Entity Lining with Graph Neural Networks

Introduction:

Notes taken by a medical practitioner during patients’ appointment serve as a rich source of information regarding patient’s health condition, diagnosis, and treatment. Advancements in natural language processing (NLP) methodologies allow us to process and utilize this information, contained in medical notes, for various medical decision problems. Information extraction tasks i.e., entity recognition and relation extraction, are the foundational steps in the processing the textual information available in medical notes. One prevalent issue of usage of abbreviations, shorthand forms, and synonyms (a non-standard form of any medical term) in medical notes (for symptoms, diagnosis, treatment, drug etc.) has proven to be an impediment for information extraction tasks. In this proposed study, we will present a graph neural network (GNN) methodology for disambiguating the non-standard mentions of medical terms.

The disambiguation task or the entity linking task aims to assign a standardized label to every medical term appearing in a medical note (or any text having mentions of medical terms), e.g., heart attack will be linked to myocardial infarction. This task requires an external knowledge source; a collection of standardized names/codes; and the mentions of medical terms appearing in the input text will be linked to one element of this external knowledge source. Effective utilization of an external source of knowledge by combining sequential semantic information (input text) with non-sequential information (knowledge base) is one added advantage of GNN. In the proposed study, we focus on implementation of method presented in Wu et.al. [1] for medical term disambiguation and experiments to improve the performance.

Problem definition:

The input data for the problem under consideration are collection of text documents and a knowledge base (or knowledge graph). With the help of these data, our aim to develop a model which will provide us links between entity mentions in the text documents to the knowledge graph nodes i.e., to standardized medical term. Hence, we propose to develop a model as follows,

and are the inputs to the model. is the set of all text documents (say medical notes), hence,

Let, stand for an entity mention and let be the set of all entity mentions found the given set of documents , therefore,

The other input is the knowledge graph,

Lastly, is the set of links between entity mentions and standardized medical term corresponding to a specific entity mention. Therefore,

Data:

As mentioned in the problem definition the proposed model development will need two data inputs, and i.e., set of text documents and a knowledge graph. In Wu et.al., authors have utilized Unified Medical Language System (UMLS) for the knowledge graph input and have used following datasets for the input of text documents, [1]

1. MIMIC-III

MIMIC-III is collection of de-identified health-related data of 40,000 patients who visited/stayed in critical care unit of Beth Israel Medical Center between 2001 to 2012 [2]. This dataset contains 26 tables of various features linked to each other. Importantly, MIMIC-III contains medical notes for all patients during their stay at the medical center. We plan to use these medical notes (text documents) as an input to our model.

1. ShARe

Similar to MIMIC-III, ShARe dataset is a subset of MIMIC-II dataset [3]. ShARe data set is a base evaluation dataset for SemEval-2014 Task 7 . The seventh task in SemEval-2014 has two subtasks, first, identification of disorder mentions (entity detection) and second, normalization of disorder mentions (entity linking). The data for first subtask consists of the medical notes and span of the disorder mention as the target variable. While the data for second subtask consists of disorder span and their corresponding ID in the UMLS knowledge base.

Authors of Wu et.al also have developed and tested their model on MDX, BioCDR and NCBI datasets but for this study we will focus on MMIC-III and ShARe. In case any permission delays are observed for the access of MIMIC-III and ShARe, we will resort to Twitter dataset for model development and testing.

Evaluation:

For testing performance of the model, we will adopt accuracy, precision, recall and F1-score as evaluation metrics and calculate these metrics for the test dataset. We will most likely use 70% of the entire test documents for model training and 15% for validation and testing, each.

We first start with extraction entities from every document available,

Once we have all entity mentions for every document, we generate a candidate set for each of the entity mention with the help of the external knowledge source,

After candidate set generation, the main step of disambiguation takes place. First we calculate similarity score between entity mention and each candidate and chose the one candidate maximizing the score,

As described in Wu et.al., entity linking task can be broken down into following steps,

1. Candidate entity generation
2. Candidate entity disambiguation
3. Linking result

Evaluation metrics:

1. Accuracy

Information extraction is a critical building block in the larger task of automated knowledge graph generation. Information extraction aims at detection entities and linking pairs of entities. Various deep learning methodologies have been developed to detect and link the entities in the text. Even though these models encode the semantic information they are missing the non-semantic information required to effectively detect ambiguous mentions of entities and links between them.

Here we have mention that the study is specifically about entity normalization.

Graph construction:

As the input for the IE task usually is a text piece, the nodes can be either words or sentebces for the longer documents. The links then capture the relationships between the words or the relations between the sentences. For this study we will construct the graph with words as nodes and then create links between the words based on an external knowledge source such as UMLS..