A Document-Level Text Simplification Dataset for Japanese

Yoshinari Nagai*, **Teruaki Oka***, **Mamoru Komachi**†
*Tokyo Metropolitan University
6-6 Asahigaoka, Hino, Tokyo, Japan
nagai-yoshinari@ed.tmu.ac.jp, teruaki-oka@tmu.ac.jp
†Hitotsubashi University
2-1 Naka, Kunitachi, Tokyo, Japan
mamoru.komachi@r.hit-u.ac.jp

Abstract

Document-level text simplification, a task that combines single-document summarization and intra-sentence simplification, has garnered significant attention. However, studies have primarily focused on languages such as English and German, leaving Japanese and similar languages underexplored because of a scarcity of linguistic resources. In this study, we devised JADOS, the first Japanese document-level text simplification dataset based on newspaper articles and Wikipedia. Our dataset focuses on simplification, to enhance readability by reducing the number of sentences and tokens in a document. We conducted investigations using our dataset. Firstly, we analyzed the characteristics of Japanese simplification by comparing it across different domains and with English counterparts. Moreover, we experimentally evaluated the performances of text summarization methods, transformer-based text simplification models, and large language models. In terms of D-SARI scores, the transformer-based models performed best across all domains. Finally, we manually evaluated several model outputs and target articles, demonstrating the need for document-level text simplification models in Japanese.

Keywords: Corpus, Text Simplification, Summarization, Wikipedia

1. Introduction

Text simplification is the task of rewriting hard-to-understand text. Most text simplification studies focus on intra-sentence simplification, including lexical simplification (Paetzold and Specia, 2017), simplification with edit operation labels (Dong et al., 2019), and sentence splitting (Narayan et al., 2017). Apart from that, document-level text simplification has garnered significant attention. Several studies have been conducted on this topic, such as creating datasets (Xu et al., 2015), proposing an automatic evaluation metric (Sun et al., 2021), and designing dedicated simplification models (Cripwell et al., 2023). The document-level text simplification task involves improving the readability of the document. The relationship between text length and reading comprehension has been investigated in the case of English (Commander and Stanwyck, 1997; Daley and Rawson, 2018). Baker (2023) demonstrated a positive correlation between overall length (word counts) and average text difficulty based on the ratings of the model essays by the subjects. Therefore, summarization and simplification tasks are frequently combined to create document-level text simplification (Blinova et al., 2023).

Many document-level text simplification datasets have been created to facilitate studies on this topic. News-domain datasets use aligning existing news articles of different complexity levels. Newsela corpus (Xu et al., 2015) was built using Newsela[1] in English, which includes the original articles and multiple simplified versions of those articles. It has been used in several document-level simplification studies (Sun et al., 2023; Zhong et al., 2020). Other languages that use news magazines or telegrams are the 20 Minuten dataset (Rios et al., 2021) in German and DSim (Klerke and Søgaard, 2012) in Danish. Furthermore, in the encyclopedia domain, most datasets use Wikipedia-Simple Wikipedia as a comparable text for simplification. D-Wikipedia (Sun et al., 2021), SWiPE (Laban et al., 2023), and the Klexikon Dataset (Aumiller and Gertz, 2022) are datasets for English and German, respectively. Although the number of datasets available for this task is increasing, few languages are available.

In this study, we created the first **JA**panese **DO**cument-level text **S**implification dataset for the news and encyclopedia domains (JADOS)[2]. The Mainichi Japanese Daily Newspaper (MN) in the news domain and the Japanese Wikipedia[3] in the encyclopedia domain were our sources for complex documents. To obtain simple documents for the former, we used the Mainichi Elementary School Newspaper (MEN), which is a newspaper for elementary school students. Conversely, we prepared the simpler version of the latter manually because of the absence of datasets such as Simple Wikipedia that can provide comparable Japanese texts. Sentence alignment IDs and simplification

---

[1] `https://newsela.com/`

[2] JADOS is available at `https://github.com/tmu-nlp/JADOS`.

[3] `https://ja.wikipedia.org`

LREC-COLING 2024, pages 459–476
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

operation labels are annotated to each complex document sentence in JADOS, which distinguishes it from document-level text simplification datasets for other languages. Excluding the simplification operation labels found in the Wikipedia corpus, these assignments were also conducted manually. We assume that JADOS is primarily used for document-level text simplification tasks. Furthermore, JADOS can support sentence-level text simplification and classify the simplification operation labels for each complex sentence (Cripwell et al., 2023) using the annotated information.

The primary contributions of this study are summarized as follows:

- We created JADOS, the first document-level Japanese text simplification dataset based on newspaper and Wikipedia articles.
- We analyzed the differences in simplification strategies between the news and encyclopedia domains within JADOS and compared them with English text simplification datasets.
- We established the baselines for the document-level Japanese text simplification task using JADOS.

2. Related Works and Background

2.1. Text Simplification Datasets

2.1.1. Non-Japanese Datasets

News domain. Numerous news domain datasets use Newsela, which consists of original news articles and four simplified versions rewritten by professional editors for pre-college students of different grades. For document-level datasets, corpora covering multi-level simplification have been created, such as the English Newsela corpus (Xu et al., 2015), and Newsela-es (Spanish text extracted from this corpus). Moreover, at the sentence level, sentences from article pairs in different levels from the Newsela corpus were aligned using a neural conditional random field aligner, yielding Newsela-AUTO (Jiang et al., 2020).

The OneStopEnglish (Vajjala and Lučić, 2018) corpus is derived from a MacMillan Education-managed website for English language learners. This website hosts articles with the same content rewritten at three different levels of difficulty, enabling multi-level simplification at the document level. In addition to English, the 20 Minuten dataset (Rios et al., 2021) in German contains full-text articles, summaries, and simplified versions from the Swiss news magazine 20 Minuten. Furthermore, DSim (Klerke and Søgaard, 2012) is created in Danish using news telegrams and articles simplified for individuals who are dyslexic or learning Danish[4]. The Chinese dataset CSS (Yang et al., 2023) focuses on sentence-level analysis, where complex sentences are extracted from the PFR corpus, which contains articles from the People's Daily newspaper, while simple sentences are manually created by asking annotators rather than relying on existing newspaper articles.

Several datasets in the news domain use existing parallel data created by professional editors, expecting a high-quality simplification. However, certain datasets, such as the Newsela corpus, have restrictions on their public release, necessitating a usage license.

Encyclopedia domain. The English datasets in the encyclopedia domain were created using comparable texts from English Wikipedia[5] and Simple English Wikipedia[6]. For sentence-level corpora, the PWKP (Zhu et al., 2010) and the Woodsend and Lapata (2011) dataset use the term frequency–inverse document frequency (tf.idf) for sentence alignment between article pairs in these comparable English texts. Furthermore, C&K-1 (Coster and Kauchak, 2011) and C&K-2 (Kauchak, 2013) use dynamic programming for the same purpose. The three datasets mentioned earlier were combined to create WikiLarge (Zhang and Lapata, 2017); except for C&K-1.

However, Xu et al. (2015) raised concerns regarding the low simplification quality of Simple English Wikipedia and the low sentence alignment accuracy in PWKP (Zhu et al., 2010). TurkCorpus (Xu et al., 2016) adopted a subset of PWKP (Zhu et al., 2010) simplified by multiple annotators rather than Simple Wikipedia. Furthermore, by manually simplifying complex sentences in this corpus, small yet high-quality datasets were created, such as HSplit (Sulem et al., 2018), which focuses on sentence splitting, and ASSET (Alva-Manchego et al., 2020), which covers various simplifications (lexical paraphrasing, sentence splitting, and compression). Wiki-AUTO (Jiang et al., 2020) was built using a methodology similar to Newsela-AUTO (Jiang et al., 2020) to reduce the proportion of defective sentence pairs compared with WikiLarge.

English document-level datasets have been created, such as D-Wikipedia (Sun et al., 2021) and SWiPE (Laban et al., 2023). A children-tailored encyclopedia called Klexikon[7] is available online in German. Klexikon dataset (Aumiller and Gertz, 2022) was created using the German Wikipedia[8]

---

[4] `https://www.dr.dk/ligetil`

[5] `https://en.wikipedia.org/`

[6] `https://simple.wikipedia.org`

[7] `https://klexikon.zum.de`

[8] `https://de.wikipedia.org`

and its Simple Wikipedia counterpart, this encyclopedia.

In the encyclopedia domain, the creation of large datasets is possible because of comparable texts from Wikipedia. However, establishing linguistic resources for simplification tasks is a significant challenge because new simplified articles must be created in many languages where Simple Wikipedia is not available.

2.1.2. Japanese Datasets

SNOW T15 (Maruyama and Yamamoto, 2018) is a lexical simplification dataset created by manually rewriting 50,000 sentences extracted from the Tanaka corpus[9] using a defined core vocabulary of 2,000 words. Additionally, SNOW T23 (Katsuta and Yamamoto, 2018) was introduced to increase the size of the previous dataset. JADES (Hayakawa et al., 2022) dataset is intended for independent non-native Japanese speakers and focuses on lexical and grammatical simplification. This dataset consists of pairs of sentences extracted from the Japanese-English development and test subsets in the WMT20 news translation task (Barrault et al., 2020), rewritten by experts using the profiles of the former Level 3 of the Japanese-Language Proficiency Test. The aforementioned datasets are all tailored for simplification at the intra-sentence level rather than at the document level.

