Generating contextual text embeddings for emergency department chief complaints using BERT

David Chang
Yale University, david.chang@yale.edu

Generating Contextual Text Embeddings for Emergency Department Chief Complaints using BERT

David Chang[1], Woo Suk Hong[2], Andrew Taylor, MD[2]

[1]Computational Biology and Bioinformatics Program, Yale University, New Haven, CT; [2]Department of Emergency Medicine, Yale University

Abstract

Emergency department care is guided by chief complaints, but the secondary use of chief complaint data in daily operational decision and research is hampered by its form and representation. Recent advances in natural language processing (NLP) provide an opportunity to address many of the challenges of chief complaint data. We use Bidirectional Encoder Representations from Transformers (BERT), a novel NLP method, to learn contextual embeddings for chief complaints. We show that these embeddings can be used to accurately predict provider-assigned chief complaint labels and that semantically similar chief complaints are mapped to nearby points in the embedding space.

Keywords
Natural language processing; emergency department; chief complaints; text classification

Introduction

What is chief complaint?

The chief complaint (also referred to as reason for encounter) is a concise statement describing the main reason for a patient's visit. It's the second step of medical history taking after identification and demographics information.

Chief complaints are usually entered into the electronic health record (EHR) as free-text, and providers can then assign a label/category (sometimes called presenting problems) based on the chief complaints. The presenting problem labels for chief complaints are not always available, and the process of manually assigning labels to chief complaints can be time-consuming.

Examples of typical chief complaints and presenting problems

Chief complaint #1: "feeling pressure on chest and left arm pain…"
Presenting problem #1: CHEST PAIN

Chief complaint #2: "diarrhea, vomiting, stomach pain…"
Presenting problem #2: ABDOMINAL PAIN

What are we trying to do?

We want to develop a model to automatically map chief complaint text to structured presenting problem categories with high accuracy.

Why does it matter?

There's a lot of useful information contained in these chief complaints, but without more structured labels it's often difficult to utilize all that information for secondary research, for example in operational decisions or quality improvement initiatives. Having a state-of-the-art information extraction model that can automatically labels these notes to more usable forms can facilitate downstream use cases.

Methodology and Results

The dataset from the Yale Emergency Department has approximately 1.9 million entries of chief complaint notes over a period of a few years. While the original number of categories is 1103, we end up with 434 after excluding sparse labels that comprise less than 0.01% of the total data. In order to map these free-text notes into their respective chief complaint categories, we used a clinical version of BERT for multi-class text classification. Figure 1 visualizes some of the inner workings of BERT's self-attention mechanism for encoding text into vector embeddings.

Figure 2 shows the results in terms of top-5 accuracies across different thresholds for excluding sparse labels (see legend). We can see that even at top-4 predictions, we were able to pass 90%, and the performance continues to increase as k increases. This demonstrates the capacity of BERT for classifying pieces of text, especially as we narrow down the space of available labels. One thing to note is that the rapid improvement in performance from top-1 to top-4 predictions is indicative of the fact that many of the labels are semantically similar (sometimes exact synonyms).

