Measuring the Appropriateness of Automatically Generated Phrasal Paraphrases

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

The most critical issue in generating and recognizing paraphrases is developing a wide-coverage paraphrase knowledge base. To attain the coverage of paraphrases that should not necessarily be represented at surface level, researchers have attempted to represent them with general transformation patterns. However, this approach does not prevent spurious paraphrases because there is no practical method to assess whether or not each instance of those patterns properly represents a pair of paraphrases. This paper argues on the measurement of the appropriateness of such automatically generated paraphrases, particularly targeting at morpho-syntactic paraphrases of predicate phrases. We first specify the criteria that a pair of expressions must satisfy to be regarded as paraphrases. On the basis of the criteria, we then examine several measures for quantifying the appropriateness of a given pair of expressions as paraphrases of each other. In addition to existing measures, a probabilistic model consisting of two distinct components is examined. The first component of the probabilistic model is a structured N-gram language model that quantifies the grammaticality of automatically generated expressions. The second component approximates the semantic equivalence and substitutability of the given pair of expressions on the basis of the distributional hypothesis. Through an empirical experiment, we found (i) the effectiveness of contextual similarity in combination with the constituent similarity of morpho-syntactic paraphrases and (ii) the versatility of the Web for representing the characteristics of predicate phrases.

Key Words: paraphrasing, appropriateness as paraphrases, morpho-syntactic paraphrase, predicate phrase, phrasal variants, semantic equivalence, substitutability, grammaticality

1 Introduction

One of the common characteristics of human languages is that a concept can be expressed with several different linguistic expressions. Handling such synonymous expressions in a given language, i.e., **paraphrases**, is one of the key issues in a broad range of natural language processing (NLP) tasks (Inui and Fujita 2004). For example, the technology for recognizing whether or not a given pair of expressions are paraphrases boosts the recall of information retrieval, information extraction, and question answering. The technology also plays an important role in

[†] School of Systems Information Science, Future University-Hakodate
[††] Graduate School of Engineering, Nagoya University

aggregating plenty of uninhibited opinions about products and services available on the Web: both the consumers and producers benefit from the summary. On the other hand, a system that generates paraphrases of a given expression is useful for text-transcoding tasks, such as machine translation and summarization. Such a system would also be extremely beneficial to people by proposing alternative expressions for writing assistance, simplifying texts for reading, and reducing homonyms for improving text-to-speech quality.

The most critical issue in generating and recognizing paraphrases is developing resources that cover a wide range of paraphrases. One way of attaining such coverage was proposed by Fujita and Inui (2006): first categorize paraphrases into several classes by the knowledge required and generality, and then separately develop resources for each class. They divided paraphrases into the following four classes.

(1) a. **Lexical paraphrases**
Emma *burst into tears* and he tried to *comfort* her.
⇔ Emma *cried* and he tried to *console* her. (Barzilay and McKeown 2001)

b. **Morpho-syntactic paraphrases**
Employment *showed a sharp decrease* in October.
⇔ Employment *decreased sharply* in October. (Iordanskaja et al. 1992)

c. **(Pure) Syntactic paraphrases**
It was his best suit that John wore to the dance last night.
⇔ John wore *his best suit* to the dance last night. (Dras 1999)

d. **Inferential paraphrases**
There was no chance it would endanger our planet, astronomers said.
⇔ NASA emphasized that there was never danger of a collision. (Dolan et al. 2004)

Examples (1a), (1b), and (1c) have potential to be explained on the basis of linguistic knowledge only, while some kinds of world knowledge is necessary to identify the equivalence of example (1d). Lexical and morpho-syntactic paraphrases involve changing the constituent words, and thus have more variation than syntactic paraphrases. On that account, we believe that building resources for lexical and morpho-syntactic paraphrases is essential for generating and recognizing paraphrases robustly.

With the same line of thinking, most of the previous work on generating and recognizing paraphrases has been dedicated to developing resources for these classes. Paraphrase knowledge for these classes is typically represented with pairs of expressions that satisfy the following criteria.

Criterion 1. Semantically equivalent

Criterion 2. Substitutable in some contexts

Examples of such paraphrase knowledge are shown in (2) and (3).

(2) a. comfort $\Leftrightarrow$ console
b. burst into tears $\Leftrightarrow$ cried (Barzilay and McKeown 2001)

(3) a. show a sharp decrease $\Leftrightarrow$ decrease sharply
b. be in our favor $\Leftrightarrow$ be favorable to us (Fujita et al. 2007)

The pairs of expressions in (2) exemplify atomic knowledge for lexical paraphrases. As they cannot be generalized into patterns, a huge amount of fully lexicalized paraphrase knowledge should be stored statically to generate and recognize this class of paraphrases[1]. On the other hand, morpho-syntactic paraphrases, such as verb alternation, nominalization, and paraphrasing of light-verb construction, exhibit a degree of generality as shown in (3). It is therefore reasonable to represent them with a set of general transformation patterns such as those shown in (4)[2].

