Assignment 8: Mini Project - Summarization of Medical Articles

Adhvait Ananthan Srinath
Chalmers University of Technology
adhvait@chalmers.se

Poh Shi Qian
Chalmers University of Technology
shiqia@chalmers.se

1 Introduction

This report describes the implementation of a text summarization model for medical journals, drawing upon concepts from **Natural Language Processing (NLP)** and leveraging various **AI tools** and techniques. In medical research, it is critical to accurately and succinctly summarize extensive scholarly literature. With a constant stream of research publications, medical professionals and researchers rely heavily on efficient summarization methods to stay current on the latest developments. Our goal is to evaluate the accuracy and efficacy of these techniques by looking at keyword extraction and performance metrics like BERTScore. By scrutinizing extracted keywords and performance indicators, we hope to provide useful insights into the quality of medical journal summaries and their fidelity to the original research articles. This study aims to improve our understanding of the efficacy of text summarization methodologies in the domain of medical literature, resulting in improved access to and comprehension of pivotal medical research discoveries.

2 Data Preparation

2.1 Data Sourcing

The data for our project was obtained from the PubMed dataset, which serves as an extensive archive of biomedical literature and research papers.

PubMed is a database of millions of citations and abstracts covering a wide range of topics, including clinical medicine, basic research, public health, and biomedical engineering. It is maintained by the National Centre for Biotechnology Information (NCBI). As such, not

only is the dataset credible, but it is also a significant resource for our study on medical journals because it covers a wide range of issues related to medical research. By using the PubMed dataset, we were able to access a wide variety of academic articles and abstracts. This enables the project to cover a wide range of text summarization.

2.2 Data Loading

The files are in jsonlines format, which is being opened and read with the use of jsonlines. The dataset is then printed out in the form of a dataframe with the use of the Panda library. Each row in the dataframe corresponds to a medical journal from PubMed. The following are the columns of the dataframe: 'article_id': str, 'article_text': List[str], 'abstract_text': List[str], 'labels': str, 'section_names': List[str], 'sections': List[List[str]].

2.3 Data Preprocessing

The data is being preprocessed using the preprocess_text_with_ner function. This function prepares text by converting it to lowercase, eliminating punctuation, and tokenizing it into individual words. This ensures that data is normalized by converting it to a consistent format. We preprocess the data by performing the following steps:

1. Lowercasing - All text is converted to lowercase to ensure that the same word with different cases is treated identical.
2. Removing punctuation - This is done to reduce the noise. Punctuation, special characters, or irrelevant symbols are removed to focus on the essential content when summarizing the text.
3. Part-of-Speech Tagging (POS) - Each token is tagged with its part-of-speech, providing information about its grammatical role.
4. Named Entitiy Recognition (NER) - To identify named entities such as person names, locations, and organizations
5. Removing Stopwords and Named Entities - Removes common stopwords and named entities to focus on meaninful words.

Lastly, to ensure that the medical journal contains sufficient words for the summarization model to perform optimally, journals with less than 400 words are removed from the dataset before feeding into the summarization model.

2.4 Text Summarization

A pre-trained model is utilized to generate summaries via the Hugging Face transformers library. The Hugging Face library plays a pivotal role in text summarization tasks by granting access to pre-trained models through an intuitive pipeline interface. These models excel at processing input text, crafting summaries, and managing text length constraints as

necessary, thereby serving as a valuable resource for natural language processing tasks focused on summarization.

However, many pre-trained language models, including those provided by Hugging Face, have limitations on the maximum length of input text they can process. To fit the input text within the model's maximum sequence length, the article_text in the dataset is truncated before proceeding to feed the data into the model.

Once the input text is prepared through preprocessing and potential truncation, it is passed through the summarization pipeline established with Hugging Face. Concise summaries based on the truncated input text are then generated within this pipeline.

