Eulerian at BioLaySumm: Preprocessing Over Abstract is All You Need

Satyam Modi*
Indian Institute of Technology, Delhi
smodi50448@gmail.com

T Karthikeyan*
Indian Institute of Technology, Delhi
tkarthikeyanai@gmail.com

Abstract

In this paper, we present our approach to the BioLaySumm 2024 Shared Task on Lay Summarization of Biomedical Research Articles at BioNLP workshop 2024 (Goldsack et al., 2024). The task aims to generate lay summaries from the abstract and main texts of biomedical research articles, making them understandable to lay audiences. We used some preprocessing techniques and finetuned Flan-T5 models for the summarization task. Our method achieved an AlignScore of 0.9914 and a SummaC metric score of 0.944. Notably, we scored the highest on the Factuality metric, composed of AlignScore and SummaC, among all the teams.

1 Introduction

Research in every domain has increased significantly, making it challenging for cross-domain researchers to keep track of terminologies outside their expertise. Providing layman summarization in biomedical research addresses this issue. This task is particularly important given the growing volume of biomedical literature, which makes manual summarization impractical. Automated lay summarization can significantly enhance the reach and impact of scientific findings by making them accessible to a wider audience, including patients, healthcare providers, policymakers, and the general public.

The BioLaySumm 2024 Shared Task on Lay Summarization of Biomedical Research Articles is designed to advance the development of automated systems capable of generating accurate and coherent lay summaries from biomedical articles. This task utilizes two separate datasets, focusing on generating summaries that maintain the essence and factuality of the original research while being understandable to a lay audience.

*These authors contributed equally to this work.

2 Related Work

Past works in summarisation has been along two directions: extractive summarisation and abstractive summarisation. Extractive summarisation involves selecting and extracting key phrases, sentences, or segments directly from the original text to create a summary while in abstractive summarisation the summary is generated by creating new sentences that convey the key information from the original text. Recent works like PEGASUS (Zhang et al., 2020a) uses transformer like models with a self supervised objective for summarisation. In recent years, most of the work on abstractive summarisation has been based on treating the task as a sequence-to-sequence task and using pretrained encoders (Liu and Lapata, 2019).

In this work, we explore on the usage of LLMs for biomedical articles summarisation. Specifically, we use Flan-T5 model (Chung et al., 2022) for finetuning it for our use case by treating summarisation as a sequence-to-sequence task.

3 Datasets

The task included two datasets, PLOS and eLife (Goldsack et al., 2022). PLOS is the larger dataset derived from Public Library of Science, comprising 24,773 instances for training and 1,376 for validation whereas the eLife dataset was derived from the peer-reviewed eLife journal and it contains 4,346 instances for training and 241 for validation. The test data used for evaluation consisted of 142 articles each of PLOS and elife datasets.

4 Methodology

4.1 PoA(Preprocessing over Abstract)

The PoA(Preprocessing over Abstract) involves extracting the initial sentences from the research paper which mainly comprises of the abstract and provide a concise overview of the study. Then we

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 826–830

apply a regular expression to remove content with parentheses, braces and brackets. These segments often contain supplementary details that can be omitted for a lay audience. This preprocessing step aims to improve readability without compromising the core information.

4.2 Finetuning Flan T5 Models

In our experiments, we fine-tuned various versions of the Flan-T5 model to enhance their performance in summarizing biomedical research articles. Input was the preprocessed abstract obtained from the PoA technique(Section: 4.1) and output was the summary provided in the training data. We began with the Flan-T5 small model, initially fine-tuning it on the PLOS dataset alone.

Next, we expanded the training data to include both PLOS and eLife articles, aiming to improve the model's generalization and robustness. By incorporating a larger and more diverse dataset, we hypothesized that the model would generate more accurate and comprehensive summaries.

We then progressed to fine-tuning the Flan-T5 base model, also using the combined PLOS and eLife datasets. The base model, being larger and more complex than the small model, was expected to capture more intricate patterns and dependencies in the data.

In our final experiment, we applied a cosine scheduler during the fine-tuning of the Flan-T5 base model with the combined datasets. The cosine scheduler adjusts the learning rate dynamically, aiming to improve convergence and model performance by reducing the learning rate gradually, which helps in avoiding overfitting and ensuring better generalization.

5 Experiments and Results

5.1 Hyperparameters for reproducibility

All experiments utilized a batch size of 25, a max input token length of 512, and a max output token length of 300. The learning rate was set to 1e-3. These experiments were conducted on a single NVIDIA A100 40GB GPU for 25 epochs. The code[1] used in this research is publicly accessible.

