Team Fusion@SU @ BC8 SympTEMIST track: Transformer-based Approach for Symptom Recognition and Linking

Georgi Grazhdanski[1], Sylvia Vassileva[1,*], Ivan Koychev[1] and Svetla Boytcheva[1,2]

[1]FMI, Sofia University St. Kliment Ohridski, [2]Ontotext

*Corresponding author: svasileva@fmi.uni-sofia.bg

Abstract

This paper presents a transformer-based approach to solving the SympTEMIST named entity recognition (NER) and entity linking (EL) tasks. For NER, we fine-tune a RoBERTa-based (1) token-level classifier with BiLSTM and CRF layers on an augmented train set. Entity linking is performed by generating candidates using the cross-lingual SapBERT XLMR-Large (2), and calculating cosine similarity against a knowledge base. The choice of knowledge base proves to have the highest impact on model accuracy.

Introduction

SympTEMIST is a shared task and dataset for detecting and normalizing symptoms, signs, and findings in Spanish clinical texts (3). It consists of three subtasks – named entity recognition (subtask 1), entity linking (subtask 2), and multilingual entity linking (subtask 3)[1]. Our team participated in subtasks 1 and 2. Previous challenges for named entity recognition in Spanish clinical texts have shown that using a Spanish RoBERTa (4) classifier with a CRF head on top achieves very good results – 79.85% (5) and 75.68% F1 score (6). Further, adding a BiLSTM layer also improves the performance and achieves 79.46% F1 on NER for clinical procedures in Spanish (5). For the task of entity linking, using cosine similarity search with cross-lingual SapBERT (2) embeddings is a common approach utilized by the top three teams in the MedProcNER challenge[2] (6, 7, 8).

Data

SympTEMIST Dataset

The SympTEMIST corpus (3) contains 1,000 clinical case reports in Spanish annotated with symptom mentions and normalized to SNOMED-CT codes. The corpus consists of a fully annotated train set, a smaller test set, as well as a gazetteer of SNOMED-CT codes and different aliases. The train set consists of 750 documents, 11,899 sentences, and 343,243 tokens. The test set has 250 documents, 3,986 sentences, and 114,536 tokens. The train set contains 3,484 annotated entities (2,438 unique), with 1,534 unique entity codes. 59 mentions have no SNOMED-CT code assigned, the rest have a single corresponding code. There is 1 nested

[1] https://temu.bsc.es/symptemist/
[2] https://temu.bsc.es/medprocner/

mention. The Spanish SympTEMIST gazetteer contains a total of 164,817 aliases for terms in multiple categories, including findings, disorders, morphologic abnormalities, and others.

UMLS Language Pre-training Dataset and Knowledge Base

To evaluate the effect of further pre-training, we compile an additional dataset consisting of Spanish UMLS (9) synonyms of all terms included in the SympTEMIST gazetteer, for a total of 337,039 aliases. We use that as a language pre-training dataset.
We also use UMLS to compile a knowledge base using all aliases in UMLS Spanish SNOMED-CT which correspond to codes found in the SympTEMIST gazetteer, combined with the gazetteer itself and the train set for subtask 2. The knowledge base consists of 289,734 aliases.

Methods

Subtask 1: Named Entity Recognition

Clinical report texts are split into sentences using the SPACCC Sentence Splitter[3], as about 33% exceed the 512 input token limit of the employed models. For the NER subtask we perform token classification following the IOB2 annotation scheme (10), using a transformer-based model with a linear layer and a conditional random field (CRF) on top of the last transformer layer. We also experiment with adding a two-layer BiLSTM before the linear layer.

In addition to sentence splitting, we apply some modest data augmentation, by replacing entity mentions with synonyms from the Spanish UMLS, for an additional 1,672 sentences.

Classification Model Selection

We experiment with the following base models for the token classifier – PlanTL-GOBES/roberta-base-biomedical-clinical-es (Spanish RoBERTa) (4) and CLIN-X-ES (11). The first model is a RoBERTa-based language model, trained on a large Spanish biomedical-clinical combined corpus of more than 1B tokens, containing data from Medical crawler, Scielo, Wikipedia Life Sciences, mespen Medline, PubMed and others[4]. Systems based on this model have achieved competitive results on previous Spanish biomedical-clinical tasks. We further pre-train the model on the UMLS Language Pre-training Dataset. The second model, CLIN-X-ES, is a cross-lingual language model, based on XLM-RoBERTa (large), pre-trained on the MeSpEN (12) dataset, and Spanish clinical documents from the Scielo archive[5]. No additional pre-training is done.

