Controlled and Balanced Dataset for Japanese Lexical Simplification

Tomonori Kodaira **Tomoyuki Kajiwara** **Mamoru Komachi**
Tokyo Metropolitan University
Hino City, Tokyo, Japan
{kodaira-tomonori, kajiwara-tomoyuki}@ed.tmu.ac.jp, komachi@tmu.ac.jp

Abstract

We propose a new dataset for evaluating a Japanese lexical simplification method. Previous datasets have several deficiencies. All of them substitute only a single target word, and some of them extract sentences only from newswire corpus. In addition, most of these datasets do not allow ties and integrate simplification ranking from all the annotators without considering the quality. In contrast, our dataset has the following advantages: (1) it is the first controlled and balanced dataset for Japanese lexical simplification with high correlation with human judgment and (2) the consistency of the simplification ranking is improved by allowing candidates to have ties and by considering the reliability of annotators.

1 Introduction

Lexical simplification is the task to find and substitute a complex word or phrase in a sentence with its simpler synonymous expression. We define complex word as a word that has lexical and subjective difficulty in a sentence. It can help in reading comprehension for children and language learners (De Belder and Moens, 2010). This task is a rather easier task which prepare a pair of complex and simple representations than a challenging task which changes the substitute pair in a given context (Specia et al., 2012; Kajiwara and Yamamoto, 2015). Construction of a benchmark dataset is important to ensure the reliability and reproducibility of evaluation. However, few resources are available for the automatic evaluation of lexical simplification. Specia et al. (2012) and De Belder and Moens (2010) created benchmark datasets for evaluating English lexical simplification. In addition, Horn et al. (2014) extracted simplification candidates and constructed an evaluation dataset using English Wikipedia and Simple English Wikipedia. In contrast, such a parallel corpus does not exist in Japanese. Kajiwara and Yamamoto (2015) constructed an evaluation dataset for Japanese lexical simplification[1] in languages other than English.

However, there are four drawbacks in the dataset of Kajiwara and Yamamoto (2015): (1) they extracted sentences only from a newswire corpus; (2) they substituted only a single target word; (3) they did not allow ties; and (4) they did not integrate simplification ranking considering the quality.

Hence, we propose a new dataset addressing the problems in the dataset of Kajiwara and Yamamoto (2015). The main contributions of our study are as follows:

- It is the first controlled and balanced dataset for Japanese lexical simplification. We extract sentences from a balanced corpus and control sentences to have only one complex word. Experimental results show that our dataset is more suitable than previous datasets for evaluating systems with respect to correlation with human judgment.
- The consistency of simplification ranking is greatly improved by allowing candidates to have ties and by considering the reliability of annotators.

Our dataset is available at GitHub[2].

2 Related work

The evaluation dataset for the English Lexical Simplification task (Specia et al., 2012) was an-

[1] http://www.jnlp.org/SNOW/E4
[2] https://github.com/KodairaTomonori/EvaluationDataset

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop, pages 1–7,
Berlin, Germany, August 7-12, 2016.

sentence	「技を出し合い、気分が**高揚する**のがたまらない」とはいえ、技量で相手を上回りたい気持ちも強い。 Although using their techniques makes you feel **exalted**, I strongly feel I want to outrank my competitors in terms of skill.						
paraphrase list	盛り上がる come alive	高まる 高ぶる raised, excited	上がる up	高揚する exalted	興奮する excited	熱を帯びる heated	活性化する revitalized

Figure 1: A part of the dataset of Kajiwara and Yamamoto (2015).

notated on top of the evaluation dataset for English lexical substitution (McCarthy and Navigli, 2007). They asked university students to rerank substitutes according to simplification ranking. Sentences in their dataset do not always contain complex words, and it is not appropriate to evaluate simplification systems if a test sentence does not include any complex words.

In addition, De Belder and Moens (2012) built an evaluation dataset for English lexical simplification based on that developed by McCarthy and Navigli (2007). They used Amazon's Mechanical Turk to rank substitutes and employed the reliability of annotators to remove outlier annotators and/or downweight unreliable annotators. The reliability was calculated on penalty based agreement (McCarthy and Navigli, 2007) and Fleiss' Kappa. Unlike the dataset of Specia et al. (2012), sentences in their dataset contain at least one complex word, but they might contain more than one complex word. Again, it is not adequate for the automatic evaluation of lexical simplification because the human ranking of the resulting simplification might be affected by the context containing complex words. Furthermore, De Belder and Moens' (2012) dataset is too small to be used for achieving a reliable evaluation of lexical simplification systems.

