A Japanese News Simplification Corpus with Faithfulness

Toru Urakawa, Yuya Taguchi, Takuro Niitsuma, Hideaki Tamori
The Asahi Shimbun Company
5–3–2 Tsukiji, Chuo-ku, Tokyo 104–8011 Japan
urakawa-t@asahi.com, taguchi-y2@asahi.com, niitsuma-t@asahi.com, tamori-h@asahi.com

Abstract

Text Simplification enhances the readability of texts for specific audiences. However, automated models may introduce unwanted content or omit essential details, necessitating a focus on maintaining faithfulness to the original input. Furthermore, existing simplified corpora contain instances of low faithfulness. Motivated by this issue, we present a new Japanese simplification corpus designed to prioritize faithfulness. Our collection comprises 7,075 paired sentences simplified from newspaper articles. This process involved collaboration with language education experts who followed guidelines balancing readability and faithfulness. Through corpus analysis, we confirmed that our dataset preserves the content of the original text, including personal names, dates, and city names. Manual evaluation showed that our corpus robustly maintains faithfulness to the original text, surpassing other existing corpora. Furthermore, evaluation by non-native readers confirmed its readability to the target audience. Through the experiment of fine-tuning and in-context learning, we demonstrated that our corpus enhances faithful sentence simplification.

Keywords: corpus, simplification, faithfulness

1. Introduction

Text Simplification (TS) is the modification of a text to make it easier to read and understand while retaining its original meaning. The primary objective of TS is to enhance the text's accessibility to a broader audience (Belder and Moens, 2010; Rello et al., 2013; Devaraj et al., 2021). Among this audience are non-native readers (Yano et al., 1994), and TS is anticipated to aid in conveying daily news or vital topics such as social welfare to them. In such crucial information dissemination, enhancing readability while ensuring faithfulness to the original content is essential.

TS is predominantly approached through a sentence-level sequence-to-sequence (seq2seq) framework, drawing parallels with Natural Language Generation tasks such as machine translation. However, in automatic summarization, considered a type of translation task, there is a noted phenomenon called *hallucination* where irrelevant or contradictory content is generated (Maynez et al., 2020). The field of sentence simplification has also seen an increased emphasis on faithfulness. Current research has found that errors frequently appear in both the corpora itself and the generated outputs (Devaraj et al., 2022).

Therefore, we created a new simplified corpus dedicated to applications where faithfulness is essential, following the sentence-level format of many conventional approaches. Our corpus was developed with the cooperation of experts in language education, following guidelines that balance readability and faithfulness, resulting in 7,075 sentence pairs. Through our analysis of the corpora, we confirmed that our dataset preserves the fundamental content of the original text, including names of persons, dates, or cities. Manual evaluation further showed that our corpus maintains a high degree of faithfulness, outperforming other existing corpora. Moreover, the readability of the corpus for the target audience was verified through assessments by non-native readers. We conducted the experiment using our corpus to finetune a pretrained model and apply the Few-shot method to a large language model. Through this, we confirmed that our corpus aids in automatic TS while ensuring faithfulness to the original content. This corpus will be made publicly available[1].

2. Related Work

Faithfulness in Simplification The issue of faithfulness and factuality in system-generated content has been discussed in the context of summarization tasks (Cao et al., 2020). However, recent studies have underscored the importance of preserving the content expressed in the original text within simplification tasks (Guo et al., 2018; Laban et al., 2021). Devaraj et al. (2022) emphasized the understudied nature of factual accuracy in TS and presented a taxonomy for errors in faithfulness, categorizing them as insertion, deletion, or substitution. Through the examination of texts from the English corpora like WikiLarge and Newsela (Zhang and Lapata, 2017; Xu et al., 2015), alongside generated samples, they observed frequent errors in both corpora and generated outputs.

Simplification Corpora In TS, the approach of treating simplification as a translation problem using seq2seq has gained prominence, leading to the

[1] `https://cl.asahi.com/api_data/simplification`

introduction of numerous parallel corpora in various languages (Ryan et al., 2023). Within Japanese TS, diverse corpora are characterized by distinct methodologies and objectives. SNOW (Maruyama and Yamamoto, 2018; Katsuta and Yamamoto, 2018) is characterized by manual simplification, employing only fundamental vocabulary. JADES (Hayakawa et al., 2022) is a dataset created to evaluate TS in Japanese targeting non-native readers. Furthermore, MATCHA (Miyata et al., 2024)[2] consists of simplified inbound tourism articles created by experts and manually performed sentence alignment. Particularly relevant to the purpose of our application is the study by The Japan Broadcasting Corporation (NHK) (Goto et al., 2015), which pioneered the simplification of Japanese news content. However, the corpus created in this study is not publicly available.

