CS6301: Project Proposal
NER on Medical Data using LLM

03/24/2023

1 Task

The task is to implement Named Entity Recognition (NER) using large-scale language models like BERT(Bidirectional Encoder Representations from Transformers) or ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) on medical literature.

The objective of Named Entity Recognition (NER) is to extract named entities from text, which include diseases, symptoms, treatments, and medications in the medical domain. Medical data comes in diverse forms such as electronic health records, clinical notes, and medical literature. Clinical notes, containing unstructured text with medical terminology, are especially advantageous for NER as they provide insights into the patient's medical history, symptoms, and treatment.

NER for medical data involves recognizing and extracting named entities related to medical terms from clinical notes or other medical text types. NLP techniques such as deep learning models, statistical models, and rule-based systems can be employed for NER with medical data. The aim is to accurately identify and extract pertinent named entities from the text to support downstream tasks such as clinical decision support and information retrieval.

2 Data

The i2b2 (Informatics for Integrating Biology and the Bedside) data set is a well-known data set in the field of natural language processing (NLP) and specifically for clinical NLP. It contains de-identified clinical notes and reports from various hospitals, including radiology reports, discharge summaries, and progress notes. The data set includes annotations for medical entities such as diseases, treatments, and medications.

The data set is in XML format and has been preprocessed to remove personally identifiable information. It contains over 1,000 clinical documents with over 3,000 annotated medical entities. The annotations are provided in the IOB format (Inside, Outside, Beginning), which is a common format used for entity recognition tasks.

We plan to use this data set as a baseline data set to train large-scale language models such as BERT and ELECTRA for named entity recognition on medical data. The data set will be split into training, validation, and test sets, and we will use a combination of supervised and unsupervised learning techniques to fine-tune the models on this data.

Using this data set as a starting point will allow us to develop and test our methodology for named entity recognition on medical data, with the goal of eventually applying it to other data sets and real-world applications in healthcare.

3 Methodology

The methodology we propose is to use Named Entity Recognition (NER) techniques to extract medical entities from the i2b2 data set, which contains clinical notes and reports from various hospitals. The aim is to identify entities such as disease names, medical procedures, medications, and their attributes (dosages, frequencies, etc.) mentioned in the notes.

To achieve this, we will use large-scale language models like BERT (Bidirectional Encoder Representations from Transformers) or ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) that have been pretrained on vast amounts of text data, including medical literature. These models can capture the context and nuances of the text, which is crucial in accurately identifying medical entities.

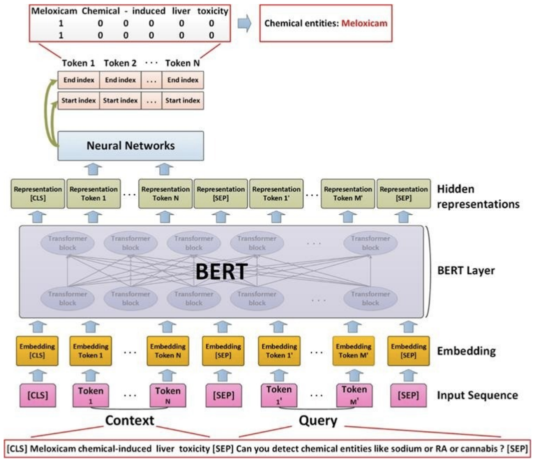
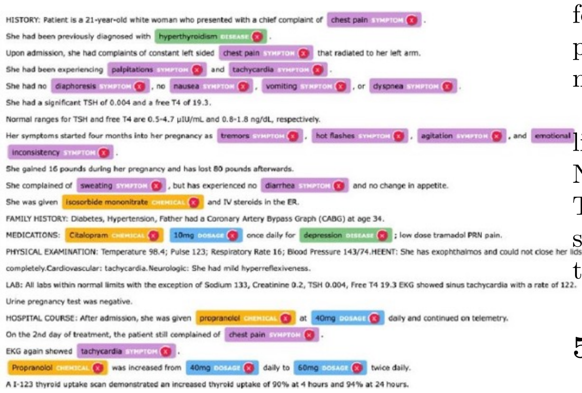


Figure 1: *Sample Block Diagram for NER using LLM*

The first step is to preprocess the data by tokenizing the text and converting it into a format that can be inputted into the language model. We will then fine-tune the model on the i2b2 data set by training it to recognize medical entities. The model will be trained using a combination of supervised and unsupervised learning techniques.

Next, we will evaluate the performance of the model on a test set of annotated data. We will measure the precision, recall, and F1-score of the model to assess its accuracy in identifying medical entities. We may also perform error analysis to identify the most common types of mistakes made by the model.



