PicusLab @ BC8 SympTEMIST track: Disambiguating Entity Linking Candidates with Question Answering

Michele Cirillo[1], Vincenzo Moscato[1,2] and Marco Postiglione[1,2,*]

[1]Department of Electrical Engineering and Information Technology (DIETI), University of Naples "Federico II", Via Claudio 21, Naples, Italy

[2]CINI Consorzio Interuniversitario Nazionale per l'Informatica, Rome, Italy

*Corresponding author: Tel: +39 392 874 8881, E-mail: marco.postiglione@unina.it

Abstract
In the field of biomedical informatics, entity linking plays a pivotal role in enhancing search capabilities, integrating heterogeneous data, and fostering advanced semantic understanding. Nonetheless, there is a significant lack of linguistic resources specifically designed for developing entity linking frameworks. Most existing datasets and concept aliases in primary ontologies are predominantly in English. This scarcity poses significant challenges in generating potential entities for linkage and in disambiguating among candidate entities, particularly for under-resourced languages. In this work, we describe our contribution to the BioCreative SympTEMIST shared task, which focuses on the detection and normalization of symptoms, signs and findings in medical documents in Spanish. Our methodology employs a pre-trained Spanish RoBERTa model in tandem with a cross-lingual SapBERT model for candidate generation, followed by a disambiguation phase utilizing a Question Answering-based module. Extending our approach to the multilingual sub-task, we demonstrate its adaptability. We conducted experiments on an internal test set mirroring the SNOMED codes distribution of the training set, which underscored the efficacy of our approach. Nonetheless, the challenge results highlight a need for additional investigation to tailor the framework to unforeseen codes.

Introduction
Utilizing Named Entity Recognition (NER) and Entity Linking (EL) to extract structured metadata from text documents forms the foundation for various downstream tasks, such as predicting diseases (1) and repurposing treatments (2). Notably, comprehensive ontologies like SNOMED CT have been crafted for the clinical field, facilitating seamless communication among software systems and aiding numerous clinical applications (3).

Clinical ontologies offer the potential for detailed semantic annotation of text documents, but their complex nature also presents challenges for automatic annotation systems. Selecting the right mappings from text to ontology concepts is context-dependent and often ambiguous, varying subtly with different application domains. Additionally, Entity Linking (EL) is resource-constrained in several aspects. First, annotated datasets cover only a minuscule portion of concepts from prevalent biomedical terminologies like the Unified Medical Language System (UMLS) (4). Second, most terms in these ontologies are primarily available in English, with limited representation in other languages. Hence, datasets annotated with entity mentions and their

corresponding ontology mappings are rare yet crucial for advancing the field of biomedical EL, especially for non-English languages.
In this work, we describe our contribution to the SympTEMIST shared task (5), which aims to the detection and normalization of symptoms in Spanish clinical notes. Our proposed EL framework begins by producing multiple candidate links. Subsequently, these links are disambiguated using a Question-Answering (Q/A) model.

Materials

The SympTEMIST dataset comprises 1,000 clinical case reports in Spanish annotated with references to symptoms, signs, and findings. These annotations are standardized to SNOMED CT.

Subtask 2: SymptomNorm (Symptom Normalization & Entity Linking)

The dataset provided by challenge organizers contains 3,484 samples from 304 unique clinical case reports. The training set contains a total of 1,535 unique concepts. The top-20 concepts and the relative occurrences in training data are shown in Figure 1.

