

Study on Text Comprehension and MCQA in Spanish

Pablo Baggetto^{1,2}, Sofía Ramos³, Joan García³ and José Ramón Navarro^{1,3}

Servicios de Análisis de Datos Avanzados (SADA), Instituto Tecnológico de Informática (ITI). Camino de Vera S/n, Valencia, Spain

Abstract

In this work we explore different approaches for the text comprehension and MCQA task in *ReCoRES: Reading Comprehension and Reasoning Explanation for Spanish*, part of IberLef 2022. Specifically, we explore the use of encoder models, generative models, clue generation systems and dataset expansion. In our experiments, the best model was a pretrained MT5 model finetuned on an expanded multilingual dataset. With this approach we obtained an accuracy in the test set of 72.54% which gave us the second position in the competition only 3.37% behind the first place.

Keywords

MCQA, MT5, Transformers

1. Introduction

This project has been carried out for the participation in the IberLEF 2022 competition. Specifically for subtask 1 of the task *ReCoRES: Reading Comprehension and Reasoning Explanation for Spanish* [1]. This consisted in answering reading comprehension questions by choosing among five options (indicated by letters A to E) the correct answer.

Our proposal investigates different approaches found in the literature applied to the English language and tries to obtain the best solution adapted to Spanish taking into account the lack of resources.

For this task we compared three different approaches to the problem. Using encoder models, using generative encoder-decoder models and using a generation of clues system with an encoder-decoder to help choosing the correct option. Furthermore, we tested augmenting the data using auxiliary datasets in other languages and how the difference of the language affects the performance of the models.

¹Corresponding author.


²This author contributed the most.


³These authors contributed equally.

IberLEF 2022, September 2022, A Coruña, Spain.

✉ pbaggetto@iti.es (P. Baggetto); sramos@iti.es (S. Ramos); joangarcia@iti.es (J. García); jonacer@iti.es (J. R. Navarro)

ORCID 0000-0002-5119-3282 (J. García); 0000-0002-6692-5941 (J. R. Navarro)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. Data processing and analysis

2.1. Processing

The competition data were given in 459 YAML files. For each file there was a single text and one or more questions. For each question the text of the question itself, the available options, the correct answer and the reason why it was chosen were provided (Figure 1).

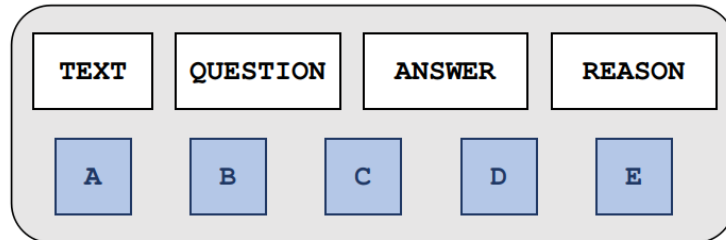


Figure 1: Structure of a datum

To create the data table, we broke the data down by question. The resulting table contains 1,818 rows (one per question). After analyzing the table, 7 rows were removed due to mistakes that made it not processable.

Data was divided in a training set and an evaluation set. As the test data of the competition was unlabeled, we evaluated every model with the evaluation set.

2.2. Analysis

Some of the questions in the text did not contain option 'E' and others lacked an explanation of the correct answer. These problems were solved by imputing an empty string where necessary.

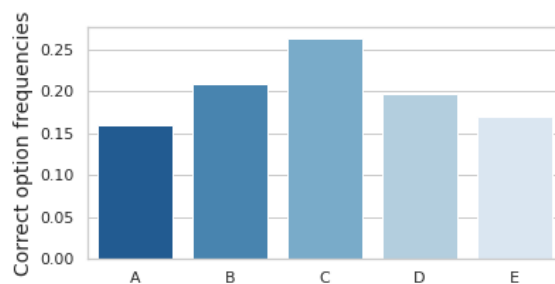


Figure 2: Frequency of the options.

Another problem with the data is the lack of a uniform frequency distribution among the data. There are many more C's than A's as can be seen in Figure 2. This will put the baseline performance of always choosing C around 25%.

Regarding the length of the texts, most of them do not exceed 500 words (see Figure 3). This is very important since many models support up to 512 tokens. For this reason, it was decided not to explore information retrieval methods for reducing the text to a relevant paragraph of the appropriate size.

