ISCSITR - INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN HEALTHCARE INFORMATION SYSTEM (ISCSITR - IJSRHIS)

Vol.4, Iss. 1, January-June, 2023, pp. 1-7.

https://iscsitr.com/index.php/ISCSITR-IJSRHIS

Journal ID: 2351-4421

Leveraging Natural Language Processing for Automated Extraction of Patient Information from Unstructured Clinical Notes in Electronic Health Records

Kiran Acharya,

Nepal.

Abstract

Unstructured clinical notes embedded within Electronic Health Records (EHRs) hold critical insights for patient care and decision-making. However, the narrative nature of these notes limits the ease of data retrieval, integration, and real-time analytics. In this paper, we explore the application of Natural Language Processing (NLP) techniques for the automated extraction of structured patient information from unstructured clinical notes. Framed in the 2022 context, where EHR adoption and AI tools have matured, this paper evaluates NLP pipelines involving Named Entity Recognition (NER), clinical ontologies (UMLS, SNOMED CT), and machine learning models.

We present both a conceptual architecture and a proof-of-concept system trained on MIMIC-III datasets. The NLP pipeline uses hybrid rule-based and deep learning components to extract diagnoses, medication events, and temporal relationships. Evaluations show a notable improvement in precision and recall compared to previous heuristic systems. This automation holds promise in clinical decision support, population health research, and administrative documentation

Keywords: Natural Language Processing, Clinical Notes, EHR, Information Extraction, Medical AI, Named Entity Recognition, Clinical Ontology**.**

How to cite this paper: Kiran Acharya. (2023) Leveraging Natural Language Processing for Automated Extraction of Patient Information from Unstructured Clinical Notes in Electronic Health Records. *ISCSITR - INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN HEALTHCARE INFORMATION SYSTEM (ISCSITR - IJSRHIS)*, 4(1), 1–7.

URL: https://iscsitr.com/index.php/ISCSITR-IJSRHIS/article/view/ISCSITR-IJSRHIS_04_01_001
Published: 24th Apr 2023

1. Introductıon

1.1 Background and Significance

Modern hospitals generate immense volumes of unstructured text in the form of discharge summaries, progress notes, and radiology reports. These free-text narratives are vital for contextual understanding but pose significant challenges for downstream clinical and analytical tasks. Manual chart reviews are time-consuming, error-prone, and inefficient. Automating the extraction of meaningful patient data from unstructured sources would transform EHR usability.

Natural Language Processing (NLP), a subfield of AI, enables machines to interpret human language. In healthcare, it provides an opportunity to extract key clinical concepts—symptoms, diagnoses, medications—from unstructured text. By transforming narrative notes into structured data, NLP enables real-time analytics, cohort identification, and supports clinical decision-making with greater precision.

1.2 Objectives

This paper presents a comprehensive overview and implementation of an NLP system tailored for clinical text extraction from EHRs. Our objectives are:

- To design an NLP pipeline that leverages pre-trained medical models and domain-specific ontologies.
- To evaluate system performance on standard benchmark datasets (e.g., MIMIC-III).
- To compare modern transformer-based models with traditional rule-based approaches in extracting clinically relevant entities.

2. Literature Review

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics*, 17(1), 128–144.

2.1 Foundations of Clinical NLP

Prior to 2022, significant progress had been made in applying NLP to clinical records. Early systems like cTAKES (Savova et al., 2010) and MetaMap were heavily rule-based and relied on dictionaries like UMLS for mapping terms. These tools demonstrated the feasibility of automating information extraction from clinical notes, particularly for problem lists and medication records. However, they lacked scalability and adaptability across institutions and specialties.

The integration of statistical learning—support vector machines, CRFs—improved adaptability, but required extensive manual feature engineering. These methods were later supplemented by deep learning, where models like BiLSTM-CRFs and CNNs improved entity recognition tasks. However, a consistent challenge was generalization and the contextual ambiguity of clinical language, often varying in semantics and structure across notes.

2.2 Transformer Models and Deep Learning Advances

By 2020, transformer models such as BERT (Devlin et al., 2018) and domain-specific variants like BioBERT and ClinicalBERT began outperforming traditional models in many medical NLP benchmarks. ClinicalBERT, trained on MIMIC notes, achieved state-of-the-art results on concept extraction, negation detection, and temporal classification. These models could understand long-range dependencies and ambiguous phrasing common in clinical narratives.

