Towards Machine Comprehension of Arabic Text

Ahmad Magdy Eid, Nagwa El-Makky and Khaled Nagi

Computer and Systems Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt

Keywords: Arabic Question Answering Systems, Machine Comprehension, Deep Learning, Machine Translation, Post-editing, Semi-supervised Learning.

Abstract: Machine Comprehension (MC) is a novel task of question answering (QA) discipline. MC tests the ability of the machine to read a text and comprehend its meaning. Deep learning in MC manages to build an end-to-end paradigm based on new neural networks to directly compute the deep semantic matching among question, answers, and the corresponding passage. Deep learning gives state-of-the-art performance results for English MC. The MC problem has not been addressed yet for the Arabic language due to the lack of Arabic MC datasets. This paper presents the first Arabic MC dataset that results from the translation of the SQuAD v1.1 dataset and applying a proposed approach that combines partial translation post-editing and semi-supervised learning. We intend to make this dataset publicly available for the research community. Furthermore, we use the resultant dataset to build an end-to-end deep learning Arabic MC models, which showed promising results.

1 INTRODUCTION

Question answering is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automatically answer questions posed by humans in a natural language.

QA systems mainly depend on corpus size availability besides processing power. As computer processing power and corpora increase, not only more researchers became interested in the field but also commercial service suppliers too.

Generally, natural languages are complex and ambiguous, where there are no specific rules or terminology imposed on users. Moreover, natural languages are more prone to human errors, such as grammatical, spelling, or punctuation mistakes. Natural language complexity and ambiguity makes the problem of understanding and answering natural language questions more challenging and requires many steps to account for these challenges.

Recently, many types of research have emerged to solve the question answering problem for domain-specific questions and general ones using deep learning. General QA systems are challenging since they retrieve information from data sources, imposing no restrictions on question types or fields. Deep learning is providing state-of-the-art performance for English question answering systems, thanks to the availability of large and high-quality QA corpora.

Unfortunately, we can not say the same for other languages like Arabic. Despite the significant number of people who use Arabic daily (more than 200 million people) and countries that consider Arabic as their first language (more than 30 countries), there are very few attempts to use deep learning in Arabic question answering. To the best of our knowledge, this is the first work to use deep learning in the Arabic machine comprehension task.

In general, few attempts were made to investigate Arabic QA due to the reasons given below.

1. Arabic morphological difficulties (diacritics, inflection, etc.).
2. Lack of large Arabic question answering corpora.

Arabic has an entirely different orthography based on standard Arabic script going from right to left. Letters in Arabic have different shapes depending on their position in the word, and some characters like ”Alef” may have ”Hamza” above or below. Also, Arabic letters can have a diacritic sign above or below the character, which may change the meaning of the whole word. Since the diacritic sign may affect word’s meaning, non-diacritic Arabic text leaves a vast room for ambiguity, which shows the need for more sophisticated preprocessing for semantically based techniques.

Arabic is also one of the highly derivational and inflectional languages. A word can be broken down

Eid, A., El-Makky, N. and Nagi, K.
Towards Machine Comprehension of Arabic Text.
DOI: 10.5220/0008065402820288
In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019)*, pages 282-288
ISBN: 978-989-758-382-7

into three parts as in 1

$$Word = prefix(es) + lemma + suffix(es) \tag{1}$$

The prefixes can be articles, prepositions or conjunctions; whereas the suffixes are generally objects or personal/possessive anaphora. Both prefixes and suffixes can be combined, and thus, a word can have zero or more affixes. Figure 1 shows an example of the composition of an Arabic word.

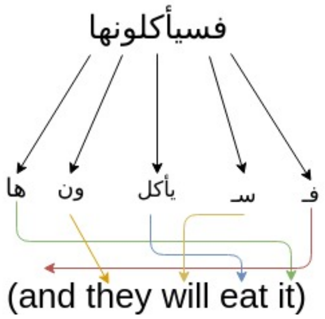

Figure 1: Example Arabic Inflection (Abdelbaki et al., 2011).

A large and high-quality QA corpus is one of the critical factors of the current improvement of the English QA systems based on deep neural networks. As the number of instances in a corpus increases, data-intensive deep learning models can be used to improve performance. In the case of Arabic, massive and high-quality corpora are rare.

