Group Name: MetaMining

Member (single): Chaochao Zhou (project leader); NetID: cz76; Email: cz76@illinois.edu

Topic: Mining of Literature Data for Meta-analysis

1. Background

A meta-analysis is a statistical analysis that combines the results of multiple scientific studies. Meta-analysis is particularly attractive in some studies. For example, in human clinical studies, a common difficulty to face is that only a few subjects can be recruited. There are many reasons to cause a small sample size, such as detriments in experiments (side-effects or complications of medicine or treatments, high-dose radiation exposure, etc.) and heavy workload required for complex experimental setups and data processing. Based on the experimental data with a small sample size, it typically results in a lack of statistical power in rejecting null hypotheses in statistical analyses. Correspondingly, the conclusion from statistical analyses could be biased.

Meta-analyses can be performed when there are multiple scientific studies addressing the same question, with each individual study reporting measurements that are expected to have some degree of error. However, meta-analysis may require researchers to extract/collect sufficient data by reviewing numerous publications from different groups. Such process is tedious, time-consuming, and error-prone. It is expected that an automatic pipeline is available to mine data from a literature database. However, such relevant studies and techniques were less reported.

2. Objectives

The overarching goal of this project is to develop a pipeline to automatically mine data from clinical or biomedical literature obtained from the PubMed database (https://pubmed.ncbi.nlm.nih.gov/). In particular, I will focus on detecting the association of numbers with lexical units (e.g., words or phrases).

3. Challenges

There are several challenges in this task to find associations of numbers and texts, including:

1) Multiple numbers may exist in a single sentence, so it requires to find the corresponding pairs of numbers and measures. For example:

> *"From April 2016 to October 2017, 106 patients were treated with the FlowTriever System at 18 U.S. sites."*
> [https://pubmed.ncbi.nlm.nih.gov/31072507/]

2) Key measures may be represented by abbreviations. For example:

> *"At 48 h post-procedure, average RV/LV ratio reduction was 0.38 (25.1%; $p < 0.0001$)."* [https://pubmed.ncbi.nlm.nih.gov/31072507/]

3) A number is associated with multiple syntactic structures hierarchically. For example:

> *"An MRI was performed for 80% of patients, showing severe DWI lesion for 28% of patients and associated FLAIR hyperintensity for 58% of patients."* (Note that 28% and 58% belongs to 80% of patients who received MRI) [https://pubmed.ncbi.nlm.nih.gov/23684343/]

4. Plans about Approach and Tools

The project was considered to divide into two stages, but these two stages will be closely interacted in order to incrementally expand datasets and improve prediction accuracy.

In the first stage, I plan to initially create a labeled dataset from about 100 abstracts for training and testing; the resulting examples would consist of sentences from the 100 abstracts. As a pilot study, I will collect paper abstracts in a literature database. Although the abstracts are very short compared to full-length articles, the abstracts include the most important quantitative data/results of a paper. Through this process, I will further find and summarize association patterns that were not listed in *Section 3*. Furthermore, I will attempt to develop an annotation interface for the convenience to label texts and incrementally expand the dataset. During this stage, I will primarily adopt text retrieval tools, such as **MeTA** in Python.

In the second stage, I will learn associations of numbers with labeled lexical units. As a pilot study, I will focus on tackling the most direct associations, which would be the cornerstone for further mining hierarchical relations (see *Section 3*) in the future. Basically, machine learning will be involved in this development stage. I anticipate that tools like **BERT**, **PyTorch** in Python and association-rule based models will be employed. I will also develop measures to evaluate the prediction accuracy using the test dataset. The learning process will closely interact with data expansion, by incorporating human assistance in labeling to boost prediction accuracy.

5. Expected Outcomes and Workload

I aim that the accuracy of machine learning to predict number and text pairs is greater than 70%. To that end, I will create an initial dataset from 100 abstracts of clinical or biomedical papers. By careful reading, I will also explore more association patterns. Therefore, I predict that the main workload lies in data preparation (Stage 1: 20 hours), compared to machine learning algorithm development and validation (Stage 2: 10 hours).