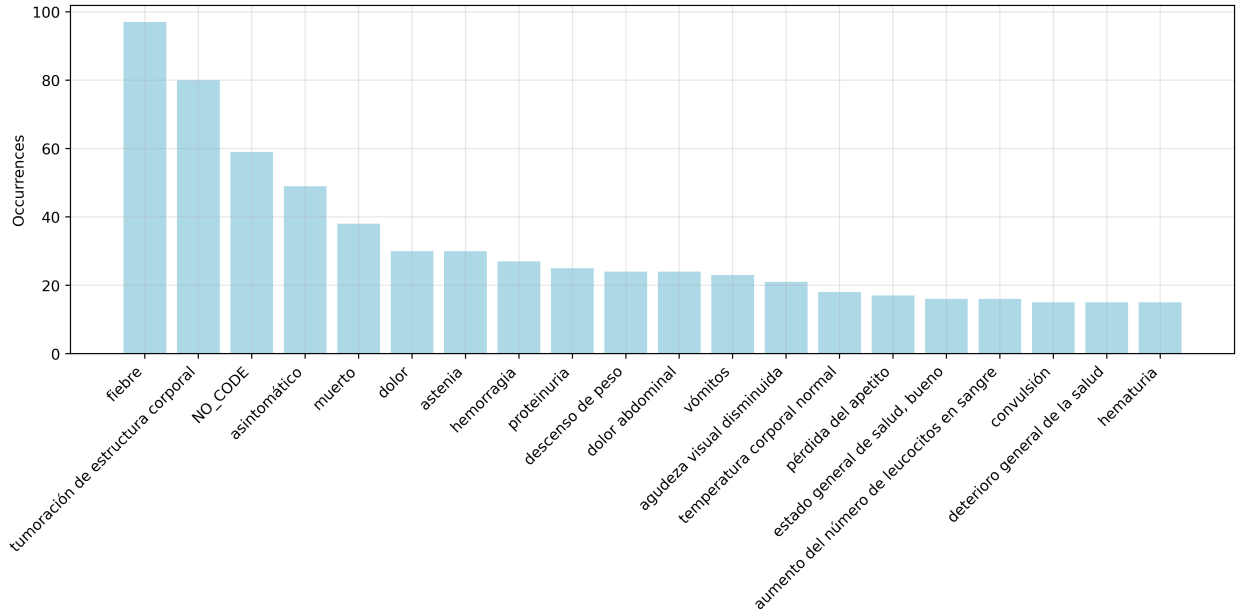

Figure 1. Top-20 occurring concepts in SympTEMIST training data.

Subtask 3: SymptomMultiNorm (Experimental English/multilingual Symptom Normalization)

This subtask aims to advance entity linking and clinical concept normalization methodologies across multiple languages, specifically: English, Portuguese, French, Italian, and Dutch. In our experiments, we take the Italian language in consideration. It contains 1,924 samples from 294 case reports. The training set contains a total of 823 unique concepts.

Gazetteer

The gazetteer contains main terms and synonyms from the relevant branches of SNOMED CT for the grounding of symptom mentions. It contains all the codes that may appear in the test set. Please

note that we have expanded the gazetteer by adding brief descriptions of the available codes. These descriptions were generated by inputting the following prompt into GPT-3.5:

> You possess expertise within the medical domain. Write a short description (one sentence) for the medical concept "<concept name>" (SNOMED CT code: <concept code>).

Here, concept name and concept code are retrieved from the gazetteer. Please note that although we have presented an English prompt for reader clarity, the original prompt – and consequently, the results – were in Spanish.

Methods

The workflow of the proposed method is shown in Figure 2. In a nutshell, for every given input sample, we generate a corresponding set of candidate codes. This is achieved through exploiting the cosine similarities between the embedding of the input mention and the embeddings of mentions found in the training set and gazetteer. Subsequently, we employ a multiple-choice Question/Answer (Q/A) module to disambiguate among the candidate links. This module utilizes the sentence and the descriptions of codes as potential answers.

