# An Automated Approach to Identify Key Medical Phrases in Patient Notes

Jinyoung Hur[*], Taewan Kim[*], and Qiwei Lin[*]

[*]Department of Computer Science, University of Chicago
[1]{jinahur, taewank, qiweilin}@uchicago.edu

## Abstract

In this work, we experiment with different approaches that utilize domain-specific BERT and Bidirectional Long Short Term Memory (Bi-LSTM) to map key phrases in patient notes to important medical concepts. Using a collection of patient notes from Step 2 Clinical Skills examination data from National Board of Medical Examiners (NBME), we train medical concept extractors that achieve satisfactory performance on mapping clinical concept features to word sequences expressed in clinical patient notes. Our best model achieves 0.928 F1 score on the validation set and 0.797 on the test set.

## 1 Introduction

Natural Language Processing (NLP) has played a crucial role in understanding unstructured data and predictive analysis. It has been utilized in biomedical and clinical text mining to obtain and analyze information. Contextualized word embedding models, such as BERT (Devlin et al., 2018) and domain-specific pre-trained BERT (Alsentzer et al., 2019), have also been applied to various NLP tasks such as classification and named entity recognition. The work of Lybarger et al focused on extracting patient's social determinants of health (SDOH) from hospital readmission notes (Lybarger et al., 2021). Their work allows clinicians to make precise decisions that could drive better health outcomes (Blizinsky KD, 2018). On the other hand, clinicians are also interested in identifying important medical concepts in patient notes using NLP models. Physicians spend a lot of time on practice writing patient notes before they are licensed, and the assessment on the skill of writing patient notes is a time-consuming task that requires feedback from other doctors. The appropriate use of NLP model could reduce a significant amount of time spent by physicians and serve as a more transparent and interpretable evaluation method for medical trainees.

In this work we experiment with different approaches that utilize domain-specific BERT such as Bio-Clinical BERT and Bidirectional Long Short Term Memory (Bi-LSTM) to develop an automated approach to map key phrases in patient notes to important medical concepts (features). Using a collection of patient notes from Step 2 Clinical Skills examination data from National Board of Medical Examiners (NBME), we train medical concept extractors that achieve satisfactory performance on mapping clinical concept features to words expressed in clinical patient notes.

## 2 Related Work

### 2.1 Domain-Specific Contextualized Text Embedding

Using contextualized word representation model like BERT obtains state-of-the-art performance on many NLP tasks. BERT, however, was pre-trained on general corpora like BooksCorpus and English Wikipedia. These corpora, although have a large vocabulary size, might not perform well on downstream clinical NLP tasks because domain-specific proper nouns and terms in medical notes are less frequent in general corpora. Also, BERT suffers from fixed input restriction and thus difficult to apply for medical notes (Hsu et al., 2020).

It is increasingly common to use pre-trained contextualized embedding developed on medical corpora. Examples include Med-BERT (Rasmy et al., 2021), G-BERT (Shang et al., 2019), Bio-BERT (Lee et al., 2020), and Clinical BERT (Alsentzer et al., 2019). These pre-trained representation models differ in the size and type of the training corpus.

Evidence suggests that pre-training BERT with domain-specific corpora improves the performance on clinical NLP tasks. Med-BERT (Rasmy et al., 2021) is pre-trained on a structured electronic health records (EHR) dataset of approximately

28.5M patients and fine-tuned to outperform models without contextualized embedding on disease prediction tasks. G-BERT (Shang et al., 2019) combines Graph Neural Networks and BERT to represent medical codes and achieve state-of-art performance on medication recommendation. Bio-BERT is pretrained on PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC), along with English Wikipedia and BooksCorpus to better understand complex biomedical texts. It outperforms previous models with only general BERT on tasks such as medical question answering and relation extraction (Lee et al., 2020). Clinical BERT and its variants use all note types text data in MIMIC-III to train BERT-base model, while Discharge Summary BERT only uses the discharge summary text in MIMIC-III for training (Alsentzer et al., 2019). Recent work (Alsentzer et al., 2019) compares Clinical BERT and its variants with Bio+Clinical BERT (initialized with Bio-BERT and trained with the same data as Clinical BERT) on several clinical NLP tasks. Bio+Clinical BERT and Bio+Discharge Summary BERT obtain improvements on accuracy over BERT and Clinical BERT. They achieve 82.7% accuracy on MedNLI and high F1 scores on i2b2 2006 (94%), i2b2 2010 (87%), and i2b2 2012 (78.9%). However, Bio-BERT has the leading performance on i2b2 2006 (94.8%). Overall, the results inform us that Bio-BERT, Bio-Clinical BERT, and Bio-Discharge Summary BERT are the appropriate models to start with for clinical text embedding when analyzing DBME-Notes data [1].

