

VBD-NLP at BioLaySumm Task 1: Explicit and Implicit Key Information Selection for Lay Summarization on Biomedical Long Documents

Phuc Xuan Phan, Tri Huu Tran, Hai-Long Trieu

VinBigData JSC, Hanoi, Vietnam

{v.phucpx, v.trith1, v.length12}@vinbigdata.com

Abstract

We describe our systems participated in the BioLaySumm 2023 Task 1, which aims at automatically generating lay summaries of scientific articles in a simplified way so that its content becomes easier to comprehend for non-expert readers. Our approaches are based on selecting key information by both explicit and implicit strategies. For explicit selection strategies, we conduct extractive summarization based on selecting key sentences for training abstractive summarization models. For implicit selection strategies, we utilize a method based on a factorized energy-based model, which is able to extract important information from long documents to generate summaries and achieve promising results. We build our systems using sequence-to-sequence models, which enable us to leverage powerful and biomedical domain pre-trained language models and apply different strategies to generate lay summaries from long documents. We conducted various experiments to carefully investigate the effects of different aspects of this long-document summarization task such as extracting different document lengths and utilizing different pre-trained language models. We achieve the third rank in the shared task (and the second rank excluding the baseline submission of the organizers).

1 Introduction

Lay summarization is a crucial task and gaining increasing attention due to its potential to provide accessible and digestible scientific information to the general public (Guo et al., 2021). The task involves summarizing technical and specialized content into a readable format for non-expert readers. This task is particularly relevant in biomedical fields, where research findings have significant implications for public health (Vinzberg et al., 2023). In order to help broaden access to technical texts and progress toward more usable abstractive summarization models in the biomedical domain, the BioLaySumm 2023 shared task (Goldsack et al.,

2023) has been organized for lay summarization task on biomedical research articles.

The challenges of this lay summarization task are in two folds: 1) input texts are full articles containing up to 10k sentences, which require models to capture long dependencies and extract key information fragments to generate summaries; 2) the lay summarization task requires us to generate summaries which not only convey the main meaning of the articles but also non-expert vocabularies for readers.

We build our systems based on sequence-to-sequence models with different key information selection strategies to solve the lay summarization task on biomedical long documents. Our abstractive summarization systems are built using sequence-to-sequence (seq2seq) architectures, which have shown state-of-the-art (SOTA) performance in recent abstractive summarization models (Lewis et al., 2019; Zhang et al., 2019; Liu et al., 2022). In order to deal with the issues of long documents, we focus on two key information selection strategies. Specifically, for the first strategy, we explicitly select key sentences as the input for training abstractive summarization models. For the second strategy, long documents are used as inputs, and important information is implicitly extracted based on the factorized energy-based model to generate summaries, in which we utilize a model called FactorSum (Fonseca et al., 2022), which has shown to be effective in long document abstractive summarization. Furthermore, our systems are initialized by BioBART (Yuan et al., 2022) to take advantage of the biomedical domain pre-trained language model. We evaluate our systems by conducting experiments on different aspects such as the effects of sequence length selection, the pre-trained language models, and applying the SOTA model (Liu et al., 2022).

We obtain the best performance with the implicit selection FactorSum models and BioBART, sepa-

Data	Train	Validation	Test
#samples	24,773	1,476	142
max-length	26,647	20,394	18,154
avg-length	6,750	6,738	9,048
min-length	750	755	4,097

Table 1: Data statistics of PLOS dataset. *max-length*, *min-length*, and *avg-length* correspond to the maximum length, minimum length, and average length of articles in the dataset.

rately trained on the two datasets, i.e., PLOS and eLife of the shared task. For the final results on the test set, we achieve the third rank in average scores. For separate metrics, our systems outperform the top teams in three of seven metrics, i.e., the relevance metrics (rouge-1 and rouge-L), and the readability metric(DCRS).

2 Data

For the BioLaySumm 2023 Task 1, the shared task provided two separate datasets, i.e., PLOS and eLife (Goldsack et al., 2022; Luo et al., 2022), containing biomedical journal articles and corresponding expert-written lay summaries.

PLOS: The Public Library of Science (PLOS) is an open-access publisher that hosts influential peer-reviewed journals across all areas of science and medicine. The journals in question focus specifically on Biology Computational Biology, Genetics, Pathogens, and Neglected Tropical Diseases.

eLife: eLife is an open-access peer-reviewed journal with a specific focus on biomedical and life sciences. Similarly to PLOS, these digests aim to explain the background and significance of a scientific article in a language that is accessible to non-experts.

As the data statistics presented in Tables 1 and 2, this task is challenging when we need to generate summaries from long documents with an average from 6k to 10k sentences in each document, and even more than 25k sentences in maximum. This requires models dealing with capturing long-range dependencies to extract important fragment information while avoiding out-of-memory issues.

For the data sizes, PLOS contains 24k samples while eLife contains only 4k samples.

Data	Train	Validation	Test
#samples	4346	241	142
max-length	28,308	23,050	27,427
avg-length	10,159	9,989	12,260
min-length	322	3,393	3,310

Table 2: Data statistics of eLife dataset.

