Computing Paraphrasability between Syntactic Variants of Predicate Phrases

Atsushi Fujita[†1] and Satoshi Sato[†1]

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

述語句統語的異形間の言い換えらしさの計算手法

藤 田 篤[†1] 佐 藤 理 史[†1]

近年，言い換えに関する知識を自動的に獲得する試みが多く報告されているが，そのような言い換え知識が保証するのは特定の文脈における同義性に過ぎない．このため，獲得した知識を用いる際に，言い換えの関係にない表現を誤って認識・生成してしまうことが問題となる．そこで本稿では，日本語述語句の統語的異形を対象とし，述語句とそこから自動的に生成された統語的異形の間の「言い換えらしさ」を計算する手法について検討する．ここで言う「言い換えらしさ」とは，(i) 各表現の文法的・意味的適格性，(ii) 二つの表現間の意味の同義性，(iii) 置換可能な文脈の存在性をまとめた概念を指す．提案手法では，語の類似度計算に用いられてきた分布類似度の考え方を述語句間の類似度計算に応用して，上記の「言い換えらしさ」を近似する．その際に用いる述語句の文脈的素性は，検索エンジンを通じて得られるウェブページの断片（スニペット）から抽出する．3 種類の評価実験を通じて，各述語句の文脈的素性は語を対象とする場合よりも顕著に疎であるにもかかわらず，それらの分布を直接比較することが言い換えらしさを計算する上で有効であることを確認した．

キーワード： 言い換え，統語的異形，述語句，分布類似度

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[13]. 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)[2], and morpho-syntactic paraphrases, e.g., (2)[7] have drawn interests of many researchers.

(1) a. burst into tears ⇔ cried
  b. comfort ⇔ console
(2) a. be in our favor ⇔ be favorable for us
  b. show a sharp decrease ⇔ decrease sharply

Interestingly, those paraphrases have been represented, in most research, with the similar templates, such as shown in (3)[16] and (4)[11].

(3) a. X wrote Y ⇔ X is the author of Y
  b. X solves Y ⇔ X deals with Y
(4) a. N_1 V N_2 ⇔ N_1's V-*ing* of N_2
  b. N_1 V N_2 ⇔ N_2 be V-*en* by N_1

The weakness of these templates is that they should be applied only in some contexts. In other words, the lack

†1 Graduate School of Engineering, Nagoya University
名古屋大学大学院工学研究科

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.*[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[11] and synchronous tree adjoining grammar[6]. These grammars, however, lack the information of applicability conditions.

Word association within phrases has been an attractive topic. Meaning-Text Theory (MTT)[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[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[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[12]. Among a number of models based on this hypothesis, two algorithms are referred to as the state-of-the-art. DIRT[16] collects paraphrase rules consisting of a pair of paths between two nominal slots based on point-wise mutual information. TEASE[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

$\Rightarrow Y$ announced the arrest of X

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

They are actually called inference/entailment rules, and paraphrase is defined as bidirectional inference/entailment relation⋆1. 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[2),18], corresponding articles from multiple news sources[3,5], and bilingual corpus[1,25]. 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[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.*[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/

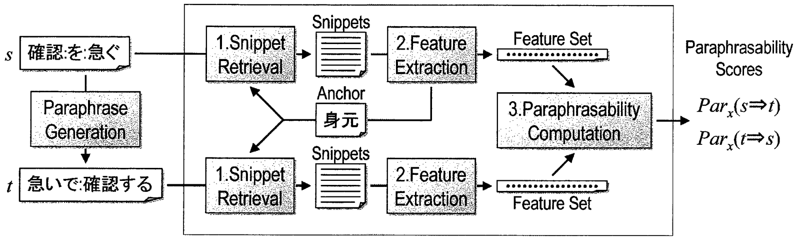

Figure1 Overview of the proposed method.

hypothesis[12]. Geffet and Dagan[9] have extended it to the distributional inclusion hypothesis for recognizing the direction of lexical entailment. Weeds *et al.*[24], on the other hand, have pointed out the limitations of lexical similarity and syntactic transformation, and have proposed to directly compute the distributional similarity of pair of subparses based on the distributions of their modifiers and parents. We think it is worth examining if the Web can be used as the source for extracting features of phrases.

