HULAT-UC3M at BiolaySumm: Adaptation of BioBART and Longformer models to summarizing biomedical documents

Adrián González, Paloma Martínez
Computer Science and Engineering Department
Universidad Carlos III de Madrid
100494633@alumnos.uc3m.es, pmf@inf.uc3m.es

Abstract

This article presents our submission to the BioLaySumm 2024 shared task: Lay Summarization of Biomedical Research Articles. The objective of this task is to generate summaries that are simplified in a concise and less technical way, in order to facilitate comprehension by non-experts users. A pre-trained BioBART model was employed to fine-tune the articles from the two journals, thereby generating two models, one for each journal. The submission achieved the 12th best ranking in the task, attaining a meritorious first place in the Relevance ROUGE-1 metric.

1 Introduction

In the context of the rapidly expanding quantity and complexity of biomedical literature, the ability to effectively and accurately summarise documents has become crucial for researchers and healthcare professionals. In this regard, Natural Language Processing (NLP) technologies have emerged as promising tools to address this need. The objective of BioLaySumm 2024 shared task (Goldsack et al., 2024) is the simplification of biomedical research articles playing a vital role in making information more comprehensible to non-experts thus enabling a wider audience to understand and use medical information effectively.

Concerning generating summaries, there are a number of different approaches that can be employed. One such approach is the extractive model, which involves selecting the most important sentences from the original text and incorporating them directly into the summary. These models were the first to emerge and the most widely used until the abstractive models came onto the scene. These models have the capacity to comprehend the content of the input text and generate summaries that may include new sentences and expressions that are not present in the original text (Nallapati et al., 2017)(Widyassari et al., 2022). The first paper to describe an abstractive summarisation model was (Cohan et al., 2018), and from that moment on, they began to gain greater relevance and were used more frequently than the extractive models. In this paper, we will employ abstractive models.

In our participation in the BioLaySumm 2024 shared task, we utilise existing large language models (LLMs), such as Bio-BART (Yuan et al., 2022), which is a biomedical variant of the BART model (Lewis et al., 2020), and Longformer Encoder-Decoder (LED) (Beltagy et al., 2020), to train our models for the generation of summaries from the provided articles. The summaries generated by the various models were evaluated in accordance with the metrics provided by the organisers. (ROUGE(1,2 and L) (Lin, 2004), BERTScore (Zhang et al., 2020), FKGL (Kincaid et al., 1975), DCRS (Chall and Dale, 1995), CLI (Coleman and Liau, 1975), LENS (Maddela et al., 2023), AlignScore (Zha et al., 2023) and SummaC (Laban et al., 2021)). The experiment that yielded the most favourable results was the one that used the BioBART pre-trained model. This model was used to train two models, one for each of the journals from which the articles were obtained. These models were used to generate the abstracts for each journal.

This release achieved excellent results in the Relevance metric of the shared task, with the highest score in the ROUGE-1 metric. However, the Readability and Factuality metrics yielded less impressive outcomes, resulting in a final ranking of 12th place. Nevertheless, this remains a satisfactory performance, as it places the team in the top half of the table of all participants.

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 780–785
August 16, 2024. ©2024 Association for Computational Linguistics

PLOS Dataset			
data	**train**	**validation**	**test**
size	24733	1376	142
avg-length	6754.09	6741.48	6939.28
min-length	748	751	1587
max-length	26643	20423	18477

Table 1: Data statistics of PLOS dataset. Size corresponds to the number of articles present in the dataset. The min-length and max-length values correspond to the minimum and maximum length of the words in the dataset. Finally, avg-length corresponds to the average word length of all texts in the dataset.

eLife Dataset			
data	**train**	**validation**	**test**
size	4346	241	142
avg-length	10200.27	10031.25	8909.15
min-length	324	3408	2492
max-length	28696	23048	16880

Table 2: Data statistics of eLife dataset.

2 Method

2.1 Dataset

In order to participate in the BioLaySumm 2024 shared task, all participants are provided with two datasets containing biomedical research articles, the expert abstract, the name of the article sections and finally the keywords of each article. The first dataset contains approximately 25,000 articles from the Public Library of Science (PLOS), while the second dataset contains approximately 5,000 articles from the journal eLIfe. Details of the dataset are provided in (Goldsack et al., 2022).