Most studies on document-level text simplification in Japanese (Sugai et al., 2020; Morita et al., 2023) have been conducted using article pairs from NHK NEWS WEB[10] and NHK NEWS WEB EASY[11] to create a pseudo-parallel dataset. To cater to foreigners and children, NHK NEWS WEB EASY provides news articles rewritten by journalists and Japanese language teachers to simplify the structures and expressions used in the originals. Although this parallel dataset is expected to be of high quality in terms of simplification, it is not available in the public domain for linguistic resources because of copyright restrictions.

2.2. Preliminary investigation

We conducted a preliminary investigation to analyze Japanese document level simplification operations. To this end, we compared 50 article pairs with identical content extracted from articles published in the MN and MEN in 2019. We defined the inter-sentence simplification operations following Petersen and Ostendorf (2007) based on the corresponding sentence pairs. Our defined operations are detailed in Table 1. Table 2 lists the frequency ratios of these operations in the compared article pairs. The current Japanese sentence-level text simplification datasets do not address the operations `Merge`, `Delete`, and `Insert`, which are specific to document-level simplification. `Delete` was the most frequently used operation in these article pairs. Therefore, similar to English and German, we expected Japanese document-level simplification to require summarization in addition to intra-sentence simplification operations.

Edit	One-to-one correspondence between complex and simple sentences.
Split	Split a single complex sentence into multiple simple ones.
Merge	Merge complex multiple sentences into a single simple one.
Delete	A sentence that exists in a complex document but not in a simple one.
Insert	A sentence that appears only in a simple document, e.g., as supporting detail.

Table 1: Simplification operations in Japanese.

	Edit	Split	Merge	Delete	Insert
Ratio(%)	41.12	6.18	6.18	43.82	2.70

Table 2: Ratios of each simplification operation.

3. Japanese Document-Level Text Simplification Dataset

This section describes JADOS, a Japanese dataset for document-level text simplification in the news and encyclopedia domains. Details of the dataset creation are provided in Appendix 12.2.

3.1. Article Data

3.1.1. Mainichi Corpus

Leading national newspapers in Japan, MN, and MEN, are used as comparable texts for Japanese text simplification. The target audience makes a significant distinction: MN articles are aimed at a general audience, whereas MEN articles are mainly aimed at elementary school children.

Source document. Employing MN as the source documents, the JADOS news domain corpus was created using MN-MEN article pairs, which differed from those used in the preliminary investigation.

Target document. We collected target documents with the same content as the source documents of the MEN articles. According to the preliminary investigation's findings, articles with matching content in MN and MEN have a high rate of surface

---

[9] `http://www.edrdg.org/wiki/index.php/Tanaka_Corpus`

[10] `https://www3.nhk.or.jp/news/`

[11] `https://www3.nhk.or.jp/news/easy/`

word alignment. Therefore, we used the bag-of-words method rather than an embedding-based approach for article alignments.

We generated a frequency vector of words obtained from the morphological analysis of the MN and MEN articles on each day of publication and weighted them using tf.idf. Subsequently, similarity cosine scores between MN articles published one week before and after the publication date of the MEN article were used to automatically extract pairs of articles. Automatically aligned pairs were manually checked to confirm that the content matched[12]. We selected 400 pairs of articles, 50 pairs annually, from the MN and MEN datasets from 2013 to 2020. These selections were based on a similarity score threshold of at least 0.75, established through manual inspection of the article pair's correspondence.

3.1.2. Wikipedia Corpus

We used the Japanese Wikipedia as the source document for the encyclopedia domain. In contrast to the Mainichi corpus, there is no Japanese equivalent to Simple Wikipedia. Therefore, we intended to create a pilot version of the Simple Japanese Wikipedia using the selected articles.

Source document. The source documents comprised the overview sections of 1,944 articles selected as Featured articles[13] (秀逸な記事) or Good articles[14] (良質な記事) as of December 22, 2022. We employed these articles, which were expected to be of higher quality than the other articles because they were peer reviewed.

Target document. Tanaka et al. (2012) administered a reading comprehension test to children using Japanese general news scripts and document-level simplified scripts. Consequently, they demonstrated that simplified articles with few difficult words and low word and sentence counts significantly enhanced elementary school students' comprehension. Therefore, JADOS was designed to integrate document summarization and simplification, aligning with similar datasets for this task in other languages.

The target documents were created through a summarization step followed by a simplification step, as described below. We refer to this sequence of steps as the "simplification process." We assigned two native Japanese-speaking workers through an outside contractor and obtained two simplified documents from each article.

Summarization step.

The summarization step enhances the document's readability. We established the following guidelines to obtain iterative revisions of summarization during the simplification process, making them available for reference, such as Newsela (Xu et al., 2015).

- Summarize the output to be 40% to 60% of the input character count.
- Iterate through this process until the output is reduced to 150 characters[15] or fewer. If the output exceeds this limit, proceed to the simplification step.

Simplification step.

The simplification step enhances the sentence-level comprehension of the document. We established the following guidelines to simplify the source documents intended for elementary school students, aligning them with the simplification strategies from MN to MEN.

- Perform sentence-level text simplification using the simplification operations in Table 1.
- Because appropriate simplification requires knowledge of the article's content, refer to the entire Wikipedia article in question, not just the extracted overview section.
- Paraphrase Kanji characters excluded from the elementary school curriculum. Paraphrase loan words or provide supplementary explanations.
- Readability assessment for elementary school students was submitted to the workers.

3.2. Sentence Alignment and Simplification Operation Annotation

Mainichi corpus. During the pre-processing stage, the articles in JADOS were segmented into sentences, including those ending with "。，！，？", by both manual processing and Bunkai (Hayashibe and Mitsuzawa, 2020), a tool for sentence-boundary disambiguation. We instructed the workers to compare the MN-MEN article pairs and assign the simplification operations listed in Table 1 to each sentence in the source documents, along with the "alignment IDs" of the corresponding sentences

---

[12] In article alignments using bag-of-words because the content may be updated depending on the time of publication (e.g., the damage situation changes), exact matches are not always successful.

[13] `https://ja.wikipedia.org/wiki/Wikipedia:秀逸な記事`

[14] `https://ja.wikipedia.org/wiki/Wikipedia:良質な記事`

[15] This is the length of the Mainichi News Flush article, hoping that we can control the length of the generated output of our simplification model in the future.

label	align. IDs	sent. ID	sentence
I	3	0	ｉＰＳ細胞：さい帯血から　再生医療、コストや時間圧縮　京大研究所 iPS cells: From Umbilical Cord Blood. Regenerative Medicine, Cost, and Time Compression. Research Institute of Kyoto University.
E	1	1	再生医療に使う備蓄用ｉＰＳ細胞を、京都大ｉＰＳ細胞研究所が、新生児のさい帯血を使って作製した。 The Center for iPS Cell Research and Application at Kyoto University created a stockpile of iPS cells from newborn baby umbilical cord blood for use in regenerative medicine.
D		2	ｉＰＳ研への取材で分かった。This was discovered through iPS Cell Institute interviews.
E	2	3	成人の血液から作った場合に比べて遺伝子変異が少なく、高品質という。 It has fewer genetic mutations and higher quality than those made from adult blood.
		1	医療に使うためのｉＰＳ細胞を、京都大学ｉＰＳ細胞研究所が、生まれたばかりの赤ちゃんのへその緒の血液を使って作りました。
		2	大人の血液から作った場合に比べ、高品質だということです。
		3	さい帯血には、血液のもとになる細胞がたくさん含まれています。 Numerous cells in umbilical cord blood are the foundation of blood.