### 2.2 Medical Named Entity Recognition

Named entity recognition (NER) is an area where important nouns and proper nouns in a text is located and categorized (Zitouni, 2014). Identifying medical named entities and relations from medical records in unstructured text is critical in extracting certain hidden information in the diagnosis, supporting medical research, and making treatment decisions (Demner-Fushman, 2009; Liang, 2019). Yet, complexity of medical text and precise normalization of extracted named entities by mapping them to concepts make building a practical NER system difficult.

Studies show that systems based on neural networks provide the best performance for NER in medical notes (Florez, 2018). Algorithms such as Support Vector Machines (SVMs) and Convolutional Neural Networks (CNN) are commonly used in NER. For sequence problems, however, Recurrent Neural Networks (RNN) models are considered more appropriate as these models can classify the input sequence, accounting for the long time dependencies (M. Liwicki and Schmidhuber, 2007). Still, simple RNN model can face vanishing gradients issue (Bengio, 1994). Accordingly, Long Short-Term Memory (LSTM), an another RNN architecture, is recommended on the state of the art, given how it uses a short memory connection along the sequence that can partially resolve vanishing gradients issue. The model can further improve its performance by: 1) feeding the network with an appropriate input representation to provide closer vectors among related words; and 2) adding additional features for its input, such as character-level features from each word extracted using CNN or LSTM, then concatenating character and word representations (Chiu and Nichols, 2015; Z. Liu and Xu, 2017).

In this work, we will use NER to identify specific medical concepts, including symptoms, in clinical patient notes. Depending on sequence problem of our case, we would consider using bidirectional LSTM model. In addition, Conditional Random Fields (CRFs) that takes every neighbour word in a fixed window of words can be applied to NER (Z. Liu and Xu, 2017). We could potentially implement a CRF algorithm after the bidirectional LSTM output.

### 2.3 Related Work on Event Extraction

Previous work (Lybarger et al., 2018) shows that neural multi-task learning outperforms discrete models that require hand-engineered features and other baselines. However, its corpus is relatively small and homogeneous in source text (only includes notes from one institution), which casts doubts on the generalizability of the model. More recent work (Lybarger et al., 2021) gives us the basic understanding on how to utilize deep learning models for event extraction. Its framework of event extraction shows satisfactory performance on Trigger and Labeled argument prediction. However, their model cannot extract multiple events of the same type, as it only predicts a single event of a specific type per sentence. In order to extract multiple features of the same type for our work, we can use token-level classifiers to extract locations

[1] Resources for these pre-trained BERT

of important medical phrases within a patient note. Additionally, they use an entire sentence as the input to BERT, and find it hard to identify the span of events that spread across multiple sentences.

## 3 Data Overview and Exploratory Data Analysis

The dataset NBME-Notes is from the Step 2 Clinical Skills examination of the United States Medical Licensing Examination (USMLE), a major medical licensure exam co-hosted by Federation of State Medical Boards (FSMB) and National Board of Medical Examiners (NBME). This exam measures a trainee’s ability to identify pertinent clinical facts during encounters with standardized patients who are trained to portray a typical clinical case. After meeting with the patient, the medical trainee documents the relevant medical facts of the encounter in a patient note. Trained physicians who score these notes will then look for the presence of certain key concepts or features relevant to the case as described in a rubrics.

The dataset contains physician’s notes on 42,146 patients with 1,000 of them annotated. Annotations provide the start and end character-level index of the corresponding medical phrase. Table 1 shows a sample annotation and the corresponding phrases in the patient notes.

