Parallel Corpus Filtering Based on Semantic Similarity and Surface Dissimilarity for Japanese Text Simplification with LLMs

Daisuke Maekawa[†] **Tomoyuki Kajiwara**[†‡] **Takashi Ninomiya**[†]
[†]Ehime University [‡]The University of Osaka
{maekawa@ai.cs., kajiwara@cs., ninomiya.takashi.mk@}ehime-u.ac.jp

Abstract

We are focusing on low-cost fine-tuning for large language models (LLMs) in Japanese text simplification. LLMs have achieved high performance even with fine-tuning on small parallel corpora in tasks such as machine translation and dialogue response generation. In this study, we propose a method of parallel corpus filtering for text simplification and investigate how much the number of sentence pairs for fine-tuning LLMs can be reduced. Experimental results on Japanese corpora in three domains revealed that the ability to perform text simplification tasks can be acquired even from a very small corpus of 16 to 64 sentence pairs. Although more parallel corpora are needed to acquire domain knowledge, our method outperformed full fine-tuning while reducing the training corpus by approximately 70%.

Keywords: Parallel Corpus Filtering, Japanese Text Simplification, Large Language Models

1. Introduction

Text simplification (Alva-Manchego et al., 2020) is a task of paraphrasing a given sentence to make it easier to understand. This technology supports the reading comprehension for non-native speakers (Petersen and Ostendorf, 2007) and children (De Belder and Moens, 2010). In this study, we tackle sentence simplification in Japanese with large language models (LLMs) (Llama Team, 2024; Fujii et al., 2024; Okazaki et al., 2024).

LLMs acquire most of their knowledge and abilities during pre-training (Zhou et al., 2023). Therefore, even fine-tuning from a small parallel corpus sometimes achieves high performance. For example, 32 sentence pairs are used in machine translation (Zhu et al., 2024), and 1,000 sentence pairs are used in dialogue response generation (Zhou et al., 2023). However, the number of sentence pairs required for fine-tuning LLMs in the text simplification task is still unknown.

Obviously, it is not enough to fine-tune on just a small parallel corpus. Previous studies (Zhou et al., 2023; Chen et al., 2024) have emphasized that the corpora for fine-tuning LLMs are more important in quality than quantity. As we will show empirically in Section 4, randomly reducing parallel corpora leads to performance degradation in the text simplification task as well. To efficiently fine-tune LLMs, we need methods for carefully selecting high-quality sentence pairs in the target task.

An existing method for parallel corpus filtering in text simplification (Hatagaki et al., 2022) is designed to remove a small amount of noisy sentence pairs. In contrast to traditional text generation models that require large-scale training corpora, since we target LLM-based models, we focus on selecting a small number of high-quality sentence pairs.

In this study, to reduce the cost of fine-tuning LLMs, we propose a method of parallel corpus filtering that extracts a small amount of high-quality sentence pairs from a parallel corpus for text simplification. First, we explain why methods based on semantic similarity (Chaudhary et al., 2019; Batheja and Bhattacharyya, 2022; Hatagaki et al., 2022), which have been employed in parallel corpus filtering for text simplification and machine translation, are not suitable for this task. Then, we propose a method that also considers surface dissimilarity, which encourages aggressive simplification.

Experimental results on Japanese text simplification in three domains show that the proposed method improves the performance of text simplification models compared to existing parallel corpus filtering methods. Detailed analysis revealed that when fine-tuning on the parallel corpus extracted by our method, even with a very small number of sentence pairs (16 to 64), LLMs can acquire the ability of text simplification. Although more corpora are needed to acquire domain knowledge sufficiently, fine-tuning with 4,000 sentence pairs, which reduced the number of sentence pairs by about 70%, achieved higher performance than full fine-tuning.