[1]Available at https://github.com/tkarthikeyan132/PoA

5.2 Evaluation Metrics

The submission was evaluated across three dimensions: relevance, readability, and factuality. Relevance is measured through metrics including Rouge-1, Rouge-2, Rouge-L (Lin, 2004) and BERTScore (Zhang et al., 2020b). Readability is assessed via the Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), CLI (Coleman Liau Index) , Dale-Chall Readability Score (DCRS) (Dale and Chall, 1948) and LENS (Maddela et al., 2023). Factuality is measured utilizing AlignScore (Zha et al., 2023) and SummaC (Laban et al., 2021). The scores calculated for each metric are the average of those calculated independently for the generated lay summaries of PLOS and eLife. The aim is to have higher relevance and factuality scores. All the readability scores must be low except the LENS metric.

5.3 Main Results

The evaluation of various Flan-T5 models and the PoA technique yielded several notable observations, which are summarized below:

5.3.1 Flan-T5 Small: PLOS vs. PLOS + eLife Data

When comparing the Flan-T5 small model trained on PLOS data alone to the same model trained on combined PLOS and eLife data, it was observed that the latter configuration was beneficial across all ROUGE scores and readability metrics, indicating better performance in capturing relevant content and readability. However, this enhancement comes at the cost of factuality metrics, as demonstrated by a decrease in AlignScore and SummaC values.

5.3.2 Flan-T5 Small vs. Flan-T5 Base: Combined Data

Comparing the Flan-T5 small and Flan-T5 base models, both trained on the combined PLOS and eLife datasets, revealed that the base model exhibited superior performance in almost all the relevance and readability metrics, with the exception of the DCRS metric, which did not show improvement. Despite these gains, the factuality metrics (AlignScore and SummaC) were compromised in the Flan-T5 base model compared to the small model.

		Relevance				Readability				Factuality	
Model	**Training data**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore**	**FKGL**	**DCRS**	**CLI**	**LENS**	**AlignScore**	**SummaC**
PoA	N/A	**0.4302**	**0.1327**	**0.3965**	**0.8571**	15.5542	11.1486	17.2919	37.4521	**0.9914**	**0.944**
Flan-T5 small	PLOS	0.3935	0.1152	0.3589	0.8479	14.832	11.3634	16.8313	48.7148	0.9369	0.8732
Flan-T5 small	PLOS + eLife	0.4035	0.1166	0.371	0.8451	14.7954	10.7561	16.5336	48.4619	0.9173	0.8538
Flan-T5 base	PLOS + eLife	0.4228	0.1255	0.3879	0.8511	**14.2915**	10.7817	**16.1177**	52.1659	0.8858	0.8024
Flan-T5 base	PLOS + eLife	0.4277	0.1297	0.3942	0.8501	15.0451	**10.6537**	16.6125	**52.3009**	0.9122	0.8385

Table 1: Inference Results of Flan-T5 Models

5.3.3 Flan-T5 Base: With vs. Without Cosine Scheduler

When analyzing the impact of incorporating a cosine learning rate scheduler in the training of the Flan-T5 base model with combined data, it was evident that the scheduler contributed to better readability and factuality metrics. Improvements were noted in DCRS and LENS, while FKGL and CLI metrics became little worse, which are also indicators of readability, were slightly compromised. This suggests that the scheduler helps in fine-tuning the model to better balance readability and factual accuracy.

5.3.4 PoA Technique Performance

Interestingly, the PoA (Preprocessing over Abstract) technique, which does not involve any training, outperformed all Flan-T5 models in terms of relevance and factuality metrics. This technique's performance in ROUGE scores and factuality assessments (AlignScore and SummaC) was superior, highlighting its effectiveness in generating concise and accurate summaries directly from the abstracts. However, the readability scores were lower, likely because abstracts are inherently complex and may not be easily readable by a lay audience.

These findings are detailed in Table 1 illustrating the performance metrics across different models and configurations.

6 Conclusion

The comparative analysis of various Flan-T5 models and the PoA technique for summarizing biomedical research articles has yielded insightful findings. The Flan-T5 small model showed enhanced relevance and readability metrics when trained on combined PLOS and eLife datasets, though at the expense of factuality. The Flan-T5 base model further improved relevance and readability metrics but also compromised factuality. Introducing a cosine learning rate scheduler to the Flan-T5 base model improved readability and factuality metrics, indicating a better balance in model performance.

Notably, the PoA technique, despite not involving any training, outperformed all Flan-T5 models in relevance and factuality metrics, demonstrating its effectiveness in generating accurate and concise summaries from abstracts. These results underscore the importance of training strategies in developing effective summarization models, while also highlighting the potential of simple preprocessing techniques like PoA.

7 Future Scope

The future scope of this research includes augmenting the training datasets to encompass a broader range of biomedical text per article, thereby enhancing the model's generalizability across diverse terminologies and styles. Advanced fine-tuning techniques such as mixed precision training and curriculum learning could be explored to further improve performance in relevance, readability, and factuality. Tailoring models for specific sub-domains within biomedical research could improve accuracy and relevance for specialized fields. Moreover, creating comprehensive evaluation frameworks that consider user satisfaction and practical utility alongside traditional metrics will be essential. Addressing these avenues can significantly advance the effectiveness and applicability of summarization models for biomedical research articles.

