UZH_Pandas at SimpleText2024: Multi-Prompt Minimum Bayes Risk with Diverse Prompts

Notebook for the SimpleText Lab at CLEF 2024

Andrianos Michail[1,*,†], Pascal Severin Andermatt[1,†] and Tobias Fankhauser[1]

[1] University of Zurich, Zurich, Switzerland

Abstract

This paper serves as a summary of further experiments of the paper "SimpleText Best of Labs in CLEF-2023: Scientific Text Simplification Using Multi-Prompt Minimum Bayes Risk Decoding" [1], adapted to the SimpleText2024 Shared Task 3.1 dataset. We observe how candidate simplifications generated by the off-the-shelf Llama3 perform differently depending on the prompt, and whether Minimum Bayes Risk (MBR) re-ranking is beneficial with underperforming candidates. Finally, on a small sample, we investigate the agreement of simplification candidate re-rankings between MBR and a human annotator.

Keywords

Scientific Text Simplification, Generative Language Models, Minimum Bayes Risk Decoding, Multi Prompt Ensembling, Prompt Engineering, Large Language Models, SimpleText@CLEF-2024

1. Introduction

Automatic simplification of complex text and, even more precisely, scientific abstracts, remains challenging. While LLMs have been shown to be adequate for text simplification, there appears to be a large variation in performance across different domains and prompting strategies [2]. We present the extended results of the further evaluations of the paper [1] on the SimpleText2024 shared task [3]. Our main contribution in this summary is to report the results of different prompting strategies in the test set and to examine the agreement between the Minimum Bayes Risk re-ranking choices and the candidate selected by a human.

2. Methodology

We perform the simplifications with off-the-shelf Llama3 [4] 8B model, using the prompts in Table 1. Further to the plain prompts, we also experiment with variations of the prompts where we provide the simplification model with intermediate definitions of complex terms during inference.

Table 1
The plain prompt templates used to generate the simplifications.

Target	Prompt
P1: General	Simplify the following scientific sentence to make it more understandable for a general audience:
P2: 5Y	Simplify the following scientific sentence. Explain it as if you were talking to a 5-year-old, using simple words and concepts:

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*] Corresponding author.

[†] These authors contributed equally.

andrianos.michail@cl.uzh.ch (A. Michail); pandermatt@ifi.uzh.ch (P. S. Andermatt); tobias.fankhauser@outlook.de (T. Fankhauser)

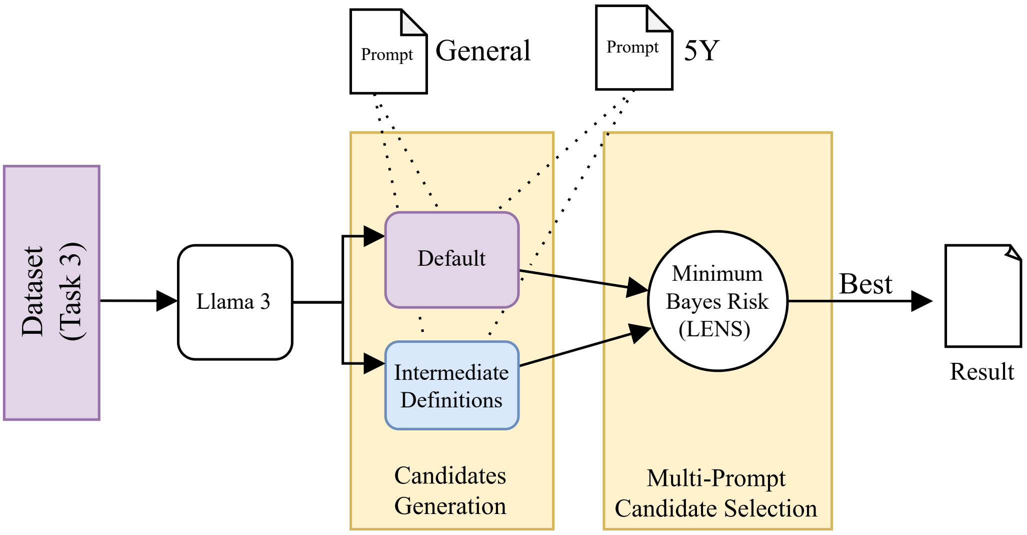

Figure 1: Complete schematic of the Simplification pipeline. For extended details, refer to [1]

These definitions are generated by the same LLM in a separate session. We refer to the simplifications generated with this approach as being generated through *Intermediate Definitions (ID)*.

We ablate by selecting the best candidate using Minimum Bayes Risk [5, 6, 7] with LENS [8] as the utility function results in better performance. The complete schematic is illustrated in Figure 1.

3. Results

In Table 2 we show the simplification evaluations of each individual prompt, together with the evaluations of simplifications selected by Minimum Bayes Risk. The evaluation metrics generally agree on the ranking of the systems. The clear exception is that the simplifications receive exceptionally high FKGL [9] when the model is prompted by *Intermediate Definitions (ID)*, because the definitions are defined within the sentence. However, due to the extremely low FKGL score of the 5Y prompt, we know that the model is over-simplifying the text, probably omitting some important details of the source text. The limitation of these prompts is also reflected in the SARI [10], demonstrating its appropriateness as an evaluation metric.

Contrary to previous results [1], simplifications selected by Minimum Bayes Risk received worse ratings than the two best performing prompts. We hypothesize that this is due to the overshooting of simplifications generated by the 5Y prompt, which acts as a negative utility to select the best candidate, demonstrating the dependency of the approach on the source distribution of candidates.

