Noun Paraphrasing Based on a Variety of Contexts

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 paraphrase nouns along the contexts of sentence input on the basis of a variety of contexts obtained from a large-scale corpus. The proposed method only uses the number of types of context, not word frequency or co-occurrence frequency features. This method is based on the notion that paraphrase candidates appear more commonly with target words in the same context. The results of our experiment demonstrate that the approach can produce more appropriate paraphrases than approaches based on co-occurrence frequency and pointwise mutual information.

1 Introduction

Although extensive and various forms of text data are easily available in the present age, in order for readers to gather information effectively, they need technology that overcomes any differences in their linguistic competence. For example, technology that buries the difference in the linguistic competence of foreign language learners, children, the elderly, and disabled persons is useful (Inui and Fujita, 2004). We present our research on paraphrasing to control language at the elementary school level in order to simplify texts for children. We believe that vocabulary simplification for children can be realized by paraphrasing text according to Basic Vocabulary to Learn (BVL) (Kai and Matsukawa, 2002) . BVL is a collection of words selected on the basis on a lexical analysis of elementary school textbooks. It contains 5,404 words that can help children write expressively.

As previous work indicated, there are lexical paraphrases that define statements from a Japanese dictionary (Kajiwara et al., 2013). The definition statements from the Japanese dictionary explain a given headword in several easy words. Therefore, lexical simplification and paraphrasing that conserves a particular meaning are expected by paraphrasing the headword with the words in the definitions. However, definition statements are short sentences that consist of several words. Consequently, there are few paraphrase candidates, and natural paraphrasing is difficult even if we use certain dictionaries together. In addition, the definition statement as a whole is equivalent to the headword; there is no guarantee that any individual word extracted from the definition statement can paraphrase the headword.

We propose lexical paraphrasing based on a variety of contexts obtained from a large corpus without depending on existing lexical resources from such a background. The proposed method is not dependent on language, thus it can perform lexical paraphrases using a corpus of arbitrary languages. In this paper we examine and report on Japanese nouns.

2 Related Works

As paraphrase acquisition from a corpus, a study with a parallel corpus and comparable corpus has been performed. Barzilay and McKeown paraphrase text using plural English translations made from the same document (Barzilay and McKeown, 2001). In addition, Shinyama and Sekine paraphrase using plural newspaper articles that report the same event (Shinyama and Sekine, 2003). In a text sim-

28th Pacific Asia Conference on Language, Information and Computation pages 644–649

plification task, Coster and Kauchak create a parallel corpus that matches English Wikipedia and Simple English Wikipedia, and they perform text simplification using the framework of statistical machine translation (Coster and Kauchak, 2011). However, the technique of using these parallel corpora and comparable corpora is problematic in terms of the accuracy of alignment of corresponding expressions and quantity of the corpora that can be used. For example, for Japanese, there is no large-scale parallel corpus in which simplification is possible for use in the framework of statistical machine translation. In this paper, we generate paraphrases using only a single-language corpus so as not to come under these influences.

In their research with paraphrasing based on the similarity of the context obtained from a non-parallel corpus, Marton et al. propose a method for paraphrasing unknown words to improve machine translation systems (Marton et al., 2009). They select candidate words with a context common to the subject. Moreover, they calculate cosine similarities of their feature vectors based on the co-occurrence frequency of subjects. Bhagat and Ravichandran extract paraphrases from a massive, 25-billion word corpus (Bhagat and Ravichandran, 2008). They regard English word 5-gram as one phrase, and they generate feature vectors using pointwise mutual information (PMI) scores. They then select the best phrase-paraphrase pairs based on their cosine similarity.

Our proposed method is different from these methods in that it does not use co-occurrence frequency or word frequency of conventional features. We focus on the variety of context. Assuming that successful paraphrases have context that is common with their subject, we select paraphrases based only on the number of types of context.

