Computing Paraphrasability between Syntactic Variants of Predicate Phrases

Atsushi Fujita† and Satoshi Sato†

In a broad range of natural language processing tasks, large-scale knowledge-base of paraphrases is anticipated to improve their performance. The key issue in creating such a resource is to establish a practical method of computing semantic equivalence and syntactic substitutability, i.e., paraphrasability, between given pair of expressions. This paper addresses the issues of computing paraphrasability, focusing on syntactic variants of predicate phrases.

Our model estimates paraphrasability based on traditional distributional similarity measures, where the Web snippets are used to overcome the data sparseness problem in handling predicate phrases. Several feature sets are evaluated through empirical experiments.

Keywords: paraphrase, syntactic variants, predicate phrase, distributional similarity.

1. Introduction

One of the common characteristics of human languages is that the same concept can be expressed by various linguistic expressions. Such linguistic variations are called paraphrases. Handling paraphrases is one of the key issues in a broad range of natural language processing (NLP) tasks. In information retrieval, information extraction, and question answering, technology of recognizing if or not the given pair of expressions are paraphrases is desired to gain a higher coverage. On the other hand, a system which generates paraphrases for given expressions is useful for text-transcoding tasks, such as machine translation and summarization, as well as beneficial to human, for instance, in text-to-speech, text simplification, and writing assistance.

Among various types of paraphrase phenomena, lexical paraphrases, e.g., (1)\(^1\), and morpho-syntactic paraphrases, e.g., (2)\(^1\) have drawn interests of many researchers.

\[ a. \text{burst into tears} \iff \text{cried} \]
\[ a. \text{comfort} \iff \text{console} \]

Interestingly, those paraphrases have been represented, in (3)\(^1\) and (4)\(^1\).

\[ a. \text{be in our favor} \iff \text{favorable for us} \]
\[ a. \text{be favorable for us} \iff \text{be favorable for us} \]

(3) a. \(X\) wrote \(Y\) \iff \(X\) is the author of \(Y\)

b. \(X\) solves \(Y\) \iff \(X\) deals with \(Y\)

The weakness of these templates is that they should be applied only in some contexts. In other words, the lack
of applicability conditions for slot fillers may lead incorrect paraphrases. One way to specify the applicability condition is to enumerate correct slot fillers. For example, Pantel et al.\textsuperscript{19} have harvested instances for the given paraphrase templates based on the co-occurrence statistics of slot fillers and lexicalized part of templates (e.g. “deal with” in (3b)). Yet, there is no method which assesses semantic equivalence and syntactic substitutability of resultant pairs of expressions.

In this paper, we propose a method of directly computing semantic equivalence and syntactic substitutability, i.e., paraphrasability, particularly focusing on automatically generated morpho-syntactic paraphrases (henceforth, syntactic variants) of predicate phrases. While previous studies have mainly targeted at words or canned phrases, we treat predicate phrases having a bit more complex structures.

This paper addresses two issues in handling phrases. The first is feature engineering. Generally speaking, phrases appear less frequently than single words. This implies that we can obtain only a small amount of information about phrases. To overcome the data sparseness problem, we investigate if the Web snippet can be used as a dense corpus for given phrases. The second is the measurement of paraphrasability. We assess how well the traditional distributional similarity measures approximate the paraphrasability of predicate phrases.

2. Related work

2.1 Representation of paraphrases

Several types of morpho-syntactic paraphrases, such as passivization and nominalization, have been represented with some grammar formalisms, such as transformational generative grammar\textsuperscript{13} and synchronous tree adjoining grammar\textsuperscript{6}. These grammars, however, lack the information of applicability conditions.

Word association within phrases has been an attractive topic. Meaning-Text Theory (MTT)\textsuperscript{17} is a framework which takes into account several types of lexical dependencies in handling paraphrases. A bottleneck of MTT is that a huge amount of lexical knowledge is required to represent various relationships between lexical items. Jacquemin\textsuperscript{14} has represented the syntagmatic and paradigmatic correspondences between paraphrases with context-free transformation rules and morphological and/or semantic relations between lexical items, targeting at syntactic variants of technical terms that are typically noun phrases consisting of more than one word. We have proposed a framework of generating syntactic variants of predicate phrases\textsuperscript{7}. Following the previous work, we have been developing three sorts of resources for Japanese.

