Text Summarization of Medical Reports using Large language Models

Aashi Goyal aashiarv@usc.edu

Sanjana Parakh svparakh@usc.edu

Vedika Jadhav vedikaya@usc.edu

Richa Maurya richamau@usc.edu

Raj Patel rajashif@usc.edu

Abstract

Our project proposes an approach to summarize all medical reports of a person using natural language processing (NLP) techniques. With the increasing volume of medical data, summarization is becoming an essential task for healthcare professionals who need to extract relevant information from large volumes of data quickly. We propose to make this task simpler using Large Language Models (LLMs) and making it easier to process the data of a particular patient.

1 Introduction

The volume of documents available is constantly growing, necessitating a corresponding effort in the area of computerized processing and user-specific access to pertinent information. Text Summarization (ATS) is a technique that captures different information from a combination of reports, and condenses it such that content is well represented with no loss of important details by simplifying the tasks.

Text summarization can be an extremely helpful tool for physicians and clinicians in medicine. Clinical documents are often verbose and extensive, thus requiring considerable time and effort to understand. Summarizing these documents in a fast and reliable way would help physicians serve and deliver better services and drastically reduce human efforts to understand these reports. A summary could provide them with the key information and they can then actively decide whether they need to invest more time in reading a particular document.

Summarizing clinical documents is a particularly challenging task as the margin for rephrasing and interpretation is narrow. Parsing over text documents and generating accurate summarizes can be optimized and be made reliable with the use of Natural Language Processing algorithms and techniques. Therefore, the objective of this project is to develop an effective method for summarizing clinical documents.

2 Related work

Recently, many works are presented based on text summarization.

Summarizing long documents has always been a task, authors in [1]. have made an approach of training transformers from scratch without the need of pre-training and scaling long documents of over 6000 words. They worked on four dataset, arXiv, Pubmed, Bigpatent and NewsRoom resulting in extractive models that are lightweight and simple and performed better than TLMs that are only conditioned on the introduction in the abstractive section.

Examining recent research on text summarization for biomedical data gave a brief idea about existing methods out there. In [2] review summarization based on different metrics, and application areas. They collected data from WoS, IEEE, and ACM digital libraries and identified 3.5% of 801 studies that met the inclusion criteria. Common approaches were single document, generic extractive summarization. The use of machine learning techniques was reported in 16 studies, and Rouge was the evaluation metric in 26 studies and highlighted the use of Transformer based methodologies to address challenges in biomedical text summarization. Authors Nadif and Role in [3] gave a survey on the potential for assisting human experts in the exploration and extraction of relevant information.These embeddings have successfully interacted with common supervised modules, improving their performance.

In [4], the authors discuss the need for universal summarization frameworks due to the increase in digital text information. The proposed model in the paper outperforms other methods on clinical notes from the MIMIC-III dataset and can be integrated into a decision-support system to better interpret clinical information.

In [5], the authors propose a method for improv-

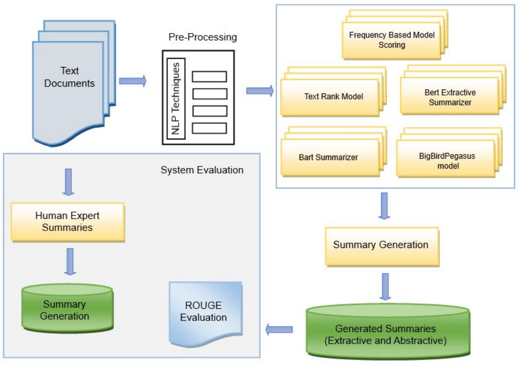

Figure 1: System Architecture

ing the quality of abstractive text summarization using a pointer-generator network and a pre-trained word embedding layer. The experimental results on the Daily Mail/CNN dataset show that the proposed model with Fasttext pre-trained achieves higher ROUGE-1, ROUGE-2, and ROUGE-L measures compared to the baseline model

3 Problem description

The goal of this study is to address the problem of lengthy medical reports that are dense with technical information, making it difficult for patients and healthcare professionals to quickly find the information they need. Previous studies have been conducted using abstractive and extractive models. However, there is still a gap in determining which one is the most appropriate. Sometimes a particular model provides a good summary for one type of document, such as general medicine, but fails for other domains due to complex terms, such as radiology, neurology, and others. As part of this project, we are testing both the extractive and abstractive text summarization models on a single document. The goal of the comparison is to determine which models are most effective in generating accurate and coherent summaries of input text.

