Named Entity Recognition for Clinical Patient Notes

Ayush Tripathi, Kishan N. Murthy, Paritosh Singh,
Prakash Parajuli, Siddharth G. Byale
University of Southern California, 90018, Los Angeles, California
ayusht@usc.edu, knarasim@usc.edu, paritosh@usc.edu
parajuli@usc.edu, sbyale@usc.edu

1 Introduction

Physicians have a long history of writing patient notes documenting medical history, physical exam findings, possible diagnoses, and follow-up care. These documents are rich in information, and extracting meaningful details from these clinical notes can aid in the accurate diagnosis of diseases. It is also beneficial for Clinical Research, when the information from the patient notes is extracted and analyzed at scale.

Named Entity Recognition (NER) is a part of information retrieval and natural language processing that tries to recognize and categorize named entities referenced in unstructured text. Clinical Named Entity Recognition (CNER) is an essential task to identify and classify key clinical terms in electronic medical records (Lei et al., 2014). Generally, it is a sequence labeling problem where entity boundary and category labels are jointly predicted.

Question answering (QA) is a field of information retrieval and natural language processing that tries to provide an answer to a given question automatically. Typically, a model answers a given question when a necessary context in the form of reading comprehension is provided.

Our project aims to build a neural network model that identifies and annotates specific clinical concepts in English patient notes using NER and QA techniques. Our approach is to extract key information from the patient notes in the form of a sequence labeling task where we predict the entity boundary, given a list of clinical concepts and patient notes. Here a single clinical concept is treated as a question, and the patient notes provide the necessary context while the model predicts entity boundaries.

The developed model can also greatly benefit exams like the United States Medical Licensing Examination (USMLE) Step 2, where extracted features from the clinical notes can potentially be evaluated against the rubric.

2 Related Work

Before the popularity of machine learning, Rule-based NER approaches using dictionary resources were widespread and performed relatively well for simple contexts (Eftimov et al., 2017). However, due to the simplistic nature of models, they offer limited performance to complex texts. In recent times, Machine learning methods and Deep learning methods have become popular approaches to solving the NER. Usually, machine learning methods rely on manual feature engineering, while deep learning approaches are generally end-to-end.

Traditionally, NER has been solved using classical machine learning approaches such as Hidden Markov Model (HMM) (Zhang et al., 2004), Maximum Entropy Markov Model (MEMM) (McCallum et al., 2000), and Conditional Random Field (CRF) (McCallum and Li, 2003). Later, researchers started using hybrid models which combine classical machine learning with deep learning techniques. One such popular architecture combines Bidirectional Long Short-Term Memory (Bi-LSTM) (Hochreiter and Schmidhuber, 1997) and CRF. Researchers further combined word embedding with BiLSTM-CRF architectures to improve performance (Lample et al., 2016). Recently with the introduction of transformer architecture (Vaswani et al., 2017), researchers have created models with combined transformers, Bi-LSTM, and CRF architectures to maximize the performance (Devlin et al., 2018).

Despite the scarce nature of medical corpora, several models have been trained in clinical and biomedical texts. Lee et al. (2019) trained a BERT model on PubMed abstracts and PMC full-text articles from scratch. Li et al. (2019) finetuned BERT model on Medication, Indication, and Adverse Drug Events (MADE) 1.0 corpus, the National Center for Biotechnology Information (NCBI) disease corpus, and the Chemical-Disease Relations (CDR)

corpus.

For CNER tasks, Xu et al. (2018) have proposed using Bi-LSTM and CRF model for the Medical Named Entity Recognition task on the NCBI disease corpus. Zhang et al. (2019) solved breast cancer NER with pre-trained BERT finetuned on Chinese clinical text. Schneider et al. (2020) created BioBERTpt, a neural language model using BERT architecture pretrained in the Portuguese Language to solve CNER tasks. Wu et al. (2020) used a Bi-LSTM, CRF, and Attention for NER and Intent Analysis on Chinese medical questions obtained from the health community website. Li et al. (2020) performed CNER with combined Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), Bi-LSTM, and CRF for Chinese Clinical Text.

Variations of transformer architectures like BERT, Generative Pretrained Transformer (GPT), Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019), Decoding-enhanced BERT with disentangled attention (DeBERTa) (He et al., 2020), and others have enabled us to reach exceptional performance in many NLP tasks, including QA and NER.