(4) a. show a X_A Y_N $\Leftrightarrow$ $v(Y_N)$ $adv(X_A)$
show a sharp decrease decrease sharply
b. be in X_N Y_N $\Leftrightarrow$ be $adj(Y_N)$ to $obj(X_N)$
is in our favor is favorable to us

The generalization enables us to attain higher coverage, keeping the knowledge manageable.

Various methods have been proposed to acquire paraphrase knowledge (see Section 2.2) where pairs of existing expressions are collected from the given corpus, taking the above two criteria into account. However, another issue arises when paraphrase knowledge is generated from the patterns for morpho-syntactic paraphrases, such as shown in (4), by instantiating variables with specific words. For example, neither of the following instances of pattern (4a) is appropriate.

(5) a. (statistics) show a gradual decline (of something) $\Leftrightarrow^{\neq}$ (statistics) decline gradually
b. (the data) show a specific distribution $\Leftrightarrow^{*}$ (the data) distribute specifically

The two phrases in (5a) are not equivalent ($\neq$), and the right-hand phrase of (5b) is not even grammatical, as indicated by the asterisk ($*$). As exhibited by these examples, excessive generalization produces an enormous number of spurious paraphrases. To avoid this problem in addition to criteria 1 and 2, the following criterion should be adopted.

Criterion 3. Both expressions are grammatical

We introduce the notion of "**appropriateness as paraphrases**" as the degree to which a pair of expressions satisfies the aforementioned three criteria, and examine several measures for quantifying it. While recent studies have tended to collect fully lexicalized paraphrase knowledge

[1] Paraphrases of idiomatic and literal phrases, such as "kick the bucket" $\Leftrightarrow$ "die," should also be included in this class of paraphrases (Fujita and Inui 2006).

[2] X and Y each denote a variable. The subscript of a variable denotes its part-of-speech. $v(\cdot)$, $adj(\cdot)$, $adv(\cdot)$, and $obj(\cdot)$ are functions that return verb, adjective, adverb, and objective case of the given word, respectively.

as shown in (2) and (3), we focus on morpho-syntactic paraphrases generated from transformation patterns such as those shown in (4). In particular, we deal with morpho-syntactic paraphrases of predicate phrases (henceforth, **phrasal variants**) in Japanese, such as follows[3].

(6) a. 確認-を 急ぐ ⇔ 急いで 確認する
checking-ACC to hurry ⇔ in a hurry to check
to hurry checking (something) ⇔ to check (something) in a hurry

b. 部屋-が 暖かく-なる ⇔ 部屋-が 暖まる
room-NOM be warm-to become ⇔ room-NOM to warm
the room becomes warm ⇔ the room becomes warm

c. 再現性-の 検証-を 行う ⇔ 再現-できる-かどうか-を 検証する
reproducibility-GEN verification-ACC to do ⇔ to reproduce-able-whether-ACC to verify
to conduct a verification of the reproducibility ⇔ to verify whether it is reproducible

“Predicate phrase” in this paper refers to a sub-parse governed by a predicate, e.g., a verb and an adjective, and thus has a structure that is bit more complex than words, word sequences, and paths in dependency parses. Furthermore, syntactic heads of phrasal variants sometimes belong to different syntactic categories as the above examples exhibit.

In this paper, we examine several measures to quantify the appropriateness of given automatically generated pair of predicate phrases. Our first challenge in this paper is to investigate the applicability of the distributional hypothesis (Harris 1968) to the computation of semantic equivalence and substitutability between predicate phrases. Another challenge is the data sparseness problem. Generally speaking, phrases appear less frequently than words. This implies that we may obtain only a small amount of contextual information for a given phrase. We therefore investigate the availability of the Web as a corpus. In addition to apply existing measures that are built upon the distributional hypothesis, we propose and examine an novel probabilistic model consisting of two distinct components. The first component of our probabilistic model quantifies the grammaticality, i.e., criterion 3, of each of the given phrases using a structured N-gram language model. The second component approximates the semantic equivalence and substitutability of the given pair of phrases, i.e., criteria 1 and 2, on the basis of the distributional hypothesis. Through an empirical experiment, we clarify the possibility and effectiveness of explicitly assessing the grammaticality of the given pair of phrases.

In the next section, we review the literature on developing paraphrase knowledge and the characteristics and drawbacks of the existing measures that have been used for acquiring paraphrase knowledge. The proposed probabilistic model is then presented in Section 3, where the

[3] Abbreviations: ACC (accusative case), COP (copula), DAT (dative case), GEN (genitive case), NEG (negation), NOM (nominative case), PAR (particle), PASS (passive), PUNC (punctuation mark), and TOP (topic).

grammaticality and similarity factors are derived from a conditional probability. The setting of empirical evaluation is described in Section 4, and then the results and analyses are shown in Sections 5 and 6. Following an error analysis in Section 7, Section 8 summarizes this paper.

