Selecting Proper Lexical Paraphrase for Children

Tomoyuki Kajiwara Hiroshi Matsumoto Kazuhide Yamamoto
Nagaoka University of Technology
{kajiwara, matsumoto, yamamoto}@jnlp.org

Abstract

We propose a method for acquiring plain lexical paraphrase using a Japanese dictionary in order to achieve lexical simplification for children. The proposed method extracts plain words that are the most similar to the headword from the dictionary definition. The definition statements describe the headword using plain words; therefore, paraphrasing by replacing the headword with the most similar word in the dictionary definition is expected to be an accurate means of lexical simplification. However, it is difficult to determine which word is the most appropriate for the paraphrase. The method proposed in this paper measures the similarity of each word in the definition statements against the headword and selects the one with the closest semantic match for the paraphrase. This method compares favorably with the method that acquires the target word from the end of the definition statements.

Keywords: Lexical Simplification, Lexical Paraphrase.

1. Introduction

In the current information age, a various readers have easy access to diverse text data. To achieve information transmission and gathering effectively, we must address the gap in readers' linguistic skills. The gap of linguistic skills results from differences in age, such as between children and adults, as well as from differences in expert knowledge. In the effort to bridge this gap, and also to facilitate better communication with foreign language speakers[8] and people with disabilities, technology can play an important role.

To investigate how technology can be applied toward bridging the gap in readers' linguistic skills, we simplify the text of newspaper articles containing words that pose difficulties in communication, especially for elementary school students. Children are still developing their language skills, and as such, they have smaller vocabularies than adults. In this paper, we perform text simplification for children by paraphrasing selected newspaper articles using only words found in Basic Vocabulary to Learn (BVL)[3].

BVL is a collection of words selected based on a lexical analysis of elementary school textbooks. It contains 5,404 words that can help children write expressively. We define words not included in BVL as Difficult Words (DWs) and those in BVL paraphrased from DW as Simple Words (SWs).

Paraphrasing newspaper articles using words that children can understand makes a great contribution to reading assistance for young students.

2. Related Works

Although there are some methods [10] proposed for automatically acquiring paraphrasable expressions from Web pages, the quality of the results are still unsatisfactory. Hence typical methods use thesauri or dictionaries. Thesaurus is a language resource that contains semantically classified vocabulary words. Methods that utilize a thesaurus have an advantage in that they can measure the semantic relatedness between words (i.e., the distance between meanings). Japanese dictionaries are another language resource that provides the definition of a given lemma. Methods that utilize a dictionary have an advantage in that they are able to acquire simplified text. The aim of this study is to simplify the text of newspaper article through paraphrasing based on the use of a Japanese dictionary.

Fujita et al. [1] and Mino and Tanaka [9] paraphrased the headword of a noun in a dictionary as the headword of another noun by assessing the similarity of the definitions for the two. Yet, as also reported by Mino and Tanaka, the target words acquired by this method are not simpler than the original words. We paraphrase by taking advantage of Japanese dictionary characteristics, namely that "The definition statements are simpler than the headwords" [9], because our aim is lexical simplification.

Kaji et al. [3] assumed that the definition statement has an inflectable word as a nominative if the headword is inflectable, and the nominative is placed at the end of the definition statement. Then, they proposed a method for paraphrasing inflectable words. Mino and Tanaka assumed that the last segment of the main sentence in the definition statement represents the meaning of the headword, and they proposed a method for paraphrasing nouns. Kajiwara and Yamamoto [4] assumed that the target word is the same part-of-speech as headword and is placed at the end of the definition statement. They proposed a method for paraphrasing both nouns and inflectable words.