2. Related Work

Parallel corpus filtering (Koehn et al., 2018, 2019, 2020) has primarily been studied in the context of machine translation tasks, which utilize large-scale parallel corpora. Unsupervised methods include those based on the cosine similarity of multilingual sentence encoders such as LASER (Artetxe and Schwenk, 2019) and LaBSE (Feng et al., 2022) (Chaudhary et al., 2019; Batheja and Bhat-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1110–1116
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tacharyya, 2022). Supervised methods include approaches based on machine learning that use features like translation probabilities (Sánchez-Cartagena et al., 2018) and deep learning methods that utilize multilingual masked language models like XLM-R (Conneau et al., 2020; Zaragoza-Bernabeu et al., 2022). In this paper, we will consider parallel corpus filtering suitable for text simplification, a text-to-text generation task within the same language, while comparing it to prior research in machine translation.

3. Proposed Method

To reduce the cost of fine-tuning LLMs, we propose a method of parallel corpus filtering for text simplification. If high-performance models can be trained from small parallel corpora, it could lead to solutions for the low-resource problem in this task.

3.1. Preliminary Experiment

Following previous studies on parallel corpus filtering for text simplification (Hatagaki et al., 2022) and machine translation (Chaudhary et al., 2019; Batheja and Bhattacharyya, 2022), we apply parallel corpus filtering based on cosine similarity of sentence embeddings as a preliminary experiment.

Method We first filtered out sentence pairs with low cosine similarity using LaBSE (Feng et al., 2022) from the training set of SNOW (Maruyama and Yamamoto, 2018; Katsuta and Yamamoto, 2018), a Japanese text simplification corpus. Then, we fine-tuned Swallow (Fujii et al., 2024; Okazaki et al., 2024), a Japanese LLM, on that corpus applying LoRA (Hu et al., 2022).

Results Unfortunately, LLMs fine-tuned on sentence pairs with high semantic similarity became overly conservative simplification models that output the input sentences as is. For example, an LLM fine-tuned on 20% of the 82,300 sentence pairs reproduced the input sentence verbatim in 90% of the cases on the evaluation set. As shown in Figure 1, many sentence pairs with high semantic similarity (cosine similarity using LaBSE) also have high surface similarity (BLEU (Papineni et al., 2002)). This leads to fine-tuning of LLMs often targeting the identical sentence pairs. Furthermore, as can also be seen from Figure 1, since sentence pairs with extremely high surface similarity dominate, a similar trend was observed even when the parallel corpus was randomly reduced.

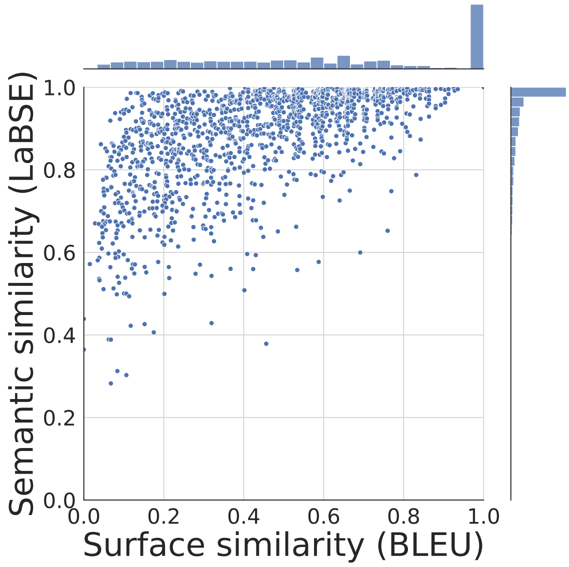

Figure 1: Distribution of semantic and surface similarity between sentence pairs in the SNOW corpus before applying any parallel corpus filtering.

3.2. Parallel Corpus Filtering Based on Semantic and Surface Similarity

To address the conservative nature of text simplification models fine-tuned on randomly selected sentence pairs or those with high semantic similarity, our proposed method also incorporates surface dissimilarity. The low surface similarity is related to more editing between sentences; therefore, fine-tuning on sentence pairs with low surface similarity leads to a preference for more aggressive simplification. However, sentence pairs with low surface similarity are often noisy pairs due to alignment errors. Therefore, we design the proposed method to select sentence pairs with high semantic similarity while low surface similarity.