Table 3: Excerpts from our Mainichi corpus article. The upper part denotes an excerpt from an annotated MN article, whereas the lower part denotes an excerpt from the MEN article corresponding to an upper part article.

label	align. IDs	sent. ID	sentence
I		0	夜長姫と耳男 Yonagahime to Mimio
E	1	1	『夜長姫と耳男』（よながひめとみみお）は、坂口安吾の短編小説。 “Yonagahime and Mimio” (Yonagahime to Mimio) is a short story by Ango Sakaguchi.
E	2	2	飛騨の匠の弟子である耳男と、無邪気さと残酷さを併せ持つ長者の娘・夜長姫を中心として説話風に語られており、同じく説話風に書かれた『桜の森の満開の下』と並ぶ傑作として評価されている。 The novel is a narrative-style story about Mimio, a disciple of Hida’s artisan, and Yonagahime, the daughter of the chief, who is both innocent and cruel, and it is considered a masterpiece, much like “Under the Full Bloom of The Cherry Blossom Forest,” which was also written in the same style.
D		3	作品執筆の背景には、随筆「飛騨・高山の抹殺—安吾の新日本地理・中部の巻—」（『安吾新日本地理』の一篇）などに描かれた、安吾の古代史とこの地方への興味・関心がある。 “The Eradication of Hida and Takayama: Ango’s New Geography of Japan, Chubu no Maki” (a part of ”Ango’s New Japan Geography”).
		1	『夜長姫と耳男』（よながひめとみみお）は、坂口安吾の短編小説。
		2	飛騨の匠の弟子である耳男と、無邪気さと残酷さを併せ持つ長者の娘・夜長姫を中心として説話風に語られており、同じく説話風に書かれた『桜の森の満開の下』と並ぶ傑作として評価されている。
		3	<Delete>
		1	『夜長姫と耳男』（よながひめとみみお）は、坂口安吾の短編小説です。
		2	飛騨（ひだ）の匠の弟子である耳男と、無邪気さと残こくさを併せ持つ長者の娘・夜長姫を中心として説話風に語られており、『桜の森の満開の下』と並ぶけっ作として評価されています。

Table 4: Excerpts from our Wikipedia corpus article. The top part denotes an annotated source document, the middle denotes an iterative extractive summary, and the bottom denotes the manually created simplified target document.

in the simple article[16]. Table 3 presents excerpts from the annotated MN and MEN articles.

Wikipedia corpus. To create a parallel dataset in the encyclopedia domain comparable to the Mainichi corpus, we applied identical preprocessing to both the source Wikipedia articles and the manually simplified target documents. In the Wikipedia corpus, sentence alignment information between the original Wikipedia articles and our simplified versions was manually recorded in the “alignment IDs” column during the simplification process. Furthermore, the simplification operation labels were automatically assigned based on their correspondence. Table 4 lists excerpts from annotated source documents from the Wikipedia corpus, an extractive summary during the simplification process, and one of the two target documents.

3.3. Dataset Analysis

This sub-section describes the detailed characteristics of JADOS, incorporating the statistics provided in Table 5 and other relevant metrics.

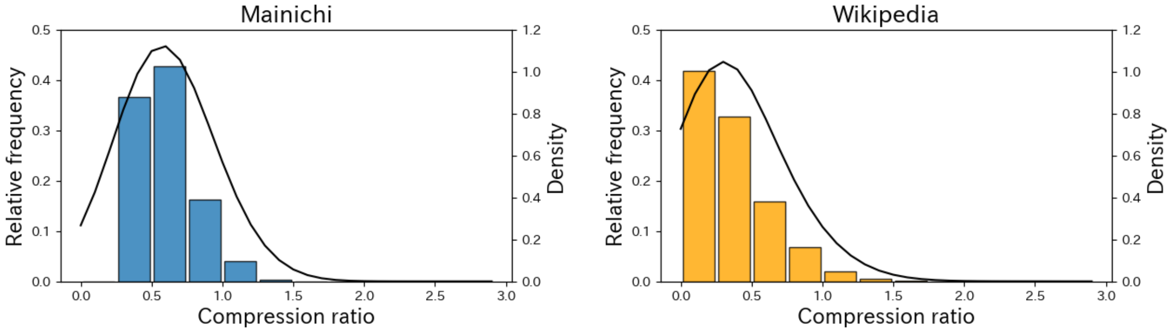
Simplification labels. A common feature observed in both corpora is the predominant use of the `Delete` operation. However, there are significant differences in the proportion of operations, except for `Delete`, particularly concerning `Split` and `Merge`. In the Wikipedia corpus, `Merge` is more common frequently (Mainichi: 25.82% and Wikipedia: 31.58%) than `Split` (Mainichi: 13.56% and Wikipedia: 0.29%). This trend results from Wikipedia target documents that frequently condense key information into single sentences rather than preserving the original text to meet the character count constraints of the simplification step.

Average sizes. In all cases, the average article sizes of source documents in the Wikipedia corpus

---

[16] To record the “sent. ID” of `Insert` in a simple article, a row with “sent. ID” of 0 is provided, and the sentence column contains the article heading.

Figure 1: Relative frequency of document-level compression ratio using the kernel density estimator.



are consistently higher than those in the Mainichi corpus. This discrepancy is probably caused by the space constraints inherent to newspaper articles, as Wikipedia articles do not have such constraints. Conversely, the target documents in the Mainichi corpus are longer because of the character count limit imposed on the Wikipedia corpus. The Wikipedia corpus ($27.19 \rightarrow 20.51$) shows a significant decrease in word count per sentence compared with the Mainichi corpus ($25.60 \rightarrow 24.58$). However, no discernible difference is observed in the target documents' character count per sentence (Mainichi: 45.09, and Wikipedia: 44.69).

		Mainichi	Wikipedia
Original articles (Source)		400	1,944
Reference articles (Target)		400	3,888
Total # of labels	Insert	72	547
	Edit	1,435	10,103
	Delete	1,548	30,324
	Split	337	45
	Merge	642	4,936
Ave. # of sentences	Source	9.90	11.68
	Target	5.86	3.27
Ave. # of words	Source	253.41	317.63
	Target	139.63	67.07
Ave. # of characters	Source	476.13	704.51
	Target	265.11	146.14
		Newsela	EnWiki
Ave. # of sentences	Source	48.86	70.55
	Target	53.30	8.02
Ave. # of words	Source	1,167.06	1,809.75
	Target	609.09	143.07

Table 5: Statistics of JADOS.

	Mainichi	Wikipedia
Source	2.07 (0.65)	2.07 (1.09)
Target	2.47 (0.72)	2.57 (1.03)
% of simpler articles	76.75%	72.99%

Table 6: Readability scores of JADOS. Values in brackets are the population standard deviation.

Compression ratio. To analyze changes in text length at the document level resulting from simplification, we computed the compression ratio by dividing the character count of the target document by that of the source document. Figure 1 illustrates the distribution of the compression ratios of character counts in JADOS and the kernel density estimator following Xu et al. (2015). The compression ratios for both datasets are consistent and followed a normal distribution, indicating uniformity in the simplification practices.

Document-level readability. To evaluate the readability of JADOS, we computed the readability scores using jReadability[17], a Japanese sentence difficulty identification system. The average sentence length and percentage of word classes included were used in multiple regression analysis to derive the readability score formula. Table 6 lists the jReadability scores of JADOS. Across both domains, the target documents are more readable than their source counterparts. Therefore, the manually created Wikipedia target documents were also rewritten in simpler terms.

Sentence-level simplification. We evaluated the extent of rewriting for simplification at the intra-sentence level of JADOS sentence pairs annotated `Edit`, `Split`, or `Merge` in two ways. First, we present the dependency tree depth ratio to assess the complexity of the sentence structure. Syntactic parsing was performed using CaboCha (Taku Kudo, 2002) to determine the depth of each sentence. The ratio was calculated by dividing the depth of the target sentence by that of the source sentence. If there were several sentences, the maximum depth was selected. Second, the Levenshtein distance, which calculates the lowest cost required to convert one string into another through three operations: "Insertions, Deletions, and Substitutions," was used to evaluate the differences in the strings between sentence pairs.

---

[17] `http://jreadability.net`

		Mainichi			Wikipedia		
		Edit	Split	Merge	Edit	Split	Merge
Ave. Dependency tree depth ratio		1.02	1.04	1.02	0.83	0.81	0.87
Ave. of Levenshtein distance	Insertions	6.39	30.37	4.33	2.81	20.53	2.85
	Deletions	10.32	7.26	41.43	27.04	28.11	71.04
	Substitutions	7.60	20.08	15.84	3.06	23.18	6.83

Table 7: Statistics of our Mainichi and Wikipedia corpora at the intra-sentence level.

Table 7 lists the results of the evaluation metrics. The dependency tree depth analysis indicates that applying one of the simplification operations to sentences in the target documents in the Wikipedia corpus results in simpler sentence structures that are easier to interpret. Furthermore, the Levenshtein distance analysis demonstrates that Deletions were frequently used in both domains during `Edit` and `Merge` operations, which generate a single sentence through sentence-level simplification. With regard to `Split`, there are significant differences in the proportion of Deletions across domains. These results indicates that each domain possesses unique characteristics. The Mainichi corpus demonstrates a proactive approach to lexical simplification, as evidenced by its high frequencies of Insertions and Substitutions. Conversely, the Wikipedia corpus emphasizes structural simplification through content selection and deletion, as demonstrated by its high Deletions rate and shallow dependency tree depth.

Comparison to English. We present statistics from the Newsela corpus (Newsela) and English Wikipedia (EnWiki) for comparison to investigate the differences between English and Japanese simplification. Based on the average number of words in Table 5, despite the differences in article sizes, the compression ratios in MN-MEN and Newsela are comparable. Furthermore, we focused on rewriting only the overview sections of Japanese Wikipedia articles, while Simple English Wikipedia rewrote the entire article, achieving a compression ratio twice as high as ours.

