Using AI to Simplify Japanese Documents for Language Learners

Jack Ning
Johns Hopkins University, jning5@jh.edu

Abstract - **This research presents a novel approach to simplifying Japanese texts for language learners using AI, specifically leveraging a BART-based model fine-tuned on NHK news articles. The challenge of reading complex native-level Japanese texts often deters learners, and existing beginner materials either lack engagement or fail to match learner needs. By applying advanced text simplification techniques, this study aims to make Japanese texts more accessible while retaining essential information. The results reject the efficacy of the proposed method, achieving a significant improvement in summarization quality, with a SARI score of 67.178. However, there is also a notable increase in difficulty scores. These findings suggest that further AI research is needed to effectively bridge the gap between native-level texts and learner comprehension to offer a promising solution for enhancing language acquisition.**

Introduction

Japanese language learners often face significant challenges when reading native-level texts due to their complexity, which can hinder comprehension and discourage continued learning. There is limited availability of level-appropriate Japanese texts for learners. While beginner materials exist, they are often either dry, such as textbooks, or targeted at different demographics, like children's media, leading to a lack of engagement and interest.

This research aims to leverage AI to simplify Japanese texts while retaining essential information, thereby enhancing comprehension and engagement. Learners can gradually build their language skills and confidence by simplifying the texts.

My BART-based Japanese text simplification tool aims to reduce the complexity of Japanese texts while preserving key information. The goal is to achieve a SARI score of at least 0.5 and increase the Japanese readability score by two points. The SARI score evaluates fluency and meaning retention compared to the original and human-simplified texts (Xu et al., 2016), while the readability score assesses text difficulty (Sato, 2014). By simplifying texts, we aim to maintain learner engagement and reduce the likelihood of disengagement due to challenging content.

Related Work

Current methodologies in Japanese text simplification mainly focus on lexical and syntactic approaches, which simplify individual words or phrases but may overlook broader context, potentially resulting in losing meaning (Kajiwara et al., 2020; Zetsu et al., 2023). While these methods have merits, they may not fully address the complexities of document-level simplification.

In contrast, researchers have made substantial progress in English text simplification through 4 types of simplification: lexical, syntactic, discourse, and stylistic simplification (Espinosa-Zaragoza et al., 2023). The automatic simplification process typically consists of two stages: (1) the simplification plan, which involves deciding which linguistic aspects to simplify, such as identifying complex words or sentences, and (2) the simplification stage, where the plan is implemented, such as splitting long sentences. Approaches to these tasks are categorized into rule-based, data-driven, and hybrid methods. Rule-based methods use predefined linguistic rules, while data-driven methods rely on models trained to generate their own simplifications, capturing broader contexts and structures for comprehensive simplification tasks. Hybrid approaches combine the two.

Although most research focuses on sentence-level simplification, this paper addresses document-level simplification, which involves multiple sentences and paragraphs. Previous attempts at document-level simplification have often resulted in poor coherence due to iteratively simplifying sentences (Cripwell et al., 2023). Researchers have explored more holistic approaches such as sentence deletion, insertion, and reordering to address this. Recent English-language advancements in sentence-level features and document-level context have shown promising results.

This paper aims to address gaps in Japanese text simplification by adopting recent English strategies and introducing a corpus for document-level simplification. By leveraging these advanced methodologies, I seek to enhance the effectiveness of Japanese text simplification and improve overall document coherence.

While datasets like JADES (Japanese Dataset for the Evaluation of Simplification) and SNOW T15/T23 exist, they focus on modifying individual sentences rather than comprehensive text simplification (Hayakawa et al., 2022, pp. 179–187).

Method

Machine Learning Model

The BART model (Bidirectional and Auto-Regressive Transformers) was utilized to simplify Japanese news articles. Specifically, the ku-nlp/bart-base-japanese version, a transformer-based model pre-trained on Japanese Wikipedia, was employed along with the Juman++ and SentencePiece tokenizer. BART, developed by Facebook AI researchers like Mike Lewis, is a denoising autoencoder known for its efficacy in summarization and translation tasks. In this project, BART was fine-tuned on both regular and easy NHK articles (Lewis, 2019).

Additional models and tokenizers were incorporated to evaluate different metrics. The SBERT model was used to assess the text similarity to human text, and the readability metric used the MeCab tokenizer.

