Evaluation Dataset and System for Japanese Lexical Simplification

Tomoyuki Kajiwara
Department of Electrical Engineering
Nagaoka University of Technology
Nagaoka City, Niigata, Japan
kajiwara@jnlp.org

Kazuhide Yamamoto
Department of Electrical Engineering
Nagaoka University of Technology
Nagaoka City, Niigata, Japan
yamamoto@jnlp.org

Abstract

We have constructed two research resources of Japanese lexical simplification. One is a simplification system that supports reading comprehension of a wide range of readers, including children and language learners. The other is a dataset for evaluation that enables open discussions with other systems. Both the system and the dataset are made available providing the first such resources for the Japanese language.

1 Introduction

Lexical simplification is a technique that substitutes a complex word or phrase in a sentence with a simpler synonym. This technique supports the reading comprehension of a wide range of readers, including children (Belder and Moens, 2010; Kajiwara et al., 2013) and language learners (Eom et al., 2012; Moku et al., 2012).

The recent years have seen a great activity in this field of inquiry, especially for English: At the SemEval-2012 workshop, many systems were participating in the English lexical simplification task (Specia et al., 2012), for which also an evaluation dataset was constructed. Other resources for statistical learning of simplified rules were built, drawing on the Simple English Wikipedia (Zhu et al., 2010; Horn et al., 2014), e.g. several parallel corpora aligning standard and simple English (Zhu et al., 2010; Kauchak, 2013)[1,2] and evaluation datasets (Specia et al., 2012; Belder and Moens, 2012)[3,4].

On the other hand, there have been no published resources on Japanese lexical simplification so far. Such resources had to be created and made public, for the sake of readers in need of reading assistance, as well as to accelerate the research on this topic. Therefore, we have constructed and published a Japanese lexical simplification system (SNOW S3) and a dataset for evaluation of the system (SNOW E4). These resources are available at the following URL:

http://www.jnlp.org/SNOW

2 Previous Work

Two datasets for evaluation of English lexical simplification have been published. Both were constructed by transforming a lexical substitution dataset, which was constructed in an English lexical substitution task of SemEval-2007 workshop (McCarthy and Navigli, 2007).

2.1 McCarthy Substitution Dataset

The English lexical substitution task of SemEval-2007 requires that the system finds words or phrases that one can substitute for the given target word in the given content. These target words are content words, and their details are shown in Table 1. These contexts are selected from the English Internet Corpus, which is a balanced and web-based corpus of English (Sharoff, 2006). This dataset consists of 2,010 sentences, 201 target words each with 10 sentences as contexts. Five annotators who are native English speakers proposed at most three appropriate substitutions for each of the target words within their contexts. When an appropriate paraphrasable word did not occur, the annotator propose paraphrasable phrases.

An example from this dataset is provided below. As a paraphrase of the adjective "bright" in this context, three annotators proposed "intelligent", another three annotators proposed "clever", and one annotator proposed "smart".

Context: During the siege, G. Robertson had ap-

[1] http://www.cs.pomona.edu/~dkauchak/simplification/
[2] https://www.ukp.tu-darmstadt.de/data/
[3] http://www.cs.york.ac.uk/semeval-2012/task1/
[4] http://people.cs.kuleuven.be/~jan.debelder/lseval.zip

Proceedings of the ACL-IJCNLP 2015 Student Research Workshop, pages 35–40,
Beijing, China, July 28, 2015.

Dataset	Sentence	Noun(%)	Verb(%)	Adjective(%)	Adverb(%)
McCarthy / Specia	2,010	580 (28.9)	520 (25.9)	560 (27.9)	350 (17.4)
De Belder	430	100 (23.3)	60 (14.0)	160 (37.2)	110 (25.6)
Ours (SNOW E4)	2,330	630 (27.0)	720 (30.9)	500 (21.5)	480 (20.6)

Table 1: Size of the dataset

pointed Shuja-ul-Mulk, who was a bright boy only 12 years old and the youngest surviving son of Aman-ul-Mulk, as the ruler of Chitral.

