Team BIT.UA @ BC8 SympTEMIST Track: A Two-Step Pipeline for Discovering and Normalizing Clinical Symptoms in Spanish.

Richard A. A. Jonker[1], Tiago Almeida[1,*], Sérgio Matos[1], João Almeida[1]

[1]IEETA/DETI, LASI, University of Aveiro, Portugal
*Corresponding author: E-mail: tiagomeloalmeida@ua.pt

Abstract
This paper presents the participation of the Biomedical Informatics and Technologies group (BIT) from the University of Aveiro in the SYMPTEMIST task at BioCreative VIII, with a primary focus on biomedical entity recognition and normalization tasks. We leverage a transformer-based solution with MCRF for entity recognition and hybrid semantic search approach for the normalization. Both our methods achieved top-performing scores, especially, our best entity recognition submission achieved 0.7369 F1 (3.69 points above median), while our best submission for normalization achieved 0.5890 (5.90 points above median). Code to reproduce our submissions is available at https://github.com/ieeta-pt/BC8-SympTEMIST.

Introduction
Despite significant advancements in healthcare text mining and Natural Language Processing (NLP), the automatic detection and normalization of clinical symptoms in textual data remains underexplored. The SYMPTEMIST (1) challenge aims to fill this gap by encouraging the creation of systems for autonomous symptom mention detection and normalization within Spanish clinical text. In this work, we follow the well-established principles for biomedical entity discovery and normalization outlined in (2), and further explored for Spanish medical procedures in (3,4). With more detail, we employed a transformer-based model coupled with a Masked Conditional Random Field (MCRF) for entity detection. For normalization, we implemented a two-step search approach: initially conducting an exact match search, followed by a semantic search to link the medical codes to the entities.

Methodology
The dataset comprises a collection of Spanish documents containing biomedical entities, specifically symptoms and signs. In the first task, the objective is to identify these entities, while in task 2, the goal is to link these identified entities to SNOMED-CT codes. Finally, subtask-3 is an experimental task in which multilingual entities need to be mapped to the same SNOMED codes.

Subtask 1 – Symptoms, Signs and Findings Recognition
We approach the entity recognition problem as a sequence labeling task, where (sub)-tokens are classified as part of an entity or not. To facilitate this, we have adopted the BIO tagging schema.

Our models are built upon previous work by Almeida et al. (2,3), which employs a transformer architecture incorporating a Masked Conditional Random Field (MCRF) (5) as the classification

layer. Similarly, we also incorporate data augmentation during training. This model comprises three essential components: a transformer-based model trained in the Spanish language, an encoder layer, and a classification head.

1. **Transformer model**: We employ a fine-tuned RoBERTa-based model designed for medical Named Entity Recognition (NER) in Spanish text (lcampillos/roberta-es-clinical-trials-ner) (6). This model proved to be the most effective in our experiments.
2. **Encoder layer**: For the encoder layer, we primarily utilize a BiLSTM layer.
3. **Classification head**: Masked CRF layer, which is similar to a regular CRF layer but includes a mask over its transitional weights. This modification is employed to enforce constraints encoded within the BIO tagging schema.

Given the limitations of transformer models in handling longer sequences (usually up to 512 tokens), we split larger documents to fit within the input size constraints. To enhance the performance on split documents, we retain a right and left context region that is not decoded but aids the model in understanding the previous and subsequent sequences. After some initial testing, a context size of 4 tokens was determined as optimal and was used.

Like previous work (3), we employ data augmentation techniques to improve the model's generalization and overall performance. Two augmentation techniques are utilized:

- **Random Token Replacement**: This technique involves randomly replacing tokens in the input sequence with any other valid token from the vocabulary (2,3,7).
- **Random Token Replacement with Unknown**: This technique is a variation of the previous one, where randomly selected tokens in the input sequence are replaced by the '[UNK]' special token (2,3).

The objective of data augmentation techniques is to encourage the model to consider contextual information when making predictions. By replacing certain tokens, the model is forced to examine neighboring tokens to determine whether the replaced token should be part of an entity. Additionally, models are trained with varying percentages of random replacement.

