

A Japanese News Simplification Corpus with Faithfulness

Toru Urakawa, Yuya Taguchi, Takuro Niitsuma and Hideaki Tamori (The Asahi Shimbun Company)
LREC-COLING 2024

- In Text Simplification, automated models may introduce unwanted content or omit essential details.
- Furthermore, existing simplified corpora contain instances of low faithfulness to the original text.
- Motivated by this issue, we present a new Japanese simplification corpus designed to prioritize faithfulness.

Corpus Design

Data and Annotators

- 690 news articles were simplified by 30 experts in Japanese language education.

Simplification Level

- JLPT N3, indicating the ability to understand everyday Japanese.

Faithfulness vs. Readability

- Established guidelines to handle a trade-off of Faithfulness vs. Readability.

まずは年 300 台の有効活用をめざす。

まずは一年に 300 台をうまく使うことを目標にします。

We aim to effectively utilize **300 units annually** at first.

The first goal is to successfully use **300 units a year**.

この日の閣僚会合で要請を決めた。

この日の閣僚の会議でお願いすると決めました。

The request was decided at **the ministerial** meeting on this day.

At a meeting of **the ministers** on this day, they decided to ask for it.

出産は帝王切開になる。

子どもは、帝王切開で産みます。

The birth will be by **cesarean section**.

The child is delivered by **cesarean section**.

	Ours	MATCHA	SNOW	JADES
#Sentence Pairs	7,075	16,000	84,300	3,907
Data Source	News Articles	Tourism Articles	Textbook	News Articles
% of Entity Retention	77.7	70.7	63.8	54.8
-PERSON	93.5	74.4	84.3	78.4
-DATE	87.2	77.6	67.5	82.9
-CITY	88.1	82.0	83.2	77.7

☑ Faithfulness

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

☑ Readability

% of simp. is easier	Avg. of readability
78.2	4.01

- Conducted a manual evaluation to assess faithfulness in each corpus following the methodology proposed in Devaraj et al. (2022)
- Our corpus had the highest percentage of scores marked as 0 (indicating highest faithfulness) and the lowest percentage rated 2 (indicating lowest faithfulness) in all operations.
- Our simplified sentences are generally more readable for nonnative readers.
- Some commented that the original text was already easy to understand.

Generation with Our Corpus

- Conducted experiments to assess our corpus' impact on system generation.
 - Split our data into train:valid:test = 8:1:1.
 - Fine-tuned the BART model using train data, and also employed training data in applying the 3-shot to GPT-3.5.
- Our corpus aids in fine-tuning BART and in providing Few-shot examples to GPT- 3.5, enabling both models to generate faithful simple sentences.

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

Conclusion

- Constructed a Japanese simplification corpus that is faithful to the original content.
- Verified faithfulness and readability through human evaluations by both native and non-native speakers.
- Demonstrated that our corpus improves faithful sentence simplification.