Gold-Standard: intelligent 3; clever 3; smart 1;

2.2 Specia Simplification Dataset

The English lexical simplification task of SemEval-2012 requires that the system ranks the target word and its several paraphrases according to how *simple* they are in the context. *Simple* means that the word is easy to understand for many people, including children and non-natives.

This dataset was annotated by fluent but non-native English speakers (college freshmen). The Trial dataset used four annotators, and the Test dataset used five annotators. These annotators ranked target words and their several paraphrases according to how simple they were in contexts from the lexical substitution dataset described in Section 2.1. Next, the ranks received from each annotator were integrated into the dataset. Finally, the gold-standard annotations were generated by averaging the annotations from all annotators.

An example from this dataset is provided below. When the following ranking was obtained from four annotators in a context, the ranks of "clear" were 1, 2, 1, 4, and the average rank was 2. Similarly, the average rank of each word calculated. Thus, the rank of "light" is 3.25, that of "bright" is 2.5, that of "luminous" is 4, and that of "well-lit" is 3.25. The final integrated ranking is obtained by rearranging the average ranks of these words in the ascending order, as shown below.

1: {clear}{light}{bright}{luminous}{well-lit}
2: {well-lit}{clear}{light}{bright}{luminous}
3: {clear}{bright}{light}{luminous}{well-lit}
4: {bright}{well-lit}{luminous}{clear}{light}
Gold: {clear}{bright}{light,well-lit}{luminous}

2.3 De Belder Simplification Dataset

De Belder and Moens (2012) also created a simplification dataset. They deleted enough simple target words included in the Basic English combined word list[5] from the lexical substitution dataset described in the Section 2.1 at first. As a result of deleting, the number of target words narrowed down from 201 to 43. Five annotators ranked these 43 target words and their several paraphrases according to how simple they were in the context.

These annotators were recruited using the Amazon Mechanical Turk[6]. De Belder and Moens requested annotators who were located in the U.S. and had completed at least 95% of their previous assignments correctly.

In the end, the rank from each annotator was integrated into the dataset. In this dataset, the noisy channel model was used in order to take account of the rank and reliability of each annotator.

3 Constructing Japanese Lexical Substitution Dataset

We have constructed a dataset for evaluation of Japanese lexical simplification. First, a Japanese lexical substitution dataset was constructed using the same method as McCarthy and Navigli (2007).

3.1 Selecting Target Words

We define target words as the list of content words (nouns, verbs, adjectives, and adverbs) that are common to two Japanese word dictionaries (IPADIC-2.7.0[7] and JUMANDIC-7.0[8]) in order to select the standard target words at first. Next, the following words were deleted from these words.

- Words that are already simple enough
- Words that have no substitutions
- Words that are a part of a compound word
- Words that are a part of an idiomatic phrase
- Low frequency words

We define simple words as words in Basic Vocabulary to Learn (Kai and Matsukawa, 2002), which is a receptive vocabulary for elementary school students. Words that are not registered

[5] http://simple.wikipedia.org/wiki/Wikipedia:Basic_English_combined_wordlist
[6] https://www.mturk.com
[7] http://sourceforge.jp/projects/ipadic/releases/24435/
[8] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

$$\frac{\sum_{p_1,p_2 \in P} \frac{p_1 \cap p_2}{p_1 \cup p_2}}{|P|} \quad (1) \qquad \frac{\sum_j (rank_i(w_j) - \overline{rank_i})(rank_{ave}(w_j) - \overline{rank_{ave}})}{\sqrt{\sum_j (rank_i(w_j) - \overline{rank_i})^2 \sum_j (rank_{ave}(w_j) - \overline{rank_{ave}})^2}} \quad (2)$$

in SNOW D2 (Yamamoto and Yoshikura, 2013) are defined as words that have no substitutions. Low frequency words are defined as words that occurred less than 10 times over the 15 years in Japanese newspapers[9].

In the end, 250 words (nouns and verbs 75 each, adjectives and adverbs 50 each) were chosen as a target words at random.

3.2 Providing Paraphrases

An annotator provided several paraphrases for each target word in 10 contexts. These contexts were randomly selected from newspaper article. When providing a paraphrase, an annotator could refer to a dictionary but was not supposed to ask the other annotators for an opinion. When an annotator could not think of a paraphrase, they were permitted to supply no entry.

