ANLY521: Computational Linguistics and Advanced Python Group Project

Group members: Ryan Walter(rcw76), Clara Richter(cr1100), Hyuksoo Shin(hs1062)

1. Topic

"Using NLP to better understand medical documents."

- Exploring NLP models to convert complex medical terms into less technical (layman's terms)

2. Overview

Medical text (such as doctor and nurse notes) can be challenging to understand because of the technical terms. Our project aims to transform medical text into something clearer to the average person. Therefore, we are adapting natural language processing (NLP) to create a simplified version that maintains the most important details from a complex source.

Specifically, we will use medical information, such as patient transcripts, employing NLP techniques, like named entity recognition (NER) and summarization, to convert medical transcriptions to something more understandable to the layperson.

3. Literature review

There are two main ways to overcome complicated technical language, the use of simplification/summarization and NER. Both methods address the issue in separate ways. Simplification/summarization tries to transform the text into something that is easier to understand. While NER aims to add context while preserving the integrity and original meaning of the text. Each method aids the reader in understanding complex documents and text.

Simplification/Summarization

To overcome the complexity of technical vernacular, the first method is through simplification and summarization. The use of simplification first used simple methods to simplify but has shifted to using more complex methods and increasingly specific and complex datasets for individual tasks.

Initially the use of synonyms and paraphrases was used in simplification by converting words into their simpler synonyms resulting in a more readable sentence. However, this sometimes led to errors in grammatical fashion such as incorrect tenses and errors in meaning such as near synonyms not resulting in similar sentence meanings when used interchangeably (Inui et. al, 2003).

The next step seen in simplification was in the using online sources and the use of machine learning. Wikipedia and the corresponding simple Wikipedia form pairs for models to train on, however it struggles to simplify highly specific technical text through this method (Zhang et. al, 2017). Shifting towards more technical and specific applications requires more robust and specific datasets to form the basis of a trained model. A specific dataset example is the use of abstracts paired with plain language summaries written by the corresponding authors from research articles. This method follows similar steps as the Wikipedia method; however, the plain language summary is more of a summary for those who understand the lexicon then a simplification for a common person (Devaraj et. al, 2021).

The method that has the greatest potential for results is forming a highly specific custom dataset for the task at hand. These methods and dataset form what is called Simplified Technical English (STE) which allows the reader to understand the text at a fundamental level. STE consists of writing rules specifying grammar and style, and a dictionary which controls meanings of words and the context they can be used (Knezevic, 2015). This idea led to the formation of ASD-STE100, a guide in technical English for the use in the aerospace industry to simplify maintenance and training manuals for readers around the globe. All methods described above play an important role in simplification but are best used for their specific scale i.e custom STE for specific applications while synonyms for more general applications.

The use of simplification in healthcare is a complex issue. However, the combinations of multiple methods such as a custom STE using SNOWMED codes, and the training based on abstracts and their plain language summaries combines the best of specifics with the ability to summarize ideas.

NER

Attempting to make technical documents intelligible to the common reader is a difficult challenge, particularly in the healthcare industry. Thus, new methods are being and need to be further adapted to address these issues. Previous research has shown that a course for undergraduate medical students can increase the awareness of using plain language in patient communication. In addition, the use of plain language in written communication was greatly improved by translating medical reports for real patients (Bittner et. al, 2015).

In 2019, BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) was introduced. It was adapted from the pre-trained language model BERT and can perform NER and other biomedical NLP tasks with high accuracy (Lee et. al, 2019).

The GENIA project is an NLP project that focuses on developing text-mining tools for biological and medical research. The project began in 1999 as a collaboration between the University of Tokyo, the University of Manchester, and the National Center for Biotechnology Information (NCBI).

4. Dataset

The dataset utilized in this project is a CSV file that is accessible on Kaggle and was obtained by scraping data from mtsamples.com. The resulting dataset is in text format, and each transcription is classified into a particular medical specialty. Additionally, the dataset used in this project includes a column for relevant keywords extracted from each transcription. These keywords can be used as input for the NLP techniques to improve the accuracy and relevance of the summarization process and as the entities to label in the NER process.

Link:https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions

5. Models and Validation Methods

In this project we will attempt to use the two methods described in the literature review, simplification/summarization, and NER.

For the simplification/summarization we will attempt to use synonym replacement and term replacement using dictionaries. Evaluation will use a manual scoring using a scale of 1-10 of readable, understandable, and simplicity.

For the NER processing we will be using scispaCy, a Python package that contains spaCy, a modeling tool used for processing biomedical, scientific, or clinical text. There are several NER models from scispaCy that specifically deal with labeling biomedical data (ex. en_ner_bionlp13cg_md, en_ner_bc5cdr_md, en_ner_jnlpba_md, en_ner_craft_md).

To evaluate the model NER accuracy, we will divide the annotated text dataset into subsets for model training, validation, and testing. Once a spaCy NER model has been trained, we can call the model testing dataset to get the model accuracy performance results. From there we can determine the precision, recall, and f1 scores for the tokens, entities, and labels.

6. Conclusion

This project proposes a novel approach to transform medical text into more accessible language using NLP techniques. The proposed methodology involves the use of NER and summarization to extract medical entities and generate a summary of the transcription in more accessible language. The proposed model will be evaluated using qualitative and quantitative evaluation metrics to ensure its effectiveness. The resulting model will have important applications in improving health literacy and enabling better communication between patients and healthcare providers.

Works Cited

Bittner, Anja, et al. “Translating medical documents into plain language enhances communication skills in medical students - A pilot study.” *Patient education and counseling* 98(9) (2015): 1137-1141

Devaraj, Ashwin, et al. "Paragraph-level simplification of medical texts." *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. Vol. 2021. NIH Public Access, 2021.

Inui, Kentaro, et al. "Text simplification for reading assistance: a project note." *Proceedings of the second international workshop on Paraphrasing*. 2003.

Knezevic, Jezdimir. "Improving quality of maintenance through Simplified Technical English." *Journal of Quality in Maintenance Engineering* 21.3 (2015): 250-257.

Lee, Jinhyuk, et al. “BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining.” *Bioinformatics*, vol. 36, no. 4, 2019, pp. 1234–1240., https://doi.org/10.1093/bioinformatics/btz682.

Tateisi, Yuka, et al. "Part-of-Speech Annotation of Biology Research Abstracts." *LREC*. 2004.

Zhang, Xingxing, and Mirella Lapata. "Sentence simplification with deep reinforcement learning." *arXiv preprint arXiv:1703.10931* (2017).