Transformer-based multilabel NER using Wikipedia corpora in multiple languages

Yelyzaveta Ahapova, Johann Frei, Frank Kramer

Angaben zur Veröffentlichung / Publication details:

Ahapova, Yelyzaveta, Johann Frei, and Frank Kramer. 2025. "Transformer-based multilabel NER using Wikipedia corpora in multiple languages." In Intelligent health systems – from technology to data and knowledge: proceedings of MIE 2025, edited by Elisavet Andrikopoulou, Parisis Gallos, Theodoros N. Arvanitis, Rosalynn Austin, Arriel Benis, Ronald Cornet, Panagiotis Chatzistergos, et al., 878–79. Amsterdam: IOS Press. https://doi.org/10.3233/shti250488.

Intelligent Health Systems – From Technology to Data and Knowledge
E. Andrikopoulou et al. (Eds.)

doi:10.3233/SHTI250488

Transformer-Based Multilabel NER Using Wikipedia Corpora in Multiple Languages

Yelyzaveta AHAPOVA [a,1], Johann FREI [a] and Frank KRAMER [a]
[a] *IT-Infrastructure for Translational Medical Research, University of Augsburg, Germany*
ORCiD ID: Johann Frei https://orcid.org/0000-0003-0323-0904, Frank Kramer https://orcid.org/0000-0002-2857-7122

Abstract. The high cost of manual data labeling and privacy concerns result in a considerable dearth of medical annotations in non-English texts. Recent work by Frank and Kramer [1] introduces an unsupervised approach for constructing an ontology-annotated corpora from Wikipedia[2] for German medical NER. We evaluate the proposed approach across English, German, Spanish, and French for medication and diagnosis entity recognition. Our multilabel corpora yield notable improvements in German medication detection under sparse annotations compared to the baseline, with consistent performance across other languages.

Keywords. natural language processing, named entity recognition, medical NER, medication extraction, diagnosis extraction

1. Introduction

Named Entity Recognition (NER) refers to the task of identifying specific entities in text and assigning them to predefined classes [2]. In healthcare, NER is used to extract clinical entities from electronic health records (EHRs). The scarcity of open-source medical NER models remains a major challenge due to the limited availability of clinical texts for training such models. Leveraging ontology structures addresses this issue and enables the creation of a large, domain-specific dataset [1]. This study aimed to assess the performance of a transformer-based multilabel NER model on Wikipedia-sourced corpora in English, German, Spanish, and French, using gold-standard clinical datasets. The results were promising across all languages, even in conditions with sparse annotations.

2. Methods

We utilized Wikipedia and Wikidata[3] to generate language-specific corpora by extracting sentences containing medication and diagnosis labels using a custom-defined SPARQL ruleset and a web tool[4] from the baseline work. To fine-tune and evaluate NER mod-

[1]Corresponding Author: Yelyzaveta Ahapova; E-mail: yelyzaveta.ahapova@uni-a.de
[3]https://www.wikidata.org
[4]https://ontocorpus.misit-augsburg.de/

els across languages in multilabel setup, we adapted the open-source pipeline from the same baseline work. The NER training consisted of two stages: initial training on generic, monolingual models with dynamic token scaling, followed by domain-specific fine-tuning on imputed datasets without scaling. Evaluation datasets for each language were selected based on the availability of relevant entities and represent a wide range of sub-domains.

3. Results

The evaluation results presented in Tables 1 and 2 are divided based on the availability of drug and diagnosis label classes within the external datasets.

Table 1. Evaluation metrics only for drug labels.

test	dataset	Pr	Re	$F1_{drug}$	$F1_{baseline}$
en	ADE-Corpus-v2	0.7738	0.8562	0.8129	-
de	GERNERMED++	0.9004	0.6163	0.7317	0.7999
	CARDIO:DE	0.7279	0.6556	**0.6898**	0.6447
es	PharmaCoNER	0.7011	0.6528	0.6761	-
	DrugTEMIST	0.5983	0.8595	0.7055	-

Table 2. Evaluation metrics for drug and diagnosis labels. Overall F1-score is computed as micro-scores for both labels.

test	dataset	$F1_{drug}$	$F1_{diag}$	$Pr_{overall}$	$Re_{overall}$	$F1_{overall}$	$F1_{baseline}$
de	GPTNERMED	0.8175	0.6734	0.8881	0.6689	0.7631	0.8545
	BRONCO150	0.7300	0.7020	0.8477	0.6104	0.7097	0.7792
	GGPOnc 2	**0.5564**	0.5397	0.6000	0.4962	0.5432	0.3552
es	CT-EBM-SP	0.7500	0.6360	0.8851	0.5612	0.6869	-
fr	QUAERO-MEDLINE	0.5740	0.7114	0.7028	0.6419	0.6709	-

Multilabel NER demonstrates clear advantages for sparse label frequency corpora, as evidenced by the increased F1-scores on both the GGPOnc 2 and CARDIO:DE datasets with scarce annotations compared to the baseline. Expanding the range of label categories improved model adaptability and minimized reliance on extensive annotations, while a carefully crafted annotation ruleset ensured more consistent entity coverage. Additionally, the results also remain promising across non-German languages, given the diversity in the evaluation datasets.

4. Discussion and Conclusions

Despite the growing popularity of Large Language Models (LLMs), the unsupervised transformer-based approaches used in this study perform notably well, especially due to their resource efficiency and dedicated architecture for NER tasks. We demonstrated the potential of leveraging ontology-annotated corpora for medical NER in sparse, non-English settings. By integrating Wikidata-defined label classes through a custom annotation scheme, this approach provides broad applicability across a range of clinical and other domain-specific areas. Furthermore, the resulting models can be easily replicated without privacy concerns or restrictions on model sharing, as the methodology relies solely on publicly available open-source data.

References

[1] Frei J, Kramer F. Creating Ontology-annotated Corpora from Wikipedia for Medical Named-entity Recognition. In: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing. Bangkok, Thailand: Association for Computational Linguistics; 2024. p. 570-9.

[2] Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investigationes. 2007;30(1):3-26.