Five annotators for every fifty sentences were recruited using crowdsourcing service[10]. On average, each of these annotators contributed 5.38 paraphrases.

3.3 Merging All Annotations

Each annotator's result was evaluated, and all the results were merged into one dataset. Five new annotators for every fifty sentences were recruited through the crowdsourcing service. We adopted the paraphrases that more than three annotators rated appropriate by answering the question, "Is this paraphrase appropriate?" in the affirmative. When an annotator rated a paraphrase as inappropriate, they were shown the following two criteria.

1. A paraphrase is inappropriate if the sentence becomes unnatural as a result of the substitution of this paraphrase for the target word.
2. A paraphrase is inappropriate if the meaning of the sentence changes as a result of the substitution of this paraphrase for the target word.

An average of 4.50 lexical paraphrases were accepted. However, 170 sentences (17 target words) that all paraphrases have been evaluated to be inappropriate were discarded.

[9]http://www.nikkeibookvideo.com/kijidb/

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

Since we have sets of paraphrases for each target word and annotator, pairwise agreement was calculated between each pair of sets ($p_1, p_2 \in P$) from each possible pairing (P) according to the Equation (1), following previous research (McCarthy and Navigli, 2007). Inter-annotator agreement is 66.4%.

An English translation of an example from the dataset is provided below. As a paraphrase of the noun "appeal" in this context, one annotator proposed "advocate", another annotator proposed "exert", and three annotators proposed "promote".

Context: You can appeal for proud batting power.
Gold-Standard: advocate 1; promote 3; exert 1;

4 Transforming into Lexical Simplification Dataset

4.1 Ranking Paraphrases

These target words and their several paraphrases were ranked according to how simple they were in the context from the dataset that we built (as discussed in Section 3) in order to transform it into a dataset for evaluation of lexical simplification. The same annotators as those mentioned in section 3.3 worked on this task.

Finally, the total number of annotators is 500. Some 250 annotators provided paraphrases, others evaluated and ranked these paraphrases.

Inter-annotator agreement was calculated by Spearman's rank correlation coefficient, following previous research (Belder and Moens, 2012). Spearman's rank correlation coefficient is defined as in the Equation (2), where $\overline{rank_i}$ is the average rank of the words given by annotator i. To extend this equation to one annotator versus other annotators, we define the rank assigned by the $rank_{ave}$ to be the average of the ranks given by the other annotators. This agreement is 33.2%[11].

4.2 Merging All Rankings

All annotators' work results were merged into one dataset. The rank of each word was decided based

[11]While this score is apparently low, the highly subjective nature of the annotation task must be taken into account (Specia et al., 2012).

	all	%	noun	%	verb	%	adj	%	adv	%
1. # context pairs	10,485	-	2,835	-	3,240	-	2,250	-	2,160	-
2. # 1. with same list	1,593	15	789	28	348	11	168	7	288	13
3. # 2. with different rankings	948	60	344	44	262	75	129	77	213	74
4. # 3. with different top word	463	49	214	62	130	50	51	40	68	32

Table 2: Context dependency ratio

on the average of the rank from each annotator, following the previous research (Specia et al., 2012). The same rank is assigned to words that have the same average. In this study, the same annotator performed both the evaluation of paraphrases and their ranking. Therefore, any word that an annotator judged as an inappropriate paraphrase was not ranked. The minimum rank is assigned to these words that were not ranked at the time of the calculation of the average rank.

An English translation of an example from the dataset is provided below. When the following ranking was obtained from five annotators in a context, the ranks of "appeal" were 1, 2, 4, 2, 2, and the average rank was 2.2. Similarly, the average rank of "promote" is 2.2, that of "advocate" is 2.6, and that of "exert" is 3. The final integrated ranking is obtained by rearranging the average ranks of these words in the ascending order.