Table 1: Sample Annotation of NBME Patient Notes

| Feature | Annotation |
|---|---|
| Lightheaded | ’222 258’ |
| Text | |
| this time had chest pressure | |
| and felt as if he were going to pass out | |

On Average, these patient notes contain 211.67 tokens (median = 217), and the longest note contains 298 tokens. These length statistics indicate that we could pass an entire note into a BERT model without the need to first separate notes at sentence-level. Prior work (Lybarger et al., 2021) suggests that breaking notes into sentences and passing individual sentences into BERT might make it hard for the model to capture entities that span across more than one sentence.

1,000 annotated notes are equally distributed across ten medical cases. The distribution of cases in the annotated notes is thus very different from that in the overall dataset that include unlabeled notes (Fig 1). These medical cases are scenarios (such as symptoms, complaints, concerns) the standardized patients presents to the exam taker (medical student, resident or physician). Annotated notes contain around 14,300 annotations on 143 features (medical concepts). Table 2 provides a summary of the annotated datasets by case, but the mapping from case numbers to exact scenario names is not provided in the raw data.

Figure 1: The Distribution of Cases in the Overall Dataset (N = 42,146)

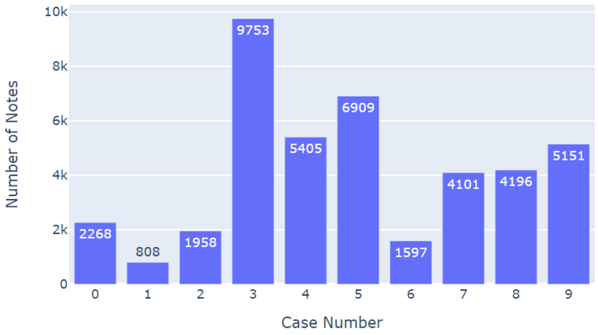


Table 2: Summary Statistics of the Annotated Dataset

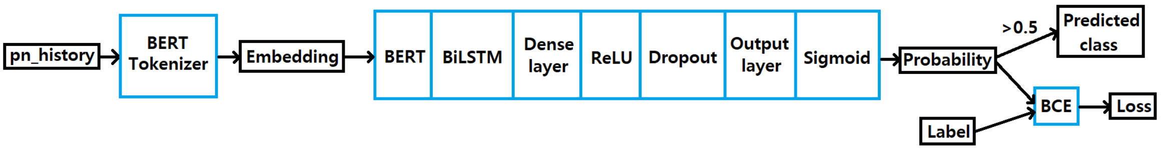
| Case | # Tokens | # Features | # Labels |
|---|---|---|---|
| 0 | 209.33 | 13 | 9.98 |
| 1 | 217.48 | 13 | 9.42 |
| 2 | 225.88 | 17 | 9.98 |
| 3 | 204.11 | 16 | 11.98 |
| 4 | 203.12 | 10 | 7.46 |
| 5 | 210.29 | 18 | 11.74 |
| 6 | 212.39 | 12 | 9.71 |
| 7 | 215.58 | 9 | 6.92 |
| 8 | 224.19 | 18 | 11.90 |
| 9 | 194.26 | 17 | 9.92 |

## 4 Methods

### 4.1 Pre-Processing

The data consists of three separate files: features, patient notes, and train data files. Features file contains the description of each feature, and patient notes file has patient notes for each case recorded by the test taker. With train file, we can get unique feature and patient note ID and key annotations in patient notes and their corresponding start, end string-level indices, which are converted to token-level labels for training models. To make our data more interpretable, train data is merged with features and patient notes files on feature and patient notes ID. Annotation and location

Figure 2: The model structure and the flow of the input data in the model



columns store the element(s) in a list, but the data is saved as string type (e.g., "['162, 190']") that hinders us from iterating over annotation locations for each case. We pre-process annotation and location columns by replacing with a list of integer elements (e.g., [[162, 190]]) to make annotation locations easily accessible.

We design the label matrix $L$ as the probability of each embedding token having a specific feature. The model predicts this probability as $P$ and the binary cross entropy between $P$ and $L$ is computed as loss. Specifically, the size of $L$ and $P$ is $(num\_batch, num\_token, num\_feature + 1)$ where the extra index in feature dimension captures the case when a given token does not have any feature. For each token, 1 is assigned to the feature index if it has that feature, and 1 is assigned to the last feature index if it does not have any feature. The final output of the model is the class matrix $C$ (same size as $P$) having 1 for indices with $P > 0.5$ and 0 for indices with $P <= 0.5$.

