Extracting Paraphrases from Definition Sentences on the Web

Chikara Hashimoto[*] **Kentaro Torisawa**[†] **Stijn De Saeger**[‡]
Jun'ichi Kazama[§] **Sadao Kurohashi**[¶]
[*†‡§] National Institute of Information and Communications Technology
Kyoto, 619-0237, JAPAN
[*¶] Graduate School of Informatics, Kyoto University
Kyoto, 606-8501, JAPAN
{[*] ch, [†] torisawa, [‡] stijn, [§] kazama}@nict.go.jp
[¶] kuro@i.kyoto-u.ac.jp

Abstract

We propose an automatic method of extracting paraphrases from definition sentences, which are also automatically acquired from the Web. We observe that a huge number of concepts are defined in Web documents, and that the sentences that define the same concept tend to convey mostly the same information using different expressions and thus contain many paraphrases. We show that a large number of paraphrases can be automatically extracted with high precision by regarding the sentences that define the same concept as parallel corpora. Experimental results indicated that with our method it was possible to extract about 300,000 paraphrases from 6×10^8 Web documents with a precision rate of about 94%.

1 Introduction

Natural language allows us to express the same information in many ways, which makes natural language processing (NLP) a challenging area. Accordingly, many researchers have recognized that automatic paraphrasing is an indispensable component of intelligent NLP systems (Iordanskaja et al., 1991; McKeown et al., 2002; Lin and Pantel, 2001; Ravichandran and Hovy, 2002; Kauchak and Barzilay, 2006; Callison-Burch et al., 2006) and have tried to acquire a large amount of paraphrase knowledge, which is a key to achieving robust automatic paraphrasing, from corpora (Lin and Pantel, 2001; Barzilay and McKeown, 2001; Shinyama et al., 2002; Barzilay and Lee, 2003).

We propose a method to extract phrasal paraphrases from pairs of sentences that define the same concept. The method is based on our observation that two sentences defining the same concept can be regarded as a parallel corpus since they largely convey the same information using different expressions. Such definition sentences abound on the Web. This suggests that we may be able to extract a large amount of phrasal paraphrase knowledge from the definition sentences on the Web.

For instance, the following two sentences, both of which define the same concept "osteoporosis", include two pairs of phrasal paraphrases, which are indicated by underlines ① and ②, respectively.

(1) a. Osteoporosis is a disease that ① decreases the quantity of bone and ② makes bones fragile.
b. Osteoporosis is a disease that ① reduces bone mass and ② increases the risk of bone fracture.

We define *paraphrase* as a pair of expressions between which entailment relations of both directions hold. (Androutsopoulos and Malakasiotis, 2010).

Our objective is to extract phrasal paraphrases from pairs of sentences that define the same concept. We propose a supervised method that exploits various kinds of lexical similarity features and contextual features. Sentences defining certain concepts are acquired automatically on a large scale from the Web by applying a quite simple supervised method.

Previous methods most relevant to our work used parallel corpora such as multiple translations of the same source text (Barzilay and McKeown, 2001) or automatically acquired parallel news texts (Shinyama et al., 2002; Barzilay and Lee, 2003; Dolan et al., 2004). The former requires a large amount of manual labor to translate the same texts

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1087–1097,
Portland, Oregon, June 19-24, 2011.

in several ways. The latter would suffer from the fact that it is not easy to automatically retrieve large bodies of parallel news text with high accuracy. On the contrary, recognizing definition sentences for the same concept is quite an easy task at least for Japanese, as we will show, and we were able to find a huge amount of definition sentence pairs from normal Web texts. In our experiments, about 30 million definition sentence pairs were extracted from 6×10^8 Web documents, and the estimated number of paraphrases recognized in the definition sentences using our method was about 300,000, for a precision rate of about 94%. Also, our experimental results show that our method is superior to well-known competing methods (Barzilay and McKeown, 2001; Koehn et al., 2007) for extracting paraphrases from definition sentence pairs.

Our evaluation is based on bidirectional checking of entailment relations between paraphrases that considers the context dependence of a paraphrase.

Note that using definition sentences is only the beginning of our research on paraphrase extraction. We have a more general hypothesis that sentences fulfilling the same *pragmatic* function (e.g. definition) for the same topic (e.g. osteoporosis) convey mostly the same information using different expressions. Such functions other than definition may include the usage of the same Linux command, the recipe for the same cuisine, or the description of related work on the same research issue.

Section 2 describes related works. Section 3 presents our proposed method. Section 4 reports on evaluation results. Section 5 concludes the paper.