Lastly, we use an ensemble technique at the entity level to boost performance by leveraging the insights of multiple models. To do this, we perform an exact matching process for each annotated entity, comparing the entities predicted by different models. A majority voting scheme is used to select the entities that best represent the document.

Subtask 2 – Symptoms, Signs and Findings Normalization

To perform the normalization of symptoms to SNOMED CT codes, we employ a multistage code discovery technique. In the initial step we perform exact matching against the training data and subsequently against the SNOMED CT codes themselves. This is followed by a semantic search approach using cosine similarity. Leveraging the multilingual SapBERT (8) model, we calculate embeddings for the relevant SNOMED CT terminologies, the training data, and the annotated entities. With these embeddings, we compute the cosine similarity from each entity against all the codes and then pick the top-1 above a predefined threshold. By combining these two approaches, we have developed a robust system, all without requiring specific training. Additionally, we also examined the priority of exact matching, finding that prioritizing training data generally produced better results.

Subtask 3 – Multilingual Symptoms, Signs and Findings Normalization
To address the multilingual task, we adopted an approach involving the translation of all texts into Spanish and applied the methods developed for subtask 2. For the translation task, we leveraged the existing models developed by Helsinki-NLP/Opus-MT (9). In nearly all cases, we had direct translation models available, with one exception being the translation from Portuguese to Spanish. In this case, we utilized an intermediate translation step from Portuguese to Catalan and then from Catalan to Spanish. Notably, there were no validation tests conducted for this task.

Results

In this section, we discuss the results of the runs submitted for the competition. In Subtask 1, the exact match F1 score is the primary evaluation metric, with overlapping F1 as a secondary metric. In Subtasks 2 and 3, accuracy is the sole metric used to assess the model performance.

Subtask 1 – Symptoms, Signs and Findings Recognition
We submitted five different models, four of which are ensemble models, and the other is the best-performing model based on the validation data. The top-performing model (2-best) on the training data utilized UNK augmentation, with a 50% chance of altering 20% of the tokens. The remaining models include an ensemble model comprising all models (1-all), an ensemble of the top 5 models (3-top5). Lastly, the last two ensembles encompass all models with UNK augmentation (5-unk) and all models with random augmentation (4-random). Based on the validation results, augmentation yielded benefits compared to models without augmentation. The results for these models can be seen in **Table 1**.

Run	***P***	***R***	***F1***	***O-P***	***O-R***	***O-F1***
1-all	0.7473	0.7258	0.7364	**0.8816**	0.8563	**0.8688**
2-best	0.7315	0.7274	0.7294	0.8675	**0.8628**	0.8651
3-top5	0.7411	0.7258	0.7334	0.8757	0.8576	0.8665
4-random	**0.7469**	0.7271	**0.7369**	0.8786	0.8553	0.8668
5-unk	0.7426	**0.7287**	0.7356	0.8759	0.8595	0.8676
Mean	-	-	0.6100	-	-	0.8200
Median	-	-	0.7000	-	-	0.8400

Table 1: Results of the submitted runs in the competition, including Precision, Recall, F1 and the overlapping scores for these metrics.

All our models performed significantly above the mean, with at least a two-percentage point lead over the median values. Our best-performing model was the ensemble of all random models. All ensemble models outperformed the best-performing individual model, showcasing the robustness of the ensemble technique, primarily due to an increase in precision. We achieved the highest overlapping F1 score in the competition with the ensemble of all models.

Subtask 2,3 – Symptoms, Signs and Findings Normalization
In Task 2, we submitted five runs, all of which were slightly different variations of the same methodology. We varied the cosine threshold from 0.65, which performed best in validation (text_snomed_065), to 1 (text_snomed_1) and to 0 (text_snomed_0). We further explored the impact of model size; the initial four runs employed the large variant of SapBERT, while the final run utilized the base size (text_snomed_065_b). Additionally, we investigated the influence of altering the order of exact dictionary matching (snomed_text_065). In Subtask 3, we used the

same models, except for one of them. The table containing the results for Subtask 2 and 3 can be found in **Table 2**, where the 'ES' column corresponds to the submission for Subtask 2.

