NCUEE-NLP at BioLaySumm Task 2: Readability-Controlled Summarization of Biomedical Articles Using the PRIMERA Models

Chao-Yi Chen, Jen-Hao Yang, and Lung-Hao Lee
Department of Electrical Engineering
National Central University
No. 300, Zongda Rd., Zhongli Dist., Taoyan City 32001, Taiwan
110581007@cc.ncu.edu.tw, 111521054@cc.ncu.edu.tw, lhlee@ee.ncu.edu.tw

Abstract

This study describes the model design of the NCUEE-NLP system for BioLaySumm Task 2 at the BioNLP 2023 workshop. We separately fine-tune pretrained PRIMERA models to independently generate technical abstracts and lay summaries of biomedical articles. A total of seven evaluation metrics across three criteria were used to compare system performance. Our best submission was ranked first for relevance, second for readability, and fourth for factuality, tying first for overall performance.

1 Introduction

Automatic text summarization is a NLP task that aims to generate concise, relevant, and informative summaries of a source document. The goal is to provide readers with a quick and easy way to understand the essential content of a document without having to read the entire text. Different from extractive summarization techniques that select important sentences from the original texts to create a summary, abstractive summarization methods learn a semantic representation of the source content to generate a summary that conveys the meaning of original texts in a more concise way.

Biomedical publications usually contain a wide range of health-related topics and may be of interest to researchers, medical experts, journalists and even members of the general public. However, such articles tend to include considerable amounts of technical jargon which can present a significant comprehension barrier to non-expert readers. Abstractive summarization models have the potential to help broaden access to biomedical articles by highlighting key information in less technical language.

The BioLaySumm shared task (Goldsack et al., 2023) at the BioNLP 2023 workshop seeks to generate more readable summaries of biomedical articles (i.e., a "lay summary"). We participated in the Task 2 focused on readability-controlled summarization, where given the main text of a biomedical article as input, the goal is to develop a single model to generate both a technical abstract and a lay summary.

This paper describes the **NCUEE-NLP** (**N**ational **C**entral **U**niversity, Dept. of **E**lectrical **E**ngineering, **N**atural **L**anguage **P**rocessing **L**ab) system for the BioLaySumm Task 2 (Goldsack et al., 2023). Our solution explores the use of a pretrained PRIMERA model and fine-tuning on the downstream summarization task for technical abstract and lay summary generation. Each participating team was allowed to submit a maximum of three runs. The three evaluation criteria contained seven metrics used for performance comparisons. Our best performing submission ranked first for relevance, second for readability, and fourth for factuality, and tied first for overall performance.

The rest of this paper is organized as follows. Section 2 describes the NCUEE-NLP system for BioLaySumm Task 2. Section 3 presents the results and performance comparisons. Conclusions are finally drawn in Section 4.

2 The NCUEE-NLP System

Figure 1 shows our NCUEE-NLP system architecture for the BioLaySumm Task 2. We use the PRIMERA model (Xiao et al., 2022), which is a pre-trained model focused on multi-document summarization, as our main architecture to

