Enhancing Clinical Data Capture: Developing a Natural Language Processing Pipeline for Converting Free Text Admission Notes to Structured EHR Data

Patrick Styll[1,*], Wojciech Kusa[1] and Allan Hanbury[1]

[1]*Data Science Research Unit (E194-04), Technische Universität Wien, Favoritenstraße 9-11, 1040 Vienna, Austria*

Abstract

Automating the extraction of essential patient information from clinical texts, such as admission notes, can significantly enhance the entry of this data into Electronic Health Records (EHR), thereby enhancing workflow efficiency and supporting improved patient care and healthcare management. To address this issue, we introduce a Natural Language Processing (NLP) pipeline designed to (i) automatically extract patient data via Named Entity Recognition (NER), (ii) normalize the extracted data to correspond to codes in official medical ontologies, and (iii) coerce the data into EHR format using Health Level 7's (HL7) Fast Healthcare Interoperability Resources (FHIR) standard. By adhering to these widely used standardized formats, the pipeline output can be immediately integrated into the Hospital Information System (HIS).

To achieve this, we propose a newly labeled dataset comprising 255 notes from unlabelled datasets published by the Text Retrieval Conference's (TREC) Clinical Trials tracks. Finally, we utilize SapBERT for the normalization of extracted entities and employ the FHIR standard as a basis to generate Electronic Health Records (EHRs).

Keywords

Clinical Named Entity Recognition, SapBERT, FHIR, Electronic Health Records, ICD-10, NDC

1. Introduction

The increasing volume of clinical text data presents both challenges and opportunities for the healthcare sector [2]. Extracting meaningful information from these texts, such as personal patient data, is critical for applications in patient care, clinical research and healthcare management [3]. Admission notes, written by doctors when a new patient is admitted to the hospital, include essential patient details such as gender, age, and various medical conditions. Currently, doctors manually input this information into an Electronic Health Record (EHR), creating a bottleneck in the medical workflow. To address this issue, we introduce a Natural Language Processing (NLP) pipeline designed to (i) automatically extract patient data via Named Entity Recognition (NER), (ii) normalize the extracted data to correspond to codes in official medical ontologies, and (iii) coerce the data into EHR format using Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) Standard. By adhering to these widely used standardized formats, the pipeline output can be immediately integrated into the Hospital Information System (HIS).

In Section 2, we present related work, while in Section 3, we introduce the first step of the pipeline, where we extract relevant information from clinical texts, i.e. patient related information from admission notes. We introduce our specific goal and evaluation metrics used. Furthermore, we propose a newly labeled dataset comprising 255 entries from unlabelled datasets published by the Text Retrieval Conference (TREC) Clinical Trials tracks. We show origins of the data, justify our labelling techniques and present insights from our Exploratory Data Analysis (EDA). We train and fine-tune several Bidirectional Encoder

NL4AI 2024: Eighth Workshop on Natural Language for Artificial Intelligence, November 26-27th, 2024, Bolzano, Italy [1]

[*] Corresponding author.

[†] These authors contributed equally.

patrick.styll@tuwien.ac.at (P. Styll); wojciech.kusa@tuwien.ac.at (W. Kusa); allan.hanbury@tuwien.ac.at (A. Hanbury)
https://github.com/Padraig20 (P. Styll); https://wojciechkusa.github.io/ (W. Kusa);
https://informatics.tuwien.ac.at/people/allan-hanbury (A. Hanbury)
0000-0003-4420-4147 (W. Kusa); 0000-0002-7149-5843 (A. Hanbury)

Representations from Transformers (BERT) [4] models, evaluate their performance and explore different techniques to further enhance them. In Section 4, we handle both the second and third step of the pipeline. Firstly, we introduce the official medical ontologies we use, which are the 10^{th} revision of the International Statistical Classification of Diseases and Related Health Problem (ICD-10) and the National Drug Code (NDC). We also explain the underlying technology in normalizing the extracted information via Self-Aligning Pretrained BERT (SapBERT) [5] [6]. Secondly, we reveal how we used the HL7 FHIR standard to coerce all extracted and normalized information into an EHR. Finally, in Section 6 we conclude our research, discuss and summarize our findings and present a web-interface showcasing the whole integrated workflow.