2 Related work

2.1 Representation of paraphrase knowledge

Paraphrase knowledge for certain classes of paraphrases can be represented with a set of general transformation patterns. This makes the knowledge manageable and attains higher coverage of paraphrases. Following the transformation grammar (Harris 1957), many researchers have attempted to develop transformation patterns (Mel'čuk and Polguère 1987; Dras 1999; Jacquemin 1999; Fujita et al. 2007).

Lexical derivation has so far been the central topic in dealing with phrasal variants, because it is indispensable for both generating and constraining the instantiation of general transformation patterns. Meaning-Text Theory (Mel'čuk and Polguère 1987) is one such framework, which incorporates several types of lexical dependencies to deal with various paraphrases. A problem inherent in this theory is that a huge amount of lexical knowledge is required to represent the more than 60 types of relationships between lexical items. Jacquemin (1999) represented the syntagmatic and paradigmatic correspondences between paraphrases with context-free transformation rules and morphological and/or semantic relations between lexical items, targeting at morpho-syntactic paraphrases of technical terms that are typically noun phrases consisting of more than one word. Following the spirit of these previous studies, Fujita et al. (2007) proposed a framework for generating phrasal variants and developed resources in Japanese: a set of general transformation patterns and dictionaries for handling lexical derivations.

However, no model in this approach is capable of preventing spurious paraphrases because there is no practical method of measuring the appropriateness of their instantiations.

2.2 Automatic paraphrase acquisition

Since the late 1990's, the task of automatically acquiring paraphrase knowledge has drawn the attention of an increasing number of researchers. They are tackling the problem, particularly focusing on accuracy, although they have tended to notice that it is hard to acquire paraphrase knowledge that ensures full coverage for various paraphrases from existing text corpora alone. To date, two streams of research have evolved: one acquires paraphrase knowledge from parallel/comparable corpora, while the other uses a regular corpus.

Several alignment techniques, which imitate those devised for machine translation, have been proposed to acquire paraphrase knowledge from parallel/comparable corpora. Various sorts of parallel/comparable corpora have been used as a source of paraphrase knowledge, such as multiple translations of the same text (Barzilay and McKeown 2001; Pang et al. 2003; Ibrahim et al. 2003), corresponding articles from multiple news sources (Shinyama et al. 2002; Barzilay and Lee 2003; Dolan et al. 2004), and bilingual corpora (Wu and Zhou 2003; Bannard and Callison-Burch 2005). Unfortunately, this approach may not cover sufficiently wide variety of paraphrases due to the difficulty of obtaining a parallel/comparable corpus that contains all phrasal variants of all predicate phrases.

In the second stream, i.e., paraphrase acquisition from a regular corpus or the Web, the distributional hypothesis (Harris 1968) has been espoused. The similarity of two expressions computed based on this hypothesis is called distributional similarity. This stream of measurement has the following three essential elements.

Representation of context: To compute the similarity, a given expression is first represented with a set of expression that co-occur with it in a given corpus. Expressions that co-occur with the given expression, such as adjacent words (Barzilay and McKeown 2001), adjacent character sequences (Yoshida et al. 2008), complements of predicates (Torisawa 2002), modifiers/modifiees (Yamamoto 2002; Weeds et al. 2005), and indirect dependencies (Hagiwara et al. 2008a) have so far been examined. For the sake of convenience, we refer to those expressions as (contextual) features.

Feature weighting: To precisely compute the similarity, the weight for each feature is adjusted. Point-wise mutual information (Lin 1998) and Relative Feature Focus (Geffet and Dagan 2004) are well-known examples of methods for determining the weights. A comparative study has been done by Hagiwara et al. (2008b).

Similarity measures: To convert two feature sets into a scalar value, several measures have been proposed, such as cosine, Lin's measure (Lin 1998), and Kullback-Leibler (KL) divergence and its variants. Some of them will be explained in Section 2.3.

While most researchers have extracted only fully lexicalized pairs of words or word sequences, two prominent algorithms have used dependency parsers for collecting template-like knowledge that contains variable slots, as shown in (7) and (8).

(7) a. X wrote $Y \Leftrightarrow X$ is the author of Y

b. X solves $Y \Leftrightarrow X$ deals with Y (Lin and Pantel 2001)

(8) a. X prevents $Y \Rightarrow X$ decreases the risk of Y

b. X goes back to $Y \Rightarrow Y$ allows X to return (Szpektor et al. 2004)

One of the algorithms, DIRT (Lin and Pantel 2001), collects pairs of paths in dependency parses that connect two nominal entities. It utilizes point-wise mutual information between paths and contextual features for weighting and filtering operative features. The similarity score in DIRT is symmetrical for the given pair of paths and thus its results are symmetric as in (7). On the other hand, TEASE algorithm (Szpektor et al. 2004) discovers dependency sub-parses that are similar to a given transitive verb from the Web using the sets of representative instances of subject and direct object of the given verb. As a result of taking the direction into account, it outputs asymmetric patterns, such as exemplified in (8). Their approach acquires various types of relations between event mentions: not only synonymous pairs of expressions, but also causal and temporal relations. The patterns that represent directional relationship between expressions are thus called inference/entailment rules. The notion of paraphrase is defined as a bidirectional inference/entailment relation.

