LHS712EE at BioLaySumm 2023: Using BART and LED to summarize biomedical research articles

Quancheng Liu,[1*] **Xiheng Ren,**[1*] and **V.G.Vinod Vydiswaran**[2]
[1]Department of Computational Medicine and Bioinformatics, University of Michigan, USA
[2]Department of Learning Health Sciences, University of Michigan, USA
[*]These authors contributed equally to this work.
{quanch, xhren, vgvinodv}@umich.edu

Abstract

As part of our participation in BioLaySumm 2023, we explored the use of large language models (LLMs) to automatically generate concise and readable summaries of biomedical research articles. We utilized pre-trained LLMs to fine-tune our summarization models on two provided datasets, and adapt them to the shared task within the constraints of training time and computational power. Our final models achieved very high relevance and factuality scores on the test set, and ranked among the top five models in the overall performance.

1 Introduction

Biomedical research articles are vital sources of information based on latest scholarly findings in the health domain, ranging from acute and chronic health diagnoses, population health, advances in cellular, molecular, and pharmaceutical technologies, and long-term impacts of recent global pandemics. However, due to the highly technical jargon and professional language used in these articles, it can be challenging for laypersons who didn't receive formal training in biomedical sciences to understand the contents. Summarization of biomedical text in layperson-friendly language can provide a solution by automatically generating concise and readable summaries of technical documents (Goldsack et al., 2023). These lay summaries can be used to communicate the key findings of a study to a broader audience, including patients, their care partners, and members of the general public.

Text summarization has been extensively studied by researchers in the past. There are two main categories of text summarization techniques – extractive summarization and abstractive summarization (Widyassari et al., 2022). Extractive summaries only include key phrases and sentences directly selected from the original text, while abstractive summaries consist of new, generated sentences that summarize the original content (Widyassari et al., 2022). Although abstractive summarization is more challenging and complex than extractive summarization, it has the potential to convey more information and better meet human needs (Widyassari et al., 2022). In recent years, most of the research on abstractive summarization was inspired by the encoder-decoder architecture of deep neural network models. This includes the use of pre-trained encoders proposed by Liu and Lapata (2019) and the development of multi-task encoder-decoder models, as suggested by Xu et al. (2020).

As part of our participation in the BioLaySumm 2023 shared tasks on lay summarization of biomedical research articles, we investigated the use of large language models (LLMs) to generate concise, relevant, factual, and easily readable summaries (Lee, 2023). We utilized existing LLMs, including Bidirectional and Auto-Regressive Transformer (BART) (Lewis et al., 2019) and Longformer Encoder-Decoder (LED) (Beltagy et al., 2020), to train our summarization model on the provided datasets. Despite having limited time and computational resources, we investigated several approaches to fine-tune the LLMs for this task. Further, we retrained our model over the abstract sections of articles in the PLOS dataset provided as a reference to generate the technical abstracts for the readability-controlled summarization task. Our final summarization model achieved strong performance in relevance and factuality, and was ranked among the top five models for lay summarization of biomedical literature.

2 Methods

2.1 Data Description

All registered participants in the BioLaySumm 2023 shared tasks were provided two datasets containing biomedical research articles and expert-

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 620–624

written lay summaries for Task 1, the Lay Summarization task. The first dataset contained 24,773 articles from the Public Library of Science (PLOS) for training and 1,376 for validation (Luo et al., 2022). The median length of the articles in the PLOS training dataset was 6,577 words. The second dataset, eLife, contained 4,346 articles for training and 241 for validation (Goldsack et al., 2022). The articles in eLife training dataset were longer, with the median length of the articles as 9,837.5. Task 2 on readability-controlled summarization utilized only the PLOS dataset.

2.2 BART Model Training for Lay Summarization (Task 1)

To generate lay summaries from long articles for Task 1, we selected models that follow the sequence-to-sequence (seq2seq) architecture. After a comprehensive review of existing literature on state-of-the-art seq2seq models, we chose the Bidirectional and Auto-Regressive Transformer (BART) model (Lewis et al., 2019). BART combines a bidirectional encoder and an autoregressive decoder. BART models have been shown to perform exceptionally well when fine-tuned for text generation tasks such as summarization and translation, as well as language comprehension tasks like text classification and question answering (Lewis et al., 2019). For this participation, we used `facebook/bart-large-xsum`, the BART model implementation variant that was pretrained by Facebook/Meta specifically targeting text summarization (Facebook, 2020). The model training and evaluation was conducted on one NVIDIA GeForce RTX 4090 GPU with the mamory capacity of 24GB.

To adapt BART for the specific task, we fine-tuned the model using the following parameters: we set the epochs to 3, number of beams to 4, maximum encoder length to 1024, maximum decoder length to 512, minimum decoder length to 100, length penalty to 2, learning rate to 1e-4, weight decay to 0.01, evaluation strategy as "steps", per-device train batch size as 4, and per-device evaluation batch size as 4. We employed several optimization techniques to fit the model into the 24GB memory limit. First, we used mixed-precision training with `fp16` (16-bit floating point) to reduce memory consumption and improve training speed. We leveraged NVIDIA's Apex library as our `fp16` backend, which simplified mixed-precision training and offered additional performance optimizations. We set the "gradient accumulation steps" parameter to 4, which allows us to accumulate gradients from multiple mini-batches before performing a single optimization step. This technique helped reduce memory usage by reducing the frequency of weight updates. Finally, we employed "gradient checkpointing", which trades computation time for memory by storing only a subset of intermediate values during the forward pass and recomputing them as needed during the backward pass. This technique further reduced memory usage, allowing the model to fit within the memory size limit of 24GB.