The quality of a pair consisting of complex sentence x and simple sentence y is defined as follows:

$$Q(x, y) = \sqrt{(1 - \text{LaBSE}(x, y))^2 + \text{BLEU}(x, y)^2}$$

Here, LaBSE$(\cdot)$ denotes semantic similarity based on cosine similarity computed using LaBSE[1] (Feng et al., 2022), and BLEU$(\cdot)$ denotes surface similarity computed using SacreBLEU[2] (Papineni et al., 2002). This score represents the Euclidean distance from the upper-left corner in Figure 1, balancing semantic similarity and surface dissimilarity. We exclude sentence pairs from the corpus in descending order of this score.

[1]https://huggingface.co/sentence-transformers/LaBSE
[2]https://github.com/mjpost/sacrebleu

4. Evaluation

We apply our parallel corpus filtering to Japanese text simplification corpora and reveal the required corpus size by fine-tuning LLMs and pre-trained sequence-to-sequence models.

4.1. Settings

Data We used MATCHA[3] (Miyata et al., 2024) and SNOW[45] (Maruyama and Yamamoto, 2018; Katsuta and Yamamoto, 2018), parallel corpora for Japanese text simplification, for both training and evaluation. Additionally, we also used JADES[6] (Hayakawa et al., 2022), a small-scale corpus for Japanese text simplification, for evaluation. Table 1 shows the statistics for each corpus.

Models We constructed text simplification models by fine-tuning Swallow[7] (Fujii et al., 2024; Okazaki et al., 2024), a Japanese LLM, and Japanese BART[8] (Lewis et al., 2020), a sequence-to-sequence model pre-trained on Wikipedia. We employed greedy decoding for text generation.

Hyperparameters We evaluated the performance of text simplification models while reducing the number of sentence pairs for fine-tuning. During fine-tuning, we set the batch size to 16 sentence pairs, the learning rate to 5×10^{-5}, and used AdamW (Loshchilov and Hutter, 2019) as the optimizer. We employed early stopping, which terminates training if the cross-entropy loss on the validation set does not improve for three epochs. For fine-tuning LLMs, we applied LoRA[9] (Hu et al., 2022) with a rank of $r = 16$, scaling factor $\alpha = 16$, and dropout rate of 0.05.

Comparison Methods We compared our proposed method with the following six parallel corpus filtering methods. For evaluation, we used SARI[10] (Xu et al., 2016) and BERTScore[11] (Zhang et al., 2020). In Japanese text simplification, SARI correlates with human evaluations in terms of simplicity, while BERTScore in terms of fluency and synonymity (Hayakawa et al., 2022).

[3] https://github.com/EhimeNLP/matcha
[4] https://www.jnlp.org/GengoHouse/snow/t15
[5] https://www.jnlp.org/GengoHouse/snow/t23
[6] https://github.com/naist-nlp/jades
[7] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.2
[8] https://huggingface.co/ku-nlp/bart-large-japanese
[9] https://github.com/unslothai/unsloth
[10] https://github.com/feralvam/easse
[11] https://github.com/Tiiiger/bert_score

	Training	Validation	Evaluation
MATCHA	14,000	1,000	1,000
SNOW	82,300	2,000	700
JADES	-	-	3,907

Table 1: Number of sentence pairs in each corpus.

w/o Filtering This baseline uses the whole parallel corpus for fine-tuning.

0-shot This baseline uses the prompt "Please simplify the input sentence" without any fine-tuning. Note that this is only used for LLMs.

Random It randomly selects sentence pairs.

Bicleaner-AI This is a supervised method for parallel corpus filtering in machine translation[12] (Zaragoza-Bernabeu et al., 2022), but applied to text simplification corpora.

COS This is a method without surface similarity derived from our proposed method. That is, sentence pairs are selected in descending order of semantic similarity.

BLEU This is a method without semantic similarity derived from our proposed method. That is, sentence pairs are selected in ascending order of surface similarity.

4.2. Results

Experimental results for SARI are shown in Figure 2. Each subgraph plots the number of sentences for fine-tuning on the horizontal axis and the evaluation score on the vertical axis. Solid and dotted lines represent LLM and BART, respectively.