These describe the selection of target words from the end of the definition statement in the dictionary. As shown in Figure 1, however, appropriate target words are not always found at the end of definitions. In Figure 1, the dictionary definition of "大詰め (final stage)" is "芝居の最後の場面 (the last scene of the play)." The end of the definition statement is "場面 (scene)." However, the DW "大詰め (final stage)" cannot be paraphrased as "場面 (scene)." In this example, paraphrasing with the SW "最後 (last)" is correct. The original phrase "大詰めの大一番 (big match of the final stage)" is paraphrased as "最後の大一番 (big match of the last)." Therefore, we propose a better method for identifying target words from within a definition statement.

definition：【大詰め】芝居の**最後**の場面
paraphrase：**大詰め**の大一番 → **最後**の大一番

Figure 1: Example of a word that cannot be paraphrased as the end portion of the definition statement

Multiple target word candidates can be acquired by making use of the entire definition statement. Therefore, a process is needed for selecting the most appropriate target words. In the study of the selection of target words, researchers employ various methods such as assessing semantic similarity based on data from a thesaurus [7] or using the statistical information from large resources based on the distributional hypothesis [3][6]. Thesauruses provide hierarchical semantic classifications of words. By measuring the semantic distance between words in the thesaurus, it is possible to measure the proximity of meaning between words. Furthermore, according to the distributional hypothesis [2], words with similar meanings are often used in similar contexts. Based on this hypothesis, Lapata et al. and Keller et al. reported that the plausibility determination of the expression can be achieved by utilizing co-occurrence frequency and n-gram. In this paper, in order to maintain as much of the original meaning as possible in the paraphrase, we select the SW with the highest similarity to the DW (as determined using a thesaurus).

3. Proposed Method

3.1 Acquisition of the Target Word Candidates

As shown in Figure 2, the target word candidates are selected according to the following steps.

1. DWs are extracted from the input (i.e., the original sentence). DWs are content words that do not appear in BVL. A content word is one whose part-of-speech is identified as either a noun, verb, adjective, or adverb. In Figure 2, the DW “教授 (professor)” is included in the original sentence “教授はどうなのだろう (What would the professor have in mind?).”
2. The original DW is located in the Japanese dictionary. Figure 2 shows that Japanese dictionaries give four different definition statements for “教授 (professor)”: “教授という地位の人 (people with the status of professor),” “教授という地位 (status of professor),” “学問や技などを教えること (teaching learning and skill),” and “大学の先生 (university teacher).”
3. The definition statements of headwords are analyzed by the Japanese language morphological analyzer MeCab(5), and words are extracted if they are the same part-of-speech as the headword. In Figure 2, DW “教授 (professor)” is a noun. Therefore, seven nouns are extracted: “教授 (professor),” “地位 (status),” “人 (people),” “学問 (learning),” “わざ (skill),” “大学 (university),” and “先生 (teacher).”

DWs are removed, and only SWs are retained. In Figure 2, “教授 (professor)” and “地位 (status)” are DWs. Therefore, five SWs are obtained as target words: “人 (people),” “学問 (learning),” “わざ (skill),” “大学 (university),” and “先生 (teacher).”

