STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LXVII**, Number 1, 2022
DOI: 10.24193/subbi.2022.1.03

ROMANIAN QUESTION ANSWERING USING TRANSFORMER BASED NEURAL NETWORKS

DIACONU BOGDAN-ALEXANDRU AND LÁZÁR-LŐRINCZ BEÁTA

Abstract. Question answering is the task of predicting answers for questions based on a context paragraph. It has become especially important, as the large amounts of textual data available online requires not only gathering information but also the task of findings specific answers to specific questions. In this work, we present experiments evaluated on the XQuAD-ro question answering dataset that has been recently published based on the translation of the SQuAD dataset into Romanian. Our best-performing model, Romanian fine-tuned BERT, achieves an F1 score of 0.80 and an EM score of 0.73. We show that fine-tuning the model with the addition of the Romanian translation slightly increases the evaluation metrics.

1. Introduction

Question answering (QA) refers to answering questions based on a context paragraph. The answers are of variable length and contain segments of the provided context paragraph. QA is a natural language processing (NLP) task as it entails the automatic understanding of the text.

The importance of question answering has been discussed for more than two decades, as online information has become widespread and users raised the need for not only gathering information from large collections of documents, but also answering specific questions [7].

The task of QA has been addressed with various approaches such as the ones based on Information Retrieval/Information Extraction (IR/IE), restricted domain systems, or rule-based systems [6]. Most IR based systems are returning a list of top-ranked documents or passages as responses to a query. In the following step, the IE system parses the questions and documents yielding the

Received by the editors: 9 December 2021.

2020 *Mathematics Subject Classification.* 68T07, 68T50.

1998 *CR Categories and Descriptors.* I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Language models*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Language parsing and understanding*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Text analysis* .

Key words and phrases. question answering, deep learning, Transformer, Romanian.

interpretation of each word, using several resources like Named Entity Tagging, Template Element, Template relation, Correlated Element, and General Element. The limitation of this system is the fact that it can only answer yes/no type of questions and wh-type of questions (such as when, where, what, etc.). Restricted domain systems, as the name suggests, restrict the domain of questions and the size of the knowledge base. In this system, a question is linguistically inspected by the Heart of Gold architecture. The semantic representations are then interpreted, and a question object that contains a proto query is produced. From this, an instance of a specific database or ontology query is created. An answer object is generated from the result(s) returned by the queried knowledge source. This object forms the foundation for subsequent natural language answer generation. Rule-based systems are an extension of the IR-based QA systems. For each type of question, it generates rules for the semantic classes like who, when, what, where, and why type questions. Rule-Based QA systems initiate parse notations and create training cases and test cases throughout the semantic model.

However, more recently the state of the art results are achieved with neural network models, especially with the fully attention-based Transformer [12] model. This neural architecture is commonly applied to NLP tasks, as it is capable of modeling long-range dependencies. The unidirectionality of the standard language Transformer models limits the options for the architectures used at pre-training. BERT (Bidirectional Encoder Representations from Transformers) is a model proposed by [4] with the purpose of reducing this unidirectionality by using a "masked language model" (MLM) for pre-training. The MLM objective allows the representation to analyze the context both to the left and the right, which enables the pre-training of the deep bidirectional Transformer.

Question answering models were proposed for several languages as a result of the availability of datasets that provide training data for these models. However, for under-resourced languages such as Romanian, to the best of our knowledge, the first baseline model for QA is described in [5]. In our work, we aim to analyze a model for Romanian QA for the newly introduced dataset by [5]. The rest of the paper is structured as follows: Section 2 presents the related work, Section 3 details the method describing the dataset and training architectures, in Section 4 the experiments and results are presented, and conclusions are summarized in Section 5.

2. Related work

Artetxe et al. [1] introduced a new Cross-lingual Question Answering Dataset (XQuAD) for a better understanding of the cross-lingual generalization ability of the models described in the paper. The XQuAD dataset includes the translation of the paragraphs, questions, and answers from the SQuAD v1.1 dataset [11] into ten languages. The authors also showed that neither a shared vocabulary nor joint pre-training is necessary for multilingual models.

In the work of Xue et al. [16], a new token-free, byte-to-byte pre-trained model is proposed. The authors compared the new model, ByT5 with another model that uses the T5 [10] framework, mT5 (the multilingual variant of a T5 architecture), introduced by Xue et al. [17]. The T5 is a unified framework that converts all text-based language problems into a text-to-text format. These models were tested on multiple tasks included in the GLUE [14] or SuperGLUE [13] benchmarks, as well as on a subset of tasks included in the Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark [8]. The XTREME multi-task benchmark is proposed for evaluating multilingual representations through their cross-lingual generalization competencies for 40 languages and 9 tasks building upon existing benchmark datasets such as XQuAD, XNLI [3], TyDi QA [2] and many others. The results favor ByT5 on small models and mT5 on large models for the XQuAD dataset, the best scores are of the mT5 with an F1 of 85.2 and Exact Match (EM) of 71.3.