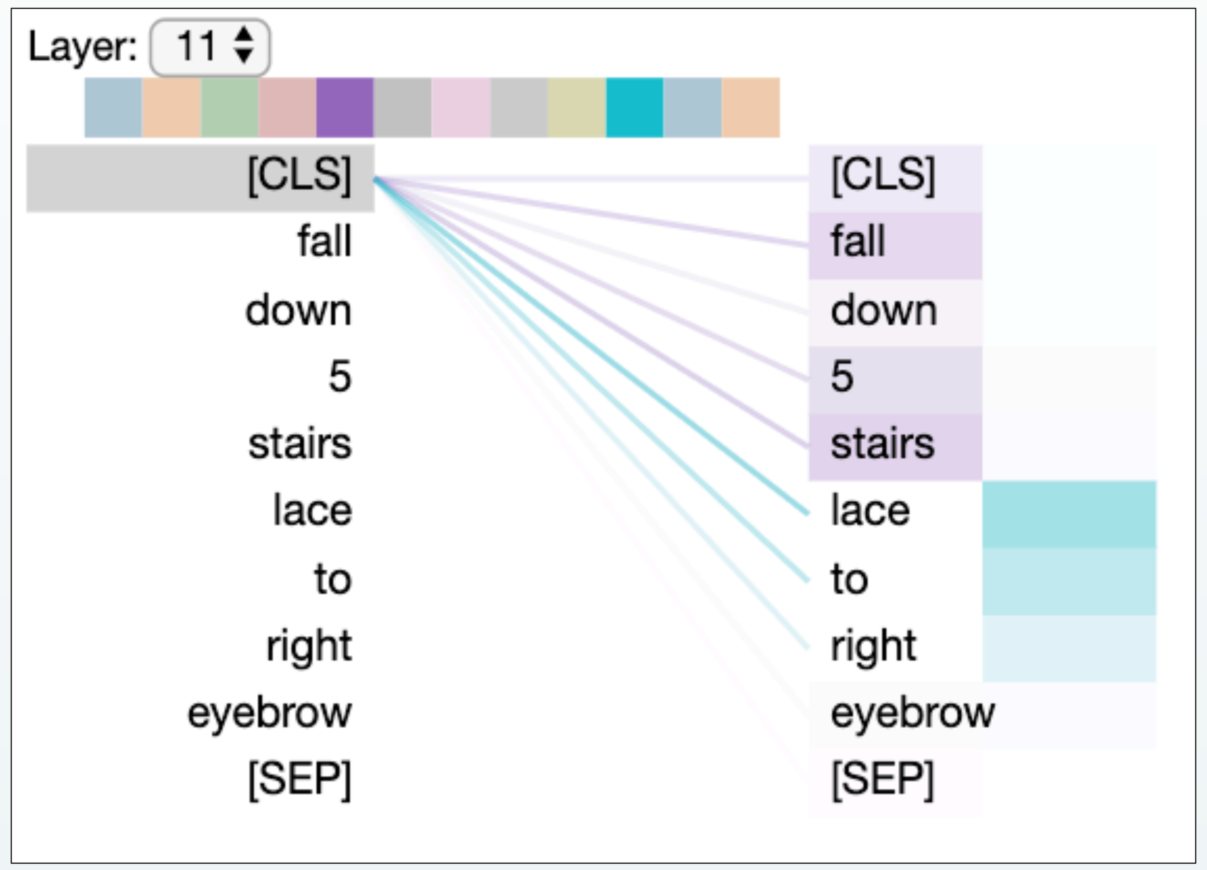

Figure 1. Bertviz visualization of BERT's self-attention mechanism.