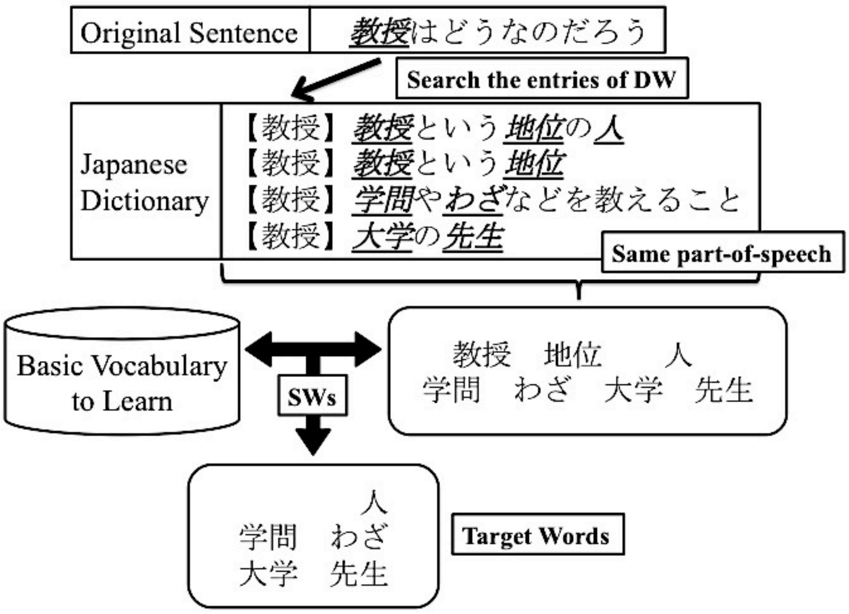

Figure 2: Target word selection by the proposed method

3.2 Selection of the Proper Target Word

In the proposed method, SWs with the highest similarity scores relative to the DW are selected for the purpose of maintaining as much of the original meaning as possible.

Japanese WordNet(1) is used to measure the similarity of meaning between words. WordNet is a language resource that includes a hierarchically described set of synonyms. Using WordNet allows us to measure the similarity of meaning as a distance between words belonging to sets of synonyms. If two or more SWs have the highest similarity score, one is selected at random.

4. Comparative Methods

4.1 Acquisition of the Target Word Candidates

As shown in Figure 3, Kajiwara and Yamamoto’s approach [4] is used as comparative method for selecting target word candidates. In this method, target word candidates are selected according to the following steps.

1. DWs are extracted from the input (i.e., the original sentence). In Figure 3, DW “教授 (professor)” is included in the original sentence “教授はどうなのだろう (What would the professor have in mind?).”

2. The original DW is located in the Japanese dictionary. Figure 3 shows that Japanese dictionaries give four different definition statements for "教授 (professor)": "教授という地位の人 (people with the status of professor)," "教授という地位 (status of professor)," "学問や技などを教えること (teaching learning and skill)," and "大学の先生 (university teacher)."
3. The definition statements of headwords are analyzed by the Japanese language morphological analyzer MeCab, and words are extracted from the end of sentences if they are the same part-of-speech as the headword. In Figure 3, DW "教授 (professor)" is a noun. Therefore, four nouns are extracted: "地位 (status)," "人 (people)," "わざ (skill)," and "先生 (teacher)." Note that "教授 (professor)," "学問 (learning)," and "大学 (university)" are also nouns; however, according to Kajiwara and Yamamoto (2013), target words are limited to words from the end of definition statements.
4. DWs are removed, and only SWs are retained. In Figure 3, "地位 (status)" is a DW. Therefore, three SWs are obtained as target words: "人 (people)," "わざ (skill)," and "先生 (teacher)."

In contrast to the method proposed here, Kajiwara and Yamamoto's method describes the acquisition of only one target word from the end of the definition statement.