To examine changes in sentence length due to intra-sentence-level text simplification, we extracted corresponding sentence pairs based on sentence alignments in JADOS. Across all datasets, simplification resulted in shorter average sentence lengths. Compared with our Wikipedia corpus ($27.20 \rightarrow 20.50$) and EnWiki ($24.71 \rightarrow 20.06$), Newsela compressed sentences to less than half their original length ($27.81 \rightarrow 11.30$), while the sentence length of Mainichi remained ($25.58 \rightarrow 23.85$). Although our Wikipedia has a slightly higher sentence compression than EnWiki, the target lengths are nearly the same.

4. Experiments

This section analyzes the performance of several existing models to establish baseline results. The experimental details are provided in Appendix 12.3.

4.1. Datasets

We created experimental datasets using JADOS. Each MN article, used as a source document in the Mainichi corpus, was assigned a publication-type code (e.g., society, editorial, and sports). Similarly, the Featured and Good articles sourced from the Wikipedia corpus were classified into ten categories, including philosophy and history. Stratified sampling was used to divide JADOS according to the publication-type code and category distribution for each domain (train:dev:test=8:1:1). No distinction was made during the experiments between the use of articles from the Wikipedia corpus and the iterative summaries created during the simplification process as source-target document pairs and iterative summaries-target document pairs.

4.2. Models

We selected two frequently used baseline models for the summarization tasks: the LEAD-3 and ROUGE-2 oracle. Additionally, we used the Luhn (Luhn, 1958) and LexRank (Erkan and Radev, 2004) algorithms as important sentence extraction methods.

We used two transformer-based models, the Japanese BART-based model[18] pretrained on Wikipedia and the T5 model[19] pretrained on mC4[20] and Wikipedia. We fine-tuned these models on the JADOS training dataset.

Furthermore, we conducted experiments with single-shot prompting using three GPT-based large language models: gpt-3.5-turbo-0613[21],

---

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

[19] https://huggingface.co/megagonlabs/t5-base-japanese-web

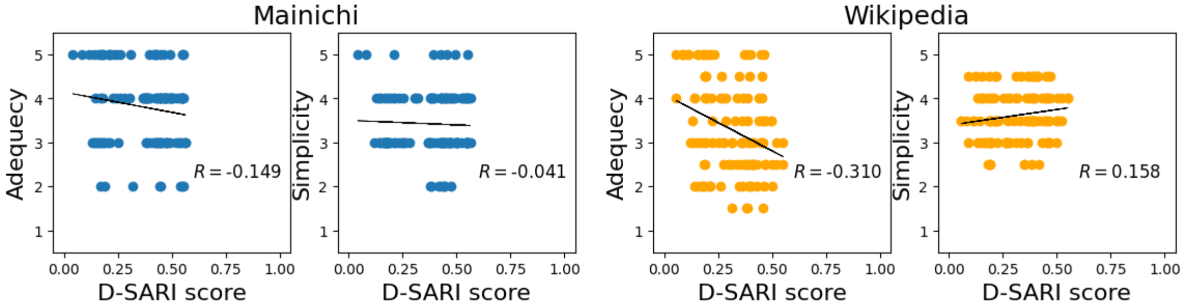
[20] https://huggingface.co/datasets/mc4

[21] https://platform.openai.com/docs/models/gpt-3-5

	Mainichi					Wikipedia				
	Ave. Chars	D-SARI	D_{add}	D_{del}	D_{keep}	Ave. Chars	D-SARI	D_{add}	D_{del}	D_{keep}
LEAD-3	152.12 (30.84)	35.34	0.00	70.53	35.48	185.89 (62.57)	31.59	0.00	62.06	32.71
R-2 oracle	208.88 (40.52)	46.61	0.35	81.44	58.04	177.24 (47.86)	40.03	0.23	72.52	47.34
Luhn	315.68 (58.90)	21.17	0.18	39.14	24.20	225.89 (74.51)	25.64	0.06	52.83	24.03
LexRank	277.15 (47.38)	23.74	0.00	44.47	26.74	198.61 (64.28)	29.14	0.05	58.95	28.42
BART	232.62 (37.60)	41.90	19.06	64.37	42.28	153.08 (19.95)	48.94	29.85	74.79	42.19
T5	195.48 (35.52)	42.50	16.66	69.32	41.52	147.32 (23.71)	46.58	26.68	73.85	39.19
gpt-3.5	429.39 (136.75)	19.61	16.84	29.48	12.50	393.32 (177.98)	21.70	14.65	38.23	12.23
line	108.65 (101.74)	27.32	1.91	65.63	14.42	94.76 (112.22)	28.39	2.55	68.27	14.34
rinna	118.15 (91.07)	28.69	1.42	65.12	19.51	93.60 (113.02)	26.37	2.81	64.67	11.64

Table 8: Average number of characters in the outputs and automatic evaluation results on the Mainichi and Wikipedia test datasets. Values in brackets in "Ave. Chars" represent the population standard deviation.

Figure 2: Scatter plots of D-SARI scores against human evaluation scores.



rinna-3.6B SFT-v2[22], and line-3.6B SFT[23].

4.3. Evaluation Metrics

To evaluate the outputs of each model, we used D-SARI (Sun et al., 2021), a metric derived for document-level text simplification tasks from SARI (Xu et al., 2016). SARI is an automatic evaluation metric that compares a system output with gold references and the input while considering the aspects of "add, keep, and delete." D-SARI applies penalties to these SARI scores based on the number of words and sentences in the output and reference to adjust for the document-level simplification task. We also presented the average number of output characters for each model to investigate whether document length can be controlled through JADOS training.

Moreover, we manually evaluated the outputs of several models selected from 4.2 and the references of 40 articles in each domain on a five-point Likert scale regarding Adequacy (whether the meaning of the source document is preserved) and Simplicity (overall readability and comprehensibility of the article). Native Japanese speakers majoring in computer science conducted human evaluations. The Mainichi corpus was assigned to one person, whereas the Wikipedia corpus was assigned two people to thoroughly investigate the quality of the created target documents.

5. Discussion

Automatic evaluation. Table 8 lists the average character count and the D-SARI scores (including "add, keep, and delete" scores) for the outputs of each model. The ROUGE-2 oracle generated summaries using a greedy approach based on ROUGE-2 scores with gold references and achieved attained the highest D-SARI score among summarization methods in both domains, indicating the upper bound for these methods. This implies that the target documents within JADOS were not necessarily extracted from the beginning of the source documents (as with LEAD-3). Furthermore, by replicating the target documents using important sentence extraction methods such as Luhn and LexRank might include important sentences from the source documents, and sentences intended to support comprehension for elementary school students may be included.

The transformer-based models, which were fine-tuned on JADOS, outperformed summarization methods in terms of D_{add} because they could handle both summarization and simplification. Therefore, the BART and T5 models had the highest scores in their respective domains, excluding the baseline methods. Experiments on the Wikipedia

---

[22] `https://huggingface.co/rinna/japanese-gpt-neox-3.6b-instruction-sft-v2`

[23] `https://huggingface.co/line-corporation/japanese-large-lm-3.6b-instruction-sft`

	Mainichi		Wikipedia		
	Ade.	Sim.	Ade.	Sim.	Ave. Chars
Reference	3.88	3.50	3.30	3.76	147.62
R-2 oracle	3.65	3.30	3.39	3.46	176.80
T5	3.80	3.50	3.15	3.73	138.47
gpt-3.5	3.95	3.48	3.38	3.65	365.12

Table 9: Results from the human evaluation and average character count of the 40 articles in the Wikipedia test dataset.

corpus demonstrated that the output lengths of both models were approximately 150 characters, consistent with the length constraint imposed by the simplification process. Furthermore, the population standard deviations are smaller than those of the Mainichi corpus, indicating that the output lengths are controlled in these models.

The GPT-based models achieved low D-SARI scores similar to the Luhn and LexRank primarily because they generated outputs through single-shot prompting without parameter updates. gpt-3.5 generated outputs with longer average character counts and showed reluctance toward summarization. The outputs of line and rinna outputs were shorter, typically below 150 characters, D_{add} demonstrated a similar degree of extractive summarization methods, indicating reluctance toward extensive rewriting. These distinctions between the models were clear.

Human evaluation. Table 9 lists the results of the human evaluation of reference articles and model outputs. In terms of Adequacy, gpt-3.5 outputs, characterized by their reluctance to summarize, tend to receive higher scores than other methods in both domains because of their preservation of the source document contents. In the Mainichi corpus, the T5 model generated shorter outputs than the ROUGE-2 oracle, but it still received high Adequacy evaluations, possibly because the fine-tuning process captured the characteristics of the training dataset into the outputs. The Wikipedia corpus scored lower than the Mainichi corpus, primarily because of shorter character counts, indicating that important content selection occurs, particularly in the T5 model.

In terms of Simplicity, the T5 model outperformed the other methods, achieving human scores close to the reference in both domains. The ROUGE-2 oracle scored the lowest because extractive summarization methods failed to simplify the input text. Although the gpt-3.5 model scored low in automatic evaluation metrics, it did not show significant differences compared with the other methods. Considering that university students conducted the evaluation, the evaluation criteria for article difficulty may have differed from those for elementary school students. Therefore, the article length may not have significantly influenced the evaluation.

