Entity Linking 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. But prevalent 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. Such standardization of the medical terms is critical for language understanding and downstream NLP tasks such knowledge graph completion, summarization, question-answering.

Medical term normalization requires an external knowledge source; a knowledge graph (KG); and codes/definitions/standard names contained in this KG will act as the target definitions for the ambiguous medical terms in the given text. For effective utilization of KG, it is essential to combine sequential semantic information (of input text) with non-sequential information (structural information in KB). While language models provide a great tool to capture the sematic information, they fail to capture the structural information (or non-sequential information). GNNs can be leveraged to overcome this gap to improve the performance of information extraction tasks.

In the proposed study, we focus on implementation of method presented in Vretinaris et.al.1 for medical term disambiguation. Vretinaris et.al.1 combine both the semantic and the structural information with a GNN approach for medical term normalization.

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 Vretinaris et.al. 1, authors have utilized Unified Medical Language System (UMLS) for the knowledge graph input and have used following datasets for the input of text documents 2,

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 3. 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 4. ShARe data set is a base evaluation dataset for SemEval-2014 Task-7 4. 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 Vretinaris et.al.1 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 5,6. In case any permission delays are observed for the access of MIMIC-III and ShARe, we will resort to Twitter dataset available under the (SMM4H)-2017 task for text classification and concept normalization for model development and testing 7.

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.

References:

1. Vretinaris, A., Lei, C., Efthymiou, V., Qin, X. & Özcan, F. Medical Entity Disambiguation Using Graph Neural Networks. *Proc. ACM SIGMOD Int. Conf. Manag. Data* 2310–2318 (2021) doi:10.1145/3448016.3457328.

2. Wu, G., He, Y. & Hu, X. Entity Linking: An Issue to Extract Corresponding Entity with Knowledge Base. *IEEE Access* **6**, 6220–6231 (2018).

3. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).

4. Pradhan, S., Elhadad, N., Chapman, W., Manandhar, S. & Savova, G. SemEval-2014 Task 7: Analysis of Clinical Text. 54–62 (2015) doi:10.3115/v1/s14-2007.

5. Li, J. *et al.* BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *academic.oup.com*.

6. Doğan, R., Leaman, R., informatics, Z. L.-J. of biomedical & 2014, undefined. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Elsevier*.

7. A, S. *et al.* Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *J. Am. Med. Inform. Assoc.* **25**, 1274–1283 (2018).