art MRC systems. Other languages have also followed this approach[6]. However, the NLP community has focused mainly on improving system performance on popular benchmarks such as SQuAD and has not focused on improving results on benchmarks with limited coverage, which we address in this research paper.

3. Data

MRC tasks are usually divided into four categories: cloze style, multiple-choice, span prediction, and free form (Liu et al., 2019). The Qur'an QA 2022 shared task[7] belongs to the span prediction category where the MRC system needs to select the correct beginning and end of the answer text from the context. The event organisers provided the QRCD (Quran Reading Comprehension Dataset), which contained 1,093 tuples of question-passage pairs that are coupled with their extracted answers to constitute 1,337 question-passage-answer triplets. QRCD is a JSON Lines (JSONL) file; each line is a JSON object that comprises a question-passage pair and its answers extracted from the accompanying passage. Figure 1 shows a sample training tu-

---

[1] The Github project is available on `https://github.com/DamithDR/QuestionAnswering`

[2] The Python package is available on `https://pypi.org/project/quesans/`

[3] The pre-trained models are available on `https://huggingface.co/Damith/AraELECTRA-discriminator-SOQAL` and `https://huggingface.co/Damith/AraELECTRA-discriminator-QuranQA`

[4] The docker image is available on `https://hub.docker.com/r/damithpremasiri/question-answering-quran`

[5] SQuAD leaderboard is available on `https://rajpurkar.github.io/SQuAD-explorer/`

[6] Korean MRC leaderboard is available on `https://korquad.github.io/`

[7] More information on the Qur'an QA 2022 shared task is available on `https://sites.google.com/view/quran-qa-2022/`

ple. The distribution of the dataset into training, development and test sets is shown in Table 1.

```
{
  "pq_id": "33:60-62_400",
  "passage": "لئن لم ينته المنافقون والذين في قلوبهم مرض
    والمرجفون في المدينة لنغرينك بهم ثم لا يجاورونك فيها إلا
    قليلا. ملعونين أينما ثقفوا أخذوا وقتلوا تقتيلا. سنة الله
    في الذين خلوا من قبل ولن تجد لسنة الله تبديلا.",
  "surah": 33,
  "verses": "60-62",
  "question": "متى يحل الإسلام دم الشخص؟",
  "answers": [
    {
      "text": "لئن لم ينته المنافقون والذين في قلوبهم مرض
        والمرجفون في المدينة لنغرينك بهم",
      "start_char": 0
    }
  ]
}
```

Figure 1: Sample Json object from the QRCD dataset (Malhas and Elsayed, 2020)

Dataset	%	Q-P Pairs	Q-P-A Triplets
Training	65%	710	861
Development	10%	109	128
Test	25%	274	348
All	100%	1,093	1,337

Table 1: Shared Task Data Composition. Column Q-P Pairs shows the number of Question Passage pairs, Column Q-P-A Triplets shows the number of Question Passage Answer triplets in the dataset

SOQAL contains two Arabic MRC datasets; Arabic Reading Comprehension Dataset (ARCD) (Mozannar et al., 2019), composed of 1,395 questions posed by crowdworkers on Wikipedia articles, and a machine translation of the SQuAD (Mozannar et al., 2019) containing 48,344 questions. SQuAD is widely used as the standard dataset in English MRC tasks, therefore using the machine translation of the same dataset will be helpful for the learning process. Compared to QRCD, SOQAL is a large dataset and both of these datasets belong to the span prediction MRC category. Therefore, they can be used to perform transfer learning which we describe in Section 4.

4. Methodology

With the introduction of BERT (Devlin et al., 2019), transformer models have achieved state-of-the-art results in different NLP applications such as text classification (Ranasinghe and Hettiarachchi, 2020), information extraction (Plum et al., 2022) and event detection (Giorgi et al., 2021). Furthermore, the transformer architectures have shown promising results in SQuAD dataset (Zhang et al., 2021; Zhang et al., 2020; Yamada et al., 2020; Lan et al., 2020). In view of this, we use