The following are some examples of the available Arabic QA corpora. The work in (Aouichat and Guessoum, 2017) proposes a corpus of 2000 factoid question/answer pairs. (Benajiba et al., 2007) translated 200 CLEF and TREC factoid question/ answer pairs to Arabic. Also, (Abouenour, 2011) introduced another dataset by manually translating a collection of 2,264 TREC and CLEF questions into Arabic.

There are no Arabic datasets specially crafted for machine comprehension, which encourages us to propose an Arabic MC corpus. In this paper, we present an Arabic MC corpus built from automatically translating the English Stanford Question Answering dataset SQuAD v1.1 (Rajpurkar et al., 2016) and then applying a proposed approach that combines partial translation post-editing and semi-supervised learning. Using the resultant Arabic MC corpus, we apply an end-to-end MC deep learning model which obtains promising results.

The rest of the paper is organized as follows. A short survey of the related work in the field of QA systems, including Non-Arabic MC systems and Arabic QA systems is given in section 2. In Section 3, we present our new Arabic QA corpus and explain how we constructed this QA corpus. We provide the baseline model architecture in section 4. The system evaluation is presented in section 5. Section 6 concludes the paper and offers some ideas for future work.

2 RELATED WORK

2.1 Non-arabic MC using Deep Learning

Most of the work in MC supports English only. Traditional MC based on question/answer pairs relies on a pipeline of NLP models, which heavily depend on linguistic annotation, and semantic parsing. The rapid growth of MC datasets for the English language made it possible to train large end-to-end neural networks models.

The attention mechanisms give the state-of-the-art performance in MC deep learning models. A variety of English model structures and attention types appeared in literature such as (Cui et al., 2017), (Xiong et al., 2016), (Seo et al., 2016), (Clark and Gardner, 2018).

Typically attention-based deep learning models for MC encode question and passage using embedding techniques and then identify answer based on an attention function. BiDAF (Seo et al., 2016) proposed a co-attention mechanism where the question vector and context vector are used to compute similarity using context-to-query attention and query-to-context attention, which was state-of-the-art at its time. Since that time, many derivational versions appeared such as (Liu et al., 2017), (Wang et al., 2017) and (Wang et al., 2018).

(Wang et al., 2017) introduced the self-attention mechanism to refine the representation by matching the passage against itself, to better capture the global passage information. (Wang et al., 2018) improved results on SQuAD dataset by determining the relationship between question and passage, with a hierarchical strategy which makes answer boundary clear with the refined attention mechanism.

Currently (Devlin et al., 2018) (BERT) model provide the state-of-the-art for English MC task.BERT is a pre-trained deep learning bidirectional representation from unlabeled text by considering both left and right context in all layers.

The work in (Lee et al., 2018) is an example of creating a non-English MC dataset based on SQuAD v1.1. The authors study semi-automated creation of a Korean MC corpus by automatically translated SQuAD and a QA system bootstrapped on a small set of manually annotated Korean question/answer pairs. The authors apply the BiDAF model (Seo et al., 2016) to the created dataset and achieve promising results.

We are going to use BiDAF model (Seo et al., 2016) in this paper as our baseline model as it is one of the high-performance, open source available models.

2.2 Arabic Question Answering Systems

To the best of our knowledge, Arabic MC is not addressed in the literature. Traditional Arabic QA was introduced in the 1990s, but good results were not reported until QARAB (Hammo et al., 2002). QARAB extracted answers of the questions from its knowledge base, which is merely a collection of Al-Raya newspaper published from Qatar. QARAB used sophisticated Natural Language Processing (NLP) techniques such as lexicon based stemming, POS tagging, named entity recognition to identify the top 10 possibly relevant passages from the knowledge base.

(Benajiba et al., 2007) presented ArabiQA which used NLP and information retrieval (IR) techniques. ArabiQA is fully oriented to the modern Arabic language.

A QA system named DefArabicQA (Trigui et al., 2010) was developed to identify exact and accurate definitions about organizations using Web resources. The system was designed using a shallow linguistic analysis but had no language understanding capability. The system identified candidate definitions using a set of manually created lexical patterns categorized these candidate definitions using heuristic rules and ranked them using a statistical approach. The system was tested using 2000 snippets returned by Google search engine, Wikipedia (Arabic version), and a set of 50 organization definition questions.

IDRAAQ (Abouenour et al., 2012) is one of the robust Arabic QA systems which is based on query expansion and passage retrieval. It works on enhancing the quality of retrieved passage based on the given question.