In the tables 1 and 2 we can see the different statistics of the two journals (PLOS and eLife), in them we can see for each split its length of texts, the average number of words in each split, as well as the maximum and minimum length. The average length of articles varies depending on the journal to which they belong. For example, the average length of articles in the eLife journal is 10,200 words, while the average length of articles in PLOS is 6,754 words. In addition, there are notable differences in the length of the abstracts. The average length of an eLife abstract is twice that of a PLOS abstract, at 382 words versus 194, respectively.

2.2 Models

In order to generate the summaries, a number of approaches were tested, with two Transformer models being employed: Longformer (LED) and BioBART.

Longformer

Upon examination of the datasets in the previous study, it became evident that the articles were relatively lengthy. This prompted the decision to utilise a model that could process a substantial number of tokens as input. Consequently, the Longformer model, specifically the LED (Longformer Encoder - Decoder) variant, was selected (Beltagy et al., 2020). This model follows a sequence-to-sequence architecture (seq2seq) and is based on Transformer-base models. However, these are limited to short input sequences due to the exponential growth in computational complexity with the length of the inputs. Longformer models address this issue by introducing a mechanism whereby the complexity grows linearly in relation to the inputs. For the experiments, the pre-trained model *allenai/led-base-16384*[1] was utilised, which is capable of supporting inputs of up to 16,000 tokens. This is feasible due to the fact that it was initiated from a BART-base model. However, the BART model is only capable of processing texts up to 1,000 tokens. Consequently, the embedding matrix from the BART-base was copied and replicated 16 times in order to enable the Longformer model to process texts up to 16,000 tokens.

Bio-BART

Given the nature of this biomedical article summarisation and simplification task, it was deemed appropriate to utilise a model that has been pre-trained in this specific domain. Consequently, the BioBART model was employed (Yuan et al., 2022), as it has already demonstrated its efficacy in tasks of a similar nature and was employed in last year's task such as in (Phan et al., 2023). This model is based on a base BART model that has been pre-trained on a corpus of biomedical texts, rendering it an optimal choice for biomedical tasks. The model for the experiments is the pre-training *GanjinZero/biobart-v2-large*[2].

[1] https://huggingface.co/allenai/led-base-16384

[2] https://huggingface.co/GanjinZero/biobart-v2-base

3 Experiments

3.1 Evaluation Measures

Submissions for the shared task are evaluated according to three distinct criteria: relevance, readability and factuality.

- The relevance measure assesses the extent to which the generated abstract contains the key information from the original article. Four metrics will be employed to evaluate this: ROUGE-1 ↑, ROUGE-2↑, ROUGE-L ↑ (Lin, 2004) and BERTScore↑ (Zhang et al., 2020).

- Readability is a measure of the readability of the generated abstract, with the objective of ensuring that it is as understandable as possible for humans. In evaluating the readability of the abstract, four metrics are employed: Flesch-Kincaid Grade Level (FKGL) ↓ (Kincaid et al., 1975), Dale-Chall Readability Score (DCRS)↓ (Chall and Dale, 1995), Coleman-Liau Index (CLI) ↓ (Coleman and Liau, 1975) and LENS ↑ (Maddela et al., 2023).

- Factuality is the extent to which the generated summary is accurate and based on verifiable facts. For this, two metrics will be employed: AlignScore ↑ (Zha et al., 2023) and SummaC ↑ (Laban et al., 2021).

The objective of the relevance and factuality measures is to maximise the metrics, while in relevance we seek to minimise them, except for the LENS metric, which, like the previous ones, we seek to maximise.

3.2 Experiments

Three distinct experiments were conducted utilising the two previously trained models.

Longformer

The pre-trained *allenai/led-base-16384* model is employed in the experiments, which is capable of supporting inputs of up to 16,384 tokens. In this experiment, a single model will be trained on the texts of the two journals, and the summaries will be generated from the same model. Consequently, the training of the model employs the texts of the two journals. Despite the maximum input capacity of the model being 16,384 tokens, the texts are limited to those below 12,000 words due to identified constraints. Nevertheless, the training is based on more than 20,000 texts.