4 Methodology

Our model architecture is depicted in Figure 1.

4.1 Corpus Preparation

We created our own dataset by scraping the website, https://www.mtsamples.com. It is an open-access publicly available database of unstructured and anonymized health records. Each of these reports have been transcribed. The corpus included reports from many different medical specialities and they have been tagged. Overall, our dataset includes over 5000 sample medical reports.

4.2 Pre-Processing

We needed to pre-process the extracted data to better suit our purpose. The initial steps involved cleaning the dataset. We had to drop certain samples from our dataset since they had missing values in the description section of the report.

We experimented with various methods of striping for the words. We removed extra spaces, URLs, and '—' characters. We found converting it to lowercase does not help as medical reports will have a lot of capsized letters.

We also did not remove the stop words as for context this big, we think it is not going to help much and for technical writings like Medical report logs, there are little stop words, to begin with.

4.3 Summarization Algorithms

We explore methods to generate extractive as well as abstractive summaries for medical reports in an attempt to create a model that would generate precise summaries.

- **Extractive Summarization** : This method of summarizing involves selecting the most important sentences or phrases from a longer piece of text and presenting them as a summary.

- **Abstractive Summarization** : This method involves generating sentences that is based on the original text. The final output is original and does not include any kind of combination of the input sentences.

Due to the sensitive nature of documents we are working with and their vulnerability to incorrect interpretation, we decided to mainly focus on extractive summarization techniques though we did explore the possibility of using an abstractive model to generate qualitatively better summaries.

4.4 Models

1. **Frequency Based Model Scoring** : This is an extractive summarization model. It is a frequency-driven approach based on common words that are often repeated and do not carry salient information. We used the Spacy NLP pipeline that returns a tokenized object which is used to create a dictionary of words

and their frequencies. Each sentence is then scored based on the relative sum of word frequencies. The final summary contains the k sentences with the largest scores.

2. **Text Rank** : Textrank is a graph-based ranking algorithm like Google's PageRank algorithm. It measures the relationship between two or more words. TextRank revolves around the idea of representing the concerned lexical unit in the form of a graph and later uses the famous PageRank algorithm to rank those chunks. We used the python summa library to generate the summaries.

3. **Bert Extractive Summarizer** : Bert is a transformer based model. In this method we first tokenize the input sentences and then feed it to the BERT model to get the sentence embeddings. Subsequently, we cluster the sentence embeddings using K-means and pick the sentences closest to the centroid. We used the ELBOW method to determine the optimal number of sentences in the final summary.

4. **Bart Summarizer** : After extensive research and performance comparison, for our use cases we used the facebook/bart-large-cnn summarizer. The BART model is pre-trained on English language, and fine-tuned on CNN Daily Mail. BART is a transformer encoder-encoder (seq2seq) model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder. It contains 1024 hidden layers and 406 million parameters.

   We performed the summary generation on our corpus using this pretrained model initially. We also decided to fine-tune this model on our corpus of data to produce better results.

   Since we did not have human annotated summaries, we decided to use the 'transcription' as the text to be summarized while the 'description' is the target that we want to obtain. We used the Blurr library for training the model.

   The tuning process involved creating a mini-batch of inputs and targets, and specifying the hyper-parameters such as maximum sequence length of 256, batch size 2, learning rate 4e-5 and an ADAM optimizer. We used the 1cycle policy for fitting the model. We used a variety of performance metrics such as rouge score, bertscore precision, recall etc. to evaluate the model during training.

5. **BigBirdPegasus model (large)** : It is a sparse attention based transformer which extends Transformer based models, much longer sequences. It can handle sequences up to a length of 4096.

5 Experimental results

5.1 Dataset

The dataset was not publicly available and therefore had to be scraped from mtsamples.com and complied.

5.2 Baseline

We need to ensure that the model is performing on a wide range of records as expected. To ensure this we need to have a reference of document summaries. The issue with baseline is resolved by working with different implementations to see if most will return similar summaries. Additionally, we are also manually writing medical report summaries with help from medical professionals to evaluate the automatically generated summaries.

