Project Proposal and Literature Review

Arav Parikh and Kaitlyn Bedard
Department of Computer Science and Engineering
University of Connecticut

***Abstract*—Clinical decision making (i.e. the formulation of diagnoses) is a critical part of what doctors do on a day-to-day basis; however, it is a rather demanding task given that it requires them to have an in-depth understanding of disease descriptions and identify hidden patterns among the symptoms. While there is no doubt that these doctors possess this knowledge, the time it takes to recall such information detracts from their ability to spend time and directly work with their patients to achieve better clinical outcomes. With the rise of AI/ML models, though, the automation of the task of diagnosis prediction is not only readily achievable, but perhaps even preferable in the sense that it might allow patients to receive diagnoses without directly consulting a doctor, thereby reducing undue strain on the healthcare system and saving doctor visits for treatments and emergencies. In the following literature review, we explore past datasets and models used for this and other similar use cases with the hope of identifying areas of augmentation and improvement for our own project.**

Literature Review

Starting with past models based on the BERT architecture, [1] proposed the BioBERT model as the first domain-specific model in the biomedical sphere. Pre-trained on a large corpus of biomedical data based on PubMed and PMC articles and fine-tuned with several named entity recognition, relation extraction, and question answering data sets, BioBERT outperformed BERT and other state-of-the-art models on the aforementioned NLP tasks in a biomedical context. Given this underlying success with biomedical texts, it makes sense that we might utilize this model and further fine-tune it for the purpose of diagnosis prediction. On that note, though, [2] actually already proposed the Med-BERT model specifically for the task of disease prediction, but the data used to accomplish this task came in the form of electronic health records (EHRs) which are quite different than the typical textual representations passed into transformer models. Although still structured and sequential, EHRs heavily rely on universal clinical code systems used by medical professionals, meaning that both the model inputs and outputs also follow a similar schema. Unfortunately, while undoubtedly very useful in a clinical setting for those who can understand such a schema, it is difficult to see this model being used outside of that setting when more plain-text language (i.e. more easily understood English biomedical terminology) is used, which is exactly what we would like to explore in our own experimentation.

Based on the largely on the BioBERT model, [3] proposed the CORe (Clinical Outcome Representations) model. An important feature of this model is a pre-training step which teaches the model relationships between symptoms, risk factors, and clinical outcomes, in a way that emulates how doctors expand knowledge via experiences and literature. [3] The primary data used in this implementation was de-identified EHR data, in particular, discharge summaries and outcome information associated with a given admission. It was trained using articles as well, including reports from PMC, Wikipedia, and MedQuAd dataset, so unlike the Med-BERT model, the data is more easily understood to those unfamiliar with the clinical setting. The CORe model deals with the following tasks: diagnosis prediction, procedure prediction, in-hospital mortality prediction, and length of stay prediction.

The results of the CORe model are promising. Data in [3] confirms that the model improves the scores on all the chosen tasks compared to baseline models – including BioBERT Base and other branches of BioBERT models. There are also indications that the model has learned to predict diagnoses that were never directly mentioned in text. However, despite the promising results, [3] mentions some pitfalls, namely, incomplete/inconsistent labels, the existence of multiple possible outcomes, negation, and numerical data. Issues caused by negation and numerical values can be potentially be combated via semantic encoding, as per recommendation of the authors. Other issues can be potentially be handled by including more data in the training.

Project Statement

As indicated throughout this literature review, we propose a diagnosis prediction model. We will experiment with the CORe Model - Clinical Diagnosis Prediction from the Hugging Face library and attempt to replicate its functionality while using plain-text inputs and outputs rather than the model's current patient admission note input and ”multi-labeled ICD-9 code prediction” output.

The CORe model: CORe Clinical Diagnosis Prediction

Some potential sources for data are:

- DDXPlus Dataset
- Disease Symptoms — Kaggle
- Disease Symptoms Prediction — Kaggle
- Disease Symptom Knowledge Dataset

References

[1] J. Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, Sep. 2019, doi: https://doi.org/10.1093/bioinformatics/btz682.

[2] L. Rasmy et al., ”Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction,” May 2020, doi: https://doi.org/10.48550/arXiv.2005.12833

[3] B. Aken et al., ”Clinical outcome prediction from admissions notes using self-supervised knowledge integration,” April 2021, doi: https://aclanthology.org/2021.eacl-main.75.pdf