Language Model Pre-training

We evaluate the effect of further pre-training the base transformer model using the UMLS Language Pre-training Dataset and the PlanTL-GOB-ES/roberta-base-biomedical-clinical-es model with the masked language modeling objective for 4 epochs. Hyperparameter values are the same as those used for pre-training RoBERTa (base) in the original RoBERTa paper (1).

[3] SPACCC Sentence Splitter
[4] https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es
[5] Scielo archive

Subtask 2: Entity Linking

The entity linking task uses the gold entities provided by the organizers and aims to predict the correct SNOMED-CT code. For this task, all entities corresponding to more than one code (composite), are removed from the train and test sets. Furthermore, not all entity mentions have a corresponding code in SNOMED-CT (NO_CODE). There are 59 such instances in the train set. The system performs linking in two steps – first, we try to match the mention to an alias in the knowledge base (KB), and second – we use cosine similarity search on the SapBERT XLMR-Large (2) embeddings and retrieve the closest alias in the KB.

Experiments and Results

Subtask 1: Named Entity Recognition

We split the train set and use 80% for training and 20% for validation. Micro-averaged precision, recall, and F1-score are used as metrics for the NER subtask.

Model	Val P	Val R	Val F1	Test P	Test R	Test F1
Aug. Spanish RoBERTa+CRF	**0.773**	0.729	**0.750**	0.732	0.718	0.725
Aug. Spanish RoBERTa+BiLSTM+CRF	0.744	**0.732**	0.738	**0.739**	**0.725**	**0.732**
Pre-trained Aug. Spanish RoBERTa+CRF	0.749	0.721	0.735	0.715	0.720	0.718
CLIN-X-ES+CRF	0.757	0.717	0.737	0.718	0.703	0.710
Aug. CLIN-X-ES+CRF	0.722	0.704	0.713	0.724	0.699	0.712

Table 1: Subtask 1 results on the validation and test sets.

Table 1 presents the results of the different model and fine-tuning scheme combinations. Models based on the Spanish RoBERTa perform best, likely due to the fact that it is specialized for the Spanish biomedical-clinical domain. The addition of a 2-layer BiLSTM increases both recall and precision, perhaps because of its ability to consider long-term dependencies (5).

Effect of Data Augmentation

After the training dataset is split into sentences, it is augmented by randomly replacing some of the annotated mentions with a synonym from the Spanish UMLS. This results in 1,672 additional sentences (13,571 in total). Table 2 compares model performance. Fine-tuning on the augmented train set significantly improves the precision and F1 of the Spanish RoBERTa model, despite the decrease in recall. Interestingly, we observe a performance drop in the augmented CLIN-X-ES model compared to its non-augmented version on the validation set. However, on the test set, the two models are close in terms of F1.

Model	Val P	Val R	Val F1	Test P	Test R	Test F1
Spanish RoBERTa + CRF	0.730	**0.746**	0.738	-	-	-
Augmented Spanish RoBERTa + CRF	**0.773**	0.729	**0.750**	**0.732**	**0.718**	**0.720**
CLIN-X-ES + CRF	0.757	0.717	0.737	0.718	0.703	0.710
Augmented CLIN-X-ES + CRF	0.722	0.704	0.713	0.724	0.699	0.712

Table 2: Effect of data augmentation in the NER subtask.

Subtask 2: Entity Linking

We perform experiments with various knowledge bases by combining the different resources gazetteer, train set, and the UMLS synonym dataset. Furthermore, we generate augmentation to the KB by adding/removing random characters in the aliases of rare concepts with less than 5 records. For validation purposes, we use the full train set and exclude these aliases from the KB. The entity linking subtask uses accuracy as a metric to evaluate the performance.