Bio-BART

The experiment employs the *GanjinZero/biobart-v2-large* pre-training model, which is a biomedical pre-training model. However, as a bart model, it has an input limitation of 1024 tokens. Consequently, for the training process, the complete dataset is utilised, with only the initial tokens of each text being employed. This approach allows for the retention of the initial tokens, which are then used for the training process. The information retained is the abstract, which has an average length of 300 tokens plus the beginning of the introduction. The average number of tokens in these two fields is 1080, demonstrating that by utilising these two sections, we are able to retain a substantial amount of information. In contrast to the aforementioned experiment with the LED model, two distinct training sets will be employed in this instance. One will comprise articles from the PLOS journal, while the other will comprise articles from the eLife journal. This approach will result in the generation of two independent models, each of which will produce summaries of the articles in their respective test sets. The fine-tuning process will utilise both complete datasets.

Longformer + Bio-BART

Finally, in order to enhance the outcomes of the preceding experiments, we opted to integrate the two models in order to retain the most advantageous aspects of each. This integration will allow us to leverage the capacity of the LED model to process voluminous text inputs while simultaneously capitalising on the BioBART model's aptitude for biomedical simplifications. As with the BioBART model, in this experiment we will utilise two independent models, one for PLOS journal and one for eLife.

In order to achieve this, the Longformer model is first employed. The input for this model is the full articles, and the output is between 700 and 800 words, which is more than double the average length of the final summaries to be delivered. Once the first summaries have been generated by the Longformer model, they are used as input to the BioBART models, which generate the final summaries.

	Relevance				Readability				Factuality	
	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	BERTScore	FKGL↓	DCRS↓	CLI↓	LENS↑	AlignScore↑	SummaC↑
Best Score	0.487	0.156	0.454	0.867	10.459	6.760	11.044	81.205	0.930	0.902
LED	0.411	0.113	0.386	0.846	13.592	**8.810**	14.966	27.749	**0.753**	0.652
BioBART	**0.487**	**0.147**	**0.452**	**0.862**	**12.710**	10.433	14.080	49.344	0.667	**0.670**
LED + BioBART	0.456	0.131	0.426	0.857	13.025	9.605	**13.360**	**52.124**	0.580	0.540

Table 3: The results of the three experiments (LED, BioBART, LED + BioBART) are presented alongside a comparison with the best results obtained in each metric in the competition.

		Relevance				Readability				Factuality	
		ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	BERTScore	FKGL↓	DCRS↓	CLI↓	LENS↑	AlignScore↑	SummaC↑
LED	Average	0.411	0.113	0.386	0.846	13.592	8.810	14.966	27.749	0.753	0.652
	PLOS	0.421	0.142	0.393	0.855	13.389	8.956	14.850	29.155	0.784	0.701
	eLife	0.400	0.084	0.379	0.837	13.794	**8.665**	15.082	26.343	0.723	0.604
BioBART	Average	0.487	0.147	0.452	0.862	12.710	10.433	14.080	49.344	0.667	0.670
	PLOS	0.465	**0.155**	0.425	**0.863**	14.566	11.936	16.550	34.079	**0.791**	**0.827**
	eLife	**0.509**	0.138	**0.479**	0.861	10.854	8.930	**11.610**	**64.609**	0.542	0.514
LED + BioBART	Average	0.456	0.131	0.426	0.857	13.025	9.605	13.360	52.124	0.580	0.540
	PLOS	0.426	0.134	0.392	0.857	15.231	10.365	15.034	41.056	0.651	0.597
	eLife	0.487	0.127	0.459	0.856	**10.820**	8.846	11.685	63.192	0.509	0.540

Table 4: The results of the metrics in the PLOS and eLife journals for each of the three experiments are presented below.

3.3 Environment Parameters

All experiments were conducted on a Tesla T4 GPU, with a series of hyperparameters set, including a learning rate of 2e-5, a batch size of 4, and two epochs.

4 Results and discussions

The table 3 presents the outcomes of the experiments, displayed in the context of the various metrics. Furthermore, an additional row has been included, in which the best value for each metric within the competition is presented. Table 4 presents the results obtained for each journal, allowing for a more detailed analysis. Upon examination of the results, the following observations can be made.

The first of these observations is that our BioBART value in the ROUGE-1 metric is the best value in the competition. In addition to this excellent result in this metric, we can also see that in the other relevance metrics we also obtain very good results, being very close to the best results. Furthermore, an analysis of the results by journal reveals that there are minimal differences between the texts of the two groups. The journal PLOS outperforms the other texts in two metrics (ROUGE-2 and BERTScore), while eLife excels in two others (ROUGE-1 and ROUGE-L). This indicates that the model generates summaries that retain a substantial amount of relevant information. In the experiment in which we combined LED and BioBART, we also obtained very good results, which suggests that these results are due to the BioBART model.

