Final Report — An anchoring approach to medical information extraction using Clinical BERT embeddings

Alison Borenstein ALISONRB@MIT.EDU
Abby Garrett ABBYG@MIT.EDU
Hamza Tazi Bouardi HTAZI@MIT.EDU
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA, United States

1. Introduction

The ability to transform digitized clinical text (e.g., medical discharge summaries and progress notes) into structured representations of extracted medical concepts, clinical assertions and, ultimately, the relations between them is in great demand in the medical field. Such a system would enable physicians to quickly and easily identify key aspects of a patient's clinical condition and history, allowing for easier and more efficient decision making and improved quality of care. Furthermore, annotated medical corpora are expensive to create, so an automated system for text processing and extracting relevant medical information would allow for their development to be much less time consuming and much more resource efficient. There are many technical challenges associated with extracting medical concepts from unstructured data including, but not limited to: choosing the right model for the task, properly processing text data to be fed to the model, creating appropriate features for the model, and tuning the chosen model appropriately and extensively (raising the question of available computational power).

In 2010, Informatics for Integrating Biology and the Bedside (i2b2), as part of the National NLP Clinical Challenges (n2c2), issued a contest (Uzuner et al., 2011) to investigate:

- **Extraction of medical problems, tests, and treatments**
- Classification of assertions made on medical problems
- Relations of medical problems, tests, and treatments

Participants in the challenge could focus on one or more of the three tasks, but given the complexity of the three part problem (concepts to assertions to relations) and the later parts being heavily dependent on the initial step of Named Entity Recognition (NER), in our study, we concentrated on improving upon the medical concept extraction task. Traditionally, the task of concept extraction from medical annotations has been tackled with rule-based systems. Recent natural language processing (NLP) developments, however, have led researchers to explore the power of Transformers, such as Bidirectional Encoder Representations from Transformers (BERT), for the task (Devlin et al., 2018). (Alsentzer et al., 2019) developed an open source clinicalBERT which was applied to the 2010 i2b2 dataset. However, their main focus was on developing an open source tool, not a new state-of-the-art.

The goal of this project was to utilize the open source clinicalBERT model and to explore strategies for improving performance on the concept extraction task of the 2010 i2b2 challenge. Our first area of exploration was inspired by the concept of learning from noisy labels under an anchor-and-learn framework, which has been utilized previously in the healthcare realm for electronic medical record phenotyping (Halpern et al., 2016). Motivated by this approach, we decided that prior to fine-tuning for the NER task on the 2010 i2b2 dataset, we would perform additional training of the clinicalBERT model using data with noisy labels. To accomplish this, we utilized discharge summary notes from the MIMIC III database and derived labels for these notes based on exact phrase matching with medical terms from the Unified Medical Language System (UMLS). Although this additional training set was far from perfect, our hypothesis was that we could still achieve performance gains for our final task due to (1) the amount of extra data available for training on and (2) knowing that the gold-labeled i2b2 dataset labels were loosely based on UMLS semantic types. If performance gains were indeed achieved, we would have demonstrated an efficient way to learn these concept representations without having to rely on hand-annotated gold-labeled data for training. The second area of exploration we focused on was the implementation of additional architectures for the final concept classification task. While BERT is most often used with a simple Linear layer for classification, we wanted to implement bidirectional Long Short-Term Memory with a Conditional Random Field (bi-LSTM+CRF) to investigate whether or not this more complex architecture improved performance.

Although we were unable to prove through this study that a noisy label learning approach is effective for medical concept extraction, we did observe slight performance gains with a warm start model and a Linear classification layer compared to a non-warm start model with a Linear classification layer. The best performance we achieved on the final downstream NER task utilized a non-warm start model with a bi-LSTM+CRF architecture. Based on our results and findings, we believe that future research efforts could focus on (1) augmenting the size of the UMLS/MIMIC III dataset used for additional training, (2) experimenting with additional initialization methods, or (3) extending the investigation of how generalizable a model trained on noisy labels may be on other medical datasets.