Model	***ES***	***EN***	***FR***	***IT***	***NL***	***PT***
text_snomed_065	**0.5890**	**0.7250**	**0.5726**	0.6697	**0.6397**	**0.5575**
snomed_text_065	0.5659	-	-	-	-	-
text_snomed_0	0.5859	**0.7250**	**0.5726**	**0.6703**	0.6389	0.5569
text_snomed_1	0.4032	0.5265	0.3944	0.5065	0.4764	0.3254
text_snomed_065_b	0.5778	0.7137	0.5733	0.6626	0.6317	0.5542
Mean	0.2100					
Median	0.5200					

Table 2: Results of the submitted runs in subtask 2 and 3. The metric measured is accuracy.

In Subtask 2, our best-performing system prioritized direct matching from the training data, followed by SNOMED, and used a threshold of 0.65 for acceptance from semantic search. This model tied for second place in the competition, surpassing the median score by 6 percentage points. We observed a slight drop in performance when prioritizing SNOMED direct matching first, suggesting that some codes may differ from the actual SNOMED corpus, emphasizing the potential importance of context. The model with complete acceptance had a minor reduction in performance in Task 2, but it still performed well overall. As expected, setting a threshold of 1 significantly decreased performance, underscoring the importance of semantic search and highlighting the effectiveness of direct matching. Lastly, as anticipated, there was a performance drop when examining the base model size (*text_snomed_065_b*). In the experimental Track 3, there were no metrics provided for us to compare our systems. However, as expected, the conclusions drawn from Subtask 2 held true for Subtask 3. We also observed that the model's performance is comparable to that in Subtask 2 in most cases, and even better for certain languages.

Conclusion

We have described our participation in the BioCreative VIII challenge, outlining our methodology and results across various subtasks. The top performing results for both subtasks showcase the efficacy of the proposed methods. These results showed that the use of ensemble techniques and the effectiveness of data augmentation played pivotal roles in enhancing our performance in entity recognition. Following this, our prioritization of direct matching from training data and optimization of semantic search thresholds significantly contributed to our superior performance in normalization. Moreover, in Subtask 3, our adaptation to a multilingual context exhibited its applicability across various languages. These findings significantly contribute to the field of biomedical entity recognition and normalization, with broader implications for advancing biomedical information extraction systems.

Funding

This work was partially supported by national funds through the Foundation for Science and Technology (FCT) in the context of the projects DSAIPA/AI/0088/2020 and UIDB/00127/2020. Tiago Almeida is funded by FCT under the grant 2020.05784.BD.

References

1. Lima-López, S., Farré-Maduell, E., Gasco-Sánchez, L., et al. (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. *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models.*

2. Almeida, T., Antunes, R., Silva, J.F., et al. (2022) Chemical identification and indexing in PubMed full-text articles using deep learning and heuristics. *Database*, **2022**.

3. Almeida, T., Jonker, R.A.A., Poudel, R., et al. (2023) BIT.UA at MedProcNer: Discovering Medical Procedures in Spanish Using Transformer Models with MCRF and Augmentation. *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*.

4. Lima-López, S., Farré-Maduell, E., Gascó, L., et al. (2023) Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023. *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*.

5. Wei, T., Qi, J., He, S., et al. (2021) Masked Conditional Random Fields for Sequence Labeling. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Online, pp. 2024–2035.

6. Campillos-Llanos Leonardo and Valverde-Mateos, A. and C.-C.A. and M.-S.A. (2021) A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC Med Inform Decis Mak*, **21**, 69.

7. Erdengasileng, A., Li, K., Han, Q., et al. (2021) A BERT-Based Hybrid System for Chemical Identification and Indexing in Full-Text Articles. *bioRxiv*.

8. Liu, F., Vulić, I., Korhonen, A., et al. (2021) Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking. *Proceedings of ACL-IJCNLP 2021*, pp. 565–574.

9. Tiedemann, J. and Thottingal, S. (2020) OPUS-MT — Building open translation services for the World. *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.