Lay It On Me. Generating Easy-to-Read Summaries for Non-Experts

Ahmed Soliman and **Marc Wenzlawski** and **Vladislav Yotkov**

Abstract

In this study, we present an extractive-abstractive lay summarization pipeline for biomedical papers aimed at generating accessible summaries for non-experts. To achieve this, we construct a sentence-level dataset optimized for maximizing ROUGE scores, utilizing both lay summaries and full articles. We employ a BERT-based classifier for identifying the most important sentences within each article. The extracted summaries are then input into two abstractive models, Clinical-Longformer and GPT-2, which paraphrase the summaries to enhance readability. We evaluate the performance of our models using the ROUGE metric, along with readability metrics such as Flesch-Kincaid Grade Level (FKGL), Gunning Fog Score, and Automated Readability Index (ARI). We find that a ROUGE-maximizing extractive summarization approach is effective for generating extractive summaries, with the Clinical-Longformer model achieving the best results for combined ROUGE and readability scores. Our approach demonstrates the potential for generating lay-friendly summaries of biomedical papers, bridging the gap between expert knowledge and public understanding.

1 Introduction

2 Related Work

3 Methods and Datasets

3.1 Dataset

The data we used is sourced from biomedical research articles in English published in the Public Library of Science (PLOS) and eLife (Goldsack et al. 2022). The datasets (Tables 3 and 2) contain technical abstracts and lay summaries written by experts, which are part of BioLaySumm2023 shared task (Goldsack et al. 2023).

Dataset	Training	Validation
PLOS	24,773	1,376
eLife	4,346	241

Table 1: PLOS and eLife: number of articles

Dataset	Avg. Sentences	Avg. Tokens
PLOS	300	9,000
eLife	600	14,000

Table 2: PLOS and eLife: Dataset statistics

3.2 Extractor Network

Due to the extreme length of medical articles (e.g., eLife has an average of 600 sentences per article), it is not feasible to pass them directly as input to the abstractive models due to their limited maximum input size:

i. GPT-2 (Radford et al. 2019): 1,024 tokens, and
ii. Clinical-Longformer (Li et al. 2023): 4,096 tokens

To overcome this limitation, we use the BioClinicalBERT (Alsentzer et al. 2019) model, pre-trained on the MIMIC-III dataset (Johnson et al. 2016), to extract the most important sentences from the articles. For that purpose, we cast the extraction summarisation problem as supervised binary classification where the input is a sentence s and the output is a binary label indicating whether the sentence should be included in the summary c or not (i.e., 1 and 0, respectively). Due to the nature of the provided gold summaries (i.e., abstractive and lay), we generate our own sentence-level dataset by applying the ROUGE-maximisation technique (Zmandar et al. 2021; Nallapati, Zhai, and Zhou 2017) on the gold summaries and the whole articles. More formally, for each gold summary sentence s_i^k, we find the sentence s_j^k in article a_k that maximises the ROUGE-2 score between them. We then label s_j^k as 1 and the rest of the sentences in a_k as 0. Because the number of sentences in the articles is much larger than the number of sentences in the gold summaries:

i. We base our extractive binary dataset on both eLife and PLOS data to maximise the number of training samples;
ii. We further resolve the class imbalance problem by random under-sampling the majority class (i.e., 0) to match the number of samples in the minority class (i.e., 1);

Our final extractive dataset consists of 944,234 sentences with a completely balanced class distribution. Data is further split into 80-training, 10-validation and 10-testing datasets in a random stratified manner. We then fine-tune the extractive model with a batch size of 32 and a learning rate of $2e-5$ following the guidance from BERT's authors

(Devlin et al. 2019) and find that the model starts to over-fit beyond 2 epochs (see Figures 1 and 2). We also report high F1 scores of 0.767 and 0.765 on the validation and test sets, respectively.

