Automated Structuring of Medical Records (Named Entity Recognition)

Team Members:

Lin Yutian 1004881

Wang Yanbao 1004865

Problem to be Investigated:

The difficulty in efficiently accessing and analyzing unstructured medical data arises from the vast amounts of free-text records. This project will leverage Named Entity Recognition (NER) to extract categorize key medical entities such as symptoms, diagnoses, medications, procedures, and patient demographics from unstructured medical records and consolidate it into a structured table format, enhancing the accessibility and utility of medical data. The proposed system aims to streamline the extraction process and improve the organization of information within medical records.

Expected Inputs and Outputs:

- Inputs: Unstructured medical records including clinical case report details.
- Outputs: Structured tables with categorized entities such as symptoms, diagnoses, medications, procedures, and patient demographics.

Dataset to be Used:

MACCROBAT2018 & MACCROBAT2020
(https://figshare.com/articles/dataset/MACCROBAT2018/9764942)

Two datasets contains 400 source documents (in plain text, on sentence per line) and 400 annotation documents (in brat standoff format) in total. Text is from PubMed Central full-text documents but has been edited to include only clinical case report details. All annotations were created manually.

Note: We may also consider two other datasets (Currently requesting approval from authors)

- https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
- https://tianchi.aliyun.com/dataset/95414

Architecture Draft:

Initially, we plan to implement an encoder-only transformer architecture, renowned for its effectiveness in capturing contextual information within sequences.

Concurrently, we aim to develop a model based on Long Short-Term Memory (LSTM) networks. And we plan to compare the performance of two models.

What You Are Going to Deliver:

1. A software application capable of transforming unstructured medical text into structured, categorized data.
2. A detailed report outlining the methodology, architecture, entity extraction performance comparison between LSTM and transformer architectures, and use cases.
3. Source code for the complete NER pipeline, from preprocessing to entity extraction.
4. Code for recreating the trained model from a file.