2 Related Work

The existing work for paraphrase extraction is categorized into two groups. The first involves a distributional similarity approach pioneered by Lin and Pantel (2001). Basically, this approach assumes that two expressions that have a large distributional similarity are paraphrases. There are also variants of this approach that address entailment acquisition (Geffet and Dagan, 2005; Bhagat et al., 2007; Szpektor and Dagan, 2008; Hashimoto et al., 2009). These methods can be applied to a normal monolingual corpus, and it has been shown that a large number of paraphrases or entailment rules could be extracted. However, the precision of these methods has been relatively low. This is due to the fact that the evidence, i.e., distributional similarity, is just indirect evidence of paraphrase/entailment. Accordingly, these methods occasionally mistake antonymous pairs for paraphrases/entailment pairs, since an expression and its antonymous counterpart are also likely to have a large distributional similarity. Another limitation of these methods is that they can find only paraphrases consisting of frequently observed expressions since they must have *reliable* distributional similarity values for expressions that constitute paraphrases.

The second category is a parallel corpus approach (Barzilay and McKeown, 2001; Shinyama et al., 2002; Barzilay and Lee, 2003; Dolan et al., 2004). Our method belongs to this category. This approach aligns expressions between two sentences in parallel corpora, based on, for example, the overlap of words/contexts. The aligned expressions are assumed to be paraphrases. In this approach, the expressions do not need to appear frequently in the corpora. Furthermore, the approach rarely mistakes antonymous pairs for paraphrases/entailment pairs. However, its limitation is the difficulty in preparing a large amount of parallel corpora, as noted before. We avoid this by using definition sentences, which can be easily acquired on a large scale from the Web, as parallel corpora.

Murata et al. (2004) used definition sentences in two manually compiled dictionaries, which are considerably fewer in the number of definition sentences than those on the Web. Thus, the coverage of their method should be quite limited. Furthermore, the precision of their method is much poorer than ours as we report in Section 4.

For a more extensive survey on paraphrasing methods, see Androutsopoulos and Malakasiotis (2010) and Madnani and Dorr (2010).

3 Proposed method

Our method, targeting the Japanese language, consists of two steps: definition sentence acquisition and paraphrase extraction. We describe them below.

3.1 Definition sentence acquisition

We acquire sentences that define a concept (definition sentences) as in Example (2), which defines "骨

粗鬆症" (osteoporosis), from the 6×10^8 Web pages (Akamine et al., 2010) and the Japanese Wikipedia.

(2) 骨粗鬆症**とは、**骨がもろくなってしまう病気だ。
(Osteoporosis is a disease that makes bones fragile.)

Fujii and Ishikawa (2002) developed an unsupervised method to find definition sentences from the Web using 18 sentential templates and a language model constructed from an encyclopedia. On the other hand, we developed a supervised method to achieve a higher precision.

We use one sentential template and an SVM classifier. Specifically, we first collect definition sentence candidates by a template "^NP **とは**.*", where ^ is the beginning of sentence and NP is the noun phrase expressing the concept to be defined followed by a particle sequence, "**と**" (comitative) and "**は**" (topic) (and optionally followed by comma), as exemplified in (2). As a result, we collected 3,027,101 sentences. Although the particle sequence tends to mark the topic of the definition sentence, it can also appear in interrogative sentences and normal assertive sentences in which a topic is strongly emphasized. To remove such non-definition sentences, we classify the candidate sentences using an SVM classifier with a polynominal kernel ($d = 2$).[1] Since Japanese is a head-final language and we can judge whether a sentence is interrogative or not from the last words in the sentence, we included morpheme N-grams and bag-of-words (with the window size of N) at the end of sentences in the feature set. The features are also useful for confirming that the head verb is in the present tense, which definition sentences should be. Also, we added the morpheme N-grams and bag-of-words right after the particle sequence in the feature set since we observe that non-definition sentences tend to have interrogative related words like "何" (what) or "一体" ((what) on earth) right after the particle sequence. We chose 5 as N from our preliminary experiments.

Our training data was constructed from 2,911 sentences randomly sampled from all of the collected sentences. 61.1% of them were labeled as positive. In the 10-fold cross validation, the classifier's accuracy, precision, recall, and F1 were 89.4, 90.7, 92.2, and 91.4, respectively. Using the classifier, we acquired 1,925,052 positive sentences from all of the collected sentences. After adding definition sentences from Wikipedia articles, which are typically the first sentence of the body of each article (Kazama and Torisawa, 2007), we obtained a total of 2,141,878 definition sentence candidates, which covered 867,321 concepts ranging from weapons to rules of baseball. Then, we coupled two definition sentences whose defined concepts were the same and obtained 29,661,812 definition sentence pairs.

Obviously, our method is tailored to Japanese. For a language-independent method of definition acquisition, see Navigli and Velardi (2010) as an example.

3.2 Paraphrase extraction

Paraphrase extraction proceeds as follows. First, each sentence in a pair is parsed by the dependency parser KNP[2] and dependency tree fragments that constitute linguistically well-formed constituents are extracted. The extracted dependency tree fragments are called *candidate phrases* hereafter. We restricted candidate phrases to predicate phrases that consist of at least one dependency relation, do not contain demonstratives, and in which all the leaf nodes are nominal and all of the constituents are consecutive in the sentence. KNP indicates whether each candidate phrase is a predicate based on the POS of the head morpheme. Then, we check all the pairs of candidate phrases between two definition sentences to find paraphrase pairs.[3] In (1), repeated in (3), candidate phrase pairs to be checked include (① decreases the quantity of bone, ① reduces bone mass), (① decreases the quantity of bone, ② increases the risk of bone fracture), (② makes bones fragile, ① reduces bone mass), and (② makes bones fragile, ② increases the risk of bone fracture).

