# CS60075 TeamEnigma at BioLaySumm 2023 Task 1 - Lay Summarization of Biomedical Articles


| Onkar Sabnis | Sidharth V | Sagar Kumar Karn | Tushar Mohta | Yash Kulkarni |
|---|---|---|---|---|
| 18CH30018 | 18EX20029 | 19HS20039 | 19HS20055 | 18EE3AI22 |



## Abstract

The present study outlines the performance of Team Enigma in Task 1 of BioLaySumm 2023, which involved the development of lay summaries for biomedical articles. Our principal contribution lies in the refinement of transformer-based language models that were pre-trained on diverse text corpora, including scientific and biomedical literature, to generate the desired summaries. To evaluate the efficacy of our approach, we employed the widely-accepted Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics to compare the performance of various transformer-based models. [1].


## 1 Introduction

The domain of lay text summarization represents a distinct area within the realm of automatic text summarization, with a specific focus on generating lay language summaries that are readily comprehensible to a non-expert audience. The overarching objective of this endeavor is to enhance the accessibility of research findings to the general public, especially individuals with limited familiarity with academic writing across all age groups.

The task of text summarization entails a formidable challenge of capturing the essence of a document in a condensed form, utilizing fewer sentences and words. The process of extracting relevant data from unstructured text and deploying it for a summarization model can be broadly classified into two methods, namely Extractive and Abstractive approaches. While the former technique selects the most salient sentences within a given text without necessarily comprehending their meaning, the resulting summary merely represents a subset of the entire document. Conversely, the latter approach relies on advanced Natural Language Processing (NLP) techniques, such as word embeddings, to gain an understanding of the text's underlying semantics and generate a comprehensive summary.

[1]Please find the code for our work here.

## 2 Related Work

Conventional techniques for text summarization encompass the Frequency-Based Sentence Scoring approach, which represents a simple and intuitive extractive method for identifying the most pertinent sentences within a given text. In this method, the principles of information theory are employed to evaluate the relevance of each sentence in the input text based on relative frequencies. Sentences with high scores are considered informative and deemed appropriate for inclusion in the summary. However, the suitability of extractive summarization techniques for lay summarization may be limited since the source text not only needs to be summarized but also simplified to make it accessible to a non-expert audience.

A well-known method in academic literature Widyassar et al., 2019 is to compute the word frequencies on a large collection of clinical samples using TF-IDF rather than depending on internal frequencies (which only operate on long texts). When documents exhibit a repetitive structure that may be exploited to obtain additional information about the content, additional weighting systems, such as positional penalties Sarker et al., 2020, can also be devised.

Biomedical abstractive summarization has been extensively employed using encoder-decoder language models like BART, T5, and PEGASUS that are pre-trained with an objective function designed for abstractive text summarization. Deyoung et al. developed the BART-based method for multi-document summarising of medical studies for biomedical literature. They examine two

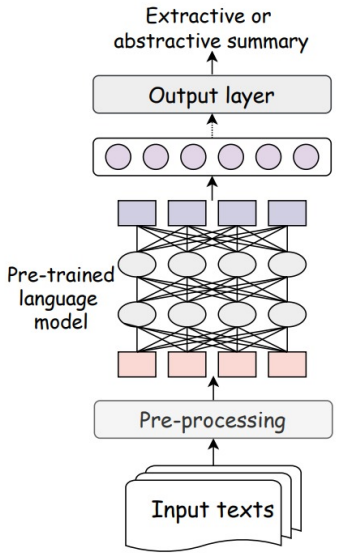


Figure 1: The framework of the fine tuning based method.

encoders for multi-document encoding. To individually encode numerous documents, one can use several BART encoders. LongformerEncoderDecoder (LED), which can encode long inputs up to 16K tokens, is another option. For COVID-19 articles, Su et al. also suggest the query-focused multi-document summarizer, which can provide abstractive and extractive summaries based on user queries. They fine tuned BART for multi-document abstractive summarization.