transformers as the basis of our methodology. Transformer architectures have been trained on general tasks like language modelling and then can be fine-tuned for MRC tasks. (Devlin et al., 2019). For the MRC task, transformer models take an input of a single sequence that contains the question and paragraph separated by a [SEP] token. Then the model introduces a start vector and an end vector. The probability of each word being the start-word is calculated by taking a dot product between the final embedding of the word and the start vector, followed by a softmax over all the words. The word with the highest probability value is considered. The architecture of transformer-based MRC model is shown in Figure 2.
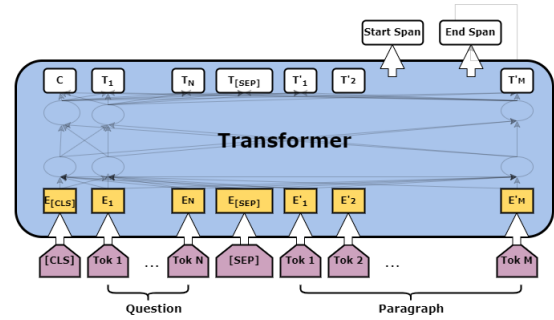

Figure 2: Transformer Architecture for MRC

We experimented with seven popular pre-trained transformer models that supports Arabic; camelbert-mix (Inoue et al., 2021), camelbert-ca (Inoue et al., 2021), mbert-cased (Devlin et al., 2019), mbert-uncased (Devlin et al., 2019), AraELECTRA-generator (Antoun et al., 2021), AraELECTRA-discriminator (Antoun et al., 2021) and AraBERTv2 (Antoun et al., 2020). These models are available in HuggingFace model hub (Wolf et al., 2020). For all the experiments we used a batch-size of eight, Adam optimiser with learning rate $2e-5$, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model, as well as the parameters of the subsequent layers, were updated. The models were trained using only training data. All the models were trained for five epochs. For some of the experiments, we used an Nvidia GeForce RTX 2070 GPU, whilst for others we used a GeForce RTX 3090 GPU. This was purely based on the availability of the hardware and it was not a methodological decision.

We further used following fine-tuning strategies to improve the performance.

4.1. Ensemble Learning

Ensemble learning is a popular technique in machine learning, where different machine learning models contribute to a single solution. As different machine learning algorithms tend to learn differently, the final predictions each one of them provides can be slightly different. However, they have the potential to contribute to the final output with ensemble learning. Usually,

ensemble learning provides better results compared to single models (Sagi and Rokach, 2018).

Transformer models that we used as the base model are prone to the random seed (Hettiarachchi and Ranasinghe, 2020). The same architecture can provide different results for different random seeds (Uyangodage et al., 2021). To avoid the impact of this, we performed self ensemble. We trained the same architecture using five different random seeds and ensembled the output files using Algorithm 1.

Algorithm 1 Ensemble Learning Algorithm for MRC

$R \leftarrow$ *all results files*
$r_i \leftarrow$ *i(th) result file*
$Q \leftarrow$ *all questions*
$q_j \leftarrow$ *j(th) question*
$A \leftarrow$ *all answers*
$a_{i,j} \leftarrow$ *answer for question j in files file i*
$a_j \leftarrow$ *all unique answers for question j in all files*
$a_{j,k} \leftarrow$ *answer k from unique answers for question j*
$a_{i,j,m} \leftarrow$ *answer m from file i to question j*
$answer_{j,k} \leftarrow$ *temporary score*
repeat
    for each $a_{j,k} \in a_j$ **do**
        repeat
            for each $a_{i,j,m} \in r_i$ **do**
                if $a_{i,j,m} = a_{j,k}$ **then**
                    $score_{j,k} \qquad \leftarrow$
$average(score_{a_{i,j,m}}, score_{a_{j,k}})$
                    $answer_{j,k} \leftarrow a_{j,k}, score_{j,k}$
                end if
            end for
        until *all items iterated in R*
        $final\_answers \leftarrow answer_{j,k}$
    end for
until *all unique answers iterated in for question j*
$answers \leftarrow$ **sort(answers)**
repeat
    for each $q_j \in Q$ **do**
        repeat
            for each $answer_{j,k} \in final_answers$
do
                $rank_{j,k} \leftarrow$ *assign rank*
            end for
        until *iterate all answers for question j*
    end for
until *iterate all questions iterated in Q*

4.2. Transfer Learning

One limitation of the QRCD dataset is that training set only contains 710 annotated QnA pairs and as a result transformer models would find it difficult to properly fine-tune their weights in the training process. A common practice to overcome this is to utilise transfer learning. The main idea of transfer learning is that we train a machine learning model on a resource rich setting, save the weights of the model and when we

initialise the training process for a lower resource setting, start with the saved weights from the resource rich setting . Transfer learning has improved results for many NLP tasks such as offensive language identification (Ranasinghe and Zampieri, 2020), machine translation (Nguyen and Chiang, 2017) and named entity recognition (Lee et al., 2018).