Adrian et al. [9] developed a system used in the Question Answering competition QA@CLEF for the ResPubliQA track in 2010. For their corpus, the JRC-Acquis and Europarl corpora were indexed, both of them containing documents in XML format. Their question analyzer performed 5 tasks: Noun Phrase chunking and Named Entity extraction, question focus identification, question type inferring, answer type identification, and identification of the keywords of the sentence. The output of these tasks was then used to help the following components: The Index Creation, Information Retrieval, and the Answer Extractor. As for the results, they obtained an accuracy of 55% on the Romanian language, 46% on English, and 30% on French. Our approach is different from the one described in [9] as we use Transformers and a different dataset.

Dumitrescu et al. [5] presents three new datasets: the Semantic Textual Similarity (RO-STS), Question Answering (XQuAD-ro), and Language Modeling (Wiki-ro) datasets. Together with five already existing datasets on the Romanian language, the authors published an open-source benchmark and

leaderboard platform for NLP tasks on Romanian language[1]. The benchmark comprises ten tasks including text categorization by topic, question answering, sentiment analysis, etc. Each task is associated with baseline results, for Question Answering, the mBERT and XLM-R Large models from [1] were used, where the XLM-R Large achieves an F1 score of 83.56.

As the Romanian component of the XQuAD dataset[2] is newly introduced, to the best of our knowledge, notable work has not been published on the XQuAD-ro.

3. Method overview

3.1. **Dataset.** The XQuAD is a benchmark dataset for evaluating cross-lingual question answering performance. It consists of an English subset of 240 paragraphs and 1190 question-answer pairs from the development set of SQuAD v1.1 [11]. The XQuAD also contains the subset's translation in eleven languages: Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, and the most recently added one, Romanian.

All files are in json format following the SQuAD dataset format. Every paragraph consists of one title and a multitude of lists of questions and answers related to the given context. The questions have a unique id and a text. The answers have the text and also the position in the context where the answer starts. Finally, context is represented by a text consisting of multiple sentences related to the topic of the title. Table 1 shows an example in Romanian of how the data is structured in the XQuAD dataset.

3.2. **Training architecture.** Because of time constraints and as well as limited hardware capacity, we decided to use a pre-trained BERT model, fine-tuned on XQuAD like data (before Romanian was added)[3] and further improve it by training on the Romanian language as well. This model has been trained on 104 languages, using 12 attention heads and it has 768 hidden neurons and 110M parameters. The languages have been chosen based on their wikipedia's size. To avoid the overfitting of the model on languages that have less content based on their wikipedia pages, the authors have performed an exponentially smoothed weighting of the data during pre-training data creation. For tokenization, they used a 110k shared WordPiece vocabulary [15]. The word counts were weighted the same way as the data, so low-resource languages were upweighted by a factor. Moreover, the model was fine-tuned on a dataset created by using data augmentation techniques (scraping, neural machine translation, etc.) to obtain more samples from the XQuAD dataset.

[1] `https://lirobenchmark.github.io/` accessed in August 2021

[2] `https://github.com/deepmind/xquad` accessed in august 2021

[3] `https://huggingface.co/mrm8488/bert-multi-cased-finetuned-xquadv1`

Context	Questions	Answer text	Answer start
Apărarea Panthers a cedat doar 308 puncte, clasându-se pe locul șase din ligă, în timp ce au dominat NFL la interceptări, în număr de 24 și s-au putut lăuda cu patru selecții la Pro Bowl.	Câte interceptări a avut apărarea Panthers în sezonul 2015?	24	134
Jucătorul principal al apărării la Pro Bowl, Kawann Short, a condus echipa la numărul de sack-uri cu 11, forțând și trei fumble-uri și recuperând două.	Cine a avut cele mai multe sack-uri în echipa Panthers?	Kawann Short	233

TABLE 1. A sample of the context, questions and answers that can be found in XQuAD-ro.

This increased the size of the dataset from a total of 13,090 question-answer pairs to 58,000 samples. We will refer to this model in the experiments section as XQuAD fine-tuned BERT.

Finally, our contribution to the model consisted in: splitting the dataset, training the model on the new Romanian data with the same number of attention heads and hidden neurons, respectively, and searching for the optimal training arguments. The dataset explained in Section 3.1 has been randomly shuffled and split into 952 question-answer pairs for training (80%), 119 for evaluation (10%), and 119 for testing (10%).