3 Our Approaches

We present two different strategies that we investigate to build our systems to solve this long-document lay summarization task.

3.1 Implicit Selection Models for Summarization

Data processing: For transformers-based models, long documents always face the problem of exponential computational cost. With hardware limitations, we preprocess the data by limiting the length of the article by taking the first n words and the last sentence of the article.

Lay Summarization: We use FactorSum (Fonseca et al., 2022) - recent promising research for abstractive summarization tasks of long documents. FactorSum also achieved state-of-the-art results on some long scientific article datasets such as PubMed and arXiv (Cohan et al., 2018). FactorSum demonstrates that disentangling content selection from the budget used to cover salient content improves the quality and capacity of abstract summaries through two steps: (1) generation of abstractive summary views covering salient information in subsets of the input document (*document views*); (2) generates an abstract summary from these views, following a budget (a threshold that limits the number of words used in the summary) and content guidance (information that guides the summarization system about what information to focus on in the summary).

3.2 Explicit Selection Models for Summarization

Extractive Summarization: We first extract important information (key sentences) before feeding it to our summarization models. Specifically, we generate two different extractive versions based on the following approaches.

- **ExSum(Lead):** We extract the first three sentences and the last sentence of each paragraph

	R-1 ↑	R-2 ↑	R-L ↑	BERT Score ↑	FKGL ↓	DCRS ↓	BARTScore ↑
Top-1 (<i>MDC</i>)	0.4822	0.1553	0.4485	0.8707	12.9370	10.2058	-1.1771
Top-2 (<i>Baseline</i>)	0.4696	0.1445	0.4371	0.8642	12.0694	10.2487	-0.8305
FactorSum_12k	0.4829	0.1469	0.4502	0.8571	12.2923	10.0862	-1.7357
FS_12k(PLOS)	0.4853	0.1711	0.4473	0.8617	14.8063	11.5870	-1.3791
FS_12k(eLife)	0.4805	0.1227	0.4532	0.8526	9.7781	8.5854	-2.0924

Table 3: The performance on the private-test set of our best system (FactorSum_12k) and compared with the top-ranked systems reported in the shared task leaderboard. FS_12k stands for FactorSum_12k model. (The best scores are in bold).

(article’s section) as the article input to train our models.

- **ExSum(Key)**: We select the abstract, conclusion, and the ExSum(Lead) results of the remaining sections.

Lay Summarization: The extractive sentences are then used to train our abstractive summarization models. We build a sequence-to-sequence model and finetuned it on the BioBART (Yuan et al., 2022).

4 Experiments

4.1 Settings

First, we use the available BioBART (Yuan et al., 2022) PrLM - a recent state-of-the-art pre-trained language model that supports the biomedical domain.

We experimented with extracting different article lengths using the first n words ($n=9k$ or $12k$ in our experiments) and the last sentence of the article. We use AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate is $5e^{-5}$. We set the generation_max_length to 512 and generation_num_beams to 4, max_source_length to 1024, max_target_length to 490 with PLOS, and 512 with eLife. We set batch_size to 2 because of the limitation of GPU memory. The gradient will accumulate every 4 iterations. The maximum number of training iterations is 50000 for all experiments on 1 GPU (NVIDIA RTX 2080Ti). During the training, we save the best model with the highest ROUGE-1 score based on the validation set. All our implementations are based on the public Pytorch implement from Transformers ¹.

4.2 Compared Systems

During the training process, we choose to build an individual model for each PLOS and eLife dataset.

To find the optimal architecture for this task within the models we have, we carefully investigate different aspects such as extracting different document lengths, utilizing different pre-trained language models, as well as using state-of-the-art architecture models. We compared the following system architectures.

- **FactorSum_(9k, 12k)**: Our sequence-to-sequence systems based on the FactorSum model (Fonseca et al., 2022), which we described in Section 3.1. We limit the article length to 9k and 12k words and the last sentence of the article as input to the FactorSum model. We only use pre-trained BioBART (Yuan et al., 2022) models for experiments with the FactorSum models.
- **ExSum(Lead), ExSum(Key) + (BioBART, LED, BRIO)**: We build two-step summarization models. First, sentences are extracted based on approaches (ExSum(Lead), ExSum(Key)) presented in Section 3.2. Then, these sentences are used as input for sequence-to-sequence abstractive summarization models based on the pre-trained BioBART (Yuan et al., 2022), LED (Beltagy et al., 2020), and (BRIO) (Liu et al., 2022), which is a SOTA model in abstractive summarization applying a ranking loss among candidate summaries.

Results are evaluated using the officially provided metrics including: *relevance* (ROUGE (1, 2, and L) (Lin and Hovy, 2003) and BERTScore (Zhang et al., 2020)), *readability* (FKGL and DCRS), and *factuality* (BARTScore (Yuan et al., 2021)). We use the best systems on the validation dataset with BARTScore (FactorSum) to generate test summaries for our submissions. The comparison results are presented in Table 4.4.