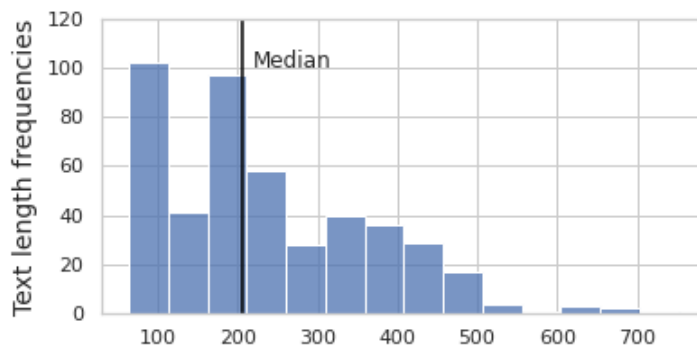


Figure 3: Frequencies of the text length in words.

3. Encoder models

Models that use only the encoder of the transformer architecture, namely BERT [2] and its derivatives, are currently state of the art in benchmarks such as RACE [3].

In order to use these models for this task, for each option of each question, data was divided in two parts. The first part contained the text and the second part contained the question with the option appended to it. Both sections were tokenized together and given to the model.

The models used were provided with the multiple choice head available in all BERT models in Huggingface Transformers [4] with the `BertForMultipleChoice` class. This gives the model a linear layer and a softmax on top of the original backbone. This setup makes the model capable of determining which of the question-answer pairs fits better with the context provided, and thus is the correct answer.

Model	Accuracy(%)	F1	Precision	Recall
BERT ES	45.5	0.457	0.461	0.455
distilBERT Multilingual	27.6	0.279	0.291	0.276
RoBERTa ES	40.7	0.414	0.431	0.407
RoBERTa ES SQAC	45.8	0.459	0.465	0.458
distilBERT ES SQUAD	45.5	0.460	0.467	0.455
BERT ES SQUAD2	45.8	0.464	0.477	0.458
RoBERTa ES SQUAD2	47.9	0.483	0.497	0.479

Table 1

Results obtained with the encoder models in different setups

Different models of the BERT family were tested in order to determine which one worked

better for the task. Results can be found in the Table 1. The model with the better score was the RoBERTa [5] model finetuned using the SQUAD 2.0 translated to Spanish[6] dataset.

4. Generative models

In this section we wanted to explore the capabilities of Encoder-Decoder models dedicated to generation of text. The model we decided to test was the MT5[7] as it is one of the most powerful multilingual generative models freely available. The main issue with this model is that it has only been trained with unsupervised data. For this reason it is untrained for the task of conditional generation, as the task requires labeled data.

To explore these models we trained 3 different versions of the MT5 using the maximum possible batch size in our GPU (NVIDIA A100 with 40GB). For each of them, we performed the training on two different setups for representing the correct solution: one giving them only the letter of the correct answer (L) and the other giving them the letter and the correct option text(L+O). For testing, we took the first character generated by the model as their answer. The results are shown in Table 2.

Model	Batch size	Accuracy(%)
MT5-base(L)	8	27
MT5-base(L+O)	8	32
MT5-small(L)	16	0
MT5-small(L+O)	16	0

Table 2

Results using generative models.

This results are much lower than those obtained with the encoder models because the MT5 model needs more training data in order to learn to generate a good conditional answer as it was not trained for the task. It also must be remarked that the MT5-small model has a 0% accuracy because it didn't generate the desired output so the metric could not be calculated.

As the results suggest that the version with the labels containing the letter and the correct option perform slightly better than the versions with only the letter, it was decided to follow this label format for all the generation models from this moment.

5. Clue Generation

For our third approach we tried a modification of the GenMC model[8]. This model was recently developed with a new approach that used encoder-decoder models to generate a clue and use it to enhance a reader for MCQA.

The MCQA model proposed by the authors of the paper used the T5 model as a pretrained base. As this model is only trained in English, we had to adapt it to the MT5 model. However, MT5, unlike T5, has not been trained with supervised data. For this reason, it is untrained for conditional generation as explained in Section 4.