3. Computing paraphrasability between predicate phrases using Web snippets

We define the concept of paraphrasability as follows:

> A grammatical phrase s is paraphrasable with another phrase t, iff t satisfies the following three:
> - t is grammatical
> - t holds if s holds
> - t is substitutable for s in some context

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):

Step 1. Retrieve Web snippets for each phrase.

Step 2. Extract features for each phrase.

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*1. 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 modifiees.

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-

---

⋆1 http://developer.yahoo.co.jp/search/
⋆2 http://mecab.sourceforge.net/
⋆3 http://chasen.naist.jp/hiki/ChaSen/
⋆4 http://chasen.org/~taku/software/cabocha/

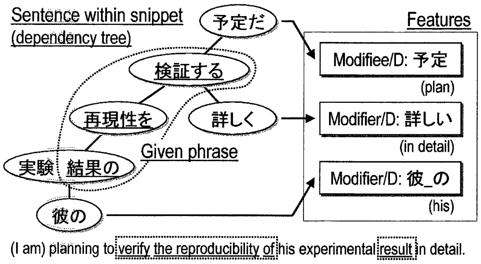

Figure2 An example of MOD feature extraction. An oval in the dependency tree denotes a *bunsetsu*.

pet. A feature is generated from a *bunsetsu*, the Japanese base-chunk, which is either modifier or modifiee of the given phrase. Each feature is composed of three or more elements:

i) Modifier or modifiee
ii) Dependency relation types (direct dependency, appositive, or parallel, c.f., RASP and MINIPAR)
iii) Base form of the head-word
iv) If they appear, case markers following nouns, auxiliary verb and verbal suffixes.

The last feature is employed to distinguish the subtle difference of meaning of predicate phrases, such as voice, tense, aspect, and modality.

As reviewed in Section 2.2, subject/object slot fillers of verb phrases (single verbs in most cases) have so far been used as features to acquire paraphrase rules[16),21)], where grammaticality of both phrases have been assumed. They have resulted pair of templates such as shown in (3) and (5); however, what actually quantified is a degree of paraphrasability between given phrases (e.g. "solves" and "deals with" in (3b)), but not templates. Hence, the pair of templates is not necessarily applicable depending on context (slot fillers), as empirically confirmed in the following work[19),22)]. In contrast, MOD consists of relatively general features of each phrases to compute paraphrasability between given phrases (not templates).

3.2.2 Feature weighting

Geffet and Dagan[8)] 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[16)] 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 paraphrasability 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 paraphrasability 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" ··· "満面"
(be smiling ··· from ear to ear)
b. "ドリブル:で:駆け:上がる" ··· "サイド"
(overlap by dribbling ··· side)
c. "良い:スタート:を:切る" ··· "幸先"
(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 Pantel[16)]:

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

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_x.

While Par_{Lin} is symmetric, it has been argued that it is important to determine the direction of paraphrase[4)]. As an asymmetric measure, we examine α-skew divergence defined by the following equation[15)]:

$$d_{skew}(t,s) = D\left(P_s \| \alpha P_t + (1-\alpha)P_s\right),$$

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

$$Par_{skew}(s{\Rightarrow}t) = \exp\left(-d_{skew}(t,s)\right).$$

$\star 1$ We estimate them simply using maximum likelihood estimation, i.e., $P_x(f) = w(x,f) / \sum_{f' \in F_x} w(x,f')$.

Table1 # of sampled source phrases and automatically generated syntactic variants.

