IEEE Xplore Part Number: CFP23BC3-ART; ISBN: 978-1-6654-5630-2

Text Summarization of Medical Documents using Abstractive Techniques

Evani Lalitha
Department of Information Technology
Sree Vidyanikethan Engineering College
Tirupati, India
evanilalitha789@gmail.com

Kasarapu Ramani
Professor, School of Computing
Mohan Babu University
(Erstwhile Sree Vidyanikethan Engineering College)
Tirupati
head-ds@mbu.asia

Dudekula Shahida
Department of Information Technology
Sree Vidyanikethan Engineering College
Tirupati, India
dudekulashahida01@gmail.com

Esikela Venkata Sai Deepak
Department of Information Technology
Sree Vidyanikethan Engineering College
Tirupati, India
deepakesikela@gmail.com

M Hima Bindu
Department of Information Technology
Sree Vidyanikethan Engineering College
Tirupati, India
mopurubindu20@gmail.com

Diguri Shaikshavali
Department of Information Technology
Sree Vidyanikethan Engineering College
Tirupati, India
shaikshavali19102001@gmail.com

***Abstract*. Medical researchers are exposed to enormous amounts of medical information in the form of medical news, clinical trial reports, research articles, etc. Researchers would need the documents' summaries that help them decide to do an in-depth study. Even though there are documents that contain abstracts, a lot of medical documents do not contain abstracts or summaries. The best solution to address this issue is abstractive text summarization. Extracting useful information and summarizing the information from medical documents inclined to the best interests of the user is a challenge that is addressed in this study. For this, several abstractive summarization techniques such as T5(Text-to-Text Transfer Transformer), BART (Bidirectional Auto-Regressive Transformer) and PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence-to-sequence) are used and based on ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics, PEGASUS is identified to perform better than the other models achieving the highest ROUGE score of 0.37.**

Keywords—Abstractive Text Summarization, Natural Language Processing, T5, BART, PEGASUS, ROUGE metrics

I. INTRODUCTION

Medical researchers are exposed to numerous medical documents. Going through such documents to find the relevant information is tiresome as researchers would need to go through a lot of documents to find documents and text that is relevant to the context.

The immediate solution is to create documents' summaries based on which the researcher can decide whether to continue reading the text or to move to the next document. While there are documents that contain abstracts, a lot of medical documents do not contain such abstracts or summaries.

With the advent of technology, natural language processing is used to solve such issues. Text summarization in NLP goes back to 1958 and since then it is widely incorporated into medical domain. NLP methods have an important place in the development of text processing tools[1]. Many approaches were proposed for effective document summarization[2-3] and it is applied in wide range of biomedical domains.

There are mainly two types of text summarization techniques which are extractive and abstractive. The first technique identifies the best sentences in the document that convey their contents and arrange them in a way that is understandable, called extractive summarization[4].

But when it comes to medical documents, generating novel sentences or paraphrasing the entire document which is called abstractive approach [5], seems better than using the method of extracting sentences and phrases from source text and putting them all together. In this study, several abstractive summarization techniques are implemented and their performance is compared with ROUGE metric.

II. RELATED WORK

There were several previous works that were concerned on solving this problem. In [6], an exhaustive survey on various text summarization methods used on different applications was done. Also, the future research avenues in medical document summarization were given. It primarily focused on scaling issues of the collection of large documents from different media, different languages, portability of new subdomains, integration of compression techniques into practical applications, and personalization issues. But the problem is that it only provided an overview of the methods, not an implementation.

In[7], an overview of automated summarization of texts in the field of biomedicine and healthcare was provided. It listed the benefits and limitations of using text summarization in the healthcare and biomedical domain by stating recent studies. It explored the evolution of text summarization applications in the healthcare and biomedical domain. It also explained how text summaries can be useful to researchers and practitioners in the field. Same as in [6], [7] also provided an overview of text summarization techniques, but not implementation.

In [8], the authors used abstractive summarization for long medical documents. Both extractive and abstractive approaches were used which were trained separately and then used sequentially to perform a mixed summarization approach. For extractive step, BERT was used and for abstractive step, BART was used. While comparing the approaches the extractive approaches are compared to abstractive approaches and their mixed approach. For abstractive approach only one technique was used which was BART.