In-Domain Evaluation with LLM While larger training corpora generally lead to higher performance, our proposed method outperformed full fine-tuning when reducing only a small amount of noisy sentence pairs. Specifically, in experimental settings using 4,000 sentence pairs from the MATCHA corpus (about 70% reduction) and 64,000 sentence pairs from the SNOW corpus (about 20% reduction), we improved simplification performance while reducing fine-tuning costs. In addition, the proposed method achieved the highest performance in most cases across all training corpus sizes. These experimental results demonstrate the effectiveness of the proposed method.

MATCHA is a corpus annotated by experts, while SNOW is annotated by non-experts. MATCHA outperformed full fine-tuning with a smaller training

[12] https://github.com/bitextor/bicleaner-ai

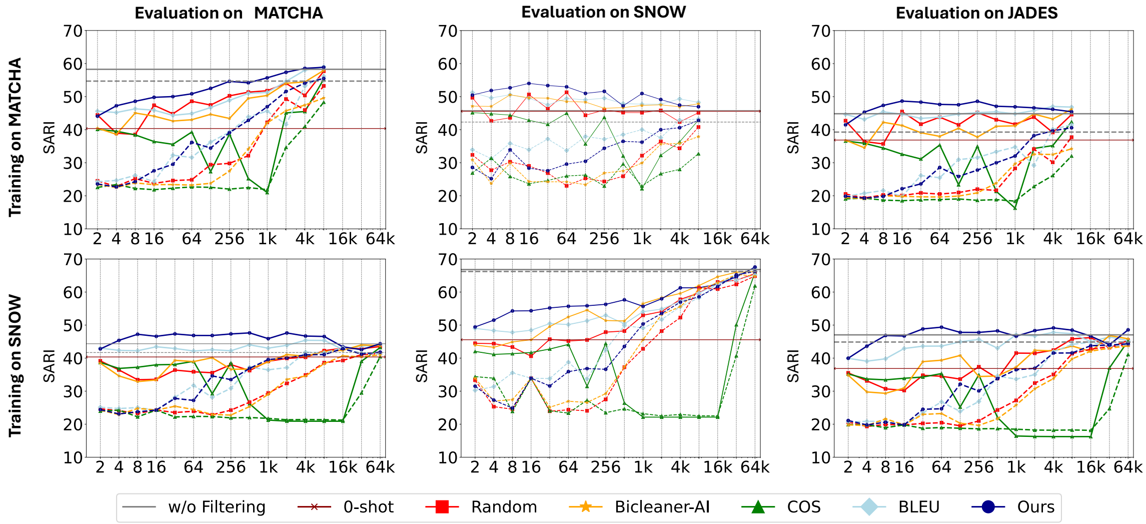

Figure 2: Evaluation results of Japanese text simplification performance with SARI for each number of sentence pairs in fine-tuning. Solid and dotted lines represent LLM and BART, respectively.

花粉症とは，植物の花粉によって引き起こされるアレルギー反応のことで，日本の場合，スギやヒノキが原因となることが多いです． (Hay fever is an allergy caused by pollen from plants, often caused by cedar and cypress in Japan.)	
LLMs (16)	花粉症は，植物の花粉が原因で起こるアレルギーです．日本では，スギやヒノキの花粉が原因になることが多いです． (Hay fever is an allergy caused by pollen from plants. In Japan, pollens from cedar and cypress are often the cause.)
LLMs (Full)	花粉症は，植物の**花粉 (pollen)** で起こるアレルギーです．日本では，スギやヒノキが原因になることが多いです． (Hay fever is an allergy caused by **pollen** from plants. In Japan, cedar and cypress are often the cause.)

Table 2: An example from the MATCHA corpus. LLMs can acquire basic ability for text simplification, such as sentence splitting, through fine-tuning with only 16 sentence pairs (16). The more fine-tuned LLMs (Full) can capture the characteristics of the MATCHA corpus, which provides English translations for complex words.

corpus size than SNOW, which may be explained by the quality of the corpus. Previous studies (Zhou et al., 2023; Zhu et al., 2024; Chen et al., 2024) have also reported that with high-quality training corpus, LLMs can achieve high performance with only small-scale fine-tuning.