Phrase type	# of tokens	# of types	th	types	Cov.(%)	Output		Ave.
$N:C:V$	20,200,041	4,323,756	1,000	1,014	10.7	1,536	(489)	3.1
$N_1:N_2:C:V$	3,796,351	2,013,682	107	1,005	6.3	88,040	(966)	91.1
$N:C:V_1:V_2$	325,964	213,923	15	1,022	12.9	75,344	(982)	76.7
$N:C:Adv:V$	1,209,265	923,475	21	1,097	3.9	8,281	(523)	15.7
$Adj:N:C:V$	378,617	233,952	20	1,049	14.1	128	(50)	2.6
$N:C:Adj$	788,038	203,845	86	1,003	31.4	3,212	(992)	3.2
Total	26,698,276	7,912,633		6,190		176,541	(4,002)	44.1

Table2 # of syntactic variants whose paraphrasability scores are computed.
Nor.HITS $\supset$ *Nor.BOW*.$*$ $\supset$ *Nor.MOD*.$*$. *Anc.HITS* $\supset$ *Anc.BOW*.$*$ $\supset$ *Anc.MOD*.$*$.
Nor.HITS $\supset$ *Anc.HITS*. *Nor.BOW*.$*$ $\supset$ *Anc.BOW*.$*$. *Nor.MOD*.$*$ $\supset$ *Anc.MOD*.$*$. X denotes the set of syntactic variants whose scores are computed based on X.

	Nor.HITS			Nor.BOW.$*$			Nor.MOD.$*$			Anc.HITS			Anc.BOW.$*$			Anc.MOD.$*$			Mainichi		
Phrase type	Output		Ave.	Output		Ave.	Output		Ave.	Output		Ave.	Output		Ave.	Output		Ave.	Output		Ave.
$N:C:V$	1,405	(489)	2.9	1,402	(488)	2.9	1,396	(488)	2.9	1,368	(488)	2.8	1,366	(487)	2.8	1,360	(487)	2.8	1,103	(457)	2.4
$N_1:N_2:C:V$	9,544	(964)	9.9	9,249	(922)	10.0	8,652	(921)	9.4	7,437	(897)	8.3	7,424	(894)	8.3	6,795	(891)	7.6	3,041	(948)	3.2
$N:C:V_1:V_2$	3,769	(876)	4.3	3,406	(774)	4.4	3,109	(762)	4.1	2,517	(697)	3.6	2,497	(690)	3.6	2,258	(679)	3.3	1,156	(548)	2.1
$N:C:Adv:V$	690	(359)	1.9	506	(247)	2.0	475	(233)	2.0	342	(174)	2.0	339	(173)	2.0	322	(168)	1.9	215	(167)	1.3
$Adj:N:C:V$	45	(20)	2.3	45	(20)	2.3	42	(17)	2.5	41	(18)	2.3	41	(18)	2.3	39	(16)	2.4	14	(7)	2.0
$N:C:Adj$	1,459	(885)	1.6	1,459	(885)	1.6	1,399	(864)	1.6	1,235	(809)	1.5	1,235	(809)	1.5	1,161	(779)	1.5	559	(459)	1.2
Total	16,912	(3,593)	4.7	16,067	(3,336)	4.8	15,073	(3,285)	4.6	12,940	(3,083)	4.2	12,902	(3,071)	4.2	11,935	(3,020)	4.0	6,088	(2,586)	2.4

Now Par_x falls within $[0, 1]$, and a larger Par_x 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[7] 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 4,002 (64.7%) phrases. Although the system generates numerous syntactic variants from a given phrase, most of them are erroneous. For example, among 159 syntactic variants that are automatically generated for the phrase "損害:賠償:を:求める" (demand compensation for damages), only 8 phrases are grammatical, and only 5 out of 8 are correct paraphrases.

Paraphrasability of each pair of source phrase and candidate is then computed by the methods proposed in Section 3. Table 2 summarizes the numbers of pairs whose features can be extracted from the Web snippets. While more than 90% of candidates were discarded due to 'No hits' in the Web, at least one candidate survived for 3,020 (48.8%) phrases. Mainichi is a baseline which counts HITS in the corpus used for sampling source phrases.

