ICB-UMA at BioCreative VIII @ AMIA 2023 Task 2 SYMPTEMIST (Symptom TExt Mining Shared Task)

Fernando Gallego[1*] and Francisco J. Veredas[1]

[1]Dept. of Computer Languages and Sciences & Research Institute of Multilingual Language Technologies, Universidad de Málaga, Málaga, Spain

*Corresponding author: Tel: +34 952 137 155, E-mail: fgdc2f3@uma.es

Abstract

These working notes summarize the contribution of the ICB research group from the University of Malaga to the BioCreative VIII Workshop @AMIA 2023, from our participation in Task 2 - SympTEMIST. Engaged in both subtasks, our approaches tackled symptom, sign, and clinical finding entities recognition (subtask 1 - SymptomNER) and their normalization to the corresponding SNOMED CT concepts (subtask 2 - SymptomNorm). For subtask 1, we analyzed the performance of some BERT-based models tailored for the nuances of Spanish clinical data. These models, specifically fine-tuned on the SymptomNER corpus, showed remarkable precision (0.804), recall (0.699), and F1-score (0.748) for the test set. For SymtomNorm subtask, we incorporated recent strategies using bi-encoder and cross-encoder models, especially SapBERT models enhanced with FAISS methods for similarity search. Finally, the model's predictions were further refined by leveraging a gazetteer with more than 150,000 concepts. Our strategy achieved 0.58 accuracy for the test set.

Introduction

The digital transformation within the medical sector has enhanced the significance of Electronic Health Records (EHR). These records represent a valuable resource for medical informatics, owing to their capacity to aggregate diverse patient data. Yet, to fully leverage this resource, the transformation of unstructured data into a structured format is imperative. Herein lies the role of Natural Language Processing (NLP): facilitating the conversion of natural language texts in EHRs into structured categories ready for analysis.

NLP's capabilities include named entity recognition (NER), which identifies specific terms within texts; normalization or linking (NEL), which standardizes these terms by using controlled vocabularies and ontologies, such as SNOMED CT (1) or UMLS (2); and clinical coding, which could be considered as a particular type of NEL that aims at classifying clinical mentions into universally recognized codes, such as ICD. Such processes enhance EHR data analysis, fostering improved clinical research and informed medical decision-making. However, a major limitation is the bias in most clinical NLP studies towards the development of models on clinical corpora in English, to the detriment of other languages such as Spanish. This bias is especially significant when considering the development of deep learning (DL) models for NLP, since large datasets in languages other than English are scarce (3).

Although challenges persist, in recent years there has been significant progress in NLP research for Spanish clinical corpora, for example in the field of clinical coding in Spanish (4,5). Clinical coding can be viewed as a subtype of NEL, which is a critical area in medical natural language processing (NLP) research. The aim of NEL is to standardize clinical mentions based on established ontologies and controlled vocabularies. In this context, recent contributions from DL approaches based on bi-encoder and cross-encoder models (6–8) are noteworthy.

Material and Methods

SympTEMIST(9) stands for SYMPtoms, signs and findings TExt MIning Shared Task. The specialized corpora supplied by the organizers of the SympTEMIST shared-task have been meticulously curated and particularly constructed to foster research and development in the realm of NER and normalization within the medical domain. These corpora distinctly emphasize the intricate nuances related to symptoms, signs, and various clinical findings written in Spanish, offering researchers an in-depth perspective into these facets.

For subtask 1 (SymptomNER: Symptoms, Signs & Findings Named Entity Recognition), the corpus supplied by the shared-task organizers consists of an extensive collection of 1,000 clinical case studies, which serves as a comprehensive repository reflecting a spectrum of medical specializations, namely pulmonology, cardiology, oncology, and urology, among others. The organizers supplied the participants with 750 documents for training, keeping the remaining 250 documents for testing.

Given the complexity of the SymptomNER corpus, a systematic approach was crucial for segmentation of the dataset. Thus, for SymptomNER, our resulting training set had 520 documents with 6,354 mention annotations, while the validation set contained 224 documents with 2,738 annotations. Notably, six documents were excluded due to insufficient annotations (see Table 1).

Table 1: distribution of the number of annotations, files, sentences and tokens for subtask 1

Dataset split	**Number of Annotations**	**Number of docs**	**Average number of sentences per doc**	**Average number of tokens per doc**
Train	6354	520	15.88	1235
Validation	2738	224	15.93	538
TOTAL	9092	744	15.91	1773

In the pursuit of the entity normalization subtask 2 (SymptomNorm: Symptom Normalization & Entity Linking), the training corpus supplied by the SympTEMIST organization consisted of 3,484 normalized mentions. Out of these, we reserved 1,028 annotations for validation, and the remaining 2,456 for training (note that 59 mentions were left out since they were annotated as NO_CODE). Furthermore, a gazetteer with 164,817 entries was supplied by the organizers, with detailed attributes like unique code, language, term, and semantic tag, supporting this subtask.

