Fast and Accurate Clinical Named Entity Recognition and Concept Mapping from Conversations

Hoo-Chang Shin, PhD, NVIDIA; Christopher Parisien, PhD; Raghav Mani; Jonathan Cohen

Introduction

The recent COVID-19 pandemic has led to a drastic increase in healthcare provider and call center volumes. Not every organization was prepared for this uptake in volume. Therefore, an automatic speech recognition model that can extract and relate key clinical concepts from clinical conversations can be very useful. We introduce a model that 1) captures a dialogue or speech, 2) identifies clinical words, and 3) maps the identified words in a standardized ontology database. These capabilities are both fast and accurate.

Hypothesis

With the advancement of language model pre-training using architectures such as BERT, the field of natural language processing has experienced a major leap in progress. Most of the successes are in general-domain applications, such as question answering and text classification. We first hypothesize that a language model that was pre-trained on a large biomedical text corpus will outperform a general language model. Secondly, an even larger model with biomedical vocabulary will perform better than a smaller model using a generic vocabulary set. Lastly, by achieving fast inference with named entity recognition and clinical concept mapping in a standardized ontology, we can support patients and providers with improved responses to patient requests.

Methods

Bidirectional Encoder Representations from Transformers (BERT) is a technique for natural language processing (NLP) pre-training. A large language model is pre-trained on a large text corpus in a self-supervised manner. Then, typically using a much smaller labeled dataset for supervised training, such a model is fine-tuned on a specific task to achieve greatly improved performance compared to earlier techniques.

For clinical named entity recognition, we fine-tune BERT on the dataset from the 2010 i2b2/VA challenge, a clinical NLP dataset that was created at a former NIH-funded National Center for Biomedical Computing (NCBC).

We first demonstrate using an existing BERT-based architecture with 110 million parameters, pre-trained on the PubMed biomedical abstract text corpus comprising of 4.5 billion words. This model, BioBERT, performs better than a general-domain BERT model pre-trained on Wikipedia and BooksCorpus. We then demonstrate using a larger BERT model, named Bio-Megatron, with 345 million parameters. By training this model on the PubMed biomedical text corpus with about 6.1 billion words, we show that using both a larger model and a biomedical vocabulary adds a further boost in performance.

We incorporate this model into an Automatic Speech Recognition system so that key clinical words can be identified in individual speech or in a conversation. Identified clinical words are then mapped into concepts in a standardized medical ontology, the Unified Medical Language System (UMLS).

Finally, NVIDIA Jarvis is used for fast inference on these large Deep Learning models..

Results

Precision, recall, and F1 scores for clinical named entity recognition (NER) are shown below, on the reserved test set from the 2010 i2b2/VA challenge. BERT, BioBERT, and Bio-Megatron refer to a general-domain BERT model, a *BERT-base* model pre-trained on PubMed abstracts, and *BERT-large* pre-trained on PubMed abstracts and full-text, respectively.

	precision	recall	F1-score

BERT-base	93.11	88.07	90.52
BioBERT	94.21	89.77	91.93
Bio-Megatron	94.79	89.46	92.05

For better clarity of the system’s output, and for better clinical guidance and downstream analysis, we map identified clinical named entities to concepts in UMLS, a standardized medical ontology. For example, the system maps the phrase "dysplasia of the left hip" into the following concepts:

N-gram	Term	CUI	Similarity	Semtypes
left hip	Left hip	C0524471	1.0	T029 - Body Location or Region
dysplasia	dysplasia	C0334044	1.0	T046 - Pathologic Function
dysplasia	Omodysplasia	C4510897	0.7	T047 - Disease or Syndrome

A further concept normalization model could be developed and applied to address any ambiguity from here. While the concept mapping in this work uses N-grams, another avenue for future research would be to use Bio-Megatron generated embeddings to evaluate semantic similarity during concept mapping.
Running deep learning models as large as BERT can be time-consuming. For example, when running NER inference on CPU, a single sentence can take up to 3 minutes. With an NVIDIA Jarvis service running on a T4 or a V100 GPU, that inference time is reduced to under 1 second. This opens significant new capabilities in systems where responsiveness to patients, clinicians, and researchers is paramount.

Conclusion

We demonstrate a fast and accurate automatic speech recognition system that can capture key clinical named entities and map them to concepts in a standardized ontology.

Statement of Impact

In a time where healthcare providers and call centers are experiencing an unprecedented increase in patient call volume, we hope our contribution will help achieve faster & better patient responses, ultimately leading to improved patient care.

Keywords

deep learning, natural language processing, automatic speech recognition