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 586–591

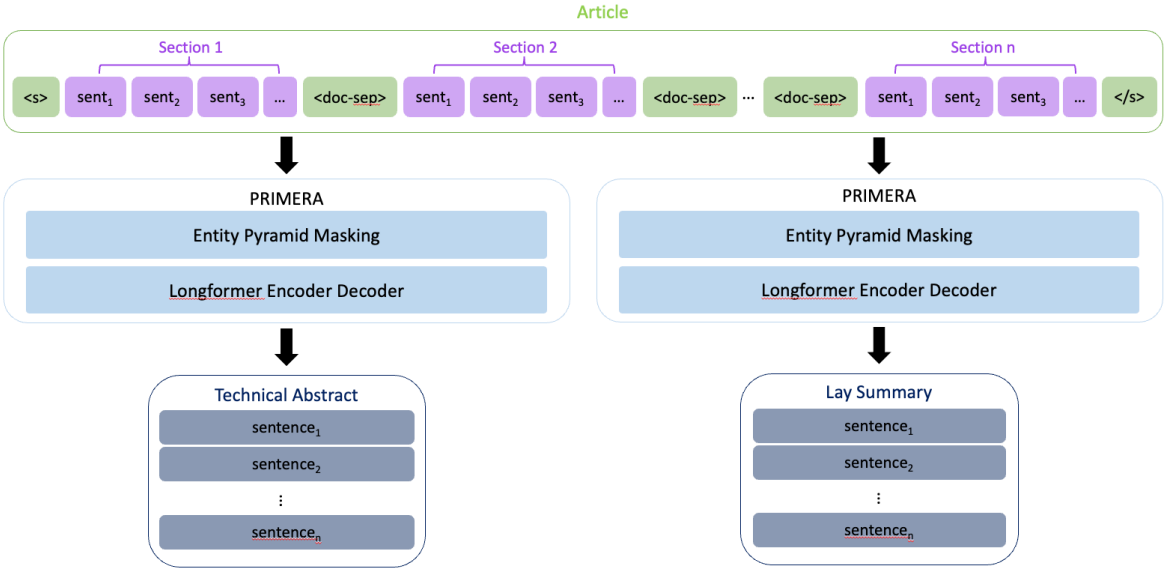

Figure 1: Our NCUEE-NLP system architecture for BioLaySumm Task 2.

independently generate a technical abstract and a lay summary for this task.

PRIMERA (Xiao et al., 2022) was designed to reduce the needs for dataset-specific architectures and large fine-tunning labeled data. Following the Gap Sentence Generation (GSG) objective used in PEGASUS (Zhang et al., 2020a), a new pretraining strategy called Entity Pyramid was proposed to train the model to identify and aggregate salient information based on entity frequency across related document clusters. Different from the importance heuristic in PEGASUS, the Entity Pyramid strategy selects masked sentences that are representative of more documents in the cluster rather than exact matching between documents. PRIMERA uses the Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) as the model initialization. The length limit of input was set at 4096 with a sliding window size of 512 for local attention. The output length limit was configured as 1024. Extensive experiments on 6 datasets from 3 different domains showed that PRIMERA outperformed other state-of-the-art models with zero-shot, few-shot, and fully-supervised settings. This motivated us to explore how the PRIMERA model can be used to effectively address this task challenge.

Biomedical articles are comprised of a sequence of sections such as Introduction, Materials, Method, Results, and Conclusions, wherein each section is regarded as a document to be delimited with a special <doc-sep> token as the input. The Acknowledgments and References sections were excluded. Tokens that exceed the input length limit were automatically discarded. It's identical to the original PRIMERA model (Xiao et al., 2022), using a local and global attention mechanism in the encoder self-attention side while using full attention on the decoder and cross-attention.

We used the training set to separately fine-tune the pretrained PRIMERA models to generate the technical abstract and lay summary. During the evaluation phase, we independently used the PRIMERA model to generate the abstract and summary as our model outputs and group them for final system result submission.

3 Performance Evaluation

3.1 Data

The datasets were mainly provided by task organizers, including a collection of biomedical research articles, technical abstracts, and expert-written lay summaries (Goldsack et al., 2022; Luo et al., 2022). The biomedical articles were published by the Public Library of Science (PLOS). A total of 24,773 training articles were used to fine-tune the pre-trained language models. The average number of tokens in the technical abstracts and lay summaries are respectively 268.11 and 194.90.

During the development phase, 1376 articles in the validation set were used to develop the system and obtain optimized parameters. The technical abstracts and lay summaries respectively averaged 271.12 and 194.51 tokens. Finally, a test set containing 142 articles was used to evaluate the system performance.

3.2 Settings

Each participating team was allowed to submit a maximum of three runs, and we selected the following models to compare system performance.

(1) PEGASUS (Zhang et al., 2020a)

We had previously used the PEGASUS transformer with promising results for the health question summarization task of MEDIQA challenge at the BioNLP 2021 workshop (Lee et al., 2021). We therefore retuned PEGAGUS for this summarization task. However, different from using the model pretrained on the XSum datasets (Narayan et al., 2018), we selected the PEGASUS-Large model pretrained on PubMed abstracts and full texts, with sampled gap sentences ratios on the both C4 (Raffel et al., 2020) and HugeNews datasets, with important sentences sampled stochastically.

(2) BART-Longformer (Beltagy et al., 2020)

The BART transformer (Lewis et al., 2020) was identified by task organizers as the baseline model, thus we used the BART-Longformer, an enhancement of BART-Large with Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) to support long document sequence-to-sequence to generate the technical abstract and lay summary.

(3) PRIMERA (Xiao et al., 2022)

This is the pretrained model mainly used in our developed NCUEE-NLP system.