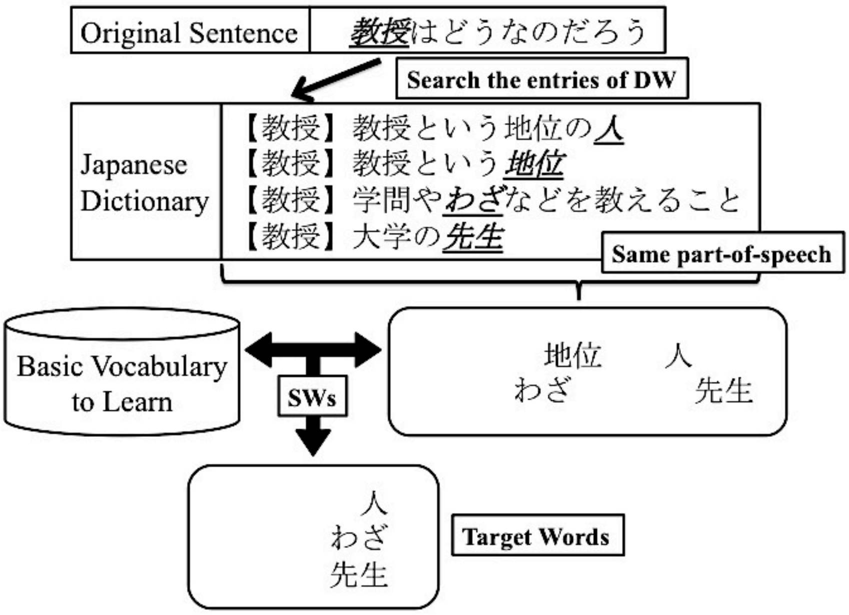

Figure 3: Target word selection by the comparative method

4.2 Selection of the Proper the Target Word

The selection of target words in the proposed method is compared with similar processes in five other methods. In addition, we compare target word selection by weighted voting, which uses a combination of these methods. Ma et al. [7] showed that weighted voting is effective in word sense disambiguation. We apply the method of weighted voting in the selection of target words in this paper, and compare it with the proposed method.

(1) Selection by frequency of the target words

We consider that if the same SW is obtained from many different definition sentences, then it is sufficiently reliable as a target word.

$$score(x) = \sum\nolimits_{X} freq(x) \qquad (1)$$

(2) Selection by co-occurrence frequency

Utilizing the co-occurrence frequencies of content words besides DWs and each SW in the same sentence, we select the most reliable SW as the target word.

$$score(x) = \sum\nolimits_{Z} freq(x, z_n) \qquad (2)$$

(3) Selection by Point-wise Mutual Information

In criterion (2), simply the summation of co-occurrence frequencies is used. For this selection criterion, in addition to the previous criteria, the Point-wise Mutual Information (PMI) criterion, which ignores the effect of single-word frequency, is utilized as well. From the calculation of co-occurrence with PMI shown in equation (3), the co-occurrence frequency can be accurately measured, even for words with high frequencies.

$$score(x) = \sum\nolimits_{Z} log \frac{freq(x, z_n)}{freq(x)freq(z_n)} \qquad (3)$$

(4) Selection by tri-gram frequency

To select SWs from the same context as DWs, tri-gram frequency is obtained. For the sentences with DWs, the frequencies of all tri-grams whose DW is replaced with a SW are obtained by using the three types of tri-grams, two surrounding words, and the DW. Then, as shown in equation (4), the score of SW x is represented using the two words before and after DW y in the source sentence $\{w_{-m}\ldots w_{-2}w_{-1}yw_{+1}w_{+2}\ldots w_{+n}\}$.

$$score(x) = freq(w_{-2}w_{-1}x) + freq(w_{-1}xw_{+1}) + freq(xw_{+1}w_{+2}) \qquad (4)$$

(5) Selection by distributional similarity

To select SWs used in contexts similar to those of DWs, we first create document vectors and then to calculate the similarity of the document vectors of DW and SW. For the similarity calculation between vectors, cosine similarity is applied. In equation (5), the similarity of two document vectors of SW x and DW y is set as the score for SW x.

$$score(x) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} \qquad (5)$$

A) Weightless voting by comparative methods

B) Weighted voting by comparative methods

Weighted voting uses the five comparative methods. Weight is an accuracy of paraphrase in that each method that has been evaluated in advance. The word with the highest score according to each criterion is selected by the five criteria. Finally, the word with the best total score is selected.

C) Weightless voting adds the proposed method to (A)

D) Weighted voting adds the proposed method to (B)

Weighted vote by all methods also adds the proposed method to the five comparative methods.

5. Experiment

5.1 Data

News sentences including one DW in each are paraphrased. DWs are words that appear more than 50 times in the Mainichi News Paper published in 2000[(8)] and are not included in BVL. The selected newspaper includes 232,038 sentences and 26,709 kinds of DWs. In total, the sample comprises 221 DWs appearing more than 50 times. Among them, 165 DWs include one or more of the paraphrasable SWs in the definition statements. After DWs with only paraphrasable candidate are excluded, the experiment data consist of 152 DWs.

We combined multiple Japanese dictionaries to increase the coverage of the paraphrasing. We used the following three dictionaries: *EDR* Japanese word dictionary[(2)], *The Challenge*, an elementary school Japanese dictionary[(7)], and *Sanseido Japanese Dictionary*[(4)].

