Predicting ICD-9 Codes Using Transfer Learning on Substructures of Patient Notes

Caroline Barker
New York University
cb3991@nyu.edu

Ellie Haber
New York University
elh391@nyu.edu

Andrew Liang
New York University
asl617@nyu.edu

Abstract

In this paper, we compare the multi-label classification performance of pre-trained embeddings fine-tuned on ICD-9 diagnosis code prediction using patient discharge summaries in the MIMIC-III dataset. The aim of our work is to analyze how the classification task is affected and possibly improved by using different word embeddings and preprocessing methodologies. We evaluate the prediction task using a case study where the top-10 most frequently occuring ICD-9 codes are classified, and use BERT-base (Devlin et al. 2019) and BioBERT (Lee et al. 2019) embeddings fine-tuned on the first 512 tokens of patient discharge summaries as baseline models. These are compared to other BERT-base and BioBERT models fine-tuned on selectively extracted sections of discharge summaries: We show that models fine-tuned on Hospital Course sections achieve a 7% increase in F1 score over our baseline models, with BioBERT HospCourse reaching the highest F1 score of 0.53, illustrating that using biomedical embeddings fine-tuned on certain subsections of MIMIC-III can improve performance for ICD-9 code prediction.

1 Introduction

1.1 Motivation

International Code for Diseases (ICD) enables a standardized coding scheme for the diagnoses and procedures contained within clinical records. The United States Healthcare pay system relies on the systematization of mapping ICD-9 codes to patient notes as a means of billing patients for various procedures and services rendered within the medical field. Machine-aided classification in this realm cuts down significant processing time, and a more accurate classifier cuts costs in damages caused by mislabeling patients (Singh et al. 2020). Our model aims to improve the reliability of automating ICD-9 code assignment of patient notes in the healthcare field by analyzing various combinations of pre-trained embeddings and preprocessing methods applied to a case study of the ICD-9 code prediction task.

1.2 Background

Recent Natural Language Processing (NLP) research contributions investigate automating text-mining and processing tasks in select fields with vocabulary specific to the domain. A focus area of prior models is clinical and biomedical data (Lee et al. 2019), where document labeling is traditionally human-resourced and prone to manual error, delays, and high costs (Singh et al. 2020).

The dual objective is to ease document handling efforts requiring knowledge of specialty language, as well as to analyze the performance of established contextualized word embedding models fine-tuned to NLP tasks in specialty domains when the model is trained on the domain's narrow corpora versus on larger general corpora. We are interested in whether the general model or the specified model performs better on tasks in the biomedical domain, and by what magnitude in NLP metrics.

Relevant research has yielded models such as MIMIC-BERT (Singh A. et al. 2020), based on the BERT architecture and pre-trained on note events/clinical notes from the MIMIC-III v1.4 critical care dataset; BioBERT, pre-trained on large biomedical corpora from PubMed abstracts; Discharge Summary BERT and Discharge Summary BioBERT, which uses BERT-base and BioBERT as baselines respectively and are trained on de-identified patient discharge summaries from MIMIC-III v1.4; and Clinical BERT and Clinical BioBERT, trained on all clinical notes from MIMIC-III v1.4 (Alsentzer et al. 2019).

Model	MedNLI	i2b2 2006	i2b2 2010	i2b2 2012	i2b2 2014
BERT	77.5%	93.9	83.5	75.9	92.8
Clinical BERT	80.8%	91.5	86.4	78.5	92.6
Discharge Summary BERT	80.6%	91.9	86.4	78.4	92.8
Bio+Discharge Summary BERT	82.7%	94.8	87.8	78.8	92.7

Table 1: Accuracy (MedNLI) and exact F1 score (i2b2) across various clinical NLP tasks. (Alsentzer et al. 2019)

Our models are a synthesis of existing models. They follow the MIMIC-BERT method of extracting the most frequent (ten) ICD-9 codes to mitigate the effects of fine-tuning on imbalanced data (some ICD-9 codes appear more frequently) and makes selective changes to which parts of patient discharge summaries to train on, altering the parameters of the Discharge Summary model process.

