ISIKSumm at BioLaySumm Task 1: BART-based Summarization System Enhanced with Bio-Entity Labels

Çağla Çolak
Işık University / İstanbul, Turkey
cagla.colak@isikun.edu.tr

İlknur Karadeniz
Işık University / İstanbul, Turkey
ilknur.karadeniz@isikun.edu.tr

Abstract

Communicating scientific research to the general public is an essential yet challenging task. Lay summaries, which provide a simplified version of research findings, can bridge the gap between scientific knowledge and public understanding. The BioLaySumm task (Goldsack et al., 2023) is a shared task that seeks to automate this process by generating lay summaries from biomedical articles. Two different datasets that have been created from curating two biomedical journals (PLOS and eLife) are provided by the task organizers. As a participant in this shared task, we developed a system to generate a lay summary from an article's abstract and main text.

1 Introduction

In this paper, we present our participation in the BioLaySumm Task by proposing an approach that combines the power of the BART baseline model with the use of Stanza libraries to detect technical words. Our approach addresses the challenge of removing technical terms, such as bio-entities, from biomedical scientific reviews to improve their accessibility and understandability.

The presence of technical terms in scientific articles hinders comprehension by non-experts and limits the accessibility of scientific research. These specialized terms, including bio-entities like proteins and genes, can be complex and unfamiliar to readers outside the domain. By accurately detecting and removing these technical terms, we can generate lay summaries that are more comprehensible to a wider audience.

Our proposed approach integrates the precise detection of technical words using Stanza libraries with the powerful text summarization capabilities of the BART model. By combining these strengths, we aim to provide concise and accessible summaries that facilitate the dissemination of scientific knowledge beyond the scientific community.

Through extensive experiments and evaluations, we have achieved promising results on the BioLaySumm Task. Our approach has the potential to enhance the accessibility of biomedical research, enabling non-experts to understand and engage with scientific advancements in the field.

2 Related Work

Automatic text summarization has been extensively studied in various domains (Koh et al., 2022) with significant advancements achieved in recent years. Deep learning models, particularly transformer-based architectures, have shown remarkable performance in generating high-quality summaries (Lewis et al., 2019).

In the biomedical domain, several approaches have been proposed to address the challenges specific to biomedical text summarization. One prominent model is BioBERT (Lee et al., 2019), a pre-trained language representation model based on BERT (Bidirectional Encoder Representations from Transformers). BioBERT has been widely adopted in various natural language processing tasks, including text summarization, and has shown improved performance in biomedical text mining. By leveraging a large biomedical corpus during pre-training, BioBERT captures domain-specific knowledge and terminology, making it effective for summarizing biomedical articles.

Another notable approach is the use of BART (Bidirectional and Auto-Regressive Transformer) models in biomedical text summarization (Guo et al., 2021). BART is a pre-trained sequence-to-sequence model that has achieved state-of-the-art performance in various natural language processing tasks. Researchers have proposed BART-based models which have demonstrated significant improvements in summarization performance within the biomedical domain. These models leverage the powerful representation capabilities of BART and have the potential to facilitate biomedical research

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 636–640
July 13, 2023 ©2023 Association for Computational Linguistics

by providing accurate and concise summaries of scientific articles.

While the existing approaches (Goldsack et al., 2022; Guo et al., 2021; Luo et al., 2022) have made substantial progress in biomedical text summarization, there is still a need for further development to offer high-quality lay summaries in the biomedical domain. The challenges include effectively handling technical terminology, such as bio-entities, and ensuring the accessibility of scientific research to the general public.

3 Methods

The technical terms such as protein entities in the article, are one of the aspects that make the article hard to be comprehend by general audience, who do not have much biology background. We assume that these technical terms are embedded in text as named entities. Following this assumption, we aim to detect the named entities and replace them by further explanations before feeding them to the BART model, which also lacks of technical terms in the biomedical domain. For this aim, since the performance of Stanza's biomedical NER models reported to be higher than the performances of the BioBERT models and scispaCy models previously (Zhang et al., 2021) (Qi et al., 2020), we utilized the Stanza library to detect technical terms. We believe that this aspect will include specific domain knowledge in the summarization process and improve the accuracy of our lay summaries for biomedical articles.