Sentence Pair (top: complex; bottom: simple)
まずは**年300台**の有効活用をめざす。 まずは**一年に300台**をうまく使うことを目標にします。 We aim to effectively utilize **300 units annually** at first. The first goal is to successfully use **300 units a year**.
この日の**閣僚**会合で要請を決めた。 この日の**閣僚**の会議でお願いすると決めました。 The request was decided at the **ministerial** meeting on this day. At a meeting of the **ministers** on this day, they decided to ask for it.
出産は**帝王切開**になる。 子どもは、**帝王切開**で産みます。 The birth will be by **cesarean sectio**. The child is delivered by **cesarean section**.

Table 1: Examples of sentence pairs from our corpus. The bolded parts indicate content retained before and after simplification according to our guidelines

3. Corpus

3.1. Corpus Design

Data and Annotators Our corpus draws from news articles owned by The Asahi Shimbun Company, covering topics such as politics, economics, incidents, sports, and local news. A total of 690 articles were simplified by 30 experts in Japanese language education. Subsequently, manual annotations of sentence pairs were conducted, with examples that did not conform to the rules either appropriately excluded or edited to comply with the rules.

Simplification Level Our target audience comprises readers studying for the JLPT N3 level[3], where they can understand Japanese used in everyday situations and comprehend paraphrased expressions for complex terms.

Faithfulness vs. Readability Maintaining maximum faithfulness between sentence pairs is the same as minimizing edits to the original text, which is a trade-off for readability. While fundamental contents such as names, dates, and monetary amounts should be retained, adding auxiliary information or explanations for complex terms, such as political jargon, can enhance readability but may reduce faithfulness. To address this, we have established guidelines to handle such terms without unnecessary simplification. The specific guidelines are as follows:

1. Translate every sentence.
2. Include all information from the original article in the translated article without any omissions or additions.
3. Simplify each sentence independently without adding auxiliary information to aid in knowledge or context.
4. Use simple words as a general rule (aiming for comprehensibility at approximately the post-beginner to N3 level).
5. Convert to a mixture of kanji and kana.
6. Uniformly use the ”desu-masu (です・ます)” form for the sentence style[4].
7. Use kanji regardless of the target simplification level.
8. For parts in the original article that use quotation marks「」, such as statements, use the same「」in the translation.
9. Write the text with spaced writing[5].

Table 1 shows examples of our corpus.

3.2. Analysis

Table 2 presents statistical information about our corpus and compares it with other corpora. To gauge how well the original content was retained, we calculated the proportion of entities appearing in the original text that also appear in the simplified text using the same expression (% of Entity Retention). Our corpus exhibits the highest rate at 77.7%. When focusing on entities such as names of persons (PERSON), dates (DATE), and city names (CITY), which are less likely to be paraphrased during editing, nearly 90% are preserved through simplification. This preservation ability is further reflected in the sentence-wise BLEU score calculation results. Our corpus has a higher average score, indicating better preservation of the lexical

[2] https://github.com/EhimeNLP/matcha

[3] https://www.jlpt.jp/about/levelsummary.html

[4] This is a polite style of speaking or writing in Japanese.

[5] Japanese text is typically not written with spaces between words, but spaces are introduced here to make word separation easier to understand.

	Ours		MATCHA		SNOW		JADES	
	comp.	simp.	comp.	simp.	comp.	simp.	comp.	simp.
Data Source	News Articles		Tourism Articles		Textbook		News Articles	
#Sentence Pairs	7,075		16,000		84,300		3,907	
#Sentences	7,280	9,627	16,143	18,605	85,076	85,051	3,940	4,741
Avg. #Words Per Sent	30.11	26.08	21.49	19.81	10.75	11.89	31.93	26.12
#Unique Entities	10,049	10,008	10,257	10,354	6,113	3,675	6,659	4,661
% of Entity Retention*	77.7		70.7		63.8		54.8	
-PERSON*	93.5		74.4		84.3		78.4	
-DATE*	87.2		77.6		67.5		82.9	
-CITY*	88.1		82.0		83.2		77.7	
Avg. BLEU*	51.2		37.3		45.8		26.7	
% of Identical	0.00		0.79		25.5		0.00	

Table 2: Statistics of our corpus compared with other corpora. Items marked with an asterisk (*) in the table represent results calculated solely from non-identical complex and simple sentence pairs

Sentence Pair (top: complex; bottom: simple)
(SNOW) 彼は**ポーカー**がとても上手だ。 彼は **5 枚のカードを基本にするゲーム**がとてもうまい。 He is very good at **poker**. He is very good at **games based on five cards**.
(JADES) **ラグビー W 杯**「ミラクル！」 **スポーツの世界試合**「すごい!」 **Rugby World Cup** ”Miracle!” **Sports World Games** ”WOW!”
(MATCHA) 備考：数あるプランの中で最も安い方法です。 **成田空港から上野駅まで**一番安い行き方です。 Note: This is the cheapest method among the many plans available. It’s the cheapest way to get from **Narita Airport to Ueno Station**.

