An Extractive-Abstractive Approach for Multi-document Summarization of Scientific Articles for Literature Review

Kartik Shinde, Trinita Roy, Tirthankar Ghosal
SciSpace, US
(kartik,trinita,tirthankar)@typeset.io

Abstract

Research in the biomedical domain is constantly challenged by its large amount of ever-evolving textual information. Biomedical researchers are usually required to conduct a literature review before any medical intervention to assess the effectiveness of the concerned research. However, the process is time-consuming, and therefore, automation to some extent would help reduce the accompanying information overload. Multi-document summarization of scientific articles for literature reviews is one approximation of such automation. Here in this paper, we describe our pipelined approach for the aforementioned task. We design a BERT-based extractive method followed by a BigBird PEGASUS-based abstractive pipeline for generating literature review summaries from the abstracts of biomedical trial reports as part of the Multi-document Summarization for Literature Review (MSLR) shared task[1] in the Scholarly Document Processing (SDP) workshop 2022[2]. Our proposed model achieves the *best performance* on the MSLR-Cochrane leaderboard[3] on majority of the evaluation metrics. Human scrutiny of our automatically generated summaries indicates that our approach is promising to yield readable multi-article summaries for conducting such literature reviews.

1 Introduction

The effectiveness of medical treatments following medical diagnosis can have both acknowledgments and contradictions with respect to various studies conducted. Prior to any medical treatment, evidence synthesis is essential to understand and stay up-to-date with medical advances from different clinical studies. A literature survey provides high-quality evidence for healthcare. However, such a task is very time-consuming if done manually.

To mitigate these issues high-quality largescale multi-document summarization datasets, e.g., The Cochrane Dataset (Wallace et al., 2021) and Multi-Document Summarization of Medical Studies (MS2) Dataset (DeYoung et al., 2021) were developed. Both the datasets consists of a wide variety of task-oriented summaries from clinical trials. To further encourage community research in multi-document summarization of biomedical reviews, the Allen Institute for Artificial Intelligence (or "AI2" for short) proposed a shared task named Multi-document Summarization for Literature Review (MSLR) 2022[4].

The MSLR shared task aims at summarizing and analyzing medical evidence from different clinical studies. The task consists of two datasets - Cochrane and MS2, which provide a brief narrative summary from the abstracts of different clinical studies communicating the main findings.

In this paper, we describe our system submission for the task. We participated in the Cochrane subtask. In our system submission, we design a pipelined approach leveraging state-of-the-art neural extractive and abstractive summarization models. Our system first extracts the vital information from the abstracts of all papers under a particular review ID and then generates an abstractive summary, with the help of pre-trained BigBird PEGASUS model (Zaheer et al., 2020), as the literature review test for that review ID.

2 Related Work

The concerned task in MSLR is a novel one and hence not much prior works were conducted except the papers that proposed the datasets. However, in this section we discuss some relevant recent works on multi-document summarization. Agarwal et al. (2011) propose an unsupervised method of using topic based clustering of fragments extracted from

[1] https://github.com/allenai/mslr-shared-task
[2] https://sdproc.org/2022/
[3] https://leaderboard.allenai.org/mslr-cochrane/submissions/public

Proceedings of the Third Workshop on Scholarly Document Processing, pages 204–209
October 12–17, 2022.