2. Related Work

For the first part of our pipeline, we largely rely on *microsoft/mdeberta-v3-base* [7] [8] as our baseline. This large, multilingual general-domain model has recently gained recognition for its effectiveness in processing medical data, making it a suitable choice for medical NER. Furthermore, in our participation [9] in the MultiCardioNER [10] task from the BioASQ [11] workshop at CLEF2024, we have found valuable insights which we make further use of in this paper. The shared task focuses on the multilingual adaptation of clinical NER systems to the cardiology domain. It includes two key tasks: disease detection in Spanish texts and drug detection across Italian, Spanish, and English texts.

In the second step during our pipeline, we use *Self-Aligning Pretrained BERT (SapBERT)* [5][6] for normalizing the extracted entities to standardized codes. SapBERT is a specialized version of the BERT model, developed specifically for biomedical and clinical text mining and designed to create high-quality embeddings of medical texts. In order to generate embeddings that are particularly well-suited for biomedical applications, the model has been exposed to large datasets of biomedical literature and clinical notes during training. As of standardized codes, we use 10^{th} *revision of the International Statistical Classification of Diseases and Related Health Problem (ICD-10)* to classify medical conditions, symptoms and medical procedures. The ICD-10 is a globally recognized medical classification system developed by the World Health Organization (WHO), and has since become a critical tool for diagnosing and classifying a wide range of diseases and health conditions. For pharmaceuticals, we have decided to make use of *National Drug Code (NDC)*, which is a unique identifier largely used in the United States for drugs and other pharmaceutical products. Established by the Food and Drug Administration (FDA), the NDC serves as a universal product identifier for human drugs.

For coercing the output into an EHR, we have decided to use *Fast Healthcare Interoperability Resources (FHIR)* in the third step of the pipeline. FHIR is a standard for exchanging healthcare information electronically, designed to enable interoperability between different healthcare systems. Since it is designed to be extensible, it allows developers to build custom applications and extensions without jeopardizing compatibility - this is a big factor on why we use FHIR as the output format of our pipeline.

3. Named Entity Recognition for Patient Admission Notes

In this section, we introduce the initial phase of our clinical text processing pipeline, focusing on the extraction of crucial patient-related information from admission notes. We look into specific objectives and evaluation metrics, and we introduce and explore a newly labeled dataset. We look into the training and fine-tuning of various BERT models, along with strategies for enhancing their performance.

3.1. Dataset Preparation and Exploratory Data Analysis

3.1.1. Data Collection

The primary dataset originates from the TREC CT/CDS topics, publicly accessible on the track's official website[1]. Each topic has a similar structure, including several diagnoses in free text format. The topics represent admission notes containing the most important patient details which a doctor takes as soon as a person is admitted to a hospital. This includes personal information and demographics, such as gender and age, but also the current medical conditions, symptoms, medications/treatments and medical procedures. The dataset makes a total of 255 entries (topics). This includes:

- **TREC CDS 2016 [12]** - each topic is split into three separate fields: note, description and summary. Since each field contains the same information in other words, they will be processed individually, creating a total of 90 topics.
- **TREC CT 2021 [13]** - 75 topics in total with one field.
- **TREC CT 2022 [14]** - 50 topics in total with one field.
- **TREC CT 2023 [15]** - preprocessed to free text in admission note style via GPT-3 - 40 topics in total. More details on preprocessing can be found in [16].

3.1.2. Data Labelling and Analysis

For simplification purposes, we have decided to focus on four different entities, encompassing the most important information which has to be extracted from admission notes.

- **Medical Conditions**
  Medical conditions describe long-term conditions, such as *diabetes mellitus* or *COVID-19*.
- **Symptoms**
  In contrast to medical conditions, these describe mostly short-term conditions, which may be indicators of medical conditions. E.g. *fever*, a symptom, is an indicator for *COVID-19*, a medical condition.
- **Medication/Treatment**
  This could either describe medicine (e.g. *Ritalin*) or treatment (e.g. *rehab*).
- **Medical Procedure**
  This includes both invasive and non-invasive procedures, such as *tracheostomy* or *MR*.

The labelling of the dataset has been done via the open-source tool doccano [17] by the author of this paper. See Table 1 for a summary of annotated data by entity type, showcasing how imbalanced the entity to non-entity ratio is. It is important to mention that this specific dataset has *not* been reviewed by domain experts.

Table 1
Statistics-Summary for all entity types, showing the imbalanced nature of the dataset. Each count represents a token, i.e. subword. The data adheres to the IBO-tagging format.