All pretrained models were downloaded from HugginFace[1]. In general, we compared the model performance differences on the downstream summarization task without further changes. We continuously fine-tuned these models using the training dataset and optimized the parameters using the validation set. On a server with a Nvidia A100 GPU (40GB memory), the hyperparameter values for our model implementation were configured as follows: batch size 2; epochs 10; learning rate 3e-5; length limit of input 4096 for BART-Longformer and PRIMERA and 1024 for PEGASUS; length limit of output 256.

3.3 Metrics

The system generated summaries evaluated across three criteria as follows:

(1) Relevance

ROUGE (Lin, 2004) includes several automatic evaluation methods to measure the similarity between summaries. ROUGE-1/2 is a uni-gram/bi-gram recall between a candidate summary of a reference summary. ROUGE-L accounts for the union Longest Common Sequence (LCS) in matching between a reference summary and every candidate summary sentence. BERTScore (Zhang et al., 2020b) leverages the pre-trained contextual embeddings from BERT (Develin et al. 2019) and matches words between the candidate and reference sentences based on cosine similarity.

(2) Readability

The Flesch Kincaid Grade Level (FKGL) is a readability formula to access the approximate reading grade level of a text, which weights the average number of words in a sentence and the average number of syllables in a word. The lower the FKGL score, the easier a piece of text is to read. The Dale-Chall Readability Score (DCRS) measures a text against a number of words considered familiar to fourth-grade American students that uses word-length to determine how difficult a word is for readers to understand it. Similarly, a lower the DCRS score means the better readability.

(3) Factuality

BARTScore (Yuan et al., 2021) conceptualizes the evaluation of generated texts as a text generation task using pretrained sequence-to-sequence models. BART (Lewis et al., 2020) is operationalized as standard model for comparison due to its superior performance in text generation. BARTScore relies on a BART's average log-likelihood of generating the evaluated summary conditional on the source document, suggesting that summarization models potentially make factual errors in the form of lower generation probability (Koh et al., 2022).

According to the BioLaySumm shared task evaluation policy, the scores presented for each metric will be the average of those calculated

[1] https://huggingface.co/google/pegasus-pubmed
https://huggingface.co/allenai/PRIMERA
https://huggingface.co/hyesunyun/update-summarization-bart-large-longformer

Model	Type	Relevance				Readability		Factuality
		ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	FKGL	DCRS	BARTScore
PEGASUS	summary	0.4241	0.1176	0.3847	0.8483	2.2680	1.0663	-1.7713
	abstract	0.4274	0.1320	0.3901	0.8439	2.3744	1.0940	-1.7717
	average	0.4257	0.1248	0.3874	0.8461	2.3212	1.0801	**-1.7715**
BART-Long	summary	0.4569	0.1323	0.4124	0.8545	2.0250	0.8782	-1.6166
	abstract	0.4631	0.1516	0.4200	0.8523	1.9912	0.9518	-1.6387
	average	0.4600	0.1419	0.4163	0.8534	**2.0081**	0.9150	-1.6277
PRIMERA	summary	0.4567	0.1338	0.4159	0.8557	2.1165	0.8603	-2.1185
	abstract	0.4697	0.1557	0.4287	0.8548	1.9996	0.8617	-2.0692
	average	**0.4632**	**0.1447**	**0.4223**	**0.8552**	2.0581	**0.8637**	-2.0939

Table 1: Submission results on the validation set.

Model	Type	Relevance				Readability		Factuality
		ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	FKGL	DCRS	BARTScore
PEGASUS	summary	0.4115	0.1049	0.3707	0.8481	2.4641	1.0598	-1.8114
	abstract	0.4292	0.1284	0.3910	0.8450	2.4261	1.0668	-1.8694
	average	0.4203	0.1166	0.3808	0.8465	2.4451	1.0633	-1.8404
BART-Long	summary	0.4376	0.1184	0.3991	0.8527	2.2225	0.9366	-1.6380
	abstract	0.4591	0.1488	0.4154	0.8531	2.1824	0.9485	-1.6919
	average	0.4483	0.1336	0.4043	0.8529	2.2025	0.9425	**-1.6649**
PRIMERA	summary	0.4376	0.1272	0.3997	0.8548	2.1465	0.9753	-2.1722
	abstract	0.4652	0.1532	0.4250	0.8542	1.9486	0.8928	-2.0482
	average	**0.4514**	**0.1402**	**0.4123**	**0.8545**	**2.0475**	**0.9340**	-2.1102

Table 2: Submission results on the test set.

independently for the generated abstracts and lay summaries. Each submission will be independently ranked based on each of three evaluation aspects (i.e., Relevance, Readability, and Factuality). Before averaging across metrics within each evaluation aspect, the min-max normalization is applied to the scores of each metric among all submissions. A submission's overall ranking is computed based on the cumulative rank across the three aspects. The lower the cumulative rank, the better the system performance.