In (Nabil et al., 2017), the authors introduced AlQuAnS QA system. AlQuAns used MADAMIRA (Pasha et al., 2014) to preprocess Arabic text and built a classifier to classify each input question to a specific category where semantic analysis is applied to retrieve relevant documents to the given question. At the final stage, they used an answer extraction module based on crafted patterns.

There are very few attempts to use deep learning in Arabic community QA. For example, in (Romeo et al., 2017), the authors address the question ranking problem in Arabic community QA forums. They use the Farasa toolkit (Abdelali et al., 2016) and use LSTM networks with an attention mechanism to choose the question part to be used in the ranker. Their results show strong performance based on their proposed approach.

3 ARABIC MC CORPUS

To build an end-to-end MC model, a large and high-quality corpus is required for the model. We previously showed that this is not the case for Arabic MC. We decided to build an Arabic MC dataset by translating SQuAD v1.1 from English to Arabic.

The SQuAD dataset is provided in JSON format, where training and development datasets share the same JSON structure, while the testing/evaluation dataset has a different structure. In case of training and development datasets, the SQuAD dataset consists of a set of paragraphs where each paragraph has a context and an array of questions on this paragraph. Each question object consists of the question text and an array of answers, where the answer object contains the answer text and ”answer start index”. The ”answer start index” is the starting index of this answer in the provided context.

The evaluation dataset is an array of JSON objects where each object holds a passage and question text. The SQuAD evaluation dataset is hidden, and only a sample file is available. For evaluating a model, one has to submit his model for evaluation to the dataset website.

While building the Arabic MC dataset, we faced two main challenges:

1. We had to choose a translating tool. We need a translating tool with high accuracy and good speed to process the SQuAD training and development datasets in a considerable time. We used Google Translate since it uses state-of-the-art artificial intelligence techniques.
2. When examining Google translation of SQuAD into Arabic, we faced another challenge. The position of answer span sometimes changes or is lost in translation, which could occur due to several reasons. For example, the number of words in the original and translated passage/ answer can be different.

3.1 Proposed Approach

As mentioned in the previous section, the position of answer span can change after translation which could be due to the difference between the number of original and translated words, the language gap, translation inaccuracy, etc.

Doing an exhaustive post-editing of the translation dataset can be very expensive, given the scarcity of post-editing tools for Arabic translation. Knowing that the translation quality varies over question/ answer pairs, we propose an approach that combines preprocessing, partial post-editing, and semi-supervised learning to create a high-quality Arabic MC dataset. The preprocessing and partial post-editing parts of the proposed method can be outlined in the following steps.

1. We preprocessed the English paragraphs before translation and added special symbols around answers in paragraphs. This technique succeeded in some but not all of the cases.
2. We used Farasa (Abdelali et al., 2016) to tokenize translated text. Also, we used Farasa in stemming for comparing text purpose only.
3. We also faced the problem of translating Named Entities (NEs), which represents a challenge for machine translation. Sometimes, NEs are not translated into Arabic, mistranslated, or translated in different ways. For example (" Super Bowl") was translated into "سوبر بول" and in some other cases to "سوبر السلطانية". We tried to unify the translation of NEs as a kind of post-editing.
4. To overcome the synonyms problem, we used the Arabic WordNet (AWN) (Belalem et al., 2014) in matching translated answer text to the corresponding paragraph text and later in the semi-supervised learning predicted answers evaluation.

After doing the above four steps, we found that around 35% of the translated question/answer pairs satisfy the exact match. These pairs are a part of a seed dataset on which a semi-supervised learning technique is to be performed. To increase the size of the seed dataset, We used human resources to post-edit another 5000 question/answer pairs.

From the previous steps, we get a seed dataset of around 40k question/answer pairs. We use a semi-supervised technique to enrich our dataset using the 40K pairs as our base, by taking the following steps which are illustrated in Figure 3:

- Build a model with the current dataset,
- Use the built model to predict answers for the questions that are not part of the seed dataset.
- Measure the accuracy of the predicted answers using the F1 measure and select the question/pairs with the highest accuracy
- Add question/answer pairs accepted from the predicted answers to our dataset.
- Repeat these steps until there are no more accepted predicted answers.