¹<https://huggingface.co/transformers/index.html>

Dataset	R-1↑	R-2↑	R-L↑	BERT Score↑	FKGL↓	DCRS↓	BARTScore↑
ExSum(Lead)							
+ BioBART	0.4592	0.1476	0.4147	0.6196	14.6584	11.4962	-2.3557
ExSum(Key)							
+ BioBART	0.4933	0.1726	0.4503	0.6388	14.9582	11.6746	-2.1166
+ LED	0.5021	0.1918	0.4624	0.6511	15.0639	11.1134	-1.6860
+ BRIO	0.4409	0.1428	0.4057	0.6324	12.1446	10.5797	-1.3334
FactorSum_9k	0.4911	0.1676	0.4492	0.8610	14.8744	11.5840	-1.3514
FactorSum_12k	0.4919	0.1693	0.4498	0.8611	14.8551	11.5465	-1.3312

Table 4: Results of our systems on the PLOS validation data.

Dataset	R-1↑	R-2↑	R-L↑	BERT Score↑	FKGL↓	DCRS↓	BARTScore↑
ExSum(Lead)							
+ BioBART	0.4875	0.1409	0.4599	0.6218	10.217	8.6879	-2.3908
ExSum(Key)							
+ BioBART	0.5007	0.1285	0.4702	0.6231	10.3992	9.0898	-2.9519
+ BRIO	0.4598	0.1095	0.4295	0.6300	12.1958	9.6356	-2.2064
+ LED	0.4835	0.1314	0.4588	0.6217	10.6531	8.0722	-2.0177
FactorSum_9k	0.5009	0.1349	0.4698	0.8520	9.9070	8.5992	-2.0932

Table 5: Results of our systems on the eLife validation data

4.3 Test Results

Table 3 shows the best results submitted on the leaderboard for PLOS, eLife, and both datasets using the FactorSum_12k model. Although our BARTScore metric is lower compared to teams ranked higher (Top-1, Top-2), we have achieved better results in other metrics such as ROUGE-1, ROUGE-L, and DCRS. We also show detailed results for each PLOS and eLife dataset. Overall, our model achieves positive results in evaluation metrics: *relevance*, *readability*, and *factuality*.

4.4 Analyses

Table 4 shows the evaluation results of different experiments on the validation set of the PLOS dataset. The ExSum(Key) + BioBART model achieved the best results in the ROUGE-1, ROUGE-2, ROUGE-L, and DCRS metrics. The ExSum(Lead) + BioBART model achieved the best result in the FKGL metric. Meanwhile, the FactorSum_12k model achieved the best results in the BERTScore and BARTScore metrics. For the PLOS dataset, our submission was chosen based on selecting the best model with the BARTScore metric, which is the FactorSum_12k. Furthermore, we can also see that the FactorSum-*nk* models also show evenly results across metrics used to evaluate the share-task.

Table 5 shows the evaluation results of differ-

ent experiments on the validation set of the eLife dataset. For this eLife dataset, we have not yet experimented with the FactorSum_12k model because of time limitations during the competition. The ExSum(Key) + BioBART model achieved the best results in the ROUGE-L metric, while ExSum(Lead) + BioBART model achieved the best results in the ROUGE-2 metric. Besides, ExSum(Key) + LED achieved the best results in the DCRS and BARTScore metrics, and FactorSum_9k has the best results in the ROUGE-1, BERTScore, and FKGL metrics. Overall on both PLOS and eLife datasets, we can see that FactorSum-*nk* models seem to have the most potential results, which is why we selected it to submit to the leaderboard.

5 Conclusion

In this paper, we describe and propose our approach to solving the lay summarization task in the BioNLP Workshop and Shared Tasks @ ACL 2023. We have also proposed different experiments on aspects such as extracting different document lengths and utilizing different pre-trained language models, as well as using state-of-the-art architecture have been proposed for Abstractive Summarization. Experimental results also show our model achieves positive results for the lay summarization task.

References

- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#).
- 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](#). *NAACL 2018*, page 615–621.
- Marcio Fonseca, Yftah Ziser, and Shay B. Cohen. 2022. [Factorizing content and budget decisions in abstractive summarization of long documents](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6341–6364, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- 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.
- Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. [Automated lay language summarization of biomedical scientific reviews](#). *AAAI 2021*.
- 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](#).
- Chin-Yew Lin and Eduard Hovy. 2003. [Automatic evaluation of summaries using n-gram co-occurrence statistics](#). *Proceedings of HLT-NAACL*, pages 71–78.
- Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. [Brio: Bringing order to abstractive summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). *ICLR 2019*.
- 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.
- Oliver Vinzelberg, Mark David Jenkins, Gordon Morrison, David McMinn, and Zoe Tieges. 2023. [Lay text summarisation using natural language processing: A narrative literature review](#).
- Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaying Zhang, Yutao Xie, and Sheng Yu. 2022. [Biobart: Pretraining and evaluation of a biomedical generative language model](#). *BioNLP 2022@ ACL 2022*, page 97.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. [Bartscore: Evaluating generated text as text generation](#). *NeurIPS 2021*.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. [Pegasus: Pre-training with extracted gap-sentences for abstractive summarization](#).
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). *ICLR2020*.