### 4.2 Models

We experiment with the following four models:

- Bio-Clinical BERT + Bi-LSTM
- Bio-Discharge BERT + Bi-LSTM
- Bio-Clinical BERT
- Distil-BERT

The general model structure is shown in Fig 2. For the two simpler model without Bi-LSTM, we make the following adjustment: 1) remove the Bi-LSTM layer, the dense layer and the Relu layer; 2) increase dropout; 3) train with fewer epochs. We hypothesize that domain-specific BERT pre-trained on medical corpora will outperform models trained on general corpora, and that more complex models with Bi-LSTM can outperform simpler models. These models generates token-level predictions that are later converted to string positions.

### 4.3 Training

We use 85% of the 1000 patient notes for training and the rest 15% for validation. The testing set, which is larger than the training and validation set combined, is held out by Kaggle.

Accuracy and loss measures are adjusted in the training process to better accommodate the prediction task. These measures are only calculated using tokens that have true or predicted labels not equal to zero. In doing so, we penalize false positive and false negative errors but avoid inflating the accuracy by counting true negative cases where models successfully predict that a token is not part of any medical phrase. The training is implemented with Pytorch on the Google Colab environment with GPU and we use Adam (Kingma and Ba, 2014) as the optimizer for all models. Hyperparameter values can be found in Table 3.

### 4.4 Evaluation

For given patient notes in the validation and testing data, models gives the predicted class matrix $C$, where each row is a one-hot vector. The locations of tokens predicted as having a feature (i.e. indices with $C_{ij} = 1$), are extracted from $C$ to determine the feature class. These locations of tokens are used to identify the start and end string position in the original text for the associated medical features. This process is done by getting the locations of each token embedding from the offset mapping of the tokenizer. Table 4 shows a synthetic sample label-prediction pair and the corresponding performance metrics. For each patient-feature pair, we score the character-level prediction as one of the three following metrics:

- True Positive (TP): if a character is within both a ground-truth and a prediction
- False Negative (FN): if a character is within a ground-truth but not a prediction
- False Positive (FP): if a character is within a prediction but not a ground truth

Table 3: Model Hyperparameters

| | Bio-C + LSTM | Bio-D + LSTM | Bio-C | Distil |
|---|---|---|---|---|
| Batch size | 16 | 16 | 16 | 16 |
| Learning rate | 1e-4 | 1e-4 | 1e-4 | 1e-4 |
| # epochs | 60 | 60 | 25 | 50 |
| LSTM hiddens ize | 256 | 256 | 256 | 256 |
| Dropout | 0.1 | 0.1 | 0.5 | 0.5 |

TP, FN, FP are used to compute Micro-F1 score on the pooled and case-specific validation dataset. Kaggle only reports the Micro-F1 on the whole test set, so we cannot compute case-specific performance on the test set.

Table 4: Sample Labels and Predictions at Character-level

| Label | Prediction | # TP | # FN | # FP |
|---|---|---|---|---|
| 0 3; 3 5 | 2 5; 7 9; 2 3 | 3 | 2 | 2 |

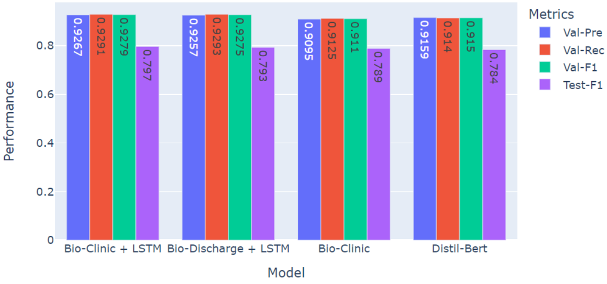
# 5 Results

## 5.1 Model Comparison

Fig 3 summarizes the performance of 4 four different models on the validation set and the test set. Domain-Specific BERT + Bi-LSTM models have higher F1 scores on the validation and test set. However, we also observe that the promising performance on the validation set (F1 score ≈ 0.928) does not generalize to the test set (F1 score ≈ 0.8), and our best result is lower than the highest score 0.89. This indicates that these two models might overfit on training and validation data. We also notice that the performance of two simpler models is comparable to the two complex model, and we are able to obtain this level of performance within fewer epochs. This pattern does not support our hypothesis that using more complex model architectures like Bi-LSTM introduces substantial improvement on prediction tasks.