Table 2
Results of the evaluation of the SimpleText2024 Shared Task, Task 3.1, presented in descending order according to the SARI score. Other participants are omitted for brevity.

run_id	Sample Size	FKGL↓	SARI↑	BLEU↑	Comp. ratio	Sent. splits	Lev. sim.	Ex. copies	Lex. comp.
Reference Texts	578	08.86	100.00	100.00	0.70	1.06	0.60	0.01	8.51
Best Run (Elsevier)	578	10.33	43.63	10.68	0.87	1.06	0.59	0.00	8.39
General	578	11.24	39.28	05.67	0.88	0.98	0.52	0.00	8.45
General through ID	578	21.36	38.29	03.13	1.93	0.99	0.46	0.00	8.86
Minimum Bayes Risk (LENS)	578	07.79	36.72	03.65	0.72	0.98	0.46	0.00	8.25
5Y through ID	578	19.30	36.53	02.27	1.76	1.01	0.45	0.00	8.87
5Y	578	05.94	34.91	02.29	0.66	0.99	0.43	0.00	8.17
Source Texts	578	13.65	12.02	19.76	1.00	1.00	1.00	1.00	8.80

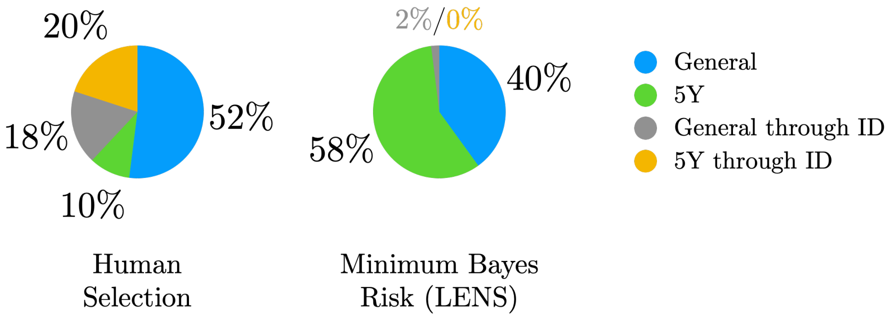

Figure 2: Selection rate for simplification candidates selected through a Human (left) and Minimum Bayes Risk (right).

3.1. Human Preference Selection

We investigate the selection process of Minimum Bayes Risk (LENS) by comparing it to how a human would select the best candidate for simplification.

Out of 50 human annotated selections, we visualize the percentage of examples selected from each source prompt in Figure 2. We see that the human selected about 38% of the simplification candidates generated by intermediate definitions, with the qualitative impression that they improve the clarity of complex terms, making them easier to read. In contrast, Minimum Bayes Risk (LENS) selected mainly (58%) samples from the 5Y prompt, which was the least selected by the human with a selection rate of only 10%, due to the qualitative impression that the candidates lacked important details from the source. In general, the cross-annotator agreement between Minimum Bayes Risk and human selection is quite low, with a Cohen's $\kappa = 0.14$.

4. Limitations

In our study, we only examine the behavior of Minimum Bayes Risk within a limited set of different prompts. In reality, Minimum Bayes Risk using LENS may be limited by the source candidate pipelines or the utility function itself, LENS. Importantly, our human selection annotation study is subjective and performed on a small sample of simplifications.

5. Conclusions

This study extended previous work on scientific text simplification using Multi-Prompt Minimum Bayes Risk re-ranking applied to the SimpleText2024 Shared Task 3 dataset. Our results showed significant differences in performance between prompts, with one prompt leading to oversimplification, and finally we measured the agreement between Minimum Bayes Risk and human selection, including qualitative observations.

Acknowledgments

We express our deepest gratitude and sincere appreciation to Simon Clematide and the Department of Computational Linguistics for their unwavering support, computational resources and constructive guidance during the creation of this work. Andrianos Michail acknowledges funding by the SNSF (213585) under the "impresso 2" project.

References

[1] A. Michail, P. S. Andermatt, T. Fankhauser, Simpletext best of labs in CLEF-2023: Scientific text simplification using multi-prompt minimum bayes risk decoding, in: L. Goeuriot, G. Q. Philippe Mulhem, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, 2024.

[2] T. Kew, A. Chi, L. Vásquez-Rodríguez, S. Agrawal, D. Aumiller, F. Alva-Manchego, M. Shardlow, BLESS: Benchmarking large language models on sentence simplification, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 13291–13309. URL: https://aclanthology.org/2023.emnlp-main.821. doi:10.18653/v1/2023.emnlp-main.821.

[3] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. Di Nunzio, F. Vezzani, J. D'Souza, J. Kamps, Overview of the CLEF 2024 SimpleText track: Improving access to scientific texts for everyone, in: L. Goeuriot, G. Q. Philippe Mulhem, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, 2024.

[4] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[5] S. Kumar, W. Byrne, Minimum bayes-risk word alignments of bilingual texts, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Association for Computational Linguistics, 2002, pp. 140–147. URL: https://aclanthology.org/W02-1019. doi:10.3115/1118693.1118712.

[6] S. Kumar, W. Byrne, Minimum bayes-risk decoding for statistical machine translation, in: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, Association for Computational Linguistics, Boston, Massachusetts, USA, 2004, pp. 169–176. URL: https://aclanthology.org/N04-1022.

[7] M. Müller, R. Sennrich, Understanding the properties of minimum bayes risk decoding in neural machine translation, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 259–272. URL: https://aclanthology.org/2021.acl-long.22. doi:10.18653/v1/2021.acl-long.22.

[8] M. Maddela, Y. Dou, D. Heineman, W. Xu, LENS: A learnable evaluation metric for text simplification, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 16383–16408. URL: https://aclanthology.org/2023.acl-long.905. doi:10.18653/v1/2023.acl-long.905.

[9] R. Flesch, Marks of readable style; a study in adult education., Teachers College Contributions to Education (1943).

[10] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415.