In[9] a data-driven approach was implemented using abstractive summarization of statements. Local attention-based model was used that can conditionally generate each summary word given an input sentence. The author used data that can be easily trained end-to-end and can be scaled to enormous amounts of training data. Compared to some strong baselines the model showed significant improvement in DUC-2004 performance joint task. However, it provided a unified model approach. It only produced a one-line summary. It does not apply abstractive text summarization to medical documents.

In[10], the authors dealt with solving two challenges, the first one was to detect the concepts that convey the main topic well, and the second was to generate new paraphrased sentences from the essential concepts. Here also, they used a mixed approach where the first challenge was solved using extractive approach and the second challenge using abstractive approach. For abstractive step they used T5 model. Here, only one abstractive approach was used and that was also used in the mixed approach. Comparison was done only with the proposed model.

The terminology used in the medical field differs from the general terminology under which the earlier approaches are commonly used and tested.As can be seen from above, there is very little work on abstract summarization techniques and even less on medical document summarization.

In this study, the approach is to try and collect some works that are latest in the field of abstractive summarization and apply them on medical documents and perform a comparative study to identify the best abstractive text summarization.

III. DATA SETS

The datasets used in this paper are extracted from pre-processed SUMPUBMED dataset[11] derived from PUBMED. SUMPUBMED is a dataset used for text summarization. The dataset contains 32,689 text documents. It contains five folders namely, text, shorter abstract, line text, scripts and abstract.

IV. APPROACHES

Text summarization is the process of summarizing a given amount of text and the ability to present the context of the entire text in a few sentences without removing the essence of the text under consideration. In this paper, Transformer based abstractive techniques are implemented as follows:

A. T5

T5 (Text-To-Text Transfer Transformer) is a neural network-based model for natural language processing tasks such as machine translation, summarization, and text generation. It was developed by Google Research as reported in [12]. The T5 model is based on the Transformer architecture, which uses a self-service mechanism to process input text. This allows the model to efficiently handle long-term dependencies and understand the context of the text.

B. BART

BART, short form of Bidirectional Auto-Regressive Transformer is a neural network-based model for natural language processing tasks, like machine translation, text generation and text summarization[13]. The BART model is based on a transformer architecture that uses self-awareness mechanisms to process input text. This allows the model to efficiently handle long-term dependencies and understand the context of the text. BART is trained as a denoising autoencoder, meaning it is trained to reconstruct the original input text from its corrupted version. This pre-training step allows the model to learn a better representation of the input text, which can be fine-tuned on specific tasks with relatively little task-specific data.

C. PEGASUS

The Pegasus model was proposed in [14]. As in the case of summarization, it also removes/masks important sentencesfrom the input text and generates them as a single sequence of output from the rest of the sentences. It achieves the aggregate performance of SOTA as measured by human evaluation as well as ROUGE for all the 12 downstream tasks.

The proposed model, PEGASUS, masks multiple whole sentences instead of smaller contiguous text spans unlike BART and T5. Rather than selecting sentences randomly, it deterministically selects important sentences. It generates only masked sentences as single output sequence rather than reconstructing complete sequences of input as in T5. It focusses entirely on downstream summarization (generative) tasks and does not evaluate on NLU (Natural Language Understanding) classification tasks.

V. EXPERIMENTAL SETUP

A. Data

The medical articles used in this paper are extracted from SUMPUBMED dataset. The dataset contains 32,689 text documents. It contains five folders from which we use two folders namely text and abstract. Currently less than 10 articles are used against the models selected to evaluate for comparative study.

B. Models

Altogether five models are used, three comprising of variations of T5 approach and a variation of BART named BART-Large-CNN and a variation of PEGASUS named Google/Pegasus-xsum.

T5-Small isthe smallest version of the T5 model and has the least number of parameters (60 million). It is suitable for tasks that do not require a high level of understanding of the text.T5-Base is another version of T5 model that has beentrained with more parameters (220 million) and is capable ofhandling more complex tasks that require a higher level of understanding ofthe text. T5 large is the largest version of the T5 model, with the most number of parameters (770 million). It is capable of handling large amounts of data and more

complex tasks that require a high level of understanding of the text.

A variation of the BART model is BART-Large-CNN which is pre-trained on the English language. As in the case of all the other pre-trained models, they differ in the corpus used for pretraining them. In this case, the model is pretrained and fine-tuned on CNN Daily Mail as can be seen in [8].