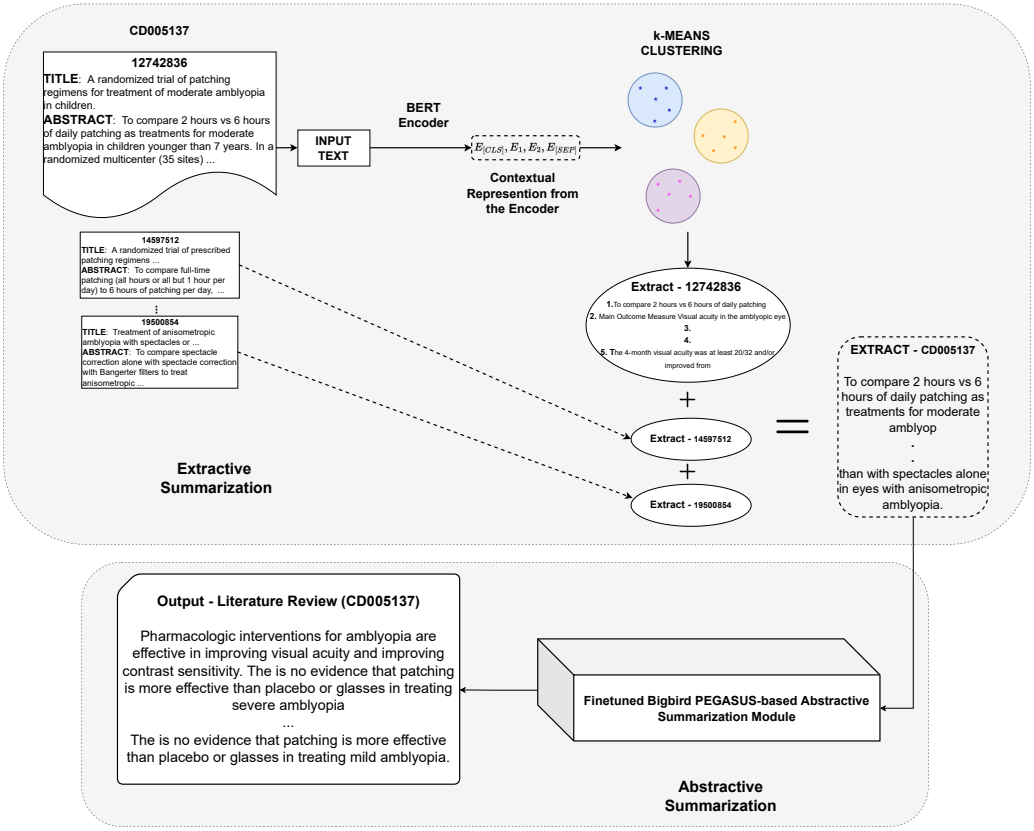

Figure 1: Workflow of the hybrid model - Original Abstracts from different PMIDs under a Review ID (*Top Left*), Combined Extractive summary being used as input for the abstractive summarizer (*Right*), Generated Output as a Literature Review Text (*Bottom*).

each co-cited article.

These fragments are ranked by relevance via a query generated from the context surrounding the co-cited list of papers. Multi-document summarization techniques can be broadly categorized into graph based (Mihalcea and Tarau, 2004; Meena et al., 2014; Hariharan and Srinivasan, 2009; Ge et al., 2011; Nguyen-Hoang et al., 2012), cluster based (Schlesinger et al., 2008; Meena et al., 2014; Gupta and Siddiqui, 2012) term frequency based (Salton, 1989; Fukumoto and SUGIMURA, 2004), context based (Sonawane et al., 2019), and latent semantic analysis based methods (Varma, 2019; Steinberger et al., 2004). Zakowski et al. (2004) describes a PICO (**P**opulation, **I**ntervention, **C**omparator and **O**utcome) framework for systematic review research. The study gives an account of the population that is being studies, what intervention was studied, what the intervention was compared to and what was the outcome. As an extension of PICO, DeYoung et al.; Fabbri et al. groups and identifies overall findings in reviews. However, multi document summarization needs expansion in the biomedical domain so as to reduce time and cost for addressing the delay in creating and updating reviews, thereby needing automation (DeYoung et al., 2021). Studies like (Marshall et al., 2016; Tsafnat et al., 2014) make an attempt at such automation tasks. Further, DeYoung et al. explore the use of Bi-directional and Auto-Regressive Transformers (Lewis et al., 2019) based approach on the MS2 Dataset. Pertaining to the peculiarities of the task, we formulate a hybird extractive-abstractive approach using a BERT-based extractive summarizer with K-means Clustering and a BigBird-PEGASUS based abstractive summarizer. Our system achieved the best performance among all the participating systems with a ROUGE-L score of 0.1969.

3 Dataset Description

The MSLR2022 shared task consists of two subtasks based on the Cochrane dataset (Wallace et al., 2021) and the MS2 dataset (DeYoung et al., 2021). In the Cochrane dataset there are approximately 4.5K systematic reviews of all trials relevant to a given clinical question, compiled by members of the Cochrane Collaboration. The dataset consists of the summarized systematic reviews along with the titles and abstracts of, on an average, 10 clinical trials each. The average length of the abstracts of

Model	ROUGE-L	ROUGE-1	ROUGE-2	BERTscore F1	Delta EI Avg. Divergence	Delta EI Macro F1
BERT+PEGASUS	**0.1969**	**0.2622**	0.0574	0.8590	0.2234	0.3011
ittc2	0.1837	0.2464	**0.0692**	**0.8762**	0.2195	0.3089
ittc1	0.1787	0.2413	0.0643	0.8729	0.2880	0.3375
BART	0.1760	0.2397	0.0671	0.8632	0.2081	0.3348
Longformer BART	0.1755	0.2387	0.0655	0.8641	0.2345	0.3316

Table 1: Evaluation Scores of different models in the MSLR2022 Cochrane Subtask.

the included trials is 245 words and the target summary is of the average length of 75 words. MS2 is a dataset containing 20K medical systematic reviews from approximately 470K studies collected from PubMed, created as an annotated subset of the Semantic Scholar research corpus. The MS2 dataset is much larger than the Cochrane dataset, but the latter contains cleaner data. For this shared task, the inputs and the target summaries are oriented in the same format which is then split into train, dev and test.