To summarize, we fine-tuned the BART model for generating lay summaries of both PLOS and eLife datasets, employing optimization techniques such as mixed-precision training, gradient accumulation, and gradient checkpointing to accommodate memory constraints. However, the BART model had a significant limitation that it could only handle input sequences up to 1024 tokens (Lewis et al., 2019). As a result, the model considered only the first 1024 tokens of each article and discarded any remaining tokens. This might lead to a loss of information needed for the optimal lay summary.

2.3 LED Model Training for Lay Summarization (Task 1)

To overcome the limitations of the BART model in generating lay summaries of biomedical research articles, we also investigated the effectiveness of Longformer models. Longformer (Beltagy et al., 2020) is a transformer-based model designed specifically to process long sequences. Unlike traditional transformer models that are limited to processing only short sequences because the computational complexity scales quadratic with sequence length, Longformer models introduce an attention mechanism that scales linearly with sequence length, enabling them to handle documents with thousands of tokens more efficiently (Beltagy et al., 2020). This is achieved through a combination of local windowed attention and task-motivated global attention. The Longformer's attention mechanism can be used as a drop-in replacement for standard self-attention in transformer models. Beltagy et al. (Beltagy et al., 2020) pre-trained Longformer models and fine-tuned them on various downstream tasks, and observed that they consistently outperformed RoBERTa on long document tasks, and set new state-of-the-art results on WikiHop and

TriviaQA datasets. WikiHop (Welbl et al., 2018) and TriviaQA (Joshi et al., 2017) are both reading comprehension datasets used to train and evaluate language models.

We further adapted Longformer models for the lay summarization task by using the Longformer Encoder-Decoder (LED) variant generated by `allenai/led-large-16384-arxiv`. LED is designed to support long text, generative, sequence-to-sequence tasks and has proven to be effective on the arXiv summarization dataset (Institute, 2023). We configured the input token size to 8192, which is 8 times larger that the BART model described in Section 2.2. We used similar techniques and settings as those described for the BART model above to optimize memory usage. Due to the memory limitation of our graphic card, we could only set the batch size for the LED model to 1. Further, we only trained the LED model on the PLOS dataset because of time constraints.

2.4 BART Model Training for Technical Abstract Generation (Task 2)

In Task 2, the goal was to generate both lay summaries and technical abstracts for the articles in the PLOS dataset. As Tasks 1 and 2 shared the same PLOS dataset, we set about to test the generality of our proposed models for Task 2 as well. First, we utilized the PLOS model trained in Task 1 as-is to generate both technical abstracts and lay summaries in Task 2. After reviewing the results of our initial submission, we developed a second model by retraining the model with the abstract sections of research articles in the PLOS dataset as a reference. Due to limited time availability, we only retrained the BART model for Task 2, since it is more computationally efficient compared to the LED model.

2.5 Summary of submitted runs

We selected the models for submission based on their performance in the validation phase. To evaluate each model, we gathered metrics including the training loss, validation loss, Rogue-1, Rogue-2, and Rogue-L with a min-max normalization (we made the loss negative for this normalization because the better model had the lower loss). Then, we assessed the validation performance by calculating a weighted sum of the following normalized scores:0.1*training loss, 0.3* validation loss, 0.2*Rogue-1, 0.2*Rogue-2, and 0.2*Rogue-L.

Aspect	Metric	Task 1	Baseline	Task 2	Baseline
Relevance	ROUGE-1	0.985	0.807	0.772	0.000
	ROUGE-2	0.872	0.709	0.569	0.000
	ROUGE-L	0.978	0.826	0.834	0.000
	BERTScore	0.889	0.856	1.000	1.000
	Rank	**1**	**9**	**2**	**4**
Readability	FKGL	0.521	0.264	0.619	1.000
	DCRS	0.631	0.640	1.000	0.954
	Rank	**18**	**8**	**3**	**4**
Factuality	BARTScore	0.906	1.000	0.857	1.000
	Rank	**2**	**1**	**2**	**1**

Table 1: Performance of the final system on Task 1 and 2 after min-max normalization

We submitted two runs for Task 1. The first submission was generated from the two fine-tuned BART models. For the second (final) submission, the lay summaries for the PLOS dataset were generated from a fine-tuned LED model and those for the eLife dataset were generated from a fine-tuned BART model. For Task 2, the first submission was generated directly from the PLOS BART model used in Task 1. For the second submission, we retrained the PLOS BART model with the abstract sections of articles in the PLOS dataset as a reference.