2. Related Work

In the 2010 i2b2 challenge, nine of the top 10 systems of the competition used Conditional Random Fields (CRFs) to determine the concept boundaries and Support Vector Machines (SVMs) to classify the concept type as a problem, test, or treatment. While 22 systems were developed for concept extraction in the 2010 competition (Uzuner et al., 2011), since then, researchers have tried to improve the obtained results still using hand-engineered features and CRF-based NER systems (Wang et al., 2018; Gurulingappa et al., 2010). Others have tested Neural Network based approaches using a bi-LSTM on top of a CRF (Florez et al., 2018; Zhu et al., 2018). Even more recently, this bi-LSTM+CRF architecture was outperformed by other researchers who tackled the concept extraction problem using Transformer-based methods (Si et al., 2019) by pre-training the BERT architecture on MIMIC III clinical notes data. (Si et al., 2019) pre-trained off-the-shelf BERT models and

used additional bi-LSTM layers to obtain an **Exact F1-Score of 0.9025**, corresponding to the current state-of-the-art (SOTA).

Additional recent contributions to this work by (Alsentzer et al., 2019) involved pre-training biomedical-text-derived BERT models, BioBERT (Lee et al., 2019), on clinical notes and discharge summaries to demonstrate the value of specialized BERT models. In our work, we aimed to combine the strengths of various models and architectures and to focus on improvement of the downstream task through extensive fine-tuning. While previous research has utilized a bi-LSTM+CRF architecture with non-contextual embeddings or BERT (with either a simple Linear classification layer or with bi-LSTM layers), to our knowledge, implementation of BERT with a bi-LSTM+CRF architecture for NER on the 2010 i2b2 dataset has not been completed. Furthermore, previous researchers tackling this challenge have not explored additional supervised training on medical corpora with noisy labels, and so we explored this by utilizing work from Halpern et al. (2014), which introduced the process of using anchors to generate a labeled dataset. Through this endeavor, we investigated the possibility of reducing or eliminating the need for gold-standard labeled data.

3. Methods

For our work, we used state-of-the-art, Transformer-based methods for Named Entity Recognition. The particular architecture we chose to use to obtain contextual word representations was Bidirectional Encoder Representations from Transformers, also known as BERT (Devlin et al., 2018). This is a general purpose NLP model, with which more or less any NLP task could be performed if fine-tuned properly. A transformer (Vaswani et al., 2017), which is the principal building block of BERT, is a non-recurrent Sequence-to-Sequence NLP model which uses an encoder-decoder based architecture as well as self-attention to enable the model to observe entire sequences, whereas other representation algorithms like GLoVE (Pennington et al., 2014) or Word2Vec (Mikolov et al., 2013) would learn a single representation for a given word without being able to grasp any long-term dependencies in long sentences. Given the length of clinical notes, using BERT seemed like a logical choice.

Our approach to the concept extraction problem had two parts. First, we used pre-trained BERT embeddings on clinical data provided by (Alsentzer et al., 2019) and performed additional training on a dataset with noisy labels derived by identifying anchors within MIMIC III discharge summaries. Second, we fine-tuned the architecture on the 2010 i2b2 dataset. We detail the methods related to the two parts of our study below.

3.1. "Post"-Pre-Training on MIMIC III Data

As explained in further detail in the Data section below, the 2010 i2b2 challenge provided participants with a training set of 349 documents and a testing set of 447 documents. Hand-labeling data of this type is a time-consuming task, so there were limitations to the amount of data previous researchers have used to fine-tune their models for the final NER task. We leveraged the wealth of data available in MIMIC III, in the form of discharge notes, to perform additional training for the NER task. The MIMIC III database contains millions of medical notes, approximately 60,000 of which are discharge summaries. Using

discharge summaries alone would provide us with 74 times the amount of data provided in the challenge. However, due to time and computational power constraints, we were only able to utilize 1,000 of these summaries. Even still, this subset provided us with **more than twice** as much data as what was provided by the 2010 i2b2 challenge. Our hypothesis was that utilizing this additional training data would enhance performance on the final concept extraction task.