Table 3: Examples of sentence pairs from existing corpora. The parts that compromise faithfulness are in bold

content of the original text compared to other corpora. MeCab[6] (ipadic) was utilized for word count calculation, GiNZA[7] for named entity extraction, and SacreBLEU (Post, 2018) for sentence-wise BLEU score computation. Table 3 shows examples of low-faithfulness sentence pairs from existing corpora.

3.3. Evaluation

Faithfulness Evaluation by Native Readers We conducted a manual evaluation to assess faithfulness in each corpus following the methodology proposed in Devaraj et al. (2022). Five annotators evaluated 30 randomly extracted sentence pairs from each corpus. Each annotator assessed whether each operation (insertion, deletion, substitution) altered the main idea conveyed by the original text, scoring as follows: 0 for no change or trivial change, 1 for a non-trivial change preserving the main score, 2 for changes affecting the main idea, and -1 for incomprehensible content.

Category	**Dataset**	**0**	**1**	**2**	**-1**
Insertion	Ours	96.6	3.3	0.0	0.0
	SNOW	96.6	3.3	0.0	0.0
	JADES	96.6	0.0	0.0	3.3
	MATCHA	92.8	7.1	0.0	0.0
Deletion	Ours	100	0.0	0.0	0.0
	SNOW	100	0.0	0.0	0.0
	JADES	62.9	3.7	29.6	3.7
	MATCHA	96.2	0.0	3.7	0.0
Substitution	Ours	79.1	20.8	0.0	0.0
	SNOW	67.8	28.5	3.5	0.0
	JADES	7.4	59.2	29.6	3.7
	MATCHA	57.1	35.7	7.1	0.0

Table 4: Manual evaluation of Insertion, Deletion, and Substitution error (%) in our corpus and other corpora. 0: no/trivial change; 1: nontrivial but preserves main idea; 2: does not preserve main idea; -1: undiscernedable

Table 4 presents the results of this evaluation, indicating our corpus had the highest percentage of scores marked as 0 (indicating highest faithfulness) and the lowest percentage rated 2 (indicating lowest faithfulness) in all operations. The manual evaluation also confirmed the high faithfulness in our corpus.

Readability Evaluation by Non-Native Readers While our corpus maintains high faithfulness, but readability for the target audience is crucial. Thus, we conducted readability evaluation with five non-native annotators. They assessed one hundred randomly selected sentence pairs, providing binary evaluations on whether the simplified sentences were easier to understand than the complex ones, along with rating the comprehensibility of the sim-

[6] https://taku910.github.io/mecab/
[7] https://megagonlabs.github.io/ginza/

plified sentences on a scale of 1 (difficult to understand) to 5 (easy to understand). Additionally, annotators provided free-form comments for each pair and the overall annotation.

The results in Table 5 indicate that our simplified sentences are generally more readable for non-native readers. Although binary evaluation results may appear modest, comments from annotators often indicated comprehension of the complex sentences, likely influencing their evaluations. Some annotators highlighted that specialized terms affected readability. However, as mentioned in Section 3.1, overly simplifying these terms might compromise faithfulness. As proposed in Tanaka et al. (2018), providing readers with a dictionary can be a practical solution.

% of simp. is easier	Avg. of readability
78.20	4.01

Table 5: Manual evaluation by non-native readers

4. Experiment

We conducted experiments to assess our corpus' impact on system generation faithfulness. After splitting our dataset into an 8:1:1 ratio for training, validation, and test data, we fine-tuned the pretrained BART model (Lewis et al., 2020) using training data, and also employed training data in applying the Few-shot method to GPT-3.5. We then evaluated the sentences generated from the test data by comparing them with Zero-shot generation by GPT-4. We used the Japanese BART base model[8] and `gpt-3.5-turbo-0613` and `gpt-4-0613` provided by OpenAI[9]. The prompts provided to GPT-3.5 and GPT-4 included the guidelines introduced in Section 3.1. In appendix A, we present a detailed experimental setup in the fine-tuning of the BART model and the prompts provided for GPT-3.5 and GPT-4.

Table 6 displays the results from the automatic evaluation, with the BART model outperforming others in BLEU and SARI (Xu et al., 2016) scores. Table 7 presents faithfulness evaluation results conducted by the same five annotators as in Section 3.3. They evaluated the output sentences from each model for 20 randomly selected inputs from the test data. BART and GPT-3.5, using our dataset, exhibited fewer faithfulness errors (scored as 2) than GPT-4 Zero-shot across all operations. Neither BART nor GPT-3.5 had substitution errors rated 2, indicating that such errors appear to be mitigated by our data. While deletions in GPT-3.5 received scores of 0 or 1, BART and GPT-4 encountered errors rated 2, with frequencies of 5.2% and 10.0%, respectively. Notably, there were no deletion errors in our corpus (as shown in Table 4), but we observed them when we fine-tuned BART in this corpus. We speculate that this might be a characteristic inherent in the BART model.