3 Proposed Method

In this paper, noun paraphrasing is achieved based on the variety of contexts extracted from a large corpus. According to Harris's Distributional Hypothesis (Harris, 1954), first, the nouns used in a context similar to the input sentence are extracted from the corpus. Then, the context similarity for each extracted noun and the noun in the input sentence is calculated utilizing the case-frame dictionary. An abstract of the proposed method is illustrated in Figure 1.

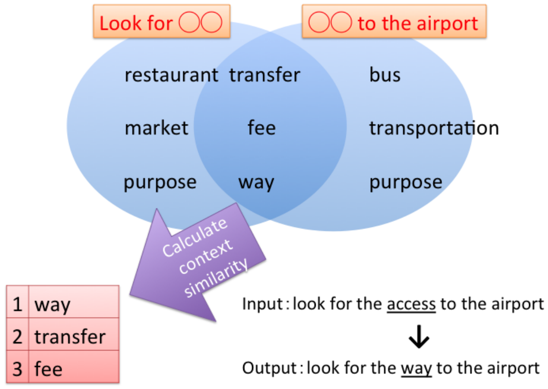
Look for ◯◯
◯◯ to the airport
restaurant
transfer
bus
market
fee
transportation
purpose
way
purpose
Calculate context similarity
1 way
2 transfer
3 fee
Input : look for the access to the airport
↓
Output : look for the way to the airport

Figure 1: Noun paraphrasing in the proposed method.

3.1 Extraction of Paraphrase Candidates

In this method, we hypothetically define the pre-phrase and post-phrase of the target noun as the context; nouns used in a similar context are extracted from the corpus.

First, the input sentence is divided into two different contexts: *pre*-context and *post*-context. Then, the input sentence is searched through each corpus. The common nouns found at the end (tail) of the pre-context and at the start (head) of the post-context are extracted.

For example, when the phrase look for the access to the airport is given as an input sentence and the word access is the paraphrase target word, the pre-context is look for X and the post-context is X to the airport. Both contexts are searched through the corpus for any phrases that have the exact same phrases next to the X for any other nouns, and the replaceable nouns for X are extracted. In the example shown in Figure 1, the pre-context and post-context have the words *transfer*, *fee*, and *way* in common.

3.2 Selection of Paraphrase Candidates

This paper forms two hypotheses and defines Equation (1) to obtain high values for similar context nouns to paraphrase a given target word.

$$sim(n_t, n_c) = com(n_t, n_c) * \log(\frac{N}{var(n_c)}) \tag{1}$$

$$cooccurrence(w_i, w_j) = \sum_{s_n \in S} freq_n(w_i, w_j) \tag{2}$$

$$pmi(w_i, w_j) = \log(\frac{cooccurrence(w_i, w_j) \sum_{s_n \in S} \sum_{w_m \in s_n} freq_n(w_m)}{\sum_{s_n \in S} freq_n(w_i) \sum_{s_n \in S} freq_n(w_j)}) \tag{3}$$

$$cos(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}||\vec{v}|} \tag{4}$$

1. When the paraphrased target word and the paraphrase candidate have the maximum possible number of common contexts, the paraphrasability increases.

2. When the paraphrase candidates have several different contexts, the paraphrasability decreases.

In the Equation (1), n_t is the paraphrase target noun, n_c is the paraphrase candidate noun, $com(n_t, n_c)$ is the number of types of common contexts, N is the sum of the number of contexts, and $var(n_c)$ is the unique number of contexts in which n_c is used. For the first term, if the number of different common contexts is large, the value also becomes larger. For the latter term, the fewer the number of contexts for the paraphrase candidate, the larger its value becomes. Hence, a high $sim(n_t, n_c)$ indicates that two contexts are similar.

According to the distribution hypothesis, the word of the similar meaning is used in the similar context. The first term of the Equation (1) expresses that context is similar so that there is much common context. However, the word used in many contexts, such as *boss* and *start*, cannot be said to be that the context resembles the paraphrase target noun even if $com(n_t, n_c)$ are large. Therefore we filter it in the latter term of the Equation (1) and lower score of the paraphrase candidate noun used in much context.