Subtask 1: SymptomNER

For the SymptomNER subtask, we leveraged the capabilities of the XLM-RoBERTa (10) and RoBERTa (3) models to tackle NER tasks. Furthermore, the XLM-Roberta model, an advanced version of the BERT architecture (11), is recognized for its multi-language capabilities, primarily due to its use of the Transformer architecture trained on multilingual corpora, which provides deep contextual representations of input data in different languages.

Our approach with this model involved segmenting input texts into individual sentences, processed at sub-word level. Retaining sentence markers was crucial during both training and validation.

In our study we have compared the performance of 5 different BERT-based models when tackling SymptomNER subtask. On the one hand, these 5 models have been trained and evaluated in an independent manner. Thus, the BSC-Bio-es (12) and RoBERTa-Base-Biomedical-es (13) models are both based on the RoBERTa architecture. For its part, XLM-R-Galén (4) utilizes the XLM-RoBERTa architecture, while mBERT-Galén (4) and Beto are based directly on BERT. Remarkably, to build both XLM-R-Galén and mBERT-Galén, continual pre-trained was utilized on the dataset of our private oncological corpus Galén (4). On the other hand, these trained models have been put together in this study into two different ensembles, following a majority vote approach. A first ensemble (icb-uma-ensemble-1) consisted of the combination of these 5 models. The second ensemble (icb-uma-ensemble-2) comprised only BSC-Bio-es, Roberta-biomedical-es, and XLM-R-Galén, after being trained.

During the experimental phase of hyperparameter adjusting, we made iterative adjustments to batch sizes (from 8 to 64) and learning rates (from 1e-05 to 5e-05). Initial results supported the conclusion that ensemble approaches could improve the performance given by the individual models.

Subtask 2: SymptomNorm

For the SymptomNorm subtask, we utilized the SapBERT-XLM-R-large (6) model. This model is an evolution of the BERT architecture, with features tailored for multilingual contexts. Its ability to link entities to unique concepts across languages is derived from its design and extensive training. Based on the Transformer architecture, SapBERT-XLM-R-large effectively leverages complex contextual details, enhancing its entity linking accuracy. The proposed methodology includes using the SapBERT-XLM-R-large to extract the embeddings from the mentions in the training dataset. To improve candidate selection, we used the Facebook AI Similarity Search (FAISS) (14). This led to an increase in precision metrics by roughly 3%. We also modified the model's inference mechanism, prioritizing decisions from the provided gazetteer supplied by the organizers over model-generated candidates. While this might seem to risk precision, it aimed to increase system reliability in challenging entity linking situations.

Results and Discussion

The SymptomNER subtask required detecting and categorizing named medical entities within a given clinical text. In Table 2, we show the results on the validation and test sets obtained with several models trained for this NER subtask. As shown in the table, for the validation set the individual models BSC-Bio-es and Roberta-Base-Biomedical-es gave F1 scores around 0.73, showing high efficiency rates. Notably, the BSC-BIO-es model gave slightly higher precision, making it a preferred option when precision is critical. The other models analyzed on the validation set, i.e., XLM-R-Galén, mBERT-Galén, and Beto, yielded F1 scores between 0.68 and 0.69 for the validation set, showing their capability to tackle the SympTEMIST NER subtask, though they are slightly behind the top performers. These results are clearly reflected in the performance scores obtained in the test set for the three models whose results were sent to the shared task, i.e., BSC-Bio-es, Roberta-Base-Biomedica-es and XLM-R-Galén, with the first one giving the best efficiency rates. Significantly, ensemble techniques have demonstrated to yield enhanced performance metrics when compared to individual models (see Table 2). The icb-uma-ensemble-1 ensemble achieved a precision of 0.804 for the test set and a F1 score of 0.748. In contrast, the icb-uma-ensemble-2 continued to exhibit high performance rates, closely competing with the leading ensemble.

For the SymptomNorm NEL subtask, the goal was to link named entities with their respective SNOMED CT codes. Given the complexity of SNOMED CT terminologies, models must have a close understanding of the details and nuances of the mentions to yield accurate predictions. We used the accuracy-@-top-k metric (with k=1) for evaluation, to measure the model's ability to identify the most relevant SNOMED CT code based on its primary prediction. This metric assesses the model's accuracy without considering multiple predictions.

Table 2: Performance results of models on the NER task.

	Validation set (mean ± std)			**Test set**		
Model name	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
BSC-Bio-es	0.723 ± 0.002	0.742 ± 0.003	0.705 ± 0.005	0.729	0.744	0.713
RoBERTa-Base-Biomedical-es	0.720 ± 0.001	0.735 ± 0.002	0.705 ± 0.005	0.716	0.729	0.704
XLM-R-Galén	0.688 ± 0.002	0.728 ± 0.002	0.653 ± 0.005	0.694	0.714	0.682
mBERT-Galén	0.681 ± 0.003	0.715 ± 0.003	0.650 ± 0.006	–	–	–
Beto	0.681 ± 0.002	0.703 ± 0.002	0.661 ± 0.005	–	–	–
icb-uma-ensemble-1	0.744 ± 0.003	0.813 ± 0.002	0.686 ± 0.005	**0.748**	0.804	0.699
icb-uma-ensemble-2	0.743 ± 0.002	0.795 ± 0.003	0.697 ± 0.005	<u>0.746</u>	0.790	0.707

As shown in Table 3, SapBERT-XLM-R-Large, when combined with FAISS similarity search, gave the best precision results (0.65) among the three scenarios analyzed on the validation set. The precision achieved by the SapBERT-XLM-R-Large model in the validation set (0.62) is comparable with that obtained with the combination of SapBERT-XLM-R-Large with FAISS and the subsequent use of the gazetteer as a look-up table for candidate search. For this subtask, we only contributed with the results obtained with the SapBERT-XLM-R-Large+FAISS+Gazetteer model, which achieved accuracy of 0.58.