1: {appeal}{promote}{advocate}{exert}
2: {advocate}{appeal}{promote}{exert}
3: {promote}{exert}{advocate} #appeal
4: {exert}{appeal}{advocate}{promote}
5: {promote}{appeal}{advocate} #exert
Gold: {appeal, promote}{advocate}{exert}

4.3 Properties of the dataset

In 1,616 (69.4%) of the sentences, a target word can be replaced by one or more simpler words. In 420 (18.0%) of the cases, there is also one or more words that are equally complex. In 1,945 (83.5%) of the cases, there are words that are more complex. The average number of substitutions is 5.50. The average number of levels of difficulty is 4.94.

Table 2 shows how the relative simplicity of the target words and their paraphrases is context dependent. Only 15.2% of all context-pairs which share the target word have the same list of paraphrases. This shows that the meaning of many target words changed slightly in different contexts. In addition, 59.5% of combinations with the same list of paraphrases have different ranks of difficulty. This shows that the difficulty of a word also changes slightly in different contexts. Among these, 48.8% is even different in the simplest word.

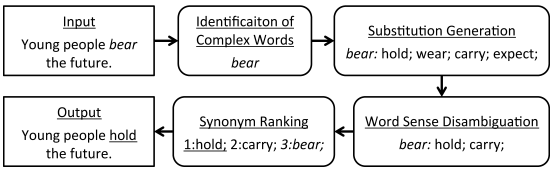

Figure 1: Outline of lexical simplification system

5 Constructing Japanese Lexical Simplification System

We have also constructed a lexical simplification system using four typical mechanisms of lexical simplification (Shardlow, 2014) shown in Figure1. We expect the standard system to be used as a baseline of Japanese lexical simplification. We also expect that the system can support the reading comprehension of a wide range of readers.

5.1 Identification of Complex Words

An input sentence is first analyzed by the Japanese morphological analyzers MeCab-0.993 (Kudo et al., 2004)[12] and IPADIC-2.7.0, and content words that are not included in the list of simple words are extracted as complex words. These complex words are not part of a compound word or an idiomatic phrase.

In this study, simple words are defined as the Basic Vocabulary to Learn; compound words are defined as the lists of entries from Japanese Wikipedia[13] and the Compound Verb Lexicon[14]; finally, idiomatic phrases are defined as the list of Japanese idiomatic phrases made by Sato (2007).

5.2 Substitution Generation

Several paraphrases are enumerated as candidates of a simple word for each complex word. These lexical paraphrases were selected from several Japanese lexical paraphrasing databases such as SNOW D2 (Yamamoto and Yoshikura, 2013),

[12]https://code.google.com/p/mecab/
[13]http://dumps.wikimedia.org/jawiki/
[14]http://vvlexicon.ninjal.ac.jp/

Precision	Recall	F-measure
0.89	0.08	0.15

Table 3: Performance of the system

Noun	Verb	Adjective	Adverb
62	65	3	0

Table 4: POS of the simplified target words

Japanese WordNet Synonyms Database[15], Verb Entailment Database[16], and Case Base for Basic Semantic Relations[16], following previous research (Kajiwara and Yamamoto, 2014).

5.3 Word Sense Disambiguation

If, given the context of the sentence, the list of suggested paraphrases for a complex word contains words that are improper in this context, these improper paraphrases are removed from the list. An input sentence receives a predicate-argument structure analysis using the Japanese predicate-argument structure analyzer SynCha-0.3 (Iida and Poesio, 2011)[17], and the predicate (verb or adjective), the arguments (nouns) and grammatical relations (case makers such as "*ga* (subject)", "*o* (object)", "*ni* (indirect object)") are extracted as a set of the form {predicate, relation, argument}.

Either the predicate or one of the arguments is identified as a complex word. A list is of candidate substitutions is generated for that word, followed by a list of sets of the form {predicate, relation, argument}, where the candidate substitutions are used instead of the complex word (so there will be as many of these sets as there are candidate substitutions). These new sets are checked against the Kyoto University Case Frame[18]. If the set is found there, the candidate substitution counts as a legitimate substitution; if the set is not found, the candidate substitution is not counted as a legitimate substitution. Kyoto University Case Frame is the list of predicate and argument pairs that have a case relationship, and it is built automatically (Kawahara and Kurohashi, 2006) from Web texts.