Even though substantial work has been done in medical NER, extraction models on English clinical patient notes are lacking. We propose designing a neural network model that identifies and annotates the text to find specific clinical concepts based on English clinical patient notes. This is performed by combining both QA and NER techniques in the transformer model.

3 Method

3.1 Datasets

The data utilized in this project comes from the USMLE Step 2 Clinical Skills exam, which is a medical licensing exam. This exam assesses a trainee's ability to notice relevant clinical data during interactions with standardized patients. Each test taker sees a Standardized Patient, a person trained to simulate a clinical situation, during the exam. The test taker records the pertinent information of the interaction in a patient note after conversing with the patient. There are three data files in the dataset. The first data file contains a list of annotations for each individual case, whereas the second data file contains patient history notes for all patient numbers. Finally, the train data is organized into 6 columns and 14300 rows. The six columns are id, patient number, case number, annotation, and annotation location in the patient notes. When data of all the files are combined, we get the dataset with annotations and their locations in the patient history for each patient history.

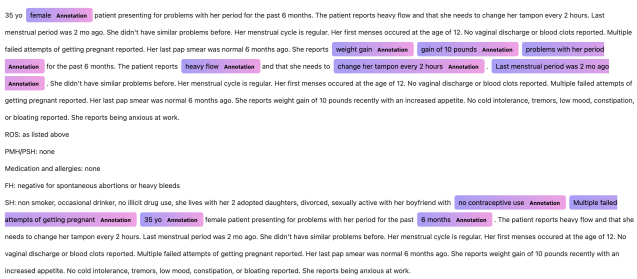

Figure 1: **Annotation in Patient Notes**

3.2 BERT

Bidirectional Encoder Representations (BERT) from Google achieved exceptional performance in a wide variety of NLP tasks, including Question Answering (SQuAD v1.1), Natural Language Inference (MNLI), and others. BERT is a language model which is bi-directionally trained. This means we can now have a deeper sense of language context and flow compared to the single-direction language models. Instead of predicting the next word in a sequence, BERT uses a novel technique called Masked LM(MLM): it randomly masks words in the sentence, and then it tries to predict them. Masking means that the model looks in both directions and uses the full context of the sentence, both left and right surroundings, to predict the masked word. Unlike the previous language models, it takes both the previous and next tokens into account simultaneously.

3.3 RoBERTa

The Robustly Optimized BERT-Pretraining Approach (RoBERTa) is a variant of the BERT model. To improve the training technique, RoBERTa eliminates the Next Sentence Prediction (NSP) assignment from BERT's pre-training and replaces it with dynamic masking, in which the masked token changes with time. The perplexity of the MLM objective and end-task accuracy have improved when a model is trained with big mini-batches. Furthermore, it was found that training BERT on larger datasets enhances its performance significantly. As a result, the training data was increased to 160GB of uncompressed text.

3.4 DeBERTa

The Robustly Optimized BERT-Pretraining Approach (RoBERTa) is a variant of the BERT model. To improve the training technique, RoBERTa eliminates the Next Sentence Prediction (NSP) assignment from BERT's pre-training and replaces it with dynamic masking, in which the masked token changes with time. The perplexity of the MLM objective and end-task accuracy have improved when a model is trained with large mini-batches. Furthermore, it was found that training BERT on larger datasets enhances its performance significantly. As a result, the training data was increased to 160GB of uncompressed text.

4 Experiment

4.1 Experimental Setup

Dataset Setup We get our dataset by combining three different data files. Using the stratified K fold technique, the dataset is divided into five folds. Three of the five folds were used for training the model, one for validation, and one for testing.

Data Preprocessing Data Preprocessing is the most important step in any NER challenge. Cleaning the data is required before passing it to any deep learning model. We deleted any unnecessary white spaces before and after the text, as well as special characters that contributed no value to the corpus, from patient history text data. Furthermore, we discovered a typographical problem in the annotation text, so we went over all of the annotations and corrected them. In order to reflect changes in the annotation, the annotation location was modified correspondingly.

Tokenizer After the data has been preprocessed, the dataset must be tokenized. Every transformer-based model uses a different tokenization technique and different special tokens. However, they all follow the same approach. The sentences to the transformer are given as a Question and Answer pair, where the feature text is the question, and the patient history is the content. The question and the content sentences are separated by a special separator token. Additionally, the tokenizer takes input parameters such as max length of string, truncation, and padding value.