4 Experiment

4.1 Experimental Object

To test our proposed method, we conducted an experiment using the Web Japanese N-gram (Kudo and Kazawa, 2007). The Web Japanese N-gram includes the word N (1 to 7)-grams parsed by the Japanese language morphological analyzer MeCab (Kudo et al., 2004). Each N-gram appears more than 20 times in 20 billion sentences in Web text. We considered that the longest 7-gram data is a sentence and used all 570,204,252 sentences. In addition, we selected 1,365,705 sentences where the head was a noun and the tail was the original form of a verb. In the experiment we used most-frequent 200 sentences as a target. Also, nouns at the beginning of sentences are excluded. In addition, we used MeCab to determine the parts of speech.

4.2 Experimental Procedure

We calculated distributional similarity using the Kyoto University case frame (KCF) (Kawahara and Kurohashi, 2009) data on the extracted nouns. KCF is the predicate and noun pair that has a case relationship, and it is built automatically (Kawahara and Kurohashi, 2005) from 1.6 billion Web texts. In the experiment, we used all 34,059 predicates and 824,639 nouns. In addition, we assumed that these predicates are contexts and calculated their distributional similarity using Equation (1).

4.3 Evaluation

To evaluate the proposed method, we compared it with related paraphrasing methods based on distributional similarity. We selected nouns included in the top 10 similarities from 200 input sentences; in addition, we extracted the paraphrasing target as described in Section 4.1 using our proposed method, the method by Marton et al. (2009), and the method by Bhagat and Ravichandran (2008). Three evaluators selected one noun each to paraphrase with a paraphrasing target in an input sentence.

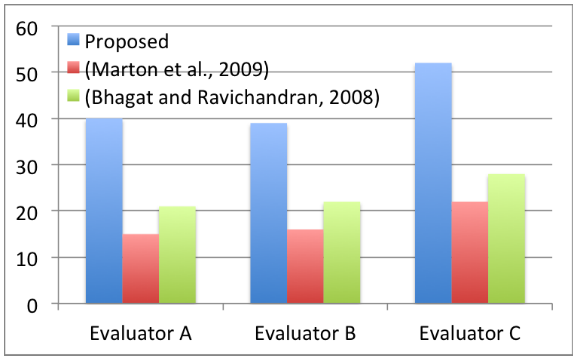

Figure 2: Number of paraphrasable nouns to first place of similarity.

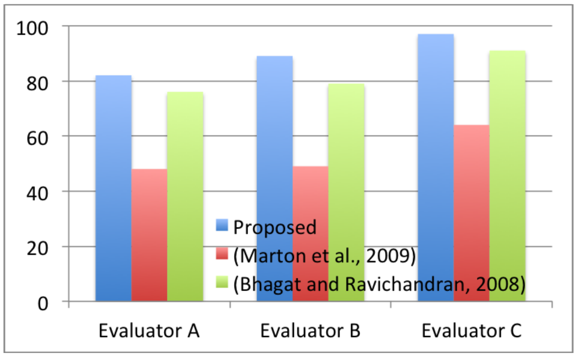

Figure 3: Number of paraphrasable nouns to the 10th place of similarity.

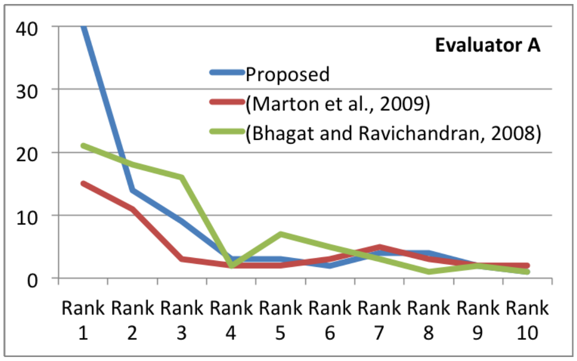

Figure 4: Relationships by order of similarity and number of paraphrasable nouns.