Conversely, an analysis of the Readability metrics reveals that the optimal outcome is achieved when the two models are combined. However, when the Dalle-Chall Readability Score (DCRS) metric is considered, the LED model exhibits significantly superior performance. Furthermore, this metric presents an intriguing phenomenon: the results in the BioBART model are quite poor, with a score of 1.5 points above our best result. This is a significant drawback for the model in terms of its final score. In contrast to the previous observation regarding relevance, the texts of the journal eLife obtain much better results than those of the journal PLOS.

With regard to the Factuality metrics, the BioBART model yielded the most favourable results, with the exception of the eLife journal, where the outcomes were considerably less favourable. Consequently, the average score was reduced, resulting in the LED model, which is more balanced, achieving better results in the AlignScore metric.

The findings of this study indicate that while the information is well-maintained, as evidenced by the relevance metrics. The PLOS journal articles contain more accurate information but are more challenging to comprehend. This discrepancy may be attributed to the smaller abstracts (175-220 words), which may have a detrimental impact on the readability metrics.

The BioBART model is the most effective in terms of relevance, outperforming all the metrics in this category thanks to its specific biomedical training. Although the combined Longformer+BioBART model improves readability, it loses accuracy due to the double simplification of the content. On the other hand, the Longformer model, although it obtained good results in some metrics, did not stand out in any of them; this could be an effect of having trained a single model with the texts of the two journals.

4.1 Selection of approach

Following the completion of the three experiments and the analysis of the results obtained from the various metrics, it was determined that the most optimal approach would be to utilise the BioBART model, as it yielded the most favourable outcomes in six out of the ten metrics, with at least one in each of the three categories.

5 Conclusions

This paper presents our participation in the BioLaySumm 2024 shared task, which aimed to generate lay summaries of large biomedical documents. In this task, we trained two different models (LED and BioBART) from which we generated three different experiments. Upon completion of the task, we observed that the best results were obtained by training two BioBART models (one for the PLOS journal articles and another for the eLife articles). This is our final submission to the competition, which resulted in a 12th-place finish. Our performance was particularly noteworthy in the ROUGE-1 metric, where we achieved first place, as well as in the Relevance metrics.

As future work, we would have liked to experiment with other models that we found interesting, particularly trained with medical data, such as medical mT5 (García-Ferrero et al., 2024). With respect to the models we have presented, we would like to continue working with them to improve the results in the Readability and Factuality metrics, in which we have not obtained such good results. We would like to study what happened in generating not adequate summaries by conducting an analysis of errors. We believe that managing specific medical terminology would help to generate more lay-oriented medical terms to ascertain the efficacy of the keyword translation from the original text to the summary. In the event that this process is not executed correctly, due to the inherent complexity of the keywords, an external dataset comprising words from the biomedical field and a translation into simpler expressions could be employed as a preprocessing step for the texts prior to training. See the open-access and collaborative (OAC) consumer health vocabulary[3] (CHV) as an example of a medical lay-oriented vocabulary.

Limitations

Our best result is obtained by using a BioBART model, which restricts the input of words to a maximum length of 1024 tokens. This represents the initial and most significant limitation encountered, given that the dataset comprises lengthy texts. Consequently, this limitation precludes the training of models with all available information, which would result in enhanced outcomes. Another limitation identified was the use of the Tesla T4 GPU. The extensive training data required for this device resulted in lengthy training times, which impeded the development of the models.

Acknowledgments

This work was supported by ACCESS2MEET project (PID2020-116527RB-I0) supported by MCIN AEI/10.13039/501100011033/.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

J.S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *Preprint*, arXiv:1804.05685.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. Medical mt5: An open-source multilingual text-to-text llm for the medical domain. *Preprint*, arXiv:2404.07613.

[3] https://biomedinfo.smhs.gwu.edu/chv-files

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.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. Lens: A learnable evaluation metric for text simplification. *Preprint*, arXiv:2212.09739.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Phuc Phan, Tri Tran, and Hai-Long Trieu. 2023. VBD-NLP at BioLaySumm task 1: Explicit and implicit key information selection for lay summarization on biomedical long documents. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 574–578, Toronto, Canada. Association for Computational Linguistics.

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

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model. *Preprint*, arXiv:2204.03905.

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

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