The model trained on the dataset presented above for 10 epochs with a checkpoint on every epoch. Afterward, the best model based on the performance of the F1 score on the evaluation set has been selected. Due to hardware limitations, we used 6 batches to train upon. We used a weight decay of 0.01 and 500 steps of warm-up. In the experiments section, we will refer to this model as Romanian fine-tuned BERT.

4. Experiments and Results

The two most dominant metrics used in question answering tasks are the F1 and EM scores. F1 is calculated by computing the harmonic mean of the precision and recall. Precision is the number of shared words between the prediction and the ground truth divided by the total number of words in the prediction. Recall is the ratio of the number of shared words between the prediction and the ground truth to the total number of words in the ground

truth. Equation 1 represents the formula for calculating the F1 score. EM (Exact Match) is the ratio of the number of predictions, that exactly match the characters of the correct answer to the total number of predictions.

$$F1 = \frac{2}{precision^{-1} + recall^{-1}} \tag{1}$$

We have used the *sklearn*[4] library for calculating F1 and the necessary scores used in its formula. The loading of models, training, and predictions were facilitated by the *transformers* [5] library.

The experiments consisted in computing the accuracy, F1, and EM metrics, the last two being the most important, for both the XQuAD fine-tuned BERT and Romanian fine-tuned BERT models on the test set presented in 3.2. Table 4 presents the results.

Examining the results, there is not much of an improvement from the XQuAD fine-tuned BERT to the Romanian fine-tuned BERT. Most likely, that is caused by the small size of the dataset. Compared to the SQuAD dataset [11] which has 100,000+ question-answer pairs, the XQuAD has only one-hundredth of that amount for one specific language. As consequence, the 952 question-answers pairs are not enough to train the model for more than a few epochs without overfitting.

Model	Dataset	Accuracy	F1	EM
XQuAD fine-tuned BERT	XQuAD-ro	0.80	0.79	0.71
Romanian fine-tuned BERT	XQuAD-ro	0.80	0.80	0.73

Table 2. Models and their computed metrics

5. Conclusions

In this paper we have presented question answering experiments performed on the newly published XQuAD-ro dataset. The first experiment evaluated the model on the Romanian dataset without fine-tuning it on the Romanian language, while the second experiment reports the results after fine-tuning with

[4]https://scikit-learn.org/ accessed in November 2021

[5]https://huggingface.co/transformers/ accessed in November 2021

the additional data. We plan to submit our models to the LiRo benchmark after publication.

Comparing the zero-shot model used with the baseline offered by the XQuAD official repository [6], we got a higher score on the EM metric, 0.71 compared to their best model with 0.69, but we got a lower score on the F1 metric, 0.79 compared to their best model with 0.83.

The low difference in the F1 and EM metrics between the two models is due to the small amount of data that the Romanian language has at its disposal for the Question Answering task. To overcome this barrier, data augmentation techniques could be used to enhance the size of the training set and reduce overfitting. Machine translation could also be used on other datasets for a higher training set. The issue with the latter option is that the accuracy of the model is very dependent on the quality of the translation.

References

[1] M. Artetxe, S. Ruder, and D. Yogatama. On the Cross-lingual Transferability of Monolingual Representations. *arXiv preprint arXiv:1910.11856*, 2019.

[2] J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020.

[3] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov. XNLI: Evaluating Cross-lingual Sentence Representations. *arXiv preprint arXiv:1809.05053*, 2018.

[4] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] S. D. Dumitrescu, P. Rebeja, B. Lorincz, M. Gaman, A. Avram, M. Ilie, A. Pruteanu, A. Stan, L. Rosia, C. Iacobescu, et al. LiRo: Benchmark and leaderboard for Romanian language tasks. 2021.

[6] P. Gupta and V. Gupta. A Survey of Text Question Answering Techniques. *International Journal of Computer Applications*, 53(4), 2012.

[7] L. Hirschman and R. Gaizauskas. Natural Language Question Answering: The View from Here. *natural language engineering*, 7(4):275–300, 2001.

[8] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. In *International Conference on Machine Learning*, pages 4411–4421. PMLR, 2020.

[9] A. Iftene, D. Trandabat, M. Husarciuc, and M. A. Moruz. Question Answering on Romanian, English and French Languages. In *CLEF (notebook papers/LABs/workshops)*, 2010.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[6] https://github.com/deepmind/xquad accessed in November 2021

[11] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. 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, Nov. 2016. Association for Computational Linguistics.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention Is All You Need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[13] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*, 2019.

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

[15] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*, 2016.

[16] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *arXiv preprint arXiv:2105.13626*, 2021.

[17] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

Babeş-Bolyai University, Department of Computer Science, 1 M. Kogălniceanu Street, 400084 Cluj-Napoca, Romania

Email address: beata.lorincz@ubbcluj.ro