An exploratory approach to the SympTEMIST challenge

José Castaño[1*] , Bruno Cruz Franchi[1] , Sonia Benitez[1], Carlos Otero[1], Daniel Luna[1]
[1]Departmento de Informática en Salud, Hospital Italiano de Buenos Aires, Argentina.
*Corresponding author: E-mail: jose.castano@hospitalitaliano.org.ar

Abstract

This paper describes the approach of the HIBA team in the SympTEMIST shared tasks at the BioCreative VIII challenge and workshop held at the Amia 2023 Symposium. This challenge provides an annotated corpus that enables different systems to evaluate their predictions on the detection of "symptom" mentions in electronic health records and their normalization to SNOMED-CT codes. We submitted our predictions for the entity mention detection using a transformer-based approach, obtaining an F1 measure of 0.71.

Introduction

Natural Language Processing (NLP) techniques have been used to extract and access information in clinical documents. This information can be structured to enable a variety of applications in the healthcare domain, to facilitate decision-making or alleviate routine time-consuming tasks. The Biomedical Text Mining Unit at the Barcelona Supercomputing Center has been organizing several challenges, providing the domain datasets for the corresponding evaluation. These datasets have been annotated following a consistent methodology. The SympTEMIST dataset and challenge follow several previous ones, such as DisTEMIST [1], CanTEMIST [2], MedProcNER [3]. In most of these challenges, the best F1 measure results for Named Entity Recognition (NER) have been in the mid 0.70s and the best Entity Linking (EL) or Entity Normalization results in the low 0.50s.

In similar problems of other domains, the involved tasks are also NER and EL. Many techniques have been used in the past such as dictionary or gazetteer-based, rule-based or machine learning, and any combination of them. Deep learning and transformation-based learning technologies have been extensively used, and have emerged as dominant methodologies in recent years, yielding good results. The SympTEMIST challenge (symptoms, signs, and findings Text Mining Shared Task) proposes these two tasks, Symptoms-entities (NER) and Symptoms-linking (EL), including a multilingual sub-task). The cover name used is Symptoms, but it does not correspond to a category in a given ontology. Our approach was exploratory on different techniques and more focused on what were the characteristics of the dataset, and possible variations on resources and well known techniques.

The SympTEMIST Dataset

The SympTEMIST corpus data [4] is complementary to the DisTEMIST and MedProcNER datasets, as the three of them use the same document collection: 1000 files corresponding to

clinical cases. It has been randomly divided into a training set, 750 clinical cases, and a test set of 250 cases.
Similarly to the DisteMIST challenge, there were three sub-tasks, the first sub-task, (NER), required identifying the named entities corresponding to the cover category Symptom (SINTOMA). The second sub-task, (EL) required to provide for each named entity recognized the corresponding SNOMED-CT concept. The training files also provided information about whether the corresponding named entity was an EXACT description for the SNOMED-CT concept (2609 instances) or a NARROW description (819 instances). In other words, the entity mention is not a direct mapping to the SNOMED-CT code. Information on whether the term was an abbreviation or needed context to disambiguate the meaning was also provided. However, none of these parameters were required to be submitted. They might be used as input to the training algorithm.
The training set for sub-task 1 contained 9093 entity mentions, but there were only 6178 unique ones, The entity linking training set for sub-task2 involved only 305 clinical cases with 3485 entity mentions and 1536 unique ones.

Exploration of the SympTEMIST Dataset using Terminology Resources

The Hospital Italiano of Buenos Aires (HIBA), has a Spanish interface terminology where each term is mapped to SNOMED-CT as its reference vocabulary, we described the HIBA terminology and our approach to DisteMIST at the CLEF 2022 [5].
We used a simple NER and EL system in Python in order to use the terminology. In our approach, we employed open-source libraries, in a standard pipeline of tokenization, parsing, and NER (spaCy , MedspaCy, Quickumls). This system allows to recognize those terms that exist in the controlled vocabulary. It also allows to select terms using UMLS types, or HIBA terminology codes. HIBA terminology codes are mapped to SNOMED-CT codes.

When we applied the approach we had used with the DisTEMIST datasets to SympTEMIST datasets, we noticed that the results were not very satisfactory. At the DisTEMIST entity-mention task, we obtained an F1 measure of 0.49. It seemed very difficult to reach even that measure. In particular, from the analysis of the UMLS types that our system mapped to the training terms, the two on the top rank were expected (T033 Finding and T184 Sign or Symptom), but the following types in the rank seemed to be out of the scope of the task, such as T047 Disease or T046 Pathologic Function. We also noticed that our system mapped to a larger set of UMLS types, which made it more difficult to figure out if the terminology resources could be used in any way, either by themselves or in combination with a learning algorithm.