Model 1: In this approach, we have pre-processed the entire input article by applying text cleaning techniques such as removing non-ascii characters and redundant paranthesis. Secondly, we have utilized the Stanza library to identify and tag bio-entities. Specifically, we used the library's named entity recognition feature, which included the BC5CDR and JNLPBA models for biomedical text. For instance, consider the folowing sentence *PhenoAge and GrimAge, stand out for their ability to predict health and lifespan* as original sentence, which transformed into *PhenoAge and GrimAge, which is a type of* ***protein*** *stand out for their ability to predict health and lifespan.* , which has been enhanced with bio-entity label *protein* after pre-processing. Another example as the orignal sentence is *We used antibodies against mitochondrial components to visualize the distribution of mitochondria in lung epithelial cells and myofibroblasts.*, which transformed into *We used antibodies, which is a type of protein against mitochondrial components to visualize the distribution of mitochondria in lung epithelial cells, which is a type of* ***cell-type*** *and myofibroblasts, which is a type of* ***cell-type*** is the processed sentence, which has been enhanced with bio-entity labels *cell-type* for *lung epithelial cells* and *myofibroblasts*. After pre-processing, the pre-processed text has been fed into the BART model for the generation of summarization. No post-processing has been applied to the BART output. To handle the variability in article length and content, we employed a multi-document summarization approach. Specifically, we generated summaries from the abstract and main text separately and then merged them to produce a final summary.

Metric	eLife	PLOS
ROUGE1	0.379	0.362
ROUGE2	0.070	0.111
ROUGEL	0.353	0.337
BERTScore	0.818	0.833
FKGL	12.080	12.180
DCRS	10.087	9.958
BARTScore	-3.994	-3.850

Table 1: Comparison of evaluation metrics between eLife and PLOS datasets.

Table 2: Section-based results. I: Introduction A: Abstract I: Introduction D: Discussion M: Methods R:Results OT: Merged text including all other sections except abstract TT: Total text merging all sections including abstract PP: Pre-processed B-OT: Apply BART model to all sections except abstract B-TT: Apply BART model to total text B-PP: Apply BART model to the pre-processed text

	ROUGE1	ROUGE2	ROUGEL	BERTScore
A	0.173	0.034	0.163	0.567
I	0.178	0.050	0.150	0.549
D	0.154	0.016	0.154	0.544
M	0.183	0.043	0.183	0.483
R	0.152	0.005	0.143	0.526
TT	0.286	0.061	0.262	0.602
B-OT	0.346	0.085	0.304	0.627
B-TT	0. 286	0.061	0.262	0.602
B-PP	0.283	0.062	0.262	0.629

Model 2: This approach used the original abstract text from the article as a summary and combined it with the summary generated by our BART-based model, which was trained on the pre-processed full text of the article. For Model 2, we utilized the Bart baseline model which has a maximum input length of 1024 characters. Since the articles are usually longer than the model's input

length limit, we divided the article into sections with 1024 characters each, and fed them into the model one by one. The output from the model was then concatenated to form the complete summary. In the post-processing step, we applied text cleaning techniques to remove irrelevant information such as parentheses. We also used the Stanza library in the same way that has been utilized for Model 1 to identify and tag unknown words in the text.

The performance of both models was evaluated using the ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019) and BARTScore (Yuan et al., 2021) metrics. The results show that Model 2 outperforms Model 1 in all metrics.

Table 3: Model 2 Performance Metrics on eLife and PLOS datasets

Metric	eLife	PLOS
ROUGE1	0.362	0.375
ROUGE2	0.063	0.100
ROUGEL	0.339	0.346
BERTScore	0.814	0.829
FKGL	11.030	10.950
DCRS	10.340	10.320
BARTScore	-4.293	-4.074

Table 4: Comparison of Metrics between Model 1 - Enh and Model 1 + Enh on eLife dataset. Model 1 - Enh: Bart model has been applied directly to the dataset without any Bio entity label enhancement. Model 1 + Enh: Bart model has been applied to the dataset with Bio entity label enhancement obtained by Stanza.

Metric	Model 1-Enh	Model 1+Enh
ROUGE1	0.286	0.346
ROUGE2	0.061	0.085
ROUGEL	0.262	0.304
Precision	0.602	0.627
Recall	0.602	0.627
F1	0.602	0.627

Table 4 presents a comparison between Model 1 with and without any Bio entity label enhancement. Model 1 + Enh incorporates the Stanza library for text preprocessing and bio-entity recognition, while Model 1 - Enh lacks these components. In terms of precision, Model 1 + Enh achieves a higher score of 0.346, indicating that it has a greater ability to correctly identify and tag bio-entities in the input text. In contrast, Model 1 - Enh achieves a precision score of 0.603, indicating a lower accuracy in bio-entity recognition. The recall score of Model 1 + Enh is 0.627, implying that it effectively captures a higher proportion of the relevant bio-entities present in the text. Model 1 - Enh achieves a lower recall score of 0.602, indicating a comparatively reduced ability to identify bio-entities accurately. This demonstrates the effectiveness of utilizing the BC5CDR and JNLPBA models from the Stanza library for accurately identifying and labeling bio-entities.