5.3 Evaluation Protocols

In the context of medical report summarization evaluating model summaries is tough as the factual correctness of the summary is the paramount criteria. Additionally, the coherence and fluency of the summaries is extremely important as well.

Content based evaluation metrics such as ROUGE scores rate the summaries based on the provided summaries. We worked along with medical professionals to manually summarize a small subset of the medical reports in order to evaluate them against the model generated summaries. We will be using the ROUGE score as the metric to evaluate the quality of our models.

ROUGE Score : It considers the recall and precision between the candidate (model-generated) and reference (manually-annotated) summaries to evaluate the quality of the text summarization system. It basically measures the overlap between the two summaries.

5.4 Results

The quantitative measure used for evaluation of the generated document summaries is the Rouge Score. The BigBirdPegasus model is not included in the quantitative evaluation as when exploring

Frequency Based Model Scoring : Because I told him that I did not feel it was ethical to just put him on the radical regimen that him and his friend devised, we compromised and elected to go back to Temodar in a low dose daily type regimen. I will look at this as a positive sign because I think radiation is the one therapy from which he can get a reasonable response in the long term.

Bert Extractive Summarizer : He comes in for an urgent visit because of increasing questions about what to do next for his anaplastic astrocytoma. Emphasizing this once again, in addition, to recommending steroids I once again tried to convince him to undergo radiation. If he tolerates this for one week, we then agree that we would institute another one of the medications that he listed for us.

Big Bird Pegasus : a man with a primary glioblastoma multiforme (gbm) presented for radiation therapy.<n> initially, the patient was to receive radiation to the base of the brain with a dose of 50 gy ; however, because of an apparent lack of response, additional chemotherapy was administered.<n> subsequently, with progression of the disease, the patient was to receive radiation to the base of the brain with a dose of 25 gy ; however, because of an apparent lack of response, additional chemotherapy was administered.<n>

Bart Summarizer (Fine - tuned) : Consult for urgent visit because of increasing questions about what to do next for his anaplastic astrocytoma. The patient has clearly been extremely ambivalent about this therapy for reasons that are not immediately apparent. It is clear that his MRI is progressing and that it seems unlikely at this time that anything other than radiation would be particularly effective.

Table 1: Qualitative evaluation of summaries generated by different models

Average Rouge F1 - Scores			
Models	Rouge-1	Rouge-2	Rouge-L
Frequency Based Model Scoring	0.132	0.018	0.124
Text Rank	0.336	0.160	0.325
Bert Extractive Summarizer	0.365	0.178	0.352
Tuned Bart Summarizer	0.455	0.380	0.411

Table 2: Rouge Scores (F1) of all models

the summaries generated by the model we realized that there were instances where the model was hallucinating which could be extremely detrimental especially in medical applications. A qualitative assessment of the different models (Table 1) shows how varied the medical summaries could be depending on the model. The BigBirdPegasus model shows completely rogue results in this case.

Based on the reference summaries we got the following Rouge scores (Table 2). Our fine tuned BART model gave us the best results among all the models which is also reflected in the qualitative analysis (Table 2) as it provides the most comprehensive summary.

6 Conclusion and Future Scope

This paper aims at comparing the quality of different extractive and abstractive text summarization models specifically in the medical domain. According to the experimental results that we have, the Fine-Tuned Bart Summarizer outperformed other models such as Frequency Based model Scoring, Text Rank, pre-trained Bart Extractive Summarizer, and BigBirdPegasus in terms of ROUGE-1, ROUGE-2, and ROUGE-L measures. We believe that if the models are further trained on larger medical datasets specific to the task the overall results achieved could be improved.

Our future research plans include studying sentence features to enhance the attention's understanding of sentence-level knowledge. Future work may be more bankable if it combines abstractive and extractive summarization using neural network language generative models.

Github Link

Please find the link to our code repository below: https://github.com/raj0823/NLP_Project

Division of labor

We have all collaboratively worked towards the project. Though we all have individually focussed on specific parts, we had weekly team meets to ensure everyone is on the same page and is equally heard. Each individual's contribution to the project was of equal value.

