Biomedical Lay Summarization

Said Bolluk[1]

[1]CS 549, Ozyegin University

Abstract

Biomedical articles are essential for detecting and curing health-related issues and can benefit from interdisciplinary collaboration. However, their complex and technical nature often makes them challenging for those outside the healthcare industry to understand. To address this, we propose a simplification framework using a transformer-based model to summarize biomedical articles in plain language. We generated summaries from biomedical articles and evaluated their performance based on relevance, readability, and factuality. Fine-tuning significantly improved the model's performance. Our best model showed competitive relevance scores but produced complex and ambiguous summaries. Future work includes post-processing the generated summaries to enhance readability, utilizing domain-specific language models to improve accuracy, and parallelizing the fine-tuning process across multiple machines to reduce computational time.

Introduction

Biomedical articles are invaluable in detecting and curing health-related issues. They can be utilized in interdisciplinary projects where individuals outside the healthcare industry can bring fresh perspectives to these problems. However, the complex format and extensive technical terms often make these articles challenging for people from other fields to understand. This highlights the need for improved interdisciplinary collaboration. Despite the potential benefits of such collaboration, there is still significant room for improvement in the biomedical domain. This is evident from the fact that 80% of references in biomedical articles come from medical education or clinical and health services research journals [1]. This insularity suggests that medical education researchers are predominantly focused on and selective about the sources they incorporate into their academic work.

Simplifying biomedical articles for audiences from different domains can help quickly detect problems and derive more comprehensive solutions. To achieve this, we propose a simplification framework that summarizes biomedical articles in a plain and non-technical manner using a transformer model, a type of deep learning architecture designed explicitly for sequence-to-sequence tasks, such as language translation, text summarization, and language modeling [2]. Two distinct approaches are utilized in the summary generation. The generated summaries are then evaluated based on relevance, readability, and factuality scores against target summaries. In the following chapters, we will explain the materials and methods used to create the simplification framework, discuss the generated summaries' performance, and present this study's limitations and future directions.

Methodology

This section provides details about the material and methods used to create simplification framework for generating lay summaries of biomedical articles. The methodology framework is illustrated in Figure 1.

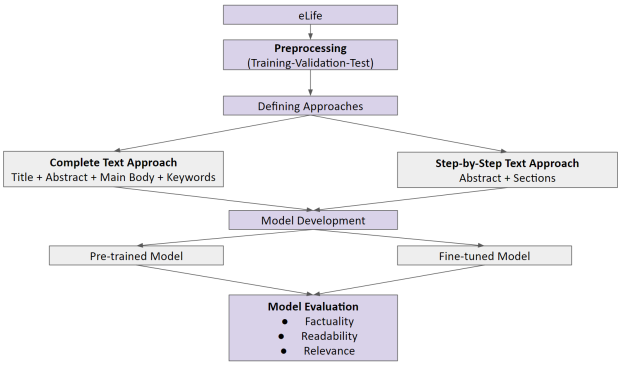

Figure 1: *Methodology workflow*

Dataset Collection: eLife

The primary data source is the sample articles from eLife Magazine. This open-access journal publishes peer-reviewed research across various fields in the life sciences, including biology, medicine, and health sciences. All articles are freely accessible to ensure wide dissemination of high-quality scientific content [3]. The processed articles from the eLife dataset contain biomedical journal articles alongside expert-written lay summaries [4]. This dataset is popularly used in simplification tasks of biomedical articles, such as in [5]. There are 4376 training, 241 validation, and 142 test articles in the complete dataset.

Data Preprocessing

Each biomedical articles is in JSON format with its title, sections, abstract, keywords, and target summary. We performed several preprocessing steps on the biomedical articles in the eLife dataset to ensure standardization and noise reduction:

- First, we tokenized the text using a pre-trained tokenizer, named "facebook/bart-large-cnn" [6], which effectively segmented the articles into individual tokens.
- Next, we converted all text to lowercase to ensure consistency in word representation. We then removed stop words, such as 'and', 'the', and 'is', to eliminate commonly occurring but linguistically insignificant words.
- Finally, we removed special characters, including non-alphanumeric characters, punctuation, and symbols, to enhance the clarity and coherence of the text for subsequent analysis.

Defining Approaches

Summary generation of a biomedical article can be performed in various ways. Different sections of an article contain information about the proposed methodology of that article with distinct levels of detail and number of words. For example, the abstract section is the already compacted version of the entire article with less detail. On the other hand, the methodology can be very complex, with exhaustive equations and terminology. Therefore, each section might contribute to the final lay summary with different weights. With this idea, we developed two approaches to generating lay summaries of biomedical articles.