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.*[22]:

Q_{sc}: Is s a correct phrase in Japanese?

Q_{tc}: Is t a correct phrase in Japanese?

Q_{s2t}: Does t hold if s holds and can t substituted for s in some context?

Q_{t2s}: 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.Gen: This investigates how well a correct candidate is ranked first among candidates for a given phrase 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.

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.

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_{sc} and Q_{tc} 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_{s2t} and Q_{t2s}, 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_{s2t}, 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[*1]. 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_{sc} 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

$\star 1$ $p < 0.05$ in 2-sample test for equality of proportions.

Table3 Precision for 200 candidates (Ev.Gen).

	Strict		Lenient	
Model	Nor.$*$	Anc.$*$	Nor.$*$	Anc.$*$
Mainichi	77 (39%)	- -	101 (51%)	- -
HITS	84 (42%)	83 (42%)	120 (60%)	119 (60%)
BOW.Lin	82 (41%)	85 (43%)	123 (62%)	124 (62%)
BOW.skew	86 (43%)	87 (44%)	125 (63%)	124 (62%)
MOD.Lin	91 (46%)	91 (46%)	130 (65%)	131 (66%)
MOD.skew	92 (46%)	90 (45%)	132 (66%)	130 (65%)
HAR.Lin	90 (45%)	90 (45%)	129 (65%)	130 (65%)
HAR.skew	93 (47%)	90 (45%)	134 (67%)	131 (66%)

Table4 Distribution of paraphrasability scores and lenient precision (Ev.Gen).

	Nor.BOW		Anc.BOW	
$Par(s{\Rightarrow}t)$	Lin	skew	Lin	skew
0.9-1.0	11/ 12 (92%)	0/ 0 -	17/ 18 (94%)	2/ 2 (100%)
0.8-1.0	45/ 49 (92%)	1/ 1 (100%)	45/ 50 (90%)	6/ 6 (100%)
0.7-1.0	72/ 88 (82%)	7/ 7 (100%)	73/ 92 (79%)	10/ 11 (91%)
0.6-1.0	94/127 (74%)	11/ 11 (100%)	83/113 (74%)	12/ 13 (92%)
0.5-1.0	102/145 (70%)	13/ 13 (100%)	96/128 (75%)	14/ 15 (93%)
0.4-1.0	107/158 (68%)	13/ 14 (93%)	103/145 (71%)	21/ 22 (96%)
0.3-1.0	113/173 (65%)	25/ 26 (96%)	114/166 (69%)	31/ 32 (97%)
0.2-1.0	119/184 (65%)	40/ 41 (98%)	121/186 (65%)	49/ 50 (98%)
0.1-1.0	123/198 (62%)	74/ 86 (86%)	124/200 (62%)	82/ 99 (83%)
0.0-1.0	123/200 (62%)	125/200 (63%)	124/200 (62%)	124/200 (62%)
Variance	0.052	0.031	0.061	0.044

	Nor.MOD		Anc.MOD	
$Par(s{\Rightarrow}t)$	Lin	skew	Lin	skew
0.9-1.0	2/ 2 (100%)	0/ 0 -	7/ 7 (100%)	1/ 1 (100%)
0.8-1.0	10/ 10 (100%)	0/ 0 -	12/ 13 (92%)	2/ 2 (100%)
0.7-1.0	13/ 14 (93%)	0/ 0 -	17/ 18 (94%)	6/ 6 (100%)
0.6-1.0	20/ 21 (95%)	1/ 1 (100%)	27/ 28 (96%)	9/ 9 (100%)
0.5-1.0	31/ 32 (97%)	6/ 6 (100%)	36/ 37 (97%)	10/ 10 (100%)
0.4-1.0	42/ 44 (96%)	11/ 11 (100%)	51/ 53 (96%)	12/ 12 (100%)
0.3-1.0	61/ 68 (90%)	12/ 12 (100%)	61/ 68 (90%)	13/ 14 (93%)
0.2-1.0	81/ 92 (88%)	13/ 13 (100%)	82/ 94 (87%)	18/ 19 (95%)
0.1-1.0	105/133 (79%)	17/ 18 (94%)	104/126 (83%)	24/ 25 (96%)
0.0-1.0	130/200 (65%)	132/200 (66%)	131/200 (66%)	130/200 (65%)
Variance	0.057	0.014	0.072	0.030

correct paraphrases. However, it caused errors in some cases; for example, since N_1 was the semantic head in (7), dropping it caused an error.