4 Methodology

Multi-document summarization aims to have a summary with maximum coverage and cohesiveness with less redundant data from the given set of papers pertaining to a topic. Sequence-to-sequence models do not perform well with large input sizes (Zaheer et al., 2020). Hence, we choose to leverage an extractive-abstractive summarization technique in our approach, to summarize biomedical reviews of correlated papers. In extractive summarization, we select a pre-decided number of statements from a given text as a relatively shorter representation of the entire text. We choose the Lecture Summarizer model in order to extract the most important sentences. This extraction is done by using a clustering algorithm on a set of embeddings, which are basically the contextual representations of sentences obtained from a BERT encoder. Hence, this also assists in maintaining some sort of coherence withing the input text.

We primarily use the provided abstracts as inputs to the extractive summarizer. For the titles that do not have any abstract, we use the titles as the inputs instead. We shorten these inputs to have at most five sentences from every different paper within a given Review_Id. We use BERT Extractive Summarizer (Miller, 2019), a model that performs extractive text summarization on lecture transcripts. We pass the abstracts separately to this model. The model first generates the contextual embeddings of the the input sentences. Further, the K-means clustering algorithm is used to find the k-sentences closest to the cluster's centroids. We proceed with the top 5 sentences from the cluster. The workflow of the model is provided in Figure 1.

For every Review_Id, we join the short extracts from different papers under that particular ID, and use the resulting sequence as the input sequence for training the abstractive summarization module. These shortened extracts, put together with the target summaries from original Cochrane dataset, give us a new data. We choose the BigBird PEGASUS model from (Zaheer et al., 2020), and finetune it on this newly obtained dataset. This model uses global attention and random attention on the input sequences apart from sparse-attention, which theoretically approximates to full attention. This sparse-attention mechanism can handle sequences of length up to 8x compared to what was possible prior to this and simultaneously reduces the quadratic dependency to linear, hence making the model suitable to learn using longer input sequences.

We finetune the model from the checkpoint '*google/bigbird-pegasus-large-pubmed*' using the newly created data for 6 epochs with a batch size of 4 and an initial learning rate of 2e-5 accompanied by FP16 precision training. The final output of the abstractive summarization module is the 'Related Works' text corresponding to the research topic aligned with a particular Review_Id. Figure 1 shows the workflow of our hybrid extractive-abstractive system.

5 Result and Analysis

The task realizes ROUGE (-1,2,L) (Lin, 2004), BERTScore F1 (Zhang et al., 2019), along with Delta EI Average Divergence and Macro F1 to be best suitable metrics for evaluation. Hence, to monitor the training, we use ROUGE as the basis of evaluation. Table 1 shows the comparison among all the participant teams on the Cochrane subtask where our best submission ranks first in ROUGE-L (0.1969) and ROUGE-1 (0.2622) scores.

ROUGE scores do not sufficiently measure the factual correctness of statements. Table 2 shows a

Review ID	Model Generated Summary
CD007066	There is some evidence that aliskiren 300 mg is superior to placebo in lowering blood pressure in patients with hypertension. The data are based on a single study and therefore we can not draw any conclusions about the relative efficacy of aliskiren 300 mg versus placebo.
CD005616	Devain disease is a common cause of pain in women of childbearing age. The evidence is limited and the use of cortisone injections in devain disease is not currently recommended.
CD007926	Menopausal hormone therapy is effective in the treatment of women with advanced or recurrent endometrial cancer. The is insufficient evidence to recommend the use of hormonal therapy alone or in combination with other hormonal agents.
CD002869	There is insufficient evidence to support or refute the effectiveness of any intervention to improve maternal and neonatal outcomes. The evidence is limited, and the results are not consistent across studies. The evidence is limited, and the results do not support the use of any intervention to improve maternal and neonatal outcomes.

Table 2: Example outputs of the hybrid model on the Cochrane dataset.

Review ID	Error outputs
CD004366	There is insufficient evidence to support the use of exercise as a treatment for depression. The is insufficient evidence to support the use of exercise as a treatment for depression. The is insufficient evidence to support the use of exercise as a treatment for depression.
CD010256	There is no evidence to support the use of aminophylline in the treatment of acute asthma. The is no evidence to support the use of salbutamol in the treatment of acute asthma. The is no evidence to support the use of aminophylline in the treatment of acute asthma.