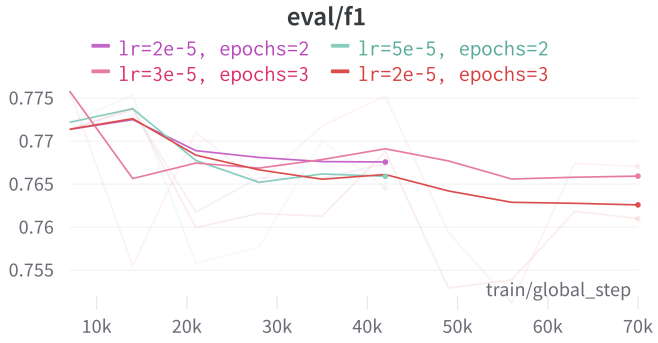

Figure 1: BioClinicalBERT: Evaluation F1

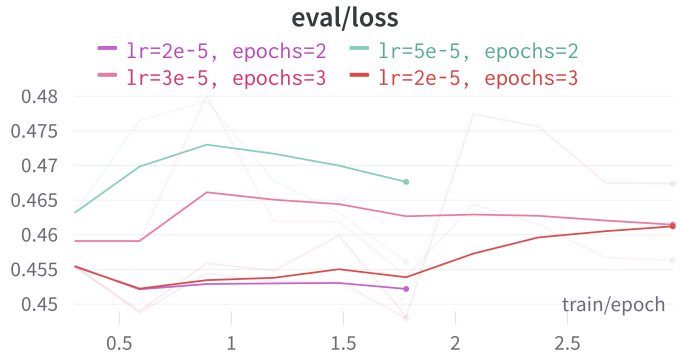

Figure 2: BioClinicalBERT: Evaluation Loss

We then use the BioClinicalBERT model to predict the probability of each sentence in the article being *summarising*. The top 10 sentences with the highest probability are selected and concatenated to produce the final extractive summary. We arrive at this number after analysing the token distribution and finding that 10 sentences is a reasonable number to fit within the maximum input size of the GPT-2 abstractive model (i.e., 1,024 tokens split between the 10 sentences and their lay paraphrases). While we are aware that this can cause the *dangling anaphora phenomenon* (Lin 2009), we use the extracted text only as an intermediate step fed into the abstractive models which paraphrase it into lay language.

3.3 Abstractive Network

Once the extractive summary is generated, we train the abstractive models on the lay summaries and the extractive summaries. For this, we compare two models: GPT-2 (Radford et al. 2019) and Clinical-Longformer (Li et al. 2022). We fine tune both models separately on eLife and PLOS. This is done due to the difference in structure and the average number of tokens in the lay summaries between the two datasets (i.e., 450 and 800 for PLOS and eLife, respectively). Hyperparameters are set based on widely used values in the literature (Li et al. 2022; Radford et al. 2019; Devlin et al. 2019).

3.3.1 Clinical Longformer Abstractor

The Clinical Longformer (Li et al. 2023) is a transformer-based model that is pre-trained on the MIMIC-III dataset (Johnson et al. 2016) and can process up to 4,096 tokens in a single input sequence. This is achieved by the implementation of a sparse attention mechanism that allows more computationally efficient processing of long-range dependencies. We fine-tune the Clinical Longformer as a sequence-to-sequence task on pairs of (a) gold lay summaries and (b) ROUGE-maximising training data described in Section 3.2.