Another worth-noting result is that the domain-specific BERT model outperforms general BERT on identifying medical phrases only by a very small margin. The simple model with Bio-Clinical BERT slightly outperforms the one with Distil BERT on the test set (0.789 vs 0.784), but not on the validation set (0.911 vs 0.915). Unlike (Alsentzer et al., 2019; Rasmy et al., 2021), We did not observe substantial advantages of using domain-specific BERT.

Figure 3: Model Performance on String-level Prediction



## 5.2 Case-Specific Results

We present the predictive performance of the best model, Bio-Clinical BERT + LSTM, on each case in Table 5. This model achieves relatively good performance on most of the case in the validation set. The highest F1 score (0.952) is observed on case 6 and 8. We also investigate feature-level performance for two cases on which the model underperforms, case 1 and case 5. It turns out that the model has a low performance on identifying associated-nausea (Recall = 0.38, F1 = 0.53) and Associated-throat-tightness (Recall = 0.61, F1 = 0.75).

Table 5: Predictive Performance of Bio-Clinical BERT Bilstm Model on the Validation Set

| case_num | Precision | Recall | Micro-F1 |
|---|---|---|---|
| 0 | 0.851 | 0.919 | 0.884 |
| 1 | 0.916 | 0.914 | 0.915 |
| 2 | 0.937 | 0.935 | 0.936 |
| 3 | 0.937 | 0.945 | 0.941 |
| 4 | 0.919 | 0.948 | 0.933 |
| 5 | 0.922 | 0.834 | 0.876 |
| 6 | 0.932 | 0.974 | 0.952 |
| 7 | 0.926 | 0.969 | 0.947 |
| 8 | 0.963 | 0.941 | 0.952 |
| 9 | 0.930 | 0.954 | 0.942 |

### 5.3 Error Analysis

We conduct an error analysis on the instances where we made the most false-positive errors to identify potential patterns in errors. In Fig 4, we find that the model is not good at predicting long annotations with gaps in between, although it is questionable why this long sequence that describes the same medical fact is separated into multiple annotations. Second, the model tends to strongly associate some words with certain categories. For example, every occurrence of the word 'palpitation' is marked as being part of the feature 'Episode of Heart Racing'. Although the association is indeed valid, this shortcut taken by the model also leads to some errors. Additionally, we also find in Fig 5 that the model cannot fully capture the negation in the sequence. Although it corrects mark the phrase 'chest pain', it fails to capture the negation 'denies' that appears much earlier in the sequence. This incompleteness might propagate erroneous understanding into downstream tasks.

## 6 Limitations

Our best model achieves a promising performance on the validation set but clearly underperforms on the testing set. This performance gap indicates potential overfitting. There are two potential reasons for overfitting. First, the distribution of cases and features are different in the test set. As we describe earlier, the 1,000 labeled notes have equally distribution across ten cases but the overall dataset has a very unequal distribution. Case 3 has more than 9,000 notes, which is ten times more than that in case 1. This difference in the distribution might explain why our model fail to generalize its performance to the test set, if the test set has a distribution similar to that in the overall dataset. Second, the validation set has a relatively small size (N = 150) and it is possible that we happen to obtain an easy-to-predict validation set.

