IIC/UC3M @ BC8 SympTEMIST track: RigoBERTa and Domain Adaptation for SympTEMIST Subtask1

Guillem García Subies[1,2,*], Álvaro Barbero Jiménez [1] and Paloma Martínez Fernández[2]

[1]Instituto de Ingeniería del Conocimiento, Madrid, Spain
[2]Computer Science and Engineering Department, Universidad Carlos III de Madrid, Leganés, Spain

*Corresponding author: E-mail: guillem.garcia@iic.uam.es

Abstract

In this paper we summarize our participation in the SympTEMIST Subtask1 (SymptomNER) in the BioCreative 2023 workshop. Our main goal is to achieve a good result using the RigoBERTa model and try more sophisticated techniques such as domain adaptation.

Introduction

The main goal of this study is to obtain a good Named Entity Recognition (NER) model for the SympTEMIST Subtask1 (SymptomNER). This task is a token classification task focused on detecting disease symptoms in clinical reports extracted from the SPACC corpus (1).

To achieve our goal, we focus on a simple yet powerful strategy: we use only high-quality language models for the Spanish language such as RigoBERTa (2) and xlm-roberta-large (3) and we fine tune them optimizing their metaparameters with a grid search. We also tried a domain adaptation of the RigoBERTa model. We avoided using gazetteers or any kind of heuristics on purpose given that most of the times it is not feasible to use them in real life applications due to their high complexity and development cost.

Corpus and task

SympTEMIST (4) is a shared task and resource initiative focused on detecting and normalizing symptoms, signs, and findings in medical documents written in Spanish. It aims to enhance text mining capabilities in the medical field for Spanish documents.

Subtask1 of SympTEMIST, known as SymptomNER (Symptoms, Signs & Findings Named Entity Recognition), involves the automatic identification of mention spans of symptoms in clinical reports written in Spanish. Participants are required to use the SympTEMIST corpus, which contains manually labeled mentions, as training data to develop systems capable of providing the character offsets (start and end positions) for all symptom entities mentioned in the text. The primary evaluation metrics for this subtask are precision, recall, and F-score, measuring the system's accuracy in recognizing these medical entities in the text.

The metric used to rank the results will be the Non-overlapping F-score, which means that the complete entity must be predicted correctly and partial guesses or predictions longer that the one in the gold standard will not count towards the metric.
The corpus has annotation guidelines created and curated by experts in the field and the manual annotation process has been revised and postprocessed which means the corpus has a good quality and could be further expanded in the future.

Experiments

For the experiments we will be using RigoBERTa2, an improved version of the original model which has proven to be consistently good, especially in the clinical domain. According to Subies et al. (5), RigoBERTa2 obtained the best results in a benchmark across 12 clinical domain datasets for the Spanish language. RigoBERTa2 was tested along with other Spanish language models like BETO (6) and MarIA (7), other domain specific models like bsc-bio-ehr-es (8) and Galén (9) and also multilingual models like xlm-roberta-large and mDeBERTaV3 (10).

Apart from RigoBERTa, and following the recommendations by Agerri et al. (11) we trained xlm-roberta-large model as a baseline to check that real improvements were made. The training procedure consisted of adding a dense neural network layer on the top of the base model and training them together.

For every trained model we made a grid search to select the best learning rates, dropouts and batch sizes. All the parameter configurations were trained with the train data and the ones that performed the best with the validation data were used to train a model with both train and validation data. Every model was trained until 20 epochs were completed or until the validation score didn’t improve for four consecutive epochs. All the training process was performed on a single NVIDIA A100 80GB GPU.

The parameters used for the grid search are the following: classifier_dropout = [0, 0.1, 0.15, 0.2, 0.21, 0.22, 0.25], learning_rate = [8e-6, 1e-5, 1.5e-5, 2e-5, 3e-5, 4e-5, 5e-5, 9e-5], batch_size = [16, 32], warmup_steps = [0], weight_decay = [0], warmup_ratio = [0].