The above knowledge falls between that in (2), which is fully lexicalized, and that in (4), which is almost fully generalized. As a way of enriching such template-like knowledge, several linguistic clues, such as fine-grained classification of named entities (Sekine 2005), have been utilized. Although these clues restrict the resultant paraphrases to those appearing in particular domains, they enable us to collect paraphrases accurately. Pantel et al. (2007) introduced the notion of Inferential Selectional Preference (ISP) and collected expressions that would fill those slots. ISP can capture broader variety of paraphrases than the above two; however, it cannot distinguish antonym relations. Bhagat et al. (2007) proposed a method of determining the direction of inference/entailment between given two templates.

As mentioned in Section 1, the aim of the studies reviewed here is to collect pairs of existing expressions likely to be paraphrases. Therefore, they need not take the grammaticality of expressions into account.

2.3 Existing measures of the appropriateness as paraphrases

The appropriateness of the given pair of expressions as paraphrases has so far been estimated on the basis of the distributional hypothesis (Harris 1968), particularly targeting at relatively short expressions, such as words, word sequences, and sub-parses. Geffet and Dagan (2005) extended it to the distributional inclusion hypothesis for recognizing the direction of lexical entailment. Weeds et al. (2005), on the other hand, pointed out the limitations of lexical similarity and syntactic transformation and proposed directly computing the distributional similarity between the given pair of sub-parses using the distributions of their modifiers and modifiees. To date, however, no model has been established that takes into account all of the three aforementioned criteria. With the ultimate aim of building an ideal model, we present an overview of the characteristics and drawbacks of the three existing measures closely related to our work, leaving a comprehensive comparison of various measures to (Weeds 2003).

2.3.1 Focusing on the intersection of contextual features

One way of implementing distributional similarity is to quantify the overlap of the feature sets of each expression. For example, Lin (1998) proposed a symmetrical measure:

$$Par_{Lin}(s \Leftrightarrow t) \quad = \quad \frac{\sum_{f \in F_s \cap F_t} \bigl(w(s,f) + w(t,f)\bigr)}{\sum_{f \in F_s} w(s,f) + \sum_{f \in F_t} w(t,f)}, \tag{1}$$

where F_s and F_t denote sets of features with positive weights for words s and t, respectively.

Although this measure has been widely cited and has so far performed the task reasonably well, its symmetry seems unnatural (see example (8)). Moreover, it may not work robustly when we deal with general predicate phrases because it is not feasible to enumerate all phrases to determine the weight of features $w(\cdot,\cdot)$.

2.3.2 Divergence-based modeling

Some researchers have modeled the distributional similarity with the divergence of probability distributions. The skew divergence, a variant of KL divergence, was proposed in (Lee 1999) on the basis of an insight: the substitutability of one word for another need not be symmetrical. The divergence is given by the following formula:

$$\begin{aligned} d_{skew}(t,s) \quad &= \quad D\bigl(P_s \| \alpha P_t + (1-\alpha) P_s\bigr), \\ D\bigl(P_1(X) \,\|\, P_2(X)\bigr) \quad &= \quad \sum_{x \in X} P_1(x) \log \frac{P_1(x)}{P_2(x)}, \end{aligned}$$

where P_s and P_t are the probability distributions of features of the given original and substituted words s and t, respectively. How accurately P_t approximates P_s is calculated on the basis of KL divergence D, where α $(0 \leq \alpha \leq 1)$ is a parameter; it expresses KL divergence when $\alpha = 1$. The divergence can be recast into a similarity score via, for example, the following function.

$$Par_{skew}(s \Rightarrow t) \quad = \quad \exp\bigl(-d_{skew}(t,s)\bigr). \tag{2}$$

This measure offers an advantage over Par_{Lin} (Equation (1)): the weight of each feature is determined on the basis of probability theory. However, the optimization of α is difficult because the optimal value varies with the task and even the data size.

2.3.3 Translation-based conditional probability

Bannard and Callison-Burch (2005) proposed a probabilistic model for acquiring paraphrases of word sequences. The likelihood of a word sequence t as a paraphrase of the given word sequence s is calculated by the following formula:

$$P(t|s) \quad = \sum_{f \in tr(s) \cap tr(t)} P(t|f)P(f|s),$$

where $tr(e)$ stands for a set of word sequences in foreign language that are aligned with e in a given bilingual corpus. Parameters $P(t|f)$ and $P(f|s)$ are estimated using the given bilingual corpus. A large-scale bilingual corpus may enable us to acquire a large amount of paraphrase knowledge accurately. However, it is not feasible to build (or obtain) a bilingual corpus in which all the instances of phrasal variants in one language are translated to at least one same expression in the other side of language.

