IITR at BioLaySumm Task 1:Lay Summarization of BioMedical articles using Transformers

Venkat Praneeth Reddy
Indian Institute of Technology Roorkee
baddam_vpr@cs.iitr.ac.in

Pinnapu Reddy Harshavardhan Reddy
Indian Institute of Technology Roorkee
pinnapu_rhr@cs.iitr.ac.in

Karanam Sai Sumedh
Indian Institute of Technology Roorkee
karanam_ss@cs.iitr.ac.in

Raksha Sharma
Indian Institute of Technology Roorkee
raksha.sharma@cs.iitr.ac.in

Abstract

This paper presents our submission to the shared task-Lay Summarization of Biomedical Research Articles at BioNLP-2023 workshop(Goldsack et al., 2023). The purpose of this task is the summarize biomedical articles in a concise and less technical way increasing their readability and their reach to lay audiences. In this paper, we use BART-based summarization techniques. We used labels of sentences to improve the performance of our model. Our model achieved a rouge-1 score value of 42.89% and an FKGL score of 10.7901 in relevance and readability parameters respectively.

1 Introduction

Scientific paper's growth has increased a lot in recent times. Papers in various domains are used to share research data. So it is important to summarize these papers so as to decrease the workload of the researchers and also to reduce the gap between the public and researchers. Technical summary still is a difficult text to understand for non-research people and can lead to misinterpretation of information. In the context of Bio-medical articles the need for Lay summarization is higher because of the more dynamic and difficult terminology and also information misinterpretation having a direct impact on Human lives.

The shared task of BioNLP Lay Summarization of Biomedical Research Articles aims to improve the tools used for Lay summarization for training models which give realistic lay summaries. It has tasks in which data is to be trained on two datasets eLife and PLOS which are two large biomedical article datasets, both varying in sizes of summary and readability of summaries(Goldsack et al., 2022)(Luo et al., 2022). Our approach tries to build a model to keep the lay summary relevant to the original model and improve its readability.

2 Related Work

The models which were existing earlier are in general extractive models which include sentence selection or hybrid models selecting sentences and then summarizing them .(Cohan et al., 2018) was the first paper describing a model abstractive summarization.

LaySumm subtask of the CL-SciSumm 2020(Chandrasekaran et al., 2020) shared task series also had submissions that use abstractive summarization. The data insufficiency was found in the task as the sufficient number of annotated lay summaries were not present at the time of the task which will help in getting a more relevant and better lay summary. Many other models that are trained for Lay Summarization also faced the same problems due to a lack of data or data summaries being too large and not suitable for lay summary training.

Dimsum Lay summarization (Yu et al., 2020) used the CL-SciSumm dataset and a BART(Lewis et al., 2019) baseline along with sentence labels as external supervision signals, data augmentation and got better results of relevance and readability.

Corpora for lay summarization(Goldsack et al., 2022) introduced the dataset with the articles of eLife and PLOS each with annotated summaries to improve the existing literature of bio-medical articles available to train a lay summarization model.

3 Datasets

There were two datasets in the task eLife and PLOS. PLOS is the larger of the two datasets, containing $24,773$ instances for training and $1,376$ for validation (the same for each subtask). eLife contains $4,346$ instances for training and 241 for validation. Each dataset contains train, validation, and test datasets in the form of JSON files.

Each sample contains articles with the following structure:

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 625–628

Label	PLOS	eLife
Background	58.11	55.03
Objective	0.54	0.47
Methods	6.24	6.23
Results	17.86	18.23
Conclusions	17.26	18.83

Table 1: Mean percentage of each label within dataset lay summaries(Goldsack et al., 2022)

- **Id**-denoting the id of the article.
- **Year**-year in which article is published.
- **Title**-title of the article.
- **Sections**-Various sections of article which include:
  - **Introduction**- Gives the context and initial idea of the article.
  - **Results and methods**-These sections discuss the methods used in the article and their results.
  - **Discussions**-These sections discuss the inferences from the results and reasons for the observations.
  - **Abstract**-This section highlights the important points in the article.
  - **Summary**-Contains the annotated lay summary of the article.
- **Headings**-Contains information about all the headings of the article.
- **Keywords**-Contains the keywords of the article.

The eLife has an average summary length of around 300-350 words whereas the summary length of PLOS is way shorter i.e. around 160-200 words. The PLOS dataset summary is also closer to its abstract whereas eLife summaries are a more simplified version of text i.e. more readable.(Goldsack et al., 2022)

The table 1 shows that the most significant contributions of the lay summary in PLOS and eLife come from the Introduction, and abstract sections.

4 Data Pre-processing

Datasets as mentioned were in the JSON format with various headings as keys of JSON dictionary objects. These datasets are converted into a CSV file with each heading as a column. The irrelevant information columns like title, id, year are removed from the CSV file. Also as per table 1, sections like results and discussion don't contribute to the final lay summary. So even those columns are removed from CSV files. The information thus present in the CSV file is only an introduction, abstract, and summary.

A new text column that contains a fraction(k) of the Introduction and its concatenation with the abstract is made which is helpful to train the model with Introduction+abstract as the input text. K is a hyper-parameter and it was found out by experimentation that the optimal value was around 0.6. The text is truncated to a max length of 1024 tokens due to the capacity of the BART model.

