Abstractive Approaches To Multidocument Summarization Of Medical Literature Reviews

Rahul Tangsali *
rahuul2001@gmail.com

Aditya Vyawahare *
aditya.vyawahare07@gmail.com

Aditya Mandke †
amandke@ucsd.edu

Onkar Litake †
olitake@ucsd.edu

Dipali Kadam ‡
ddkadam@pict.edu

Pune Institute of Computer Technology, India

Abstract

Text summarization has been a trending domain of research in NLP in the past few decades. The medical domain is no exception to the same. Medical documents often contain a lot of jargon pertaining to certain domains, and performing an abstractive summarization on the same remains a challenge. This paper presents a summary of the findings that we obtained based on the shared task of Multidocument Summarization for Literature Review (MSLR). We stood fourth in the leaderboards for evaluation on the MS^2 and Cochrane datasets. We finetuned pre-trained models such as BART-large, DistilBART and T5-base on both these datasets. These models' accuracy was later tested with a part of the same dataset using ROUGE scores as the evaluation metrics.

1 Introduction

The last few decades have witnessed a wide range of research applications in the field of natural language processing, especially text summarization. Text summarization has been applied in a number of domains including healthcare and medicine. With the tremendous amounts of big data getting generated in the medical industry each day, there is a need realized for effective techniques to summarize the data for further purposes. With the exponential rise in data getting accumulated in hospital databases and medical research labs, the need is increasing correspondingly. Text summarization in the healthcare domain has enabled far-reaching benefits for medical professionals. Effective summarization techniques help researchers and other individuals to parse long documents effectively, and gain valuable insights in shorter time periods.

The history of text summarization in NLP dates back to 1958, when the first paper on text summarization was published. Since then, its incorporation in healthcare has been widely done. Text mining and NLP methods have played an essential role in developing automatic text processing tools (Fleuren and Alkema, 2015). Automatic text summarization, thus proves to be an effective means of gaining valuable information from large documents and reports. In the medical domain, many approaches have been proposed for effective document summarization(Mishra et al., 2014) (Moradi and Ghadiri, 2019). Subfields in the biomedical domain where summarization is used include medical literature(Moradi and Ghadiri, 2016), evidence-based medical care (Fiszman et al., 2009), clinical notes(Moen et al., 2016), and drug information extraction(Fiszman et al., 2006).

Summarization approaches are broadly classified as abstractive and extractive. In extractive summarization(Gupta and Lehal, 2010), important sentences from the text are directly extracted and put into the summary, whereas for abstractive summarization(Moratanch and Chitrakala, 2016), new sentences depicting the summary of the topic are formed. Summarization approaches based on the number of documents can be classified as single document and multi-document(more than one documents are searched). In this paper, we present our findings obtained from performing multi-document summarization on the MS^2(DeYoung et al., 2021a) and Cochrane(Wallace et al., 2020a) datasets.

We finetune a few models on the MS^2 and Cochrane datasets, and research upon the best possible hyperparameters that could give us good results. We experimented with the BART-large model (Lewis et al., 2020) provided by Facebook AI on HuggingFace, the CNN version of the DistilBART model (Shleifer and Rush, 2020), and T5-base model (Raffel et al., 2020a) for text summarization. We preprocessed the inherently messy data provided, and generated summariza-

* equal contribution
† equal contribution
‡ equal contribution

Proceedings of the Third Workshop on Scholarly Document Processing, pages 199–203
October 12–17, 2022.

MS^2 (Provided Dataset)	Total input studies	Target summaries
Train	323608	14191
Validation	49002	2021
Test	42723	-
Cochrane (Provided Dataset)	**Total input studies**	**Target summaries**
Train	40497	3752
Validation	5033	470
Test	5678	-

Table 1: Statistics of the dataset used for training

tions on the same. We have experimented and compared the results of the aforementioned models. The datasets were provided by AllenAI. We have used the ROUGE evaluation metric (Lin, 2004) for comparing summarization accuracies.