For PEGASUS,google/Pegasus-xsum is used. It is a specific model that gives generated abstracts based on the article.Which means the model reads the article text and writes a suitable headline. The difference between the variations in Pegasus models or any other models relies on the number of parameters and the data based on which they are pretrained. In case of google/Pegasus-xsum, it is pre-trained on C4 (Colossal and Cleaned version of Common Crawl) and HugeNews corpuses.

C. Performance metrics

For evaluation metrics, ROUGE metric is used which is the abbreviation of Recall-Oriented Understudy for Gisting Evaluation scores, are used to compare the performance of models chosen for this comparative study[15].ROUGE-N metric is used to measure the n-gram overlaps between abstract and generated summaries. ROUGE-L is used to obtain a score of the summary based on the longest common sequence between generated summary and abstract. ROUGE-N Recall is the original ROUGE metric, while ROUGE-NPrecision is the metric which focuses more on precision instead of recall. ROUGE-N F1 score is the harmonic mean between precision and recall. They can be somewhat similar to BLEU (BiLingual Evaluation Understudy) score, but lack the length scoring penalty.

VI. RESULTS

Abstractive text summarization is well suited for medical documents that contain the terminology which is difficult to replicate when using extractive text summarization.

As can be seen in the Table I. The ROUGE-1, ROUGE-2 and ROUGE-L values each with precision, recall and F1 score are obtained for the medical documents for comparative study. The document thus considered is divided into chucks of 300 words each and fed into the summarizer. The summary generated from the entire document is then refed into the summarizer to obtain the final summary.

The ROUGE metric is applied to the obtained summary to the reference abstract of the considered medical document. The values obtained in the scale of 0 to 1 state that ROUGE-1 recall is highest for PEGASUS with a ROUGE score of 0.37 followed by BART with a ROUGE score of 0.31, as can be seen in Table 1. ROUGE-1 precision is highest for BART with a ROUGE score of 0.37 followed by PEGASUS with a ROUGE score of 0.35. ROUGE-1 F1 Score is highest for PEGASUS with a ROUGE score of 0.36 followed by BART with a ROUGE score of 0.33.

The ROUGE-2 Recall is highest for PEGASUS with a ROUGE score of 0.16 followed by BART with a ROUGE score of 0.11. The ROUGE-2 Precision is highest for T5 base with a ROUGE score of 0.18. In case of ROUGE-2 F1 Score, PEGASUS achieves highest score of 0.15.

The ROUGE-l Recall is found to be highest for PEGASUS with the score of 0.31. The ROUGE-l Precision is highest for BART with ROUGE score of 0.35 followed by PEGASUS with a score of 0.29. In the case of ROUGE-L F1 Score, BART has the highest achieving the score of 0.32 followed by PEGASUS with the ROUGE score of 0.30.