For this task, we first trained a transformer-based MRC model on SOQAL dataset which contained more training data compared to the QRCD dataset as mentioned in Section 3. Then when we started the training for QRCD dataset we started from the saved weights from the SOQAL dataset.

5. Results and Discussion

In this section, we report the experiments we conducted and their results. As advised by the task organisers, we used partial Reciprocal Rank (pRR) score to measure the model performance. It is a variant of the traditional Reciprocal Rank evaluation metric that considers partial matching. We also report Exact Match (EM), and F1@1 in the results tables, which are evaluation metrics applied only to the top predicted answer. The EM metric is a binary measure that rewards a system only if the top predicted answer matches exactly one of the gold answers. In comparison, the F1@1 metric measures the token overlap between the top predicted answer and the best matching gold answer. The reported results are for the dev set.

As can be seen in Table 2, camelbert-mix model produced the best results with 0.549 pRR value. This was closely followed by camelbert-ca and AraELECTRA-discriminator. Transformer models built specifically for Arabic generally outperformed multilingual models.

Model	pRR	EM	F1@1
AraELECTRA-discriminator	**0.516**	0.303	0.495
AraELECTRA-generator	0.355	0.339	0.324
camelbert-mix	**0.549**	0.193	0.529
camelbert-ca	**0.535**	0.119	0.516
mbert-cased	0.425	0.321	0.405
mbert-uncased	0.440	0.220	0.424
AraBERTv2	0.501	0.294	0.472

Table 2: Results of different transformer models without ensemble learning or transfer-learning. Column **pRR** shows the partial Reciprocal Rank score, Column **EM** shows results for exact match and Column **F1@1** shows F1@1 score. The top three results are highlighted in Bold.

To answer our **RQ1**, we performed self ensemble learning. Table 3 shows the results of different models with results ensemble. Even though there was a slight improvement in AraELECTRA-discriminator, the overall impact for the results from the ensemble was very low. And we noticed that some of the models had performed less when using ensemble. However, the results were stable compared to the single models. Therefore, we

used self ensemble learning even though it did not contribute to improving the results. With these findings, we answer our **RQ1**, ensemble models do not provide better results compared to single models; however, they provide more consistent results.

Model	pRR	EM	F1@1
AraELECTRA-discriminator	**0.528**	0.321	0.500
AraELECTRA-generator	0.364	0.128	0.335
camelbert-mix	**0.520**	0.303	0.497
camelbert-ca	**0.495**	0.239	0.467
mbert-cased	0.438	0.220	0.417
mbert-uncased	0.424	0.220	0.399
AraBERTv2	0.475	0.239	0.436

Table 3: Results of different transformer models with self ensemble learning. Column **pRR** shows the partial Reciprocal Rank score, Column **EM** shows results for exact match and Column **F1@1** shows F1@1 score. The top three results are highlighted in Bold.

To answer our **RQ2**, we performed transfer learning from SOQAL (Mozannar et al., 2019) to QRCD dataset as mentioned in Section 4. We only conducted the experiments for the best model from the self ensemble setting. As can be seen in the results in Table 4, transfer learning improved the results for AraELECTRA-discriminator. Without transfer learning, AraELECTRA-discriminator scored only 0.528 pRR, while with transfer learning, it provided 0.616 pRR. We did not observe improvements in other transformer models. However, the 0.616 pRR we got with performing transfer learning with AraELECTRA-discriminator was the best result for the dev set. With this, we answer our **RQ2**, other Arabic MRC resources such as SOQAL (Mozannar et al., 2019) can be used to improve the results for Qur'an MRC. We believe that this finding will be important to the researchers working on low-resource MRC datasets.

Model	pRR	EM	F1@1
AraELECTRA-discriminator	0.616	0.394	0.609
camelbert-mix	0.520	0.284	0.494
AraBERTv2	0.430	0.138	0.412

Table 4: Results of different transformer models after transfer learning. Column **pRR** shows the partial Reciprocal Rank score, Column **EM** shows results for exact match and Column **F1@1** shows F1@1 score.