Work Distribution	
Name	Task
Aashi Goyal	Data Collection & Pre-processing
Vedika Jadhav	Model Integration
Richa Maurya	Model Integration
Sanjana Parakh	Model Fine-Tuning
Raj Patel	Text Summarization & Evaluation

References

[1] Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.748. URL https://aclanthology.org/2020.emnlp-main.748.

[2] Andrea Chaves, Cyrille Kesiku, and Begonya Garcia-Zapirain. Automatic text summarization of biomedical text data: A systematic review. *Information*, 13(8), 2022. ISSN 2078-2489. doi: 10.3390/info13080393. URL https://www.mdpi.com/2078-2489/13/8/393.

[3] Mohamed Nadif and François Role. Unsupervised and self-supervised deep learning approaches for biomedical text mining. *Briefings in Bioinformatics*, 22(2):1592–1603, 02 2021. ISSN 1477-4054. doi: 10.1093/bib/bbab016. URL https://doi.org/10.1093/bib/bbab016.

[4] Neel Kanwal and Giuseppe Rizzo. Attention-based clinical note summarization. *CoRR*, abs/2104.08942, 2021. URL https://arxiv.org/abs/2104.08942.

[5] Dang Trung Anh and Nguyen Thi Thu Trang. Abstractive text summarization using pointer-generator networks with pre-trained word embedding. In *Proceedings of the 10th International Symposium on Information and Communication Technology*, SoICT '19, page 473–478, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450372459. doi: 10.1145/3368926.3369728. URL https://doi.org/10.1145/3368926.3369728.

[6] Seyed Vahid Moravvej, Mohammad Meleki, M Sartkahti, and Mehdi Joodaki. Efficient gan-based method for extractive summarization. pages 287–298, 07 2022. doi: 10.22061/JECEI.2021.8051.475.

[7] Hao Xu, Yanan Cao, Ruipeng Jia, Yanbing Liu, and Jianlong Tan. Sequence generative adversarial network for long text summarization. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 242–248, 2018. doi: 10.1109/ICTAI.2018.00045.

[8] Rupal Bhargava, Gargi Sharma, and Yashvardhan Sharma. Deep text summarization using generative adversarial networks in indian languages. *Procedia Computer Science*, 167:147–153, 2020. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.03.192. URL https://www.sciencedirect.com/science/article/pii/S1877050920306578. International Conference on Computational Intelligence and Data Science.

[9] Ayham Alomari, Norisma Idris, Aznul Qalid Md Sabri, and Izzat Alsmadi. Deep reinforcement and transfer learning for abstractive text summarization: A review. *Computer Speech Language*, 71:101276, 2022. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2021.101276. URL https://www.sciencedirect.com/science/article/pii/S0885230821000796.

[10] Sunhye Kim and Byungun Yoon. Multi-document summarization for patent documents based on generative adversarial network. *Expert Systems with Applications*, 207:117983, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.117983. URL https://www.sciencedirect.com/science/article/pii/S0957417422012118.

[11] Jelmer M. Wolterink, Anirban Mukhopadhyay, Tim Leiner, Thomas J. Vogl, Andreas M. Bucher, and Ivana Išgum. Generative adversarial networks: A primer for radiologists. *RadioGraphics*, 41 (3):840–857, 2021. doi: 10.1148/rg.2021200151. URL https://doi.org/10.1148/rg.2021200151. PMID: 33891522.

[12] Narges Nazari and Mohammad Amin Mahdavi. A survey on automatic text summarization. *Journal of AI and Data Mining*, 7:121–135, 2019.

[13] C Friedman, P O Alderson, J H Austin, J J Cimino, and S B Johnson. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc*, 1(2):161–174, March 1994.

[14] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Med Image Anal*, 58:101552, August 2019.

[15] Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, and Hongyan Li. Generative adversarial network for abstractive text summarization. *CoRR*, abs/1711.09357, 2017. URL http://arxiv.org/abs/1711.09357.

[16] Seyed Vahid Moravvej, Abdolreza Mirzaei, and Mehran Safayani. Biomedical text summarization using conditional generative adversarial network(cgan). *CoRR*, abs/2110.11870, 2021. URL https://arxiv.org/abs/2110.11870.