5 System Overview

5.1 Baseline

We used the Facebook BART-based model as our baseline.BART(Lewis et al., 2019) is a transformer encoder-decoder (seq2seq) model with a bidirectional (BERT-like) encoder and an auto-regressive (GPT-like) decoder (Vaswani et al., 2017).

BART is very effective when it is fine-tuned with domain-specific datasets for text-generation and summarization tasks. BART is pre-trained by corrupting text with an arbitrary noising function and learning a model to reconstruct the original text. We used BART fine-tuned on the CNN/DailyMail dataset (Hermann et al., 2015) as baseline model.

5.2 Fine Tuning

We have fine-tuned the baseline with both eLife and PLOS datasets. We used the abstract+Fraction of Introduction as our input and summary as our output. We have performed training for around 25000 iterations and chose checkpoints with the least validation loss. We also evaluated the BART model which is not fine-tuned on the datasets to get a base score.

5.3 Metrics

We have used the Rouge1, Rouge2, and RougeL values for evaluating the relevance of our summary with respect to the actual article and we have used FKGL, and DCRS scores for measuring the readability of our summaries.

5.4 Libraries and Language

Our experiments are conducted on a supercomputer environment with the Conda 3.7 module environment and Python 3.7.2 language. Also, Pandas = 1.3.5 and transformers = 4.28.1 are used.

5.5 Testing

Testing is done by taking article sections as input to the fine-tuned model and summary sections as reference summaries to calculate the rouge scores. During testing due to the large size of article data and 1024 token capacity of the BART model summarizer input we had to give the input in the form of chunks i.e the article is divided into chunks of less than 1024 tokens and then pass the input. The max-length and min-length parameters of the summarizer are adjusted so that the output summary length is in the range of the average summary length of the dataset (different for PLOS and eLife).

6 Experiments

We did experiments with different parameters as follows:

- BART(Abstract-eLife): BART is chosen as the baseline model and the abstract of the eLife article is taken as input to the BART model.
- BART(Introduction-eLife): BART is chosen as the baseline model and the Introduction of the eLife article is taken as input to the BART model.
- BART(Introduction+abstract-eLife): BART is chosen as the baseline and part of the Introduction along with the abstract is given as input.
- T5(Introduction+abstract-eLife): T5 is chosen as the baseline and part of the Introduction along with the abstract is given as input.

For the hyper-parameters, we have considered a batch size of 4 due to the memory constraints of GPU. We have performed 15 epochs and chose a model with the least validation loss. We have taken the implementation of BART from hugging-face transformers.[1]

Model	Rouge-1	Rouge-2	Rouge-l
BART-base	27.47	8.27	28.03
BART+abs	35.76	11.83	32.73
BART+intro	38.49	11.72	35.75
BART+intro+abs	41.81	12.17	38.14
t5+intro+abs	38.47	11.27	35.03

Table 2: results for various experiments on elife dataset

Model	Rouge-1	Rouge-2	Rouge-l
Pre-trained on elife	42.81	12.82	39.26
Pre-trained on plos	41.47	12.91	38.80

Table 3: results for PLoS dataset with BART as base model and pretraining with different dataset

7 Results

We have calculated the Rouge1, Rouge2, and RougeL values for all the experiments. We can see that BART baseline without any fine-tuning gives Rouge1 value of around 27% showing that without fine-tuning the summarization might be more technical and fine-tuning is required for summarization to be lay.

We can see from the table that using the data of the sections of Introduction and abstract combined with BART baseline gives better results than using any of them separately probably due to Introduction not having relevant key points of the article and abstract size in eLife being very small compared to the summary.

Also, t5 results are a bit less accurate than the BART model showing BART model performance when fine-tuned with proper domain-related datasets gives very good results Also, the PLOS dataset tested with the eLife model has almost the same rouge values as the PLOS model so considering the PLOS model has less readability as compared to the eLife model we considered the eLife model for both the datasets for the final submission.

Ranking Phase: We have submitted the BART model fine-tuned with the eLife dataset. Our scores for relevance and readability are 42.81% for Rouge1, 32.7% for Rouge2, 10.7 for FKGL score, and 8.85 for DCRS. The ranks obtained for relevance and readability are 17 and 1 respectively.

8 Conclusion

Our results show that the BART model fine-tuned gives the best results as compared with other models. Also, the PLOS dataset has an abstract near to

[1] https://github.com/huggingface/transformers

its summary so using the eLife model for summarization gives more readability.

Limitations

We can improve our model in the future by using the keywords given in the dataset to enhance the quality of the summaries generated by using keyword embedding in addition to the fine-tuned model.

References

Muthu Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita Waard. 2020. Overview and insights from the shared tasks at scholarly document processing 2020: Cl-scisumm, laysumm and longsumm. pages 214–224.

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. *arXiv preprint arXiv:1804.05685*.

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.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

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*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tiezheng Yu, Dan Su, Wenliang Dai, and Pascale Fung. 2020. Dimsum@ laysumm 20: Bart-based approach for scientific document summarization. *arXiv preprint arXiv:2010.09252*.