Model	BLEU	SARI
BART	**57.20**	**57.90**
GPT-3.5 3-shot	54.38	56.32
GPT-4 Zero-shot	29.66	40.02

Table 6: Automatic evaluation of system generation

Category	Model	0	1	2	-1
Insertion	BART	100	0.0	0.0	0.0
	GPT-3.5 3-shot	100	0.0	0.0	0.0
	GPT-4 Zero-shot	95.0	0.0	0.0	5.0
Deletion	BART	94.7	0.0	5.2	0.0
	GPT-3.5 3-shot	95.0	5.0	0.0	0.0
	GPT-4 Zero-shot	80.0	5.0	10.0	5.0
Substitution	BART	94.7	5.2	0.0	0.0
	GPT-3.5 3-shot	100	0.0	0.0	0.0
	GPT-4 Zero-shot	36.8	47.3	10.5	5.2

Table 7: Manual evaluation of Insertion, Deletion, and Substitution error (%) in system generation

5. Conclusion

In this paper, we introduced a sentence-level Japanese news simplification corpus with high faithfulness. By manually simplifying newspaper articles following guidelines that consider the trade-off between readability and faithfulness, we compiled a dataset of 7,075 pairs. Our analysis and evaluations demonstrated superior faithfulness in our corpus compared with existing ones. Furthermore, we confirmed its readability to the target audience through evaluation by non-native readers. We also demonstrated that our corpus aids in fine-tuning BART and in providing Few-shot examples to GPT-3.5, enabling both models to generate faithful simple sentences. In future work, we aim to create and evaluate a document-level dataset.

6. Acknowledgments

We thank ABK COLLEGE's faculty and students for their invaluable cooperation in developing and evaluating our dataset. We are also grateful to our research advisors, Professor Kentaro Inui from MBZUAI & Tohoku University and Professor Naoaki Okazaki from Tokyo Institute of Technology, for their feedback and support during our research.

[8] `https://huggingface.co/ku-nlp/bart-base-japanese`

[9] `https://openai.com/blog/openai-api`

7. Bibliographical References

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Isao Goto, Hideki Tanaka, and Tadashi Kumano. 2015. Japanese news simplification: tak design, data set construction, and analysis of simplified text. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 462–476, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Akio Hayakawa, Tomoyuki Kajiwara, Hiroki Ouchi, and Taro Watanabe. 2022. JADES: New text simplification dataset in Japanese targeted at nonnative speakers. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 179–187, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *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)*, pages 6365–6378, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Takumi Maruyama and Kazuhide Yamamoto. 2018. Simplified corpus with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Rina Miyata, Hyuga Koretake, Hiroki Yamauchi, Daiki Yanamoto, Tomoyuki Kajiwara, Takeshi Ninomiya, and Yasuhiro Nishiwaki. 2024. Matcha: Parallel corpus for japanese text simplification based on professionally simplified articles. *Journal of Natural Language Processing*, 31(2). To appear.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help? text simplification strategies for people with dyslexia.

Michael Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the*

61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

Hideki Tanaka, Tadashi Kumano, Isao Goto, and Hideya Mino. 2018. Easy japanese news production support system. *Journal of Natural Language Processing*, 25(1):81–117. (in Japanese).

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yasukata Yano, Michael H. Long, and Steven J. Ross. 1994. The effects of simplified and elaborated texts on foreign language reading comprehension. *Language Learning*, 44:189–219.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

A. Experimental Setup

A.1. Model Training

For the training of our model, we employed the Hugging Face Transformers library, using the following configuration parameters:

- **Optimization:** Standard Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e - 08$
- **FP16 Training:** Enabled, with optimization level O1 for mixed precision training
- **Batch Size:** 8 per device for training, 4 per device for evaluation
- **Gradient Accumulation Steps:** 2
- **Learning Rate:** 5×10^{-5}
- **LR Scheduler:** Linear, with warmup ratio of 0.2

A.2. Prompt Input

Figure 1 illustrates the prompts input to the large language models.

```
You are an editor who is about to modify complex
text(s) into simple text(s).
Do not omit any content and rewrite it simpler.
Output should be in Japanese.

Note: Please adhere to the following rules.
1. Translate every sentence.
2. Include all information from the original article
...
9. Write the text with spaced writing

Complex: {Input Complex Sentence}
Simple:
```

Figure 1: Prompt input for GPT-3.5 / 4. In the three-shot approach, input three pairs as examples in the format Complex: Simple:.