The subsequent sections of this report include model architecture, data for pre-training, data for fine-tuning. Further sections provides a detailed account of the experimental settings, baselines employed, as well as an in-depth comparison and analysis of the results obtained. Finally, last section furnishes a conclusive summary of our findings.

## 3 Model Architecture

Our efforts towards achieving the summarization objective have involved the application of multiple transformer models, which have been subjected to fine-tuning on pre-existing models. While Figure 1 provides an overview of the general framework that has been employed for fine-tuning purposes, Figure 2 furnishes a detailed illustration of the encoder-decoder architecture that has been utilized for pre-training BART and similar models.

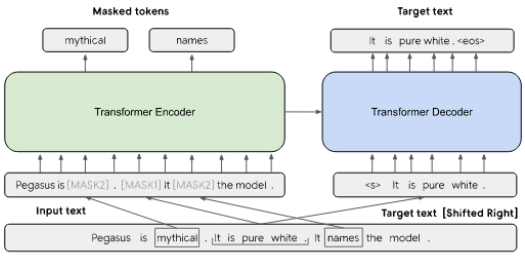


Figure 2: Encoder-decoder architecture for fine tuning

## 4 Data

### 4.1 Data for fine tuning

For model training and evaluation purposes, two distinct datasets have been employed, both of which are derived from biomedical journals - PLOS and eLife. Each dataset comprises full scientific articles accompanied by manually-created lay summaries.

PLOS, being an open-access publisher, hosts a range of peer-reviewed journals spanning various domains in science and medicine. On the other hand, eLife is a peer-reviewed journal that focuses specifically on biomedical and life sciences and is also open-access.

| Dataset | # Docs | Doc | Summary | |
|---|---|---|---|---|
| | | # words | # words | # sents |
| **PLOS** | 27,525 | 5,366.7 | 175.6 | 7.8 |
| **eLife** | 4,828 | 7,806.1 | 347.6 | 15.7 |

## 5 Experiments and Results

### 5.1 Transformer Language Models used

Multiple transformer-based models, namely T5-small, LED-base-16384, LED-base-8192, Distill-PEGASUS-xsum, and BART-base-cnn, were fine-tuned to generate lay summaries. Among these, LED-base-16384 and LED-base-8192 are longformer models that outperform transformers for longer sequences. This is due to the self-attention operation of transformer-based models, which scales quadratically with the sequence length, making them unable to process long sequences. LED-base-16384 and LED-base-8192 were initialized from the BART-base model.

### 5.2 Training, validation and Test Sets

For each sub-task, the training and the test sets are the same as those provided for the competition. The trial data given for each sub-task is taken to be the validation data.

| Model | Unsupervised objective |
|---|---|
| **BART** | Token Masking, Token deletion, Token infilling Sentence Infilling, Document Rotation |
| **T5** | Concatenated Spans Prediction |
| **Pegasus** | Gap Sentence Prediction, MLM |

Table 1: Self-supervised modeling objectives for the models employed in our experiments (Devlin et. al, 2017)

| Model | Rouge-1 | Rouge-2 |
|---|---|---|
| **T5-small** | 0.231 | 0.089 |
| **LED-base 16384** | 0.324 | 0.113 |
| **LED-base 8192** | 0.310 | 0.087 |
| **Distill-Pegasus-xsum** | 0.387 | 0.119 |
| **BART-base-CNN** | **0.463** | **0.134** |

Table 2: Comparing ROUGE scores of different models for eLife Dataset

| Model | Rouge-1 | Rouge-2 |
|---|---|---|
| **T5-small** | 0.243 | 0.102 |
| **LED-base 16384** | 0.341 | 0.122 |
| **LED-base 8192** | 0.297 | 0.106 |
| **Distill-Pegasus-xsum** | 0.401 | 0.127 |
| **BART-base-CNN** | **0.486** | **0.144** |

Table 3: Comparing ROUGE scores of different models for PLOS(Public Library of Science) Dataset

### 5.3 Hyperparameter Setting

We tuned the models around the following values of the hyperparameters. The values were decided based on literature study and previous works associated with fine-tuning of that model: learning_rate: 5e-05, train_batch_size: 4, eval_ba tch_size: 4, seed: 42, optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08, lr_scheduler_type: linear, num_epochs: 10, mixed_precision_training: Native AMP.