2.4 Post-process for automatic paraphrase generation

When paraphrase knowledge, i.e., a pair of expression, is generated from general transformation patterns by filling their variables with specific words, the grammaticality of the expressions should be assessed in addition to computing their semantic equivalence and substitutability. To the best of our knowledge, no work has taken all of these criteria into account. However, apart from our goal, i.e., automatic generation of paraphrase knowledge, there have been several studies on measuring the appropriateness of applying paraphrase knowledge in a specific context: how correct a sentence that is partially/entirely paraphrased is.

Fujita et al. (2004) pointed out that case assignments of verbs tend to be incorrect in paraphrasing Japanese sentences irrespective of the class of paraphrases and applied a language model to the assessment of case assignments as a post-process for paraphrase generation. Their model, however, cannot evaluate the semantic equivalence of the resultant pairs of expressions. Quirk et al. (2004) built a paraphrase generation model from a monolingual comparable corpus on the basis of a statistical machine translation framework, where the language model was used to quantify the grammaticality of the translations, i.e., generated expressions. The translation model, however, is not suitable for generating phrasal variants, because it learns word/phrase alignments at the surface level. To cover all phrasal variants, we require a non-real comparable corpus in which all instances of phrasal variants have a chance of being aligned. Furthermore, because the translation model optimizes the word/phrase alignment at the sentence level, the substitutability of the aligned pairs of word/phrase sequences cannot be explicitly guaranteed.

3 A probabilistic model for measuring appropriateness

3.1 Formulation with conditional probability

As described in Section 1, we regard sub-parses governed by predicates as predicate phrases. Our goal is to establish a measure that computes the appropriateness of a given pair of automatically generated predicate phrases as paraphrases on the basis of the following three criteria.

Criterion 1. Both expressions are semantically equivalent

Criterion 2. Both expressions are substitutable in some contexts

Criterion 3. Both expressions are grammatical

Let s and t be the source (original) and target (paraphrased) predicate phrase, respectively. We introduce the following two assumptions.

Assumption 1. s is given and grammatical: As a way of generating all the instances of phrasal variants from the given transformation pattern, it is reasonable to instantiate one side of the pattern with existing predicate phrases to generate the other side. We adopt this approach, assuming that the existing predicate phrases are grammatical.

Assumption 2. s and t do not co-occur: Provided that s and t are paraphrases, they are members of a set of paradigmatic expressions, so do not co-occur.

On the basis of Assumption 1, the appropriateness of the given pair of expressions as paraphrases is formulated with conditional probability $P(t|s)$, as in (Bannard and Callison-Burch 2005). It is then transformed as follows on the basis of Assumption 2:

$$\begin{aligned} P(t|s) &= \sum_{f \in F} P(t|f)P(f|s) \\ &= \sum_{f \in F} \frac{P(f|t)P(t)}{P(f)} P(f|s) \\ &= P(t) \sum_{f \in F} \frac{P(f|t)P(f|s)}{P(f)}, \qquad (3) \end{aligned}$$

where F denotes a set of features. The first factor $P(t)$ is the **grammaticality factor**, which quantifies the degree to which the third criterion is satisfied. Note that we assume that the given s is grammatical. The second factor $\sum_{f \in F} \frac{P(f|t)P(f|s)}{P(f)}$, on the other hand, is the **similarity factor**, which approximates the degree to which the first and second criteria are satisfied by summing up the overlap of the features of s and t. The characteristics and advantages of this probabilistic model are summarized as follows.

- The model is asymmetric[4].
- Grammaticality is explicitly assessed by $P(t)$.
- There is no need to enumerate all the phrases. s and t are merely the given conditions.
- No heuristic is introduced. The weight of the features can be determined by conditional probabilities $P(f|t)$ and $P(f|s)$ and marginal probability $P(f)$.

The rest of this section explains our implementation of each factor in turn, taking Japanese as the target language.

3.2 Grammaticality factor

The factor $P(t)$ of Equation (3) quantifies how a predicate phrase t is grammatical using a statistical language model. Unlike English, in Japanese, predicates such as verbs and adjectives do not necessarily determine the order of their complements and adjuncts, although they have some preferences. For example, both sentences in (9) are grammatical.

(9) a.

彼-は	**グラス-を**	**いつも**	**左手-で**	持つ.
he-TOP	glass-ACC	always	left hand-with	to take-PUNC

He always holds glasses with his left hand.

b.

彼-は	**いつも**	**左手-で**	**グラス-を**	持つ.
he-TOP	always	left hand-with	glass-ACC	to take-PUNC

He always holds glasses with his left hand.