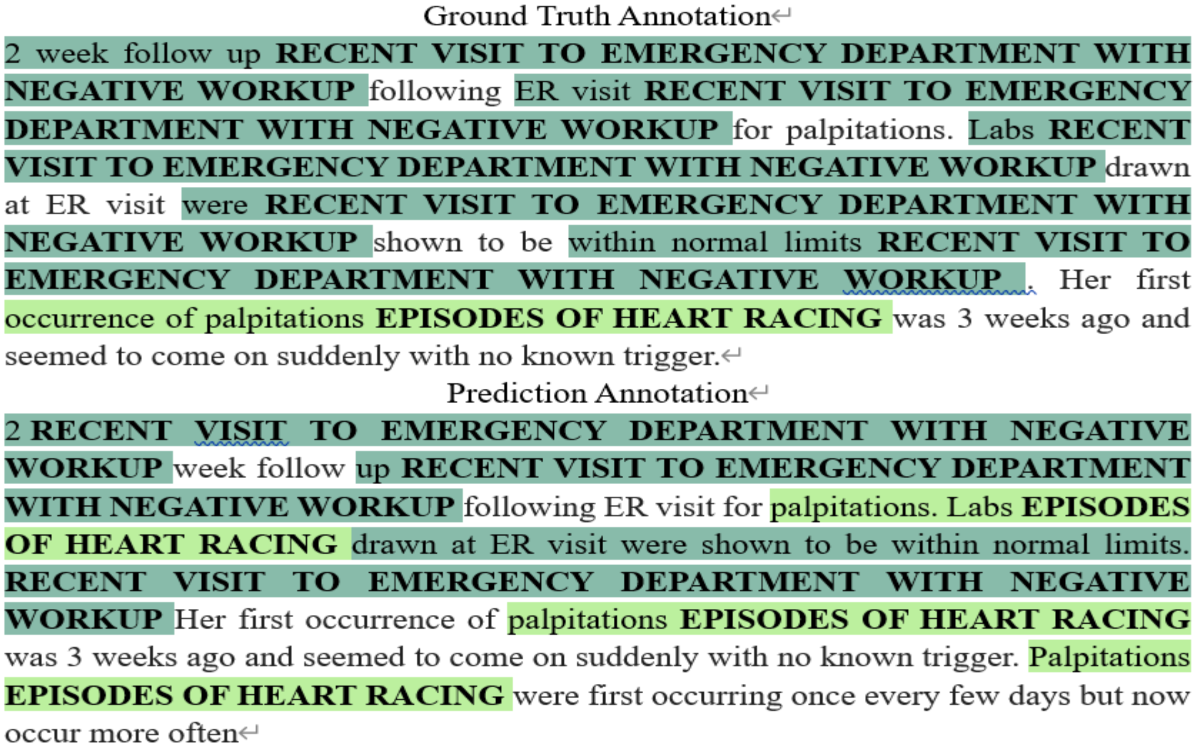
Another limitation of our implementation is that our models only use labeled data in the training process while the majority of the data (97.6%) is unlabeled. Others [2] have shown that it is possible to generate more labeled data using exact matching by regular expression. This approach, however, might shift the distribution of the training set. We examine the labeled data generated with this approach and find that it is easier to create annotations for simple features like gender, age, and family history but much harder to do so for complex features like Associated-throat-tightness. Additionally, this approach cannot effectively capture features that include negations or have large heterogeneity in expression. Thus, we did not include this additional annotated data into modeling. Recent work (Lybarger et al., 2021) shows that data argumentation approaches that use active learning to select unlabeled data could improve the downstream performance. Models trained with this augmented dataset outperform models with other techniques like random selection or using only the original annotated dataset. We consider data argumentation with active learning as a more structured and preferable approach since it utilizes unlabeled data that contains more information. However, we do not implement it because labeling selected instances requires domain expertise about clinical scenarios, which is not possessed by this group of researchers. Future work can explore whether the data argumentation technique in Lybarger et al. (2021) generalizes to this task.

## 7 Conclusion

In this work, we develop an automated approach to support physicians in quickly identifying important medical facts in patient notes. Among the four models that we train, Bio-Clinical BERT + Bi-LSTM model achieves the highest F1 score of 0.928 on validation set, a satisfactory performance on mapping important clinical features to words in clinical patient notes. However, we conclude that there is space to improve our models to achieve better performance on this task. In-depth investigation on case-specific result illustrates that the model poorly predicts phrases associated with nausea and throat tightness. Noting that the performance of domain-specific BERT model does not outperform that of general BERT model significantly, we could utilize different contextual embedding models, such as RoBERTa and SpanBERT. In addition, incorporation of unlabeled data selected by active learning will increase the number of our training data that could improve our downstream performance, identifying key phrases in patient notes.