We also tried to adapt RigoBERTa to the biomedical domain. To do so, we performed one epoch of training with all the background data available for the competition, similar to Gururangan et al. (127and with an effective batch size of 2048.

After all that, we noticed that a lot of predictions had trailing punctuation signs at the end of the entity and, given that the original training data did not, we performed a simple postprocessing of the predictions in order to remove them.

In the following table we can see the results for the test set calculated with Overlapping F-score. In this case partial entities count towards the goal, contrary to Non-Overlapping F-score.

model	Score
xlm-roberta-large	0. 8162
RigoBERTa	0. 8236
RigoBERTa-domain-adaptation	0. 8057
RigoBERTa-postprocess (no domain adapt)	0. 8318

Table1: Overlapping F-scores for the main experiments performed.

RigoBERTa outperformed the baseline, however it did not obtain any gains from the domain adaptation, despite having performed quite well with the validation data. For the domain adaptation we used the same parameters used for the original pretraining of RigoBERTa and we hypothesize that his may have hurt the training process because the corpus used was much smaller and homogeneous.

model	Score
xlm-roberta-large	0. 3215
RigoBERTa	0. 3471
RigoBERTa-domain-adaptation	0. 3567
RigoBERTa-postprocess (no domain adapt)	0. 5062

Table2: Non-Overlapping F-scores for the main experiments performed.

In the table above the Non-Overlapping F-scores are presented. As the models were optimized mostly with Overlapping F-score in mind, the results are notably worse, however a great performance boost is obtained with the postprocessing.

Conclusion

In this paper we have seen that very good results can be obtained with simple approaches and without the need for heuristics such as gazetteers, just with machine learning. RigoBERTa has been proven to be better than other approaches which makes it a great starting point for any problem suited for encoders in Spanish language.

As for future work, right now there is a need for Spanish biomedical domain models, so gathering a big biomedical corpus and adapting RigoBERTa to the domain would be a very interesting feat to achieve. In the NER side, the models should improve by adding a CRF layer and a BiLSTM one like in Lample et al. (13).

Funding

This work was supported in part by the Instituto de Ingeniería del Conocimiento and R&D&i ACCESS2MEET (PID2020-116527RB-I0) project supported by MCIN AEI/10.13039/501100011033/. Additionally, this work has been supported by "Intelligent and interactive home care system for the mitigation of the COVID-19 pandemic" PRTR-REACT UE project.

References

1. Ander Intxaurrondo. (2018). SPACCC (2019-02-01) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.2560316

2. Serrano, A. V., Subies, G. G., Zamorano, H. M., Garcia, N. A., Samy, D., Sanchez, D. B., ... & Jimenez, A. B. (2022). RigoBERTa: A State-of-the-Art Language Model For Spanish. *arXiv preprint arXiv:2205.10233*.

3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

4. Lima-López, S., Farré-Maduell, E., Gasco-Sánchez, L., Rodríguez-Miret, J. and Krallinger, M. (2023) Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In: Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models.

5. Subies, G. G., Jiménez, Á. B., & Fernández, P. M. (2023). A Survey of Spanish Clinical Language Models. *arXiv preprint arXiv:2308.02199*.

6. Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J. (2023). Spanish pre-trained bert model and evaluation data. *arXiv preprint arXiv:2308.02976*.

7. Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., ... & Villegas, M. (2021). Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.

8. Carrino, C. P., Llop, J., Pàmies, M., Gutiérrez-Fandiño, A., Armengol-Estapé, J., Silveira-Ocampo, J., ... & Villegas, M. (2022, May). Pretrained biomedical language models for clinical NLP in Spanish. In Proceedings of the 21st Workshop on Biomedical Language Processing (pp. 193-199).

9. López-García, G., Jerez, J. M., Ribelles, N., Alba, E., & Veredas, F. J. (2021). Transformers for clinical coding in spanish. *IEEE Access*, *9*, 72387-72397.

10. He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

11. Agerri, R., & Agirre, E. (2022). Lessons learned from the evaluation of Spanish Language Models. *arXiv preprint arXiv:2212.08390*.

12. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

13. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.