To transition these 1,000 notes to a noisy dataset, we used UMLS semantic types to label MIMIC III notes for NER (with the same labels as i2b2) based on exact phrase matching, followed by a frequency analysis which resulted in removing a few nonsensical labels that were common throughout the documents (i.e. "his" was originally labeled "B-Problem" but was changed to "O" following our frequency analysis). This framework is similar to the anchor variable approach used by (Halpern et al., 2014) in that the presence of a UMLS concept constituted application of the entity label for which that concept belonged. The benefit of such an approach is that, once a pipeline is defined, manual labeling of data is no longer needed before training a classifier. Upfront, however, this task required significant effort to i) determine which UMLS terms to match on, ii) develop a pipeline for matching, and iii) format the labels of matched words and phrases to match the 2010 i2b2 format. With this dataset (which we call UMLS/MIMIC III), we treated the anchors as noisy labels to further train clinicalBERT for NER. In doing so, we experimented with both a Linear and a bi-LSTM+CRF top-layer architecture. We considered this added "post"-pre-training step as a warm start of weights for the prediction we were ultimately interested in.

3.2. Downstream NER on 2010 i2b2 Data

In the "post"-pre-training step described above, we split the noisy-labeled dataset into a training and validation set and conducted a grid search over hyperparameters to determine which parameters led to the best validation performance. We then saved and used the two best versions of the "post"-pre-trained clinicalBERT as the final models, with updated (or "warm-started") weights, to perform NER on the 2010 i2b2 dataset. The main focus during this step of our study was to investigate how different classification architectures affected performance. Thus, we once again experimented with both a linear layer and a bi-LSTM+CRF architecture for classification. Furthermore, as described in greater detail in our Results section, we also explored the effects of fine-tuning, so we conducted a grid search to observe which parameters led to optimal performance for the final task.

4. Data

For our first analysis effort, we utilized the MIMIC III Database and the Unified Medical Language System (UMLS). The MIMIC III database is a freely-available database that contains de-identified health-related data compiled from over 40,000 patients that were admitted to critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database contains approximately 60,000 patient discharge summaries. The UMLS is a set of files containing health and biomedical vocabularies. Combining these two data sources, we developed a pipeline for processing the MIMIC III discharge notes data such that we matched the output label format of the 2010 i2b2 dataset, where each word was

labeled as a medical problem, treatment, test, or as not belonging to any of these medical concepts. We followed a standard "inside, outside, beginning" (IOB) format for tagging tokens in chunks. To determine which terms in the discharge summaries were labeled with which entity, we used vocabulary from specific UMLS semantic types (as detailed in the Appendix), which resulted in a label vocabulary of 800,000 medical problems, 700,000 treatments, and 100,000 tests. Then, for our additional training task, we used a subset of 1,000 "noisy-labeled" discharge summaries. We separated these discharge summaries into two datasets:

- A training set, consisting of 900 documents with 371,887 concept labels
- A validation set, consisting of 100 documents with 41,363 concept labels

For the second step of our analysis, we utilized datasets from the 2010 i2b2/VA n2c2 Relations Challenge. The data for this challenge included discharge summaries from Partners Healthcare and from Beth Israel Deaconess Medical Center (MIMIC II Database). Discharge summaries and progress notes from University of Pittsburgh Medical Center were also included in the datasets. Medical records were randomly split by institution and document type. All records were fully de-identified and manually annotated with concept, assertion, and relation information. For the original challenge a training set of 349 documents and a testing set of 447 documents were released, but the current release of the dataset contains only 170 training documents and 256 testing documents. We separated this data into three datasets:

- A training set, consisting of 152 documents with 29,679 concept labels
- A validation set, consisting of 18 documents with 5,079 concept labels
- A testing set, consisting of 256 documents with 64,811 concept labels

Given un-annotated text from patient reports similar to those described above, we aimed to develop a well-performing system that can extract text that corresponds to medical concepts (named entities, which consist of: medical problems, treatments, and tests). The descriptions of these three named entities can be found below and were drawn from the Concept Annotation Guidelines provided by i2b2 with the 2010 data. Each of the concepts are loosely based on certain UMLS semantic types, but may also include instances not covered within UMLS.

- **Medical Problems** are phrases that contain observations made by patients or clinicians about the patient's body or mind that are thought to be abnormal or caused by a disease.
- **Treatments** are phrases that describe procedures, interventions, and substances given to a patient in an effort to resolve a medical problem.
- **Tests** are phrases that describe procedures, panels, and measures that are done to a patient or a body fluid or sample in order to discover, rule out, or find more information about a medical problem.