This motivates us to use structured N-gram language models (Habash 2004) for quantifying the grammaticality of predicate phrase. Given a predicate phrase t, its grammaticality $P(t)$ is estimated on the basis of its dependency structure $T(t)$, assuming a $(N-1)$-th order Markov process for generating $T(t)$:

$$P(t) \quad = \quad \left[\prod_{i=1 \ldots |T(t)|} P_d\big(c_i | d_i^1, d_i^2, \ldots, d_i^{N-1}\big) \right]^{1/|T(t)|},$$

where $|T(t)|$ stands for the number of nodes in $T(t)$. d_i^j denotes the direct ancestor node of the i-th node c_i, where j is the distance from c_i; for example, d_i^1 and d_i^2 are the parent and grandparent nodes of c_i, respectively. The normalization factor $1/|T(t)|$ is introduced for canceling out the length bias of the given phrase.

Then, a concrete definition of dependency structure is given. Widely-used Japanese dependency parsers consider a morpheme sequence consisting of at least one content morpheme followed by a sequence of function morphemes, if any, as a node called a "*bunsetsu*." The chunks

[4] Note that our formulation does not take inclusion of features into account: cf. (Yamamoto 2002; Geffet and Dagan 2005; Bhagat et al. 2007).

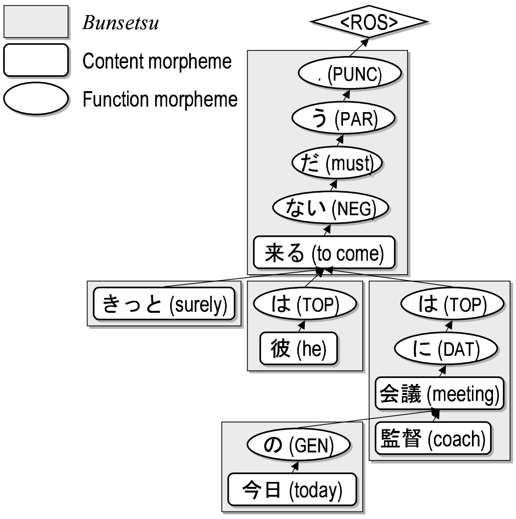

Fig. 1 MDS of the sentence in (10).

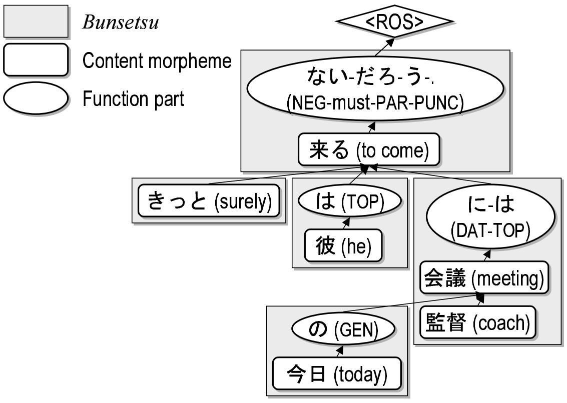

Fig. 2 PWDS of the sentence in (10).

of sentences in (9) and (10) exemplify these nodes.

(10) きっと 彼-は 今日-の 監督-会議-に-は 来-ない-だろ-う-.
surely he-TOP today-GEN coach-meeting-DAT-TOP to come-NEG-must-PAR-PUNC
He will surely not come to today's coach meeting.

As a *bunsetsu* can be quite long, involving many morphemes, to regard it as a node will make the model complex. We therefore use the following two versions of dependency structures whose nodes are smaller than a *bunsetsu*.

MDS: Morpheme-based dependency structure (Takahashi et al. 2001) regards a morpheme as a node. The MDS of the sentence in (10) is shown in Figure 1.

PWDS: MDS cannot assess the collocation between content morphemes when a number of function morphemes appear between them. Pseudo-word-based dependency structure (PWDS) with $N \geq 3$ can do it by regarding the sequence of function morphemes as a single node, in addition to MDS, as exemplified in Figure 2.

For both of the above models, dependencies between nodes are determined on the basis of the *bunsetsu* dependencies obtained by using a morphological analyzer and dependency parser.

- The rightmost node of $bunsetsu_i$ depends on the syntactic head of $bunsetsu_j$ on which $bunsetsu_i$ depends; for example, "の (GEN)" depends on "会議 (meeting)" in both Figures 1 and 2. The rightmost node of the final *bunsetsu* depends on a special node "⟨ROS⟩ (root-of-sentence)."
- Other nodes depend on succeeding nodes of the *bunsetsu*.