2.2 Acquiring paraphrase rules

Since the late 1990’s, the task of automatic acquisition of paraphrase rules has drawn the attention of an increasing number of researchers. Although most of the proposed methods do not explicitly eliminate morpho- syntactic paraphrases, their output tends to be lexical paraphrase.

Previous approaches to this task are two-fold. The first group espouses the distributional hypothesis\textsuperscript{12}. Among a number of models based on this hypothesis, two algorithms are referred to as the state-of-the-art. DIRT\textsuperscript{16} collects paraphrase rules consisting of a pair of paths between two nominal slots based on point-wise mutual information. TEASE\textsuperscript{21} discovers binary relation templates from the Web based on sets of representative entities for given binary relation templates. These systems often output directional rules such as exemplified in (5).

(5) a. $X$ is charged by $Y$ \quad $\Rightarrow$ $Y$ announced the arrest of $X$

b. $X$ prevent $Y$ \quad $\Rightarrow$ $X$ lower the risk of $Y$

They are actually called inference/entailment rules, and paraphrase is defined as bidirectional inference/entailment relation\textsuperscript{11}. While the similarity score in DIRT is symmetric for given pair of paths, the algorithm of TEASE considers the direction.

The other utilizes a sort of parallel texts, such as multiple translation of the same text\textsuperscript{21,18}, corresponding articles from multiple news sources\textsuperscript{3,5}, and bilingual corpus\textsuperscript{1,23}. This approach is, however, limited by the difficulty of obtaining parallel/comparable corpora.

2.3 Acquiring paraphrase instances

As reviewed in Section 1, paraphrase rules generate incorrect paraphrases, because their applicability conditions are not specified. To avoid the drawback, several linguistic clues, such as fine-grained classification of named entities and coordinated sentences, have been utilized\textsuperscript{20,23}. Although these clues restrict phenomena to those appearing in particular domain or those describing coordinated events, they have enabled us to collect paraphrases accurately. The notion of Inferential Selectional Preference (ISP) has been introduced by Pantel et al.\textsuperscript{19}. ISP can capture more general phenomena than above two; however, it lacks abilities to distinguish antonym relations.

2.4 Computing semantic equivalence

Semantic equivalence between given pair of expressions has so far been estimated under the distributional

\*1 See http://nlp.cs.nyu.edu/WTEP/
phrasability of phrases, i.e., the sparseness of features for words. This raises a crucial problem in computing paraphrasability, taking Japanese as the target language.

Step 3. Compute their paraphrasability as distributional similarity between their features.

Step 2. Extract features for each phrase.

Step 1. Retrieve Web snippets for each phrase.

Most previous studies on acquiring paraphrase rules have evaluated resultant pairs from only the second viewpoint, i.e., semantic equivalence. Additionally, we assume that one of a pair \( t \) of syntactic variants is automatically generated from the other \( s \). Thus, grammaticality of \( t \) should also be assessed. We also take into account the syntactic substitutability, because head-words of syntactic variants sometimes have different syntactic categories.

Given a pair of predicate phrases, we compute their paraphrasability in the following procedure (see also Figure 1):

1. **Step 1.** Retrieve Web snippets for each phrase.
2. **Step 2.** Extract features for each phrase.
3. **Step 3.** Compute their paraphrasability as distributional similarity between their features.

The rest of this section elaborates on each step in turn, taking Japanese as the target language.

3.1 Retrieving Web snippets

In general, phrases appear less frequently than single words. This raises a crucial problem in computing paraphrasability of phrases, i.e., the sparseness of features for given phrases. One possible way to overcome the problem is to take back-off statistics assuming the independence between constituent words\(^{19,23}\). This approach, however, has a risk of involving noises due to ambiguity of words.

We take another approach, which utilizes the Web as a dense source of examples instead of a limited size of corpus. For each of the source and target phrases, we retrieve snippets via the Yahoo API\(^+\). The maximum number of snippets is set to 500.

3.2 Extracting features

The second step extracts the features for each phrase from Web snippets. We have some options for feature set, feature weighting, and snippet collection.

3.2.1 Feature sets

To assess a given pair of phrases against the definition of paraphrasability, the following three sets of features are examined.

**HITS:** A phrase must appear in the Web if it is grammatical. The more frequently a phrase appears, the more likely it is grammatical.

**BOW:** A pair of phrases are likely to be semantically similar, if the distributions of words surrounding the phrases are similar.

**MOD:** A pair of phrases are likely to be substitutable with each other, if they share a number of instances of modifiers and modifies.