Marton et al. (2009) produce a feature vector by co-occurrence frequency with a noun and the context, and they calculate vector similarity by cosine. On the other hand, Bhagat and Ravichandran (2008) produce a feature vector by PMI with a noun and the context and calculates vector similarity by cosine. Both methods define nouns and verbs in dependency relationships to the context and produce feature vectors using Web Japanese N-gram. We define the co-occurrence frequency in Equation (2), PMI in Equation (3), and cosine similarity in Equation (4).

In the equations, $s_n \in S$, $w_m \in s_n$, $w_m \in W$, S is the set of sentences, W is the set of words, $freq_n(w_m)$ is the appearance frequency of word w_m in sentence n, $freq_n(w_i, w_j)$ is the co-occurrence frequency of word w_i and w_j in sentence n and $\vec{u}$ and $\vec{v}$ are the feature vectors.

5 Experiment Results

Figure 2 and Figure 3 show the evaluation results of the experiment described in Section 4, with a paraphrase of 200 sentences. The Fleiss's Kappa coefficient of three evaluators is 0.61. Thus, the agreement degree between raters is high enough.

Figure 2 shows the number of nouns evaluated as the possible paraphrase for each method.

On one hand, (Marton et al., 2009) applied the idea that the frequently co-occurring context is the important context. On the other hand, (Bhagat and Ravichandran, 2008) argued that the biasedly co-occurring context is important. Therefore, (Marton et al., 2009)'s method depends solely on high frequency words, whereas (Bhagat and Ravichandran, 2008)'s method relies on low frequency words. Hence, for (Marton et al., 2009)'s method, the word thing is suggested as the paraphrase candidate for 100 combinations out of 200 combinations. For (Bhagat and Ravichandran, 2008), the counter words, which are words that describe the number of items, are suggested as paraphrase candidates a significant number of times.

The proposed method does not rely on the frequency of the context; therefore, such an effect is disregarded as possible, and as a result, our method obtains high scores.

Figure 3 shows the number of nouns evaluated as possible candidates for paraphrases for the top 10 nouns of similarity. When observing the top 10 nouns, the results of (Bhagat and Ravichandran, 2008)'s method are close to the results of the proposed method. Figure 4 shows the rankings of similarities and the relationship of the number of possi-

Table 1: English translation of paraphrases generated by the proposed method.

Owner's [recognition → permission] is required.
Proceeding the [subject → problem] as important matter.
Generous [fee → price] is offered.
National agriculture's [advance → growth] is obstructed.
Education's [expansion → strengthening] are the examples.

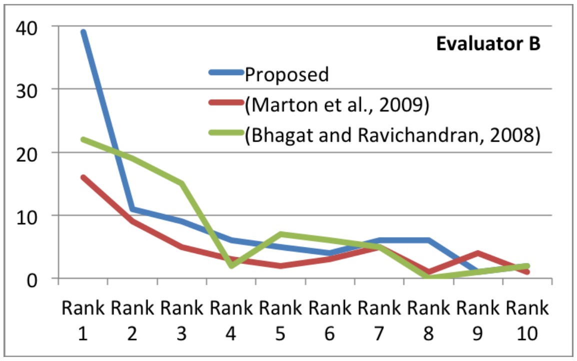

Figure 5: Relationships by order of similarity and number of paraphrasable nouns.

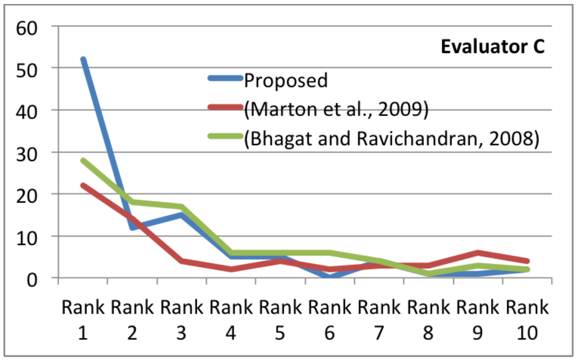

Figure 6: Relationships by order of similarity and number of paraphrasable nouns.