(3) a. Osteoporosis is a disease that ① decreases the quantity of bone and ② makes bones fragile.
b. Osteoporosis is a disease that ① reduces bone mass and ② increases the risk of bone fracture.

[1] We use SVM^{light} available at `http://svmlight.joachims.org/`.

[2] `http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html`.

[3] Our method discards candidate phrase pairs in which one subsumes the other in terms of their character string, or the difference is only one proper noun like "toner cartridges that Apple Inc. made" and "toner cartridges that Xerox made." Proper nouns are recognized by KNP.

f1	The ratio of the number of morphemes shared between two candidate phrases to the number of all of the morphemes in the two phrases.
f2	The ratio of the number of a candidate phrase's morphemes, for which there is a morpheme with small edit distance (1 in our experiment) in another candidate phrase, to the number of all of the morphemes in the two phrases. Note that Japanese has many orthographical variations and edit distance is useful for identifying them.
f3	The ratio of the number of a candidate phrase's morphemes, for which there is a morpheme with the same pronunciation in another candidate phrase, to the number of all of the morphemes in the two phrases. Pronunciation is also useful for identifying orthographic variations. Pronunciation is given by KNP.
f4	The ratio of the number of morphemes of a shorter candidate phrase to that of a longer one.
f5	The identity of the inflected form of the head morpheme between two candidate phrases: 1 if they are identical, 0 otherwise.
f6	The identity of the POS of the head morpheme between two candidate phrases: 1 or 0.
f7	The identity of the inflection (conjugation) of the head morpheme between two candidate phrases: 1 or 0.
f8	The ratio of the number of morphemes that appear in a candidate phrase segment of a definition sentence s_1 and in a segment that is NOT a part of the candidate phrase of another definition sentence s_2 to the number of all of the morphemes of s_1's candidate phrase, i.e. how many extra morphemes are incorporated into s_1's candidate phrase.
f9	The reversed ($s_1 \leftrightarrow s_2$) version of **f8**.
f10	The ratio of the number of parent dependency tree fragments that are shared by two candidate phrases to the number of all of the parent dependency tree fragments of the two phrases. Dependency tree fragments are represented by the pronunciation of their component morphemes.
f11	A variation of **f10**; tree fragments are represented by the base form of their component morphemes.
f12	A variation of **f10**; tree fragments are represented by the POS of their component morphemes.
f13	The ratio of the number of unigrams (morphemes) that appear in the child context of both candidate phrases to the number of all of the child context morphemes of both candidate phrases. Unigrams are represented by the pronunciation of the morpheme.
f14	A variation of **f13**; unigrams are represented by the base form of the morpheme.
f15	A variation of **f14**; the numerator is the number of child context unigrams that are adjacent to both candidate phrases.
f16	The ratio of the number of trigrams that appear in the child context of both candidate phrases to the number of all of the child context morphemes of both candidate phrases. Trigrams are represented by the pronunciation of the morpheme.
f17	Cosine similarity between two definition sentences from which a candidate phrase pair is extracted.

Table 1: Features used by paraphrase classifier.

The paraphrase checking of candidate phrase pairs is performed by an SVM classifier with a linear kernel that classifies each pair of candidate phrases into a paraphrase or a non-paraphrase.[4] Candidate phrase pairs are ranked by their distance from the SVM's hyperplane. Features for the classifier are based on our observation that two candidate phrases tend to be paraphrases if the candidate phrases themselves are sufficiently similar and/or their surrounding contexts are sufficiently similar. Table 1 lists the features used by the classifier.[5] Basically, they represent either the similarity of candidate phrases (**f1-9**) or that of their contexts (**f10-17**). We think that they have various degrees of discriminative power, and thus we use the SVM to adjust their weights. Figure 1 illustrates features **f8-12**, for which you may need supplemental remarks. English is used for ease of explanation. In the figure, **f8** has a positive value since the candidate phrase of s_1 contains morphemes "of bone", which do not appear in the candidate phrase of s_2 but do appear in the other part of s_2, i.e. they are extra morphemes for s_1's candidate phrase. On the other hand, **f9** is zero since there is no such extra morpheme in s_2's candidate phrase. Also, features **f10-12** have positive values since the two candidate phrases share two parent dependency tree fragments, *(that increases)* and *(of fracture)*.

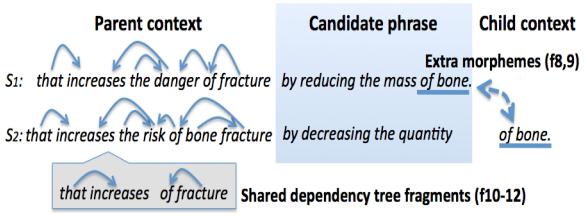


Figure 1: Illustration of features **f8-12**.