To extract BOW features from sentences including the given phrase within Web snippets, a morphological analyzer MeCab\(^-2\) was firstly used. However, it excessively labels out-of-vocabulary words including symbol sequences as “deverbal noun,” which is wrong in most cases, and thus makes features noisy. Therefore, ChaSen\(^-3\) is finally employed.

To collect MOD features, a dependency parser CaboCha\(^-4\) is used. Figure 2 depicts an example of extracting MOD features from a sentence within Web snip-
3.2.2 Feature weighting

Geffet and Dagan\(^9\) have reported on that the better quality of feature vector (weighting function) leads better results. So far, several weighting functions have been proposed, such as point-wise mutual information\(^8\) and Relative Feature Focus\(^8\). While these functions compute weights using a small corpus for merely re-ranking samples, we are developing a measure that assesses the par-phasability of arbitrary pair of phrases, where a more robust weighting function is necessary. Therefore we directly use frequencies of features within Web snippets as weight. Normalization will be done when the par-phasability is computed (Section 3.3).

3.2.3 Source-focused feature extraction

Independent collection of Web snippets for each phrase of a given pair might yield no intersection of feature sets even if they have the same meaning. To obtain more reliable feature sets, we retrieve Web snippets by querying the phrase AND the anchor of the source phrase. The “anchored version” of Web snippets is retrieved in the following steps:

Step 2-1. Determine the anchor using Web snippets for the given source phrase. We regarded a noun which most frequently modifies the source phrase as its anchor. Examples of source phrases and their anchors are shown in (6).

Step 2-2. Retrieve Web snippets by querying the anchor for the source phrase AND each of source and target phrases, respectively.

Step 2-3. Extract features for HITS, BOW, MOD. Those sets are referred to as Anc*, while the normal versions are referred to as Nor*.

(6) a. “笑みを浮かべる” . . . “面带
overlap by dribbling . . . side)
(make a good start . . . good sign)

3.3 Computing paraphrasability

Paraphrasability is finally computed by two conventional distributional similarity measures. The first is the measure proposed by Lin and Panell\(^6\):

\[
Par_{Lin}(s\Rightarrow t) = \frac{\sum_{f \in F_s \cap F_t} (w(s, f) + w(t, f))}{\sum_{f \in F_s} w(s, f) + \sum_{f \in F_t} w(t, f)},
\]

where \(F_s\) and \(F_t\) denote feature sets for \(s\) and \(t\), respectively. \(w(x, f)\) stands for the weight (frequency in our experiment) of \(f\) in \(F_s\).

While \(Par_{Lin}\) is symmetric, it has been argued that it is important to determine the direction of paraphrase\(^6\). As an asymmetric measure, we examine \(\alpha\)-skew divergence defined by the following equation\(^13\):

\[
d_{skew}(t, s) = D (P_t || \alpha P_t + (1 - \alpha) P_s),
\]

where \(P_s\) denotes a probability distribution estimated\(^1\) from a feature set \(F_s\). How well \(P_t\) approximates \(P_s\) is calculated based on the KL divergence, \(D\). The parameter \(\alpha\) is set to 0.99, following tradition, because the optimization of \(\alpha\) is difficult. To take consistent measurements, we define the paraphrasability score \(Par_{skew}\) as follows:

\[
Par_{skew}(s\Rightarrow t) = \exp (-d_{skew}(t, s)).
\]

\(^1\) We estimate them simply using maximum likelihood estimation, i.e.,

\[
P_s(f) = w(x, f) / \sum_{f' \in F_s} w(x, f').
\]
Now *Par* falls within [0, 1], and a larger *Par* indicates a more paraphrasable pair of phrases.

4. Experimental setting

We conduct empirical experiments to evaluate the proposed methods. Settings are described below.

4.1 Test collection

First, source phrases were sampled from a 15 years of newspaper articles (Mainichi 1991-2005, approximately 1.5GB). Referring to the dependency structure given by CaboCha, we extracted most frequent 1,000+ phrases for each of 6 phrase types. These phrases were then fed to a system proposed in our preceding paper to generate syntactic variants. The numbers of the source phrases and their syntactic variants are summarized in Table 1, where the numbers in the parentheses indicate that of source phrases paraphrased. At least one candidate was generated for 962 source phrases (distributional similarity measures), in total 15 models are compared. We sampled following three sets of pairs for evaluation:

### Table 1

<table>
<thead>
<tr>
<th>Phrase type</th>
<th>src</th>
<th><em>Nor</em>.HITS</th>
<th>src</th>
<th><em>Nor</em>.BOW</th>
<th>src</th>
<th><em>Anc</em>.HITS</th>
<th>src</th>
<th><em>Anc</em>.BOW</th>
<th>src</th>
<th><em>Anc</em>.MOD</th>
<th>src</th>
<th>Mainichi</th>
</tr>
</thead>
<tbody>
<tr>
<td>N : C : V</td>
<td>3,796,351</td>
<td>2,013,682</td>
<td>107,105</td>
<td>63</td>
<td>88,040</td>
<td>966</td>
<td>91.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N : C : V1 : V2</td>
<td>325,964</td>
<td>213,923</td>
<td>15,102</td>
<td>12.9</td>
<td>75,344</td>
<td>982</td>
<td>76.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N : C : Adv : V</td>
<td>1,209,265</td>
<td>923,475</td>
<td>11,097</td>
<td>9.8</td>
<td>8,281</td>
<td>523</td>
<td>15.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adv : N : C : V</td>
<td>378,617</td>
<td>233,952</td>
<td>20,104</td>
<td>14.1</td>
<td>128</td>
<td>50</td>
<td>2.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N : C : Adv</td>
<td>788,038</td>
<td>203,845</td>
<td>86,100</td>
<td>31.4</td>
<td>3,212</td>
<td>992</td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>26,982,276</td>
<td>7,912,633</td>
<td>6,190</td>
<td>176,341</td>
<td>4,002</td>
<td>44.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Output Ave.**

<table>
<thead>
<tr>
<th>Phrase type</th>
<th>src</th>
<th><em>Nor</em>.HITS</th>
<th>src</th>
<th><em>Nor</em>.BOW</th>
<th>src</th>
<th><em>Anc</em>.HITS</th>
<th>src</th>
<th><em>Anc</em>.BOW</th>
<th>src</th>
<th><em>Anc</em>.MOD</th>
<th>src</th>
<th>Mainichi</th>
</tr>
</thead>
<tbody>
<tr>
<td>N : C : V</td>
<td>1,405</td>
<td>489</td>
<td>2.9</td>
<td>1,402</td>
<td>488</td>
<td>2.9</td>
<td>1,396</td>
<td>488</td>
<td>2.9</td>
<td>1,368</td>
<td>487</td>
<td>2.8</td>
</tr>
<tr>
<td>N : C : V1 : V2</td>
<td>9,544</td>
<td>964</td>
<td>9.9</td>
<td>9,249</td>
<td>922</td>
<td>9.9</td>
<td>8,652</td>
<td>921</td>
<td>9.4</td>
<td>7,437</td>
<td>897</td>
<td>8.3</td>
</tr>
<tr>
<td>N : C : Adv : V</td>
<td>3,769</td>
<td>876</td>
<td>4.3</td>
<td>3,406</td>
<td>774</td>
<td>4.4</td>
<td>3,109</td>
<td>762</td>
<td>4.1</td>
<td>2,517</td>
<td>697</td>
<td>3.6</td>
</tr>
<tr>
<td>Adv : N : C : V</td>
<td>690</td>
<td>359</td>
<td>1.9</td>
<td>506</td>
<td>247</td>
<td>2.0</td>
<td>475</td>
<td>233</td>
<td>2.0</td>
<td>342</td>
<td>174</td>
<td>2.0</td>
</tr>
<tr>
<td>N : C : Adv</td>
<td>45</td>
<td>20</td>
<td>2.3</td>
<td>45</td>
<td>20</td>
<td>2.3</td>
<td>42</td>
<td>17</td>
<td>2.5</td>
<td>41</td>
<td>18</td>
<td>2.3</td>
</tr>
<tr>
<td>Total</td>
<td>16,912</td>
<td>3,593</td>
<td>4.7</td>
<td>16,912</td>
<td>3,593</td>
<td>4.7</td>
<td>15,073</td>
<td>3,285</td>
<td>4.6</td>
<td>12,940</td>
<td>3,083</td>
<td>4.2</td>
</tr>
</tbody>
</table>

**Output Ave.**

4.2 Criteria of paraphrasability

To assess by human the paraphrasability discussed in Section 3, we designed the following four questions based on those proposed by Szpektor et al.