3 Problems in previous datasets for Japanese lexical simplification

Kajiwara and Yamamoto (2015) followed Specia et al. (2012) to construct an evaluation dataset for Japanese lexical simplification. Namely, they split the data creation process into two steps: substitute extraction and simplification ranking.

During the substitute extraction task, they collected substitutes of each target word in 10 different contexts. These contexts were randomly selected from a newswire corpus. The target word was a content word (noun, verb, adjective, or adverb), and was neither a simple word nor part of any compound words. They gathered substitutes from five annotators using crowdsourcing. These procedures were the same as for De Belder and Moens (2012).

During the simplification ranking task, annotators were asked to reorder the target word and its substitutes in a single order without allowing ties. They used crowdsourcing to find five annotators different from those who performed the substitute extraction task. Simplification ranking was integrated on the basis of the average of the simplification ranking from each annotator to generate a gold-standard ranking that might include ties.

During the substitute extraction task, agreement among the annotators was 0.664, whereas during the simplification ranking task, Spearman's rank correlation coefficient score was 0.332. Spearman's score of this work was lower than that of Specia et al. (2012) by 0.064. Thus, there was a big blur between annotators, and the simplification ranking collected using crowdsourcing tended to have a lower quality.

Figure 1 shows a part of the dataset of Kajiwara and Yamamoto (2015). Our discussion in this paper is based on this example.

Domain of the dataset is limited. Because Kajiwara and Yamamoto (2015) extracted sentences from a newswire corpus, their dataset has a poor variety of expression. English lexical simplification datasets (Specia et al., 2012; De Belder and Moens, 2012) do not have this problem because both of them use a balanced corpus of English (Sharoff, 2006).

Complex words might exist in context. In Figure 1, even when a target word such as "高揚する (feel exalted)" is simplified, another complex word "技量 (skill)" is left in a sentence. Lexical simplification is a task of simplifying complex words in a sentence. Previous datasets may include multiple complex words in a sentence but target only one complex word. Not only the target word but also other complex words should be considered as well, but annotation of substitutes and simplification ranking to all complex words in a sentence produces a huge number of patterns, therefore takes a very high cost of annotation. For example, when three complex words

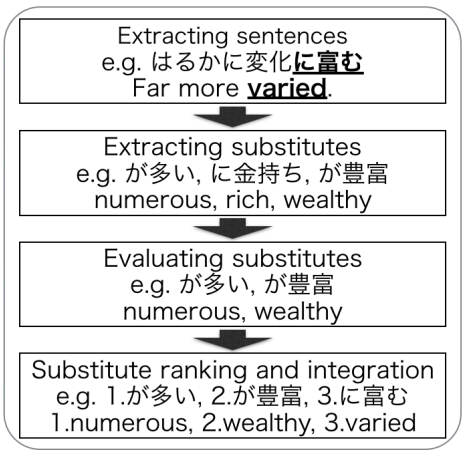

Figure 2: Process of constructing the dataset.

which have 10 substitutes each in a sentence, annotators should consider 10^3 patterns. Thus, it is desired that a sentence includes only simple words after the target word is substituted. Therefore, in this work, we extract sentences containing only one complex word.

Ties are not permitted in simplification ranking. When each annotator assigns a simplification ranking to a substitution list, a tie cannot be assigned in previous datasets (Specia et al., 2012; Kajiwara and Yamamoto, 2015). This deteriorates ranking consistency if some substitutes have a similar simplicity. De Belder and Moens (2012) allow ties in simplification ranking and report considerably higher agreement among annotators than Specia et al. (2012).

The method of ranking integration is naïve. Kajiwara and Yamamoto (2015) and Specia et al. (2012) use an average score to integrate rankings, but it might be biased by outliers. De Belder and Moens (2012) report a slight increase in agreement by greedily removing annotators to maximize the agreement score.

4 Balanced dataset for evaluation of Japanese lexical simplification

We create a balanced dataset for the evaluation of Japanese lexical simplification. Figure 2 illustrates how we constructed the dataset. It follows the data creation procedure of Kajiwara and Yamamoto's (2015) dataset with improvements to resolve the problems described in Section 3.

