UNED at QuALES 2022: Testing the Performance of Transformer-Based Language Models for Spanish Question-Answering*

Alvaro Rodrigo, Anselmo Peñas

NLP & IR group at UNED, Madrid, Spain

Abstract

QuALES at IberLef 2022 is a shared task oriented to extractive Question Answering, where some questions may not have a correct answer inside the given contexts. In our participation in this task, we have tested several state-of-the-art transformer-based models. We have observed that models fine-tuned on the training split obtain better results than other models already fine-tuned on another similar dataset. Besides, we have seen how a combination of different models could outperform individual models. Now, we want to explore the results of other models and perform a deeper analysis.

Keywords

Extractive Question Answering, Spanish Language Models, Voting Schemes

1. Introduction

QuALES at IberLef 2022 is a shared task focused on extractive Question Answering (QA) on Spanish documents [1]. Systems receive pairs of questions and news articles about the Covid-19 domain and must return the shorter span of text answering each question. There are some questions without answers in the given text. For that questions, systems must return an empty answer. So, this format is similar to the task developed in other popular datasets as for example, SQuAD [2]. However, SQuAD provides a larger training dataset with more than 100k pairs, while QuALES provides only 1800 pairs. Thus, the amount of training data is a challenge for this task.

In the last few years, the best results for the QA task have been obtained by transformer-based models or ensembles of such models [3]. This is why, in this work, we have tested the performance of several BERT-based models for this task. More in detail, we have tested different Spanish models, and a multilingual model. Besides, we have tested two combinations of these models. Our objective is to study the real performance of current state-of-the-art models that are available for any user.

Our results are quite lower than the best scores obtained in the task. We think that current models might not be used with just fine-tuning when dealing with a difficult task in a new dataset. So, we must perform more experiments and deeper analysis to obtain meaningful

IberLEF 2022, September 2022, A Coruña, Spain.

alvarory@lsi.uned.es (A. Rodrigo); anselmo@lsi.uned.es (A. Peñas)

0000-0002-6331-4117 (A. Rodrigo); 0000-0002-7867-0149 (A. Peñas)

CEUR Workshop Proceedings (CEUR-WS.org)

conclusions.

The rest of this paper is structured as follows: Section 2 describes the approaches used in this work. Section 3 gives the details of each run submitted to the task, while we discuss the results in Section 4. Finally, we give the conclusions and future work in Section 5.

2. Approaches

Although we have employed similar models, we have tested three main approaches for this task. These approaches depend on the way each model is trained and how the final output is obtained. We give details of each approach in the following subsections.

2.1. Fine-tuning on the Training Set

The first approach is based on fine-tuning different models on the dataset provided by the organizers. We have fine-tuned each model for 10 epochs on the training set, given that we detected a loss in performance when fine-tuning for 11, or more, epochs. Our training dataset contains both training and development splits of the task, with a total number of 1800 question-answer pairs.

The models tested in this approach are:

- PlanTL-GOB-ES-roberta-bne [4]: these are RoBERTa models trained with data from the National Library of Spain (BNE). We have tested with both base and large models. We wanted to test the performance of this Spanish model, comparing the results of base and large versions.
- bert-base-spanish-wwm-cased [5]: this is a Spanish BERT-base model called BETO and trained on a large Spanish corpus. We wanted to test another Spanish model and compare it with the Spanish models described above.

We wanted to test a multilingual model, but we could not do it for time reasons.

2.2. Models Trained on other Datasets

This approach is based on predicting answers by using models trained on other datasets. More in detail, we have taken models already fine-tuned on other datasets and made predictions on the test split provided in this task.

With this approach, we wanted to study the transferability among different datasets. Since this is a task with a new, and small, dataset, we desire to study if the task could be tackled by reusing other systems trained for the same task.

The models included in this approach are:

- PlanTL-GOB-ES-roberta-bne-sqac: these are the RoBERTa models described in Section 2.1, which have been fine-tuned on the Spanish Question Answering Corpus (SQAC) dataset. This dataset contains 18k question-answer pairs. We have tested both base and large models, which are provided to the general public in a version fine-tuned on SQAC. Unfortunately, the output given by these models could not be read by the submission

system. Thus, we could not send any runs from these models and we only used them for the third approach.
- bert-multi-cased-finetuned-xquadv1: this is a multilingual BERT model fine-tuned for QA on the XQuAD dataset [6], which is a cross-lingual QA dataset. Our objective in using this model was to test a multilingual model trained on a multilingual dataset.

2.3. Combined Models

The third approach is based on combining different models for returning a single output. The objective is to improve the performance of individual systems by rewarding the outputs more repeated across the different outputs.

We have implemented two voting schemes, where we return the most predicted answer for each answer. The two voting schemes are:

- Shortest answer: this voting scheme rewards shortest answers. That is, given two answers a_1 and a_2, where a_1 is contained in a_2, a_1 receives two votes while a_2 receives one vote. Here we follow the evaluation criteria of the task, where shortest answers are prefer.
- Longest answer: this voting scheme rewards longest answers. Then, given two answers a_1 and a_2, where a_1 is contained in a_2, a_2 receives two votes while a_1 receives one vote. We have tested this scheme to study the differences in results with respect to the other scheme.

