A Document-Level Text Simplification Dataset for Japanese

Yoshinari Nagai[1], Teruaki Oka[1], Mamoru Komachi[2]
[1]Tokyo Metropolitan University, [2]Hitotsubashi University

LREC-COLING 2024

Introduction

Text simplification:

- The task of rewriting hard-to-understand text with reduced lexical and grammatical complexity.
- Most text simplification studies focus on intra-sentence simplification.

トルコの最大都市イスタンブールで5月末に始まった反政権デモは2日夜から3日にかけて再燃、国内200カ所以上に拡大した。

トルコの最大都市イスタンブールで5月末から、反政権デモが続いています。
2日夜から3日にかけて、デモは国内200か所以上に拡大。

Document-level text simplification task:

- The document-level text simplification task involves improving the readability of the document.
- Several studies have been conducted on this topic.
  - creating datasets, proposing an automatic evaluation metric, and designing dedicated simplification models

Facilitating studies

Japanese text simplification datasets

usage data	Created datasets	
	Document level	Sentence level
Tanaka corpus	-	SNOW T15, SNOW T23
The Japanese-English development and test subsets in the WMT20 news translation task	-	JADES

Most studies on document-level text simplification in Japanese：

- Using article pairs from NHK NEWS WEB and NHK NEWS WEB EASY to create a pseudo-parallel dataset. (Sugai et al., 2020; Morita et al., 2023)
- It is not available in the public domain for linguistic resources because of copyright restrictions.

Main contributions

- 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.
- We established the baselines for the document-level Japanese text simplification task using JADOS.

Preliminary investigation

Usage data:

- Mainichi Japanese Daily Newspaper (MN)
- Mainichi Elementary School Newspaper (MEN)

Defined 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.

Creating a Japanese dataset for document-level text simplification

Mainichi corpus (news domain):

usage data (source/target)	the MN and MEN datasets from 2013 to 2020.
article alignment	a cosine similarity score threshold of at least 0.75 using bag-of-words
dataset size	400 pairs of articles (50 pairs annually)

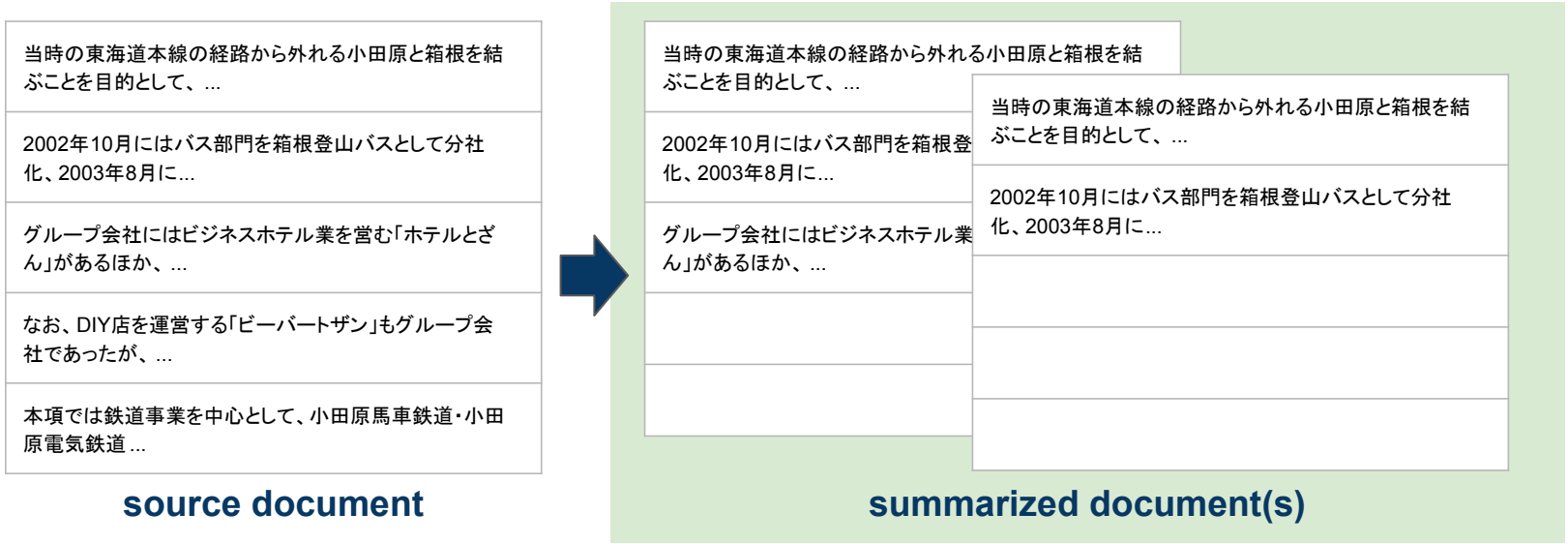
Wikipedia corpus (encyclopedia domain):