2 Related Work

2.1 BERT-base

The BERT-base model is a multilayer model that incorporates bi-directional transformers and masked language modeling to create contextualized word embeddings. It is pre-trained on Wikipedia and BooksCorpus, and is shown to substantially improve performance on many downstream NLI tasks (Devlin et al. 2019). The BERT-base-uncased model we used consists of 12 transformer blocks, 12 attention heads, 110 million hyperparameters, and a dropout rate of 0.1. We chose this model since bidirectional representation learning captures complex relationships within clinical texts.

2.2 BioBERT

BioBERT is an adaptation of BERT which was introduced as a domain-specific language representation model for the biomedical domain. It is a BERT model trained specifically on biomedical data from PubMed abstracts and PMC full-text articles. Recent studies have found that fine-tuning universally pre-trained models like BERT on specific domain corpora tends to produce higher performance (Hababi et al. 2017). BioBERT significantly outperforms BERT in various biomedical text mining tasks: biomedical named entity recognition, biomedical relation extraction (0.62% and 2.80% F1 score improvements respectively), and biomedical question answering (12.24% MRR improvement) (Lee et al. 2019). We consider BioBERT in our work because it has been fine-tuned for clinical texts and previously used as a baseline alongside BERT, where it performs better on the i2b2 2006 metric for Exact F1 score (see Table 1).

2.3 Discharge Summary BERT

Discharge Summary BERT is a variant of BERT fine-tuned on discharge summaries of patient notes from MIMIC. It is at least as performant as BERT-base in the Med Natural Language Inference (MedNLI) task for accuracy (Romanov and Shivade, 2018) and F1 scores in all but one of four i2b2 named entity recognition (NER) tasks (Table 1). However, the BioBERT model yields consistently better results in comparison to Discharge Summary BERT (Alsentzer et al. 2019).

2.4 Discharge Summary BioBERT

Discharge Summary BioBERT uses BioBERT as a baseline to show the clinical performance of a model pre-trained on the biomedical domain. Similar to Discharge Summary BERT, Discharge Summary BioBERT is fine-tuned on the discharge summaries of patient notes from MIMIC. The model yielded the highest accuracy scores when applied to MedNLI, and comparable scores to BioBERT in all i2b2 tasks except for i2b2 2014 7A deidentification challenge (Alsentzer et al. 2019; Stubbs and Uzuner, 2015; Stubbs et al. 2015).

Discharge Summary BERT and Discharge Summary BioBERT truncate all discharge summaries to a maximum token length of 150 tokens; however, we truncated to 512 tokens since it is BERT's maximum token length for fine-tuning, and most discharge summaries are longer than 512 tokens.

2.5 MIMIC-BERT

MIMIC-BERT is a BERT model fine-tuned and trained on all MIMIC note events rather than solely discharge summaries. It was implemented to classify only the top 10 and top 50 ICD-9 codes to be consistent with prior work. The model achieved F1 scores of 0.858 and 0.922 for the top 10 and top 50 respectively. Due to time constraints, instead of unfreezing and retraining all of the model parameters like MIMIC-BERT, our models are fine-tuned with

Model	Section	Micro F1
Bert-Base	First 512	.46
BioBERT	First 512	.46
Bert-Base	HospCourse	.51
Bert-Base	HistoryPresIllness	.42
BioBERT	HospCourse	**.53**
BioBERT	HistoryPresIllness	.44

Table 2: Comparison of F1 micro scores between baseline models and experimental models

an additional classifier layer on top of the pre-trained BERT models (Singh et al. 2020).

2.6 Evaluating ICD-9 Assignment with MIMIC-III

In Huang et. al, 2019, several deep learning methods are explored for predicting ICD-9 codes using MIMIC discharge summaries. Huang's work describes a process of extracting records of hospital visits (by HADM or Hospital Admission IDs) that correspond to the top 10 and top 50 most frequent ICD-9 codes, since they cover a majority of the patient admissions data and many other ICD-9 codes are highly underrepresented. Huang's highest performing model was a CNN that achieved an F1 score of 0.696 for ICD-9 code prediction. Huang also utilizes only discharge summaries to train their classifiers, because discharge summaries contain unstructured text and are written after a patient's diagnosis has been reached. We follow the method of using discharge summaries and the top 10 codes as suggested by both MIMIC-BERT and Huang.

