Elaborative Text Simplification via Target Estimation using Large Language Models

Martyna Gruszka
Institute of Science Tokyo
gruszka.m.8d39@m.isct.ac.jp

Yuki Arase
Institute of Science Tokyo
arase@c.titech.ac.jp

Abstract

Text simplification is widely believed to enhance comprehension for non-native readers, and has, therefore, been the focus of extensive research. However, conventional text simplification often involves removing a considerable amount of complex content, which may reduce valuable information and potentially limit opportunities to engage with challenging concepts. In this work, we specifically address the needs of language learners by focusing on elaborative text simplification, a process involving content addition, such as providing specific explanations and clarifications that could make a text comprehensible within its context. We introduce a novel data-driven approach for guided elaboration generation, demonstrating that explicitly specifying elaboration targets leads to improved performance.

1 Introduction

Text simplification is a broad field that focuses on making text content more comprehensible and accessible to a wider audience. It achieves this through various text modifications, such as paraphrasing, word reordering, content deletion, or insertion, while retaining the original meaning [1]. It has applications in improving readability for diverse groups, including children [3], language learners [13] [11], and individuals with language-related disabilities, such as aphasia or dyslexia [2] [14].

This paper focuses on elaborative text simplification [15], which, in contrast to standard simplification methods, focuses exclusively on content addition. Its objective is to enhance the text comprehension by providing readers with additional contextual information. It involves the insertion of various types of clarifications or explanations, including definitions, examples, and background knowledge, to clarify unclear or complex terms and concepts in the text. Table 1 illustrates an example of the elaborative simplification.

Elaborative Simplification
And his wife, Maria, was inspired to get her GED. The general educational development (GED) is equal to a high school diploma. It is for adults who were unable to finish high school. Benito's path is more uncertain. He has not yet registered at the adult school.

Table 1: Example of an elaborative simplification, where the highlighted sentence is inserted as elaboration to provide additional context and clarification.

Early studies on elaborative simplification primarily focused on definition retrieval [10] [5] [8] and the insertion of contextually relevant phrases, often referred to as entity post-modifiers [9]. Recently, a data-driven approach has emerged, treating elaboration generation as a sequence-to-sequence task [15]. Despite these advancements, previous studies have revealed several challenges. In many cases, the generated elaborations diverge from the reference content, either clarifying concepts unrelated to the target concept or addressing entirely different terms [15]. We hypothesize that this limitation arises from the lack of explicit guidance on what to elaborate upon.

To address this issue, we propose a new guided approach for elaboration generation, the **Target-specified Generation**. We identify elaboration targets by prompting the GPT-4o model [7] and subsequently input them into a language model to generate elaboration sentences.

Our experimental results demonstrate that incorporating target information alongside context sentences

enhances model performance. In addition to generating simple definitions, the models are capable of producing elaborations that involve more complex reasoning. To facilitate reproducibility and further research, we make our code publicly available at https://github.com/martgru/ElabSimp

2 Proposed Method

We build our research upon the data-driven experiment presented in the previous work [15]. A model takes as input the context sentences surrounding the elaboration sentence in the simplified document, with a given context window size, while the output consists solely of the elaboration sentence.

To generate meaningful and contextually appropriate elaborations, it is essential to accurately determine the targets of such elaborations. We determine elaboration targets for each instance in our dataset using the GPT-4o model. We prompt the GPT-4o model in ChatML format, utilizing structured output, to identify two key elements from the context sentences surrounding the elaboration sentence:

- **Target sentence**: The sentence in the context that is directly clarified by the elaboration sentence.
- **Target phrase**: The specific phrase within the context sentences that the elaboration sentence explains or provides additional information about.

Table 2 provides an example of an elaboration sentence along with its corresponding target sentence and target phrase, as identified from context by the GPT-4o [7].

In our approach, the input to the model includes not only the context sentences, but also explicitly specified targets extracted by the GPT-4o model: the target phrase, target sentence, or both.

3 Experimental Settings

Dataset In our work, we utilize the annotated version of the Newsela corpus [16] , which contains 1.3K instances of elaborative simplification [15]. For the model inputs, we adopt the original context window sizes defined in previous work [15]:

- C_{2s} : 2 prior context sentences.
- C_{4s}: 4 prior context sentences.
- C_{2s+}: 2 prior and posterior context sentences.
- C_{4s+}: 4 prior and posterior context sentences.

But there's a problem: What scientists find by looking at big cat DNA **doesn't agree with what the fossils tell them**. Scientists are hoping to figure out where big cats first appeared. But the two kinds of evidence don't point to the same place. "If you only looked at the fossil, it would suggest Africa," Tseng said. "If you only looked at DNA, it would suggest Asia."

Table 2: Example of elaborative simplification with specified elaboration targets: the elaboration is highlighted in yellow, the target sentence in blue, and the target phrase is bolded.