A conventional way of dealing with a node that has never appeared in the given corpus is to use the linear interpolation of the lower degrees of models. For example, the 3-gram conditional probability $P_d(c_i|d_i^1, d_i^2)$ is given by the following equation:

$$\begin{aligned} P_d(c_i|d_i^1, d_i^2) &= \lambda_3 P_{MLE}(c_i|d_i^1, d_i^2) + \lambda_2 P_{MLE}(c_i|d_i^1) + \lambda_1 P_{MLE}(c_i), \\ \text{s.t.} \quad \sum_j \lambda_j &= 1, \end{aligned}$$

where P_{MLE} stands for the empirical probability distribution determined by maximum likelihood estimation. The optimal mixture weights λ_j are determined via an EM algorithm using development data that are not used for estimating P_{MLE}.

3.3 Similarity factor

The similarity factor of Equation (3) computes how similar two phrases s and t are by comparing two sets of contextual features F_s of s and F_t of t. On the basis of the findings in the previous work (see Section 2.3), we use the following two types of feature sets, each of which is composed of expressions that co-occur with the given phrase p in the given corpus. Let f be a feature consisting of a tuple $\langle r, e \rangle$ of such an expression e and a relation r between p and e.

BOW: A pair of phrases is likely to be semantically equivalent if the distributions of the words surrounding the phrases are similar. The relation "`co-occur_in_the_same_sentence`" is considered as the only element of the relation set R_{BOW}.

MOD: A pair of phrases is likely to be substitutable with each other, provided they share a number of instances of modifiers and modifiees. The set of the relation R_{MOD} has two elements: "`modifier`" and "`modifiee`."

As reviewed in Section 2.2, subject/object slot fillers of verb phrases (single verbs in most cases) in English have been used as contextual features to acquire paraphrase templates, i.e., pairs of templates such as those shown in (7) and (8), where the grammaticality of their lexicalized parts have been assumed. What is actually quantified is, however, not the similarity between the whole templates, but that between their lexicalized parts and correspondences of slots. Thus, the pair of templates cannot necessarily be used as paraphrases in the given specific context (slot fillers), as empirically confirmed in their following work (Pantel et al. 2007; Szpektor et al. 2008). In contrast, the MOD features capture more unrestricted characteristics of each phrase, which enables us to compute the similarity between phrases of arbitrary sizes.

3.3.1 Web snippet retrieval

In general, phrases appear less frequently than single words. This raises a crucial problem in computing the similarity between phrases, i.e., the sparseness of contextual features. One possible way to overcome the problem is to take back-off statistics assuming the independence between constituent words (Torisawa 2006; Pantel et al. 2007). This approach, however, has a risk of involving noise due to the ambiguity of those words.

We take another approach that uses the Web as a corpus instead of a corpus of limited size. Given a phrase p, the snippets of Web pages (henceforth, Web snippets) containing p are retrieved via Yahoo! Web search API[5] by issuing quoted p as a query. The resultant Web snippets can be considered as a dense example collection of contextual information for p.

3.3.2 Feature extraction

Both BOW and MOD features of the given phrase are extracted from the sentences within the Web snippets. To collect BOW features, first, the content morphemes, i.e., nouns, verbs, adjectives, and adverbs, in the sentences that include the phrase are extracted using morphological analyzer ChaSen[6]. Each feature is then composed of the base form of such a content morpheme and its part-of-speech. Note that we exclude morphemes that are labeled as proper nouns, such as person names and location names, because we expect that they will unworthily decrease the possibility of overlap due to their much greater variation than common words.

On the other hand, to collect MOD features, we carried out structural matching between the given phrase and the sentences within the Web snippets, where the *bunsetsu*-based dependency structure is determined using ChaSen and CaboCha[7]. Figure 3 depicts an example of extracting the MOD features from a sentence that includes the given phrase. As shown in the figure, each feature is composed of the following four elements extracted from a *bunsetsu* that is either a modifier or a modifiee of the given phrase.

- Modifier or modifiee
- Relation type (Depending, Appositive, or Parallel)
- Base form of the syntactic head
- Several types of function morphemes, if any

[5] http://developer.yahoo.co.jp/search/, version 1.0. The API provides up to 1,000 Web snippets for a query.

[6] http://chasen.naist.jp/hiki/ChaSen/, version 2.3.3 with IPADIC version 2.7.0.
We first used MeCab (http://mecab.sourceforge.net/, version 0.96). However, it excessively labels out-of-vocabulary morphemes including symbol sequences as "deverbal nouns," which is wrong in most cases, and thus makes the features noisy.

[7] http://chasen.org/~taku/software/cabocha/, version 0.53.