Finally, note that the COS baseline exhibited tendencies distinct from other methods. Within the first 1,000 sentence pairs in the MATCHA corpus, i.e., the top 1,000 pairs by semantic similarity, and within the first 16,000 pairs in the SNOW corpus, increasing the training corpus size did not improve simplification performance. This suggests that sentence pairs with excessively high semantic similarity may hinder training for text simplification.

In-Domain Evaluation with BART When BART had access to a large-scale parallel corpus, i.e., over 16,000 sentence pairs in the SNOW corpus, it performed comparably to LLMs. However, its performance degraded significantly as the training corpus size decreased. In the MATCHA corpus, BART also achieved performance equivalent to full fine-tuning through fine-tuning on only 4,000 sentence pairs. This suggests that a high-quality parallel corpus can significantly reduce training costs by our method. For the COS baseline, similar trends were observed as in the experiments for LLMs.

Out-of-Domain Evaluation with LLM Unlike in-domain evaluation, increasing the training corpus size did not necessarily improve simplification performance. In our method, fine-tuning with only 16 sentence pairs on the MATCHA corpus and 64 pairs on the SNOW corpus achieved the highest performance. Further training did not improve simplification performance. These experimental results imply that the ability to perform text simplification task can be acquired even from a very small number of sentence pairs, such as several dozen.

On the other hand, the training corpus size and simplification performance were proportional for in-domain evaluation. As shown in Table 2, this implies that acquiring domain knowledge requires a large-scale training corpus. In addition, similar to the in-domain evaluation, the fact that the proposed method generally outperforms other methods demonstrates its effectiveness.

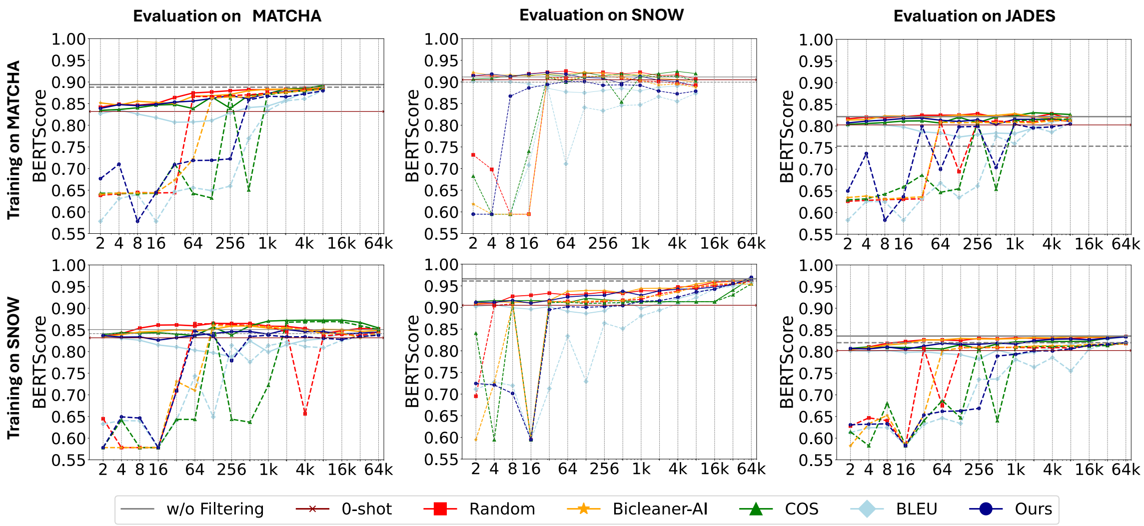

Figure 3: Evaluation results for fluency and synonymity in Japanese text simplification with BERTScore. Solid and dotted lines represent LLM and BART, respectively.

Out-of-Domain Evaluation with BART Similar to the in-domain evaluation, simplification performance deteriorated in BART in proportion to the reduction in training corpus size. Therefore, unlike LLMs, it is difficult to fine-tune BART using a very small number of sentence pairs.

Analysis of BERTScore Analysis of SARI suggested that our proposed method enables LLMs to acquire simplification ability even with a small training corpus. To complement our experiments, we also evaluated using BERTScore in Figure 3, which is a better metric for fluency and synonymity.