Baselines As a baseline, we compare our approach to a previous method that directly generate elaborations from the surrounding context [15]. In addition, we also compare our method to the one that indicates the position within the context text where the generated elaboration should be inserted. In this setting the input to the model consists of context sentences with the position of the elaboration sentence marked by a specialized tag token, i.e., `<explanatory sentence>`.

Generation Models For elaboration generation, we employed LLaMA 3.2 3B model [4] (`meta-llama/Llama-3.2-3B`) available via the Hugging Face library.[1)2)] The model was fine-tuned for 3 epochs with a learning rate of $1e-6$ and a batch size of 32. All instructions provided to the LLaMa model were formatted according to the standard Alpaca format [6]. Elaborations were generated using beam search with 4 beams, ensuring deterministic outputs by avoiding sampling.

Evaluation Metrics Elaboration generation is a relatively new task in the field of text simplification, and as of now, no specific metric has been developed to assess the quality of such content additions. In our work, we adopt the BLEU metric [12], which has been used in previous studies to evaluate elaborations, for the sake of comparison. We also evaluate the generated elaborations using BERTScore [18] and BARTScore [17], which are more capable of capturing semantic similarity and contextual relevance.

1) https://huggingface.co/meta-llama/Llama-3.2-3B
2) We also experimented with the BART-base model; however Llama-3.2 demonstrated better performance.

	BLEU-2 Score					BERT Score					BART Score				
Context	base	pos	p	s	p+s	base	pos	p	s	p+s	base	pos	p	s	p+s
C_{2s}	**9.9**	9.6	8.8	9.4	9.2	0.50	0.50	0.50	0.50	**0.51**	−3.33	−3.34	**−3.29**	−3.31	−3.25
C_{2s+}	8.0	9.6	**10.8**	7.0	10.5	0.48	0.51	0.51	0.48	**0.52**	−3.39	−3.27	**−3.20**	−3.37	−3.21
C_{4s}	9.5	9.6	9.9	10.1	**10.2**	0.50	0.50	0.50	0.50	**0.51**	−3.34	−3.34	**−3.26**	−3.31	−3.28
C_{4s+}	8.3	8.4	**9.3**	6.4	8.4	0.48	0.50	**0.51**	0.47	0.50	−3.36	−3.33	**−3.21**	−3.37	−3.24

Table 3: Results across all settings: **base**: baseline setting where the model receives only context sentences as input, pos: **position-specified**, p: **target phrase**, s: **target sentence**, and p+s: **target sentence with target phrase**. For each context setting, the highest scores for BLEU-2, BERTScore, and BARTScore are bolded.

Method	Input	Output
pos	The search team is in a race against time. The black box's battery will last only 30 to 45 days. **<explanatory sentence>** The odds of finding it in time are low. Searchers will need to find pieces of the plane.	**Predicted:** After that, it will stop working. **Reference:** After that, the signal will go silent.
pos	Munro grew up in Wingham, a traditional Canadian town west of Toronto. She tended to write about her experience there and her life after the social revolution of the 1960s. **<explanatory sentence>**	**Predicted:** The revolution was a time of change. **Reference:** It was a time when people began to break free of traditional roles..
p	It would be able to **download live video**. These would come from remote-controlled airplanes, or drones, flying overhead. Tiny motors in the exoskeleton would give the soldier extra strength. He'd be able to easily run and jump while carrying 100 or more pounds.	**Predicted:** The soldier would see what was happening in real time. **Reference:** This would allow a soldier to know what was happening nearby.
p+s	Then about 6 million years ago another big change occurred. Big cats **split into several different species**.	**Predicted**: They became different kinds of cats. **Reference:** They became lions, tigers, jaguars and leopards..
p	"This is a big one", said Les Grober. He is with the **Water Resources Control Board**. "We get many issues before us. But this is a very dry year."	**Predicted:** It is a state agency. **Reference:** The board decides what to do about the water..

Table 4: Examples of target-specified and position-specified generation outputs: (1) an accurate explanation directly related to the preceding sentence, (2) overly general elaboration, (3) an accurate speculation on the outcome of certain action, (4) an elaboration that fails to provide relevant examples, and (5) an accurate elaboration that addresses a different implicit question. The target sentence is highlighted in blue, and the target phrase in bold.

4 Experimental Results

Final results are presented in Table 3.

The position-specified generation setting demonstrated slight improvements. This approach was especially effective in generating elaborations that involved simple explanations directly related to the preceding sentence, as shown in the first row of Table 4. However, there were also instances where the generated elaborations either lacked meaningful content, or were overly general. An example of this issue is illustrated in the second row of Table 4.