We use a crowdsourcing application, Lancers,[3]

[3]http://www.lancers.jp/

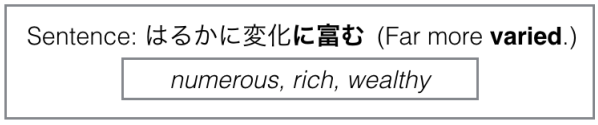

Figure 3: Example of annotation of extracting substitutes. Annotators are provided with substitutes that preserve the meaning of target word which is shown bold in the sentence. In addition, annotators can write a substitute including particles.

to perform substitute extraction, substitute evaluation, and substitute ranking. In each task, we requested the annotators to complete at least 95% of their previous assignments correctly. They were native Japanese speakers.

4.1 Extracting sentences

Our work defines complex words as "High Level" words in the Lexicon for Japanese Language Education (Sunakawa et al., 2012).[4] The word level is calculated by five teachers of Japanese, based on their experience and intuition. There were 7,940 high-level words out of 17,921 words in the lexicon. In addition, target words of this work comprised content words (nouns, verbs, adjectives, adverbs, adjectival nouns, *sahen* nouns,[5] and *sahen* verbs[6]).

Sentences that include a complex word were randomly extracted from the Balanced Corpus of Contemporary Written Japanese (Maekawa et al., 2010). Sentences shorter than seven words or longer than 35 words were excluded. We excluded target words that appeared as a part of compound words. Following previous work, 10 contexts of occurrence were collected for each complex word. We assigned 30 complex words for each part of speech. The total number of sentences was 2,100 (30 words $\times$ 10 sentences $\times$ 7 parts of speech). We used a crowdsourcing application to annotate 1,800 sentences, and we asked university students majoring in computer science to annotate 300 sentences to investigate the quality of crowdsourcing.

4.2 Extracting substitutes

Simplification candidates were collected using crowdsourcing techniques. For each complex word, five annotators wrote substitutes that did not

[4]http://jhlee.sakura.ne.jp/JEV.html

[5]Sahen noun is a kind of noun that can form a verb by adding a generic verb "suru (do)" to the noun. (e.g. "修理 repair")

[6]Sahen verb is a sahen noun that accompanies with "suru". (e.g. "修理する (do repair)")

Dataset	balanced	lang	sents.	noun (%)	verb (%)	adj. (%)	adv. (%)	outlier
De Belder and Moens (2012)	yes	En	430	100 (23.3)	60 (14.0)	160 (37.2)	110 (25.6)	excluded
Specia et al. (2012)	yes	En	2,010	580 (28.9)	520 (25.9)	560 (27.9)	350 (17.6)	included
Kajiwara and Yamamoto (2015)	no	Ja	2,330	630 (27.0)	720 (30.9)	500 (21.5)	480 (20.6)	included
This work	yes	Ja	2,010	570 (28.3)	570 (28.3)	580 (28.8)	290 (14.4)	excluded

Table 1: Comparison of the datasets. In this work, nouns include *sahen* nouns, verbs include *sahen* verbs, and adjectives include adjectival nouns.

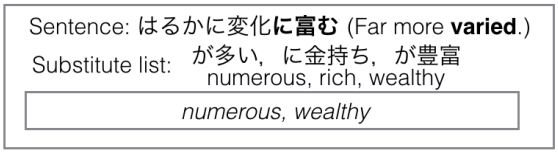

Figure 4: Example of annotation of evaluating substitutes. Annotators choose substitutes that fit into the sentence from substitutes list.

Figure 5: Example of annotation of ranking substitutes. Annotators write rank in blank. Additionally, they are allowed to write a tie.

change the sense of the sentence. Substitutions could include particles in context. Conjugation was allowed to cover variations of both verbs and adjectives. Figure 3 shows an example of annotation.

To improve the quality of the lexical substitution, inappropriate substitutes were deleted for later use, as described in the next subsection.

4.3 Evaluating substitutes

Five annotators selected an appropriate word to include as a substitution that did not change the sense of the sentence. Substitutes that won a majority were defined as correct. Figure 4 shows an example of annotation.

Nine complex words that were evaluated as not having substitutes were excluded at this point. As a result, 2,010 sentences were annotated, as described in next subsection.

4.4 Ranking substitutes

Five annotators arranged substitutes and complex words according to the simplification ranking. Annotators were permitted to assign a tie, but they could select up to four items to be in a tie because we intended to prohibit an insincere person from selecting a tie for all items. Figure 5 shows an example of annotation.