1) Complete Text Approach: The Complete Text approach is the baseline approach proposed in this study. We merged each article's sections under a final string named Merged Text. This text includes the title, abstract, and body paragraph, referring to the sub-sections of the article and the keywords.

2) Step-By-Step Approach: In contrast, the Step-By-Step Approach assigns varying importance to different sections of an article. Here, we established an empty string named Main String and appended the plain version of the abstract section. Subsequently, utilizing a predetermined maximum token length for generated summaries, we sequentially summarized each body paragraph subsection and added them to the Main String. Upon incorporating the summary of the last subsection into the Main String, we obtained a condensed version of the original article. A final summarization was then performed on this compacted text to derive the ultimate article summary.

Model Development

To generate summaries from the biomedical articles, we utilized a transformer model called Bidirectional and Auto-Regressive Transformers (BART) [6]. BART operates by utilizing a combination of bidirectional and auto-regressive architectures (Figure 2). In the bidirectional component, it learns to understand context from both directions of a sequence and capture dependencies and relationships between words. Additionally, BART employs an auto-regressive approach during generation, where it predicts the next token in a sequence based on the previously generated tokens. This auto-regressive mechanism enables BART to produce coherent and contextually appropriate outputs. Such skills of BART make it practical for tasks such as text generation, translation, and summarization.

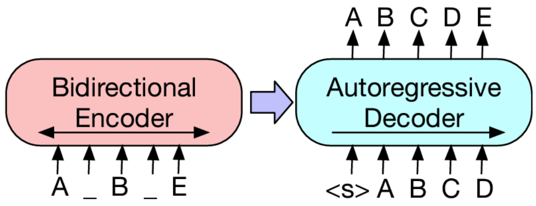

Figure 2: *BART architecture. Retrieved from: [6]*

However, the pre-trained version of the BART model called "facebook/bart-large-cnn", available at the Huggin Face platform, is a generic model that covers many natural language processing (NLP) tasks from various domains. We need a BART model trained explicitly over the biomedical articles to obtain adequate and refined summaries for the biomedical domain. This process, known as fine-tuning, represents a common practice in NLP, and it allows models to adapt to domain-specific nuances and produce more accurate outputs (Figure 3).

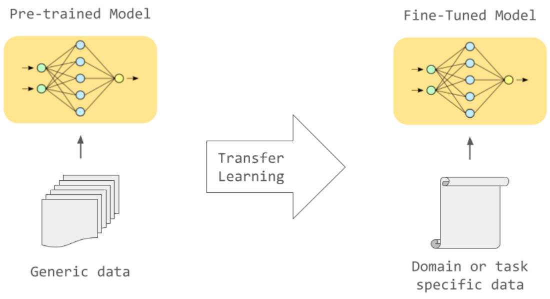

Figure 3: *Fine-tuning illustration. Retrieved from: [7]*

Hyperparameter Tuning

Before tuning the pre-trained BART model using biomedical articles, we must determine its optimal hyperparameters. BART is a large transformer-based model with numerous parameters, including the weights and biases the model learns during training. Hence, training a BART model for generating summaries can be time-consuming. In this sense, we only evaluated three different learning rates using ten epochs for training.

We utilized the Complete Text approach when training the model. The merged text of each article served as the input features, while the target summary served as the output label during training. In each epoch, the BART model was fitted to the training data, and summaries were generated utilizing the merged text from the validation data. These generated summaries were then compared to the target labels of the validation data to compute the Cross-Entropy Loss, which assessed their similarity. Notably, our hyperparameter tuning treated the task as a classification problem. As illustrated in Figure 4, the optimal learning rate identified through this process was $5e-06$ after training for six epochs.

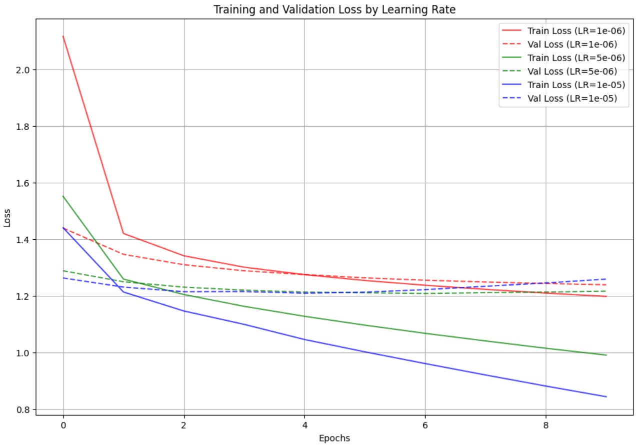

Figure 4: *Hyperparameter tuning of the BART model*

Fine-Tuning

After completing the hyperparameter tuning process, we retrained the BART model using the optimal hyperparameter settings. Subsequently, we saved the resulting fine-tuned model in our local directory for summary generation tasks, employing the Complete Text and Step-By-Step approaches.