2 Dataset Description

2.1 MS^2 (Multi-Document Summarization of Medical Studies)

The MS^2 (Multi-Document Summarization of Medical Studies) dataset (DeYoung et al., 2021b) is derived from documents and summaries from systematic literature reviews constructed from the papers in the Semantic Scholar literature Corpus (Ammar et al., 2018). Systematic literature reviews are a type of biomedical paper that compiles results from many different studies. The MS^2 dataset uses clustering before splitting into train, validation and test to avoid the learning of the test data during training. For each review, sentences were classified into 2 categories: Target sentences which contained information about the findings or summary of the paper and background sentences which described the research question. The statistics of the data provided are given in Table 1.

2.2 Cochrane Dataset

The Cochrane dataset (Wallace et al., 2020b) consists of the systematic reviews, created by the Cochrane collaboration, along with the title and abstract of the trials summarized by these reviews. The reviews summarized about 10 trials on average. The abstracts of the systematic reviews contained an average length of 75 words. The dataset statistics provided by the organizers are given in Table 1.

3 Data Preparation

The MS^2 and Cochrane datasets were provided to us in the CSV format. The input dataset consisted of the following columns: "ReviewID", "PMID", "Title" and "Abstract", whereas the target dataset consisted of the following columns: "ReviewID" and "Target". For the MS^2 dataset, additional 'Reviews-Info' files were included, which consisted of background information associated with the review. However, we didn't utilize them for training purposes.

In data preprocessing, the reviews present in the MS^2 and Cochrane datasets contain unnecessary delimiters and redundant line breakers, which made it necessary to clean them, before they could be passed to the model. We used simple Pandas preprocessing(Mckinney, 2011) on the CSV files, and cleaned these reviews into simple plain text which could be passed to the model.

We mapped each of the documents corresponding to a particular review ID, to the corresponding target summary in the target dataset, thus establishing a many-to-one relationship between the abstracts and the targets. We then removed all the other columns which were unnecessary for summarization ("Background", "Title", etc). Newly formed dataframes, consisting of the source texts (multiple documents merged together for each review ID) and the target text (target summaries) were formed and passed for preprocessing.

We used the pretrained BART-base tokenizer provided by Facebook AI for the BART-large and DistilBART models, whereas for the T5-base model training, the t5-base tokenizer was used. Both of these tokenizers are available open-source on the HuggingFace[1] model hub.

4 Experiments

4.1 Training Details

For training the models we used the Simple Transformers [2] library, an API used for transformer mod-

[1]https://huggingface.co
[2]https://simpletransformers.ai/

System/Model	rougeL	rouge1	rouge2	RougeLsum
facebook/bart-large	0.1449	0.2139	0.0349	0.172
sshleifer/distilbart-cnn-12-6	0.1377	0.2082	0.0298	0.1347
t5-base	0.1139	0.1762	0.1830	0.1179

Table 2: Scores recorded on the MS^2 dataset.

System/Model	rougeL	rouge1	rouge2	RougeLsum
facebook/bart-large	0.1751	0.2638	0.0576	0.1775
sshleifer/distilbart-cnn-12-6	0.1821	0.2898	0.0503	0.1820
t5-base	0.1549	0.2278	0.0319	0.1549

Table 3: Scores recorded on the Cochrane dataset.

els (Vaswani et al., 2017), which provides built-in support for various natural language processing tasks including text summarization.

We trained our models on the Nvidia K80 GPU which has a GPU RAM of 15 gigabytes. CUDA was utilized for effective computing, and making the training and evaluation processes faster. All the models were trained on 10 epochs, with training and validation losses measured over time for each epoch.

We trained the BART-large and the DistilBART-CNN models on the datasets, by instantiating Seq2Seq models (Sutskever et al., 2014) and arguments provided by Simple Transformers. We later modified some of the arguments by making the maximum length for each sequence equal to 140. Due to limited RAM available on the CUDA used, we faced memory errors. Hence, after each epoch, the weights directory was overwritten for memory availability. Maximum sequence length for the tokenized sequences of each input document was set to 512. For T5 (Text-To-Text Transfer Transformer), we used the t5-base models (Raffel et al., 2020b), after providing t5-base tokenization, and trained them with the same aforementioned hyperparameters.