In the comparative method for selection, co-occurrence frequencies of content words and content word frequency are obtained using the 7-gram from the Web Japanese N-gram[(6)]. Web Japanese N-gram includes the word N(1 to 7)-grams parsed by MeCab. Each N-gram appears more than 20 times in 20 billion sentences in Web text. The acquisition of co-occurrence frequency or creating a document vector uses the longest 7-gram data. Additionally, the word frequency used for calculation of PMI is acquired from 7-gram data to match the co-occurrence frequency. Tri-gram frequency is from the tri-gram.

5.2 Procedure

The target words are acquired by each method with 152 DWs. In selection of proper target word, DWs are split into 52 DWs and 100DWs. First, 52 DWs are used in order to select the proper target word by the proposed method and five comparative methods. Based on the rate of correct answers, a proper target word is selected by weighted voting of 100 DWs.

Three subjects that do not include the author and coauthor are evaluated. When two or more subjects in the three subjects are judged that the SW can be replaced with DW in the original sentence, the SW is the correct answer. Kappa coefficients of the subjects are 0.617, 0.600, and 0.662, respectively.

5.3 Results

Tables 1 and 2 and Figures 4 and 5 show the accuracies of the paraphrases. For the selection of the target word, the proposed method using WordNet similarity is the most efficient. At this point in the analysis, the proposed method has a level of accuracy similar to the comparative methods.

Table 1: Accuracy of each selection for 52 DWs

Method of selection	Method of acquisition	
	Proposed (%)	K&Y2013 (%)
Baseline: Randomness	32.2	41.5
Proposed: WordNet similarity	69.2	65.4
(1) Frequency	40.4	40.4
(2) Co-occurrence	32.7	38.5
(3) Point-wise Mutual Information	30.8	51.9
(4) 3-gram frequency	50.0	53.8
(5) Distributional similarity	40.4	48.1

Table 2: Accuracy of each combinational selection for 100 DWs

Method of selection	Method of acquisition	
	Proposed (%)	K&Y2013 (%)
Baseline: Randomness	30.2	39.2
Proposed: WordNet similarity	60.0	58.0
A) Weightless voting by comparative methods (1)-(5)	45.0	55.0
B) Weighted voting by comparative methods (1)-(5)	44.0	60.0
C) Weightless voting adds the WordNet similarity to (A)	54.0	60.0
D) Weighted voting adds the WordNet similarity to (B)	60.0	62.0

Table 3 shows the percentage of the possible SWs among the acquired target word candidates. Multiple SWs are acquired for each target word, and in some cases, multiple SWs may be the correct answer. The number of included paraphrasable target words in Table 3 is the number of DWs that acquire more than one word that can be the correct answer. The number of correct answers is slightly better than that produced by the proposed method, which selects target words from the entire definition statement.