3 Methodology

3.1 Dataset

The Massachusetts Institute of Technology's Medical Information Mart for Intensive Care (MIMIC) is an open database that contains approximately 2 million de-identified patient notes and 6,918 diagnostic codes based on a diverse set of Intensive Care Unit patient data. Our work uses MIMIC-III v1.4, which is associated with over 61,532 ICU admissions to the Beth Israel Deaconess Medical Center between 2001 to 2012. MIMIC-III contains structured and unstructured data, such as prescribed medications, admissions statistics, procedures and tests administered, and patient notes in the form of free text.

We extracted data from the NOTEEVENTS and DIAGNOSES_ICD tables, and only worked with patient notes categorized as discharge summaries. We narrowed the scope of our classification task to the top 10 most frequently occuring ICD-9 codes, reducing our dataset size to 30,201 discharge summaries. We split them into 80-10-10 training, validation, testing sets.

3.1.1 Data Preprocessing

We found that some hospital admissions are assigned multiple discharge summaries, and extracted only the most recent summary associated with each such admission so that we would not be using multiple summaries referring to the same event.

We further cleaned the data to remove numbers, punctuation, and symbols (Huang et al. 2019). We also truncated all summaries to contain a maximum sequence length of 512 tokens since that is the upper limit for BERT. Only 11.7% of the discharge summaries meet this requirement, meaning that the majority of the patient notes cannot be fully processed during note-to-code mapping. Using the discharge summaries in a transformer model serves as a challenge due to the high variance in length distribution among summaries, capping the information extracted from our data into the model.

Rather than send in the first 512 tokens from each discharge summary, as implemented in our baseline models, we select from discrete subsections within the summaries. We extract the "Brief Hospital Course" section, which is a summary of the patient's symptoms and hospital stay, as well as the "History of Present Illness" section, which summarizes all the current symptoms a patient displays prior to and at the start of their admission to the hospital.

4 Results

4.1 Baseline

Our baselines are BERT-base and BioBert trained on the first 512 tokens of each discharge summary, resulting in F1 scores of 0.46 for both models. We tuned our hyperparameters based on recommendations for the pre-trained BERT model (Devlin et al. 2019). Due to memory constraints, we could not run our models with a batch size of 16 or 32, so we used a batch size of 12 instead. After tuning, we used four epochs and a learning rate of 1e-5 for the Adam optimizer to compensate for our smaller batch size. We used the same hyperparameters for each model in order to compare model performance

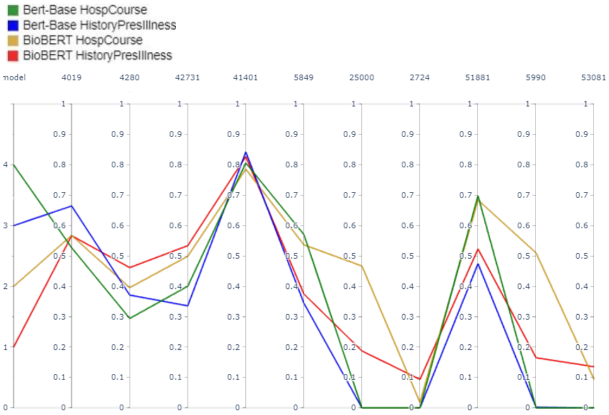

Figure 1: Comparison of ICD-9 code prediction accuracy for top 10 codes

and the effects of utilizing specialty embeddings and different note subsections.

4.2 Subsection Experimentation

We next trained the BioBERT and BERT-base models on two different subsections of the discharge summaries: History of Present Illness (HPI) and Hospital Course (HC). Both models were fine-tuned for multilabel classification using the first 512 tokens from these two subsections. In comparison to our baseline models, BioBERT fine-tuned on HC yielded 7% higher F1-scores, whereas BioBERT fine-tuned on HPI had a 2% decrease in F1-score. For the BERT-base models, we observed a similar trend, where the model fine-tuned on HC achieved a higher F1-score than the baseline model, and BERT-base fine-tuned on HPI achieved a lower score than the baseline. In comparing HPI and HC, we observed that the difference in ICD-9 code prediction performance between BioBERT and BERT-base varies by approximately 2%, with BioBert yielding higher F1-scores (Table 2).

Compared to our other models, BioBERT fine-tuned on HC achieved the highest F1 micro score of 0.53 when run on a separate test dataset. We also observed a significant difference in performance between models trained on HC versus HPI (Table 2). Between BioBERT models, BioBERT for HC had an 9% higher F1-score than that of BioBERT for HPI. Similarly, BERT-base for HC had a 9% higher F1-score than that of BERT-base for HPI.