We have also tried the following features, which we do not detail due to space limitation: the similarity of candidate phrases based on semantically similar nouns (Kazama and Torisawa, 2008), entailing/entailed verbs (Hashimoto et al., 2009), and the identity of the pronunciation and base form of the head morpheme; N-grams (N=1,2,3) of child and parent contexts represented by either the inflected form, base form, pronunciation, or POS of mor-

[4] We use SVM^{perf} available at `http://svmlight.joachims.org/svm_perf.html`.

[5] In the table, the parent context of a candidate phrase consists of expressions that appear in ancestor nodes of the candidate phrase in terms of the dependency structure of the sentence. Child contexts are defined similarly.

Original definition sentence pair (s_1, s_2)	Paraphrased definition sentence pair (s'_1, s'_2)
s_1: Osteoporosis is a disease that **reduces bone mass** and makes bones fragile.	s'_1: Osteoporosis is a disease that **decreases the quantity of bone** and makes bones fragile.
s_2: Osteoporosis is a disease that **decreases the quantity of bone** and increases the risk of bone fracture.	s'_2: Osteoporosis is a disease that **reduces bone mass** and increases the risk of bone fracture.

Figure 2: Bidirectional checking of entailment relation ($\rightarrow$) of $p_1 \rightarrow p_2$ and $p_2 \rightarrow p_1$. p_1 is "**reduces bone mass**" in s_1 and p_2 is "**decreases the quantity of bone**" in s_2. p_1 and p_2 are exchanged between s_1 and s_2 to generate corresponding paraphrased sentences s'_1 and s'_2. $p_1 \rightarrow p_2$ ($p_2 \rightarrow p_1$) is verified if $s_1 \rightarrow s'_1$ ($s_2 \rightarrow s'_2$) holds. In this case, both of them hold. English is used for ease of explanation.

pheme; parent/child dependency tree fragments represented by either the inflected form, base form, pronunciation, or POS; adjacent versions (cf. **f15**) of N-gram features and parent/child dependency tree features. These amount to 78 features, but we eventually settled on the 17 features in Table 1 through ablation tests to evaluate the discriminative power of each feature.

The ablation tests were conducted using training data that we prepared. In preparing the training data, we faced the problem that the completely random sampling of candidate paraphrase pairs provided us with only a small number of positive examples. Thus, we automatically collected candidate paraphrase pairs that were expected to have a high likelihood of being positive as examples to be labeled. The likelihood was calculated by simply summing all of the 78 feature values that we have tried, since they indicate the likelihood of a given candidate paraphrase pair's being a paraphrase. Note that values of the features f8 and f9 are weighted with -1, since they indicate the unlikelihood. Specifically, we first randomly sampled 30,000 definition sentence pairs from the 29,661,812 pairs, and collected 3,000 candidate phrase pairs that had the highest likelihood from them. The manual labeling of each candidate phrase pair (p_1, p_2) was based on bidirectional checking of entailment relation, $p_1 \rightarrow p_2$ and $p_2 \rightarrow p_1$, with p_1 and p_2 embedded in contexts.

This scheme is similar to the one proposed by Szpektor et al. (2007). We adopt this scheme since paraphrase judgment might be unstable between annotators unless they are given a particular context based on which they make a judgment. As described below, we use definition sentences as contexts. We admit that annotators might be biased by this in some unexpected way, but we believe that this is a more stable method than that without contexts. The labeling process is as follows. First, from each candidate phrase pair (p_1, p_2) and its source definition sentence pair (s_1, s_2), we create two paraphrase sentence pairs (s'_1, s'_2) by exchanging p_1 and p_2 between s_1 and s_2. Then, annotators check if s_1 entails s'_1 and s_2 entails s'_2 so that entailment relations of both directions $p_1 \rightarrow p_2$ and $p_2 \rightarrow p_1$ are checked. Figure 2 shows an example of bidirectional checking. In this example, both entailment relations, $s_1 \rightarrow s'_1$ and $s_2 \rightarrow s'_2$, hold, and thus the candidate phrase pair (p_1, p_2) is judged as positive. We used (p_1, p_2), for which entailment relations of both directions held, as positive examples (1,092 pairs) and the others as negative ones (1,872 pairs).[6]

We built the paraphrase classifier from the training data. As mentioned, candidate phrase pairs were ranked by the distance from the SVM's hyperplane.

4 Experiment

In this paper, our claims are twofold.

I. Definition sentences on the Web are a treasure trove of paraphrase knowledge (Section 4.2).

II. Our method of paraphrase acquisition from definition sentences is more accurate than well-known competing methods (Section 4.1).

We first verify claim II by comparing our method with that of Barzilay and McKeown (2001) (BM method), Moses[7] (Koehn et al., 2007) (SMT method), and that of Murata et al. (2004) (Mrt method). The first two methods are well known for accurately extracting semantically equivalent phrase pairs from parallel corpora.[8] Then, we verify claim

[6] The remaining 36 pairs were discarded as they contained garbled characters of Japanese.