Entity Type	**Medication**	**Symptom**	**Procedure**	**MedCond**
B-count	370	900	231	1080
I-count	168	538	127	642
Total Subwords	32942			

There were several issues while labelling the data. The term *medical condition* is not entirely clear and subject to interpretation. For instance, we can observe the relationship between *medical condition* and *symptom*. E.g., a *fever* is not a medical condition - it is a response to medical condition or disease.

[1] http://trec-cds.org/

The same goes for *dysuria*, being the subsequent response to e.g. UTIs (Urinary Tract Infections), a collection of various medical conditions. On the other hand the question arises whether injuries can be seen as medical conditions. In fact, injuries, such as a broken arm, are not considered medical conditions - injuries themselves are their own category in the medical language, which are, however, not included in this analysis.

3.2. Model Training and Evaluation

For evaluating the models, we used entity-level evaluation metrics [18], consistent with our previous submission for MultiCardioNER [9], specifically using $F1_{avg}$. Since we are working with a highly imbalanced dataset, entity-level evaluation provides a more accurate assessment of NER performance.

For the NER step of our pipeline, we have decided to use **microsoft/mdeberta-v3-base** [7] [8] instead of models with less parameters such as **google-bert/bert-base-multilingual-cased** [19] or specialized models as **alvaroalon2/biobert_diseases_ner** [20], since we were most successful with it in previous experiments dealing with medical NER [9]. For hyperparameter-tuning, we used a 70:15:15 split; we observed that changes in certain parameters led to large performance differences in the model. These include the learning rate, where higher values (i.e. 0.1) lead to worse results ($F1_{avg}$ of ≈ 0.8); low values (i.e. 0.0001) also led to bad performance, suggesting that a certain balance is required. Similar behaviour can be observed for the batch size, where 16 appears to be the optimum. $F1_{avg}$ no longer significantly changes after about 10 epochs, and even drops, showing signs of overfitting the training data. In the end, the parameters we achieved from tuning and therefore used for training are 16 for batch size, 0.01 for learning rate and 128 for the input size of the model, running with the SGD optimizer for 10 epochs. These parameters gave us a final validation $F1_{avg}$ of $\mathbf{85.6}\%$ and training $F1_{avg}$ of $\mathbf{89.1}\%$. The training history of the final model can be observed in Figure 1. Table 2 demonstrates the optimized results for all entity types. Unfortunately, the metrics for surgical procedures are relatively low compared to the other metrics - this is largely due to the fact that surgical procedures are rather scarce (see Table 1) and offer a more diverse vocabulary. More data would be crucial to receive better results.

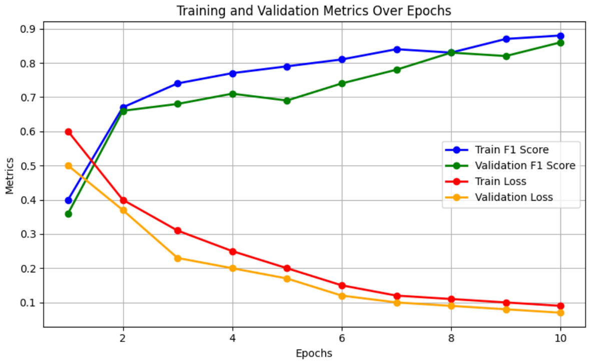

Figure 1: Training for Medical Condition with optimized Parameters.

3.2.1. Effect of Data Augmentation

As can be seen in section 3.1.2, there exist great imbalances in the relative ratio between entities and non-entities. In order to tackle this problem and increase model accuracy, we have decided to even these modalities out via data augmentation. In detail, we shuffle the sentences and their respective entities around in random order and thus generate new model input, essentially doubling the amount of training data. This augmentation has only been performed on the train set, while the validation and test set were left unchanged. For sentence detection, we have used spaCy [21]. In general, this resulted in overall increased metric values, as can be seen in Table 2; bear in mind that they originate from already fine-tuned models.

Table 2
Validation results of tuned models for *Medical Condition*, *Symptom*, *Medication/Treatment*, and *Surgical Procedure* before (left) and after data augmentation (i.e., shuffling of sentences) has been performed (right).

Entity Type	$\mathbf{F1_{avg}}$
Medical Condition	85.6%
Symptom	80.0%
Medication/Treatment	75.8%
Surgical Procedure	67.2%

Entity Type	$\mathbf{F1_{avg}}$
Medical Condition	90.8%
Symptom	83.2%
Medication/Treatment	78.6%
Surgical Procedure	71.4%

4. Generation of Structured Electronic Health Records (EHRs)

In this section, we address both the second and third steps of our pipeline. We introduce the standard medical ontologies utilized in our work, and we also dive into the technology used for normalizing the extracted information, specifically focusing on the application of SapBERT. Following this, we discuss how we used the HL7 FHIR standard to integrate all extracted and normalized data into an EHR system.

