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

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

1 Background and Motivation

Scientific publications can be difficult for non-experts to understand, particularly in the biomedical field where misinformation can have direct impacts on health decisions (Islam et al., 2020).

Lay summaries could offer a solution, but they are not yet widely available. Previous studies on automatic text summarisation have not focused on biomedical research due to a lack of data (Chandrasekaran et al., 2020), but two new datasets (PLOS and eLife) have been introduced to address this issue (Goldsack et al., 2022). These datasets will be used in the BioLaySumm2023 shared task, which our team will participate in.

2 Problem Statement

We will outline the different aspects of our problem statement below:

Problem: Develop an abstractive automatic text summarisation system that can generate lay summaries for biomedical research articles, for consumption by non-experts.

Problem Type: The abstractive text summarisation task belongs to the family of Natural Language Generation (NLG) problems.

Techniques: To address the problem, the proposed system will fine-tune pre-trained models on in-domain data, specifically the BioGPT and Clinical-T5 models. The summarisation performance of the system will be evaluated using the ROUGE metric, while the readability quality will be assessed based on the RNPTC.

3 Related Work

The LaySumm task is a shared task that was a part of the CL-SciSumm 2020 shared task series, which aimed to automatically generate lay summaries for scientific articles (Chandrasekaran et al., 2020).

The dataset for the LaySumm task consisted of 572 scientific articles from various disciplines, along with manually written lay summaries provided by the authors of the articles. Participants were required to generate a lay summary of each article, which was evaluated using the ROUGE metric.

A variety of methods were employed by the participating teams, mostly centered on fine tuning PEGASUS (Zhang et al., 2020) or utilizing BART (Lewis et al., 2020). However, due to the nature of the dataset articles, none of the proposed models are specifically tailored to the biomedical domain.

Furthermore, because the biomedical domain is knowledge-intensive, it is essential that pre-training is performed on in-domain data (Jimenez Gutierrez et al., 2022) which in turn yields state-of-the-art performance for various tasks (Gu et al., 2021). Recent examples of such biomedical models are the BioGPT (Luo et al., 2022a) and the Clinical-T5 (Lehman and Johnson, 2023), pre-trained on PubMed [1] and MIMIC (Goldberger, 2000), respectively.

As such, the aim of the project is to explore the effectiveness of using pre-trained biomedical models, specifically BioGPT and Clinical-T5, in generating lay summaries for biomedical research articles given the newly released dataset as illustrated below.

4 Datasets

For this project, the data will be sourced from articles published by the Public Library of Science (PLOS) and eLife (Goldsack et al., 2022). Both datasets consist of biomedical research articles in English along with their technical abstracts and expert-written lay summaries.

The larger of the two datasets is PLOS, which has 24,773 instances for training and 1,376 instances for validation. On the other hand, eLife has a total of 4,346 instances for training and 241

[1]-`https://pubmed.ncbi.nlm.nih.gov`

instances for validation.

It is also worth noting that eLife articles contain longer expert-written lay summaries, which simplify the content to a greater extent compared to PLOS summaries.

5 Evaluation

To compare the performance of our models, similar to (Kim, 2020) and (Chaturvedi et al., 2020) we are going to use the standard summarisation metric ROUGE (Lin, 2004) based on n-gram recall between provided summary and the candidate ones. For that purpose, we will use an open-source implementation of ROUGE, available on Github[2]. As our baseline, we will utlize TextRank[3] and LexRank [4] (Goldsack et al., 2022).

Furthermore, to measure the readability of the produced lay summaries we plan to employ the ranked NP-based text complexity (RNPTC). This metric, introduced by (Luo et al., 2022b), has been proven to outperform significantly traditional readability metrics (e.g., ARI (Smith and Senter, 1967), and Coleman-Liau Index (Coleman and Liau, 1975)).

Alternatively, RNPTC surpasses their shallow features (e.g., sentence length and word characters count) by calculating the weighted sum of the probabilities of scientific jargon words estimated with a BERT (Devlin et al., 2019) pre-trained on general text.

6 Proposed Activities

Our plan of action involves five steps as described in Table 1. Specifically, we intend to fine-tune four models - two for each dataset - using BioGPT and Clinical-T5. Given the size of these models, fine-tuning will be computationally intensive and we will be seeking CSF access from the department. To make sure that we assess the project comprehensively, we will carry out evaluations at different points throughout its progress.

References

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview and insights from the shared tasks at scholarly document processing 2020: CL-SciSumm, LaySumm and LongSumm. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 214–224, Online. Association for Computational Linguistics.

Rochana Chaturvedi, Saachi ., Jaspreet Singh Dhani, Anurag Joshi, Ankush Khanna, Neha Tomar, Swagata Duari, Alka Khurana, and Vasudha Bhatnagar. 2020. Divide and conquer: From complexity to simplicity for lay summarization. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 344–355, Online. Association for Computational Linguistics.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Goldberger. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals.

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.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, Sarkar Mohammad Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, Abrar Ahmad Chughtai, and Holly Seale. 2020. Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4).

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Seungwon Kim. 2020. Using pre-trained transformer for better lay summarization. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 328–335, Online. Association for Computational Linguistics.

[2]-https://github.com/pltrdy/rouge
[3]-https://github.com/DerwenAI/pytextrank
[4]-https://github.com/crabcamp/lexrank

Activity	Any comments	Duration	Lead
Work preparation	Complete CITI Training, aquire CSF access	0.5 Week	All
Data Exploration + Baseline model	Segment the data & Employ TextRank/LexRank Baseline	1 Week	All
Fine Tune BioGPT model	Explore transfer learning hyper-parameters	1.5 Weeks	V & M
Fine Tune Clinical-T5 model		1.5 Weeks	A & V
Evaluate and Compare Models	Use the metrics discussed in the evaluation section	1.5 Week	All

Table 1: Proposed activities

Eric Lehman and Alistair Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text v1.0.0.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022a. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6). Bbac409.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022b. 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.

E A Smith and R. Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*, pages 1–14.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.