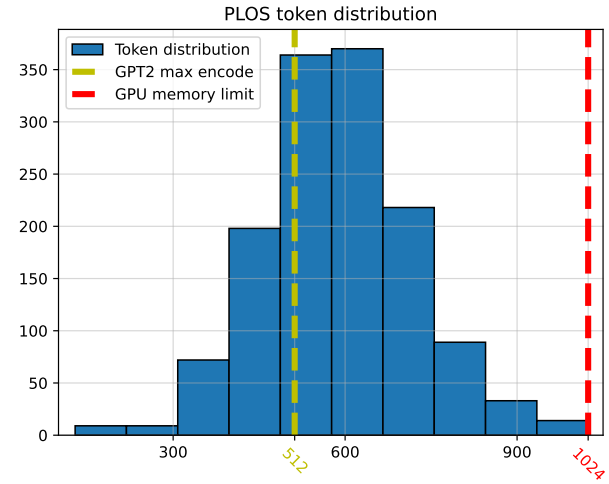

Figure 3: Token Distribution of Extracted Summaries

For the Longformer model, we experimented with window, batch, and input size to ensure that we would not run out of memory during training, as this is a common issue with Longformer models (Orzhenovskii 2021). We found that a window size of 64, batch size of 1, and input size of 1024 worked best for our dataset.

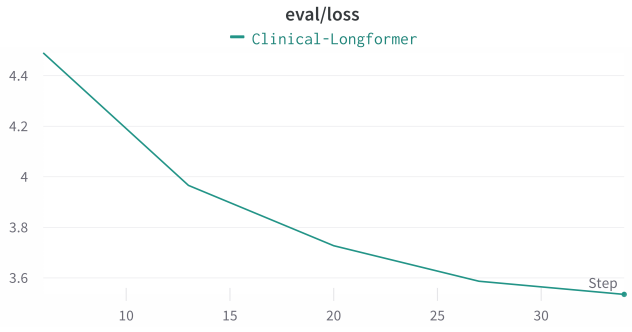

Figure 4: Longformer evaluation loss

Model	Rouge1	Rouge2	RougeL	FKGL	ARI	Gunning
Lexrank	0.334	0.085	0.164	33.59	15.41	18.50
Extractive	0.329	0.0998	0.163	10.6	25.01	26.22
GPT2	0	0	0	0	0	0
Longformer	0.289	0.062	0.143	27.33	16.89	18.44

Table 3: ROUGE and readability statistics.

3.3.2 GPT-2 Abstractor

4 Evaluation

5 Discussion and Conclusion

Bibliography

Alsentzer, Emily, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. "Publicly Available Clinical BERT Embeddings." In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78. Minneapolis, Minnesota, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-1909.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *ArXiv* abs/1810.04805.

Goldsack, Tomas, 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.

Goldsack, Tomas, 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*, 10589–604. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.724.

Johnson, A. E. W., T. J. Pollard, L. Shen, L. W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. 2016. "MIMIC-III, a Freely Accessible Critical Care Database." *Scientific Data* 3: 160035. https://doi.org/10.1038/sdata.2016.35.

Li, Yikuan, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. "Clinical-Longformer and Clinical-BigBird: Transformers for Long Clinical Sequences." https://arxiv.org/abs/2201.11838.

Li, Yikuan, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. "A Comparative Study of Pretrained Language Models for Long Clinical Text." *Journal of the American Medical Informatics Association* 30 (2): 340–47.

Lin, Jimmy. 2009. "Summarization." In *Encyclopedia of Database Systems*, 2906–10. Heidelberg, Germany: Springer.

Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou. 2017. "SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents." In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 3075–81. AAAI'17. San Francisco, California, USA: AAAI Press.

Orzhenovskii, Mikhail. 2021. "T5-LONG-EXTRACT at FNS-2021 Shared Task." In *Proceedings of the 3rd Financial Narrative Processing Workshop*, 67–69. Lancaster, United Kingdom: Association for Computational Linguistics. https://aclanthology.org/2021.fnp-1.12.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners."

Zmandar, Nadhem, Abhishek Singh, Mahmoud El-Haj, and Paul Rayson. 2021. "Joint Abstractive and Extractive Method for Long Financial Document Summarization." In *Proceedings of the 3rd Financial Narrative Processing Workshop*, 99–105. Lancaster, United Kingdom: Association for Computational Linguistics. https://aclanthology.org/2021.fnp-1.19.