5. Evaluation

In order to quantify the performance of our models, we used metrics such as Accuracy, Precision, Recall, and F-Measure (or F1-Score). These are the metrics that were used and reported in the 2010 i2b2 challenge. Given that the version of the dataset released for the challenge is different from the version we used in this study, we could not make a direct comparison with their results, however we wanted to remain consistent with their general evaluation methodology. It is worth noting that the best concept extraction system from the 2010 challenge achieved an exact F-measure of **0.852**.

The primary metrics of interest, which we will discuss later in this report, are the **Exact** Micro-Averaged F-Measure and Accuracy for all concepts together, as well as the F-Measure for problems, treatments, and tests separately. Notice that we do not report Precision or Recall for the sake of simplicity, as has been done extensively in the literature (Alsentzer et al., 2019), (Si et al., 2019). While (Alsentzer et al., 2019) and (Si et al., 2019) compute evaluation metrics at the entity level, we decided to compute metrics at the word-piece level. BERT's tokenizer tokenizes words into word-pieces, and converting the word-pieces back to whole word entities proved to be too complex given the timeline of our project (the "ner_eval" section of the (Alsentzer et al., 2019) GitHub Repository demonstrates the complexity of this process). An example of how we predicted is as follows: suppose the word "chloroquine" was separated into "chloro" and "quine" by the tokenizer: this gave us two word-pieces with two labels [B-Treatment, X]. At prediction time, we only considered the first part of the word (thus the one not corresponding to X) for exact match predictions, i.e. "chloro" for the label B-Treatment.

Given the differences in our evaluation schemes, we were not able to compare our results to those in (Alsentzer et al., 2019) or in (Si et al., 2019), as it would not be a fair comparison, especially since our evaluation methodology makes the problem even harder than it already is. Therefore, we considered the best performance we obtained using the raw version of clinicalBERT and a Linear Classifier on the i2b2 data to be our baseline performance to beat, and we compared it to:

- No UMLS/MIMIC III Warm Start, but using a bi-LSTM+CRF top layer instead of a linear layer
- UMLS/MIMIC III Warm Start using a linear layer, then downstream i2b2 using either a linear layer or a bi-LSTM+CRF
- UMLS/MIMIC III Warm Start using a bi-LSTM+CRF, then downstream i2b2 using either a linear layer or a bi-LSTM+CRF

In addition to general performance scores, we evaluated our model in terms of how much gold labeled data was required to achieve adequate performance. The MIMIC III dataset provided noisy labels to train on, but the i2b2 data was hand-labeled and thus is much more accurate. We explored how well our model performed using varying levels of the i2b2 data (0%, 20%, 50%, 70%, and 100%) in the final training step to see if gold-labeled data was indeed necessary for producing a reliable medical concept extraction model.

6. Results

6.1. Data Processing and Experimental Setup

Our project involved a few key data and modeling components: the 2010 i2b2 dataset, the UMLS/MIMIC III "noisy-labeled" dataset, and a fine-tuned clinicalBERT model. The i2b2 dataset, which began with individual files for each document, was transformed into a csv file with each row containing a word and sentences separated by a blank row. Generating a "noisy-labeled" dataset from MIMIC discharge notes to match this data format was no small task. As discussed in our Data section, to create the necessary labels, words and phrases, the discharge notes needed to be matched with terms from several UMLS semantic types. For each word, potential labels included: O, B-treatment, B-test, B-problem, I-treatment, I-test, and I-problem.

To accomplish this task, we utilized "PhraseMatcher" from the spaCy library. The algorithm we created parses each note into sentences of length 126 or 64 and labels each word according to the i2b2 label structure. Another important thing to note is that BERT's tokenizer splits words into smaller chunks of words, and we therefore had to add another label, 'X', on top of the two special labels specific to BERT's architecture ('[CLS]' and '[SEP]'). We made sure that this tokenization, which created much longer sequences (because of the additional 'X' labels), would still result in having sentences of maximum length 128 (including the special labels), each of them starting with the '[CLS]' token and potentially ending with a padding token, which is commonly labeled as 'O.' As explained earlier, evaluation is an exact match on word-piece tokens, which therefore ignores word-piece tokens with ground truth label in {'[CLS]', '[SEP]', 'X'}.

