Summarization of Long Medical Histories

Shivani Kannan
Computer Science
Courant Institute of Mathematical Sciences
New York City
sk10502@nyu.edu

December 18, 2023

Abstract

This project introduces a method for summarizing extensive medical histories using a combined extractive and abstractive approach. Initially, the extractive phase selects only pertinent sections of these histories, effectively reducing their length. This shortened content is then processed by an abstractive transformer model, which is limited by its context window length. This methodology is a foundational step towards the automated comprehension of technical language. The findings indicate that employing pre-trained transformers enhances the extractive process and potentially improves the quality of the abstractive summarization.

1 Introduction

In the field of medicine, healthcare professionals frequently confront the challenge of navigating through extensive patient histories to make informed decisions. This task can be both time-consuming and susceptible to errors. Consequently, the ability to efficiently summarize these records is not only crucial but also offers numerous benefits.

The goal of this project is to develop a summarization model capable of transforming lengthy and complex medical histories into brief, yet comprehensive summaries. Such a model will empower healthcare professionals to rapidly understand a patient's medical history, thereby facilitating more effective and informed decision-making.

For this initiative, we utilized the discharge notes from ICU patients in the MIMIC III dataset. These notes contain detailed patient histories and are instrumental in our efforts to create a foundation for summarizing medical histories, particularly for patients with diseases that are challenging to diagnose.

2 Related Work

Several pivotal works form the foundation and inspire this project. A key contribution is the MS2 dataset, designed for Multi-Document Summarization of Medical Studies[1]. This dataset is crucial for training models to summarize extensive medical literature, paralleling the task of condensing detailed medical histories.

M-FLAG, or Medical Vision-Language Pre-training, extends vision-language pre-training to the medical field[2]. It utilizes medical images and annotations to create a model adept at understanding complex medical content. This approach is directly relevant to projects that process and summarize medical texts and imagery.

Furthermore, recent advancements by Subramanian et al. (2019) have been particularly noteworthy[3]. They combined extractive and abstractive summarization techniques, initially using an LSTM model for extraction, followed by a transformer for abstractive summarization. They identified the LSTM extraction as a potential bottleneck. In response, our project aims to refine this approach by employing a pre-trained transformer model for both the extractive and abstractive phases. We anticipate that transformers, which have demonstrated significant improvements over traditional RNNs, will enhance the extractive process. Additionally, a pre-trained model in the abstractive step is expected to benefit from a richer information base, thereby improving overall summarization quality.

These contributions collectively advance the field of medical data summarization, providing valuable insights and methodologies that shape the development of our summarization model.

3 Approach

In this project, we address the challenge of summarizing lengthy technical documents through a dual-phase approach, incorporating both extractive and abstractive summarization methods.

Initially, our process involves the use of a BIOBERT-based extractive summarizer. This summarizer functions by generating token-level representations for each sentence in the document. These features are then passed on to clustering algorithm, which identifies the sentences most relevant for the summary.

Following the extractive phase, the selected sentences are provided as input to the T5 transformer model. The fine-tuned T5 model then generates a summary based on these inputs. By incorporating the extractive step, our method allows the summary to be influenced by key sentences from the entire document, effectively overcoming the limitations posed by the restricted context window.

3.1 Extractive Summarization Using BioBERT

For the extractive phase of our summarization process, we employed BioBERT ('dmis-lab/biobert-v1.1' model from Hugging Face), optimized for biomedical texts, in conjunction with clustering algorithms[4].

To manage the extensive length of medical documents, we encoded each document into several 512-token chunks using BioBERT. This chunk-based encoding strategy allows us to capture word embeddings within a broad context, resulting in more nuanced and rich embeddings. Post-encoding, these embeddings are pooled to form sentence embeddings. These sentence embeddings are then processed through an inter-sentence transformer layer. This layer is designed to extract the most significant two sentences from each chunk. By doing so, we ensure that the most relevant and information-rich sentences are selected from different parts of the document, maintaining a comprehensive representation of the original text.

To further refine the sentence selection process, we utilized the Elbow method to determine the optimal number of clusters for the K-means algorithm [5].The sentences extracted from each cluster are then aggregated and these extracted sentences serve as the input for the subsequent abstractive summarization phase.

3.2 Abstractive Summarization

The extracted sentences then served as input for the abstractive summarization phase. Here, we fine-tuned the T5 model, a text-to-text transformer, to generate cohesive and comprehensive summaries [6]. This step ensured that the final summary maintained the context and critical information of the original medical records.