Table 1 bellow, shows the distribution of UMLS types corresponding to the SympTEMIST entity mention terms, in the training set, using the HIBA, SNOMED-CT and UMLS terminological resources.

UMLS TYPE	Description	#terms
T184	Sign or Symptom	1893
T033	Finding	1851
T047	Disease or Sindrome	753
T059	Laboratory Procedure	444
T046	Pathologic Function	377
T060	Diagnostic Procedure	333
T061	Therapeutic or Preventive Procedure	146
T048	Mental or Behavioral Dysfunction	114

Table 1: Distribution of the most frequent UMLS types in the training set

Named-entity Predictions Based on Transformer Models

To recognize entity mentions we also utilized a standard NER approach using Hugging Face Transformers and a token classification approach. The encoder was initialized with weights from a Spanish clinical domain RoBERTa model. To implement a token-based NER, documents are split into sentences and tokens using a general-domain spaCy model. We adapted a code used for DisTEMIST available on Github [6]. Two base models were used: the first being, the PlanTL-GOB-ES/roberta-base-biomedical-clinical-es, henceforth (RB) and the second, an adaptation of this model, for which we used a corpus of books from the clinical domain. We call this base model roberta-base-clinical-modified henceforth (RBM).

Training and Model Selection

For model selection and estimation of generalization performance, we split the dataset into a training and validation set. We integrated the use of the Weight & Biases platform to facilitate the model training and evaluation. Therefore we automatized the optimal hyper-parameter search to maximize the F1 score, each training was run for 100 epochs. Table 2 shows the Sweep configuration for the hyperparameters tuning.

Hyperparameter	RB	RBM	Values
Label smoothing Factor	0.2	0.2	[0, 0.05, 0.1, 0.2]
Learning Rate	0.000005	0.00001	[5e-06, 1e-05, 5e-05, 1e-04, 5e-04]
Learning Rate Scheduler	Constant with warmup	Linear	Cosine with restarts, Linear, Constant with warmup
Warmup Ratio	0.05	0	[0, 0.05, 0.1]
Weight Decay	0.05	0.05	[0.05, 0.1, 0.2, 0.3, 0.5, 1]

Table 2: Hyperparameter sweep configuration using Weight & Biases and selected values

Results and Discussion

The results of the NER task on the training and test dataset were not surprising. There was some variation in the final hyperparameters selected and some lower performance in the test set results. The base model that was further adapted to the clinical domain, using textbooks, had a lower performance instead of an improvement. The results are depicted in Table 3.

Table 3: Results on Training and Test sets.

Base Model	Corpus	Precision	Recall	F1
RB	Training	0.744	0.752	0.748
RB	Test	0.727	0.698	0.712
RBM	Training	0.719	0.743	0.731
RBM	Test	0.717	0.693	0.705

There are other models and further variations that could be tried to improve performance, this will be considered in future work.

References

1. Miranda-Escalada, A., Gascó, L., Lima-López, S., Farré-Maduell, E., Estrada, D. , Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G., and Krallinger, M. (2022) Overview of DisTEMIST at BioASQ:Automatic detection and normalization of diseases from clinical texts: Results, methods,evaluation and multilingual resources, in: Working Notes of Conference and Labs of theEvaluation (CLEF) Forum. CEUR Workshop Proceedings.

2. Miranda-Escalada, A., Farré, E., and Krallinger, M., (2020) Named entity recognition, concept normal-ization and clinical coding: Overview of the cantemist track for cancer text mining in Spanish, corpus, guidelines, methods and results., IberLEF@ SEPLN (2020) 303–323

3. Lima-López, S., Farré-Maduell, E., Gascó, L., Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G. and Krallinger, M. (2023) Overview of MedProcner task on medical procedure detection and entity linking at Bioasq 2023.

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. Castaño, J., Gambarte, L., Otero C. and Luna D. (2022) A Simple Terminology-Based Approach to Clinical Entity Recognition. Working Notes of Conference and Labs of theEvaluation (CLEF) Forum. CEUR Workshop Proceedings.

6. Borchert, F. and Schapranow M. (2022) HPI-DHC @ BioASQ DISTEMIST: Spanish Biomedical Entity Linking with Cross-lingual Candidate Retrieval. and Rule-based Reranking. Working Notes of Conference and Labs of theEvaluation (CLEF) Forum. CEUR Workshop Proceedings.