6.2. Results of "Post"-Pre-Training on MIMIC III Data

As discussed, UMLS semantic types were used to label the MIMIC III dataset. Although approximately 60,000 discharge summaries exist within the MIMIC III database, due to GPU memory and runtime constraints, we were only able to perform additional clinicalBERT training using 1,000 discharge summary notes. Despite not being able to take full advantage of the wealth of data present in the MIMIC III database, we believed that if we could show improvement in performance on our final task even using just a subset of the MIMIC III notes then we would have shown the value in this additional training effort. Another limitation, however, was the lack of access to multiple GPUs: we had access to one K80 or P100 GPU with 16GB of RAM (on Google Colab), which only allowed us to use a maximum batch size of 64 for the warm start training on UMLS/MIMIC III. This setup already took approximately 2 hours to run per epoch. So, while we could have processed more data into datasets, we would not have been able to run the experiments given the computational resources we had and the timing of the project.

After creating a "noisy-labeled" data set through exact matching of terms, and after tokenizing them using the BERT tokenizer (into word-piece tokens), we split into a training set and a validation set of sentences of maximum length 128 (including padding and spe-

cial BERT tokens). The sizes of these two sets, depending on the number of discharge summaries processed, can be found in Table 1 below.

# MIMIC Notes	Train Size	Validation Size
1,000	195,853	21,762
2,000	395,973	43,998

Table 1: Train and validation size of datasets fed to clinicalBERT for post-pre-training

Note that we did not run any model on 2,000 MIMIC notes because of runtime and GPU memory issues, but this is something that should be explored if more resources are available, as we will discuss later in this report.

We retrieved a pre-trained version of clinicalBERT (Alsentzer et al., 2019) and fine-tuned it with a multiclass classification linear layer, or a bi-LSTM+CRF layer; these two possibilities were the most successful according to our literature review. We trained clinicalBERT, tuning all of its parameters, to perform NER on the labeled MIMIC III notes, and performed a grid search with the following hyperparameters:

- Maximum sentence size $\in \{128\}$ (including special tokens CLS & SEP)
- Number of Epochs $\in \{3, 4, 5\}$
- Batch size $\in \{32, 64\}$
- AdamW Optimizer: Learning Rate (LR) $\in \left\{2 \cdot 10^{-5}, 3 \cdot 10^{-5}, 5 \cdot 10^{-5}\right\}$
- Gradient Clipping (GC) $\in$ {True, False} with a maximum gradient norm fixed to 2
- Bi-LSTM Hidden Size (HS) $\in \{512, 768\}$, only used when applicable

In both cases, we evaluated Accuracy, Micro-Averaged F1-Score, and per-label F1-Score on the validation set. The results can be found in Table 2 below, and all correspond to 1,000 notes extracted from MIMIC III discharge summaries.

Hyperparameters	Top Layer	Micro-Avg F1-Score	Accuracy	Treatment F1-Score	Problem F1-Score	Test F1-Score
(4, 64, $3 \cdot 10^{-5}$, False, NA)	Linear	**0.9716**	0.9879	**0.9747**	0.9714	0.9661
(4, 64, $2 \cdot 10^{-5}$, True, NA)	Linear	0.9713	0.9876	0.9735	**0.9718**	0.9632
(4, 64, $3 \cdot 10^{-5}$, False, 512)	bi-LSTM+CRF	0.9711	**0.9882**	0.9730	0.9710	**0.9683**

Table 2: Results with BERT on the validation set of MIMIC III / UMLS terms for top 3 sets of hyperparameters (Epochs, Batch Size, LR, GC, HS)

It seems from the results above that a more complex top layer such as a bi-LSTM+CRF layer does not necessarily lead to better performance than a simpler Linear classifier, at least in terms of Micro Averaged F1-Score. Thus, we consider two different linear layers which we will reference with "Warm Start with Linear 1" to indicate the best linear layer model in Table 2, and "Linear 2" to indicate to the second best. Unfortunately, due to time constraints, we were not able to run experiments with the best bi-LSTM+CRF Warm Start (third model in Table 2) on the i2b2 data, but this is something that should be explored in future endeavors.