4 Experiments

4.1 Data

For the experiments, the discharge notes from the MIMIC III database was utilised, specifically focusing on patients who were readmitted. We grouped the data by these patients, concatenating all their entry and discharge notes. This method creates a continuous narrative of each patient's medical history, capturing the complexities and nuances of their health journey over multiple hospital visits.

A key aspect of the data processing involved extracting the "HISTORY OF PRESENT ILLNESS" section from the notes. These sections, typically written by nurses or healthcare providers, provide a concise and relevant summary of the patient's current medical condition and history. These extracted sections as the ground truth summaries for our experiments.

After the processing, the compiled train dataset consists of approximately 10,000 rows. Each row in this dataset represents a unique patient encounter, with concatenated notes serving as the input and the extracted "HISTORY OF PRESENT ILLNESS" section as the summary target.

4.2 Evaluation Metric

The model's performance is evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [7]. This evaluation includes ROUGE-N, measuring n-gram overlap between the generated summaries and reference texts, and ROUGE-L, which assesses based on the longest common subsequence. Additionally, the distinction is made between ROUGE-N Recall, focusing on recall, and ROUGE-N Precision, which emphasizes precision. These metrics, similar to BLEU scores but without incorporating a length penalty, offer a comprehensive measure of the model's effectiveness in summarizing medical documents.

4.3 Baseline and Fine-Tuning

The baseline experiments encompassed extractive, abstractive, and mixed approaches to summarization. In the extractive baseline, the BioBERT model was utilized to encode documents and generate sentence embeddings. Clustering of these sentences was performed using K-means, with a maximum of 30 clusters and a 50 percent compression rate. The sentences nearest to the cluster centroids were selected for the summary. For the abstractive baseline, the T5 model was employed. The model was tested on truncated versions of documents, limited to the first 512 tokens, to generate summaries. This abstractive approach provided a baseline for summary generation based solely on the initial sections of the documents.

In the fine-tuning phase, the T5 large model was trained on the dataset produced by the extractive summarization. The training involved using sentences extracted via the BioBERT-based extractive method, supplemented with as much of the remaining article as could fit within the T5 model's context window. The model was tasked with generating ground truth abstracts for each article. Due to constraints in time and computing resources, training was conducted on a subset of 2,000 examples from the dataset. The T5 model underwent fine-tuning for both 3 and 5 epochs. At the testing phase, the model's performance was evaluated using inputs generated by both the original BioBERT extractor and the fine-tuned T5 model, across 500 test set examples.

4.4 Results

Generated Summary	Ground Truth Summary
The 38-year-old was admitted to the NI 9224** service with close neurological monitoring. She underwent coiling of an aneurysm of the left internal capillary. She was taken to the angiography suite on [**2108-9-24**] where she underwent coiling of an aneurysm.	[**Known firstname **] [**Known lastname 57383**] is a 38-year-old female who is transferred here from an outside hospital for evaluation of questionable subarachnoid hemorrhage with a negative CT. She notes that one day prior to admission, at approximately 2:30 p.m. her head and neck "felt funny." It was not painful, and she had had a gradual onset of increasing pain and then had sudden onset 2 hours later at 4:30 in the afternoon of pounding whole head headache and neck pain that she describes as the worst of her life. She also had some photophobia. She had no fever or chills. She had no history of trauma. Lumbar puncture at an outside hospital showed tube number two 117,000 red blood cells and in tube number three 95,000 red cells. Total protein was 241, glucose was 54, and there was no xanthochromia and no opening pressure was recorded.

Table 1: Comparison of Generated Summary and Ground Truth Summary (Case 1)