The tokenizer divides the words into tokens and subtokens, which are then converted to input ids. These input ids act as indices to the word embedding table and map tokens to their respective word embeddings. Apart from the input ids, the tokenizer generates several other essential arguments. These arguments are sequence ids, offset mapping, and attention mask. The sequence ids show whether the token belongs to the question or the content. Tokens that are part of the question are assigned a value of 0, tokens that are part of the content are assigned a value of 1, and special tokens like padding and separator tokens are assigned a value of None. Offset mapping indicates each token's start and end indices in the string. Finally, the attention mask assists the model in determining which tokens present in the original text need to be labeled as a named entity.

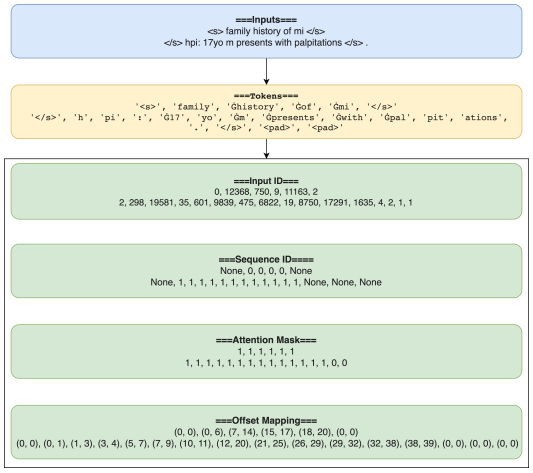

Figure 2: **Tokenization of Sentences(RoBERTa)**

Baseline Methods As a starting point, we developed a standard string matching method. It looks for a one-to-one correspondence between the feature text and the sentences in the patient history. The model's accuracy is severely poor because it is seeking for a complete match, however the annotation location of correct prediction has excellent precision.

Training Google Colab and the USC Center For Advanced Research Computing (CARC) platform were extensively used to train transformer models. The BERT and RoBERTa-Large transformer models were trained on Google Colab. The DeBERTa-v3-Large model was trained on the CARC platform with an 8-core CPU, 16 GB RAM, and one Nvidia A100 GPU

BERT The BERT model is usually an excellent place to start experimenting because of its faster training pace and lower memory usage than other

models in the transformers library. We added a dropout layer and a linear classifier to the Bert-base model to get more accurate predictions
Hyperparameters
Batch Size = 8, Learning Rate = 1e-5,
Optimizer = AdamW, Epochs = 5, Dropout = 0.2

RoBERTa RoBERTa-large model was used to extract the patient's feature information from the patient history. The tokenizer was fed the feature text and the patient history to generate the sequence ids, input ids, and attention mask. These fields are fed into our model. A dropout layer was added to the end of the roberta-model. Finally, a linear classifier was used to make the prediction.
Hyperparameters
Batch Size = 4, Learning Rate = 1e-5,
Optimizer = AdamW, Epochs = 5, Dropout = 0.2

DeBERTa After seeing improved results with the RoBERTa-large model, it was clear that we were on the right track and that DeBERTa, which improves on the BERT and RoBERTa models, needed to be trained. We added a dropout and linear classifier layer to the end of DeBERTa-v3-Large and performed the prediction.
Hyperparameters
Batch Size = 4, Learning Rate = 2e-5,
Optimizer = AdamW, Epochs = 5, Dropout = 0.2

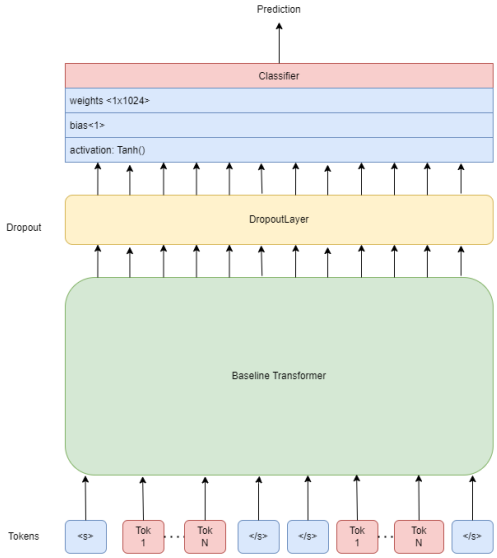

Figure 3: **General Architecture of Transformer Model**

Evaluation Protocols The developed models are evaluated using micro-averaged Precision, Recall, and F1 score based on the character spans. A character span is a pair of indexes representing a range of characters within the text. A span ***i j*** represents the characters with indices ***i*** through ***j***, inclusive of ***i*** and exclusive of ***j***. We predict a set of character spans for each instance of a feature class and patient notes and compare it with the ground-truth.