HISTORY: Patient is a 21-year-old white woman who presented with a chief complaint of chest pain SYMPTOM .
She had been previously diagnosed with hyperthyroidism DISEASE .
Upon admission, she had complaints of constant left sided chest pain SYMPTOM that radiated to her left arm.
She had been experiencing palpitations SYMPTOM and tachycardia SYMPTOM .
She had no diaphoresis SYMPTOM , no nausea SYMPTOM , vomiting SYMPTOM , or dyspnea SYMPTOM .
She had a significant TSH of 0.004 and a free T4 of 19.3.
Normal ranges for TSH and free T4 are 0.5-4.7 µIU/mL and 0.8-1.8 ng/dL, respectively.
Her symptoms started four months into her pregnancy as tremors SYMPTOM , hot flashes SYMPTOM , agitation SYMPTOM , and emotional inconsistency SYMPTOM .
She gained 16 pounds during her pregnancy and has lost 80 pounds afterwards.
She complained of sweating SYMPTOM , but has experienced no diarrhea SYMPTOM and no change in appetite.
She was given isosorbide mononitrate CHEMICAL and IV steroids in the ER.
FAMILY HISTORY: Diabetes, Hypertension, Father had a Coronary Artery Bypass Graph (CABG) at age 34.
MEDICATIONS: Citalopram CHEMICAL 10mg DOSAGE once daily for depression DISEASE ; low dose tramadol PRN pain.
PHYSICAL EXAMINATION: Temperature 98.4; Pulse 123; Respiratory Rate 16; Blood Pressure 143/74.HEENT: She has exophthalmos and could not close her lids completely.Cardiovascular: tachycardia.Neurologic: She had mild hyperreflexiveness.
LAB: All labs within normal limits with the exception of Sodium 133, Creatinine 0.2, TSH 0.004, Free T4 19.3 EKG showed sinus tachycardia with a rate of 122.
Urine pregnancy test was negative.
HOSPITAL COURSE: After admission, she was given propranolol CHEMICAL at 40mg DOSAGE daily and continued on telemetry.
On the 2nd day of treatment, the patient still complained of chest pain SYMPTOM .
EKG again showed tachycardia SYMPTOM .
Propranolol CHEMICAL was increased from 40mg DOSAGE daily to 60mg DOSAGE twice daily.
A I-123 thyroid uptake scan demonstrated an increased thyroid uptake of 90% at 4 hours and 94% at 24 hours.

Figure 2: *NER detecting the entities such as diseases in medical reports*

Finally, we will deploy the model on new, unseen data and use it to extract medical entities. This can be done in real-time, allowing healthcare professionals to quickly and accurately extract information from clinical notes and reports.

4 Experiments and Analysis

The main experiment and analysis we will conduct is named entity recognition on the i2b2 data set using large-scale language models such as BERT and ELECTRA. We will fine-tune these models on the i2b2 data set and evaluate their performance on the task of NER.

To validate our approach, we will compare the performance of our models with benchmark models such as BIOELECTRA, which is a state-of-the-art model for NER. We will measure the precision, recall, and F1-score of the models and compare them against the benchmark models.

In addition, we will test our trained models on other datasets such as MIMIC-III and ShARe/CLEF eHealth datasets, which contain clinical notes and reports from different sources. This will allow us to evaluate the performance of our models on unseen data that is different from our training data.

We will also perform error analysis on the test data to identify the most common types of mistakes made by the models. This will help us improve the models and identify areas where further training may be necessary.

Finally, we will create parser models to convert other data formats to the i2b2/n2c2 model format. This will facilitate seamless training of the models with the same remainder pipeline and allow us to train the models on other data sets easily.

Overall, the experiments and analysis we will conduct will help us verify our hypothesis that large-scale language models can be effectively used for named entity recognition on medical data and provide insights into the performance of these models on different datasets.

The study aims to develop a website that utilizes an optimized model for the extraction of Named Entities from relevant text documents. The website, once deployed, will allow users to seamlessly upload a text file and retrieve its extracted Named Entities.

5 References

- https://paperswithcode.com/paper/bioelectra-pretrained-biomedical-text-encoder

- http://www.lllf.uam.es/ESP/nlpdata/wp2/s12911-021-01395-z.pdf

- https://arxiv.org/ftp/arxiv/papers/1901/1901.08746.pdf

- https://www.sciencedirect.com/science/article/pii/S15320464210

- https://ceur-ws.org/Vol-2551/paper-04.pdf