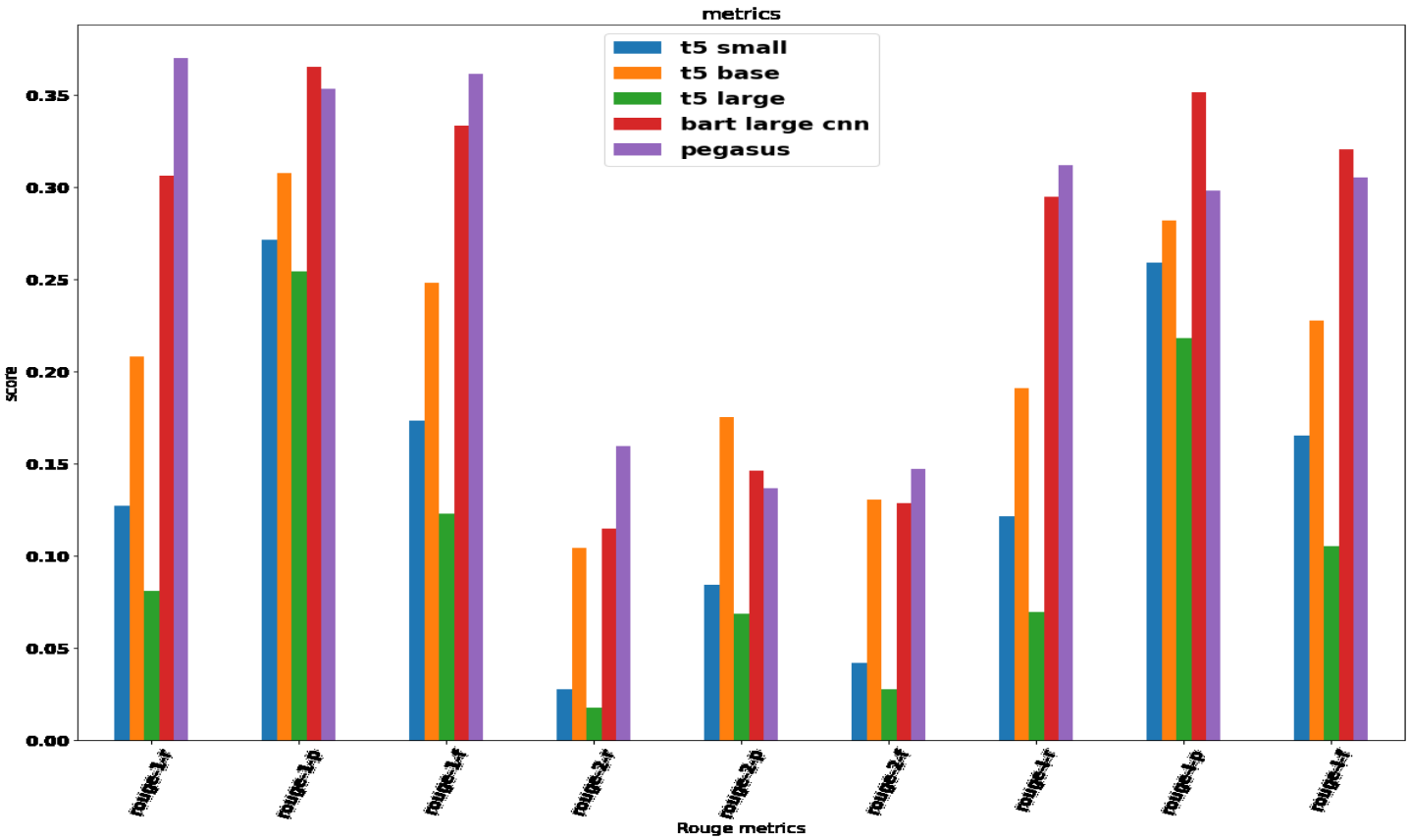

Fig.1. ROUGE performance metrics

TABLE I. Evaluation of Text Summarization using ROUGE

Performance metrics	T5 small	T5 base	T5 large	BART large cnn	Pegasus
ROUGE-1-r	0.127168	0.208092	0.080925	0.306358	0.369942
ROUGE-1-p	0.271605	0.307692	0.254545	0.365517	0.353591
ROUGE-1-f	0.173228	0.248276	0.122807	0.333333	0.361582
ROUGE-2-r	0.027778	0.104167	0.017361	0.114583	0.159722
ROUGE-2-p	0.084211	0.175439	0.068493	0.146018	0.136499
ROUGE-2-f	0.041775	0.130719	0.027701	0.128405	0.1472
ROUGE-L-r	0.121387	0.190751	0.069364	0.294798	0.312139
ROUGE-L-p	0.259259	0.282051	0.218182	0.351724	0.298343
ROUGE-L-f	0.165354	0.227586	0.105263	0.320755	0.305085

Considering all the above cases and from the graph shown in Fig.1, it is analysed that in the case of ROUGE-1 recall, ROUGE-1 F1 score, ROUGE-2 recall, ROUGE-2 F1 score, and ROUGE-l Recall PEGASUS outperforms the other approaches. If this scenario is considered, PEGASUS outperforms the other approaches in five out of nine cases of ROUGE scores and hence can be considered the best among the considered approaches. If the harmonic mean of recall and precision is considered which is F1 score, it can be seen that PEGASUS achieves highest ROUGE-1 and ROUGE-2 F1 scores and is the second highest when it comes to ROUGE-l F1 score, ifthe number of cases in which Pegasus outperforms is considered, here also PEGASUS can be said to best perform when compared to other approaches. If only the ROUGE-L score which by definition is the ROUGE score of the longest common sequence is considered, it can be seen that both BART and PEGASUS perform equally well but if only the F1 Score of ROUGE-L is considered then it can be seen that BART out performs PEGASUS with a slight difference.