[7] `http://www.statmt.org/moses/`

[8] As anonymous reviewers pointed out, they are unsupervised methods and thus unable to be adapted to definition sen-

I by comparing definition sentence pairs with sentence pairs that are acquired from the Web using Yahoo!JAPAN API[9] as a paraphrase knowledge source. In the latter data set, two sentences of each pair are expected to be semantically similar regardless of whether they are definition sentences. Both sets contain 100,000 pairs.

Three annotators (not the authors) checked evaluation samples. Fleiss' kappa (Fleiss, 1971) was 0.69 (substantial agreement (Landis and Koch, 1977)).

4.1 Our method vs. competing methods

In this experiment, paraphrase pairs are extracted from 100,000 definition sentence pairs that are randomly sampled from the 29,661,812 pairs. Before reporting the experimental results, we briefly describe the BM, SMT, and Mrt methods.

BM method Given parallel sentences like multiple translations of the same source text, the BM method works iteratively as follows. First, it collects from the parallel sentences identical word pairs and their contexts (POS N-grams with indices indicating corresponding words between paired contexts) as positive examples and those of different word pairs as negative ones. Then, each context is ranked based on the frequency with which it appears in positive (negative) examples. The most likely K positive (negative) contexts are used to extract positive (negative) paraphrases from the parallel sentences. Extracted positive (negative) paraphrases and their morpho-syntactic patterns are used to collect additional positive (negative) contexts. All the positive (negative) contexts are ranked, and additional paraphrases and their morpho-syntactic patterns are extracted again. This iterative process finishes if no further paraphrase is extracted or the number of iterations reaches a predefined threshold T. In this experiment, following Barzilay and McKeown (2001), K is 10 and N is 1 to 3. The value of T is not given in their paper. We chose 3 as its value based on our preliminary experiments. Note that paraphrases extracted by this method are not ranked.

SMT method Our SMT method uses Moses (Koehn et al., 2007) and extracts a phrase table, a set of two phrases that are *translations* of each other, given a set of two sentences that are *translations* of each other. If you give Moses *monolingual* parallel sentence pairs, it should extract a set of two phrases that are *paraphrases* of each other. In this experiment, default values were used for all parameters. To rank extracted phrase pairs, we assigned each of them the product of two phrase translation probabilities of both directions that were given by Moses. For other SMT methods, see Quirk et al. (2004) and Bannard and Callison-Burch (2005) among others.

Mrt method Murata et al. (2004) proposed a method to extract paraphrases from two manually compiled dictionaries. It simply regards a difference between two definition sentences of the same word as a paraphrase candidate. Paraphrase candidates are ranked according to an unsupervised scoring scheme that implements their assumption. They assume that a paraphrase candidate tends to be a valid paraphrase if it is surrounded by infrequent strings and/or if it appears multiple times in the data.

In this experiment, we evaluated the unsupervised version of our method in addition to the supervised one described in Section 3.2, in order to compare it fairly with the other methods. The unsupervised method works in the same way as the supervised one, except that it ranks candidate phrase pairs by the sum of all 17 feature values, instead of the distance from the SVM's hyperplane. In other words, no supervised learning is used. All the feature values are weighted with 1, except for f8 and f9, which are weighted with -1 since they indicate the unlikelihood of a candidate phrase pair being paraphrases. BM, SMT, Mrt, and the two versions of our method were used to extract paraphrase pairs from the same 100,000 definition sentence pairs.

Evaluation scheme Evaluation of each paraphrase pair (p_1, p_2) was based on bidirectional checking of entailment relations $p_1 \rightarrow p_2$ and $p_2 \rightarrow p_1$ in a way similar to the labeling of the training data. The difference is that contexts for evaluation are two sentences that are retrieved from the Web and contain p_1 and p_2, instead of definition sentences from which p_1 and p_2 are extracted. This

tences. Nevertheless, we believe that comparing these methods with ours is very informative, since they are known to be accurate and have been influential.

[9] http://developer.yahoo.co.jp/webapi/

is intended to check whether extracted paraphrases are also valid for contexts other than those from which they are extracted. The evaluation proceeds as follows. For the top m paraphrase pairs of each method (in the case of the BM method, randomly sampled m pairs were used, since the method does not rank paraphrase pairs), we retrieved a sentence pair (s_1, s_2) for each paraphrase pair (p_1, p_2) from the Web, such that s_1 contains p_1 and s_2 contains p_2. In doing so, we make sure that neither s_1 nor s_2 are the definition sentences from which p_1 and p_2 are extracted. For each method, we randomly sample n samples from all of the paraphrase pairs (p_1, p_2) for which both s_1 and s_2 are retrieved. Then, from each (p_1, p_2) and (s_1, s_2), we create two paraphrase sentence pairs (s'_1, s'_2) by exchanging p_1 and p_2 between s_1 and s_2. All samples, each consisting of (p_1, p_2), (s_1, s_2), and (s'_1, s'_2), are checked by three human annotators to determine whether s_1 entails s'_1 and s_2 entails s'_2 so that entailment relations of both directions are verified. In advance of evaluation annotation, all the evaluation samples are shuffled so that the annotators cannot find out which sample is given by which method for fairness. We regard each paraphrase pair as correct if at least two annotators judge that entailment relations of both directions hold for it. You may wonder whether only one pair of sentences (s_1, s_2) is enough for evaluation since a correct (wrong) paraphrase pair might be judged as wrong (correct) accidentally. Nevertheless, we suppose that the final evaluation results are reliable if the number of evaluation samples is sufficient. In this experiment, m is 5,000 and n is 200. We use Yahoo!JAPAN API to retrieve sentences.