Despite their success, limitations remained in explainability and computing overhead. Additionally, regulatory frameworks and clinician trust were still catching up. Many systems functioned best in retrospective analysis but had yet to demonstrate consistent impact in real-time clinical workflows. Hybrid models combining rule-based filters with neural models were increasingly recommended to balance precision, speed, and interpretability.

3. Methodology

3.1 System Architecture

The NLP pipeline implemented for this study follows a modular architecture, integrating:

- **Text preprocessing**: tokenization, lemmatization, stopword removal.
- **Named Entity Recognition (NER)**: leveraging BioBERT and SpaCy's med7.
- **Entity normalization**: mapping terms to SNOMED CT codes using UMLS Metathesaurus.
- **Negation detection**: identifying absent symptoms or negated events using NegEx.
- **Temporal relation extraction**: identifying sequence and duration of medical events.

3.2 Dataset and Tools

The **MIMIC-III** dataset served as the benchmark for training and validation. This dataset includes de-identified ICU patient data with rich textual content. The annotated corpus was created using 2,000 discharge summaries manually labeled for conditions, medications, procedures, and temporal indicators.

We evaluated three models:

- **Rule-based using MetaMap**
- **BiLSTM-CRF model**
- **BioBERT fine-tuned on the annotated corpus**

All experiments were conducted in Python using HuggingFace Transformers, scispaCy, and Scikit-learn.

4. Results and Discussion

4.1 Performance Evaluation

The **BioBERT** model outperformed traditional approaches across all extraction categories, especially in recognizing diagnosis and treatment entities. The hybrid model (BioBERT + rule-based filters) improved precision without significant trade-off in recall. Performance metrics are shown in the table below:

Table 1: Performance Comparison (F1-score %)

Entity Type	MetaMap	BiLSTM-CRF	BioBERT	BioBERT+Rules
Diagnosis	74.5	81.3	89.1	**91.4**
Medication	68.2	77.6	86.9	**89.5**
Procedures	70.4	75.2	83.5	**86.0**

4.2 Interpretability and Clinical Relevance

Clinicians favored the hybrid model as it provided traceable rationales (via rules) and contextual insights (via deep models). Visualization tools such as **SHAP** were used to explain token-level importance, improving trust. The pipeline also supported real-time integration via APIs, allowing easy embedding into existing EHR systems for automated cohort identification or alert generation.

5. Conclusion

This paper highlights the transformative potential of NLP for extracting structured information from unstructured clinical narratives. In 2022, with maturing models and high EHR penetration, implementing BioBERT-based pipelines with rule augmentation can enhance accuracy and usability. This shift enables faster research, smarter alerts, and improved documentation quality.

While results are promising, future work must address domain adaptation, multilingual data, and model auditing. A collaborative approach between informaticians, clinicians, and NLP experts will be key to achieving robust, fair, and generalizable solutions in the clinical setting.

References

[1] Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics*, 17(1), 128–144.

[2] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., &

Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5), 507–513.

[3] Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5), 760–772.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[5] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323.*

[6] Sohn, S., Clark, C., Halgrim, S., & Chute, C. G. (2012). MedXN: An open source medication extraction and normalization tool for clinical text. *Journal of the American Medical Informatics Association*, 21(5), 858–865.

[7] Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association, 11(5), 392–402.

[8] Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 18(5), 552–556

[9] Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., & Uzuner, Ö. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association, 18(5), 540–543.

[10] Liu, H., Bielinski, S. J., Sohn, S., Murphy, S., Wagholikar, K. B., Jonnalagadda, S., ... & Chute, C. G. (2013). An information extraction framework for cohort identification using electronic health records. AMIA Summits on Translational Science Proceedings, 2013, 149–153.

[11] Roberts, K., Demner-Fushman, D., Tonning, J. M., Gonzalez, G., & Karp, P. D. (2017).

Overview of the TREC 2017 precision medicine track. Proceedings of The Twenty-Sixth Text REtrieval Conference (TREC 2017).

[12] Yang, X., Bian, J., Hogan, W. R., & Wu, Y. (2017). Clinical concept extraction using transformers. BMC Medical Informatics and Decision Making, 17(Suppl 2), 1–10.

[13] Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., & Xu, H. (2017). CLAMP – A toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association, 25(3), 331–336.

[14] Wang, Y., Wang, L., Rastegar-Mojarad, M., Liu, S., Shen, F., Liu, H., & Zhu, Q. (2018). Clinical information extraction applications: a literature review. Journal of Biomedical Informatics, 77, 34–49.