Even though the results obtained do not cross the mark of even 50% of ROUGE score in any of the case but as the definition of abstractive summary states, the summary generated tries to maintain the context of the original text while generated the new text. Unfortunately there are no metrics currently found that accurately measure the summaries generated by abstractive techniques and based on

the previous works on abstractive techniques, ROUGE summary is used to evaluate the models under this comparative study.

The reason why PEGASUS outperforms other approaches in most of the cases might be because, unlike T5 and BART, PEGASUS masks more entire sentences rather than smaller contiguous text spans. It deterministically selects sentences based on importance, rather than randomly.

VII. CONCLUSION

It is difficult to go through a large number of medical documents to identify the relevant information when searching for material related to the problem at hand. In such cases, abstractive summarization helps in summarizing the entire medical document so that the physician can go through the document if it is found relevant based on the summary and skip the document otherwise. When comparing various models such as T5, BART, and PEGASUS, PEGASUS is found to perform well than the other approaches as evaluated by the ROUGE scores in most of the cases achieving a highest ROUGE score of 0.37. BART is also found to perform well as achieves a ROUGE score of 0.36.

For future work, a greater number of documents can be placed and the models' performances can be further evaluated so that more accurate answer can be obtained to which among PEGASUS and BART perform well. Although the ROUGE scores do not entirely evaluate the performance of abstractive techniques as they are best suited for extractive summarization techniques as the score it produces is based on the comparison of same words in the generated and reference summaries, it is hoped that a more relevant abstractive summarization metric is introduced to best evaluate the models at hand.

REFERENCES

[1] Fleuren, Wilco & Alkema, Wynand. (2015). Application of text mining in the biomedical domain. Methods (San Diego, Calif.). 74. 10.1016/j.ymeth.2015.01.015.

[2] Mishra, Rashmi & Bian, Jiantao & Fiszman, Marcelo & Weir, Charlene & Jonnalagadda, Siddhartha & Mostafa, Javed & Del Fiol, Guilherme. (2014). Text Summarization in the Biomedical Domain: A Systematic Review of Recent Research. Journal of biomedical informatics. 52. 10.1016/j.jbi.2014.06.009.

[3] Moradi, Milad & Ghadiri, Nasser. (2019). Text Summarization in the Biomedical Domain.https://doi.org/10.48550/arXiv.1908.02285

[4] Gupta, Vishal & Lehal, Gurpreet. (2010). A Survey of Text Summarization Extractive Techniques. Journal of Emerging Technologies in Web Intelligence. 2. 10.4304/jetwi.2.3.258-268.

[5] Moratanch, N. & Gopalan, Chitrakala. (2016). A survey on abstractive text summarization. 1-7. 10.1109/ICCPCT.2016.7530193.

[6] Summarization from medical documents: a survey by Stergos Afantenos, Vangelis Karkaletsis, Panagiotis Stamatopoulos,Artificial Intelligence in Medicine, Volume 33, Issue 2, February 2005, Pages 157-177,https://doi.org/10.1016/j.artmed.2004.07.017

[7] An exploratory study of automatic text summarization in biomedical and healthcare domain by Mukesh Kumar Rohil, Varun Magotra, https://doi.org/10.1016/j.health.2022.100058

[8] Abstractive Summarization of Long Medical Documents with Transformers by Luciano Gonzalez, Sabrina Rong Lu, William Blake Buchanan, CS224n: Natural Language Processing with Deep Learning, Stanford / Winter 2021

[9] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

[10] Graph-based Abstractive Biomedical Text Summarization by Azadeh Givchi, Reza Ramezani, Ahmad Baraani.DOI: 10.1016/j.jbi.2022.104099

[11] SUMPUBMED: Summarization Dataset of PubMed Scientific Article,Gupta, Vivek and Bharti, Prerna and Nokhiz, Pegah and Karnick, Harish. In "Proceedings of the 2021 Conference of the Association for Computational Linguistics: Student Research Workshop", Association for Computational Linguistics, 2021.10.18653/v1/2021.acl-srw.30

[12] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer by Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. https://doi.org/10.48550/arXiv.1910.10683

[13] BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension by Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer. https://doi.org/10.48550/arXiv.1910.13461

[14] PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization by Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu. https://doi.org/10.48550/arXiv.1912.08777

[15] ROUGE: A Package for Automatic Evaluation of summaries by Chin-Yew Lin. Lin, Chin-Yew. 2004a. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop