HPI-DHC @ BC8 SympTEMIST Track: Detection and Normalization of Symptom Mentions with SpanMarker and xMEN

Florian Borchert,* Ignacio Llorca and Matthieu-P. Schapranow

Hasso Plattner Institute for Digital Engineering, University of Potsdam, Germany

*Corresponding author: Tel: +49 (331) 5509-4839, E-Mail: florian.borchert@hpi.de

Abstract

Signs and symptoms of patients are frequently reported in clinical text documents. Therefore, accurate automated extraction of symptom information is essential for their integration into downstream clinical applications. In this work, we describe our contribution to the BioCreative VIII SympTEMIST shared task, a benchmark for the detection and normalization of symptom mentions in Spanish-language clinical case reports.
Our systems for subtasks 1 and 2 are built upon two state-of-the-art, open-source information extraction tools: (1) SpanMarker for named entity recognition with document-level context and (2) xMEN for normalizing symptom mentions to their corresponding SNOMED CT code.
For subtask 1, our best submitted run achieves an F_1 score of 0.7363, which exceeds the median across all submissions by more than 3pp. Our experiments underline the positive impact of including document-level context for named entity taggers. For subtask 2, our best system for entity normalization obtains an accuracy of 0.6070, an improvement of more than 8pp over the median.

Introduction

Accurate and detailed information about signs and symptoms play a critical role in medical diagnosis, treatment planning, and phenotyping (1). However, this information is frequently recorded in clinical free-text documents only and needs to be extracted and normalized to enable the semantic interoperability required for many downstream applications (2). The BioCreative VIII SympTEMIST shared task (3) provides a benchmark dataset of clinical case reports in Spanish with gold-standard annotations of symptom mentions and corresponding SNOMED CT codes (4). In this work, we present our systems for subtasks 1 and 2 of the shared task, i.e., for named entity recognition (NER) and entity linking (EL).
Our work builds upon our experience in the DisTEMIST shared task (5,6). Here, a standard Transformer-based token classification approach was the most successful strategy applied by most participants in the NER track. For subtask 1 of the current SympTEMIST shared task, we extend this line of work by incorporating document-level context to improve NER performance. To this end, we use the recently released SpanMarker library.[1] For subtask 2, we use the xMEN toolkit for cross-lingual medical entity normalization (7), combining unsupervised cross-lingual candidate retrieval and a trainable Transformer-based cross-encoder for re-ranking.

[1] https://github.com/tomaarsen/SpanMarkerNER

Material and Methods

In this section, we describe our methods for NER and EL. The source code to reproduce our experimental results is available online.[2] All Transformer-based models are initialized from the checkpoint *PlanTL-GOB-ES/roberta-base-biomedical-clinical-es,* a Spanish biomedical-clinical RoBERTa model (8).

Dataset

The SympTEMIST corpus consists of 1,000 annotated case reports in Spanish, a subset of which is available for training and evaluation. We implement a data loader for SympTEMIST within the BigBIO (9) framework. This enables a seamless integration with the open-source tools SpanMarker and xMEN. For both subtasks, we use 20% of the respective training documents as an internal validation set for model selection.

Named Entity Recognition

For subtask 1, we rely on SpanMarker, a Transformer-based framework designed to enhance NER performance by incorporating document-level context. To this end, we pre-process all documents through sentence-splitting and tokenization using spaCy and the model *es_core_news_sm*[3]. The adjacent sentences in a document, up to the maximum input length of the RoBERTa model, are used as contextual information for each sentence. Labeled spans are converted to IO-tags.

All considered models are trained with document-level context for 30 epochs, with a batch size of 32. Furthermore, we increase the maximum entity length from the default of eight to 15 tokens, to account for the large number of long annotations in the corpus. Thus, entities longer than 15 tokens are not passed to the model at training.

During inference, the input can also be provided with or without document-level context. In our experiments, we conduct a comparative analysis of the two methods, considering three different values for the learning rate (1×10^{-5}, 5×10^{-5}, and 9×10^{-5}) and two for weight decay (0.0 and 0.1). We select the optimal configurations on our internal validation set for submission.

Entity Linking

For subtask 2, we use the open-source framework xMEN, configured with a target knowledge base constructed from the SympTEMIST gazetteer (121,760 concepts), extended by Spanish and English aliases from the Unified Medical Language System (version 2023AB). This way, we obtain more than 1.08M aliases.

We use an ensemble of a TF-IDF vectorizer (10) and SapBERT (11) for unsupervised candidate generation, followed by a cross-encoder model for re-ranking. The cross-encoder is trained for 20 epochs, using batches of 64 candidates with ground truth labels and a softmax loss with rank regularization (7). We keep the checkpoint that maximizes accuracy@1 on the validation set. For our submitted runs, we consider predictions with and without the NIL (not-in-list) option. This is implemented by optionally including an additional synthetic NIL example in the candidate list, which is considered the correct candidate when the actual ground truth concept is not part of the training batch. We also evaluate different classification thresholds for abstaining from predictions, which allows for different trade-offs between precision and recall.

---

[2] https://github.com/hpi-dhc/symptemist_biocreative_2023/

[3] https://spacy.io/

Finally, we observe that all unique mention strings in the gold standard have been mapped to the same SNOMED CT code. Therefore, we perform an exact lookup of codes in the training set as a last post-processing step. This approach resulted in a small performance improvement in previous work (6), although the majority of such cases is covered through the supervised training of the cross-encoder model.

Results

In this section, we describe our experimental results.

Subtask 1: SymptomNER

The test set results for subtask 1 are shown in Table 1. Including document-level context substantially improves performance in terms of F_1 score (+4pp), mostly due to dramatically increased recall. The impact of the other hyperparameters is less pronounced. Ultimately, the best performance is attained with a learning rate of 1×10^{-5} and employing the final model checkpoint after 30 epochs of training, instead of the checkpoint that performed best on the validation set.

				Test Set		
Run	**Context**	**Checkpoint**	**Learning Rate**	**Precision**	**Recall**	**F_1 score**
1	yes	best	1×10^{-5}	0.7667	0.6995	0.7299
2	yes	last	1×10^{-5}	**0.7707**	**0.7049**	**0.7363**
3	yes	best	5×10^{-5}	0.7586	0.6956	0.7257
4	no	best	1×10^{-5}	0.7673	0.6108	0.6802
5	no	last	1×10^{-5}	0.7675	0.6189	0.6852
All teams (mean)				–	–	0.61
All teams (median)				–	–	0.7

Table 1: Results for subtask 1 (SymptomNER)

Subtask 2: SymptomNorm

For subtask 2, the results for our internal validation set and the official held-out test set are shown in Table 2. Considering the validation set results, different trade-offs in terms of precision and recall can be achieved through considering NIL options and decision thresholds. However, for the shared task, *accuracy* was the main metric, which is equivalent to recall@1 given that no concept has multiple ground-truth codes. Therefore, increased precision did not count towards the evaluation and making a prediction for each mention was the best strategy. Consequently, our second run achieved the best performance in terms of test set accuracy.

Run	NIL Option	Threshold	Validation Set			Test Set
			Precision	Recall	F_1 score	Accuracy
1	yes	–	0.7067	0.6080	0.6536	0.5848
2	no	–	0.6261	**0.6261**	0.6261	**0.6070**
3	yes	0.02	**0.8924**	0.5368	0.6704	0.5265
4	no	0.02	0.8919	0.5446	**0.6763**	0.5321
All teams (mean)			–	–	–	0.21
All teams (median)			–	–	–	0.52

Table 2: Results for subtask 2 (SymptomNorm)

Discussion and Evaluation

Obtaining better performance on the NER subtask when utilizing document-level context aligns with our assumptions. Our choice of using 15 tokens as the maximum entity length attempts to accommodate lengthy annotations, but we did not systematically optimize this hyperparameter.
For the EL subtask, the default xMEN pipeline achieves competitive performance, exceeding the median across all teams by more than 8pp accuracy. However, the absolute recall is only 60.7% and thus far below values obtained in other biomedical EL tasks, such as the Quaero corpus (7,12). We note that our system is particularly accurate for normalizing shorter entity spans. For mentions consisting of a single word, the validation set recall is 90.0%. In contrast, the recall for length 2 mentions is 72.4%, and only 42.3% for mentions with 3 or more tokens. This does not even account for the problem of multi-normalization with composite mention spans, which have not been considered in the evaluation.

Conclusion and Outlook

In this work, we described our contribution to the SympTEMIST shared task. For both the NER and EL subtasks, state-of-the-art, Transformer-based open-source tools achieve competitive performance with relatively little adaptation. In the future, we would like to improve NER and EL performance for composite entities and other long mention spans. We will also explore multilingual normalization across the wide range of languages covered by the automatically translated versions of SympTEMIST.

Funding

This work was supported by a grant from the German Federal Ministry of Research and Education [01ZZ2314N].

References

1. Steinkamp, J.M., Bala, W., Sharma, A., et al. (2020) Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes. *J. Biomed. Inform.*, **102**, 103354.
2. Lehne, M., Sass, J., Essenwanger, A., et al. (2019) Why digital medicine depends on interoperability. *npj Digit. Med.*, **2**, 1–5.
3. 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.*
4. Donnelly, K. (2006) SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud. Health Technol. Inform.*, **121**, 279–290.
5. Miranda-Escalada, A., Gascó, L., Lima-López, S., et al. (2022) Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.
6. Borchert, F., Llorca, I. and Schapranow, M.-P. (2023) Cross-Lingual Candidate Retrieval and Re-ranking for Biomedical Entity Linking. In Arampatzis, A., Kanoulas, E., Tsikrika, T., et al. (eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer Nature Switzerland, Cham, pp. 135–147.
7. Borchert, F., Llorca, I., Roller, R., et al. (2023) xMEN: A Modular Toolkit for Cross-Lingual Medical Entity Normalization. *arXiv [cs.CL]*, 2310.11275.
8. Carrino, C.P., Llop, J., Pàmies, M., et al. (2022) Pretrained Biomedical Language Models for Clinical NLP in Spanish. *Proceedings of the 21st Workshop on Biomedical Language Processing*, Association for Computational Linguistics, Dublin, Ireland, pp. 193–199.
9. Fries, J., Weber, L., Seelam, N., et al. (2022) BigBIO: a framework for data-centric biomedical natural language processing. *Adv. Neural Inf. Process. Syst.*
10. Neumann, M., King, D., Beltagy, I., et al. (2019) ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Proceedings of the 18th BioNLP Workshop and Shared Task*, Association for Computational Linguistics, Florence, Italy, pp. 319–327.
11. Liu, F., Vulić, I., Korhonen, A., et al. (2021) Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Association for Computational Linguistics, Online, pp. 565–574.
12. Wajsbürt, P., Sarfati, A. and Tannier, X. (2021) Medical concept normalization in French using multilingual terminologies and contextual embeddings. *J. Biomed. Inform.*, **114**, 103684.