## 6 Conclusion

Our study demonstrates that fine-tuning transformer models (T5, LED-base, PEGASUS, BART) - which are pre-trained on extensive language corpora - results in better metric scores for lay summarization, compared to using off-the-shelf models directly. Furthermore, our findings suggest that BART-base-CNN provides the most favorable outcomes when fine-tuned on either the elife or PLOS dataset. Our results indicate that transformer-based language models, pre-trained on distinct text corpora, exhibit vastly superior performance when fine-tuned on domain-specific datasets, as opposed to deploying pre-trained models directly.

## 7 References

1. Mishra, R.; Bian, J.; Fiszman, M.; Weir, C.R.; Jonnalagadda, S.; Mostafa, J.; Del Fiol, G. Text summarization in the biomedical domain: A systematic review of recent research. *J. Biomed. Inform.* 2014, 52, 457–467. [CrossRef] [PubMed]

2. Afantenos, S.; Karkaletsis, V.; Stamatopoulos, P. Summarization from medical documents: A survey. *Artif. Intell. Med.* 2005, 33, 157–177. [CrossRef]

3. Moradi, M.; Ghadiri, N. Text Summarization in the Biomedical Domain. arXiv 2019, arXiv:1908.02285. [CrossRef]

4. Wang, M.; Wang, M.; Yu, F.; Yang, Y.; Walker, J.; Mostafa, J. A systematic review of automatic text summarization for biomedical literature and EHRs. *J. Am. Med. Inform. Assoc.* 2021, 28, 2287–2297. [CrossRef]

5. Chaves, A.; Kesiku, C.; Garcia-Zapirain, B. Automatic Text Summarization of Biomedical Text Data: A Systematic Review. *Information* 2022, 13,

393. [CrossRef]

6. Moradi, M. Small-world networks for summarization of biomedical articles. arXiv 2019, arXiv:1903.02861.

7. Moradi, M.; Dashti, M.; Samwald, M. Summarization of biomedical articles using domain-specific word embeddings and graph ranking. *J. Biomed. Inform.* 2020, 107, 103452. [CrossRef]

8. Mridha, M.F.; Lima, A.A.; Nur, K.; Das, S.C.; Hasan, M.; Kabir, M.M. A Survey of Automatic Text Summarization: Progress, Process and Challenges. *IEEE Access* 2021, 9, 156043–156070. [CrossRef]

9. Awasthi, I.; Gupta, K. Natural Language Processing (NLP) based Text Summarization—A Survey. In Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 20–22 January 2021; ISBN 978-1-7281-8501-9.

10. Manish, S.; Disha, M. Techniques and Research in Text Summarization—A Survey. In Proceedings of the 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 4–5 March 2021.

11. Gulden, C.; Kirchner, M.; Schüttler, C.; Hinderer, M.; Kampf, M.; Prokosch, H.-U.; Toddenroth, D. Extractive summarization of clinical trial descriptions. *Int. J. Med. Inform.* 2019, 129, 114–121. [CrossRef] [PubMed]

12. Alsentzer, E. Extractive Summarization of EHR Discharge Notes. arXiv 2018, arXiv:1810.12085v1.

13. Kaur, M.; Mollá, D. Supervised Machine Learning for Extractive Query Based Summarisation of Biomedical Data. In Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), Brussels, Belgium, 31 October 2018; pp. 29–37.

14. Fiszman, M.; Rindflesch, T.C.; Kilicoglu, H. Summarizing drug information in Medline citations. *AMIA Annu. Symp. Proc.* 2006, 2006, 254–258.

15. Sackett, D.L. Evidence-based medicine. *In Seminars in Perinatology;* Elsevier: Amsterdam, The Netherlands, 1997; Volume 21, pp. 3–5.