**Qsc:** Is *s* a correct phrase in Japanese?

**Qc:** Is *t* a correct phrase in Japanese?

**Qsc2t:** Does *t* hold if *s* holds and can *t* substituted for *s* in some context?

**Qc2s:** Does *s* hold if *t* holds and can *s* substituted for *t* in some context?

4.3 Samples for evaluation

In the experiment, the harmonic mean of the scores derived from +BOW and +MOD (referred to as +HAR) is also used as a paraphrasability measure. Since +BOW, +MOD, and +HAR have two variants (distributional similarity measures), in total 15 models are compared. We sampled following three sets of pairs for evaluation:

**Ev.Ling:** This compares paraphrasability of each phrase type using the 20-best pairs for each of 6 phrase type and 14 Web-based models.

**Ev.Gen:** This assesses how well a method gives higher scores to correct candidates using the top-ranked pairs for randomly sampled 200 source phrases for each of 15 models.

**Ev.Rec:** This assesses how well a method gives higher scores to correct candidates using the 200-best pairs for each of 15 models.

5. Experimental results

5.1 Agreement of human judge

Two human assessors separately judged all of the 1,152 syntactic variant pairs (for 962 source phrases) within the union of the three sample sets. They agreed on all four
questions for 795 (68.4%) pairs. For the 963 (83.6%) pairs that passed \( Q_e \) and \( Q_s \) in both two judges, we obtained reasonable agreement ratios 86.9% and 85.0% and substantial Kappa values 0.697 and 0.655 for assessing \( Q_{2s} \) and \( Q_{2e} \), respectively.

5.2 Ev.Gen

Table 3 shows the results for Ev.Gen, where the strict precision is calculated based on the number of two positive judges for \( Q_{2s} \), while the lenient precision is for at least one positive judge for the same question. +MOD and +HAR outperformed the other models, although there was no statistically significant difference\(^\ast\). Significant differences between Mainichi and the other models in lenient precisions indicate that the Web enables us to compute paraphrasability more accurately than a limited size of corpus.

From a closer look at the distributions of paraphrasability scores of +BOW and +MOD shown in Table 4, we find that if a top-ranked candidate for a given phrase is assigned enough high score, it is very likely to be correct. The scores of Anc.+ are distributed in a wider range than those of Nor.+ preserving precision. This allows us to easily skim the most reliable portion by setting a threshold.

5.3 Ev.Rec

The results for Ev.Rec, as summarized in Table 5, show the significant differences of performances between Mainichi or +HITS and the other models. The results of +HITS supported the importance of comparing features of phrases. On the other hand, +BOW performed as well as +MOD and +HAR. This sounds nice because BOW features can be extracted extremely quickly and accurately than MOD features.

Unfortunately, Anc.+ led only a small impact on strict precisions. We speculate that the selection of the anchor is inadequate. Another possible interpretation is that source phrases are rarely ambiguous, because they contain at least two content words. In paraphrase generation, capturing the correct boundary of phrases is rather vital, because the source phrase is usually assumed to be grammatical. \( Q_s \) for 55 syntactic variants (for 44 source phrases) were actually judged incorrect.

The lenient precisions, which were reaching a ceiling, implied the limitation of the proposed methods. Most common errors among the proposed methods were generated by a transformation pattern \( N_1 : N_2 : C : V \Rightarrow N_2 : C : V \). Typically, dropping a nominal element \( N_1 \) of the given nominal compound \( N_1 : N_2 \) generalizes the meaning that the compound conveys, and thus results

\( \ast \) \( p < 0.05 \) in 2-sample test for equality of proportions.
Table 5: Precision for 200 candidates (Ex.Rec).

<table>
<thead>
<tr>
<th>Model</th>
<th>Strict</th>
<th>Lenient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nor.</td>
<td>Anc.</td>
<td>Nor.</td>
</tr>
<tr>
<td>Manichi</td>
<td>78 (79%)</td>
<td>111 (56%)</td>
</tr>
<tr>
<td>HITS</td>
<td>71 (36%)</td>
<td>93 (47%)</td>
</tr>
<tr>
<td>BOW/Lin</td>
<td>78 (89%)</td>
<td>162 (81%)</td>
</tr>
<tr>
<td>MOD/Lin</td>
<td>58 (79%)</td>
<td>192 (96%)</td>
</tr>
<tr>
<td>MOD/kwik</td>
<td>58 (79%)</td>
<td>193 (97%)</td>
</tr>
<tr>
<td>HAB/Lin</td>
<td>58 (79%)</td>
<td>191 (96%)</td>
</tr>
<tr>
<td>HAB/kwik</td>
<td>58 (79%)</td>
<td>194 (97%)</td>
</tr>
</tbody>
</table>

Table 6: Precision for each phrase type (Ev.Rec).

