EHRKit (Electronic Health Record Toolkit) and Application: ICD10 Classification

Ben Rosand, Irene Li, Jeremy Goldwasser, Matthew Zhang, Rohan Khera, Dragomir Radev

LILY Lab

Introduction

Investigations using medical texts in the EHR represent a key application of natural language processing (NLP). A common application of NLP on clinical notes is ICD10 diagnosis classification. This is the process of assigning patient diagnoses documented in notes to specific codified diagnoses. For example, during a meeting with a patient, a doctor may note that their patient experienced a loss of consciousness after being struck in the head by a falling object. The loss of consciousness/concussion would be encoded as S06.0x1A, while the blow to the head would be encoded as W20.8xxA. The codified diagnoses are used for billing, medical research and epidemiological monitoring. Further details are reported at https://www.aapc.com/icd-10/icd-10-documentation-example.aspx. This process of encoding is laborious, especially for more complex cases. It is also prone to error, as it is easy for a physician to miss a part of the coding, thereby reducing the quality of coded record and the various applications of these data. This project focuses on generating these icd10 codes from clinical notes. To this end, we build a pipeline which allows automated processing of notes and coding, in addition to model training and testing.

EHRKit

There is currently a lack of tools available for clinical health providers to utilize NLP methods easily. Tools such as huggingface and Allenai assist in parts of development, but even these tools have high learning curves. With EHRKit, we demonstrate a number of key uses of NLP methods in the EHR space. This toolkit contains a number of tools in the classification, summarization, and extraction domains. These tools are all developed for clinical health notes, and trained and tested on them. This toolkit is designed for use by doctors, and will be sent out to members of the LiLY lab for feedback today.

Dataset -- MIMIC

The electronic health record (EHR) now holds nearly all patient health information in the US and is key to both delivery of care and clinical investigation. A key component of the EHR is unstructured clinical data, including documentation by physicians at patient intake, during the course of their care, and at hospital discharge or the end of patient visits. Most of the work in unstructured medical text has been focused on MIMIC (Johnson et al. 2016), a public repository of 40,000 case records from patients admitted to the ICU at Beth Israel Deaconess Medical Center in Boston. In addition to this dataset, we built some models on proprietary Yale New Haven Hospital Data.

Dataset Split

For the icd classification task, we performed an 70/15/15 split, training, testing, and validation. The mimic corpus contained 52,509 notes, and the Yale corpus contained 7,781 notes.

CSU

Pet_Name is a 10 year old male castrated hound mix that was presented for continuation of chemotherapy
B-cell multicentric lymphoma. Pet_Name was started on CHOP chemotherapy last week and has been do
receiving doxorubicin. The owners have noted his lymph nodes have gotten much smaller. He has some loos
metronidazole. Current medications include prednisolone. Assessment: Pet_Name is in a strong partial re
physical exam. He is also doing very well since starting chemotherapy. A CBC today was unremarkable and
chemotherapy. She was dispensed oral cyclophosphamide and furosemide that the owners were instructed t

Expert annotated diseases: Malignant tumor (disorder), Disorder of haematopoietic cell proliferation, Lymph
(disorder), Neoplasm and/or hamartoma, Lymphoreticular tumor (disorder), Neoplasm, Malignant tumor of ly
related tissue (disorder), Neoplasm of haematopoietic cell type

PP

$Person_Name$ cc : recheck hypercalcemia responding to pred - iCa has dropped - is on 10 milligram twice
twice a day lasix eating human food on own feeling better blood urea nitrogen down from 89 -- > 62 today Vi
has residual brisket edema from Fluids on Wed patient : Treating symptomatically discontinue fluids since dr
pred to 10 milligram once daily and lasix 25 milligram once daily recheck on Tues pred appears to be treating
strong suspicion for LSA E. Ellis VMD wt : 62.5 lbs . temperature : 101.3f

Expert annotated diseases: Metabolic derangement, Disorder of calcium metabolism (disorder), Disorder o
vitamin D metabolism (disorder), Disorder of mineral metabolism (disorder)

PSVG

Pet_Name initially presented on 9/30 with lethargy and fever. She was diagnosed with bilateral renomega
revealed possible cyst in a small left kidney and probable ureteral obstruction with secondary hydronephros
Pet_Name was transfered to our care on the evening of 10/2.

Figure 1: Example Clinical Notes (Zhang 2019)

Models

BERT-Based Models:
For the classification task, we used several BERT-based models, specifically trained on biomedical data. Biomed-roberta (Gururangan et al. 2020), and clinical-roberta-long (Abraham et al.2020), finally for the multilabel classification we used TF-IDF embeddings underneath a single layer neural network, and Longformer (Beltagy et al. 2020).

Multiclass vs Multilabel:
Most of the work that has been performed in classification of the mimic dataset has treated it as a single label problem. They take the most common labels, and assign the first visible label to each note. In reality, each note may have more than one, sometimes up to 10 different codes. We present two models, one of which is trained and tested for single code prediction, the other of which is trained for multiple labels. Unfortunately, there were a lot of difficulties in setting up these models to work across the different data streams. Given that the focus of this work was to prepare usable code in EHRKit, the full testing of these models was put on hold. As such, there is only limited data available.

Results

Table 1. Multiclass Classifier Results

Model	Acc	F1	Precision	Recall	Data source
biomed-roberta	0.219	0.202	0.212	0.219	Yale
biomed-roberta	0.688	0.686	0.700	0.687	MIMIC
clinical-roberta-long	0.721	0.722	0.726	0.721	MIMIC

Table 2. Multilabel Classifier Results

Model	Acc	F1	Precision	Recall	Data source
tf-idf + NN	N/A	0.51	0.51	0.49	MIMIC
tf-idf + NN	N/A	0.34	0.39	0.31	Yale
longformer	0.962	0.962	0.595	0.30	Yale

Term-Frequency Inverse-Document-Frequency (TF-IDF) is a common baseline embedding method used in NLP. In TF-IDF, each document is transformed into an N-dimensional vector, in which each word corresponds to one element. N is the total number of unique words across all documents being examined. The value of the element in said vector is determined by the TF-IDF weighting

'document', 0.216
'each', 0.216
'frequency', 0.216
'idf', 0.323
'in', 0.431
'is', 0.431
'of', 0.216
'tf', 0.323
'the', 0.431
'vector' 0.216

Figure 3: TF-IDF example (Rosand 2021)

Conclusion

Currently we have tested a number of models with varying success, but even the highest performing models are still pretty limited. We also see pretty substantial drop-offs going from MIMIC models to Yale models. Unfortunately, these different datasets are encoded differently (ICD9 vs ICD10), so we cannot directly test the mimic trained model on a test set from Yale. Moreover, we see that the multiclass results are significantly better than the multilabel results. This reinforces the fact that many recent approaches to the icd classification problem have been insufficient. Nevertheless, the goal of this project is to provide a basis for future EHR NLP research, and in the github release we provide a large variety of options for people to explore.