3.4 Results

Table 1 shows the results on the validation set. The PRIMERA model performed best for all evaluation metrics in terms of both relevance and readability, while the PEGASUS transformer produced a better BARTScore for factuality.

Table 2 shows the results on the test set are very similar to those on the validation set. Similarly, the PRIMERA model outperformed the others for the same metrics. However, the BART-Longformer model outperformed PEGSUS on the test set.

In summary, we found that using the PRIMERA model for technical abstract and lay summary generation can achieve the better summarization performance in terms of both relevance and readability, but underperforms in terms of factuality.

3.5 Analysis

Summarization performance on technical abstracts usually outperformed lay summary generation, no matter which model and evaluation metric were concerned, confirming that lay summarization is more complicated than the abstract generation task.

Regarding the BARTScore for factuality evaluation criterion, almost all participating systems underperformed in this evaluation metric, even performing worse than the baseline result of the task organizers' BART-base model. We will conduct the error analysis to find possible reasons when the evaluation code and gold standard of the test set are publicly available.

3.6 Rankings

According to official rankings released by task organizers (Goldsack et al., 2023), our best submission based on the PRIMERA model ranked first for relevance, second for readability, and fourth for factuality, tying first for overall performance.

4 Conclusions

This study describes the NCUEE-NLP system for the BioLaySumm Task 2, including system design, implementation and evaluation. Each section in a biomedical article is regarded as a document and kept its original sequence in paper organization for multi-document summarization based on the PRIMERA model. We fine-tuned the pretrained language model using the training set, achieving good performance on the test set for relevance and readability, while improvements are still needed in terms of factuality. Finally, our submission tied for first place for overall performance.

Limitations

This study is our first exploration of this research topic. We use pre-trained PRIMERA, PEGASUS, and BART-Longformer models and fine-tune them for technical abstract and lay summary generation. Novelty in the techniques is the main limitation. We downloaded all pre-trained from Hugginface, which may be inappropriate for this summarization task. How to determine an excellent pre-trained model doesn't include in this study. In addition, we only fine-tuned all pre-trained models over the task-given datasets without collecting other related summarization data to enhance the model performance.

Acknowledgments

This study is partially supported by the National Science and Technology Council, Taiwan, under the grant MOST 111-2628-E-008-005-MY3.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. arXiv preprint, arXiv:2004.05150. https://doi.org/10.48550/arXiv.2004.05150

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pages 4171-4186. http://dx.doi.org/10.18653/v1/N19-1423

Tomas Goldsack, 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 22nd Workshop on Biomedical Natural Language Processing*. 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*. Association for Computational Linguistics, pages 10589-10604.

Lung-Hao Lee, Po-Han Chen, Yu-Xiang Zeng, Po-Lei Lee, and Kuo-Kai Shyu. NCUEE-NLP at MEDIQA 2021: health question summarization using PEGASUS transformers. In *Proceedings of the 20th Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, pages 268-272. http://dx.doi.org/10.18653/v1/2021.bionlp-1.30

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*. Association for Computational Linguistics, pages 7871–7880. http://dx.doi.org/10.18653/v1/2020.acl-main.703

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*. Association for Computational Linguistics, pages 74-81.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, pages 4667-4680.

Huan Yee Koh, Jiaxin Ju, He Zhang, Min 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*, Association for Computational Linguistics, pages 2682-2698.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1797-1807. http://dx.doi.org/10.18653/v1/D18-1206

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485-5551.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pages 5245–5263. http://dx.doi.org/10.18653/v1/2022.acl-long.360

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: evaluating generated text as text generation. In Proceedings of the 35th Conference on Neural Information Processing Systems. Volume 34, pages 27263-27277. https://doi.org/10.48550/arXiv.2106.11520

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:11328-11339. https://doi.org/10.48550/arXiv.1912.08777

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1904.09675