usage data (source)	the overview sections of 1,944 articles selected as Featured articles or Good articles.
dataset size	1,944 articles

- There is no Japanese equivalent to Simple Wikipedia.
- We intended to create a pilot version of the Simple Japanese Wikipedia using the selected articles.

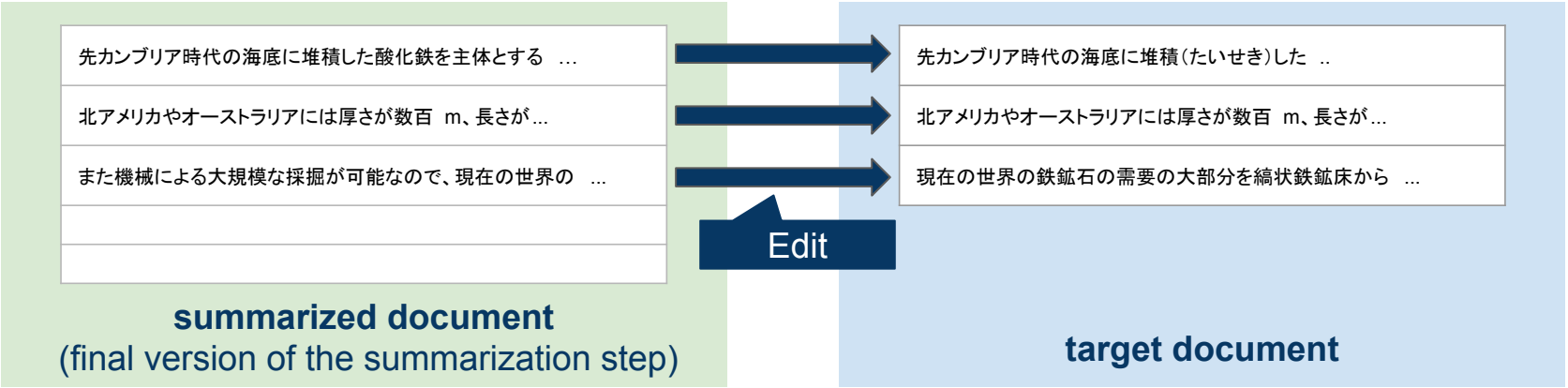
Creating target documents in the encyclopedia domain

1. Summarization step
   - 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 or fewer.
   - If the output exceeds this limit, proceed to the simplification step.



Creating target documents in the encyclopedia domain

2. Simplification step
   - Perform sentence-level text simplification using the simplification operations.
   - Because appropriate simplification requires knowledge of the article's content, refer to the entire Wikipedia article in question.
   - 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.



Excerpts from JADOS Wikipedia corpus article

	label	align. IDs	sent. ID	sentence
Wikipedia article (source)	I		0	夜長姫と耳男
	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。
	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。
	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。
summarization step			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>
simplification step (target)			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。

We annotated based on comparison with the target document.

・defined simplification operations
・corresponding sentence ID

Dataset Analysis

Statistics of each domain in JADOS

		Mainichi	Wikipedia
Total # of articles	Source	400	1,944
	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

Proportion of simplification operation labels:

- the predominant use of the *Delete* operation.
- *Split* (Mainichi: 13.56%, Wikipedia: 0.29%)
- *Merge* (Mainichi: 25.82%, Wikipedia: 31.58%)

>> The Wikipedia corpus tended to extract only the important parts of each sentence and combine them into a single sentence.

Average sizes:

- Target documents were summarized shortly.
- word count per sentence:
  - Mainichi (25.60 → 24.58)
  - Wikipedia (27.19 → 20.51)

Dataset Analysis

Readability of each domain in JADOS

jReadability：a Japanese sentence difficulty identification system

Readability Score = - 0.056 ×（average length of sentence）- 0.126 ×（% of *kango*）
- 0.042 ×（% of *wago*）- 0.145 ×（% of verbs）- 0.044 ×（% of particles）+ 11.724

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

- Across both domains, the target documents are more readable than their source counterparts.
- The manually created Wikipedia target documents were also rewritten in simpler terms.

