Group Name: MetaMining

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

Topic: Mining of Literature Data for Meta-analysis

1. The progress made

In this project, my objective was to develop a machine learning pipeline to mine the relationships between numerals and text from medical publications. By searching the database of "PubMed" (https://pubmed.ncbi.nlm.nih.gov/), I collected 2451 abstracts of publications in the recent year (from Nov 2021 to Nov 2022) using the single keyword of "thrombectomy". I randomly chose 100 abstracts, and I segmented each abstract into sentences. Then, I filtered sentences that include numerals, resulting in 521 sentences.

Furthermore, I have started to label the relations between numerals and text in each sentence. Currently, I have labeled 66 sentences. As shown in **Fig. 1** (which presents 5 labeled sentences in an abstract), I defined two relations, including the numeral-unit relation and the numeral-target relation. The numeral-unit relation is mainly used to locate the unit of a numeral; especially, I also included a percentage, e.g., "50% of patients", into the numeral-unit relation. The numeral-target relation mainly locates the measure corresponding to a numeral; if there are multiple candidate targets for a numeral, I only considered the most direct measure.

2. Remaining tasks

The remaining tasks include: 1) completing the labeling of 512 sentences including numerals in 100 abstracts; 2) developing and training a machine learning model to automatically extract the relations between numerals and text.

3. Challenges/issues being faced

For the numeral-target relations, it is possible for a numeral to correspond to multiple candidate targets in a sentence. Although I only considered the most direct relation from these relations, ambiguity cannot be avoided. I assume that a bigger training dataset can help the machine learning model more intelligently identify the most direct relation, similar to human's judgment. In the step of model training, I will perform cross validation to explore how many data are needed to train the model.

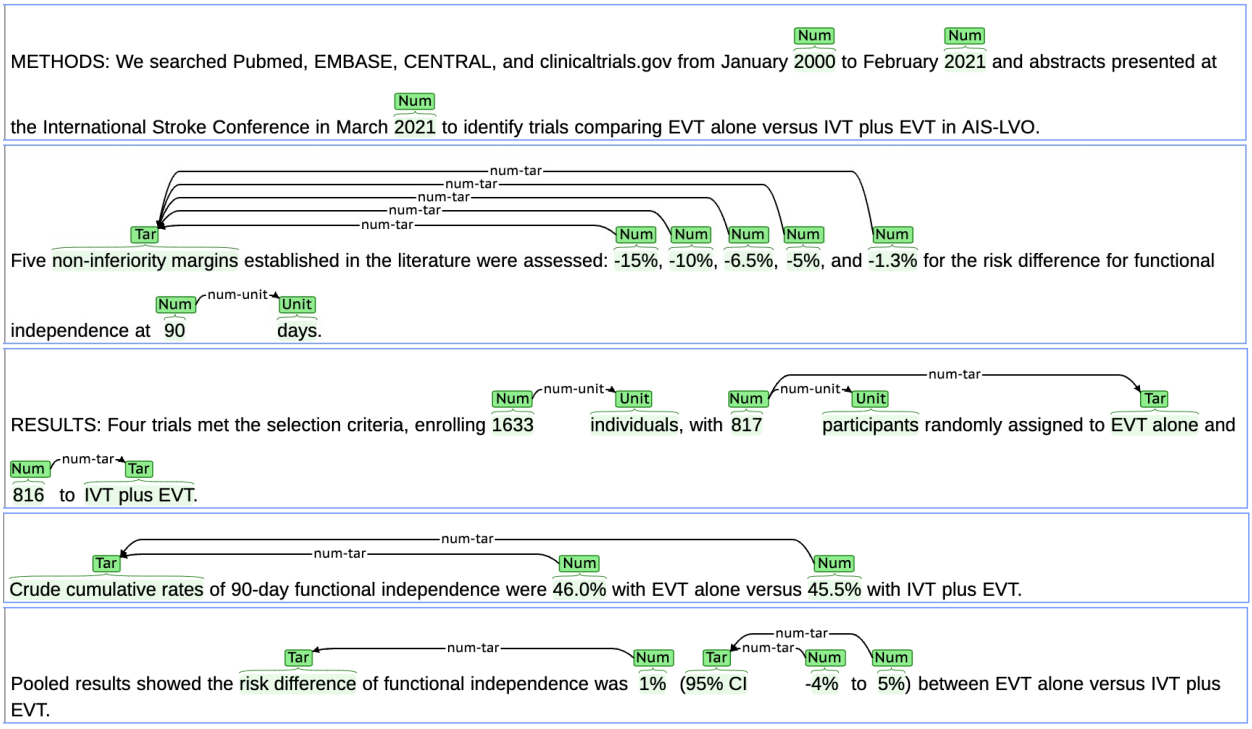

Fig. 1: The number-unit and number-target relations (denoted by "num-unit" and "num-tar", respectively) identified in 5 sentences of an abstract (PMID: 34266909). In particular, the first sentence is a case where no relations about the years were found.