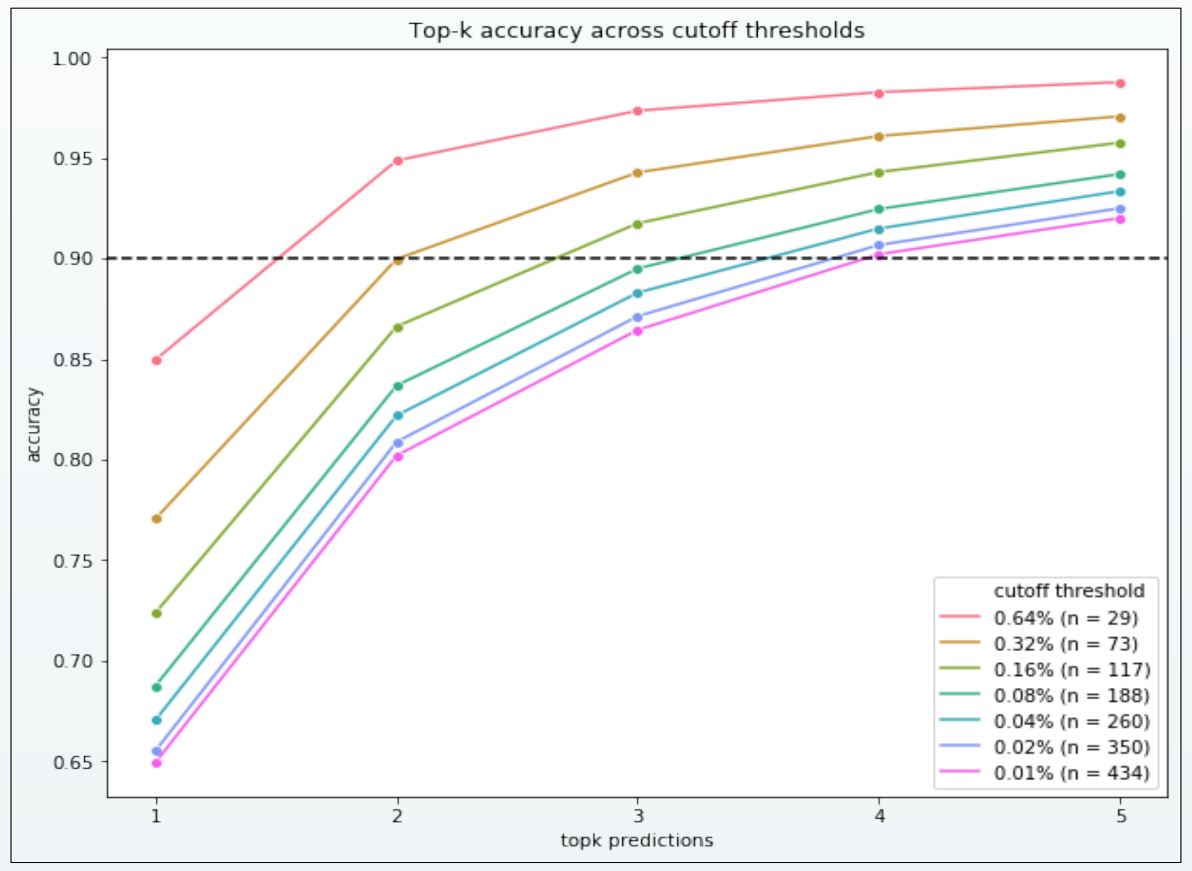

Figure 2. Top-k accuracy over different cutoff thresholds for categories.

Discussion and Conclusion

We successfully trained a BERT-based NLP model that can accurately map chief complaint free-text to structured categories while learning a clinically meaningful representation of domain-specific text in emergency medicine. Figure 3 shows a t-SNE plot of the labels as calculated as the average of the embedding. Vectors of their corresponding chief complaints. It is clear that the labels are clustered in a clinically meaningful way. Future work will consider methods to better leverage such prior knowledge about the semantic relationships between labels.

