Automated Medical Report Summarization and Terminology Explanation

CS6120: Natural Language Processing Final Project Proposal

Problem Description:

Medical reports often contain long pages of cryptic words which only medical professionals can understand, and it is a laborious task for people to search each term on the internet and understand it. This project aims to enhance the understanding of medical reports by common people by generating a summary of the long reports and adding definitions for medical terms at the end of summary.

Approach/Method:

We have split the entire problem into three subproblems as listed below:

- Summarize the medical report into a short paragraph.
- Extract the medical terms from the summary.
- Link the correct definition for the terms and attach it to the summary.

Prior to training, the clinical reports undergo essential preprocessing steps, including text cleaning, normalization, tokenization. Since the input data contains only medical reports and no personal details of the patient, data de-identification is not required. These measures ensure that the input data is appropriately formatted and standardized, contributing to the overall effectiveness of the model.

We plan to use Transformer models like BERTSum or BART for generating summary on medical reports. Regular NNs cannot match the performance of Transformer models like BART that has already been trained over extensive data. BART possess bidirectional and auto-regressive capabilities that helps in efficiently convert complex medical information into concise summaries while maintaining crucial details. Its pre-training on extensive text data ensures a deep understanding of medical terminology and context. To fine-tune BART for this project we will train it over the n2c2 data.

We have obtained clinical reports from n2c2 NLP research data sets from Harvard Medical School. This n2c2 has a plethora of dataset related to clinical reports and we will be using 2011 Coreference Challenge and 2012 Temporal Relations Challenge.

For extracting medical terms from the summary, we will be using Named Entity Recognition models from spaCy module. The spaCy NER models are effective in recognizing entities, but the issue is that these models are already trained and cannot be fine-tuned. Additionally, we plan to study BioBERT Transformer model which is another pre-trained model on clinical data. We can again utilize the n2c2 dataset for fine-tuning the BioBERT model.

And lastly for the medical term definition we will use established medical terminology databases, such as the Unified Medical Language System (UMLS), to ensure accurate and contextually relevant definitions

are associated with each identified term. Access permission has already been requested for this system and we are hopeful that we will get access.

Scope:

The system aims to improve the understanding of patients by generating concise, informative, and easily understandable medical summaries. The project encompasses diverse medical scenarios using the N2C2 NLP research data and UMLS data and its success is measured by the accuracy of summarization, effective term extraction, and linking the terms to its definition.

Datasets and Models used:

Report Summarization:

Data:

- n2c2 NLP Research data: https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
    - 2011 Coreference Challenge dataset
    - 2012 Temporal Relations Challenge dataset

Model:

- BART

Medical Term Extraction:

Data:

- n2c2 NLP Research data: https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
    - 2009 Medication Challenge
    - 2014 De-identification and Heart Disease Risk Factors Challenge

Both the above datasets have annotated information about medications which is essential for extracting medical terms from the summary

Model:

- SciSpacy: A pre-trained model from spaCy module
- BioBERT: This transformer model can be fine-tuned for entity recognition by training over the selected dataset.

Medical Term definition:

To link medical terms extracted from the summaries to their definitions, we propose leveraging existing medical terminology databases such as UMLS. The system will programmatically connect identified terms to their corresponding definitions, providing users with accurate and reliable information.

Team Members:

- Ravi Shankar Sankara Narayanan
- Prithiv Rajkumar
- Mahadharsan Ravichandran
- Bhuvan Karthik Channagiri

Citations:

- https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
- https://uts.nlm.nih.gov/uts/
- https://www.analyticsvidhya.com/blog/2023/02/extracting-medical-information-from-clinical-text-with-nlp/