Based on the results of the dev set, we selected three models for the final submission; camelbert-mix with ensemble learning but without transfer learning, camelbert-mix with transfer learning and ensemble learning and AraELECTRA-discriminator with transfer learning and ensemble learning. Table 5 shows the results that the organisers provided on the test set for our submitted models.
AraELECTRA-discriminator performed best in the test set too. The camelbert-mix mode without transfer

Model	TL	EN	pRR	EM	F1@1
camelbert-mix	✗	✓	0.290	0.084	0.258
camelbert-mix	✓	✓	0.408	0.138	0.390
AraELECTRA-discriminator	✓	✓	**0.495**	0.226	0.476

Table 5: Results of different transformer models on the test set. Column **TL** implies whether we performed transfer learning or not and the Column **EN** shows whether we performed ensemble learning. Column **pRR** shows the partial Reciprocal Rank score, Column **EM** shows results for exact match and Column **F1@1** shows F1@1 score.

learning has decreased its performance from 0.549 to 0.290, which is a 47% decrease. However, the models with transfer learning have performed comparatively high, confirming our answer to the **RQ2**.

6. Conclusion

In this paper, we have presented the system submitted by the DTW team to the Qur'an QA 2022 shared task in the 5th Workshop on Open-Source Arabic Corpora and Processing Tools. We have shown that AraELECTRA-discriminator with transfer learning from an Arabic MRC dataset is the most successful transformer model from several transformer models we experimented with. Our best system scored 0.495 pRR in the test set. With our **RQ1**, we showed that transformer models based on self ensemble provided stable results than single models in Qur'an QA task. Revisiting our **RQ2**, we showed that transfer learning could be used to improve the MRC results of the Qur'an. We believe that this finding would pave the way to enhance MRC in many low-resource domains. Our code, software and the pre-trained models have been made available freely to the researchers working on similar problems.

In future work, we would like to explore more to transfer learning. We will be exploring cross-lingual transfer learning with larger English MRC datasets such as SQuAD, as cross-lingual transfer learning has shown splendid results in many NLP tasks (Ranasinghe et al., 2021). Furthermore we will be exploring zero-shot and few-shot learning, which could benefit a multitude of low-resource languages.

7. Acknowledgements

8. References

Abadani, N., Mozafari, J., Fatemi, A., Nematbakhsh, M. A., and Kazemi, A. (2021). Parsquad: Machine translated squad dataset for persian question answering. In *2021 7th International Conference on Web Research (ICWR)*, pages 163–168.

Antoun, W., Baly, F., and Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France, May. European Language Resource Association.

Antoun, W., Baly, F., and Hajj, H. (2021). 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. Association for Computational Linguistics.

Baradaran, R., Ghiasi, R., and Amirkhani, H. (2022). A survey on machine reading comprehension systems. *Natural Language Engineering*, page 1–50.

Charniak, E., Altun, Y., Braz, R. d. S., Garrett, B., Kosmala, M., Moscovich, T., Pang, L., Pyo, C., Sun, Y., Wy, W., Yang, Z., Zeller, S., and Zorn, L. (2000). Reading comprehension programs in a statistical-language-processing class. In *Proceedings of the 2000 ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Sytems - Volume 6*, ANLP/NAACL-ReadingComp '00, page 1–5, USA. Association for Computational Linguistics.

Dasigi, P., Lo, K., Beltagy, I., Cohan, A., Smith, N. A., and Gardner, M. (2021). A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online, June. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

d'Hoffschmidt, M., Belblidia, W., Heinrich, Q., Brendlé, T., and Vidal, M. (2020). FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online, November. Association for Computational Linguistics.

Efimov, P., Chertok, A., Boytsov, L., and Braslavski, P. (2020). Sberquad – russian reading comprehension dataset: Description and analysis. In Avi Arampatzis, et al., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 3–15, Cham. Springer International Publishing.

Giorgi, S., Zavarella, V., Tanev, H., Stefanovitch, N., Hwang, S., Hettiarachchi, H., Ranasinghe, T., Kalyan, V., Tan, P., Tan, S., Andrews, M., Hu, T., Stoehr, N., Re, F. I., Vegh, D., Atzenhofer, D., Curtis, B., and Hürriyetoğlu, A. (2021). Discovering black lives matter events in the United States: Shared task 3, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 218–227, Online, August. Association for Computational Linguistics.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In C. Cortes, et al., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Hettiarachchi, H. and Ranasinghe, T. (2020). InfoMiner at WNUT-2020 task 2: Transformer-based covid-19 informative tweet extraction. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 359–365, Online, November. Association for Computational Linguistics.