Graph (a) in Figure 3 shows a precision curve for each method. *Sup* and *Uns* respectively indicate the supervised and unsupervised versions of our method. The figure indicates that *Sup* outperforms all the others and shows a high precision rate of about 94% at the top 1,000. Remember that this is the result of using 100,000 definition sentence pairs. Thus, we estimate that *Sup* can extract about 300,000 paraphrase pairs with a precision rate of about 94%, if we use all 29,661,812 definition sentence pairs that we acquired.

Furthermore, we measured precision after trivial paraphrase pairs were discarded from the evaluation samples of each method. A candidate phrase pair (p_1, p_2) is regarded as trivial if the pronunciation is the same between p_1 and p_2,[10] or all of the content words contained in p_1 are the same as those of p_2. Graph (b) gives a precision curve for each method. Again, *Sup* outperforms the others too, and maintains a precision rate of about 90% until the top 1,000. These results support our claim II.

Definition sentence pairs	Sup	Uns	BM	SMT	Mrt
with trivial	1,381,424		24,049	9,562	18,184
without trivial	1,377,573		23,490	7,256	18,139
Web sentence pairs	Sup	Uns	BM	SMT	Mrt
with trivial	277,172		5,101	4,586	4,978
without trivial	274,720		4,399	2,342	4,958

Table 2: Number of extracted paraphrases.

The upper half of Table 2 shows the number of extracted paraphrases with/without trivial pairs for each method.[11] *Sup* and *Uns* extracted many more paraphrases. It is noteworthy that *Sup* performed the best in terms of both precision rate and the number of extracted paraphrases.

Table 3 shows examples of correct and incorrect outputs of *Sup*. As the examples indicate, many of the extracted paraphrases are not specific to definition sentences and seem very reusable. However, there are few paraphrases involving metaphors or idioms in the outputs due to the nature of definition sentences. In this regard, we do not claim that our method is almighty. We agree with Sekine (2005) who claims that several different methods are required to discover a wider variety of paraphrases.

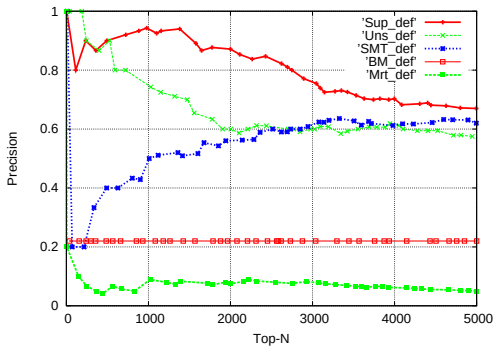
In graphs (a) and (b), the precision of the SMT method goes up as rank goes down. This strange behavior is due to the scoring by Moses that worked poorly for the data; it gave 1.0 to 82.5% of all the samples, 38.8% of which were incorrect. We suspect SMT methods are poor at monolingual alignment for paraphrasing or entailment tasks since, in the tasks, data is much noisier than that used for SMT. See MacCartney et al. (2008) for similar discussion.

4.2 Definition pairs vs. Web sentence pairs

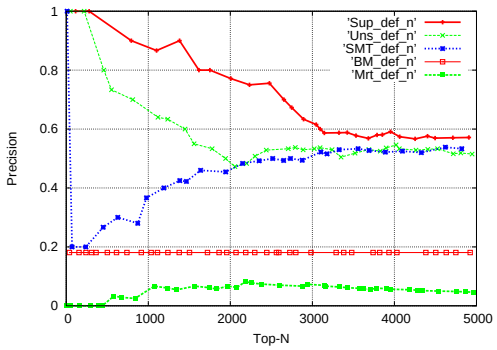
To collect Web sentence pairs, first, we randomly sampled 1.8 million sentences from the Web corpus.

[10] There are many kinds of orthographic variants in Japanese, which can be identified by their pronunciation.

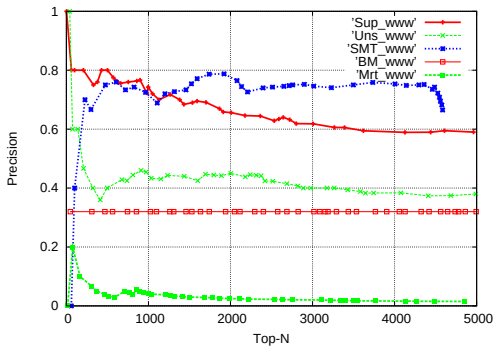
[11] We set no threshold for candidate phrase pairs of each method, and counted all the candidate phrase pairs in Table 2.