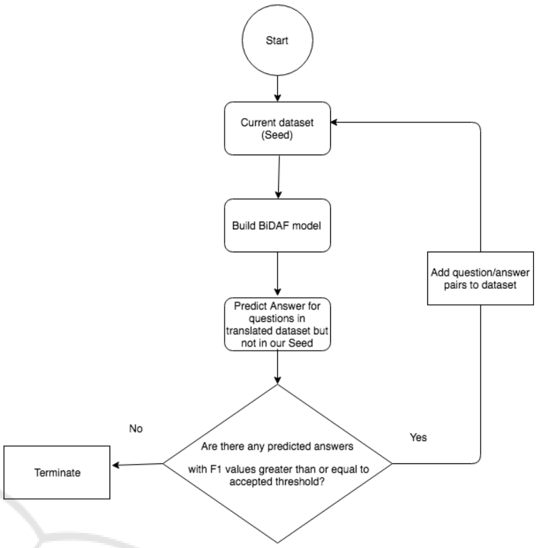
Start
Current dataset (Seed)
Build BiDAF model
Predict Answer for questions in translated dataset but not in our Seed
Are there any predicted answers with F1 values greater than or equal to accepted threshold?
No
Yes
Terminate
Add question/answer pairs to dataset

Figure 2: Semi-supervised learning workflow.

As a result of the above steps, we were able to extract 70k translated question/answer pairs from approximately 100k question/answer pairs found in English SQuAD dataset. We validated a representative sample of the resultant dataset manually. We intend to do a complete dataset validation and then make the dataset publicly available for the research community.

4 SYSTEM ARCHITECTURE

We adopt the bidirectional attention flow network (BiDAF) model (Seo et al., 2016) as our baseline for the Arabic MC task.

Figure 3 illustrate the different components. The model is a hierarchical multi-stage model which consists of six layers:

1. Character Embedding layer: maps each word to a vector space using character-level Convolutional Neural Networks (CNNs).
2. Word Embedding Layer: maps each word to a high-dimensional vector space. Instead of using GLOVE as in the original paper, we used FastText Arabic (Bojanowski et al., 2017).
3. Contextual Embedding Layer: a Bi-directional Long Short-Term Memory Network (LSTM) which model interactions between word vectors generated by previous embedding layers.

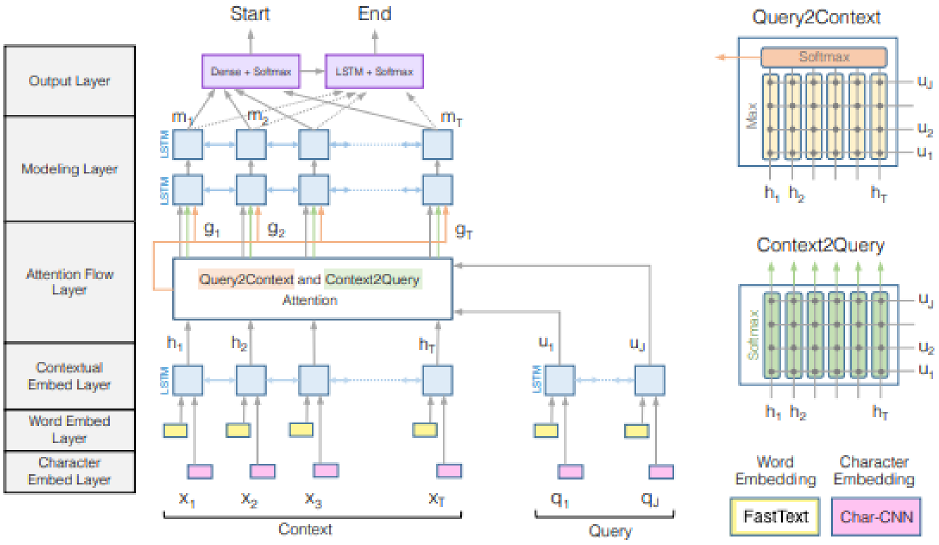

Figure 3: Modifed BiDAF model components (Seo et al., 2016).

4. Attention Flow Layer: a co-attention mechanism is applied in this layer between context-to-query and query-to-context, where the result alongside the contextual embedding layer flows to the next layer.
5. Modeling Layer: this layer also uses bidirectional LSTM to capture the interaction among the context words conditioned on the query.
6. Output Layer: predicts the answer (which is a subphrase from the context) start and end indices in the context paragraph.