The T5 model and gpt-3.5 achieved scores closer to the references than the ROUGE-2 oracle when the evaluation of Adequacy in the Wikipedia corpus was excluded. This suggests that document-level text simplification models can generate simplified documents while preserving a level of Adequacy similar to that of the references. This highlights the importance of such models.

Figure 2 illustrates the scatter plot that depicts the correlation between human evaluations and D-SARI scores. According to Liu et al. (2023), establishing a human evaluation metric that correlates with automatic evaluation metrics is required based on references. D-SARI evaluates using the reference, source document, and output, whereas human evaluation metrics are solely based on comparing the source document with the output. This probably resulted in a weak correlation.

6. Conclusion

This study addressed the lack of linguistic resources for Japanese document-level text simplification. To achieve this, we created JADOS, a document-level Japanese text simplification dataset that encompass news and encyclopedia domains. The former was extracted from comparable news articles, whereas the latter was developed from the Japanese Wikipedia by manually creating its simplified version along with their alignments. We analyzed the differences in simplification between the two domains in our dataset and compared them with those of the English datasets. Furthermore, we conducted performance evaluation experiments on existing summarization and simplification models. Therefore, transformer-based models fine-tuned on our dataset outperformed extractive summarization methods and GPT-based large language models in both domains. Moreover, human evaluation results demonstrated that a Japanese document-level text simplification model capable of generating simplified articles while preserving comprehension-critical content is required.

Although large language models did not demonstrate high simplification scores, they had favorable scores in human evaluation, indicating room for improvement. This implies that there are numerous valid approaches to simplifying a particular document in the document-level simplification task. In future work, we intend to evaluate document-level text simplification outputs with target people (i.e., elementary school children) and improve our work by leveraging existing intra-sentence text simplification datasets and models.

7. Acknowledgments

This work was supported by the National Institute of Information and Communications Technology (NICT) under the “Research and Development of externally controllable modeling of multimodal information to enhance the accuracy of automatic translation.” This study used the 日本語文章難易度判別システム (`http://jreadability.net`), a product of JSPS KAKENHI Grant Number JP25370573.

8. Ethical Considerations

We outsourced the creation of target documents for the Wikipedia corpus in 3.1.2 and the annotation for the Mainichi and Wikipedia corpus in 3.2. We requested the Mainichi corpus at a rate of approximately 10,785 JPY per day and the Wikipedia corpus at a rate of approximately 39,880 JPY per day. These costs are higher than the local average minimum wage (8,000 JPY per day).

9. Limitations

Our dataset was designed for Japanese, making it unsuitable for use with other languages. Because of the lack of a Simple Wikipedia for Japanese, we manually created the target documents within our Wikipedia corpus. Although we can anticipate high simplification quality, the dataset size is significantly smaller than that of other languages in the same domain. We anticipate that document-level simplification datasets in more languages will be available in the future. Particularly, in situations with limited linguistic resources, we hope that our manual creation process of simplified target documents can provided guidance.

Furthermore, our dataset primarily simplifies the provided text at the document level. Therefore, we do not consider whether the content is suitable for the target readers, as discussed in educational studies. We hope to advance interdisciplinary text simplification research spanning the education and linguistics fields.

10. Bibliographical References

John Baker. 2023. The effects of text length on the readability of model essays. *Gema Online Journal of Language Studies*, 23:60–73.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. 2023. SIMSUM: Document-level text simplification via simultaneous summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9927–9944, Toronto, Canada. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Nannette Evans Commander and Douglas J. Stanwyck. 1997. Illusion of knowing in adult readers: Effects of reading skill and passage length. *Contemporary Educational Psychology*, 22(1):39–52.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Nola Daley and Katherine A. Rawson. 2018. Elaborations in expository text impose a substantial time cost but do not enhance learning. *Educational Psychology Review*, 31:197–222.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Yuta Hayashibe and Kensuke Mitsuzawa. 2020. Sentence boundary detection on line breaks in Japanese. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 71–75, Online. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Hajime Morita, Kojiro Iizuka, and Kosho Kubo. 2023. [Generating Simplified Summaries Indirectly Using Lexical Constraints] Goi Seiyaku wo Kansetsuteki ni Motiita Heii na Youyaku no Seisei (in Japanese). In *29th Annual Meeting of the Association for Natural Language Processing*, pages 3043–3047.

Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. 2017. Split and rephrase. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 606–616, Copenhagen, Denmark. Association for Computational Linguistics.

Gustavo H. Paetzold and Lucia Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60(1):549–593.

Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Proc. Speech and Language Technology in Education (SLaTE 2007)*, pages 69–72.

Naito Sugai, Hitoshi Nishikawa, and Takenobu Tokunaga. 2020. [Summarization and Simplification of News Texts] News Text no Youyaku Oyobi Heiika (in Japanese). In *26th Annual Meeting of the Association for Natural Language Processing*, pages 1455–1458.

Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. Teaching the pre-trained model to generate simple texts for text simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9345–9355, Toronto, Canada. Association for Computational Linguistics.

Yuji Matsumoto Taku Kudo. 2002. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69.

Hideki Tanaka, Hideya Mino, Shinji Ochi, and Motoya Shibata. 2012. Reading comprehension test of simplified japanese in the trial web service of news web easy. *IPSJ SIG Technical Report*, 2012(9):1–9.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. Association for Computational Linguistics.

Sato Toshinori. 2015. Neologism dictionary based on the language resources on the web for mecab.

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.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Li. 2020. Discourse level factors for sentence deletion in text simplification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9709–9716.

11. Language Resource References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German dataset for joint summarization and simplification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2693–2701, Marseille, France. European Language Resources Association.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

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 (TSAR-2022)*, pages 179–187, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Yuta Hitomi, Yuya Taguchi, Hideaki Tamori, Ko Kikuta, Jiro Nishitoba, Naoaki Okazaki, Kentaro Inui, and Manabu Okumura. 2019. A large-scale multi-length headline corpus for analyzing length-constrained headline generation model evaluation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 333–343, Tokyo, Japan. Association for Computational Linguistics.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. 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).

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria. Association for Computational Linguistics.

Sigrid Klerke and Anders Søgaard. 2012. DSim, a Danish parallel corpus for text simplification. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 4015–4018, Istanbul, Turkey. European Language Resources Association (ELRA).

Tomonori Kodaira and Mamoru Komachi. 2018. The rule of three: Abstractive text summarization in three bullet points. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305.

Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. SWiPE: A dataset for document-level simplification of Wikipedia pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10674–10695, Toronto, Canada. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, 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).

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021*

Conference on Empirical Methods in Natural Language Processing, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A new dataset and empirical study for sentence simplification in Chinese. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8306–8321, Toronto, Canada. Association for Computational Linguistics.

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.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.

12. Appendices

12.1. Japanese Summarization Datasets

The summarization task is an important element of document-level text simplification (Blinova et al., 2023). Two lines of Japanese summarization datasets are available.

First, ThreeLineSummaryDataset (Kodaira and Komachi, 2018) was created by collecting pairs of Japanese articles and manually generating three-sentence summaries that were published in Livedoor News[24]. As is the case with our dataset, summary documents are also abstractive and provide a general understanding of the content of an entire article. They differ from our dataset as the length of summary documents is fixed at three and no sentence alignments are provided. Moreover, JAMUL (Hitomi et al., 2019) and XL-Sum (Hasan et al., 2021) datasets are available in the news domain.

Second, the wikiHow dataset (Japanese version)[25] was created based on the Japanese wikiHow, following Koupaee and Wang (2018). The WikiLingua dataset (Ladhak et al., 2020) also contains Japanese summarization. They are both extracted from wikiHow, and resources such as Simplified Wikipedia are not available in Japanese.

None of the aforementioned datasets involve any simplification process. The summarization task aims to retain key information while removing non-essential information, whereas the document-level text simplification contributes to simplification by adding or retaining relevant sentences. The latter helps target readers understand the text by adding explanations using easy-to-understand vocabulary. Therefore, such simplification rewrites might not be included in the summary datasets.

12.2. Dataset Creation Details

12.2.1. Mainichi corpus

Pre-processing of newspaper articles. One characteristic of MEN is its tendency to include Japanese syllabaries after Kanji characters, to aid children still learning pronunciation. As a pre-processing step, remove the round brackets and the parenthetical content from MN and MEN.

Morphological analysis. In aligning MN and MEN articles, we performed a morphological analysis to create a frequency vector for each article word. We used MeCab (Kudo et al., 2004) with mecab-ipadic-NEologd (Toshinori, 2015) for the

---

[24] `https://news.livedoor.com/`

[25] `https://github.com/Katsumata420/wikihow_japanese`

morphological analysis. We extracted verbs, adjectives, and nouns with two or more characters from the pre-processed articles. Numerals and stop words were not included.

12.2.2. Wikipedia corpus

Data collection. We gathered Wikipedia source documents using the `wikipedia` library[26] and filtered out Featured and Good articles unsuitable for corpus creation, such as those written in bulleted format. The JADOS Wikipedia corpus employed the overview sections labeled as 概要 (outline), 概説 (overview), or 概略 (broad outline) from the extracted document. If these sections were not available, the topmost section was substituted.