Target-specified generation, particularly the setting involving target phrase, proved to be beneficial. The highest BLEU-2 scores and BARTScores were achieved in this setting. In this configuration, mod-

Method	Input	Output
pos	The Internet helped Mark Zuckerberg start Facebook while still a college student. **<explanatory sentence>**	**Predicted:** Facebook is now one of the most popular websites in the world. **Reference:** He soon became a billionaire.
pos	New companies have come that need skilled workers with more education. New Haven youth want those jobs, but they do not have the education or the skills. **<explanatory sentence>** That is where New Haven Promise comes in. It will make a difference by paying for college.	**Predicted:** They do not know how to get the education they need. **Reference:** Many do not have the money to get the training they need.
p+s	There are many reasons for the population shift. Many different groups of Asian-Americans are moving to the South. The government tracks **23 different ethnic groups** that it considers Asian-American. It's hard to find one explanation for the changes.	**Predicted:** There are many different Asian-American groups. **Reference:** Groups of Chinese, Indians, Koreans and Vietnamese are just some of them.

Table 5: Examples of predicted elaborations showcasing different challenges in evaluation: (1) an elaboration that provides meaningful information but diverges from the reference, (2) an elaboration that achieves high BLEU scores due to word overlap but significantly differs in meaning, and (3) an elaboration that lacks necessary examples but achieves a high BERTScore. The target sentence is highlighted in blue, and the target phrase in bold.

els accurately generated elaborations that included not only simple definitions but also more complex reasoning, such as explaining cause-effect relationships, or speculating on the outcomes of certain actions, as illustrated by the example in the third row of Table 4.

Additionally, the setting combining the target sentence and target phrase achieved the highest BERTScores. The generated elaborations in this setting were similar to those obtained with the target phrase setting, but there were some notable differences. For example, some elaborations, particularly those intended to provide examples of a target term or concept, were abstracted, as shown in the forth row of Table 4.

Key findings from our work indicate that target-specified elaboration generation settings can improve model's performance; however, the improvements were not as significant as we initially hypothesized. While specifying the target phrase often proved to be beneficial, challenges remained in generating elaborations that accurately addressed the specified target. In many instances, the generated elaborations diverged in form and content from the references, as they tended to answer different implicit questions. An example of such elaboration is presented in the fifth row of Table 4.

5 Discussion

Currently, the evaluation of elaborative simplification is constrained by its reliance on reference-based comparisons, where predicted elaborations are evaluated solely against a predefined reference. However, we argue that this approach is not well-suited for a task that involves generating new content. Our analysis revealed many instances where the generated elaborations were relevant and provided high-quality additional information but received low scores because they differed from the reference elaborations. An example of this is shown in the first row of Table 5.

Conversely, there were cases where the generated elaborations achieved high scores by simply mirroring the words in the reference, yet they differed significantly in meaning or failed to provide sufficient information, such as examples of a given concept. These issues are illustrated in the second and third rows of Table 5, respectively.

These limitations highlight the need for a novel evaluation metric specifically designed for content addition tasks, which would account for both the relevance and quality of the added content, eliminating the reliance on reference-based comparisons.

Acknowledgments

We extend our gratitude to Junyi Jessy Li and Yating Wu from The University of Texas at Austin for providing the annotated version of the Newsela corpus. This work was supported by the JSPS KAKENHI Grant Number JP21H03564.

References

[1]Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187, 2020.

[2]John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical simplification of english newspaper text to assist aphasic readers. In *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, 1998.

[3]Jan De Belder and Marie-Francine Moens. Text simplification for children. In *In Proceedings of SIGIR workshop on accessible search systems*, 2010.

[4]Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[5]Soojeong Eom, Markus Dickinson, and Rebecca Sachs. Sense-specific lexical information for reading assistance. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 2012.

[6]Center for Research on Foundation Models (CRFM). Stanford alpaca: An instruction-following model, 2023. https://crfm.stanford.edu/2023/03/13/alpaca.html.

[7]Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[8]Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. A semantic and syntactic text simplification tool for health content. *AMIA Symposium*, 2010:366–70, 2010.

[9]Jun Seok Kang, Robert Logan, Zewei Chu, Yang Chen, Dheeru Dua, Kevin Gimpel, Sameer Singh, and Niranjan Balasubramanian. PoMo: Generating entity-specific post-modifiers in context. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

[10]Gustavo Paetzold and Lucia Specia. Anita: An intelligent text adaptation tool. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, 2016.

[11]Gustavo H. Paetzold. *Lexical simplification for non-native English speakers*. PhD thesis, University of Sheffield, 2016.

[12]Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002.

[13]Sarah E. Petersen and Mari Ostendorf. Text simplification for language learners: a corpus analysis. In *Slate*, 2007.

[14]Luz Rello, Ricardo A. Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP TC13 International Conference on Human-Computer Interaction*, 2013.

[15]Neha Srikanth and Junyi J. Li. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.

[16]Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297, 2015.

[17]Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.

[18]Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.