During the evaluation, based on ground-truth and predicted spans, we assign each character index a label as shown in Table 1. Finally, we compute an overall micro-average Precision, Recall, and F1 score from the True Positives, False Positives, and False Negative aggregated across all instances.

Table 1: Description of labels

Label	Description
True Positive	Present in both ground-truth and in prediction
False Positive	Absent in ground-truth and present in prediction
False Negative	Present in ground-truth but not in prediction

4.2 Results and Discussion

The performance of Basline, BERT, RoBERTa-Large and DeBERTa-v3-Large are shown in the Table 2.

Table 2: Model performance metrics on test set

Model	Precision	Recall	F1-Score
Baseline	36.64	12.99	22.91
BERT	75.27	72.34	71.27
RoBERTa-Large	79.89	84.98	82.35
DeBERTa-Large	88.00	88.40	88.21

5 Conclusion

We fine-tuned pre-trained BERT, RoBERTa, and DeBERTA models to annotate specific clinical phrases in English patient notes in this project. The trained model achieves an F1 score of 71.27, 82.35, and 88.21, respectively. These models could aid in the accurate diagnosis of diseases and may also help in Clinical Research. Finally, the trained model can further assist downstream tasks like automated evaluations of the United States Medical Licensing Examination (USMLE) Step 2.

Team responsibilities.

The project was executed in the following manner:

Phase 1: Perform Exploratory Data Analysis and Visualization by Ayush Tripathi and Paritosh Singh.
Phase 2: Achieve Data preprocessing and evaluate a baseline using string matching by Prakash Parajuli and Siddharth G. Byale.
Phase 3: Develop and evaluate BERT NER Model by Kishan N. Murthy and Siddharth G. Byale.
Phase 4: Design and evaluate RoBERTa NER Model Ayush Tripathi and Prakash Parajuli.
Phase 5: Develop and evaluate DeBERTa NER Model by Kishan N. Murthy and Paritosh Singh.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Tome Eftimov, Barbara Koroušić Seljak, and Peter Korošec. 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLOS ONE*, 12(6):1–32.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, and H. Xu. 2014. A comprehensive study of named entity recognition in Chinese clinical text. *J Am Med Inform Assoc*, 21(5):808–814.

F. Li, Y. Jin, W. Liu, B. P. S. Rawat, P. Cai, and H. Yu. 2019. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)-Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. *JMIR Med Inform*, 7(3):e14830.

Xiangyang Li, Huan Zhang, and Xiao-Hua Zhou. 2020. Chinese clinical named entity recognition with variant neural structures based on bert methods. *Journal of Biomedical Informatics*, 107:103422.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 591–598, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, page 188–191, USA. Association for Computational Linguistics.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Chaochen Wu, Guan Luo, Chao Guo, Yin Ren, Anni Zheng, and Cheng Yang. 2020. An attention-based multi-task model for named entity recognition and intent analysis of chinese online medical questions. *Journal of Biomedical Informatics*, 108:103511.

Kai Xu, Zhanfan Zhou, Tianyong Hao, and Wenyin Liu. 2018. A bidirectional lstm and conditional random fields approach to medical named entity recognition. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017*, pages 355–365, Cham. Springer International Publishing.

Jie Zhang, Dan Shen, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2004. Enhancing hmm-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics*, 37(6):411–422. Named Entity Recognition in Biomedicine.

Xiaohui Zhang, Yaoyun Zhang, Qin Zhang, Yuankai Ren, Tinglin Qiu, Jianhui Ma, and Qiang Sun. 2019. Extracting comprehensive clinical information for breast cancer using deep learning methods. *International Journal of Medical Informatics*, 132:103985.