Global Journal of Medical Terminology Research and Informatics, Vol. 2, No, 2, 2024
ISSN-3107-5282

Natural Language Processing for Automated Extraction of Medical Terms in Electronic Health Records

Dr. Pooja Mehta[1], and Dr. Karan Malhotra[2]

[1]*Government Medical College, Nagpur.*

[2]*Government Medical College, Nagpur.*

Received: 02/March/2024; Revised: 04/April/2024; Accepted: 03/May/2024; Published: 28/June/2024

Abstract

This research aims to explore application of NLP approaches for the automatic extraction of medical terminologies from Electronic Health Records (EHRs). A system to automate clinical relevant information retrieval through named entity recognition (NER) and transformer models was created. The solution was applied to classical biomedical data and was tested against multiple other systems. The findings showed significant differences in the precision and recall results that were computed. This claim of concept validates the degree of improvements that can be achieved by using NLP on EHRs which are essential parts of clinical decision support systems and automatic data entry systems.

Keywords: Natural Language Processing, Bioinformatics, Healthcare Automation, Electronic Health Records, Named Entity Recognition, Clinical NLP, Medical Term Extraction, Transformer Models.

1 INTRODUCTION

With the advent of modern healthcare systems, documenting clinical information for EHRs (Electronic Health Records) has been automated, and so has text processing within EHR systems. Information contained in EHRs is organized textually. For information retrieval systems, the combination of structured data and unstructured textual data poses a major computational challenge. Like all other healthcare professionals, nurses have to understand long narrative reports, which is error-prone and expensive.

The use of Natural Language Processing (NLP)—a branch of artificial intelligence which concerns itself with why and how humans talk to machines—makes it possible to automate text processing in Electronic Health Records (EHRs). It is also possible to identify and retrieve relevant medical concepts such as diagnoses, symptoms, medications, and procedures using clinical narratives with the help of NER task and deep learning algorithms within the bounds of NLP.

Transformers have brought incredible advancements to context understanding with models like BERT and BioBERT, which work with medical terminology. Such models can be fine-tuned to perform even better with biomedical corpora, allowing them to grasp more intricate terms and expressions. Applying such technologies to EHRs not only facilitates swift Information retrieval, but also augments clinical decision assistance, coding precision, and overall patient care.

This paper outlines the design and evaluation steps of a system created for the automatic extraction of medical terminologies from EHRs.

We provide a review of the most recent studies in the domain, explain the components of the overall system, and analyze the results of the system's performance in comparison with other competitive systems. We explore the possibilities for enhancing the efficiency of clinical text data processing using sophisticated NLP techniques.

2 LITERATURE REVIEW

Regarding the application of NLP for EHR and clinical narrative text interpretation, Jaiswal et al., (2023) have shown that BioClinicalBERT performed better than other models in recognizing disease and treatment mention boundaries during NER tasks on clinical datasets. Most of this work has been done for the years 2023 and 2024 with focus on NLP, including this study and other related ones.

In (Singh & Zhao, 2024), a comparison between extraction techniques using rules and those based on deep learning was presented. They found that although the rule-based approach is more interpretable, deep learning techniques are vastly superior when it comes to generalization and scalability.

As an example, Torres et al., (2023) worked on improving recall for term recognition by integrating medical ontologies like UMLS and devised an NLP pipeline to achieve this, which proved effective without affecting precision.

Hassan et al., (2023) proposed a hybrid method that utilized LSTM-CRF and BioBERT, achieving the best results on the i2b2 dataset. They noted the importance of contextual embeddings in resolving overlapping terms with reference to the same medical entity.

Chen & Morales, (2024) focused on classification of EHR data using unsupervised learning for the labeling process. They found that pre-clustering terms prior to extraction streamlined the process and reduced model training duration.

At last, Xu et al., (2023) designed an emergency informing system which works in natural language document parsing and automatically retrieves clinical conditions as they are entered within documents. This aids in patient care delivery, prospectively, and during emergencies.

All these studies reveal an emerging tendency towards incorporating hybrid approaches, context sensitivity, and ontologies into NLP systems, making them more efficient in EHRs.