5 EXPERIMENTS

In this section, we first present the used dataset. Then we show the baseline model results in comparison to the results of the BiDAF model on the original SQuAD v1.1 dataset and on the Korean dataset obtained by the approach proposed in (Lee et al., 2018). We also study the effect of the size of the resultant training set on model performance. Also, we demonstrate comparison between our baseline model and the current SQuAD (Rajpurkar et al., 2016) state-of-the-art model (Devlin et al., 2018).

5.1 Experimental Settings

We used the obtained Arabic MC dataset to train our model. The test set of SQuAD is hidden. However, the validation set is publicly available, and it is similar to the testing test (both datasets may have more than one answer for a question). So, we divided the set, obtained from translating the SQuAD validation dataset, equally in a random fashion, into a validation set used for hyper-parameter tuning and a test dataset used for evaluation. We finally have a validation set of 3.5K and a test set of 2.8 K.

Our baseline model consists of the same layers of Bidaf (Seo et al., 2016) with some changes to hyper-parameters according to hyper-parameters tuning. We use Adam optimizer, with a mini-batch size of 64 and an initial learning rate of 0.0006, where LSTM layers have 130 hidden units. For CNN character embedding, we use 100 1D filters each with a width of 5. The drop out for the phrase, model and span layers are 0.15, 0.1 and 0.25 units respectively. The model has about 4.8 million parameters. We use Allennlp framework (Gardner et al., 2017) to build our model.

For BERT (Devlin et al., 2018), we used the official implementation provided by the authors. We use the learning rate of 3e-5 and model trained for two epochs using BERT multilingual pre-trained model, which include Arabic.

5.2 Model Evaluation

We use the standard performance metrics for MC: Exact Match (EM) and a softer metric, F1 score, which measures the weighted average of the precision and recall rate at the word level. In case of a question with multiple human answers, standard SQuAD evaluation takes the maximum F1 and EM scores across the provided answers.

Table1 shows a comparison of the performance of the BiDAF model when applied to our Arabic MC dataset, to the original SQuAD dataset and to the

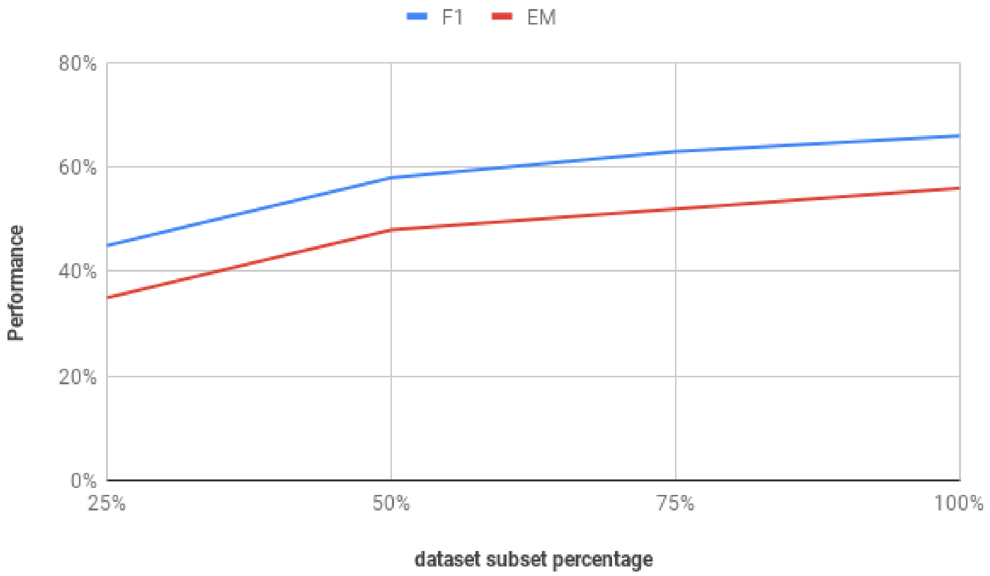

Figure 4: EM and F1 for different subsets of the Arabic MC training set.

dataset obtained by translating SQuAD into Korean language using the approach in (Lee et al., 2018). The results are promising for our Arabic MC dataset in spite of the Arabic language challenges.

Table 1: Comparison of the performance of the BiDAF model on SQuAD, Korean dataset (Lee et al., 2018) and Arabic MC datasets.

	EM	F1
Arabic MC dataset	56	66
Original SQuAD v1.1	67.97	77.32
Korean dataset	50.7	71.5

Table2 shows a comparison of the performance of the baseline model and BERT model trained on our Arabic MC dataset, which shows a significant performance improvement by applying more powerful MC model.