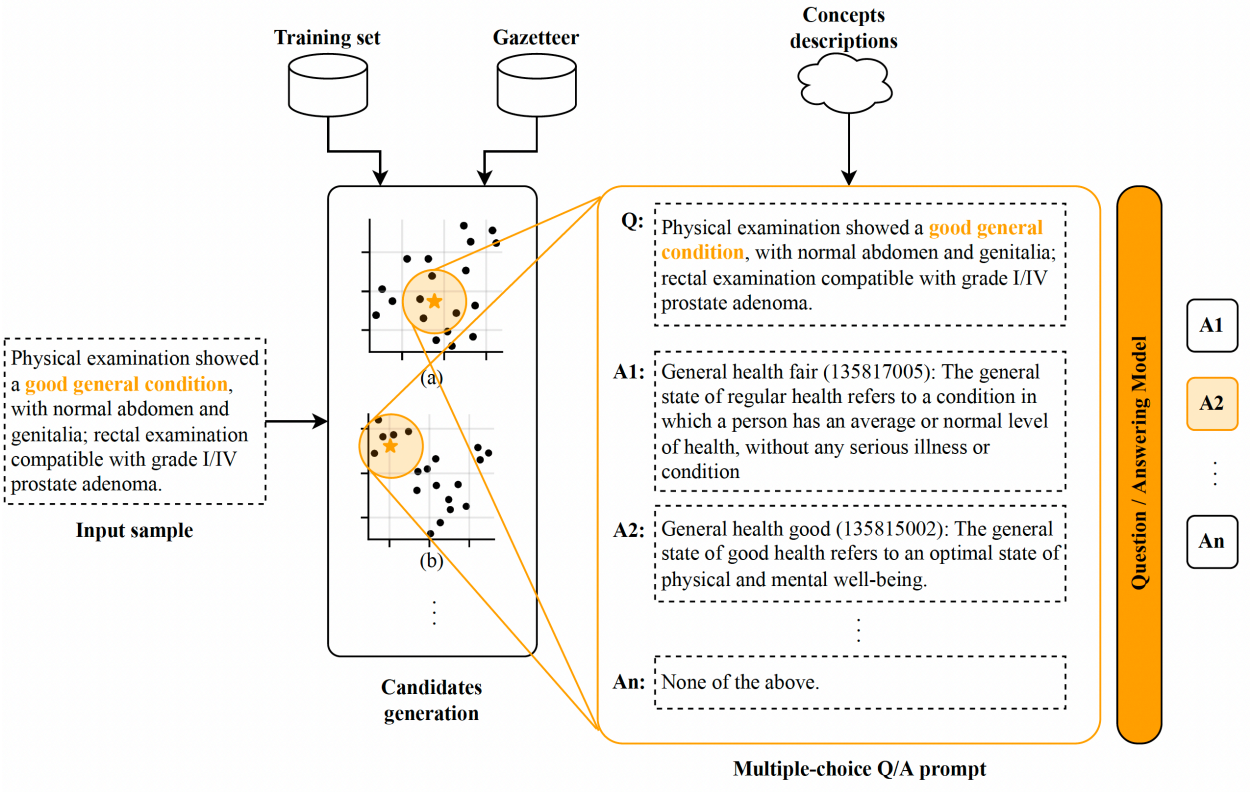

Figure 2. Methodological workflow. Please note that input sentences and candidate answers are in Spanish.

While our method can be easily extended with multiple candidate generation methods, we have tested only two methods within our participation to the SympTEMIST challenge. Specifically, we have used (a) a RoBERTa-based model trained on a biomedical-clinical corpus in Spanish (6) and

(b) a SapBERT multilingual model trained with UMLS 2020AB (7), using xlm-roberta-large as the base model for Q/A.

We employed our framework for subtask 2 (Spanish EL) and subtask 3 (experimental multilingual EL), specifically focusing on the Italian language. In light of this, we utilized GPT-3.5 to automatically translate the concept names, descriptions, and test clinical notes when needed.

Results and Discussion

In this section, we detail our results obtained on both an internal test set and the official leaderboard, thus highlighting and discussing their discrepancies.

Internal training/test splitting

We set aside 20% of the available training dataset for internal testing. The internal test set was created by stratifying the data based on the frequency of concept occurrences. However, it is important to note that this splitting strategy led us to exclude from the internal test set all concepts that appeared only once in the training set. This resulted in outcomes that did not reflect the challenge's actual results. In fact, the training set contains as many as 1,000 concepts out of 1,535 with a single occurrence, and it seems the situation with the test set is not different.

Results

In Figure 3 and Table 1, we present our findings from the internal test set. Please note that in the interest of conciseness, we will focus our presentation on the outcomes from the second subtrack. Nonetheless, the conclusions drawn are analogous and can be extrapolated to the third subtrack. Figure 3 details the performance of our system as we incorporate an increasing number of candidates into our Q/A model. Notably, the peak results, marked with a yellow triangular indicator, demonstrate the effectiveness of the Q/A-based disambiguation strategy with RoBERTa candidates. However, this approach is not beneficial for SapBERT candidates. On the other hand, Table 1 underscores the true potential of our method: by amalgamating diverse candidate generation strategies, we can harness top alternatives from each, which, while distinct, are of equal significance. Please note that before submitting our results we also used the training look-up strategy proposed by Borchert et al. (8).

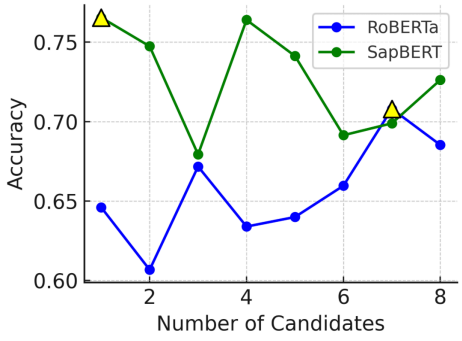

Figure 3. Accuracy of the system with different candidate generation methods as the number of candidates we consider increases. Best results are highlighted with a triangular marker.

Table 1. Results of the best alternatives for each candidate generation method and an hybrid solution that combines them.

Method	# Candidates	Accuracy
RoBERTa	7	0.708
SapBERT	1	0.765
Hybrid	7 RoBERTa + 1 SapBERT	**0.790**

Although our internal test set showed promising results, our hybrid solution did not yield improvements on the external test data. This is likely attributed to the flawed splitting strategy, which consequently influenced our decision on the number of candidates to consider.

References

1. D'Auria, D., Moscato, V., Postiglione, M., Romito, G., & Sperlí, G. (2023). Improving graph embeddings via entity linking: A case study on Italian clinical notes. *Intelligent Systems with Applications*, 17, 200161.

2. McCoy, K., Gudapati, S., He, L., Horlander, E., Kartchner, D., Kulkarni, S., ... & Mitchell, C. S. (2021). Biomedical text link prediction for drug discovery: a case study with COVID-19. *Pharmaceutics*, 13(6), 794.

3. Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121, 279.

4. Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1), D267-D270.

5. 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*

6. Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained Biomedical Language Models for Clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

7. Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

8. Florian Borchert and Matthieu-P. Schapranow. HPI-DHC @ BioASQ DisTEMIST: Spanish Biomedical Entity Linking with Pre-trained Transformers and Cross-lingual Candidate Retrieval. *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, pp. 244-258. Bologna, Italy.