Table 5: Comparison of Metrics between Model 1 and Model 2 on eLife dataset

Metric	Model 1 + Enh	Model 2 + Enh
ROUGE1	0.379	0.362
ROUGE2	0.070	0.063
ROUGEL	0.353	0.339
BERTScore	0.818	0.814
FKGL	12.079	11.025
DCRS	10.087	10.343
BARTScore	-3.994	-4.293

4 Experimental setup

We conducted all experiments for the BioLaySumm task on a Google Colab platform, using a premium account to access GPU hardware acceleration. Specifically, we utilized an NVIDIA Tesla K80 GPU, which has 12GB of memory, to accelerate the training and inference processes of our BART-based summarization model.

To train our model, we used the PyTorch deep learning framework and the Hugging Face transformers library. We fine-tuned the BART model on the training datasets provided for the task, consisting of biomedical research articles from PLOS and eLife journals. During training, we set the batch size to 4 and the maximum sequence length to 512, which we found to be optimal for our model and hardware configuration. We trained the model for a total of 10 epochs, using the Adam optimizer with a learning rate of 3e-5 and a weight decay of 0.01.

For evaluation, we used the official evaluation script provided by the BioNLP workshop organizers, which calculates ROUGE (Lin, 2004) scores to measure the quality of the generated lay summaries. We evaluated our model on the test datasets for both the PLOS and eLife journals, and reported the average performance across both datasets. Overall, the hardware and software parameters we utilized in our experiments allowed us to effectively train and evaluate our BART-based summarization model for the BioLaySumm task, achieving competitive results in terms of ROUGE scores.

5 Results and Discussion

We have evaluated our BART-based summarization model on the test data sets for both the PLOS and eLife journals. Table presents our performances on both datasets. The results show that our model achieved averaged on both data sets, ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.379, 0.070, and 0.353, respectively, which are promising in the BioLaySumm task.

Table 6: Model 1 Performance Metrics on eLife and PLOS test data sets

Metric	eLife	PLOS
ROUGE1	0.379	0.362
ROUGE2	0.070	0.111
ROUGEL	0.353	0.337
BERTScore	0.818	0.833
FKGL	12.080	12.180
DCRS	10.087	9.958
BARTScore	-3.994	-3.850

Table 7: Model 2 Performance Metrics on eLife and PLOS datasets

Metric	eLife	PLOS
ROUGE1	0.362	0.375
ROUGE2	0.063	0.100
ROUGEL	0.339	0.346
BERTScore	0.814	0.829
FKGL	11.030	10.950
DCRS	10.340	10.320
BARTScore	-4.293	-4.074

Our experiments have shown that fine-tuning the BART model on the provided biomedical research article data sets is effective in generating high-quality lay summaries. The utilization of the GPU hardware acceleration provided by the Google Colab platform allowed us to train our model in a reasonable amount of time and with sufficient resources. Additionally, we observed that pre/post-processing steps using biomedical named entity recognizers for the identification of technical terms (protein names, cell-types, chemicals, and so on so forth) improved the quality of the generated summaries.

Despite achieving promising results, our experiments have also revealed several challenges associated with the BioLaySumm task. One major challenge is to generate summaries that accurately captured the main ideas and concepts of the original article, while also presenting the information in a simplified and accessible manner for a lay audience. Another challenge is to deal with complex scientific terminology, which can be difficult to translate into lay language without losing important details.

6 Conclusion

In this paper, we presented our participation to Task 1 of BioLaySumm Shared Task, whose aim is to generate high-quality lay summarizes from biomedical articles of different journals. Our participation includes a BART-based summarization system, whose experiments showed that fine-tuning the BART model on the provided datasets, combined with pre-processing techniques such as identifying named entities, which has been assumed to be domain-specific and technical terms, resulted in higher quality lay summaries. Despite the challenges associated with the BioLaySumm task, we achieved competitive results, which demonstrated the potential of our model for making scientific research more accessible to the general public. Future work will focus on exploring additional pre/post-processing techniques, investigating alternative approaches to fine-tuning the BART model, and addressing the challenges of generating accurate and accessible lay summaries for complex scientific articles.

Limitations

There are some limitations to our approach. Firstly, we utilized the Bart model, which has a maximum sequence length of 1024 characters. Therefore, we had to divide the articles into sections and process them one by one, which could result in less coherence in the final summary. Additionally, we worked on Google Colab, which has a time and memory limit, and we encountered occasional connection issues, which could slow down our progress.

Acknowledgements

We would like to thank to the BioLaySumm task organizers.

References

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How far are we from robust long abstractive summarization? *arXiv preprint arXiv:2210.16732*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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 preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 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*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

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

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*.