These two schemes are applied over the output of all the models described in Section 2.1 and 2.2, including the output of the two *PlanTL-GOB-ES-roberta-bne-sqac* that could not be uploaded to the submission system. Thus, each voting scheme combines the output of 6 systems.

3. Submitted Runs

We have submitted the following 6 runs:

- Run 1 (roberta-base-bne): RoBERTa base model fine-tuned over the training and development collection of the task.
- Run 2 (roberta-large-bne): RoBERTa large model fine-tuned over the training and development collection of the task.
- Run 3 (beto): BETO base model fine-tuned over the training and development collection of the task.
- Run 4 (multibert-xquadv1): multilingual-BERT fine-tuned on the XQuAD dataset.
- Run 5 (voting short): voting scheme where shortest answers are rewarded over longest ones.
- Run 6 (voting long): voting scheme where the longest answers are rewarded over the shortest ones.

Thus, we have tested 3 runs based on fine-tuning models on the training dataset (runs 1, 2 and 3 described in Section 2.1), 1 run based on a model fine-tuned on another dataset (run

4 described in Section 2.2, since the other two possible runs of this approach did not work with the submission system) and 2 runs combining different models (runs 5 and 6 described in Section 2.3).

4. Analysis of Results

Systems in this task are evaluated using two metrics. The first metric is Exact Match, which is also used in other datasets like SQuAD [7]. The second metric is the macro-average F1 score of word overlap between systems' output and answers in the gold standard.

We show the results of our runs in Table 1. We have highlighted the best result for each measure. The Table also contains the best results for each measure achieved by other participants.

run	Exact Match	macro-average F1
Best system	0.5349	0.7282
run1	0.3043	0.3976
run2	0.2714	0.3385
run3	**0.3175**	0.4002
run4	0.2622	0.3745
run5	0.3136	**0.4293**
run6	0.2938	0.3962

Table 1
Results of the 6 runs submitted to the task and scores of the best performing system among the other participants. The best score for each measure, for our runs, is highlighted.

Firstly, we can see that our results are quite far from the results of the best system. Concerning our results, the best exact match score is obtained by run 3, which is a BETO model fine-tuned on the training split. The best f1 score is obtained by run 5, which is a voting scheme that rewards short answers. Anyway, most of the systems have obtained similar scores.

When comparing our systems, we can see that run 2, which is based on a RoBERTa large model, obtains worst results than runs 1 and 3, which are based on a RoBERTa and a BERT base models respectively. We think this result could be because we did not use enough data for training a large model. Thus, the large model could have overfit.

Run 4 obtains the worst result for the exact match and the second worse for macro-average f1. Although we think these scores can be a consequence of training the model with another dataset, we want to explore also the fact of using a multilingual model for this task.

Regarding the voting-based systems (runs 5 and 6), they have obtained quite good results. In fact, run 5 (voting with short answers) scores the best for f1 and the second one for the exact match. These results show that it is important to combine different systems in QA. Thus, we can take advantage of the main features of each system. Moreover, it is important to reward short answers instead long responses.

5. Conclusions and Future Work

In this paper, we have reported our participation at QuALES in the framework of IberLEF 2022. QuALES is an extractive Question Answering task where we have tested different Spanish BERT-base models.

Our results are far from the best ones in the task. However, we have observed how current state-of-the-art models behave in this task. We have seen that models fine-tuned on the training dataset obtain the best scores than models already fine-tuned on another dataset. Besides, a combination of different approaches can give the best results in this task.

Future work is oriented to a deeper analysis of current results, as well as testing more state-of-the-art models.

Acknowledgments

This work has been partially funded by the Spanish Research Agency (Agencia Estatal de Investigación), DeepInfo project PID2021-127777OB-C22 (MCIU/AEI/FEDER,UE) and the projects given to the *ISL: Intelligent Systems for Learning* group (GID2016-39) in the PID 20/21 and 21/22 calls.

References

[1] A. Rosá, L. Chiruzzo, L. Bouza, A. Dragonetti, S. Castro, M. Etcheverry, S. Góngora, S. Goycoechea, J. Machado, G. Moncecchi, J. J. Prada, D. Wonsever, Overview of QuALES at IberLEF 2022: Question Answering Learning from Examples in Spanish, Procesamiento del Lenguaje Natural 69 (2022).

[2] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for SQuAD, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 784–789. URL: https://aclanthology.org/P18-2124. doi:10.18653/v1/P18-2124.

[3] O. Ram, Y. Kirstain, J. Berant, A. Globerson, O. Levy, Few-shot question answering by pretraining span selection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3066–3079. URL: https://aclanthology.org/2021.acl-long.239. doi:10.18653/v1/2021.acl-long.239.

[4] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60.

[5] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[6] M. Artetxe, S. Ruder, D. Yogatama, On the cross-lingual transferability of monolingual representations, in: Proceedings of the 58th Annual Meeting of the Association for Computational

Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4623–4637. URL: https://aclanthology.org/2020.acl-main.421. doi:10.18653/v1/2020.acl-main.421.

[7] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392. URL: https://aclanthology.org/D16-1264. doi:10.18653/v1/D16-1264.