8 Limitations

Although we experimented with text-to-text models like Flan-T5, extending our research to autoregressive large language models such as LLaMA 3(AI@Meta, 2024) could offer different advantages and improvements in summarization tasks.
Our experiments focused on preprocessing techniques and hyperparameter tuning, but the potential of prompt tuning with advanced models like GPT-4(et al., 2023) and Gemini(Team et al., 2023)

remains unexplored. Investigating prompt engineering and tuning could enhance summarization performance.

Additionally, we combined eLife and PLOS datasets to train a single model, which may not capture the nuances of each dataset. Training separate models for each dataset could yield more specialized and effective summarization capabilities.

Furthermore, our proposed technique might be more effective when integrated into a more complex pipeline to refine the generated summaries. Future research should address these areas to enhance the robustness and applicability of summarization models.

9 Acknowledgements

The authors thank IIT Delhi HPC facility for computational resources.

A Experiments with Various Schedulers

We finetuned the Flan-T5 base model with three distinct schedulers: Cosine, Step, and Exponential. The goal was to determine the impact of each scheduler on the model's performance across multiple metrics. In Table 2, Our experiments demonstrate that the choice of scheduler can significantly impact the performance of the Flan-T5 model in terms of relevance, readability, and factuality. The Cosine scheduler performed best overall in relevance metrics, while the Step scheduler excelled in readability, and the Exponential scheduler achieved the highest factuality scores.

B Experiments with Various Learning Rates

In Table 3, We present the results of experiments conducted to evaluate the performance of the Flan-T5 base model with different learning rates. The learning rates tested in these experiments were 1e-3, 1e-4, 5e-4, and 1e-5. The learning rate of 1e-3 generally provided the best balance across relevance and readability metrics, while the learning rate of 1e-5 excelled in factuality.

C Experiments with and without Preprocessing over Abstract (PoA)

In Table 4, the experiments demonstrate that the PoA method has a nuanced impact on the performance of the Flan-T5 base model. While it slightly reduced some relevance metrics, it improved the depth of content coverage and significantly enhanced factual accuracy. The readability metrics presented mixed results, indicating that the preprocessing step altered the text complexity and structure. These findings suggest that while the PoA method can enhance certain aspects of summarization, it may need to be combined with other techniques for optimal performance across all metrics.

References

AI@Meta. 2024. Llama 3 model card.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

OpenAI Josh et al. 2023. Gpt-4 technical report.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the biolaysumm 2024 shared task on the lay summarization of biomedical research articles. In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand. 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.

Peter Kincaid, Robert P. Fishburne, 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.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Preprint*, arXiv:2111.09525.

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.

		Relevance				Readability				Factuality	
Model	**Scheduler**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore**	**FKGL**	**DCRS**	**CLI**	**LENS**	**AlignScore**	**SummaC**
Flan-T5 base	Cosine	**0.4277**	**0.1297**	**0.3942**	**0.8501**	15.0451	10.6537	16.6125	**52.3009**	0.9122	0.8385
Flan-T5 base	Step	0.4161	0.1233	0.3815	0.8495	**14.7222**	10.9538	**16.5340**	49.3804	0.9148	0.8417
Flan-T5 base	Exponential	0.3571	0.0914	0.3332	0.8252	15.3144	**8.4444**	16.7020	40.3914	**0.9294**	**0.8509**

Table 2: Inference Results of Flan-T5 Models with Various Schedulers

		Relevance				Readability				Factuality	
Model	**Learning rate**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore**	**FKGL**	**DCRS**	**CLI**	**LENS**	**AlignScore**	**SummaC**
Flan-T5 base	1e-3	**0.4277**	**0.1297**	**0.3942**	**0.8501**	15.0451	**10.6537**	16.6125	**52.3009**	0.9122	0.8385
Flan-T5 base	1e-4	0.4099	0.1205	0.3766	0.8474	14.7894	10.9620	16.5964	48.0036	0.9294	0.8642
Flan-T5 base	5e-4	0.4172	0.1231	0.3833	0.8497	**14.5144**	10.8936	**16.3596**	49.7981	0.9052	0.8308
Flan-T5 base	1e-5	0.4114	0.1189	0.3784	0.8458	15.1296	10.8945	16.8266	49.3234	**0.9388**	**0.8787**

Table 3: Inference Results of Flan-T5 Models with Various Learning Rates

	Relevance				Readability				Factuality	
PoA (Preprocessing over Abstract)	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore**	**FKGL**	**DCRS**	**CLI**	**LENS**	**AlignScore**	**SummaC**
No	**0.4323**	0.1315	**0.3989**	0.8484	**14.8658**	11.1500	**16.5704**	**50.4200**	0.9486	0.9204
Yes	0.4302	**0.1327**	0.3965	**0.8571**	15.5542	**11.1486**	17.2919	37.4521	**0.9914**	**0.944**

Table 4: Comparison of Performance with and without POA Method

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

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *Preprint*, arXiv:2305.16739.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *Preprint*, arXiv:1912.08777.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.