Generated Summary	**Ground Truth Summary**
A 73-year-old male with a history of coronary artery disease, hypertension, diabetes mellitus, and hypercholesterolemia was admitted for aortic valve replacement and coronary artery bypass graft. He underwent successful surgery without immediate complications. Postoperative recovery was smooth, with early extubation and management of heart rate and blood pressure. His diabetes was poorly controlled with preoperative medication, requiring adjustments including Lantus and Glipizide. The patient was discharged in good condition with medications including Lopressor, Aspirin, Lipitor, and Glipizide.	"The patient is a 73 year old male with a history of coronary artery disease and aortic stenosis who has had a jaw tightness with walking short distances. He has been followed by his cardiologist given his history of coronary artery disease and was discovered to have aortic stenosis. This aortic stenosis is followed by echocardiogram. The patient's coronary history is significant for percutaneous transluminal coronary angioplasty with stent to obtuse marginal one. This percutaneous transluminal coronary angioplasty was complicated by formation of a right femoral AV fistula and pseudoaneurysm which eventually required surgical repair. Cardiac catheterization in [**2196-12-7**], showed 80 percent in-stent restenosis of the obtuse marginal one which was treated with roto. Cardiac catheterization in [**2198-1-6**], showed 30 percent in-stent restenosis of obtuse marginal one. Also at that time, the patient was discovered to have a moderate to severe aortic stenosis with a mean gradient of 26 mmHg. Ejection fraction was 61 percent at the time. "

Table 2: Comparison of Generated Summary and Ground Truth Summary (Case 2)

Metric	**T5 3e**	**T5 5e**
ROUGE-1 Precision	42.5	43.2
ROUGE-1 Recall	27.3	28.1
ROUGE-1 F1	32.5	33.2
ROUGE-2 Precision	17.2	17.9
ROUGE-2 Recall	9.8	10.3
ROUGE-2 F1	12.0	12.7
ROUGE-L Precision	28.4	29.0
ROUGE-L Recall	16.2	16.8
ROUGE-L F1	19.6	20.3

Table 3: Finetuned ROUGE Scores for T5 Models

5 Discussion

The observed performance of the T5 model, particularly in its handling of longer context windows, reveals significant areas for improvement. Notably, the model tends to generate more concise summaries when faced with extended text inputs. This behavior suggests a limitation in the model's capacity to maintain coherence and relevancy over lengthy narratives. This trend is evident in the comparison of generated and ground truth summaries, where the model frequently omits crucial details or fails to capture the nuances present in the original texts. This shortfall is particularly pronounced in cases with complex medical histories or multifaceted patient information.

Furthermore, an analysis of the ROUGE scores indicates that while the model shows a reasonable degree of precision and recall, there is substantial room for enhancement. For instance, the ROUGE-1 and ROUGE-2 scores, particularly in terms of recall and F1, are indicative of the model's challenges in capturing finer details and maintaining the fidelity of the original text. The lower ROUGE-L scores also point to shortcomings in the model's ability to structure summaries in a way that closely mirrors the flow and coherence of the source material. Enhancements in the model's architecture or training methodology, such as fine-tuning with more diversified datasets or adjusting the context window size, could potentially address these limitations and improve the overall quality of the summaries generated.

6 Conclusion

This project underscores the effectiveness of using transformer models, specifically T5, in the realm of abstractive summarization. The application of BioBERT in the extractive step significantly enhanced the quality of the summaries, resulting in improved ROUGE scores compared to those reported in our initial reference paper. However, the abstractive phase faced limitations due to computational constraints, confining the training of the T5 model to just 5 epochs. Future endeavors with greater computational resources could explore more advanced models like T5-3B, potentially elevating the abstractive summarization performance.

A promising area for future work is the exploration of Recurrent Memory Transformer (RMT). This approach, capable of handling context windows up to 1 million tokens, extends the capabilities of our summarization models [8]. By incorporating RMT, we could overcome current limitations related to context window size, enabling the processing of longer documents with higher accuracy and detail. Additionally, integrating aspects of interpretability into the models would not only enhance their transparency but also improve the trust and usability of the generated summaries, particularly in sensitive domains like healthcare.

7 References

1. DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B., Wang, L. L. MS2: Multi-Document Summarization of Medical Studies. Northeastern University & AI2. https://arxiv.org/abs/2104.06486

2. Liu et al. M-FLAG: Medical Vision-Language Pre-training with Frozen Language Models and Latent Space Geometry Optimization. https://arxiv.org/abs/2307.08347

3. Subramanian, S., Li, R., Pilault, J., Pal, C. (2019). On extractive and abstractive neural document summarization with transformer language models. https://arxiv.org/abs/1909.03186

4. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. https://arxiv.org/abs/1901.08746

5. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. https://arxiv.org/abs/1910.10683

6. Marutho, D., Sarno, R., Sinaga, D. P. The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News.

7. Ganesan, K. ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks, 2019.

8. Bulatov, A., et al. Scaling Transformer to 1M tokens and beyond with RMT, 2021. https://arxiv.org/abs/1803.01937