SimpleText Task 1 Report

Ben Gaudreau and Gabrielle Akers

University of Southern Maine, Portland ME 04101, USA

Abstract. Given the task of text simplification and passage retrieval as outlined in the SimpleText@CLEF 2024 Lab, we summarize the challenges of text simplification. We then discuss previous models submitted to earlier iterations of the SimpleText Lab, and analyze their structure, along with their strengths and weaknesses. With this information, we then propose a model that is both capable of providing decent results while also being technically feasible as relatively new students of machine learning topics.

1 The Task At Hand

The SimpleText@CLEF 2024 Lab is focused on the problem of simplifying academic resources for general audiences. In a time where the Internet has allowed vast quantities of information to reach the public, the complex language and concepts of scientific research continues to be a limiting factor[1]. This lab explores the methods by which these texts may be simplified and rewritten for greater accessibility.

Task 1 of the SimpleText Lab presents a deceptively tricky question: **What is in (or out)?** Given a topic and a query, the model must provide a relevant summary from abstracts in the corpus[1]. The accuracy, complexity, and credibility of each referenced resource plays a factor in the overall quality of the result.

The SimpleText Lab has been iterated upon several times over the last few years, providing information that will be of use when constructing a model to perform the task. These past examples will be discussed in Section 2, and our approach with respect to previous iterations will be detailed in Section 3.

2 Previous Approaches

Many models from previous iterations of this lab have utilized natural language processing (NLP) models such as BERT[2, 3]. While these have historically performed rather well, other submissions in previous years have opted to use other models, to varying levels of success. Most notably, the models provided by Elsevier[4] in 2023, utilizing generative pseudo-labeling (GPL), showed stronger results compared to those that did not.

Other NLP approaches not specific to the SimpleText Lab have demonstrated other methods of passage selection that may be of use when consider a new model. For example, techniques implementing control tokens in a model[5] can provide additional context to a retrieved passage's overall complexity, which could play a role in determining the better of two similarly relevant passages. Similarly, methods of dense passage retrieval (DPR) can be incorporated into a neural models to further improve the speed and accuracy of retrieval over traditional methods[6].

Table 1. Various strengths and weaknesses of models as they relate to the task.

Model	Strengths	Weaknesses
Bi-encoder models[7]	Easy to implement, numerous fine-tuned models for various tasks	Varied performance across different testing conditions
GPL[4]	Strong performance across different testing conditions	Generative inaccuracy present, more difficult to implement
ChatGPT[8]	Can perform simplifications and translations together, can be implemented on top of another model	Requires careful pre-processing, possibility of hallucinations

3 Our Approach

Our proposed model will attempt to incorporate elements of bi-encoder models alongside other techniques to counteract the previously mentioned limitations. Specifically, a ColBERT model[9] used alongside DPR should provide a solid foundation upon which further improvements may be made. Our decision comes from our relative inexperience in programming machine learning tasks, so we plan to use this as an opportunity to develop our skills in this area of computer science.

The methods by which we will compensate for the inaccuracies displayed by bi- and cross-encoder models over subsets of the testing data, as demonstrated in the results of last year's SimpleText submissions[1], is currently unclear. We will look into solutions as we begin our development of the model.

4 Conclusion

In this report, we have laid out the task required of our model as stated by the SimpleText@CLEF 2024 Lab. By analyzing prior attempts at this task, we have furthered our own understanding of the problem and our methods of solving it. Ultimately, we have decided to build off existing work with bi- and cross-encoders and apply them in different structures in the hopes of achieving a better result. Regardless whether or not we succeed in this endeavor, the process of working through this lab will have developed our understanding of machine learning topics overall.

References

[1] Eric SanJuan et al. *Overview of the CLEF 2023 SimpleText Task 1: Passage Selection for a Simplified Summary*. 2023.

[2] Roos Hutter et al. *University of Amsterdam at the CLEF 2023 SimpleText Track*. 2023.

[3] Behrooz Mansouri et al. *AIIR and LIAAD Labs Systems for CLEF 2023 SimpleText*. 2023.

[4] Artemis Capari et al. *Elsevier at SimpleText: Passage Retrieval by Fine-tuning GPL on Scientific Documents*. 2023.

[5] Sweta Agrawal and Marine Carpuat. *Controlling Pre-trained Language Models for Grade-Specific Text Simplification*. 2023. arXiv: `2305.14993 [cs.CL]`.

[6] Vladimir Karpukhin et al. *Dense Passage Retrieval for Open-Domain Question Answering*. 2020. arXiv: `2004.04906 [cs.CL]`.

[7] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: `1810.04805 [cs.CL]`.

[8] Björn Engelmann et al. *Text Simplification of Scientific Texts for Non-Expert Readers*. 2023.

[9] Carlos Lassance et al. *A Study on Token Pruning for ColBERT*. 2021. arXiv: `2112.06540 [cs.IR]`.