ble paraphrase candidates. Although Figure 4 shows the results for Evaluator A, the tendency is the same as for Evaluator B (Figure 5) and Evaluator C (Figure 6). In the results of the proposed method, there is a significant gap in the numbers of first-ranked and second-ranked nouns. However, in the results of (Bhagat and Ravichandran, 2008)'s method, the gap is insignificant. This is because the proposed method strictly applies the paraphrase process to nouns that are exactly in the context in which they are used in the input sentence. Because (Bhagat and Ravichandran, 2008)'s method does not consider the context of the input sentence, the quality is not always guaranteed to obtain the possible best score.

For instance, given an input sentence such as *assign a maximum [penalty] of $*, the paraphrase process for *[penalty]* in both (Marton et al., 2009) and (Bhagat and Ravichandran, 2008) grants *imprisonment* the highest score. On the other hand, the proposed method shows *paying penalty* with the best score, followed by correctional fine; *imprisonment* does not even appear as a candidate.

For the input sentence, *reduce the [burdens] on the back*, in the case of paraphrasing *[burdens]*, (Bhagat and Ravichandran, 2008)'s method suggests *cost*, *expenses*, and *actual cost*, all of which are money-related; any words listed within the top 10 are not appropriate paraphrase candidates.

Meanwhile, the proposed method suggests *loads*, *stress*, *damage*, *exhaustion*, *tense*, *impact*, etc., all of which are considerably appropriate for paraphrasing. Table 1 presents a list of successful examples.

6 Conclusion and Future Work

In this paper, we showed the effectiveness of the method of paraphrasing a noun along the context of a given input sentence based on the variety of contexts obtained from a large-scale corpus. Our proposed method can paraphrase nouns depending on the context of the input sentence, and we can obtain the appropriate paraphrase independently of the appearance frequency and co-occurrence frequency of the word. This is because we select a noun that shares more contexts with the paraphrasing target in the paraphrase.

This paper discussed the validity of paraphrases using a different statistics value from frequency called the number of types of the context. Our goal is to simplify vocabulary by paraphrasing, and it considers the restriction to plain vocabularies, such as the Basic Vocabulary to Learn, to maintain the accuracy and comprehensibility of lexical paraphrasing.

References

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, pages 50–57.

Rahul Bhagat and Deepak Ravichandran. 2008. Large Scale Acquisition of Paraphrases for Learning Surface Patterns. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*, pages 674–682.

William Coster and David Kauchak. 2011. Simple Wikipedia: A New Simplification Task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 665–669.

Zellig S. Harris. 1954. Distributional Structure. *Word*, 10(23):146–162.

Kentaro Inui and Atsushi Fujita. 2004. A Survey on Paraphrase Generation and Recognition. *Journal of Natural Language Processing*, 11(5):151–198.

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

Tomoyuki Kajiwara, Hiroshi Matsumoto and Kazuhide Yamamoto. 2013. Selecting Proper Lexical Paraphrase for Children. In *Proceedings of the Twenty-Fifth Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73.

Daisuke Kawahara and Sadao Kurohashi. 2005. Gradual Fertilization of Case Frames. *Journal of Natural Language Processing*, 12(2):109–131.

Daisuke Kawahara and Sadao Kurohashi. 2009. Kyoto University’s Case Frame Data ver 1.0. Gengo Shigen Kyokai.

Taku Kudo and Hideto Kazawa. 2007. Web Japanese N-gram Version 1. Gengo Shigen Kyokai.

Taku Kudo, Kaoru Yamamoto and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*

Yuval Marton, Chris Callison-Burch and Philip Resnik. 2009. Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 381–390.

Yusuke Shinyama and Satoshi Sekine. 2003. Paraphrase Acquisition for Information Extraction. In *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP 2003)*, pages 65–71.