2.6 Evaluation Measures

The submissions for the shared tasks are evaluated on three aspects – relevance, readability, and factuality. Relevance is assessed using Rouge-1, Rouge-2, Rouge-L (Lin, 2004), and BERTScore (Zhang et al., 2019) metrics. Readability is evaluated using two measures: Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975) and Dale-Chall Readability Score (DCRS) (Dale and Chall, 1948). Factuality is measured using BARTScore that was fine-tuned by the organizers (Koh et al., 2022). Better systems have higher relevance and factuality scores and lower readability scores.

3 Results

Task 1 The performance of the two submissions for Task 1 were not significantly different from one another. Table 1 shows the scores from the second (final) submission for lay summarization on the three evaluation aspects after min-max normalization. The submitted run was ranked the best system on relevance and the second best system (following the baseline run) on factuality. Although the readability score was not as high, the overall score placed the model at fifth place among the 20 participants and the baseline.

		1^{st} Submission		2^{nd} Submission	
Aspect	Metric	lay summ	abstract	lay summ	abstract
Relevance	ROUGE-1	0.404	0.349	0.419	0.464
	ROUGE-2	0.104	0.084	0.117	0.143
	ROUGE-L	0.367	0.321	0.382	0.429
	BERTScore	0.849	0.822	0.854	0.855
	Rank			**Rank 2**	
Readablity	FKGL	2.194	3.402	2.369	2.158
	DCRS	0.945	1.835	0.870	1.002
	Rank			**Rank 3**	
Factuality	BARTScore	-0.681	-1.023	-0.973	-1.307
	Rank			**Rank 2**	

Table 2: Performance on Task 2, generating lay summaries and technical abstracts, after min-max normalization.

Task 2 While the final model performed very well on the lay summarization task (Task 1), it yielded relatively poor results in technical abstract generation compared to other submissions. The performance of our model significantly improved compared to the first submission (Table 2). After min-max normalization, the final scores for our second (final) submission for Task 2 in the three aspects are shown in Table 2. Our approach was ranked a joint first in Task 2.

4 Discussion

We leveraged the capabilities of two transformer based models – BART and Longformer – to create high-quality lay summaries or technical abstracts of biomedical research articles, and ensuring that our approach can handle the challenges associated with processing long, complex text. For Task 1, there was no significant improvement in performance after we changed the model from BART to LED. The LED models were expected to perform better because they can accept longer sentences, enhancing their ability to capture text features. However, due to time and memory constraints, we used fewer epochs and smaller batch sizes to train the LED model, which might have resulted in its similar performance to the BART model.

In Task 2, the model trained for lay summarization did not perform well in generating technical abstracts. After retraining the model on PLOS abstracts, the model was able to perform as well as the previous model. This implies that adjusting the parameters of the existing model alone does not change the readability of the generated text. On the other hand, retraining with the expected output as a reference can improve the model performance.

5 Conclusion

In our participation in the BioLaySumm 2023 shared tasks, we were able to successfully utilize pre-trained large language models – BART and LED – to generate lay summaries and technical abstracts. We did so by fine-tuning our summarization models on two different datasets, PLOS and eLife. Despite the limited time and computational resources, we were able to develop models that performed well in the relevance and factuality score of the summarizing tasks, and finally ranked as the fifth best models in the overall performance. The BioLaySumm tasks showed the potential of lay summarization models in making biomedical research accessible to a broader audience. We believe that the development of these models will continue to play a critical role in advancing healthcare and empowering individuals to make informed decisions about their health.

Limitations

Despite our efforts to enhance the efficiency and minimize the memory cost of our models, large language models still demand considerable time and memory resources, which remains a limitation of our work. Given sufficient time and computational resources, we could explore the possibility of increasing the batch sizes and running additional epochs to further optimize the model's performance. While our final system excelled in relevance and factuality aspects, it was relatively poor on readability, which represents a potential area for improvement. To improve the robustness of the model, given additional time, we would use a combination of machine learning and list-based approaches to identify arcane words and technical terms and substitute them with their easy-to-understand synonyms. With these procedures, we believe that we can make our summaries more readable.

Ethics Statement

Large Language Models (LLMs), including BART and LED, can implicitly learn biases from their training dataset. In the biomedical fields, these biases potentially include exclusion of certain groups of people who are underrepresented or misrepresented in the training data. It is important to be aware of this potential bias. Moreover, LLMs are not always accurate and reliable. Inaccuracies in

the generated summaries from LLMs could have serious consequences and impact on health and well-being of persons who trust the automated summaries of biomedical research articles generated by LLMs.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. ArXiv:2004.05150 [cs].

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Facebook. 2020. facebook/bart-large-xsum · Hugging Face.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Allen Institute. 2023. allenai/led-large-16384-arxiv · Hugging Face.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How Far are We from Robust Long Abstractive Summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682–2698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Angie Lee. 2023. What Are Large Language Models and Why Are They Important?

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:1910.13461 [cs, stat].

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability Controllable Biomedical Document Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, et al. 2022. Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1029–1046.

Weiran Xu, Chenliang Li, Minghao Lee, and Chi Zhang. 2020. Multi-task learning for abstractive text summarization with key information guide network. *EURASIP Journal on Advances in Signal Processing*, 2020(1):1–11.

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