As expected, the model generated clues that did not make sense with the question as it was very under-trained. As a consequence, the clues were detrimental for the model resulting in a performance of a 17.8% accuracy.

6. Data augmentation with RACE

As it was obvious that more data was required in order to obtain better results, we decided to add to our training set the RACE dataset. RACE is a huge dataset with over 70,000 questions that are very similar to our dataset. However, it has few but important differences that need to be addressed before using it.

The first issue is that the questions have four options instead of five. This was solved by adding a blank option and randomly shuffle the options in order to make it appear in every options with similar frequencies. The second problem, and the most important, is that the dataset is in English. In this section we wanted to explore the importance of the language of the texts and if it is beneficial for the task. For this reason we trained the best configurations of our three approaches with this new dataset. The only model that was changed was the RoBERTa trained in Spanish SQUAD 2.0 for the multilingual BERT as we thought it would be necessary for the model to have understanding of the English language.

Model	Batch size	Accuracy(%)
BERT Multilingual	8	27
MT5 base	4	49.7
GenMC	4	17.8

Table 3
Results using dataset expansion.

Results in Table 3 show multilingual BERT performed worse than its model without the RACE dataset. We think this is due to the fact that it is overfitted for the understanding of English, given the huge size of the RACE dataset in comparison to ReCoRES, and therefore it struggles in understanding the Spanish texts.

On the contrary, the MT5 increased its performance making it the best model trained up to this moment. This is probably because this model has a good interlanguage understanding that helped it increase its performance of Spanish text comprehension even if it was trained in other languages.

Finally, the GenMC model obtained the same accuracy than without the RACE dataset. However, the predictions were better but they were unhelpful to provide a correct answer. For this reason the further study of this approach was discontinued.

7. Data augmentation with translated RACE

In this section we decided to investigate how translating the RACE dataset would affect our models. To perform the translation the Google Translator API¹ using the Deep-Translator

¹<https://cloud.google.com/translate/docs/apis?hl=es-419>

library² was used.

As the translations were very time consuming and we had the deadline of the end of the competition, we could only translate 25% of the dataset. Nonetheless, this quantity is approximately 10 times the size of the competition dataset. For this reason, we conclude that this quantity was enough for our exploration.

Model	Batch size	Accuracy(%)
RoBERTa ES SQUAD2	8	48.5
MT5 base	4	50

Table 4
Results using the translated RACE.

Table 4 shows that the RoBERTa ES SQUAD2 improves significantly compared with the untranslated version. This confirms our suppositions that these kind of models are very language sensitive.

Regarding the MT5, the model presents very similar results to the ones obtained with the original RACE dataset. As the quality of the translations has worsen and the training dataset reduced to a quarter of the original RACE, we hypothesize that the translation slightly benefits the model but it is not worth the required amount of time and computation.

8. Data augmentation with several datasets

As we deduced that the MT5 model could be trained with texts and questions in other languages and benefit from them as the increase of data also increases its performance, we thought it would be a good idea to combine different datasets and train the model with them.

We used these datasets to expand our original one:

- RACE[3]: a multi-choice reading comprehension dataset, collected from middle and high school English examinations in China.
- RACE-C[9]: a multi-choice reading comprehension dataset, collected from college English examinations in China.
- DREAM[10]: is a multiple-choice Dialogue-based READING comprehension examination dataset.
- RECLOR[11]: a dataset extracted from logical reasoning questions of standardized graduate admission examinations.
- C3[12]: multiple-Choice Chinese machine reading Comprehension dataset.

With all those datasets we have a combined dataset of more than 148k questions.

After training the MT5 model with this dataset the accuracy obtained was **57%** on the validation set. This value surpasses the previous mark by 7%, which is an enormous increase in performance. For this reason we sent the predictions of the test split made with this model to the competition and achieved a 72.54%.

²<https://github.com/nidhaloff/deep-translator>

9. Conclusion

This work presents our participation in IberLef 2022 for the subtask 1 in *ReCoRES: Reading Comprehension and Reasoning Explanation for Spanish* which aimed text comprehension and MCQA in Spanish.