In addition to using the SapBERT XLMR-Large embeddings, we perform one experiment that aims to tackle the problem of long entity mentions. For each code, we determine the final score as a linear combination of its cosine similarities to the full mention, the first 75% of tokens in the mention, and the last 75% of tokens in the mention, and we select the code with the highest score. Using the train set, we find the optimal coefficients with grid search for each part to be 0.75, 0.17, 0.08. The sliding window performs about 2% better than the basic SapBERT-XLMR Large model using the same KB, which suggests that the information needed to find the correct code is more focused in the first part of the mention.

The results from entity linking experiments are shown in table 3. The best model shows 58.9% accuracy on the test set and has the richest knowledge base which includes additional data from UMLS. Most of the models show close results in the range of 56-58% accuracy. Due to a bug in the code for generating the knowledge base, the final experiment shows drastically lower results of about 1%.

Model	Knowledge Base	Val Acc	Test Acc
SapBERT XLMR-Large	Gazetteer + Train	0.514	0.588
SapBERT XLMR-Large	Gazetteer + Train + Aug.	0.533	0.565
SapBERT XLMR-Large	Gazetteer + Train + UMLS	0.524	**0.589**
SapBERT XLMR-Large+Sliding Window	Gazetteer + Train + Aug.	**0.536**	0.587
SapBERT XLMR-Large	Gazetteer + Train + UMLS	0.510	0.017

Table 3: Subtask 2 results on the validation and test sets.

Conclusion

We explore transformer-based approaches to solving the SympTEMIST named entity recognition and linking tasks. For NER, systems based on a specialized monolingual model achieve the best results. The addition of a BiLSTM layer after the last transformer layer, and train data augmentation improves performance on the test set. We employ SapBERT XLMR-Large exclusively for the entity linking subtask. The choice of a knowledge base has the highest impact on system performance – our highest accuracy model combines the SympTEMIST gazetteer, UMLS synonyms, and train set annotations. Augmenting the knowledge base to include slightly modified versions of rare mentions leads to a stable improvement on the validation set, but has no impact on performance on the test set, so this approach could be explored further.

Funding

This work was supported by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No BG-RRP-2.004-0008.

References

1. Liu,Y., Ott,M., Goyal,N., et al. (2019) Roberta: A robustly optimized bert pretraining approach. https://doi.org/10.48550/arXiv.1907.11692.

2. Liu,F., Vulić,I., Korhonen,A., et al. (2021) Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACLIJCNLP 2021*, pages 565–574.

3. 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. In: *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

4. Carrino,C.P., Armengol-Estapé,J., Gutiérrez-Fandiño,A, et al. (2021) Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a midresource scenario. https://doi.org/10.48550/arXiv.2109.03570.

5. Almeida,T., Jonker, R. A. A. , Poudel, R, et al. (2023) Discovering medical procedures in spanish using transformer models with mcrf and augmentation. In *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*.

6. Chizhikova,M., Collado-Montañez,J., Díaz-Galiano,M., et al. (2023) Coming a long way with pre-trained transformers and string matching techniques: Clinical procedure mention recognition and normalization. In *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*.

7. Zotova,E., García-Pablos,A., Cuadros,M., et al. (2023) Vicomtech at medprocner 2023: Transformers-based sequence-labelling and cross-encoding for entity detection and normalisation in spanish clinical texts. In *Working Notes of CLEF 2023 Conference and Labs of the Evaluation Forum*.

8. Vassileva,S., Grazhdanski,G., Boytcheva,S, et al. (2023) Fusion @ bioasq medprocner: Transformer-based approach for procedure recognition and linking in spanish clinical text. In *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*.

9. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70. doi: 10.1093/nar/gkh061. PubMed PMID: 14681409; PubMed Central PMCID: PMC308795.

10. Krishnan,V. and Ganapathy,V. (2005) Named entity recognition.

11. Lange,L., Adel,H., Strötgen,J, et al. (2022) iCLIN-x/i: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain. Bioinformatics, 38(12):3267–3274.

12. Villegas,M., Intxaurrondo,A., Gonzalez-Agirre,A, et al. (2018) The mespen resource for english-spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. In LREC MultilingualBIO: Multilingual Biomedical Text Processing, pages 32–39. ELRA.