Research Paper Summarizer

Afnan Mir
amm23523
afnanmir@utexas.edu

Jaykumar Patel
jnp2369
patel.jay4802@utexas.edu

Abstract

In this report, we aim to improve accessibility of information presented in complex research papers by using Natural Language Processing (NLP) methods to automatically generate summaries of research papers that are easier to read than the average technical abstract. A dataset of research papers in the medical field which have a technical abstract as well as a plain language summary associated with each paper was used to finetune a pretrained T5 model. The research paper and the technical abstract were used as inputs to the model, and the plain language summary was used as the target output. The finetuned model was evaluated using the ROUGE metric for summarization as well as various readability metrics. Our code can be found here.

1 Introduction

Research papers are groundbreaking in terms of presenting new findings, theories, and perspectives, making them essential for scientific breakthroughs. However, the knowledge presented in research papers is often inaccessible to the general public for various reasons. Reasons include putting research papers behind paywalls, using dense technical language that the average person cannot understand, and sometimes simply being far too long. This discourages those in the general public who may be interested in the topic of the research paper or those who would like to enter the research area presented in the paper from reading the paper. Though the first issue is not one that can be solved without a change in the current publishing system, the latter two issues can be addressed by using Natural Language Processing (NLP) methods to automatically generate summaries of research papers that are easier to read than the average technical abstract. This would involve finetuning a pretrained model on a dataset to convert the paper and its technical abstract into a plain language summary, which, to some extent, would take down the barrier of technical language and length.

2 Dataset and Resources

2.1 Dataset

To train and evaluate the model, a dataset of research papers in the biomedical domain was used. This dataset contains 28,124 research papers, each with an associated technical abstract and a plain language summary. The technical abstract is the abstract that is typically found in research papers, but the plain language summary is a summary that is written by the author, where they are required to highlight how the work fits into a broader context in a simple manner without complex acronyms and terminology (Luo et al., 2022). The dataset was split into the train set, the validation set, and the test set, with 26,124, 1,000, and 1,000 papers respectively. The dataset can be found here.

2.2 Model Resources

To obtain the resources to load our pretrained model and finetune it, HuggingFace was used, which is an online hub where users and/or companies are able to publish their trained PyTorch or TensorFlow models for others to use. HuggingFace provides a high-level API that allows us to load the model, load the dataset in a proper format, and train the model without worrying about the inner workings of the model.

In order to train the model efficiently, training on a local machine would not suffice. For this purpose, Google Colab was used, where the access of a GPU is provided for free for some limited amount of time.

3 Translation of non-English Terms

It is also advised to supplement non-English characters and terms with appropriate transliterations and/or translations since not all readers understand all such characters and terms. Inline transliteration or translation can be represented in the order of: original-form transliteration “translation”.

4 Length of Submission

The NAACL-HLT 2019 main conference accepts submissions of long papers and short papers. Long papers may consist of up to eight (8) pages of content plus unlimited pages for references. Upon acceptance, final versions of long papers will be given one additional page – up to nine (9) pages of content plus unlimited pages for references – so that reviewers’ comments can be taken into account. Short papers may consist of up to four (4) pages of content, plus unlimited pages for references. Upon acceptance, short papers will be given five (5) pages in the proceedings and unlimited pages for references. For both long and short papers, all illustrations and tables that are part of the main text must be accommodated within these page limits, observing the formatting instructions given in the present document. Papers that do not conform to the specified length and formatting requirements are subject to be rejected without review.

NAACL-HLT 2019 does encourage the submission of additional material that is relevant to the reviewers but not an integral part of the paper. There are two such types of material: appendices, which can be read, and non-readable supplementary materials, often data or code. Do not include this additional material in the same document as your main paper. Additional material must be submitted as one or more separate files, and must adhere to the same anonymity guidelines as the main paper. The paper must be self-contained: it is optional for reviewers to look at the supplementary material. Papers should not refer, for further detail, to documents, code or data resources that are not available to the reviewers. Refer to Appendix A and Appendix B for further information.

Workshop chairs may have different rules for allowed length and whether supplemental material is welcome. As always, the respective call for papers is the authoritative source.

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

Preparing References:
Include your own bib file like this:
`\bibliographystyle{acl_natbib}`
`\bibliography{naaclhlt2019}`

where `naaclhlt2019` corresponds to a naaclhlt2019.bib file.

References

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization.

A Appendices

Appendices are material that can be read, and include lemmas, formulas, proofs, and tables that are not critical to the reading and understanding of the paper. Appendices should be **uploaded as supplementary material** when submitting the paper for review. Upon acceptance, the appendices come after the references, as shown here. Use `\appendix` before any appendix section to switch the section numbering over to letters.

B Supplemental Material

Submissions may include non-readable supplementary material used in the work and described in the paper. Any accompanying software and/or data should include licenses and documentation of research review as appropriate. Supplementary material may report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported in the paper. Seemingly small preprocessing decisions can sometimes make a large difference in performance, so it is crucial to record such decisions to precisely characterize state-of-the-art methods.

Nonetheless, supplementary material should be supplementary (rather than central) to the paper. **Submissions that misuse the supplementary material may be rejected without review.** Supplementary material may include explanations or details of proofs or derivations that do not fit into the paper, lists of features or feature templates, sample inputs and outputs for a system, pseudo-code or source code, and data. (Source code and data should be separate uploads, rather than part of the paper).

The paper should not rely on the supplementary material: while the paper may refer to and cite the supplementary material and the supplementary material will be available to the reviewers, they will not be asked to review the supplementary material.