6.3. Results of Downstream NER on 2010 i2b2 Data

Recent advances in NLP, which have led to transformer-based models such as BERT, have enabled researchers to show that contextualized embeddings can be combined with simple downstream models to achieve performance that outperforms complex models, such as a bidirectional Long Short-Term Memory with Conditional Random Field (bi-LSTM+CRF) architecture. Thus, as a baseline, we began by performing the downstream NER task using a Linear layer on top of the last BERT layer to determine the most probable label for each token. Then, our experiments consisted of using a "warm-started" clinicalBERT with UMLS/MIMIC III data, and fine-tuning it to perform NER on the 2010 i2b2 dataset. We will discuss the results of a Linear top layer, and a bi-LSTM+CRF top layer, used to fine-tune a warm-started clinicalBERT (and the warm start model will be specified). Each table will contain the baseline model characterized by no warm start, to which we compare the results of our anchoring methods. In terms of hyperparameters fitted, we performed a grid search on the same hyperparameter space as earlier in Section 6.2 when applicable, *except for the epochs and batch size*, where we explored in $\{2, 3, 4, 5\}$ and in $\{16, 32\}$, respectively. We only report the top 3 architectures (in terms of Micro-Averaged F1-Score) for each type of top layer used.

6.3.1. Linear Multiclass Classification Layer

Here we report the top three architectures (in terms of Micro-Averaged F1-Score on the i2b2 test) using a warm start and a linear layer on top of BERT to perform classification. The first row of Table 3 below corresponds to the baseline model.

Hyperparameters	Warm Start	Micro-Avg F1-Score	Accuracy	Treatment F1-Score	Problem F1-Score	Test F1-Score
(3, 16, $2 \cdot 10^{-5}$, False, NA)	None, Baseline Linear	0.8424	0.9526	0.8525	**0.8604**	0.8621
(5, 16, $5 \cdot 10^{-5}$, False, NA)	Linear 2	**0.8436**	0.9515	**0.8609**	0.8478	**0.8655**
(5, 32, $3 \cdot 10^{-5}$, False, NA)	Linear 2	0.8431	**0.9531**	0.8608	0.8554	0.8583
(5, 16 $2 \cdot 10^{-5}$, False, NA)	Linear 2	0.8416	0.9526	0.8597	0.8518	0.8611

Table 3: Results on the test set of i2b2 2010 for top 3 sets of hyperparameters (Epochs, Batch Size, LR, GC, HS) with a Linear Multiclass Classifier

As we can see, it appears as though the Linear 2 Warm Start performs best on the i2b2 task. Although Linear 1 is the best performing model on the UMLS/MIMIC III dataset, transfer learning does not seem to agree for the i2b2 data. However, **while there is an improvement in most metrics**, it is very small, and thus the significance of our findings is inconclusive for this experiment.

6.3.2. bi-LSTM+CRF Layer

In addition to performing the downstream task utilizing the updated weights from the additional training task along with a linear classification layer, we also utilized a bi-LSTM+CRF architecture for the concept classification task. The bi-LSTM allowed for taking information from right and left contexts into account, and feeding the output of the bi-LSTM to the CRF classifier resulted in predictions for the most likely sequence of labels as opposed

to the most likely label for each word independently (as in the linear classifier case). We present the results obtained using this more complex architecture in Table 4 below; unfortunately, due to time and resources constraints, we have only been able to run the grid search with a Linear 1 Warm Start, but it could be interesting to look into Linear 2 and bi-LSTM+CRF Warm Start. The general baseline still corresponds to the same one as shown previously (first row), but we also include a baseline for a bi-LSTM+CRF layer (second row), which corresponds to no warm start, but using a bi-LSTM+CRF layer on top of the raw clinicalBERT, instead of a linear layer as in 6.3.1.

Hyperparameters	Warm Start	Micro-Avg F1-Score	Accuracy	Treatment F1-Score	Problem F1-Score	Test F1-Score
(3, 16, $2 \cdot 10^{-5}$, False, NA)	None, Baseline Linear	0.8424	0.9526	0.8525	0.8604	0.8621
(2, 16, $3 \cdot 10^{-5}$, True, 768)	None, Baseline bi-LSTM+CRF	**0.8546**	**0.9555**	**0.8683**	0.8647	0.8642
(3, 16, $3 \cdot 10^{-5}$, True, 768)	Linear 1	0.8460	0.9523	0.8547	**0.8655**	**0.8695**
(2, 16, $3 \cdot 10^{-5}$, True, 512)	Linear 1	0.8429	0.9529	0.8507	0.8464	0.8668
(5, 32, $5 \cdot 10^{-5}$, False, 512)	Linear 1	0.8428	0.9512	0.8565	0.8540	0.8513

Table 4: Results on the test set of i2b2 2010 for top 3 sets of hyperparameters (Epochs, Batch Size, LR, GC, HS) with a bi-LSTM+CRF Layer

As we can see, it seems like a linear warm start with a bi-LSTM+CRF top layer does not help improve the performance, and quite the contrary, as the bi-LSTM+CRF baseline remains the best performer of all models, except for the F-measures for the problem and test categories.