Extractive summarization. The extractive summaries created in the summarization step of the "simplification process" consistently include the article title.

Reason for manually creating a pilot version of the Simple Japanese Wikipedia. Instead of translating other languages' Wikipedia datasets into Japanese to obtain target documents for the JADOS Wikipedia corpus, we chose manual creation. This decision was made due to concerns about the simplification quality within Simple Wikipedia, as outlined in previous studies (Xu et al., 2015).

12.3. Experimental Details

12.3.1. Pre-processing of Datasets

The Mainichi and Wikipedia corpora used in the experiment underwent conversion to half-width and lowercase characters. However, the dataset prepared for BART was converted to full-width characters and tokenization using JUMAN++ (Tolmachev et al., 2018)[27]. We evaluated after combining the shared words and converting them back to half-width characters.

12.3.2. Model Settings

Text summarization methods. The number of extracted sentences was adjusted to match that of the gold references in the ROUGE-2 oracle. Meanwhile, the important sentence extraction methods (Luhn and LexRank) were determined based on the "Ave. Sent" in Table 5, resulting in six and three sentences for the Mainichi and Wikipedia corpus, respectively.

	BART	T5
learning rate	$1e^{-4}$	$7e^{-4}$
batch size	8	4
# of train epochs	40	20
optimizer	Adam	Adam
LR scheduler type	Linear	Linear
max length	1024	768
repetition penalty	8.0	8.0
# of beams	3	3

Table 10: Hyperparameters for transformer-based models. The upper is for training, and the lower is for inference.

	gpt-3.5	line	rinna
temperature	0.0	1.0	1.0
max tokens	4,096	2,048	2,048
# of beams	-	3	3
repetition penalty	-	3.0	3.0
do sample	-	True	True

Table 11: Hyperparameters for GPT-based models.

Transformer-based models. Table 10 presents the hyperparameters used in the experiments with BART and T5 models. We evaluated using the model from the epoch with the highest ROUGE-Lsum (Lin, 2004) score on the development dataset. Experiments were conducted with three seed values for each model, and Table 8 presents the macro-average D-SARI scores.

GPT-based models. Tables 14 and 15 present the prompts for each domain given to the GPT-based models. While the format of prompts differed slightly by models, we provided the same single-shot instance for each domain. Furthermore, the instruction was consistent across models and domains, concretely: "Please convert the article into a format that is easy for elementary school students to understand (小学生の理解しやすい記事に変換してください。)." Table 11 lists the hyperparameters set for the GPT-based model. The line and rinna generated texts using the decoding strategy known as beam-search multinomial sampling. Truncation was applied to longer prompts to ensure an output space of at least 1,024 tokens for gpt-3.5 and 256 tokens in line and rinna. The GPT-based model's outputs may include line breaks. Since we focused on text rewriting, we eliminated these in the automatic and human evaluations.

12.3.3. Details of Evaluation Metrics

Resources for an automated evaluation metric. For the evaluation of D-SARI, we used the script provided by Sun et al. (2021)[28]. However, as there

---

[26] https://pypi.org/project/wikipedia/

[27] https://github.com/ku-nlp/jumanpp

[28] https://github.com/RLSNLP/Document-level-text-simplification/blob/

5	The entirety of the source document's content is retained.
4	While preserving SM and CHE, certain sections include content unrelated to the subject matter.
3	Only SM and CHE are retained, while content deviating from the subject matter is removed.
2	SM and CHE have been partially removed.
1	SM and CHE are completely lost, and the content has changed to something different.

Table 12: Evaluation criteria for Adequacy.

5	An article is simpler than the source document and understandable for elementary school students.
4	An article is simpler than the source document, but certain parts are still somewhat challenging to comprehend.
3	There is no difference in terms of simplicity compared to the source document.
2	The article is more challenging to understand than the source document, although some parts might feel easier to comprehend.
1	The article is more difficult to understand than the source document. Also, it is difficult for primary school students to comprehend.

Table 13: Evaluation criteria for Simplicity.

was no support for Japanese, we conducted word and sentence counting using MeCab (Kudo et al., 2004) with ipadic-NEologd (Toshinori, 2015) and Bunkai (Hayashibe and Mitsuzawa, 2020).

Criteria for human evaluation. Tables 12 and 13 list the evaluation criteria for Adequacy and Simplicity, respectively. Regarding Adequacy, we defined criteria that concentrate on the presence of the subject matter (SM) and content helpful for elementary school students' comprehension (CHE) from the source document. We considered a rating of 3 points as the highest, ideally consisting only of the main subject and supporting text for comprehension. In terms of Simplicity, we established criteria to evaluate whether the content is easily understandable for elementary school students and whether it is simpler than the source document.

12.4. Additional Analysis of Human Evaluation

12.4.1. Agreement Degree between Workers

We measured the degree of agreement between workers of the human evaluations of the JADOS Wikipedia corpus by Cohen's κ coefficients (Cohen, 1960). The Adequacy score was 0.43, indicating a weak agreement in determining which content to retain and discard during simplification. Conversely, the Simplicity score exhibited low agreement at 0.17. The cause of this is likely due to the human metric scores being influenced by detailed judgments at the word and syntax levels, as indicated in Table 13. However, it was rare to find differences of 2 points or more between annotators (7 out of 200 cases), indicating that substantial differences were not found in human evaluations.

The correlation between Adequacy and Simplicity scores was -0.07 for the Mainichi corpus and -0.18 for the Wikipedia corpus, respectively. These results indicate that human evaluations were independent of each other.

12.4.2. Correlation with D-SARI and Compression Ratios

Figure 3 illustrates the relationship between human evaluation scores and D-SARI add, delete, and keep scores, respectively. Adequacy and D_{del} exhibit a strong correlation among them. The regression line also suggests a relationship where, as D_{del} increases to correctly selecting deletion contents, Adequacy approaches the highest rating of 3. As illustrated in Figure 2, D-SARI has similar negative correlations, where D_{del} influences the magnitude of the score.

Figure 4 presents a scatter plot of human evaluation and compression ratios. In terms of Adequacy, both domains showed positive correlations, as higher compression ratios indicated a greater probability of retaining the content from the source document. However, as indicated in the evaluation criteria in Table 12, the assessment may vary depending on whether the deleted content contributes to the main subject or aids comprehension. Therefore, a strong correlation was not observed.

For Simplicity, the Mainichi corpus showed close to no correlation, whereas the Wikipedia corpus exhibited a negative correlation. The Wikipedia articles offer a comprehensive description of the details and events related to the title. However, this result indicates that readers may find it easier to understand articles when presented with selectively chosen content, even if the amount of information provided decreases. This implies that summarization could lead to improved readability. In the Simplicity evaluation in Table 9, the notable distinction between gpt-3.5 and the reference articles or the T5 model in the Mainichi corpus suggests that the characteristics of this domain may be influencing the results.

12.5. Cross-Dataset Evaluation

We examined the generalization capabilities of the BART and T5 models fine-tuned in Section 4 to different domains. The outputs were generated