[2] see https://www.kaggle.com/wuyhbb/get-more-training-data-with-exact-match

Figure 4: Error Analysis Sample 1



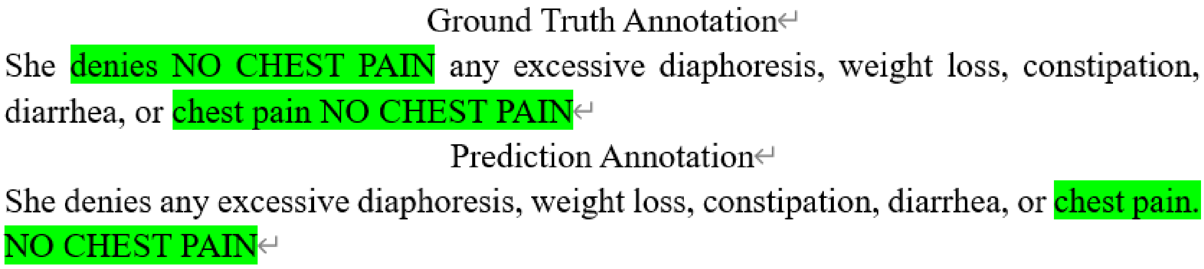
Ground Truth Annotation

2 week follow up **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** following ER visit **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** for palpitations. Labs **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** drawn at ER visit were **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** shown to be within normal limits **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** . Her first occurrence of palpitations **EPISODES OF HEART RACING** was 3 weeks ago and seemed to come on suddenly with no known trigger.

Prediction Annotation

2 **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** week follow up **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** following ER visit for palpitations. Labs **EPISODES OF HEART RACING** drawn at ER visit were shown to be within normal limits. **RECENT VISIT TO EMERGENCY DEPARTMENT WITH NEGATIVE WORKUP** Her first occurrence of palpitations **EPISODES OF HEART RACING** was 3 weeks ago and seemed to come on suddenly with no known trigger. Palpitations **EPISODES OF HEART RACING** were first occurring once every few days but now occur more often

Figure 5: Error Analysis Sample 2



Ground Truth Annotation

She denies NO CHEST PAIN any excessive diaphoresis, weight loss, constipation, diarrhea, or chest pain NO CHEST PAIN

Prediction Annotation

She denies any excessive diaphoresis, weight loss, constipation, diarrhea, or chest pain. NO CHEST PAIN

## 8 References

### References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Simard P. Frasconi P. Bengio, Y. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks (pp. 157-166*.

Bonham VL Blizinsky KD. 2018. Leveraging the learning health care model to improve equity in the age of genomic medicine. *Learning Health Systems*, 2(1).

J. Chiu and E. Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

Chapman W. W. McDonald C. J. Demner-Fushman, D. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Precioso F. Riveill M. Pighetti R. Florez, E. 2018. Named entity recognition using neural networks for clinical notes. *International Workshop on Medication and Adverse Drug Event Detection (pp. 7-15*.

Chao-Chun Hsu, Shantanu Karnwal, Sendhil Mullainathan, Ziad Obermeyer, and Chenhao Tan. 2020. Characterizing the value of information in medical notes. *arXiv preprint arXiv:2010.03574*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Chen J. Xu Z. Chen Y. Hao T. Liang, Z. 2019. A pattern-based method for medical entity recognition from chinese diagnostic imaging text. *Frontiers in Artificial Intelligence*.

Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. 2021. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *Journal of Biomedical Informatics*, 113:103631.

Kevin Lybarger, Meliha Yetisgen, and Mari Ostendorf. 2018. Using neural multi-task learning to extract substance abuse information from clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2018, page 1395. American Medical Informatics Association.

S. Fern'andez H. Bunke M. Liwicki, A. Graves and J. Schmidhuber. 2007. A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In *In Proceedings of the 9th International Conference on Document Analysis and Recognition*.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*.

Artem Shelmanov, Vadim Liventsev, Danil Kireev, Nikita Khromov, Alexander Panchenko, Irina Fedulova, and Dmitry V. Dylov. 2019. Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 482–489.

X. Wang Q. Chen B. Tang Z. Wang Z. Liu, M. Yang and H. Xu. 2017. Entity recognition from clinical texts via recurrent neural network. *BMC medical informatics and decision making, 17(2):67*.

Imed Zitouni. 2014. Natural language processing of semitic languages. *Springer*.