Table 2: Comparison of the performance of the baseline model and BERT model trained on the Arabic MC datasets.

	EM	F1
Baseline model	56	66
BERT model	67.17	77.26

Having a large Arabic MC dataset was a key factor for this promising performance. Figure 4 shows the effect of the size of the Arabic MC training set on the model performance. As we increase the size of the training set by 25% per step, the performance in terms of EM and F1 improves. It reaches its peak (with EM = 56% and F1 = 66%) when we have the full training set.

6 CONCLUSION

In this paper, we propose the first Arabic Machine Comprehension dataset. The dataset consists of 70k question/answer pairs that result from translating SQuAD v1.1 and then applying post-editing and semi-supervised learning to the translated dataset. We intend to make this dataset available to the research community.

We applied a state-of-the-art end-to-end deep learning model to the resultant dataset and obtained promising results, despite the complexity of the Arabic language. To the best of our knowledge, this is the first time to apply an end-to-end deep learning model in Arabic machine comprehension.

As future work, we intend to do a complete dataset validation before making the dataset publicly available for the research community. We will do a more hyper-parameter tuning, and we consider applying more complex deep learning models to our dataset to improve the performance results.

REFERENCES

Abdelali, A., Darwish, K., Durrani, N., and Mubarak, H. (2016). Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16.

Abdelbaki, H., Shaheen, M., and Badawy, O. (2011). Arqa high performance arabic question answering system.

In *Proceedings of Arabic Language Technology International Conference (ALTIC)*.

Abouenour, L. (2011). On the improvement of passage retrieval in arabic question/answering (q/a) systems. In Muñoz, R., Montoyo, A., and Métais, E., editors, *Natural Language Processing and Information Systems*, pages 336–341, Berlin, Heidelberg. Springer Berlin Heidelberg.

Abouenour, L., Bouzoubaa, K., and Rosso, P. (2012). Idraaq: New arabic question answering system based on query expansion and passage retrieval. volume 1178.

Aouichat, A. and Guessoum, A. (2017). Building talaa-afaq, a corpus of arabic factoid question-answers for a question answering system. In Frasincar, F., Ittoo, A., Nguyen, L. M., and Métais, E., editors, *Natural Language Processing and Information Systems*, pages 380–386, Cham. Springer International Publishing.

Belalem, G., Abbache, A., Barigou, F., and Belkredim, F. Z. (2014). The use of arabic wordnet in arabic information retrieval. *Int. J. Inf. Retr. Res.*, 4(3):54–65.

Benajiba, Y., Rosso, P., and Lyhyaoui, A. (2007). Implementation of the arabiqa question answering system's components.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Clark, C. and Gardner, M. (2018). Simple and effective multi-paragraph reading comprehension. In *ACL*.

Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., and Hu, G. (2017). Attention-over-attention neural networks for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–602. Association for Computational Linguistics.

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*.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. S. (2017). Allennlp: A deep semantic natural language processing platform.

Hammo, B., Abu-Salem, H., Lytinen, S., and Evens, M. (2002). Qarab: A: Question answering system to support the arabic language. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*.

Lee, K., Yoon, K., Park, S., and Hwang, S.-w. (2018). Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Liu, R., Hu, J., Wei, W., Yang, Z., and Nyberg, E. (2017). Structural embedding of syntactic trees for machine comprehension. In *EMNLP*.

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*.

Pasha, A., Elbadrashiny, M., Diab, M., Elkholy, A., Eskandar, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 1094–1101.

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. Association for Computational Linguistics.

Romeo, S., Martino, G. D. S., Belinkov, Y., Barrón-Cedeño, A., Eldesouki, M., Darwish, K., Mubarak, H., Glass, J. R., and Moschitti, A. (2017). Language processing and learning models for community question answering in arabic.

Seo, M. J., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603.

Trigui, O., Belguith, L. H., and Rosso, P. (2010). Defarabicqa: Arabic definition question answering system.

Wang, W., Yan, M., and Wu, C. (2018). Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1705–1714. Association for Computational Linguistics.

Wang, W., Yang, N., Wei, F., Chang, B., and Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198. Association for Computational Linguistics.

Xiong, C., Zhong, V., and Socher, R. (2016). Dynamic coattention networks for question answering. *CoRR*, abs/1611.01604.