---

`main/D_SARI.py`

ユーザー: 小学生の理解しやすい記事に変換してください。<NL>防災の日の1日、各地で防災訓練が実施された。政府の総合訓練では、東京23区を震源とするマグニチュード7・3で最大震度7の首都直下地震を想定。首相官邸で緊急災害対策本部会議を開き、災害時の対応手順や東京都との連携体制などを確認した。相模原市では、同市を震源とするm7クラスの地震を想定し、9都県市による合同訓練もあった。この日は26都道府県で訓練が実施される予定で、約234万7000人が参加する見通し。政府の総合訓練は首都機能の維持に重点を置き、午前7時10分に地震発生との想定で始まった。各閣僚は赤坂宿舎などから徒歩で首相官邸に集まり、対策本部会議に参加。都庁とのテレビ会議で、舛添要一知事から「大規模な延焼火災や建物崩壊が発生している」と被害状況などの報告があり、政府調査団の派遣を決定するなどした。安倍晋三首相は訓練の一環の記者会見で「建物の倒壊の恐れがある場合は、近隣で声を掛け合って安全を確保してください」などと呼びかけた。一方、相模原市の合同訓練はjr相模原駅前の在日米軍施設であり、東京や埼玉など4都県と千葉市など5政令指定都市の職員や相模原市民らが参加した。内閣府によると、防災週間の間、44都道府県で訓練が実施され、約1645万人の参加が見込まれる。<NL>システム: 防災の日の1日、各地で防災訓練がありました。政府の総合訓練では、東京23区を震源とするマグニチュード7.3で最大震度7の首都直下地震が、午前7時10分に発生したという想定で始まりました。各閣僚は歩いて首相官邸に集まり、対策本部会議に防災服姿で参加しました。都庁とのテレビ会議で、舛添要一都知事から被害状況などの報告を受け、政府調査団の派遣を決めるなどしました。この日は26都道府県で訓練があり、約234万7000人が参加しました。内閣府によると、防災の日を含めた防災週間の期間中、44都道府県で訓練があり、約1645万人が参加する見込みです。<NL>ユーザー: {A MN article you want to simplify.}<NL>システム:
User: Please convert the article into a format that is easy for elementary school students to understand.<NL>On the 1st of Disaster Prevention Day, disaster drills were conducted in various locations. During the government's comprehensive drill, a scenario involving a magnitude 7.3 earthquake originating from the Tokyo 23 wards, with a maximum seismic intensity of seven, directly beneath the capital, was simulated. At the Prime Minister's Office, an emergency disaster response headquarters meeting was convened to confirm disaster response procedures and coordination with the Tokyo Metropolitan Government. In Sagamihara, a joint drill involving nine prefectures and cities was executed in preparation for a magnitude 7-class earthquake centered in the city. Drills were scheduled in 26 prefectures on this day, with an estimated 2,347,000 participants. The government's comprehensive drill emphasized the continuity of capital functions and commenced at 7:10 AM under the assumption of an earthquake occurrence. Ministers convened at the Prime Minister's Office, making their way on foot from sites like the Akasaka Dormitory, to attend the response headquarters meeting. During a teleconference with the Tokyo Metropolitan Government, they received reports from Governor Yoichi Masuzoe on the extent of the damage, including a large-scale spreading fire and collapsed buildings, and decided to dispatch a government investigation team. In a press conference as part of the drill, Prime Minister Shinzo Abe urged the public to ensure their safety by communicating with each other in their neighborhoods if there is a risk of a building collapsing. On the other hand, Sagamihara City's joint drill occurred at a US military facility in Japan near JR Sagamihara Station. It was attended by officials from four prefectures, Tokyo and Saitama among them, and five ordinance-designated cities, including Chiba City, along with Sagamihara citizens. According to the Cabinet Office, drills are scheduled in 44 prefectures during Disaster Reduction Week, with an estimated 16.45 million participants.<NL>System: On the 1st of Disaster Prevention Day, disaster drills were conducted in various locations. During the government's comprehensive drill, a scenario involving a magnitude 7.3 earthquake originating from the Tokyo 23 wards, with a maximum seismic intensity of seven, directly beneath the capital at 7:10 AM, was simulated. The ministers converged at the Prime Minister's office on foot and proceeded to participate in the response headquarters meeting wearing their emergency uniforms. During a video conference with the Tokyo Metropolitan Government, they received reports on the extent of the damage from Governor Yoichi Masuzoe and decided to dispatch a government investigation team. Drills were conducted in 26 prefectures on this day, with an estimated 2,347,000 participants. According to the Cabinet Office, drills are scheduled in 44 prefectures during Disaster Reduction Week, which includes Disaster Prevention Day, with an estimated 16.45 million participants.<NL> User: {A MN article you want to simplify.}<NL>System:

Table 14: Single-shot prompting for simplifying MN articles with rinna 3.6B SFT-v2.

ユーザー: 小学生の理解しやすい記事に変換してください。 南硫黄島原生自然環境保全地域(みなみいおうとうげんせいしぜんかんきょうほぜんちいき)は、自然環境保全法に基づき1975年(昭和50年)5月17日に指定された日本の原生自然環境保全地域。 南硫黄島(東京都小笠原村)全域が指定されており、これまで人間の影響が希薄であったことにより原生の自然がよく保たれている。 原生自然環境保全地域の中では唯一全域が立入制限地区とされている。 システム: 自然環境保全法に基づき1975年5月17日に指定された日本の原生自然環境保全地域です。 南硫黄(みなみいおう)島(東京都小笠原村)全域が指定されており、これまで人間の影響が希薄(きはく)であったことにより原生の自然がよく保たれています。 原生自然環境保全地域の中では唯一全域が立入制限地区とされています。 ユーザー: {A Wikipedia article you want to simplify.}
User: Please convert the article into a format that is easy for elementary school students to understand. The Minami-Iwojima Primitive Nature Conservation Area (Minami-Iōjima gensei shizen kankyō hozen chiiki) is a conservation area established in Japan under the Nature Conservation Law on May 17, 1975 (Showa 50). Encompassing the entirety of Minami-Iwojima (located within Ogasawara Village, Tokyo), this area has been designated, and its pristine natural environment has been meticulously preserved, largely owing to minimal human intervention. Of all the areas designated for the conservation of pristine natural environments, Minami-Iwojima is unique in that access is restricted across its entire expanse. Among all the regions designated for the Primitive Nature Conservation Area, Minami-Iwojima stands out for its comprehensive restriction of access throughout its entire territory. System: The area was designated as a Primitive Nature Conservation Area in Japan under the Nature Conservation Law. Encompassing the entirety of Minami-Iwo (Minami-Iō) jima (located within Ogasawara Village, Tokyo), this area has been designated, and its pristine natural environment has been meticulously preserved, largely owing to minimal (kihaku) human intervention. Within the designated area of the primeval natural environment conservation area, Minami-Iwo Island is the only region where access is restricted throughout. Among all the regions designated for the Primitive Nature Conservation Area, Minami-Iwojima stands out for its comprehensive restriction of access throughout its entire territory. User: {A Wikipedia article you want to simplify.}

Table 15: Single-shot prompting for simplifying Wikipedia articles with line-3.6b SFT.

using test data of different domains from the training data and evaluated with D-SARI (Sun et al., 2021).

Table 16 lists the results of the cross-dataset evaluation. In both domains, evaluating the test dataset in the same domain as the training dataset yielded superior D-SARI scores. Comparable results were consistently observed for the add, delete, and keep metrics.

BART and T5 models have been pre-trained on Wikipedia. Therefore, the model fine-tuned on the Mainichi corpus exhibits a smaller difference in D-SARI scores across test dataset domains, indicating its superior generalization performance. In contrast, models fine-tuned on the Wikipedia corpus tended to struggle in the news domain. However, the output lengths are limited to around 150 characters regardless of the test dataset domain.

		train Mainichi					train Wikipedia				
		Ave. Chars	D-SARI	D_{add}	D_{del}	D_{keep}	Ave. Chars	D-SARI	D_{add}	D_{del}	D_{keep}
BART	-test Mainichi	232.62 (37.60)	41.90	19.06	64.37	42.28	151.38 (15.86)	31.74	19.39	53.36	22.48
	-test Wikipedia	256.54 (71.19)	38.99	9.57	71.28	36.13	153.08 (19.95)	48.94	29.85	74.79	42.19
T5	-test Mainichi	195.48 (35.52)	42.50	16.66	69.32	41.52	139.31 (17.85)	35.61	7.85	68.63	30.35
	-test Wikipedia	183.33 (65.84)	36.62	16.78	64.47	28.62	147.32 (23.71)	46.58	26.68	73.85	39.19

Table 16: Average number of characters in outputs and automatic evaluation results in the cross-dataset evaluation. Values in brackets in “Ave. Char” are the population standard deviation. (Grey: results presented in Table 8; black: results of cross-dataset evaluation)

Figure 3: Scatter plots of D-SARI add, delete, and keep scores against human evaluation scores.