For entitiy types *Medical Condition*, *Medication/Treatment* and *Symptom* we use the ICD-10 codes taken directly from the website for Centers for Medicare & Medicaid Services[2], and for *Medication* we use NDC codes. These were taken from the FDA's official website openFDA[3], but had to be thoroughly preprocessed for use. We used both the proprietary and non-proprietary name for matching the code. Furthermore, since each NDC code includes packaging information, which we do not extract from the text, we have arbitrarily selected one code to represent the medicine. As a result, the packaging details associated with this code may not be accurate.

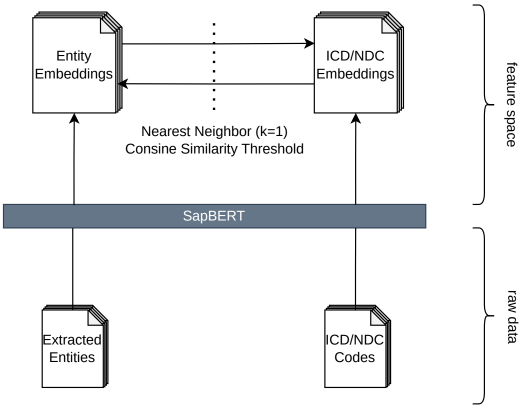

Figure 2: SapBERT integrated workflow. The raw data, i.e. the extracted entities and medical classification codes, are transformed into feature space (embeddings). From there, they are matched via nearest neighbour search with $k = 1$ based on a cosine similarity threshold.

4.1. Integration with SapBERT

The integration with SapBERT is required for medical entity normalization. In order to standardize the extracted entities, we need to connect each of them with their respective ICD-10/NDC code. In

[2] https://www.cms.gov/medicare/coding-billing/icd-10-codes
[3] https://open.fda.gov/data/ndc/

detail, we leverage SapBERT to create embeddings for ICD-10 and NDC. The core steps of the workflow (see Figure 2) include generating embeddings for these codes, performing nearest neighbor search, and determining the cosine similarity between embeddings to find the closest matches based on a pre-defined threshold. Based on experience from initial experiments, we chose 0.4 for ICD-10 codes and 0.3 for NDC codes.

4.2. Application of the FHIR Standard

The input of this part is (i) the extracted text of the NER model, (ii) the normalized entity and (iii) the corresponding ICD-10/NDC code. These triples are then grouped into the fitting FHIR resource, which represent specific types of clinical and administrative information in the FHIR standard. The HL7 organization offers a FHIR Resource Guide[4], with which it was quite simple to find and use the appropriate resources. The FHIR resource templates for each separate entity type and an example for a FHIR Resource Bundle as a final EHR can be found in the GitHub repository. It is interesting to note that the resources templates for *Medical Condition* and *Symptom* are the same, except for a note being used to highlight the difference. This once more emphasizes the content-related overlap of these entity types as described in section 3.1.2.

5. Limitations and Future Work

One of the primary limitations of this work lies in the lack of a quantitative evaluation of the mapping methodology introduced in Section 4.1. While the approach to map extracted entities to standard codes shows promise, we do not provide a formal assessment of its performance. Future iterations of this study should aim to address this gap by introducing an appropriate evaluation framework, allowing for a stronger argument regarding the effectiveness of the mapping mechanism and its applicability in real-world scenarios. Furthermore, the dataset used in this study, consisting of approximately 32943 wordpieces, poses potential challenges for the generalizability of the findings. A dataset of this size, while sufficient for a proof-of-concept, may not capture the full complexity and variability present in larger, real-world datasets. Moreover, the dataset lacks expert intervention during the labeling process, which introduces the possibility of inaccuracies in entity extraction. In future work, incorporating expert validation for at least a subset of the data would enhance the quality and accuracy of the annotations, providing a more robust foundation for the entity extraction and mapping methods.