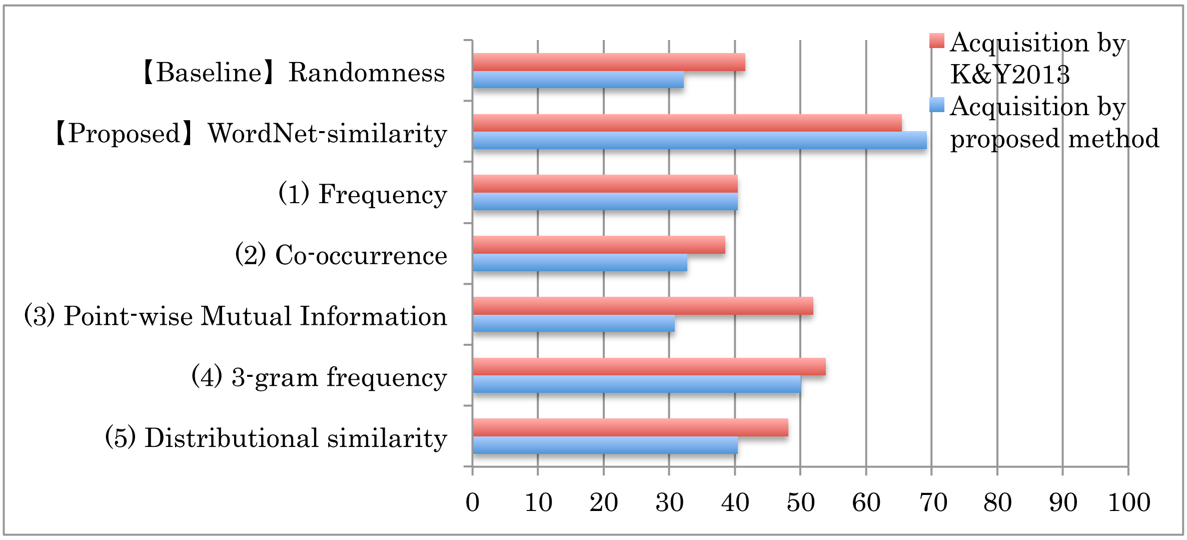

Figure 4: Accuracy of each target word selection for 52 DWs

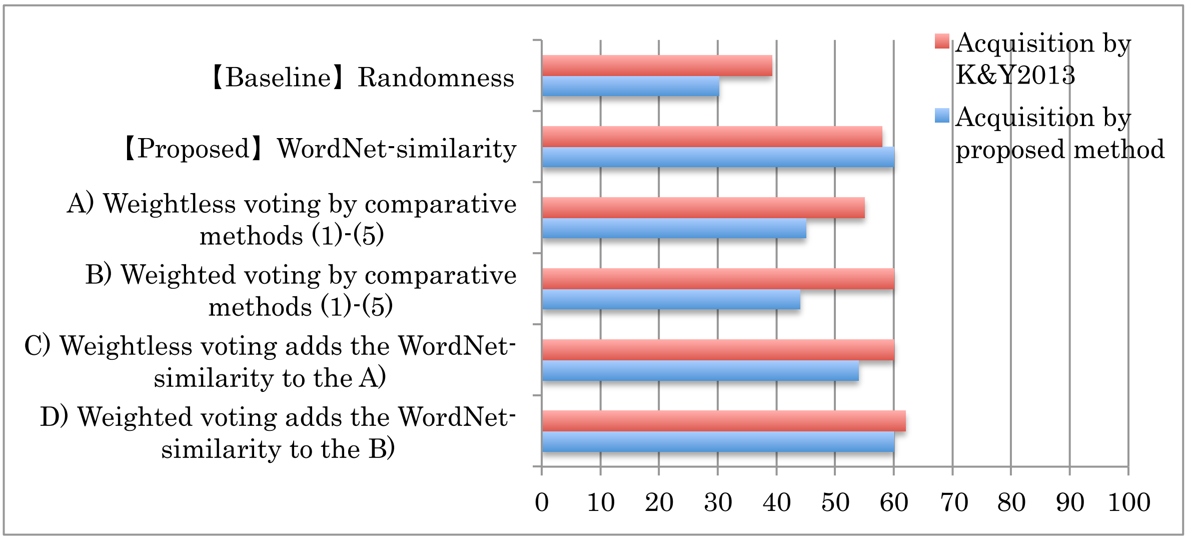

Figure 5: Accuracy of each combinational selection for 100 DWs

Table 3: Number of paraphrasable target words

Acquisition Method	Number of included paraphrasable target words	Percentage of included paraphrasable target word (%)
Proposed	153 / 221	69.2
K&Y2013	143 / 221	64.7

6. Discussion

6.1 Acquisition of the Target Word Candidates

The proposed method is able to acquire more target words than the comparative method, which includes paraphrasable SWs. Assuming that we can reliably select the target words, the proposed method can be expected to improve the accuracy of paraphrasing. This shows the potential as well as effectiveness of the proposed method, which acquires target words from the entire definition statement.