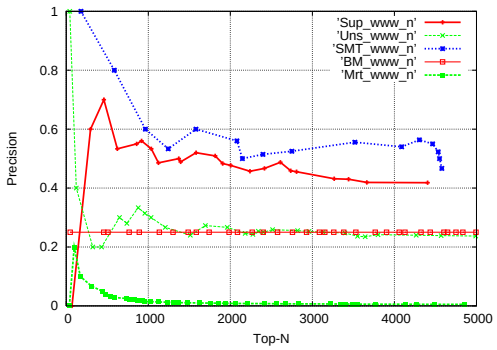
(a) Definition sentence pairs with trivial paraphrases



(b) Definition sentence pairs without trivial paraphrases



(c) Web sentence pairs with trivial paraphrases



(d) Web sentence pairs without trivial paraphrases

Figure 3: Precision curves of paraphrase extraction.

	Rank	Paraphrase pair
Correct	13	メールアドレスにメールを送る (send a message to the e-mail address) ⇔ メールアドレスに電子メールを送る (send an e-mail message to the e-mail address)
	19	お客様の依頼による (requested by a customer) ⇔ お客様の委託による (commissioned by a customer)
	70	企業の財政状況を表す (describe the fiscal condition of company) ⇔ 企業の財政状態を示す (indicate the fiscal state of company)
	112	インフォメーションを得る (get information) ⇔ ニュースを得る (get news)
	656	きまりのことです (it is a convention) ⇔ ルールのことです (it is a rule)
	841	地震のエネルギー規模をあらわす (represent the energy scale of earthquake) ⇔ 地震の規模を表す (represent the scale of earthquake)
	929	細胞を酸化させる (cause the oxidation of cells) ⇔ 細胞を老化させる (cause cellular aging)
	1,553	角質を取り除く (remove dead skin cells) ⇔ 角質をはがす (peel off dead skin cells)
	2,243	胎児の発育に必要だ (required for the development of fetus) ⇔ 胎児の発育成長に必要不可欠だ (indispensable for the growth and development of fetus)
	2,855	視力を矯正する (correct eyesight) ⇔ 視力矯正を行う (perform eyesight correction)
	2,931	チャラにしてもらう (call it even) ⇔ 帳消しにしてもらう (call it quits)
	3,667	ハードディスク上に蓄積される (accumulated on a hard disk) ⇔ ハードディスクドライブに保存される (stored on a hard disk drive)
	4,870	有害物質を排泄する (excrete harmful substance) ⇔ 有害毒素を排出する (discharge harmful toxin)
	5,501	１つのＣＰＵの内部に２つのプロセッサコアを搭載する (mount two processor cores on one CPU) ⇔ １つのパッケージに２つのプロセッサコアを集積する (build two processor cores into one package)
	10,675	外貨を売買する (trade foreign currencies) ⇔ 通貨を交換する (exchange one currency for another)
	112,819	派遣先企業の社員になる (become a regular staff member of the company where (s)he has worked as a temp) ⇔ 派遣先に直接雇用される (employed by the company where (s)he has worked as a temp)
	193,553	Ｗｅｂサイトにアクセスする (access Web sites) ⇔ WWWサイトを訪れる (visit WWW sites)
Incorrect	903	ブラウザに送信される (send to a Web browser) ⇔ パソコンに送信される (send to a PC)
	2,530	調和をはかる (intend to balance) ⇔ リフレッシュを図る (intend to refresh)
	3,008	消化酵素では消化できない (unable to digest with digestive enzymes) ⇔ 消化酵素で消化され難い (hard to digest with digestive enzymes)

Table 3: Examples of correct and incorrect paraphrases extracted by our supervised method with their rank.

We call them sampled sentences. Then, using Yahoo!JAPAN API, we retrieved up to 20 snippets relevant to each sampled sentence using all of the nouns in each sentence as a query. After that, each snippet was split into sentences, which we call snippet sentences. We paired a sampled sentence and a snippet sentence that was the most similar to the sampled sentence. Similarity is the number of nouns shared by the two sentences. Finally, we randomly sampled 100,000 pairs from all the pairs.

Paraphrase pairs were extracted from the Web sentence pairs by using BM, SMT, Mrt and the supervised and unsupervised versions of our method. The features used with our methods were selected from all of the 78 features mentioned in Section 3.2 so that they performed well for Web sentence pairs. Specifically, the features were selected by ablation tests using training data that was tailored to Web sentence pairs. The training data consisted of 2,741 sentence pairs that were collected in the same way as the Web sentence pairs and was labeled in the same way as described in Section 3.2.

Graph (c) of Figure 3 shows precision curves. We also measured precision without trivial pairs in the same way as the previous experiment. Graph (d) shows the results. The lower half of Table 2 shows the number of extracted paraphrases with/without trivial pairs for each method.