Experiments *Analysis of the performance of existing models to establish baseline results*

Dataset:

- JADOS was divided according to the publication-type code and category distribution for each domain (train:dev:test=8:1:1).

Models:

extractive summarization methods	**LEAD-3**, **ROUGE-2 oracle** : baseline models for the summarization tasks **Luhn** , **LexRank** : important sentence extraction methods
transformer-based models	**Japanese BART-based model** and **T5 model** fine-tuned on JADOS training data
GPT-based models	single-shot prompting using three GPT-based large language models (**gpt-3.5-turbo-0613**, **rinna-3.6B SFT-v2**, **line-3.6B SFT**)

Results

Automatic evaluation

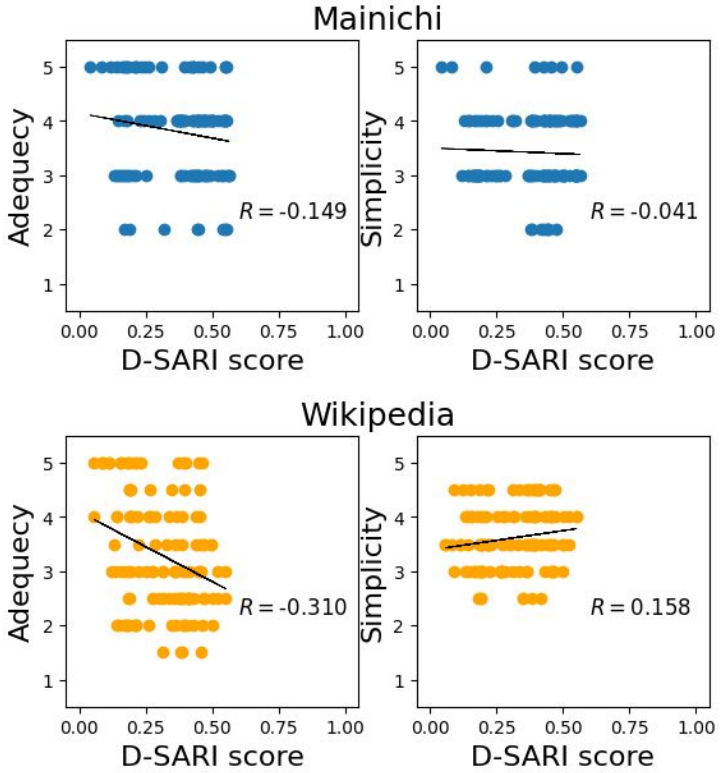
	Mainichi					Wikipedia				
	Ave. Chars	D-SARI	Dadd	Ddel	Dkeep	Ave. Chars	D-SARI	Dadd	Ddel	Dkeep
LEAD-3	152.12 (30.84)	35.34	0.00	70.53	35.48	185.89 (62.57)	31.59	0.00	62.06	32.71
ROUGE-2 oracle	208.88 (40.52)	**46.61**	0.35	81.44	58.04	177.24 (47.86)	40.03	0.23	72.52	47.34
Luhn	315.68 (58.90)	21.17	0.18	39.14	24.40	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

The BART and T5 models had the highest scores in their respective domains, excluding the baseline methods.

Results

Human evaluation

	Mainichi		Wikipedia	
	Adeq.	Simp.	Adeq.	Simp.
Reference	3.88	3.50	3.30	3.76
ROUGE-2 oracle	3.65	3.30	3.39	3.46
T5	3.80	3.50	3.15	3.73
gpt-3.5	3.95	3.48	3.38	3.65



- The T5 model and gpt-3.5 can generate simplified documents while preserving a level of Adequacy similar to that of the references.
- There was no strong correlation observed between automatic and human evaluations.

Conclusions

- We created JADOS, a document-level Japanese text simplification dataset that encompass news and encyclopedia domains.

- Performance evaluation experiments were conducted on existing summarization and simplification models.
  - Transformer-based models fine-tuned on our dataset outperformed extractive summarization methods and GPT-based large language models in both domains.

Future work:

- We intend to improve our work by leveraging existing intra-sentence text simplification datasets and models.