4.5 Integrating simplification ranking

Annotators' rankings were integrated into one ranking, using a maximum likelihood estimation (Matsui et al., 2014) to penalize deceptive annotators as was done by De Belder and Moens (2012). This method estimates the reliability of annotators in addition to determining the true order of rankings. We applied the reliability score to exclude extraordinary annotators.

5 Result

Table 1 shows the characteristics of our dataset. It is about the same size as previous work (Specia et al., 2012; Kajiwara and Yamamoto, 2015). Our dataset has two advantages: (1) improved correlation with human judgment by making a controlled and balanced dataset, and (2) enhanced consistency by allowing ties in ranking and removing outlier annotators. In the following subsections, we evaluate our dataset in detail.

5.1 Intrinsic evaluation

To evaluate the quality of the ranking integration, the Spearman rank correlation coefficient was calculated. The baseline integration ranking used an average score (Kajiwara and Yamamoto, 2015). Our proposed method excludes outlier annotators by using a reliability score calculated using the method developed by Matsui et al. (2014).

$$\frac{1}{|P|}\sum_{p_1,p_2\in P}\frac{p_1 \cap p_2}{p_1 \cup p_2} \quad (1)$$

Pairwise agreement is calculated between each pair of sets ($p_1, p_2 \in P$) from all the possible pairings (P) (Equation 1). The agreement among annotators from the substitute evaluation phase was 0.669, and agreement among the students is 0.673, which is similar to the level found in crowdsourcing. This score is almost the same as that from Kajiwara and Yamamoto (2015). On the contrary,

sentence	最も安上りにサーファーを**装う**方法は，ガラムというインドネシア産のタバコを，これ見よがしに吸うことです． The most simplest method that is **imitating** safer is pretentiously smoke that Garam which is Indonesian cigarette.						
paraphrase list	1. のふりをする professing	2. に見せかける counterfeiting	3. の真似をする, の振りをする playing, professing	4. を真似る playing	5. に成りすます pretending	6. を装う imitating	7. を偽る falsifying

Figure 6: A part of our dataset.

genre	PB	PM	PN	LB	OW	OT	OP	OB	OC	OY	OV	OL	OM	all
sentence	0	64	628	6	161	90	170	700	1	0	6	9	175	2,010
average of substitutes	0	4.12	4.36	5.17	4.41	4.22	3.9	4.28	4	0	5.5	4.11	4.45	4.3

Table 3: Detail of sentences and substitutes in our dataset. (BCCWJ comprise three main subcorpora: publication (P), library (L), special-purpose (O). PB = book, PM = magazine, PN = newswire, LB = book, OW = white paper, OT = textbook, OP =PR paper, OB = bestselling books, OC = Yahoo! Answers, OY = Yahoo! Blogs, OL = Law, OM = Magazine)

	baseline	outlier removal
Average	0.541	**0.580**

Table 2: Correlation of ranking integration.

the Spearman rank correlation coefficient of the substitute ranking phase was 0.522. This score is higher than that from Kajiwara and Yamamoto (2015) by 0.190. This clearly shows the importance of allowing ties during the substitute ranking task.

Table 2 shows the results of the ranking integration. Our method achieved better accuracy in ranking integration than previous methods (Specia et al., 2012; Kajiwara and Yamamoto, 2015) and is similar to the results from De Belder and Moens (2012). This shows that the reliability score can be used for improving the quality.

Table 3 shows the number of sentences and average substitutes in each genre. In our dataset, the number of acquired substitutes is 8,636 words and the average number of substitutes is 4.30 words per sentence.

Figure 6 illustrates a part of our dataset. Substitutes that include particles are found in 75 context (3.7%). It is shown that if particles are not permitted in substitutes, we obtain only two substitutes (4 and 7). By permitting substitutes to include particles, we are able to obtain 7 substitutes.

In ranking substitutes, Spearman rank correlation coefficient is 0.729, which is substantially higher than crowdsourcing's score. Thus, it is necessary to consider annotation method.

5.2 Extrinsic evaluation

In this section, we evaluate our dataset using five simple lexical simplification methods. We calculate 1-best accuracy in our dataset and the dataset of Kajiwara and Yamamoto (2015). Annotated data is collected by our and Kajiwara and Yamamoto (2015)'s work in ranking substitutes task, and which size is 21,700 ((2010 + 2330) × 5) rankings. Then, we calculate correlation between the accuracies of annotated data and either those of Kajiwara and Yamamoto (2015) or those of our dataset.