5.4 Synonym Ranking

All candidate words are given a degree of difficulty. The simplest word is used to replace the complex word in the input sentence, and the output sentence is generated.

In this study, Lexical Properties of Japanese (Amano and Kondo, 2000) is used for determining the degree of difficulty.

[15]http://nlpwww.nict.go.jp/wn-ja/jpn/downloads.html
[16]https://alaginrc.nict.go.jp/resources/nict-resource/
[17]http://www.cl.cs.titech.ac.jp/ ryu-i/syncha/
[18]http://www.gsk.or.jp/catalog/gsk2008-b/

5.5 Evaluation of the System by the Dataset

The performance of the lexical simplification system that was discussed in this section is estimated using the evaluation dataset that was constructed as discussed in Section 4. The performance of the system is shown in Table 3. In 146 sentences, the system converted a target word into another word; in 130 sentences, that output word was simpler than the target word defined by the evaluation dataset appropriately. In addition, the system converted 652 words in total, but only 146 words of these were the target words.

The details as to the parts of speech of the target words that got simplified appropriately are shown in Table 4. The system simplifies only the predicates and arguments extracted by the predicate-argument structure analysis. However, adverbs are not simplified since they are included in neither predicates nor arguments. In addition, an adjective may become a predicate, but it may also become part of a noun phrase by modifying a noun. The system simplifies only predicate adjectives.

An English translation of an example of several system outputs is provided below.

- It is {distributed –> dealt} to a {caller –> visitor} from foreign countries.
- {Principal –> President} Takagi of the bank presented an idea.

6 Final Remarks

We built a Japanese lexical simplification system and a dataset for evaluation of Japanese lexical simplification. Subsequently, we have published these resources on the Web.

The system can support the reading comprehension of a wide range of readers, including children and language learners. Since we have developed a standard system, we expect the system to be used as a baseline system of lexical simplification.

Furthermore, the dataset enables us to figure out system performance. This solves the problems of cost and reproducibility associated with the conventional manual evaluation and accelerates research on this topic.

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):44–50.

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 (CICLing-2012)*, pages 426–437.

Soojeong Eom, Markus Dickinson, and Rebecca Sachs. 2012. Sense-specific lexical information for reading assistance. *In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 316–325.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using wikipedia. *In Proceedings of the 52th Annual Meeting of the Associatioin for Computational Linguistics*, pages 458–463.

Ryu Iida and Massimo Poesio. 2011. A cross-lingual ilp solution to zero anaphora resolution. *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 804–813.

Mutsuro Kai and Toshihiro Matsukawa. 2002. *Method of Vocabulary Teaching: Vocabulary Table version*. Mitsumura Tosho Publishing Co., Ltd.

Tomoyuki Kajiwara and Kazuhide Yamamoto. 2014. Qualitative evaluation of available japanese resources for lexical paraphrasing. *IEICE Technical Report, NLC2014-37*, 114(366):43–48.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. *In Proceedings of the 25th Conference on Computational Linguistics and Speech Processing*, pages 59–73.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. *In Proceedings of the 51th Annual Meeting of the Associatioin for Computational Linguistics*, pages 1537–1546.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 176–183.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task10: English lexical substitution task. *In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53.

Manami Moku, Kazuhide Yamamoto, and Ai Makabi. 2012. Automatic easy japanese translation for information accessibility of foreigners. *In Proceedings of the Workshop on Speech and Language Processing Tools in Education*, pages 85–90.

Satoshi Sato. 2007. Compilation of a comparative list of basic japanese idioms from five sources. *The Special Interest Group Technical Reports of IPSJ, 2007-NL-178*, pages 1–6.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications, Special Issue on Natural Language Processing*, pages 58–70.

Serge Sharoff. 2006. Open-source corpora: Using the net to fish for linguistic data. *International 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 (SemEval-2012)*, pages 347–355.

Kazuhide Yamamoto and Kotaro Yoshikura. 2013. Manual construction of lexical paraphrase dictionary of japanese verbs, adjectives, and adverbs. *In Proceedings of 19th Annual Meeting of Association for Natural Language Processing*, pages 276–279.

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*, pages 1353–1361.