Model Evaluation

We generated summaries for the test articles using the pre-trained and fine-tuned models and the Complete Text and Step-By-Step approaches. The generated summaries must be observed in terms of relevance, readability, and factuality metrics against the target summaries. The evaluation of the generated summaries is basically a classification evaluation, which can be performed using multiple scores for each metric, such as precision, recall, and F-Score. For simplicity, we focused merely on the F-Score when determining the evaluation scores. The relevance metric measures the ability of the summary to capture the core content from the original research source articles. We evaluated the relevance scores of the generated articles using Rouge-1, Rouge-2, Rouge-L, and BERT Scores:

- **Rouge (Recall-Oriented Understudy for Gisting Evaluation)**: It calculates overlap measures for unigram (R-1), bigram (R-2), and longest common subsequence (R-L) between the generated summary and target summary.
- **BERTScore**: This metric uses embeddings from a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model to estimate the semantic similarity between the generated summary and target summary.

The subsequent evaluation metric focuses on readability, which measures the comprehensibility of a piece of text for a reader. We only employed the Flesch-Kincaid Grade Level (FKGL) among the various available readability metrics.

- **Flesch-Kincaid Grade Level (FKGL)**: This measures the readability of the text based on the average number of syllables per word and the average number of words per sentence.

The final evaluation metric is related to factuality. The factuality score of a summary emphasizes the accuracy of information conveyed in the generated summary and its alignment with the original summary. We utilized the BART Score to evaluate the factuality of the generated summaries:

- **BART Score**: BART Score measures how well the generated summary captures the factual information present in the original text by comparing the contextualized word and sentence embeddings, which are learned by a pre-trained BART model, in the generated and original texts.

We assessed the quality of the generated summaries using the evaluation metrics mentioned above. Higher scores were targeted for the relevance and factuality metrics, indicating better alignment with the target summaries and greater accuracy of information. Conversely, lower scores were desired for the readability metric, reflecting simpler and more easily understandable summaries. The scripts for computing these evaluation metrics can be found in the provided reference [8].

The overall evaluation of the tested models are provided in Table 1 and Table 2. From these tables, we see that our best model is the Fine-Tuned Model with the Complete Text Approach based on each evaluation metric. For relevance measures, the best model among the available studies achieved scores of 49.46%, 16.76%, 46.15%, and 87.07% for R-1, R-2, R-L, and Bert Score, respectively [5]. Our best model achieved scores of 47.87%, 15.18%, 20.74%, and 66.02% for R-1, R-2, R-L, and Bert Score. While our model demonstrated competitiveness against current studies in the Biomedical Lay Summarization literature [5], there is still room for improvement in our approach. Here, increasing the training examples for fine-tuning might help increase the relevance score.

Model	**R-1 (%)**	**R-2 (%)**	**R-L (%)**
Pretrained Complete	40.20	8.70	16.30
Pretrained SBS	37.00	7.30	15.70
Finetuned Complete	47.90	15.20	20.70
Finetuned SBS	46.00	14.10	19.60

Table 1: *Model evaluation Part-1*

In readability evaluation, the best model among the available studies provided an FKGL of 10.70 [5], while our best model provided an FKGL of 13.41. This indicates that our generated summaries are complex and may require a higher level of education to be fully comprehended. Post-processing the generated articles with simplifying statements for complex terms can help increase the readability score.

Lastly, our model received a BART Score of -2.34, which is significantly lower than the best-reported score of -0.83 in the leading studies [5]. This indicates potential inaccuracies in our generated summaries. We might have removed the technical terms excessively to generate simplified summaries while aiming to increase the readability and relevance scores. However, ensuring accuracy is crucial for generating academic summaries. Increasing the training data size with more biomedical articles might enable generating more accurate summaries.