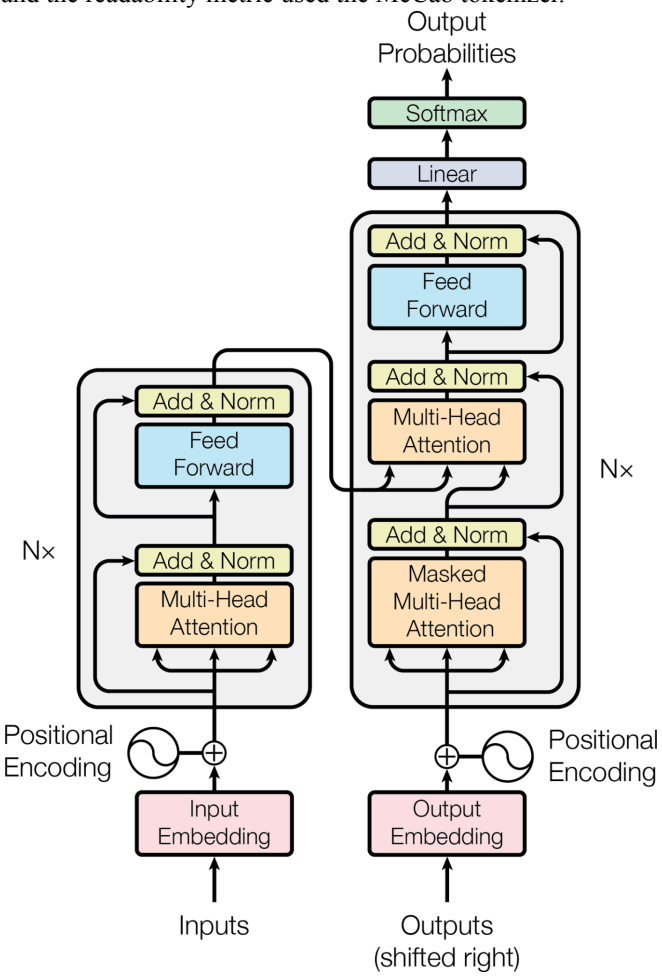

Figure 1: Transformer Architecture. The BART model uses a standard Transformer-based architecture, blending a bidirectional encoder, akin to BERT, with a left-to-right decoder, similar to GPT. The ku-nlp/bart-base-japanese model follows this architecture, utilizing 6 layers in both the encoder and decoder, consistent with the original BART model (Géron, 2019).

Datasets

This study utilizes NHK news articles. These articles were originally written for native Japanese speakers and have an easier version created specifically for foreigners learning Japanese (Tanaka, 2013). Native Japanese speakers wrote both versions. The simplification process focused on simpler syntax, removing redundancy, and using simpler vocabulary, typically at a JLPT (Japanese-Language Proficiency Test) N2-N3 level (Tanaka & Mino, 2018).

NHK conducted a study on the effectiveness of these simplified articles by having foreign language school students read ten random articles and answer related questions. There were three groups of students: pre-intermediate (L3), intermediate (L2), and advanced (L1). For each group, the simplified articles resulted in a correct answer rate increase of over 9%, and the average time to answer a question decreased by over 20%.

Additionally, the Japanese Language Education Vocabulary List is used to determine the origin of words for the readability score (Sunakawa, 2012). This data was collected based on data from 100 language textbooks and contemporary Japanese, making this dataset suited towards generating and scoring beginner text.

Evaluation Metrics

To evaluate the quality and effectiveness of the AI-based text simplification tool, I use 3 criteria: preservation of meaning, fluency, and difficulty of text. To do this, the project will employ the following metrics:

- **BLEU (Bilingual Evaluation Understudy)**: Measures the similarity between the generated simplified text and reference texts by comparing n-grams sequences. This is a metric for evaluating machine translated text. However, it is also commonly used for text simplification tasks, so it was included for comparison with existing models (Xu et al., 2016).
- **SARI (System-Agnostic Reference-based Evaluation)**: Evaluates how well the simplified text preserves meaning and fluency compared to the original text and a simpler reference text by evaluating the effect of adding, deleting, and keeping words (Xu et al., 2016).
- **SBERT (Sentence-BERT)**: Creates vector embeddings from natural language to compare documents within a corpus on a semantic level using distance metrics allowing us to assess whether the simplified text preserves the originals meaning (Reimers & Gurevych, 2019).
- **Readability Formula for Japanese Language Education**: Hasebe and Lee's formula predicts Japanese text difficulty using the following formula (Hasebe & Lee, 2015):
    - X = (mean length of sentence * -0.056) + (proportion of kango words * -0.126) + (proportion of wago words * -0.042) + (proportion of verbs * -0.145) +

(proportion of auxiliary verbs * -0.044) + 11.724

- **Kango Words**: These are words of Chinese origin that have been adopted into Japanese. They often have more formal or academic connotations.
- **Wago Words**: These are words of Japanese origin, typically used in everyday conversation and native Japanese expressions.

Experiment Setup

The models were trained and evaluated on a machine equipped with an on a machine equipped with an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory.

Implementation

The implementation of our methodology involved several key steps. Initially, I collected the NHK news articles and other relevant datasets. The data preprocessing phase included cleaning the dataset by removing instances with missing values and filtering out low-quality sentences based on their SBERT scores (<.45). Low scores indicate that the articles were unrelated to each other or did a poor job preserving the original article. Tokenization was carried out using SentencePiece for BART and MeCab for readability analysis. The data was then split into training (80%) and test (20%) sets. During model training, I used a learning rate of 5e-5, a batch size of 16, and trained for 10 epochs with the AdamW optimizer. The simplified text was generated using a batch size of 16 and a beam size of 3. The evaluation of the simplified texts involved calculating BLEU, SARI, SBERT scores, and readability metrics using pytorch, transformers, and hugging face libraries. The text generation and evaluation are conducted 10 times to show stable results.

Results and Analysis

Average Readability Score	Average	Variance
Regular Text	5.220	0.005855
Easy Text	6.881	0.001584
Simplified Text	4.717	0.018796

Table 1: The average and variance readability score for the dataset, where a lower score indicates text is more difficult to read.

Fluency and Meaning Preservation	Average	Variance
SBERT Cosine Similarity	0.780	0.000026
BLEU Score	0.145	0.000019
SARI Score	67.178	0.023649

Table 2: Average scores and variance for SBERT, BLEU, and SARI metrics applied to the simplified text. These metrics compare the generated simplified text against easier NHK articles as references, with regular NHK articles used as sources if needed.

Results are shown in Table 1 and 2. Table 1 metrics reflect text comprehension, while Table 2 concerns the simplified text's fluency and meaning preservation. Based on the results, I reject the null hypothesis. I observed that while the SARI score improved, readability decreased for the simplified text. This underscores the need for a more nuanced approach that balances SARI optimization with readability considerations. In the case of simplifying NHK articles using BART, I found that focusing solely on improving SARI scores led to reduced readability. This is because the model often removed unnecessary details and preserved more complex ideas, resulting in fewer but more intricate sentences. Such outcomes suggest that models optimized for SARI might prioritize meaning preservation at the expense of text accessibility, which is problematic for language learners. The reduced readability in the simplified NHK texts could be attributed to the model's failure to consider sentence length and vocabulary complexity—key factors in text readability. Consequently, models designed for Japanese text simplification should prioritize comprehension over exact meaning preservation, particularly for new language learners.

Moreover, while the BLEU score achieved in this study was relatively low, it is important to note that this does not necessarily reflect poor performance. BLEU is primarily designed to evaluate machine translation and assumes that the reference text is a high-quality translation. However, a low BLEU score can occur despite effective text simplification, such as when there are fewer reference points. Therefore, while BLEU provides valuable insights, its lower score should be viewed as a secondary concern rather than a definitive indicator of the model's success.

Future Works

There is a notable gap in research concerning the classification of Japanese text difficulty. To enhance text simplification efforts, it is crucial to develop more refined methods for assessing text difficulty, as the current readability metrics fall short in accounting for various factors. For instance, the existing readability formula relies on a dictionary of 17,920 words, so uncommon words not contained within the dictionary fail to negatively impact the difficulty score.

Additionally, although both SBERT and SARI scores suggest high fluency levels, these findings should be validated through human evaluations to ensure their reliability and relevance.