3 Result

3.1 Calculating BERT Score

To understand the effectiveness of text summarization better, BERTScore is computed to assess how closely the generated summaries match the reference summaries in terms of semantic similarity and content overlap. In this case, the abstract_text of the medical journals in the dataset is used as reference summaries.

BERTScore is also useful as it considers not only exact word matches but also the contextual embeddings of words, capturing their semantic significance. As such, a higher BERTScore indicates a stronger resemblance between the generated and reference summaries. This hence indicates a higher-quality summary with greater content fidelity and semantic accuracy.

As seen from the results obtained, the model has a precision of 0.8295 which indicates a relatively high proportion of accurately identified relevant words out of all identified words. In other words, the summary generated by the model can be considered to have effectively captured the pertinent information from the reference text. In addition, a value of 0.7937 is recorded for the recall, which indicates the extent to which the generated summary captures the relevant information present in the reference text. The harmonic mean of precision and recall, also known as the F1 Score is then derived to be 0.8110, ascertaining the overall performance of how well the generated summary aligns with the reference text based on a balanced assessment between precision and recall. This is done by considering both false positives (precision) and false negatives (recall). The relatively high F1 Score hence signifies the relatively good performance of the summarizer model.

3.2 Keyword Extraction

To supplement the evaluation metrics such as BERTScore, keyword extraction is done to provide additional insights into the content coverage and relevance of the generated

summary. This keyword extraction process aids in identifying the most important ideas or subjects in a document, allowing us to determine whether the important ideas from the article text are sufficiently covered in the summary that is produced.

As such, the top 10 keywords of the generated summary and abstract text are extracted and compared. This comparison makes it possible to assess how well the generated summary captures the essential ideas of the reference text by looking at the keyword overlap between it and the abstract. Overall, the number of common keywords or the similarity in keyword distributions between the two summaries serves as a quantitative gauge of the accuracy and comprehensiveness of the generated summary.

Based on the results obtained by our model, it can be seen that there is a 40% similarity in the top 10 keywords between the generated summary and the abstract text. This signifies that important keywords have been captured in the generated summary.

Additional insights can also be gained by comparing similarities in the keywords from the original article text with the entire generated summary. A result of 0.8554 was obtained, which indicates that the original keywords that also appear in the summary take up about 85.54% of the original article text. The same comparison was done between the original article text and the abstract text, which we are using as a form of reference summary for comparison in this case. It can be observed that a similar result of 76.99% was obtained for the latter, suggesting that our model performed relatively well in preserving crucial information in the generated summary.

4 Limitations and Future Direction

One of the limitations includes potentially neglecting the generalization across the wide range of medical domains that exist in the real world. It is hence possible that the evaluation metrics did not fully assess the efficacy of the summarising technique across the range of medical specialisations and subjects. Despite the PubMed dataset providing a variety of medical journals of varying medical topics, for summarization models to function well, special modifications and refinement may be required

In addition, the performance of the text summarization model in this project may vary depending on the complexity and length of the input text. Summaries of shorter and simpler articles may hence tend to be more accurate compared to those of longer and more technical publications. Given such cases, human judgments of the generated summary quality may also provide more comprehensive insights. As such, apart from the evaluation metrics that have been conducted, adding human evaluators to gauge coherence, clinical relevance, and general usefulness may be a measure taken to supplement the evaluation process.

There are also several promising future directions that can be considered. Firstly, fine-tuning the summarization model with domain-specific data could enhance generalization across diverse medical specialties. Additionally, integrating multi-domain datasets could broaden the scope of the summarization technique. Developing adaptive summarization approaches that dynamically adjust to text complexity could also lead to more accurate summaries. By taking into account these future directions, we can aim to overcome our current limitations and advance the effectiveness of medical text summarization.

5 Conclusion

In a nutshell, ongoing research and development activities in medical literature will be crucial for automated text summarizing systems in this field to improve quality and usability in the future. Such effective summarization by automated systems can greatly improve efficiency in the medical field by reducing the amount of time needed by medical professionals to search for suitable medical journals that they require.