<table>
<thead>
<tr>
<th>Phrase type</th>
<th>Strict</th>
<th>Lenient</th>
</tr>
</thead>
<tbody>
<tr>
<td>N : C : V</td>
<td>51/72 (71%)</td>
<td>64/72 (89%)</td>
</tr>
<tr>
<td>N : V : N</td>
<td>42/86 (49%)</td>
<td>60/86 (70%)</td>
</tr>
<tr>
<td>N : C : Adv.</td>
<td>33/61 (54%)</td>
<td>44/61 (72%)</td>
</tr>
<tr>
<td>AAdv : N : C : V</td>
<td>0/25 (0%)</td>
<td>4/25 (16%)</td>
</tr>
<tr>
<td>Total</td>
<td>196/615 (47%)</td>
<td>270/615 (65%)</td>
</tr>
</tbody>
</table>

Table 7: # of features.

<table>
<thead>
<tr>
<th>Phrase type</th>
<th>Nor.BOW</th>
<th>Nor.MOD</th>
<th>Anc.BOW</th>
<th>Anc.MOD</th>
</tr>
</thead>
<tbody>
<tr>
<td># features (type)</td>
<td>75,848</td>
<td>471,720</td>
<td>72,109</td>
<td>495,379</td>
</tr>
<tr>
<td>Average features (type)</td>
<td>1,332</td>
<td>231</td>
<td>1,277</td>
<td>202</td>
</tr>
<tr>
<td>Average features (tokens)</td>
<td>4,883</td>
<td>391</td>
<td>4,728</td>
<td>383</td>
</tr>
</tbody>
</table>

(9) was thus judged incorrect.
(10) s. “貴:か強い” (having high quality)
(11) s. “映画:を:見:るる” (finishing seeing the movie)

6. Discussion

As described in the previous sections, our quite naive methods have shown fairly good performances in this first trial. This section describes some remaining issues to be discussed further.

The aim of this study is to create a thesaurus of phrases to recognize and generate phrases that are semantically equivalent and syntactically substitutable, following the spirit described in our preceding paper. Through the comparisons of Nor. and Anc., we have obtained a little evidence that the ambiguity of phrases was not problematic at least for handling syntactic variants, arguing the necessity of detecting the appropriate phrase boundaries.

To overcome the data sparseness problem, Web snippets are harnessed. Features extracted from the snippets outperformed newspaper corpus; however, the small numbers of features for phrases shown in Table 7 and the lack of sophisticated weighting function suggest that the problem might persist. To examine the proposed features and measures further, we plan to use TSUBAKI, an indexed Web corpus developed for NLP research, because it allows us to obtain snippets as much as it archives.

The use of larger number of snippets increases the computation time for assessing paraphrasability. For reducing it as well as gaining a higher coverage, the enhancement of the paraphrase generation system is necessary. A look at the syntactic variants automatically generated by a system, which we proposed, showed that the system could generate syntactic variants for only a half portion of the input, producing many erroneous ones (Section 4.1). To prune a multitude of incorrect candidates, statistical language models such as proposed by Habash will be incorporated. In parallel, we plan to develop a paraphrase generation system which lets us to quit from the labor of maintaining patterns such as shown in (4). We think a more unrestricted generation algorithm will gain a higher coverage, preserving the meaning as far as handling syntactic variants of predicate phrases.

7. Conclusion

In this paper, we proposed a method of assessing paraphrasability between automatically generated syntactic variants of predicate phrases. Web snippets were utilized to overcome the data sparseness problem, and the conventional distributional similarity measures were employed to quantify the similarity of feature sets for the given pair of phrases. Empirical experiments revealed that features extracted from the Web snippets contribute to the task, showing promising results, while no significant difference was observed between two measures.

In future, we plan to address several issues such as those described in Section 6. Particularly, at present, the coverage and portability are of our interests.

Acknowledgments

This work was supported in part by MEXT Grant-in-Aid for Young Scientists (B) 18700143, and for Scientific Research (A) 16200009, Japan.

References

2) Barzilay, R. and McKeown, K.R.: Extracting paraphrases from a parallel corpus, Proceedings of the

*1 http://tsubaki(ixnlp.nii.ac.jp/se/index.cgi