However, the number of target words including paraphrasable words acquired by the proposed method differs by only 3.2 points from the number acquired by the comparative method. This shows that words at the end of definition statements are more effective than those found elsewhere.

The word that can be used to paraphrase the headword represents the central core of the meaning in definition statements. In the Japanese dictionary, central core meanings often appear at the end of the definition.

6.2 Selection of the Proper Target Word

As shown in Table 1, the selection using WordNet similarity was highly accurate, in contrast to the proposed method. As shown in Table 2, the selection accuracies by comparative methods are improved by match-up and vote.

Regarding the acquisition of target word candidates, the accuracy of voting by the five comparative methods is less than the proposed method, which uses WordNet similarity. Moreover, by combining the WordNet similarity method and five comparative methods, the voting method achieves an accuracy rate nearly equal to that of the proposed method, which uses only WordNet similarity.

The comparative methods, which use the weighted voting method without WordNet similarity, have an accuracy rate nearly equal to that of the proposed method, which uses only WordNet similarity. However, when the WordNet similarity method and five comparative methods were combined, no significant changes were observed in the accuracy rate of the weighted voting.

If the combined method obtains a nearly equal accuracy, the proposed method is better than the weighted voting method because of its simplicity. These results show that selecting the target word based on its similarity of meaning with the original word is a better method than selection by frequency or context information.

6.3 Output Analysis

There are some successful examples only produced by the proposed method; these are shown in Table 4. In these cases, the target word is located not at the end of the definition statements. The proposed method is able to acquire the target word in these cases, although they are few.

On the other hand, as shown in Table 5, the proposed method is the only one to produce certain unsuccessful examples. There are two major types of such errors: 1) The target word is selected at random because two or more SWs have the highest similarity of WordNet with original word, and 2) the non-paraphrasable word’s similarity is higher than that of the word at the end of the definition statement. For example, in the case of DW “再生 (play),” the SW “利用 (use)” has the highest WordNet similarity compared to the words from the end of the definition statement, but the SW “力 (power)” is acquired from another part, not the end of the definition statement, and its similarity is higher than the SW “利用 (use).” In this original sentence in Table 5, the DW “再生 (play)” and SW “力 (power)” are non-paraphrasable, but the DW “再生 (play)” and SW “利用 (use)” are paraphrasable.

Table 4: Successful examples from the proposed method without combination

Original	***警戒***は厳重、ピリピリしている。 ***Vigilance*** is strict, and the tension is so thick.
Paraphrase	***注意***は厳重、ピリピリしている。 ***Caution*** is strict, and the tension is so thick.
Definition Statement	【警戒】***注意***して用心すること 【vigilance】***caution*** and precaution
Original	とはいえ、勇気ある***決断***だ。 Although it is a courageous ***decision***
Paraphrase	とはいえ、勇気ある***決定***だ。 Although it is courageous ***determining***
Definition Statement	【決断】はっきりと***決定***した事柄 【decision】what was ***determined*** clearly
Original	***大詰め***の大一番 big match of the ***final stage***
Paraphrase	***最後***の大一番 big match of the ***last***
Definition Statement	【大詰め】芝居の***最後***の場面 【final stage】the ***last*** scene of the play

Table 5: Erroneous examples from the proposed method without combination

Original	主な***ポイント***をまとめた a summary of the main ***points***
Paraphrase	主な***点数***をまとめた a summary of the main ***scores***
Compared method	主な***要点***をまとめた a summary of the main ***essentials***
Definition Statements	ポイント：***要点***。***点数***。得点。地点。拠点。… Point: ***essential***. ***score***. game. spot. hub. …
Original	録画中の番組も***再生***できる I can also ***play*** the program during recording.
Paraphrase	録画中の番組も***力***できる I can also ***power*** the program during recording.
Compared method	録画中の番組も***利用***できる I can also ***use*** the program during recording.
Definition Statements	再生：廃物を再***利用***する。いったん消え失せていたものが、 ***力***や命を取り戻すこと。 Play: ***Use*** the garbage again. What was gone once again regains ***power*** and life. …