	This work	K & Y	annotated
Frequency	**41.6**	35.8	41.0
# of Users	**32.9**	25.0	31.5
Familiarity	30.4	**31.5**	32.5
JEV	**38.2**	35.7	38.7
JLPT	**42.0**	40.9	43.3
Pearson	**0.963**	0.930	N/A

Table 4: Accuracy and correlation of the datasets.

5.2.1 Lexical simplification systems

We used several metrics for these experiments:

Frequency Because it is said that a high frequent word is simple, most frequent word is selected as a simplification candidate from substitutes using uni-gram frequency of Japanese Web N-gram (Kudo and Kazawa, 2007). This uni-gram frequency is counted from two billion sentences in Japanese Web text.

Number of Users Aramaki et al. (2013) claimed that a word used by many people is simple, so we pick the word used by the most of users. Number of Users were estimated from the Twitter corpus created by Aramaki et al. (2013). The corpus contains 250 million tweets from 100,000 users.

Familiarity Assuming that a word which is known by many people is simple, replace a target word with substitutes according to the familiarity score using familiarity data constructed by Amano and Kondo (2000). The familiarity score is an averaged score 28 annotators with seven grades.

JEV We hypothesized a word which is low difficulty for non-native speakers is simple, so we select a word using a Japanese learner dictionary made by Sunakawa et al. (2012). The word in dictionary has a difficulty score averaged by 5 Japanese teachers with their subjective annotation according to six grade system.

JLPT Same as above, but uses a different source called Japanese Language Proficient Test (JLPT). We choose the lowest level word using levels of JLPT. These levels are a scale of one to five.

5.2.2 Evaluation

We ranked substitutes according to the metrics, and calculated the 1-best accuracy for each target word. Finally, to compare two datasets, we used the Pearson product-moment correlation coefficient between our dataset and the dataset of Kajiwara and Yamamoto (2015) against the annotated data.

Table 4 shows the result of this experiment. The Pearson coefficient shows that our dataset correlates with human annotation better than the dataset of Kajiwara and Yamamoto (2015), possibly because we controlled each sentence to include only one complex word. Because our dataset is balanced, the accuracy of Web corpus-based metrics (Frequency and Number of Users) closer than the dataset of Kajiwara and Yamamoto (2015).

6 Conclusion

We have presented a new controlled and balanced dataset for the evaluation of Japanese lexical simplification. Experimental results show that (1) our dataset is more consistent than the previous datasets and (2) lexical simplification methods using our dataset correlate with human annotation better than the previous datasets. Future work includes increasing the number of sentences, so as to leverage the dataset for machine learning-based simplification methods.

References

Shigeaki Amano and Kimihisa Kondo. 2000. On the NTT psycholinguistic databases "lexical properties of Japanese". *Journal of the Phonetic Society of Japan 4(2)*, pages 44–50.

Eiji Aramaki, Sachiko Maskawa, Mai Miyabe, Mizuki Morita, and Sachi Yasuda. 2013. Word in a dictionary is used by numerous users. In *Proceeding of International Joint Conference on Natural Language Processing*, pages 874–877.

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

Jan De Belder and Marie-Francine Moens. 2012. A dataset for the evaluation of lexical simplification. In *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 426–437.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–463.

Tomoyuki Kajiwara and Kazuhide Yamamoto. 2015. Evaluation dataset and system for Japanese lexical simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 35–40.

Taku Kudo and Hideto Kazawa. 2007. Japanese Web N-gram Version 1. *Linguistic Data Consoritium*.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 1483–1486.

Toshiko Matsui, Yukino Baba, Toshihiro Kamishima, and Hisashi Kashima. 2014. Crowdordering. In *Proceedings of the 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 336–347.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53.

Serge Sharoff. 2006. Open-source corpora: Using the net to fish for linguistic data. *Journal of Corpus Linguistics, 11(4)*, pages 435–462.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. SemEval-2012 task 1: English lexical simplification. In *Proceedings of the 6th International Workshop on Semantic Evaluation*, pages 347–355.

Yuriko Sunakawa, Jae-ho Lee, and Mari Takahara. 2012. The construction of a database to support the compilation of Japanese learners dictionaries. *Journal of the Acta Linguistica Asiatica 2(2)*, pages 97–115.