In the in-domain evaluation, LLMs fine-tuned with the proposed method consistently achieved high scores across all corpora, regardless of training corpus sizes. These experimental results demonstrate that the proposed method provides text simplification models capable of faithfully preserving the original meaning. In contrast, BART had a significantly worse BERTScore as the training corpus size decreased, suggesting it is less robust than LLMs under low-resource conditions.

In the out-of-domain evaluation, LLMs fine-tuned with the proposed method again achieved consistently high scores regardless of training corpus sizes, showing a similar trend to that observed with SARI. Notably, the LLM fine-tuned on SNOW achieved higher performance than full fine-tuning using only 16 to 64 sentence pairs.

In evaluations with BERTScore, there were some cases where the proposed method slightly underperformed the COS baseline. This is possibly due to the consideration of surface dissimilarity, which encourages aggressive simplification in our method. However, in contrast to the BLEU baseline, which considers only surface dissimilarity and significantly degrades synonymity, the proposed method effectively prevents semantic deviation.

Overall, these experimental results revealed that the proposed method, which balances semantic similarity and surface dissimilarity, effectively improves synonymity, fluency, and simplicity even from a very small training corpus.

5. Conclusion

We propose a method for parallel corpus filtering in text simplification that considers both semantic and surface similarity. Experimental results in Japanese show that LLMs acquire the ability to perform the text simplification task with fine-tuning on just a few dozen sentence pairs. Nevertheless, acquiring domain knowledge requires more corpora.

Acknowledgements

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425.

Bibliographical References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven Sentence Simplification: Survey and Benchmark. *Computational Linguistics*, 46(1):135–187.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond.

Transactions of the Association for Computational Linguistics, 7:597–610.

Akshay Batheja and Pushpak Bhattacharyya. 2022. Improving Machine Translation with Phrase Pair Injection and Corpus Filtering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5395–5400.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-Resource Corpus Filtering Using Multilingual Sentence Embeddings. In *Proceedings of the Fourth Conference on Machine Translation*, pages 261–266.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. AlpaGasus: Training A Better Alpaca with Fewer Data. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jan De Belder and Marie-Francine Moens. 2010. Text Simplification for Children. In *Proceedings of the SIGIR 2010 Workshop on Accessible Search Systems*, pages 19–26.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 878–891.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In *Proceedings of the First Conference on Language Modeling*.

Koki Hatagaki, Tomoyuki Kajiwara, and Takashi Ninomiya. 2022. Parallel Corpus Filtering for Japanese Text Simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability*, pages 12–18.

Akio Hayakawa, Tomoyuki Kajiwara, Hiroki Ouchi, and Taro Watanabe. 2022. JADES: New Text Simplification Dataset in Japanese Targeted at Non-Native Speakers. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability*, pages 179–187.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the Tenth International Conference on Learning Representations*.

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*, pages 461–466.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 Shared Task on Parallel Corpus Filtering for Low-Resource Conditions. In *Proceedings of the Fourth Conference on Machine Translation*, pages 54–72.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739.

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.

Llama Team. 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the Seventh International Conference on Learning Representations*.

Takumi Maruyama and Kazuhide Yamamoto. 2018. Simplified Corpus with Core Vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1153–1160.

Rina Miyata, Hinata Koretaka, Hiroki Yamauchi, Daiki Yanamoto, Tomoyuki Kajiwara, Takashi Ninomiya, and Yasuhiro Nishiaki. 2024. MATCHA: Parallel Corpus for Japanese Text Simplification Based on Professionally Simplified Articles. *Journal of Natural Language Processing*, 31(2):590–609.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. 2024. Building a Large Japanese Web Corpus for Large Language Models. In *Proceedings of the First Conference on Language Modeling*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Sarah E. Petersen and Mari Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of the 1st Workshop on Speech and Language Technology in Education*, pages 69–72.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962.

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.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner Goes Neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 55006–55021.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024. Fine-Tuning Large Language Models to Translate: Will a Touch of Noisy Data in Misaligned Languages Suffice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 388–409.