7. Conclusion

This paper demonstrates that to achieve lexical simplification for elementary school students, it is effective to paraphrase using definition sentences from multiple Japanese dictionaries and the lexical restrictions of BVL. Since the proposed method acquires target words from the full text of the definition, it may be able to select more appropriate target words than comparative methods, which make use of only the end of the definition statement. However, if the appropriate target word appears in other places (i.e., other than the end of the definition), which is the case for a few words in this experiment, the proposed method still achieves about the same level of the accuracy of paraphrase as does the comparative method.

It is necessary to select a proper target word from among several candidates that have been acquired. The results of this experiment show that the method of utilizing WordNet similarity is better than the method utilizing frequency and context information.

References

[1] Atsushi Fujita, Kentaro Inui, Hiroko Inui. 2000. An environment for constructing nominal-paraphrase corpora. *Technical Report of IEICE, TL, 100*(480): 53-60. (in Japanese).

[2] Zellig S. Harris. 1954. Distributional structure. *Word, 10*: 146-162.

[3] Nobuhiro Kaji, Daisuke Kawahara, Sadao Kurohashi, and Satoshi Sato. 2002. Verb paraphrase based on case frame alignment. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 215-222.

[4] Tomoyuki Kajiwara, Kazuhide Yamamoto. 2013. Lexical simplification and evaluation for children's reading assistance from multiple resources. *Proceedings of the 19th Annual Meeting of the Association for Natural Language Processing*, 272-275. (in Japanese).

[5] Frank Keller, Maria Lapata, and Olga Ourioupina. 2002. Using the web to overcome data sparseness. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 230-237.

[6] Maria Lapata, Frank Keller, and Scott McDonald. 2001. Evaluating smoothing algorithms against plausibility judgements. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, 346-353.

[7] Xiaojuan Ma, Christiane Fellbaum, and Perry R. Cook. 2010. A multimodal vocabulary for augmentative and alternative communication from sound/image label datasets. *Proceedings of the NAACL Human Language Technologies (HLT 2010) Workshop of Speech and Language Processing for Assistive Technologies*, 62-70.

[8] Manami Moku, Kazuhide Yamamoto and Ai Makabi. 2012. Automatic Easy Japanese Translation for information accessibility of foreigners. *Proceedings of Coling-2012 Workshop on Speech and Language Processing Tools in Education (SLP-TED), pp.85-90.*

[9] Hideya Mino, and Hideki Tanaka. 2011. Simplification of nominalized continuative verbs in broadcast news. *Proceedings of the 17th Annual Meeting of the Association for Natural Language Processing*, 744-747. (in Japanese).

[10] Kazuhide Yamamoto. 2002. Acquisition of Lexical Paraphrases from Texts. *Proceedings of 2nd International Workshop on Computational Terminology (Computerm 2002), no page numbers.*

Tools and Resources

1) Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. Enhancing the Japanese WordNet in The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009.

2) *EDR Japanese Word Dictionary*. Japan Electronic Dictionary Research Institute, Ltd. (EDR). 1995.

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

4) Hidetoshi Kenbo, Kyosuke Kindaichi, Haruhiko Kindaichi, Takeshi Shibata, and Yoshihumi Hida. 1994. *Sanseido Japanese Dictionary*. Sanseido Publishing Co., Ltd.

5) Taku Kudo. MeCab 0.993.
http://mecab.googlecode.com/svn/trunk/mecab/doc/criterion.html

6) Taku Kudo, Hideto Kazawa. Web Japanese N-gram Version 1. Published by Gengo Shigen Kyokai.
http://www.gsk.or.jp/catalog/GSK2007C/catalog.html

7) Yoshimasa Minato. 2011. *The Challenge Elementary School Japanese Dictionary*. Benesse Holdings, Inc.

8) The Mainichi Newspapers. 2000. Mainichi Shimbun CD-ROM 2000.