Table 3: Performance results (accuracy) of models on the NEL task.

Model name	**Validation set**	**Test set**
SapBERT-XLM-R-Large	0.62	–
SapBERT-XLM-R-Large+FAISS	0.65	–
SapBERT-XLM-R-Large+FAISS+Gazetteer	0.62	0.58

In conclusion, domain-adapted and multilingual language models, combined with ensemble strategies, are shown to be competitive for medical NER tasks in Spanish. On the other hand, clinical NEL is a highly complex task, where bi-encoder and cross-encoder based models are promising. As future work, we propose the use of specialized ontologies and gazetteers as tools to refine these models and adapt them to the peculiarities of clinical NEL in Spanish.

For future work, we intend to prioritize the exploration of advanced candidate reordering techniques, including the Bi-Encoder and Cross-Encoder architectures, as well as employing a classification-based approach utilizing Normalized-Temperature (NT) Softmax. Our objective is to refine and establish new benchmarks for state-of-the-art entity linkage within the medical domain.

The development of the model was carried out utilizing the PyTorch 2.0.1 and Tensorflow 2.13 frameworks, in conjunction with the Transformers library from Hugging Face 0.17.1, and the FAISS 1.7.4 library for efficient similarity search and clustering of dense vectors.

Funding

This work was supported by the Spanish *Ministerio de Ciencia e Innovación*, Project PID2020-116898RB-I00.

References

1. Benson T, Grieve G. Principles of Health Interoperability: FHIR, HL7 and SNOMED CT [Internet]. SN; 2020. 475 p. Available from: https://play.google.com/store/books/details?id=TiwEEAAAQBAJ

2. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res [Internet]. 2004 Jan 1;32(Database issue):D267-70. Available from: http://dx.doi.org/10.1093/nar/gkh061

3. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol [Internet]. 2020 Feb;145(2):463–9. Available from: http://dx.doi.org/10.1016/j.jaci.2019.12.897

4. Lopez-Garcia G, Jerez JM, Ribelles N, et al. Transformers for clinical coding in Spanish. IEEE Access [Internet]. 2021;9:72387–97. Available from: https://ieeexplore.ieee.org/abstract/document/9430499/

5. López-García G, Jerez JM, Ribelles N, et al. Explainable clinical coding with in-domain adapted transformers. J Biomed Inform [Internet]. 2023 Mar;139:104323. Available from: http://dx.doi.org/10.1016/j.jbi.2023.104323

6. Liu F, Shareghi E, Meng Z, et al. Self-Alignment Pretraining for Biomedical Entity Representations [Internet]. arXiv [cs.CL]. 2020. Available from: http://arxiv.org/abs/2010.11784

7. Zettlemoyer M, Wu L, Petroni F. Zero-shot entity linking with dense entity retrieval. Proceedings of the 2020 Conference on.

8. Zhu T, Qin Y, Chen Q, Hu B, Xiang Y. Enhancing entity representations with prompt learning for biomedical entity linking. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence [Internet]. California: International Joint Conferences on Artificial Intelligence Organization; 2022 [cited 2023 Oct 20]. Available from: https://www.ijcai.org/proceedings/2022/0560.pdf

9. Lima-López, S., Farré-Maduell, E., Gasco-Sánchez, et al. 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: BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models. 2023.

10. Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale [Internet]. arXiv [cs.CL]. 2019. Available from: http://arxiv.org/abs/1911.02116

11. Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv [cs.CL]. 2018. Available from: http://arxiv.org/abs/1810.04805

12. Carrino CP, Llop J, Pàmies M, Gutiérrez-Fandiño A, et al. Pretrained Biomedical Language Models for Clinical NLP in Spanish. In: Proceedings of the 21st Workshop on Biomedical Language Processing [Internet]. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 193–9. Available from: https://aclanthology.org/2022.bionlp-1.19

13. Carrino CP, Armengol-Estapé J, Gutiérrez-Fandiño A, et al. Biomedical and Clinical Language Models for Spanish: On the Benefits of Domain-Specific Pretraining in a Mid-Resource Scenario [Internet]. arXiv [cs.CL]. 2021. Available from: http://arxiv.org/abs/2109.03570

14. Johnson J, Douze M, Jegou H. Billion-scale similarity search with GPUs. IEEE Trans Big Data [Internet]. 2021 Jul 1;7(3):535–47. Available from: https://ieeexplore.ieee.org/abstract/document/8733051/