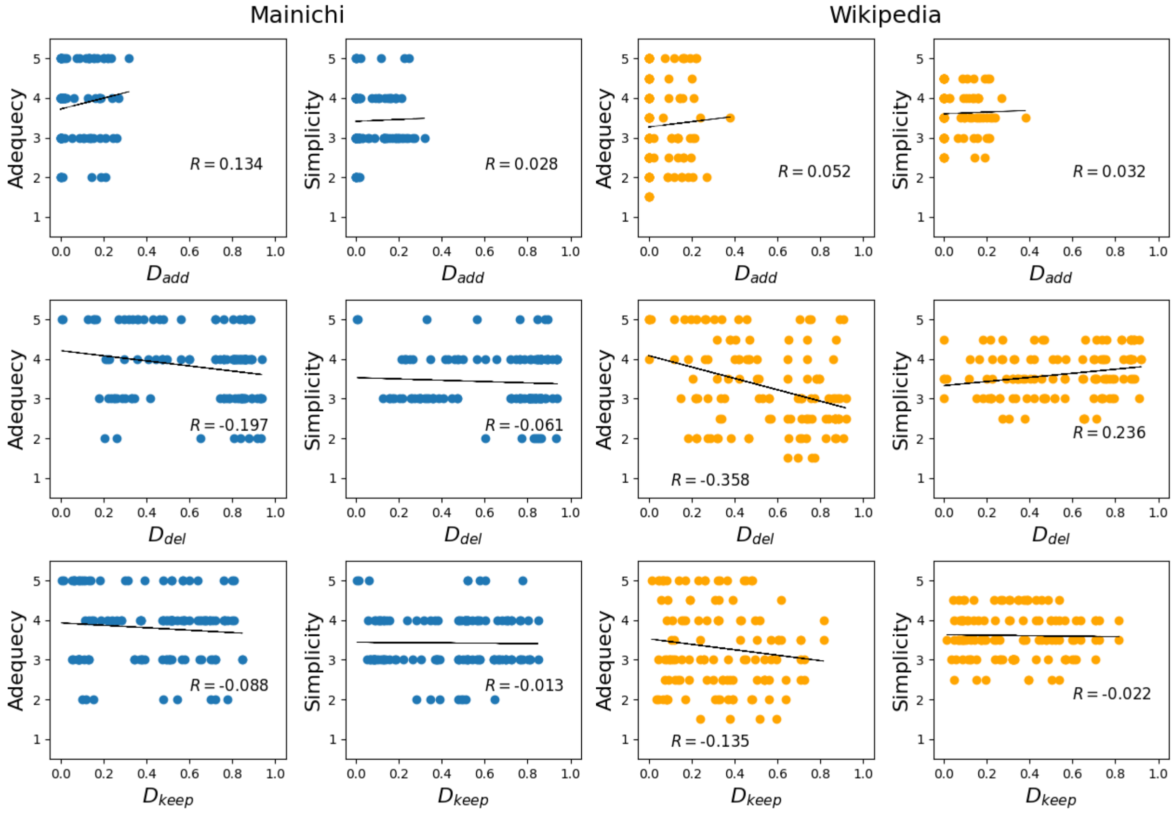
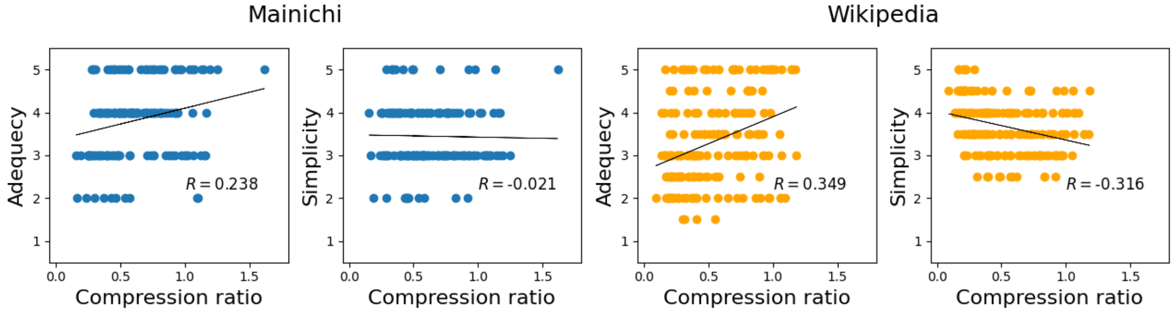


Figure 4: Scatter plots of compression ratio against human evaluation scores.



Source	「黄金虫」(おうごんちゅう / こがねむし、原題:the gold-bug)は、1843年に発表されたエドガー・アラン・ポーの短編小説。語り手とその聡明な友人ルグラン、その従者のジュピターが、宝の地図を元にキャプテン・キッドの財宝を探し当てるまでを描く冒険小説である。また厳密には推理小説の定義からは外れるものの、暗号を用いた推理小説の草分けとも見なされている。この作品は『フィラデルフィア・ダラー・ニュースペーパー』の懸賞で最優秀作となり、ポーは賞金として100ドルを得た。これはポーが単独作品で得た収入ではおそらく最高額である。「黄金虫」はポーの作品のうち、彼の存命中もっとも広く読まれた作品となり、暗号というトピックを出版界に広く知らしめる役割を果たした。 “Ogonchu” (Ogonchu / Koganemushi, original title: the gold-bug) is a short story authored by Edgar Allan Poe, initially published in 1843. This adventure novel follows the narrator, his brilliant friend Legrand, and his servant Jupiter as they embark on a quest for Captain Kidd’s treasure, guided by a treasure map. Although not strictly categorized as a mystery novel, this is considered a pioneering work in integrating cryptography. The story received the first prize in the “Philadelphia Dollar Newspaper” contest, awarding Poe a cash prize of $100. This income is likely the highest Poe ever received for a single work. “Ogonchu” is the most widely read work during his lifetime and played a significant role in popularizing the topic of cryptography to the publishing industry.
Reference 1	1843年に発表されたエドガー・アラン・ポーの短編小説です。語り手とその友人、その従者が、キャプテン・キッドの財宝を探し当てるまでを描く冒険小説です。暗号を用いた推理小説の草分けともされています。『フィラデルフィア・ダラー・ニュースペーパー』の懸賞で最優秀作となり、ポーは賞金100ドルを得ました。 This short story was authored by Edgar Allan Poe and initially published in 1843. This adventure novel follows the narrator, his friend, and his servant as they embark on a quest for Captain Kidd’s treasure. This story is considered a pioneering work in integrating cryptography. The story received the first prize in the “Philadelphia Dollar Newspaper” contest, awarding Poe a cash prize of $100.
Reference 2	「黄金虫」(おうごんちゅう/こがねむし)は、1843年に発表されたエドガー・アラン・ポーの短編小説です。語り手とその友人たちが、宝の地図を元にキャプテン・キッドの財宝を探し当てるまでを描く冒険小説です。暗号を用いた推理小説の草分けと見なされており、暗号というトピックを出版界に広く知らしめました。 “Ogonchu” (Ogonchu / Koganemushi) is a short story authored by Edgar Allan Poe, initially published in 1843. This adventure novel follows the narrator, and his friends as they embark on a quest for Captain Kidd’s treasure. This story is considered a pioneering work in integrating cryptography into mystery novels and widely popularized the topic of cryptography in the publishing industry.
BART	「黄金虫」（おうごんちゅう ／ こがねむし）は、１８４３年に発表されたエドガー・アラン・ポーの短編小説です。語り手とその聡明な友人ルグラン、その従者のジュピターが宝の地図を元にキャプテン・キッドの財宝を探し当てるまでを描く冒険小説で、彼の存命中もっとも広く読まれた作品となりました。 “Ogonchu” (Ogonchu / Koganemushi) is a short story authored by Edgar Allan Poe, initially published in 1843. This adventure novel, which follows the narrator, his brilliant friend Legrand, and his servant Jupiter as they embark on a quest for Captain Kidd’s treasure, guided by a treasure map, is the most widely read work during his lifetime.
T5	1841年に発表されたエドガー・アラン・ポーの短編小説です。語り手とその聡明な友人ルグラン、その従者のジュピターが、宝の地図を元にキャプテン・キッドの財宝を探し当てるまでを描く冒険小説で、暗号を用いた推理小説の草分けとも見なされています。「黄金虫」はポーの作品のうち最も広く読まれた作品となり、暗号というトピックを出版界に広く知らしめる役割を果たしました。 This short story was authored by Edgar Allan Poe, initially published in 1841. This adventure novel, which follows the narrator, his brilliant friend Legrand, and his servant Jupiter as they embark on a quest for Captain Kidd’s treasure, guided by a treasure map, is considered a pioneering work in integrating cryptography into mystery novels. “Ogonchu” is the most widely read work during his lifetime and played a significant role in popularizing the topic of cryptography to the publishing industry.
gpt-3.5	「黄金虫」という作品は、エドガー・アラン・ポーが1843年に発表した短編小説です。物語は、語り手とその賢い友人ルグラン、そして従者のジュピターが、宝の地図を使ってキャプテン・キッドの財宝を見つけるまでの冒険を描いています。この作品は、厳密には推理小説ではないですが、暗号を使った推理小説の先駆けとしても知られています。実は、この作品は「フィラデルフィア・ダラー・ニュースペーパー」という新聞のコンテストで最優秀作品に選ばれ、ポーは100ドルの賞金を受け取りました。これは、ポーが一人で書いた作品で得たお金の中で、おそらく最も多い額です。「黄金虫」は、ポーの作品の中でもっとも広く読まれた作品であり、暗号というテーマを出版業界に広める役割を果たしました。 “Ogonchu” is a short story authored by Edgar Allan Poe, initially published in 1843. The story follows the adventures of the narrator, his brilliant friend Legrand, and his servant Jupiter as they embark on a quest for Captain Kidd’s treasure, guided by a treasure map. Although not strictly a mystery novel, this is known as a pioneering work in integrating cryptography into mystery novels. In fact, the story received the first prize in the “Philadelphia Dollar Newspaper” contest, awarding Poe a cash prize of $100. This income is likely the highest Poe ever received for a single work he wrote independently. “Ogonchu” is the most widely read work and played a significant role in popularizing the theme of cryptography within the publishing industry.
line	エドガー・アラン・ポーの「黄金虫」は、1843年に発表された作品です。 Edgar Allan Poe’s “Ogonchu” is a story initially published in 1843.
rinna	「黄金虫」は、1843年にエドガー・アラン・ポーによって発表された短編小説であり、語り手とその聡明な友人ルグラン、その従者のジュピターが、宝の地図を元にキャプテン・キッドの財宝を探し当てるまでを描く冒険小説である。 “Ogonchu,” a short story published by Edgar Allan Poe in 1843, is an adventure novel that follows the narrator, his brilliant friend Legrand, and his servant Jupiter as they embark on a quest for Captain Kidd’s treasure, guided by a treasure map.

Table 17: Examples of outputs generated by transformer and GPT-based models in the Wikipedia corpus.