Model	**BERTs (%)**	**FKGL**	**BARTs**
Pretrained Complete	59.10	14.41	-3.45
Pretrained SBS	56.70	13.50	-3.50
Finetuned Complete	66.00	13.41	-2.40
Finetuned SBS	64.60	13.82	-2.47

Table 2: *Model evaluation Part-2*

Overall, fine-tuning improved the performance of the pre-trained model in summary generation. However, our Step-By-Step Approach could not outperform the Complete Text Approach. Summarizing the entire text at once might have better preserved the context and coherence compared to doing it section by section. Thus, the Complete Text Approach seems to capture the hierarchical structure and relationships between sections more effectively.

Discussion

A significant limitation of our work is the exclusion of the PLOS dataset in model training due to computational and time constraints. The PLOS dataset contains over two thousand training articles [5], which could potentially improve fine-tuning and enhance the quality of the generated summaries. However, incorporating these articles into the fine-tuning process would likely render the project infeasible within the allocated time and resource limits. As shown in Table 3, the total runtime for generating summaries using our current framework exceeded sixteen hours. Therefore, including the PLOS dataset was not considered feasible given these constraints.

Steps	**Runtime (mins)**
Preprocessing	5
Hyperparameter Tuning	360
Fine Tuning	72
Summary Generation	535
Model Evaluation	20
Total	992

Table 3: *Computational runtimes for each process*

Based on the insights gained from this study, several future works can be pursued to further enhance the quality and efficiency of summary generation for biomedical texts. For example, employing a Named Entity Recognition (NER) model, such as Stanza, as a post-processor can help further simplify and clarify generated summaries by accurately identifying and categorizing entities within the text. This can improve the readability and relevance of the summaries by highlighting key terms and concepts.

Additionally, utilizing language models specifically trained on biomedical texts, such as BioBART ([9]) or BioGPT ([10]), can significantly enhance the factuality of the generated summaries. These models are better equipped to understand and process biomedical literature's complex terminology and context-specific nuances.

Finally, implementing parallelization techniques to distribute the training process across multiple devices or machines can substantially reduce computational time. This can allow for more extensive fine-tuning using additional training data and enable the capture of the domain-specific patterns without being constrained by time and computational limitations. We can develop more efficient and effective summarization techniques for biomedical articles by pursuing these future works.

References

[1] Mathieu Albert et al. "Barriers to cross-disciplinary knowledge flow: The case of medical education research". In: *Perspectives on medical education* 11.3 (Oct. 2021), pp. 149–155. DOI: 10.1007/s40037-021-00685-6. URL: https://doi.org/10.1007/s40037-021-00685-6.

[2] Ashish Vaswani et al. "Attention is all you need". In: *arXiv (Cornell University)* (Jan. 2017). DOI: 10.48550/arxiv.1706.03762. URL: https://arxiv.org/abs/1706.03762.

[3] Stuart Rf King, Emma Pewsey, and Sarah Shailes. "An inside guide to eLife digests". In: *eLife* 6 (Mar. 2017). DOI: 10.7554/elife.25410. URL: https://doi.org/10.7554/elife.25410.

[4] Tomas Goldsack et al. "Making science simple: Corpora for the lay summarisation of scientific literature". In: *arXiv preprint arXiv:2210.09932* (2022).

[5] Tomsa Goldsack et al. "Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles". In: *arXiv preprint arXiv:2309.17332* (2023).

[6] Mike Lewis et al. "BART: Denoising Sequence-to-Sequence Pre-training for natural language generation, Translation, and Comprehension". In: *arXiv (Cornell University)* (Jan. 2019). DOI: 10.48550/arxiv.1910.13461. URL: https://arxiv.org/abs/1910.13461.

[7] Akshit Mehra. *Fine-Tuning tutorial: Falcon-7B LLM to a General purpose chatbot*. Feb. 2024. URL: https://www.labellerr.com/blog/hands-on-with-fine-tuning-llm/.

[8] TGoldsack. *GitHub - TGoldsack1/BioLaySumm2023-evaluation$_s$cripts*. URL: https://github.com/TGoldsack1/BioLaySumm2023-evaluation_scripts.

[9] Hongyi Yuan et al. "BioBART: Pretraining and evaluation of a Biomedical Generative Language model". In: *arXiv (Cornell University)* (Jan. 2022). DOI: 10.48550/arxiv.2204.03905. URL: https://arxiv.org/abs/2204.03905.

[10] Renqian Luo et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining". In: *Briefings in bioinformatics* 23.6 (Sept. 2022). DOI: 10.1093/bib/bbac409. URL: https://doi.org/10.1093/bib/bbac409.