Future avenues for exploring text simplification include combining different methods, such as document-level context, restricted vocabulary lists, and word-level replacement with simpler words.

Conclusion

This research successfully demonstrates the potential of AI-based text simplification for Japanese language learners by utilizing a BART-based model fine-tuned on NHK news articles. Our results show a significant improvement in text summarization quality, as indicated by the high SARI score of 67.178, suggesting that the model effectively preserves essential content while simplifying the text. However, the increase in difficulty scores highlights a critical challenge: while adept at maintaining meaning, the model often produces outputs that are still too complex for learners. This finding suggests that achieving a balance between meaning preservation and text accessibility remains a challenge in AI-driven text simplification.

Moreover, the relatively low BLEU score emphasizes the limitations of traditional machine translation metrics in evaluating text simplification tasks, particularly when the goal is more nuanced than direct translation. The findings suggest that while our approach advances simplification efforts, further research is necessary to refine the balance between simplification and readability, particularly for language learners at various proficiency levels.

In conclusion, this study confirms the efficacy of AI for Japanese text simplification but also underscores the importance of refining methods to better serve the needs of learners. Future work should explore hybrid models that integrate document-level context, restricted vocabulary lists, and human evaluations to enhance the overall effectiveness of text simplification for language acquisition.

References

Cripwell, L., Legrand, J., & Gardent, C. (2023). Context-aware document simplification. *arXiv preprint arXiv:2305.06274*.

Espinosa-Zaragoza, I., Abreu-Salas, J., Lloret, E., Pozo, P. M., & Palomar, M. (2023, September). A review of research-based automatic text simplification tools. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing* (pp. 321-330).

Géron, A. (2019). Hands-on machine learning with Scikit-learn, Keras, and Tensorflow: Concepts, tools, and techniques to build intelligent systems (2nd edition). O'Reilly Media.

Hasebe, Yoichiro and Lee, Jae-Ho (2015) 'Introducing a Readability Evaluation System for Japanese Language Education' Proceedings of the 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J), pp.19-22.

Hayakawa, A., Kajiwara, T., Ouchi, H., & Watanabe, T. (2022). JADES: New Text Simplification Dataset in Japanese Targeted at NonNative Speakers (pp. 179–187). https://doi.org/10.18653/v1/2022.tsar1.17

Kajiwara, T., Nishihara, D., Kodaira, T., & Komachi, M. (2020). Language Resources for Japanese Lexical Simplification日本語の語彙平易化のための言語資源の整備. Journal of Natural Language Processing, 27, 801–824. https://doi.org/10.5715/jnlp.27.801

Lewis, M. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Reimers, N., & Gurevych, I. (2019, November). Sentence-bert: Sentence embeddings using siamese BERT-Networks. https://arxiv.org/abs/1908.10084

Sato, S. (2014). Text readability and word distribution in Japanese (N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis, Eds.; pp. 2811–2815). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/633%3Csub%3EP%3C/sub%3Eaper.pdf

Sunakawa, Yuriko, Lee, Jae-ho, and Takahara, Mari (2012) The Construction of a Database to Support the Compilation of Japanese Learners Dictionaries, Acta Linguistica Asiatica 2(2), pp.97-115

Tanaka , H., & Mino, H. (2018, March). ニュースのためのやさしい日本語とその外国人日本語学習者への効果 | NHK技研R&D. NHK放送技術研究所. https://www.nhk.or.jp/strl/publica/rd/168/5.html

Tanaka, H. (2013, August). Trial Service of Easy Japanese News on the Web | Broadcast Technology. NHK STRL. https://www.nhk.or.jp/strl/english/publica/bt/54/4.html

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & CallisonBurch, C. (2016). Optimizing Statistical Machine Translation for Text. Transactions of the Association for Computational Linguistics, 4, 401–415. https://doi.org/10.1162/tacl_a_00107

Zetsu, T., Kajiwara, T., & Arase, Y. (2023). Controllable Text Simplification Using Lexically Constrained Decoding Based on Edit Operation Prediction編集操作予測に基づく語彙制約付きデコーディングによるテキスト平易化の難易度制御. Journal of Natural Language Processing, 30, 991–1010. https://doi.org/10.5715/jnlp.30.991