5 Discussion

In the discharge summaries, HPI is one of the first sections in order of appearance, and HC occurs last. The models fine-tuned on HPI performed worse than the baselines fine-tuned on the first 512 tokens. Since the first 512 tokens often include HPI, this may suggest that the sections prior to HPI contain more relevant information to the classification task. Moreover, since the models fine-tuned on the HC section did significantly better than the baselines and the models using HPI sections, it can be inferred that information contained within HC, which likely never appears in the first 512 tokens due to summaries typically being much longer than 512 tokens, contain more relevant information for the classification task.

The models that were fine-tuned on the same subsections have similar F1 scores; we visualized the percentage of accurate predictions per ICD-9 code for each model to see if we could discover any significant differences or interesting trends on the individual label classification level. We found that the models had relatively similar prediction accuracy trends per ICD-9 code (Figure 1). On the other hand, we believe that the difference in performance for BioBERT versus BERT-base models is significant enough that it can be attributed to the use of domain specific versus general embeddings. As shown in Figure 1, the BERT-base models obtain 0% accuracy for ICD-9 codes 25000, 2724, 5990, and 5308, while none of the ICD-9 codes are predicted with a 0% accuracy by the BioBERT models.

Moreover, for most ICD-9 codes, BioBERT tends to predict the top 10 ICD-9 codes with a consistently higher accuracy than that of the BERT-base models, although BERT-base for HC occasionally performs better than BioBERT for HPI. This indicates that the minor increase in F1-score for BioBert over BERT-base could be a result of using contextualized embeddings pre-trained on biomedical documents rather than general corpora. Furthermore, the difference in performance for models fine-tuned using the HPI section of discharge summaries versus the HC section is significant enough to conclude that using the HC section when training models for ICD-9 code prediction can yield better results. Since there is a significant increase in a model's overall performance when fine-tuned on HC as opposed to the baseline and HPI, our work demonstrates that using domain specific embeddings fine-tuned on selectively extracted subsections of MIMIC discharge summaries can yield better performance for ICD-9 code prediction.

6 Source Code

The Github repository is available at: https://github.com/elliehaber/icd_code_pred

7 Contribution Statement

Ellie Haber preprocessed and performed data manipulation to prepare the MIMIC data for models. Caroline Barker, Ellie Haber, and Andrew Liang pair-programmed to create the baseline models and prepared the pre-trained embeddings for the other models. Caroline Barker and Andrew Liang researched and compiled prior literature sources. Ellie Haber wrote the Methodology section of the paper. Andrew Liang wrote the Abstract and Introduction section of the paper. Caroline Barker wrote the Prior Literature section of the paper. Caroline Barker, Ellie Haber, and Andrew Liang peer edited each other and wrote the Results and Discussion sections of the paper.

References

A. K. Bhavani Singh, Mounika Guntu, Ananth Reddy Bhimireddy, Judy W. Gichoya, Saptarshi Purkayastha. 2020. Multi-label natural language processing to identify diagnosis and procedure codes from MIMIC-III inpatient notes. CoRR abs/2003.07507

Alexey Romanov, Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. CoRR, abs/1808.06752

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. Scientific data, 3:160035, 2016.

Amber Stubbs, Christopher Kotfila, zlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. Journal of Biomedical Informatics, 58 Suppl:S11–19.

Amber Stubbs and zlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. Journal of Biomedical Informatics, 58 Suppl:S20–29.

Emily Alsentzer, John Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. CoRR, abs/1904.03323v3

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. CoRR, abs/1810.04805

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, Volume 36, Issue 4, 15 February 2020, Pages 1234–1240

Jinmiao Huang, Cesar Osorio, Luke Wicent Sy. 2019. An Empirical Evaluation of Deep Learning for ICD-9 Code Assignment using MIMIC-III Clinical Notes. CoRR, abs/1802.02311v2

Lance Ramshaw and Mitch Marcus. 1995. Text Chunking using Transformation-Based Learning. CoRR, abs/cmp-lg/950504v1

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, Ulf Leser, Deep learning with word embeddings improves biomedical named entity recognition, Bioinformatics, Volume 33, Issue 14, 15 July 2017, Pages i37–i48