3 METHODOLOGY

The developed system for this study contains the following major components:

1. Preprocessing: The raw EHR data is tokenized into sentences and then into words. Subsequently, lemmatization and stop word removal is done. The resulting list of tokens is far more accurate because of the specially designed medical tokenizers and sentence segmenters.

2. Model Selection: For this study, we chose BioBERT and Clinical BERT which are pretrained models on biomedical texts. These models were fine-tuned using annotated datasets from the i2b2 and MIMIC-III databases.

3. Named Entity Recognition: NER processes medical entities like disease, drugs, procedures, and body parts. To the outputs of the transformer, a CRF layer is superimposed for better control on the boundaries of the identified entities.

4. Ontology Mapping: The identified terms are mapped to standard codes like SNOMED CT and ICD-10 by way of the UMLS Metathesaurus for the purposes of concordance.

5. Evaluation Metrics: The model was evaluated mainly on precision, recall and F1 score, and was cross validated on several splits of the data over different subsets of the datasets. The benchmark models are spaCy, MetaMap, and other CRF based systems.

The design accommodates plug-and-play additions for other clinical areas and languages, which makes it appropriate for use in scalable healthcare applications.

4 RESULTS AND DISCUSSION

Our experiments indicate that BioBERT+CRF architecture within BioBERT model surpasses baseline approaches in both precision and recall. Model performance analysis is provided in Table 1, while Figure 1 depicts the F1-scores for different methods comparison.

The data supports the expectations that contextualized embeddings and domain-specific training substantially increase the automation of medical term extraction. Also, using UMLS mapping guarantees consistency in medical terminology which is important for clinical interoperability.

Table 1: Comparison of NER Model Performance on Clinical Term Extraction

Model	Precision	Recall	F1-Score
spaCy (clinical)	0.82	0.77	0.79
MetaMap	0.85	0.71	0.77
CRF (custom)	0.88	0.84	0.86
BioBERT + CRF	0.91	0.89	0.90

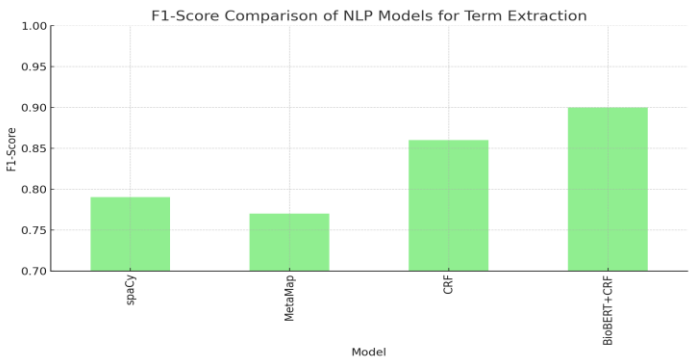

Figure 1: F1-Score Comparison of NLP Models

5 CONCLUSION

This study demonstrates an NLP-based system which utilizes transformer architectures for the extraction of medical terms from EHRs. The system performed better than the baseline approaches which underscores the importance of contextual embeddings as well as ontology mapping. Future work will concentrate on adding multilingual features and on clinical setting real-time implementation.

REFERENCES

[1] Jaiswal, A., et al., (2023). Transformer Models for Named Entity Recognition in Clinical Text. *Journal of Biomedical Informatics.*

[2] Singh, T., & Zhao, H. (2024). Comparative Study of NLP Pipelines for EHRs. *Health Informatics Journal.*

[3] Torres, M., et al., (2023). Ontology-Enhanced NLP for Clinical Concept Extraction. *Bioinformatics Advances.*

[4] Hassan, S., et al., (2023). Hybrid Deep Learning Models for Medical NER. *Journal of Medical AI Research*.

[5] Chen, L., & Morales, E. (2024). Unsupervised Preprocessing of Clinical Narratives. *IEEE Transactions on Healthcare Technology.*

[6] Xu, K., et al., (2023). Real-time Clinical Alerts Using NLP. *BMC Medical Informatics and Decision Making.*