6.3.3. Evaluation on varying levels of 2010 i2b2 data

While we did not see significant performance gains from implementing the "post"-pre-training step, or adding additional architecture, we obtained promising results in reducing the need for hand-labeled data. By utilizing "noisy-labeled" data in training a clinicalBERT model with a linear top layer[1] for NER, our model produced a much higher F1-Score with 0% of the gold standard data in training than a model without the warm start. These results suggest that a considerable amount of learned information was transferred as a result of the additional training on UMLS/MIMIC III to create a warm-started model.

Train Size	Test F1-Score with WS	Accuracy with WS	Test F1-Score no WS	Accuracy no WS
0	**0.1957**	0.6692	0.033	0.2298
2,867 (20%)	**0.7905**	0.9380	0.7765	0.9380
7,168 (50%)	0.8156	0.9467	**0.8242**	0.9497
10,035 (70%)	**0.8275**	0.9484	0.826	0.9492

Table 5: Results of different architectures with varying amounts of gold standard labeled i2b2 2010 data used in training[2]

While the score is still much lower than necessary for implementation, Table 5 above reveals the reduced need for hand-labeled data, as the warm start model outperforms the other model in all scenarios but one, thus demonstrating its superior generalization properties.

1. This model on which these experiments were realized corresponds to Linear 1 Warm Start in Tables 3, 4
2. Percentage in the Train Size column corresponds to the percentage of total training data initially available

7. Discussion

The aim of this project was to explore an anchoring approach to medical information extraction using different architectures that have proven to be efficient for NER tasks, and to try to bring substantial proof of how anchor learning (and thus transfer learning) can lead to improved downstream performance. From a technical standpoint, our experiments seem to show that a model without a warm start can perform as well (cf. Table 3), if not better (cf. Table 4), than one with a warm start obtained using the UMLS/MIMIC III dataset. Despite the performance improvement observed when using a bi-LSTM+CRF architecture in the baseline setup (vs. a Linear layer), as mentioned by (Si et al., 2019), BERT contains sufficient label correlation in and of itself, and a CRF might be unnecessary to obtain such performance gains. This observation is supported by the fact that only a bi-LSTM was utilized to obtain the current SOTA. Therefore, it seems as though our hypothesis of an anchor approach improving downstream learning cannot be confirmed from our experiments.

However, there are a few research directions that could be explored in order to make sure that this methodology does not work as we hypothesized it would, namely:

- While we have trained the warm start clinicalBERT with 1,000 MIMIC notes, we believe that using all 60,000 notes could prove very helpful for both the "post"-pre-training performance and the downstream NER task on i2b2 data. This would, however, require many more GPUs (or at the very least a few days of GPU time) and some parallelization.
- Another idea we have not had the opportunity to test is alternative weight initialization methods in the top layers, namely Xavier normal or uniform initialization (Glorot and Bengio, 2010) as it has proven to be extremely efficient for both Linear layers and LSTM cells weight initialization.
- In order to compare to current SOTA (Si et al., 2019) and other published papers, the post-processing of tokenized entries at prediction time should be consistent with the way it was during the competition. Namely, one should do predictions at a word level, instead of at the word-piece level, ignoring all word-pieces except the first one.
- Finally, one could further explore how generalizable our warm start model is on other datasets, and evaluate how much it can reduce the necessity for gold-labeled data in order to obtain good performance.

While modeling performance gains were not a significant part of this research project, the promising results from varying the level of gold standard data identify an interesting point of discussion. As seen in section 6.3.3, minimal performance gains are realized by increasing the amount of hand labeled data beyond 20%, specifically when utilizing the warm start. By implementing "post"-pre-training with the noisy dataset, using no hand-labeled notes increases the F-1 Score by 0.16 as compared to no warm start. Continuing to reduce the need for hand-labeled data increases the viability of implementing an NER model, as it reduces the expense of building an accurate model.