Note that precision figures of our methods in graphs (c) and (d) are lower than those of our methods in graphs (a) and (b). Additionally, none of the methods achieved a precision rate of 90% using Web sentence pairs.[12] We think that a precision rate of at least 90% would be necessary if you apply automatically extracted paraphrases to NLP tasks without manual annotation. Only the combination of *Sup* and definition sentence pairs achieved that precision.

Also note that, for all of the methods, the numbers of extracted paraphrases from Web sentence pairs are fewer than those from definition sentence pairs.

From all of these results, we conclude that our claim I is verified.

[12]Precision of SMT is unexpectedly good. We found some Web sentence pairs consisting of two mostly identical sentences on rare occasions. The method worked relatively well for them.

5 Conclusion

We proposed a method of extracting paraphrases from definition sentences on the Web. From the experimental results, we conclude that the following two claims of this paper are verified.

1. Definition sentences on the Web are a treasure trove of paraphrase knowledge.

2. Our method extracts many paraphrases from the definition sentences on the Web accurately; it can extract about 300,000 paraphrases from 6×10^8 Web documents with a precision rate of about 94%.

Our future work is threefold. First, we will release extracted paraphrases from all of the 29,661,812 definition sentence pairs that we acquired, after human annotators check their validity. The result will be available through the ALAGIN forum.[13]

Second, we plan to induce paraphrase *rules* from paraphrase *instances*. Though our method can extract a variety of paraphrase *instances* on a large scale, their coverage might be insufficient for real NLP applications since some paraphrase phenomena are highly productive. Therefore, we need paraphrase *rules* in addition to paraphrase *instances*. Barzilay and McKeown (2001) induced simple POS-based paraphrase rules from paraphrase instances, which can be a good starting point.

Finally, as mentioned in Section 1, the work in this paper is only the beginning of our research on paraphrase extraction. We are trying to extract far more paraphrases from a set of sentences fulfilling the same *pragmatic* function (e.g. definition) for the same topic (e.g. osteoporosis) on the Web. Such functions other than definition may include the usage of the same Linux command, the recipe for the same cuisine, or the description of related work on the same research issue.

Acknowledgments

We would like to thank Atsushi Fujita, Francis Bond, and all of the members of the Information Analysis Laboratory, Universal Communication Research Institute at NICT.

[13]http://alagin.jp/

References

Susumu Akamine, Daisuke Kawahara, Yoshikiyo Kato, Tetsuji Nakagawa, Yutaka I. Leon-Suematsu, Takuya Kawada, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara. 2010. Organizing information on the web to support user judgments on information credibility. In *Proceedings of 2010 4th International Universal Communication Symposium Proceedings (IUCS 2010)*, pages 122–129.

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pages 597–604.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL 2003*, pages 16–23.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the ACL joint with the 10th Meeting of the European Chapter of the ACL (ACL/EACL 2001)*, pages 50–57.

Rahul Bhagat, Patrick Pantel, and Eduard Hovy. 2007. Ledir: An unsupervised algorithm for learning directionality of inference rules. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP2007)*, pages 161–170.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006)*, pages 17–24.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics (COLING 2004)*, pages 350–356.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Atsushi Fujii and Tetsuya Ishikawa. 2002. Extraction and organization of encyclopedic knowledge information using the World Wide Web (written in Japanese). *Institute of Electronics, Information, and Communication Engineers*, J85-D-II(2):300–307.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 107–114.

Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Stijn De Saeger, Masaki Murata, and Jun'ichi Kazama. 2009. Large-scale verb entailment acquisition from the web. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 1172–1181.

Lidija Iordanskaja, Richard Kittredge, and Alain Polguère. 1991. Lexical selection and paraphrase in a meaning-text generation model. In Cécile L. Paris, William R. Swartout, and William C. Mann, editors, *Natural language generation in artificial intelligence and computational linguistics*, pages 293–312. Kluwer Academic Press.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006)*, pages 455–462.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 698–707, June.

Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 407–415.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 177–180.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the 2008*

Conference on Empirical Methods in Natural Language Processing (EMNLP-2008), pages 802–811.

Nitin Madnani and Bonnie Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3).

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the 2nd international conference on Human Language Technology Research*, pages 280–285.

Masaki Murata, Toshiyuki Kanemaru, and Hitoshi Isahara. 2004. Automatic paraphrase acquisition based on matching of definition sentences in plural dictionaries (written in Japanese). *Journal of Natural Language Processing*, 11(5):135–149.

Roberto Navigli and Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1318–1327.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 142–149.

Deepak Ravichandran and Eduard H. Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 41–47.

Satoshi Sekine. 2005. Automatic paraphrase discovery based on context and keywords between ne pairs. In *Proceedings of the Third International Workshop on Paraphrasing (IWP-2005)*, pages 80–87.

Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of the 2nd international Conference on Human Language Technology Research (HLT2002)*, pages 313–318.

Idan Szpektor and Ido Dagan. 2008. Learning entailment rules for unary template. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING2008)*, pages 849–856.

Idan Szpektor, Eyal Shnarch, and Ido Dagan. 2007. Instance-based evaluation of entailment rule acquisition. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 456–463.