6. Conclusion

We have established an NLP pipeline for processing free-text admission notes into EHR. In Section 3, we look into the problem of medical NER. We select the appropriate architecture and define metrics for demonstrative evaluation. We describe the data mining process, explore the data and justify our data labelling processes. Finally, we train models through various strategies and assess their performance. In Section 4, we gave insights into medical classification lists such as ICD-10 and NDC. We further showcase how we take the output of Section 3 to normalize the extracted entities via SapBERT. Finally, we show how the FHIR standard aids us in generating a standardized EHR. Furthermore, we have created a web interface to showcase all three steps of the pipeline: (i) the extracted entities inside the admission note, (ii) the normalized entities, including the extracted text, the normalized text and ICD-10/NDC code, and (iii) the automatically generated FHIR Resource Bundle representing a standardized EHR. Any code can be found in the GitHub repositories *Padraig20/Disease-Detection-NLP* and *Padraig20/EHR-Generator* for medical NER and the EHR-Generator including the SapBERT workflow as well as the Web-Interface, respectively.

[4] https://www.hl7.org/fhir/resourceguide.html

References

[1] G. Bonetta, C. D. Hromei, L. Siciliani, M. A. Stranisci, Preface to the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2024), 2024.

[2] H. Dalianis, Clinical text mining: Secondary use of electronic patient records, Springer Nature, 2018.

[3] D. Demner-Fushman, N. Elhadad, C. Friedman, Natural language processing for health-related texts, in: Biomedical Informatics: Computer Applications in Health Care and Biomedicine, Springer, 2021, pp. 241–272.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[5] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 4228–4238.

[6] F. Liu, I. Vulić, A. Korhonen, N. Collier, Learning domain-specialised representations for cross-lingual biomedical entity linking, in: Proceedings of ACL-IJCNLP 2021, 2021, pp. 565–574.

[7] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, 2021. arXiv:2111.09543.

[8] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[9] P. Styll, L. Campillos-Llanos, W. Kusa, A. Hanbury, Cross-linguistic disease and drug detection in cardiology clinical texts: Methods and outcomes, in: CLEF 2024: Conference and Labs of the Evaluation Forum, Technische Universität Wien, Spanish National Research Council (CSIC), Grenoble, France, 2024.

[10] S. Lima-López, E. Farré-Maduell, J. Rodríguez-Miret, M. Rodríguez-Ortega, L. Lilli, J. Lenkowicz, G. Ceroni, J. Kossoff, A. Shah, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of MultiCardioNER task at BioASQ 2024 on Medical Speciality and Language Adaptation of Clinical NER Systems for Spanish, English and Italian, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), CLEF Working Notes, 2024.

[11] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, M. Krallinger, N. Loukachevitch, V. Davydova, E. Tutubalina, G. Paliouras, Overview of BioASQ 2024: The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. Maria Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[12] K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, Overview of the trec 2016 clinical decision support track, in: Proceedings of the Text REtrieval Conference (TREC) 2016, National Institute of Standards and Technology (NIST), NIST, Gaithersburg, MD, 2016.

[13] I. Soboroff, Overview of trec 2021, in: Proceedings of the Text REtrieval Conference (TREC) 2021, National Institute of Standards and Technology (NIST), NIST, Gaithersburg, MD, 2021.

[14] K. Roberts, D. Demner-Fushman, E. M. Voorhees, S. Bedrick, W. R. Hersh, Overview of the trec 2022 clinical trials track, in: Proceedings of the Text REtrieval Conference (TREC) 2022, National Institute of Standards and Technology (NIST), NIST, Gaithersburg, MD, 2022.

[15] I. Soboroff, Overview of trec 2023, in: Proceedings of the Text REtrieval Conference (TREC) 2023, National Institute of Standards and Technology (NIST), NIST, Gaithersburg, MD, 2023.

[16] W. Kusa, P. Styll, M. Seeliger, O. E. Mendoza, A. Hanbury, Dossier at trec 2023 clinical trials track,

in: Proceedings of the Text REtrieval Conference (TREC), National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2023.

[17] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, X. Liang, doccano: Text annotation tool for human, 2018. URL: https://github.com/doccano/doccano, software available from https://github.com/doccano/doccano.

[18] D. S. Batista, Named-entity evaluation metrics based on entity-level, 2018. URL: https://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/, accessed: 2024-05-21.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[20] Á. Alonso Casero, Named entity recognition and normalization in biomedical literature: a practical case in SARS-CoV-2 literature, 2021. URL: https://oa.upm.es/67933/, unpublished.

[21] Explosion-AI, spaCy: Industrial-strength Natural Language Processing in Python, https://spacy.io/usage/linguistic-features#sbd, 2023. URL: https://spacy.io/, version 3.0.