(7) *s*. "出血:多量:で:死亡する"
(die due to heavy blood loss)
t. *"多量:で:死亡する" (die due to plenty)

5.4 Ev.Ling

Finally the results for Ev.Ling is shown in Table 6. Paraphrasability of syntactic variants for phrases containing an adjective was poorly computed. The primal source of errors for $Adj : N : C : V$ type phrases was the subtle change of nuance by switching syntactic heads as illustrated in (8), where underlines indicate heads.

(8) *s*. "良い:仕事:を:する" (do a good job)
t_1.$^{\neq}$"よく:仕事する" (work hard)
t_2.$^{\neq}$"仕事:を:よく:する" (improve the work)

Most errors in paraphrasing $N : C : Adj$ type phrases, on the other hand, were caused due to the difference of aspectual property and agentivity between adjectives and verbs. For example, (9*s*) can describe not only things those qualities have been improved as inferred by (9*t*), but also those originally having a high quality. Q_{s2t} for

Table5 Precision for 200 candidates (Ev.Rec).

	Strict		Lenient	
Model	Nor.*	Anc.*	Nor.*	Anc.*
Mainichi	78 (39%)	- -	111 (56%)	- -
HITS	71 (36%)	93 (47%)	113 (57%)	128 (64%)
BOW.Lin	159 (80%)	162 (81%)	193 (97%)	191 (96%)
BOW.skew	154 (77%)	158 (79%)	192 (96%)	191 (96%)
MOD.Lin	158 (79%)	164 (82%)	192 (96%)	193 (97%)
MOD.skew	156 (78%)	161 (81%)	191 (96%)	191 (96%)
HAR.Lin	157 (79%)	164 (82%)	192 (96%)	194 (97%)
HAR.skew	155 (78%)	160 (80%)	191 (96%)	191 (96%)

Table6 Precision for each phrase type (Ev.Ling).

Phrase type	Strict	Lenient
$N:C:V$	52/ 98 (53%)	69/ 98 (70%)
$N_1:N_2:C:V$	51/ 72 (71%)	64/ 72 (89%)
$N:C:V_1:V_2$	42/ 86 (49%)	60/ 86 (70%)
$N:C:Adv:V$	33/ 61 (54%)	44/ 61 (72%)
$Adj:N:C:V$	0/ 25 (0%)	4/ 25 (16%)
$N:C:Adj$	18/ 73 (25%)	38/ 73 (52%)
Total	196/415 (47%)	279/415 (67%)

Table7 # of features.

	Nor.BOW	Nor.MOD	Anc.BOW	Anc.MOD
# of features (type)	73,848	471,720	72,109	409,379
Average features (type)	1,322	211	1,277	202
Average features (token)	4,883	391	4,728	383

(9) was thus judged incorrect.

(9) s. "質:が:高い" (having high quality)
t.≠"質:が:高まる" (quality rises)

Precisions of syntactic variants for the other types of phrases were higher, but they tended to include trivial paraphrases such as shown in (10) and (11). Yet, collecting paraphrase instances statically will contribute to paraphrase recognition tasks.

(10) s. "承認:を:得る" (clear)
t. "承認:さ:れる" (be approved)

(11) s. "映画:を:見:終わる" (finish seeing the movie)
t. "映画:が:終わる" (the movie ends)

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[7]. 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[*1], 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[10] 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.

