COHERE HEALTH TAKE HOME REPORT

CAMERON CINEL

1. Problem Statement

After every interaction with a patient, a clinician records a clinical note of the encounter. This clinical note can include infomration about the patient's history, current condition, and future treatments recommended by the clinician.

One of the more important information in a clinical note is the primary diagnosis of the patient. The primary diagnosis or diagnoses is the diagnosis or diagnoses that is the most resource and time intensive for the hospital or clinician to deal with. In addition, the clinical note can include information about other conditions that the patient had before arriving to the hospital that give evidence to the diagnosis.

The goal of this task is to create a database of primary diagnoses found in a collection of clinician notes, with each of them associated with various other diseases that are underlying factors of the primary diagnosis. That is, each diagnosis should be associated with collection of other diseases that often occur with or before the diagnosis. In addition, a bonus task was attempted which was to build a system that can identify a primary diagnosis from a clinical note along with the underlying factors that give evidence for the diagnosis.

2. Dataset

The dataset provided consists of approximately 300 clinical notes for various patient visits. Identifying information was redacted from each clinical note. The clinical notes included sections such as "Chief Complaint", "History of Present Illness", "Discharge Diagnosis", and "Discharge Disposition". Note that not every clinical note had every possible section. Each clinical note was save as a `.txt` file.

In addition to each clinical note, there was a corresponding `.ann` file. Each `.ann` file included named entities recognized in the corresponding clinical note. In addition, related entities were linked together. Each entity was given a unique identifier, a category (such as "Reason" or "Drug"), and the start and end index in the clinical note `.txt` file corresponding to the identity. Each relation was given the two entities it referred to as well as the relation between them.

3. Data Cleaning

The data was loaded into three pandas DataFrames:

1. `txt_df`: a DataFrame consisting of the file name and text from the `.txt` file
2. `ent_df`: a DataFrame consisting of the file name, entity identifier, category, start index, end index, and the entity name
3. `rel_df`: a DataFrame consisting of the file name, relation identifier, the relation category, and the two different entities in the relation

We cleaned the `txt_df` by first identifying the sections of each clinical note that corresponded to the "Discharge Diagnosis", "History of Present Illness", and "Chief Complaint". We chose these sections as they would be the most correlated with the patient's primary diagnoses as well as any underlying factors the patient had that would lead to these diagnoses. Not every clinical note had each of these sections. We discarded clinical notes without a diagnosis section, as they would not give us relevant information for our database. Without a diagnosis, we could not link any factors the patient had before coming to the hospital accurately to another disease. However, we kept clinical with a diagnosis that did not have either a history or complaint section, as information about the diagnosis could still be extracted from them. After discarding notes without diagnoses, we were left with about 280 notes.

In addition, some clinical notes had their diagnoses separated into "Primary" and "Secondary" categories. As primary categories are our main interest, we extracted the primary diagnosis section from each diagnosis. For diagnoses that did not have such separation, we took the entire diagnosis as the primary diagnosis.

As far as the `ent_df` and `rel_df`, we did not use them for our solution.

4. Solution Implementation

Once we had a diagnosis, primary diagnosis, history, and complaint category for each clinical note, we use a spaCy Named Entity Recognition (NER) model to extract entities from each category. Specifically, we used a model trained on the BC5CDR corpus provided by the `scispacy` package. We chose such a model as it is able to identify entities as belonging to two different categories: `DISEASE` and `DRUG`. As our interest is in diagnoses and underlying factors, we chose to discard entities labelled as `DRUG`.

In addition, we use an Entity Linker from the `scispacy` package to link each `DISEASE` entity found to a canonical name. Specifically, we chose to use a linker linked to the Unified Medical Language System (UMLS). This linker would use a K-Nearest Neighbor search of a given disease name inside the UMLS corpus, and return the best matches above a certain threshold. For the canonical name of a given disease, we chose to use the nearest neighborhood that passed the threshold. We used this entity linker for a variety of reasons:

1. Certain entities we the same, but written in different ways. For instance, the term ”seizure” and ”seizures” both occur in our dataset and are tagged as `DISEASE`, yet are considered different terms without matching to the canonical name ”seizure”
2. Certain entities were closely linked, but technically different terms. For instance, the terms ”schizophrenia” and ”schizophrenic” are both `DISEASE` terms that appear in our dataset. However, a person is schizophrenic if and only if they have schizophrenia, so they should be considered the same term.
3. Certain entities were abbreviations of each other. For instance, the entity ”COPD” appears in our dataset, yet we should use its full name ”Chronic obstructive pulmonary disease”

Once we had canonical names for each disease occurring in the relevant sections of the clinical note, we separated them into primary diagnoses and underlying factors as follows: all diseases occurring in the primary diagnosis section were considered to be primary diagnoses and all others were considered to be underlying factors.

Finally, to create our final database, we iterated through our primary diagnoses. For each primary diagnosis, we then iterated through the diseases within it. For each of these diseases, we considered all of the other diseases in the corresponding underlying factors to be a subset of the factors for this disease. We then took the union of all of these factors for a given disease to be the full set of factors for the disease. We finally created a database of two columns: one for diseases that occur as a primary diagnosis and the other as the set of all the factors for these diseases.

The last step in our solution was to include a method to query it. Since our database has diseases in their canonical names, we have our system take in a name of a disease and use our UMLS linker to link it to its canonical name. If no such name is found, we return an empty set of factors. If we do find a name, we then check the corresponding entry of our database. If we have such a disease, we return the set of underlying factors. Otherwise we return that such a disease is not in our database.

5. Bonus Task

Our bonus task was to implement a method to extract all the primary diagnoses from a clinical note as well as evidence for each diagnosis by cross-referencing our database. We built a system that could take in a `.txt` file and give back these primary diagnoses and their relevant factors.

When given a path to a `.txt` file, our system first loads in the text from the file. We then perform the same data cleaning on the text as we did for our `txt_df` DataFrame. Namely, we split off the diagnosis, history, and complaint sections, followed by finding the primary diagnosis section. We then extract the diseases from each section using the same NER model as before and convert them to their canonical names using the UMLS linker. We then split all the found diseases into primary diagnoses and factors. For each disease in our primary diagnoses, we then query our database for the underlying factors associated to this disease. We then take an intersection of the set of underlying factors obtained from our database and the set of factors obtained from the clinical note as our evidence. We then return each disease along with their evidence.