Table 3: Observed erroraneous outputs from the model on the Cochrane dataset.

few instances with the review IDs and the generated literature review text. We can see that the generated text is coherent and does not contradict within itself. We observe that all the summaries were factually true and matched with the statements from input abstracts. Although the model generates better among other systems, a few issues still persist. Table 3 shows the most observed error case in the generation of model. We see that the model repeats the same statements multiple times. This might be attributed to the fact that a *Literature Review* OR *Related Works* section from a paper often consists of statements that are very coherent, and reinforce each other in order to establish an overall review of literature from a particular research topic. They highlight different findings, and more often than not, they have a similar gist.

For instances, consider a) "*We found only low quality evidence comparing ultra-radical and standard surgery in women with advanced ovarian cancer and carcinomatosis.*", b) "*It was unclear whether there were any differences in progression-free survival, QoL and morbidity between the two groups.*", and c) "*We are, therefore, unable to reach definite conclusions about the relative benefits and adverse effects of the two types of surgery.*". All these statements are very closely related in terms of the message they deliver. Hence, the finetuned summarizer does not account for facts, instead repeats the overall gist of the literature review.

6 Limitations

There are no ground truth summaries for lecture summarizer and therefore no metric for evaluating the outputs that we receive from the model. Due to the use of a clustering algorithm, the extractive part of out system is not readily trainable. We notice that the same model could not perform well in the subtask using the MS2 dataset. This can pertain to the long input sequences which is much greater than the Cochrane input sizes. Sequence-to-sequence models tend to not perform well with larger input sizes. Even if we shorten the input sequences, we would be losing out of essential information from the original data.

7 Conclusion

With the increasing rate of research and publications, literature reviews help keep track of the various advancements in the respective domains. Automation, although essential, also opens up new challenges including summarization over contradictory information present in different studies over a particular topic and summarization quality. Although our results show that our hybrid approach can be used for generating fluent high-quality literature review summaries, there is still significant scope for improvement. Additionally, ethical concern involving the factuality of the summaries also comes into play because deploying such a system without proper monitoring is speculative when it comes to such a high-impact domain as healthcare. This task helps us understand the challenges in multi-document summarization in the high-impact biomedical domain. The future scope of research can include trying real-world applications of such systems having proper evaluation and monitoring strategies to test the correctness of the summaries.

References

Nitin Agarwal, Ravi Shankar Reddy, GVR Kiran, and Carolyn Rose. 2011. Towards multi-document summarization of scientific articles: making interesting comparisons with scisumm. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, pages 8–15.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486*.

Jay DeYoung, Eric Lehman, Ben Nye, Iain J Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. *arXiv preprint arXiv:2005.04177*.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Jun-ichi Fukumoto and Tomoya SUGIMURA. 2004. Multi-document summarization using document set type classification. In *NTCIR*.

Shuzhi Sam Ge, Zhengchen Zhang, and Hongsheng He. 2011. Weighted graph model based sentence clustering and ranking for document summarization. In *The 4th International Conference on Interaction Sciences*, pages 90–95. IEEE.

Virendra Kumar Gupta and Tanveer J Siddiqui. 2012. Multi-document summarization using sentence clustering. In *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, pages 1–5. IEEE.

Shanmugasundaram Hariharan and Rengaramanujam Srinivasan. 2009. Studies on graph based approaches for single and multi document summarizations. *International Journal of Computer Theory and Engineering*, 1(5):1793–8201.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2016. Robotreviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201.

Yogesh Kumar Meena, Ashish Jain, and Dinesh Gopalani. 2014. Survey on graph and cluster based approaches in multi-document text summarization. In *International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014)*, pages 1–5. IEEE.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Tu-Anh Nguyen-Hoang, Khai Nguyen, and Quang-Vinh Tran. 2012. Tsgvi: a graph-based summarization system for vietnamese documents. *Journal of Ambient Intelligence and Humanized Computing*, 3(4):305–313.

Gerard Salton. 1989. Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*, 169.

Judith D Schlesinger, Dianne P O'leary, and John M Conroy. 2008. Arabic/english multi-document summarization with classy—the past and the future. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 568–581. Springer.

Sheetal Sonawane, Archana Ghotkar, and Sonam Hinge. 2019. Context-based multi-document summarization. In *Contemporary advances in innovative and applicable information technology*, pages 153–165. Springer.

Josef Steinberger, Karel Jezek, et al. 2004. Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4(93-100):8.

Guy Tsafnat, Paul Glasziou, Miew Keen Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. 2014. Systematic review automation technologies. *Systematic reviews*, 3(1):1–15.

Rashmi Varma. 2019. A hybrid approach for multi-document text summarization.

Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. 2021. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Laura Zakowski, Christine Seibert, and Wisconsin Selma VanEyck. 2004. Evidence-based medicine: answering questions of diagnosis. *Clinical medicine & research*, 2(1):63–69.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.