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

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

1) Bannard, C. and Callison-Burch, C.: Paraphrasing with bilingual parallel corpora, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.597–604 (2005).
2) Barzilay, R. and McKeown, K.R.: Extracting paraphrases from a parallel corpus, *Proceedings of the*

39th Annual Meeting of the Association for Computational Linguistics (*ACL*), pp.50–57 (2001).
3) Barzilay, R. and Lee, L.: Learning to paraphrase: an unsupervised approach using multiple-sequence alignment, *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp.16–23 (2003).
4) Bhagat, R., Pantel, P. and Hovy, E.: LEDIR: an unsupervised algorithm for learning directionality of inference rules, *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (*EMNLP-CoNLL*), pp.161–170 (2007).
5) Dolan, B., Quirk, C. and Brockett, C.: Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources, *Proceedings of the 20th International Conference on Computational Linguistics* (*COLING*), pp.350–356 (2004).
6) Dras, M.: Tree adjoining grammar and the reluctant paraphrasing of text, PhD Thesis, Division of Information and Communication Science, Macquarie University (1999).
7) Fujita, A., Kato, S., Kato, N. and Sato, S.: A compositional approach toward dynamic phrasal thesaurus, *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (*WTEP*), pp.151–158 (2007).
8) Geffet, M. and Dagan, I.: Feature vector quality and distributional similarity, *Proceedings of the 20th International Conference on Computational Linguistics* (*COLING*), pp.247–253 (2004).
9) Geffet, M. and Dagan, I.: The distributional inclusion hypotheses and lexical entailment, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 107–116 (2005).
10) Habash, N.: The use of a structural N-gram language model in generation-heavy hybrid machine translation, *Proceedings of the 3rd International Natural Language Generation Conference* (*INLG*), pp.61–69 (2004).
11) Harris, Z.: Co-occurrence and transformation in linguistic structure, *Language*, Vol. 33, No. 3, pp. 283–340 (1957).
12) Harris, Z.: *Mathematical structures of language*, John Wiley & Sons (1968).
13) 乾健太郎, 藤田　篤：言い換え技術に関する研究動向, 自然言語処理, Vol.11, No.5, pp.151–198 (2004).
14) Jacquemin, C.: Syntagmatic and Paradigmatic Representations of Term Variation, *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.341–348 (1999).
15) Lee, L.: Measures of distributional similarity, *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.25–32 (1999).
16) Lin, D. and Pantel, P.: Discovery of inference rules for question answering, *Natural Language Engineering*, Vol.7, No.4, pp.343–360 (2001).
17) Mel'čuk, I. and Polguère, A.: A formal lexicon in meaning-text theory (or how to do lexica with words), *Computational Linguistics*, Vol.13, No.3-4, pp.261–275 (1987).
18) Pang, B., Knight, K. and Marcu, D.: Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences, *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp.102–109 (2003).
19) Pantel, P., Bhagat, R., Coppola, B., Chklovski, T. and Hovy, E.: ISP: Learning Inferential Selectional Preferences, *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pp.564–571 (2007).
20) Sekine, S.: Automatic paraphrase discovery based on context and keywords between NE pairs, *Proceedings of the 3rd International Workshop on Paraphrasing* (*IWP*), pp.80–87 (2005).
21) Szpektor, I., Tanev, H., Dagan, I. and Coppola, B.: Scaling Web-based acquisition of entailment relations, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.41–48 (2004).
22) Szpektor, I., Shnarch, E. and Dagan, I.: Instance-based evaluation of entailment rule acquisition, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.456–463 (2007).
23) Torisawa, K.: Acquiring inference rules with temporal constraints by using Japanese coordinated sentences and noun-verb co-occurrences, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp.57–64 (2006).
24) Weeds, J., Weir, D. and Keller, B.: The distributional similarity of sub-parses, *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pp.7–12 (2005).
25) Wu, H. and Zhou, M.: Synonymous collocation extraction using translation information, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 120–127 (2003).