From a clinical and practical standpoint, the integration of our modeling efforts into existing clinical practice is an important consideration. From a physician's perspective, being able to identify the problems, treatments, and tests mentioned in medical notes would be especially helpful when a patient is admitted and the physician is seeking a quick, yet comprehensive sense of their medical history. If information could be provided on the number of times certain events in these categories were mentioned in their chart, a medical professional could easily get a more informative "problem list" for that patient relative to the standard problems that populate current lists in the EPIC electronic medical record system. Additionally, automatically identifying labels for concepts would allow future automated systems to learn relationships between problems and tests through their co-occurrence. Therefore, although our work is just the first step toward the clinical implications we have discussed, we believe our modeling efforts could easily be extended to create systems viable of being integrated into practice.

8. Acknowledgements & Member Contributions

We would like to thank our mentor, Chloe O'Connell, for the clinical guidance she provided us with throughout the project. Her insight was extremely valuable, especially in terms of understanding how our efforts could be integrated into existing clinical practice. We would also like to thank the entire teaching team of 6.871/HST.956 for helping us to formulate our problem and for providing continuous support toward achieving our project goals.

Regarding member contributions, all three members performed the literature review, then Abby and Alison spent the first few weeks figuring out an efficient way to pre-process all of the data (especially for the UMLS/MIMIC III task), while Hamza was implementing a general purpose pipeline to tokenize and prepare dataloaders, as well as the first clinicalBERT + linear layer model. All three members then spent a lot of time running experiments and keeping track of the results on both Google Colab and Engaging (MIT Sloan's cluster, which only has 4 GPUs available). Then, Alison and Abigail implemented the bi-LSTM+CRF layer on top of BERT, while Hamza ran more and more experiments, and paved the way for new experiments with the new architecture. Finally, all reports and presentations were created by the whole team.

9. Appendix

Concept Type	**UMLS Semantic Types**
Problem	Acquired Abnormality, Anatomical Abnormality, Cell or Molecular Dysfunction, Congenital Abnormality, Disease or Syndrome, Experimental Model of Disease, Finding, Injury or Poisoning, Mental or Behavioral Dysfunction, Neoplastic Process, Pathologic Function, Sign or Symptom, Organ or Tissue Function
Treatment	Therapeutic or Preventive Procedure, Clinical Drug, Health Care Activity
Test	Diagnostic Procedure, Laboratory Procedure, Laboratory or Test Result

Table 6: UMLS Semantic Types used for each concept

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Edson Florez, Frederic Precioso, Michel Riveill, and Romaric Pighetti. Named entity recognition using neural networks for clinical notes. In Feifan Liu, Abhyuday Jagannatha, and Hong Yu, editors, *Proceedings of the 1st International Workshop on Medication and Adverse Drug Event Detection*, volume 90 of *Proceedings of Machine Learning Research*, pages 7–15. PMLR, 04 May 2018. URL http://proceedings.mlr.press/v90/florez18a.html.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2 2010.

H. Gurulingappa, M. Hofmann-Apitius, and J. Fluck. Concept identification and assertion classification in patient health records. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, 2010.

Yoni Halpern, Youngduck Choi, Steven Horng, and David Sontag. Using anchors to estimate clinical state without labeled data. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, 2014.

Yoni Halpern, Steven Horng, Youngduck Choi, and David Sontag. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, 23(4):731–740, 04 2016. ISSN 1067-5027. doi: 10.1093/jamia/ocw011. URL https://doi.org/10.1093/jamia/ocw011.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Sep 2019. ISSN 1460-2059. doi: 10.1093/bioinformatics/btz682. URL http://dx.doi.org/10.1093/bioinformatics/btz682.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. Enhancing clinical concept extraction with contextual embedding. *CoRR*, abs/1902.08691, 2019. URL http://arxiv.org/abs/1902.08691.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang D Liu. Clinical information extraction applications: A literature review. January 2018.

Henghui Zhu, Ioannis Ch. Paschalidis, and Amir Tahmasebi. Clinical concept extraction with contextual word embedding, 2018.