Hirschman, L., Light, M., Breck, E., and Burger, J. D. (1999). Deep read: A reading comprehension system. ACL '99, page 325–332, USA. Association for Computational Linguistics.

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

Jayakody, J. A. T. K., Gamlath, T. S. K., Lasantha, W. A. N., Premachandra, K. M. K. P., Nugaliyadde, A., and Mallawarachchi, Y. (2016). "mahoshadha", the sinhala tagged corpus based question answering system. In Suresh Chandra Satapathy et al., editors, *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1*, pages 313–322, Cham. Springer International Publishing.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Lee, J. Y., Dernoncourt, F., and Szolovits, P. (2018). Transfer learning for named-entity recognition with neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May.

European Language Resources Association (ELRA).

Lehnert, W. G. (1977). *The process of question answering.* Yale University.

Liu, S., Zhang, X., Zhang, S., Wang, H., and Zhang, W. (2019). Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9(18).

Malhas, R. and Elsayed, T. (2020). Ayatec: Building a reusable verse-based test collection for arabic question answering on the holy qur'an. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(6), oct.

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)*.

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.

Nguyen, T. Q. and Chiang, D. (2017). Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Plum, A., Ranasinghe, T., Jones, S., Orasan, C., and Mitkov, R. (2022). Biographical: A semi-supervised relation extraction dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain. Association for Computing Machinery.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). 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. Association for Computational Linguistics.

Ranasinghe, T. and Hettiarachchi, H. (2020). BRUMS at SemEval-2020 task 12: Transformer based multilingual offensive language identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1906–1915, Barcelona (online), December. International Committee for Computational Linguistics.

Ranasinghe, T. and Zampieri, M. (2020). Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online, November. Association for Computational Linguistics.

Ranasinghe, T., Orasan, C., and Mitkov, R. (2021).

An exploratory analysis of multilingual word-level quality estimation with cross-lingual transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 434–440, Online, August. Association for Computational Linguistics.

Richardson, M., Burges, C. J., and Renshaw, E. (2013). MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA, October. Association for Computational Linguistics.

Riloff, E. and Thelen, M. (2000). A rule-based question answering system for reading comprehension tests. In *Proceedings of the 2000 ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Sytems - Volume 6*, ANLP/NAACL-ReadingComp '00, page 13–19, USA. Association for Computational Linguistics.

Rouws, N. J., Vakulenko, S., and Katrenko, S. (2022). Dutch squad and ensemble learning for question answering from labour agreements. In Luis A. Leiva, et al., editors, *Artificial Intelligence and Machine Learning*, pages 155–169, Cham. Springer International Publishing.

Sagi, O. and Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1249.

Suzuki, J., Sasaki, Y., and Maeda, E. (2002). SVM answer selection for open-domain question answering. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. (2017). NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August. Association for Computational Linguistics.

Uyangodage, L., Ranasinghe, T., and Hettiarachchi, H. (2021). Transformers to fight the COVID-19 infodemic. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 130–135, Online, June. Association for Computational Linguistics.

Welbl, J., Liu, N. F., and Gardner, M. (2017). Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark, September. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma,

C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). 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. Association for Computational Linguistics.

Yamada, I., Asai, A., Shindo, H., Takeda, H., and Matsumoto, Y. (2020). LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online, November. Association for Computational Linguistics.

Yen, S.-J., Wu, Y.-C., Yang, J.-C., Lee, Y.-S., Lee, C.-J., and Liu, J.-J. (2013). A support vector machine-based context-ranking model for question answering. *Information Sciences*, 224:77–87.

Zeng, C., Li, S., Li, Q., Hu, J., and Hu, J. (2020). A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences*, 10(21).

Zhang, Z., Wu, Y., Zhou, J., Duan, S., Zhao, H., and Wang, R. (2020). Sg-net: Syntax-guided machine reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9636–9643, Apr.

Zhang, Z., Yang, J., and Zhao, H. (2021). Retrospective reader for machine reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14506–14514, May.