All the above mentioned hyperparameters were giving the best possible results, and hence we proceeded with the use of the same. We finetuned the basic configurations specified in the Fairseq documentation. [3]

4.2 Evaluation Metrics

ROUGE Score (Lin, 2004), which stands for Recall-Oriented Understudy for Gisting Evaluation, was used as the evaluation metric. To calculate the rouge score we used the rouge metric provided by HuggingFace library [4]. We recorded rouge1, rouge2, rougeL and RougeLsum scores for our summaries. Rouge1 measured the overlap of unigram between the candidate and the reference summaries whereas rouge2 compared the bigram similarities between the summaries. RougeL and RougeLsum measured the Longest Common Subsequence (LCS)(Lin and Och, 2004) words between predicted and target summaries. All the Rouge scores recorded are scored out of 1; where, closer to 1 means more accurate summaries.

5 Results

For the results please refer to Table 2 and Table 3. The table contains different models which we tried for the summarization task and the ROUGE recorded on those models. For the submission of the summarization task on both datasets, we used the BART-base tokenizer and trained BART-large model provided by Facebook AI.

6 Competition Results

We obtained high rouge1 and deltaEi-macrof1 scores for the multi-document summarization task on the Cochrane dataset. We stood 5th when ranked according to rougeL metric.

For the MS^2 data summarization subtask, we stood 4th when ranked according to the rougeL metric. We attained high delta EI-avg scores for the summarization subtask.

The scores obtained in the MSLR MS^2 and Cochrane subtask are given in Table 4

[3]https://fairseq.readthedocs.io/en/latest/index.html

[4]https://huggingface.co/spaces/evaluate-metric/rouge

MSLR Subtask	rougeL	rouge1	rouge2	BERTScore	DeltaEI-avg	DeltaEI-macrof1
MS^2	0.1439	0.2060	0.0350	0.8479	0.5319	0.3558
Cochrane	0.1725	0.2468	0.0545	0.8591	0.2707	0.3789

Table 4: Rouge and BERT scores of the summarizations submitted to MSLR MS^2 and Cochrane Subtasks.

7 Conclusion

Thus, we implemented multi-document summarization of different clinical studies and their literature surveys in the medical field. We implemented various architectures and analysed their performance. Finally, we evaluated the models using ROUGE metric. We plan to explore other models and tokenization methods to provide more accurate summarizations. Also, we plan to train the models on different medical survey datasets for better results in our summarizations.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021a. Ms2: Multi-document summarization of medical studies.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021b. Ms^2: Multi-document summarization of medical studies. In *EMNLP*.

Marcelo Fiszman, Dina Demner-Fushman, Halil Kilicoglu, and Thomas C. Rindflesch. 2009. Automatic summarization of medline citations for evidence-based medical treatment: A topic-oriented evaluation. *J. of Biomedical Informatics*, 42(5):801–813.

Marcelo Fiszman, Thomas C Rindflesch, and Halil Kilicoglu. 2006. Summarizing drug information in medline citations. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, page 254—258.

Wilco W. M. Fleuren and Wynand Alkema. 2015. Application of text mining in the biomedical domain. *Methods*, 74:97–106.

Vishal Gupta and Gurpreet Lehal. 2010. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Wes Mckinney. 2011. pandas: a foundational python library for data analysis and statistics. *Python High Performance Science Computer*.

Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. 2014. Text summarization in the biomedical domain: a systematic review of recent research. *Journal of biomedical informatics*, 52:457—467.

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37.

Milad Moradi and Nasser Ghadiri. 2016. Different approaches for identifying important concepts in probabilistic biomedical text summarization.

Milad Moradi and Nasser Ghadiri. 2019. Text summarization in the biomedical domain. *ArXiv*, abs/1908.02285.

N. Moratanch and S. Chitrakala. 2016. A survey on abstractive text summarization. In *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pages 1–7.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sam Shleifer and Alexander M. Rush. 2020. Pre-trained summarization distillation.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Byron C. Wallace, Sayantan Saha, Frank Soboczenski, and Iain J. Marshall. 2020a. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization.

Byron C. Wallace, Sayantani Saha, Frank Soboczenski, and Iain James Marshall. 2020b. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Annual Symposium*, abs/2008.11293.