Our findings are that while encoder models are a very strong baseline, they perform even better if pretrained on a similar task such as QA. Best results were obtained with more powerful multilingual generative models as they improve the most when increasing available data. In contrast, encoder models show little improvement when including translated data or English data. Generative models were able to leverage the massive multilingual (Spanish, English and Chinese) dataset that we created to improve significantly over the encoder models and gave us the second position in the competition with an accuracy of 72.54%

10. Future work

This project had to be developed in a constrained time frame due to the nature of the competition. This lack of time didn't allow us to test some ideas or hypothesis that could be interesting:

- Prompt engineering: Generative models produce different results when encountering the same problem with different prompts. As shown in [13], a change in how the problem is verbalized to the model can achieve huge gains in accuracy.
- Context windows: Using information retrieval techniques it should be possible to crop the irrelevant information from the context of the question. This could be beneficial to the model as the context would be shorter and more concise.
- Bigger models: MT5, BERT and RoBERTa have bigger variants than those we have used. Bigger variants (as shown in Table 2) have shown better results than smaller ones. We believe this is expected to hold to even bigger models and produce better results.

References

- [1] M. A. Sobrevilla Cabezudo, D. Diestra, R. López, E. Gomez, A. Oncevay, F. Alva-Manchego, Overview of ReCoRES at IberLEF 2022: Reading Comprehension and Reasoning Explanation for Spanish, *Procesamiento del Lenguaje Natural* 69 (2022).
- [2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL: <https://arxiv.org/abs/1810.04805>. doi:10.48550/ARXIV.1810.04805.
- [3] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, Race: Large-scale reading comprehension dataset from examinations, *arXiv preprint arXiv:1704.04683* (2017).
- [4] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.

- [5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: <https://arxiv.org/abs/1907.11692>. doi:10.48550/ARXIV.1907.11692.
- [6] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for squad, CoRR abs/1806.03822 (2018). URL: <http://arxiv.org/abs/1806.03822>. arXiv:1806.03822.
- [7] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, CoRR abs/2010.11934 (2020). URL: <https://arxiv.org/abs/2010.11934>. arXiv:2010.11934.
- [8] Z. Huang, A. Wu, J. Zhou, Y. Gu, Y. Zhao, G. Cheng, Clues before answers: Generation-enhanced multiple-choice qa, 2022. URL: <https://arxiv.org/abs/2205.00274>. doi:10.48550/ARXIV.2205.00274.
- [9] Y. Liang, J. Li, J. Yin, A new multi-choice reading comprehension dataset for curriculum learning, in: W. S. Lee, T. Suzuki (Eds.), Proceedings of The Eleventh Asian Conference on Machine Learning, volume 101 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 742–757. URL: <https://proceedings.mlr.press/v101/liang19a.html>.
- [10] K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, C. Cardie, DREAM: A challenge dataset and models for dialogue-based reading comprehension, CoRR abs/1902.00164 (2019). URL: <http://arxiv.org/abs/1902.00164>. arXiv:1902.00164.
- [11] W. Yu, Z. Jiang, Y. Dong, J. Feng, Reclor: A reading comprehension dataset requiring logical reasoning, in: International Conference on Learning Representations (ICLR), 2020.
- [12] K. Sun, D. Yu, D. Yu, C. Cardie, Investigating prior knowledge for challenging chinese machine reading comprehension, Transactions of the Association for Computational Linguistics (2020). URL: <https://arxiv.org/abs/1904.09679v3>.
- [13] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, 2022. URL: <https://arxiv.org/abs/2205.11916>. doi:10.48550/ARXIV.2205.11916.

A. Comparison of the results

Model	Accuracy(%)
BERT ES	45.5
distilBERT Multilingual	27.6
RoBERTa ES	40.7
RoBERTa ES SQAC	45.8
distilBERT ES SQUAD	45.5
BERT ES SQUAD2	45.8
RoBERTa ES SQUAD2	47.9
MT5-base(L)	27
MT5-base(L+O)	32
MT5-small(L)	0
MT5-small(L+O)	0
GenMC	17.8
BERT Multilingual RACE	27
MT5-base RACE (L+O)	49.7
GenMC RACE	17.8
RoBERTa ES SQUAD2 RACE ES	48.5
MT5-base RACE ES (L+O)	50
MT5-base DATASETS (L+O)	57

Table 5
Accuracy comparison on our evaluation set