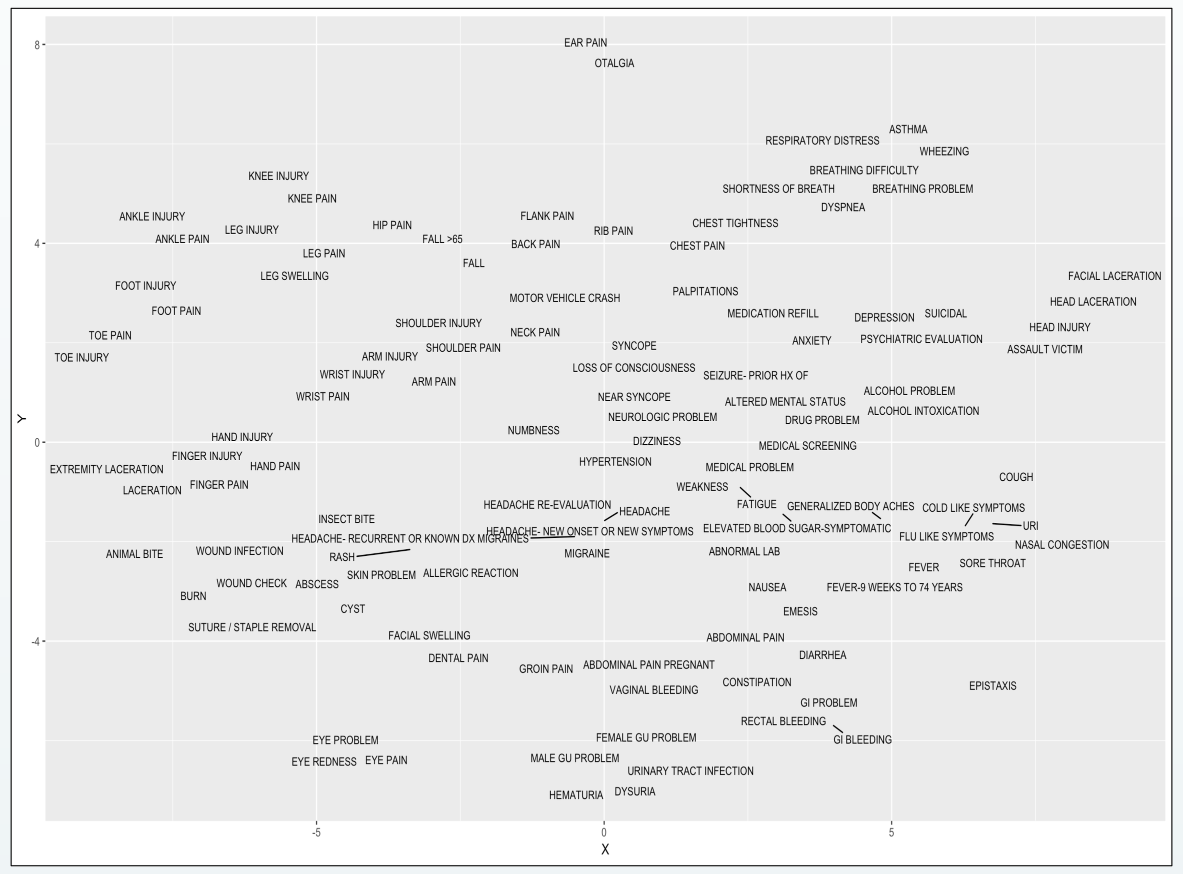

Figure 3. t-SNE plot of labels as averages of their embeddings

Error Analysis

Chief complaint text	Top k Predictions from BERT
Correctly classified at second prediction	
"right third finger injured in door"	FINGER INJURY, **HAND PAIN**
"pt comes to er with cc peice of plastic stuck to back of left ear from earing"	FOREIGN BODY IN EAR, **EAR PROBLEM**
"vomiting for days, increasing yesterday. pos home preg test on Saturday"	EMESIS, **EMESIS DURING PREGNANCY**
"both eyes swollen & itchy & tearing after his nap"	EYE SWELLING, **EYE PROBLEM**
"fall at 0300 today, rt side weakness"	FALL, **FALL>65**
Correctly classified at fifth prediction	
"Felt like heart was pounding history of cabg. missed metoprolol for about 3 days."	PALPITATIONS, RAPID HEART RATE, TACHYCARDIA, IRREGULAR HEART BEAT, **CHEST PAIN**
"2 weeks of sore throat, aches, dry cough. Denies intervention."	SORE THROAT, COLD LIKE SYMPTOMS, URI, COUGH, **FLU LIKE SYMPTOMS**
"fall down 5 stairs lace to right eyebrow"	FALL, FACIAL LACERATION, LACERATION, FALL>65, **HEAD LACERATION**
"fever to 101, diarrhea, vomiting"	FEVER-9 WEEKS TO 74 YEARS, FEVER, EMESIS, ABDOMINAL PAIN, **FEVER-8 WEEKS OR LESS**
"blister on back of foot."	BLISTER, FOOT PAIN, FOOT INJURY, FOOT SWELLING, **SKIN PROBLEM**

To the left is a table displaying examples of typical errors from the BERT model. The first 5 rows are examples that were incorrect at the first prediction but correct at the second prediction, and the last 5 rows are examples that were correct at the fifth prediction.

We can see that in a lot of cases, the predictions made by the model that are technically incorrect actually work just as well if not better than the provider-assigned ground truth. This speaks to our model's ability to really learn a meaningful representation of the text and capture clinically meaningful relationships, as demonstrated by the performance and the label plot in Figure 3.

This works nicely in a practical setting where we don't necessarily need to get the label correctly at the first try and can rely on a drop-down menu with the top-k predictions.

Lastly, the problems evident in the label space provide a strong argument for developing more data-driven and concise ontology or chief complaint categories for emergency care.

Acknowledgements

This project was made possible by NIH Training Grant 5T15LM007056-33