CPSC477/577-SP2024 Final Proposal: Biomedical Lay Summarization

Xincheng Cai
Department of Graduate School of Arts and Sciences
Yale University
New Haven, CT, 06510
xincheng.cai@yale.edu

Mengmeng Du
Department of Graduate School of Arts and Sciences
Yale University
New Haven, CT, 06510
mengmeng.du@yale.edu

Abstract

1 Problem Identification and Motivation

This project is inspired by the shared task of the BIONLP 2024 workshop : SHARED TASK ON THE LAY SUMMARIZATION OF BIOMEDICAL RESEARCH ARTICLES. This year, the workshop is particularly focused on the transparency of the generative approaches and factuality of the generated text. The primary motivation of this shared task is to generate summaries of biomedical papers for non-expert audiences, possibly through adding more background information and translating technical terminologies. To achieve this, we are considering the deployment of transformer-based models, such as RAG (Retrieval-Augmented Generation) and BART (Bidirectional and Auto-Regressive Transformers)(1). The RAG model integrates the benefits of both parametric and non-parametric memory systems, enhancing its performance in natural language processing tasks. Furthermore, we plan to evaluate the effectiveness of non-transformer models as a comparative benchmark to assess their performance against transformer-based approaches.

2 Methodology

2.1 Datasets

We are provided with two datasets, derived from PLOS and eLife biomedical journals. Datasets are pre-splitted for model training and validation. Each file is provided in .jsonl format, where each line is a JSON object with the following fields:

```
{
    "lay_summary": string, # article lay summary
    "article": string,     # article main text (abstract included)
    "headings": list,      # the article headings
```

Preprint. Under review.

```
    "keywords": list,      # keywords describing the topic of the article
    "id": string,          # article id
}
```

where the lay_summary field is the reference summary, the article field is the input to the model, and the headings, keywords, and id fields containing metadata (for optional usage). Our aim is to generate lay summaries for each article that are more readable for non-expert audiences. The “lay_summary" field will be the reference during training and validation processes.

For PLOS, there are 24,773 instances allocated for training and 1,376 for validation, while eLife comprises 4,346 instances designated for training and 241 for validation. The test data instances do not have lay summaries as reference, and each test dataset has 142 instances. If necessary, we may apply additional data (i.e., not provided by the competition) at training or inference time, so that we may improve the generalizability of our model.

2.2 Neural-NLP Approaches

The RAG model combines the advantages of both parametric and non-parametric memory systems to enhance natural language processing tasks. Components of the RAG Model Adapted for the Task:

Retriever (Non-Parametric Memory): The Dense Passage Retriever (DPR) would be tailored to pull from a dense vector index of biomedical literature. This adaptation ensures access to relevant, expert-reviewed content that can be translated into simpler terms for lay summaries. The retriever’s job is to access this external, non-parametric memory to fetch contextually appropriate background information and explain complex terms in a way that’s accessible to non-experts.

Generator (Parametric Memory): The generative component remains a pre-trained seq2seq transformer, such as BART, and it’s fine-tuned to prioritize the generation of lay summaries. Upon receiving the input article along with the documents retrieved, the generator works to synthesize both sets of information. It uses its trained parametric memory in conjunction with the context from the retrieved documents to create a summary that’s both informative and easy for non-experts to understand. The goal is to distill complex scientific information into summaries that are engaging, informative, and accessible(2).

2.3 Experimental Plan

1. Data Prepossessing: Checking the missing value to ensure that each article is paired with a lay_summary for supervised learning.

2. Model Configuration: Adjust the RAG model to use the article text as input and generate outputs that will be compared against the lay summaries. Consider utilizing headings and keywords for additional context retrieval or to inform the generation process, enhancing the model’s ability to produce relevant and understandable summaries.

3. Training: Train the model on the provided dataset, using the lay summaries as the ground truth. The training process should minimize the difference between the generated summaries and the reference lay summaries, improving the model’s accuracy and readability over time.

4. Testing: Use the testing split of the dataset to evaluate the model’s performance, focusing on metrics in the evaluation part that measure relevance, readability, and factuality.

5. Model Comparison: Compare the RAG model performance with the other model for example the BiLSTM networks model which can focus on relevant parts of the article when generating the summary of biomedical articles.

2.4 Evaluation

According to the information given by the workshop, all the submitted results will be evaluated based on three perspectives: (1) relevance, (2) readability, and (3) factuality. Therefore, we select the following evaluation metrics corresponding to the three perspectives in Table 1. ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a package that allows automatic measures to determine the quality of a summary by comparing it to other ideal human-written summaries(3). Specifically, ROUGE-1 and ROUGE-2 belong to the ROUGE-N metrics, which refer to the overlap of unigrams and bigrams between the candidate summaries and the reference summaries. ROUGE-L is based on the Longest Common Subsequence (LCS) to evaluate sentence-level structure similarity. As for readability evaluation, we want to highlight two standard metrics: Flesch-Kincaid grade level (FKGL)(4) and Coleman-Liau index (CLI)(5), which have the following formula:

Flesch-Kincaid grade level

$$0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59,$$

Coleman-Liau index

$$0.0588L - 0.296S + 15.8,$$

where L is the average number of letters per 100 words and S is the average number of sentences per 100 words. These readability metrics estimate the year of education needed to understand the generated texts, and lower scores mean higher readability. Finally, the factuality metrics, AlignScore and SummaC, are more recently introduced to detect inconsistency in summarization.

Table 1: Evaluation Metrics

	Evaluation Metrics
Relevance	ROUGE (1, 2, and L) and BERTScore
Readability	FKGL and DCRS, CLI, and LENS
Factuality	AlignScore, SummaC

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 9459–9474, Curran Associates, Inc., 2020.

[2] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019.

[3] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81, 2004.

[4] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom, "Flesch-kincaid grade level," *Memphis: United States Navy*, 1975.

[5] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring.," *Journal of Applied Psychology*, vol. 60, no. 2, p. 283, 1975.