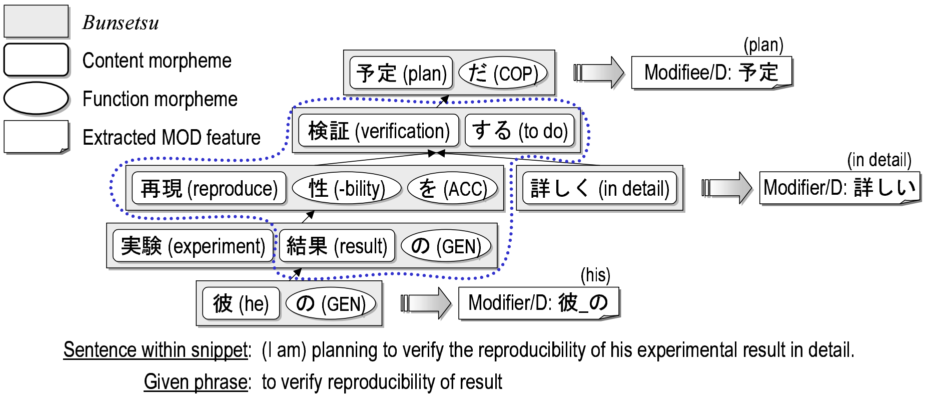

Fig. 3 An example of MOD feature extraction.

Function morphemes are incorporated to capture the subtle difference between the meanings of predicate phrases, such as voice, aspect, and modality. At present, we take into account only case markers (including genitive "の") that immediately follow nouns, and auxiliary verbs and verbal suffixes that do so for verbs and adjectives.

3.3.3 Parameter estimation

Finally, conditional probability distributions $P(f|s)$ and $P(f|t)$ are estimated. Given a phrase p, the conditional probability distribution $P(f|p)$ is determined by maximum likelihood estimation as follows, superficially assuming the mutual exclusiveness of features:

$$\begin{aligned} P(f|p) &= P(\langle r, e\rangle|p) \\ &= \frac{\mathit{freq}_{sni}(p,r,e)}{\sum_{r'\in R}\sum_{e'}\mathit{freq}_{sni}(p,r',e')}, \end{aligned}$$

where $\mathit{freq}_{sni}(p,r,e)$ stands for the frequency of an expression e that appeared with the phrase p in relation r within the Web snippets retrieved by querying p.

On the other hand, the weight for features $P(f)$ can be estimated on the basis of the following equation using a static corpus in an offline manner:

$$\begin{aligned} P(f) &= P(\langle r, e\rangle) \\ &= \frac{\sum_p \mathit{freq}_{cp}(p,r,e)}{\sum_p\sum_{r'\in R}\sum_{e'}\mathit{freq}_{cp}(p,r',e')}, \end{aligned}$$

where $\mathit{freq}_{cp}(p,r,e)$ denotes the frequency of an expression e that appeared with some expression p in relation r within the given corpus. Features are again considered to be mutually exclusive.

Table 1 Types of N-grams observed in the corpus.

$P(t)$	$N = 1$	$N = 2$	$N = 3$
MDS	320,394	10,120,469	55,356,014
PWDS	513,982	16,043,365	79,087,006

4 Experimental settings

We have conducted an experiment to assess the performance of the existing measures and the probabilistic models on the task of measuring the appropriateness of automatically generated candidates of phrasal variants. In this section, the settings of the experiment are described.

4.1 Measures examined

The probabilistic model has several options. In this experiment, we have examined all combinations of the following four options: 48 versions in total.

Implementation of $P(t)$: 3-gram language models based on MDS and PWDS were respectively built upon 15 years of Mainichi newspaper articles (1991–2005, 1.5 GB, 21 M sentences, henceforth, Mainichi) using morphological analyzer ChaSen and dependency parser CaboCha, with N being varied from 1 to 3. Table 1 shows the types of N-grams that each version of the statistical language model contains. Yomiuri newspaper articles 2005 (350 MB, 4.7 M sentences) and Asahi news paper articles 2005 (180 MB, 2.7 M sentences) were used for optimizing the mixture weights of interpolating the lower degrees of models.

The number of the Web snippets (N_S): 100, 200, 500, and 1,000

Contextual feature set: BOW, MOD, and their combination, HAR, were examined, because they are supposed to be complementary (see Section 3.3). However, they cannot be merged directly, because they are not necessarily disjoint and the frequency of each feature is used in the probabilistic framework. We therefore adopted the harmonic mean of the scores respectively derived using BOW and MOD, dealing with both scores equally.

Corpus used for estimating $P(f)$: Two different corpora were exclusively used to build two variations of $P(f)$. One was Mainichi, which was also used for building structured N-gram language models. The other was a much larger corpus consisting of 470M sentences collected from Web pages (Kawahara and Kurohashi 2006). We refer to the resultant parameter sets based on those corpora as NewsCP and WebCP, respectively. The statistics in Table 2 show a larger variation of MOD than BOW. WebCP is expected to have higher

Table 2 Types of features observed in the corpora.

$P(f)$	BOW	MOD
NewsCP	75,881	5,600,069
WebCP	90,189	44,018,226

coverage and a smoother probability distribution, although the Web corpus may contain a lot of ungrammatical expressions, and a morphological analyzer and dependency parser would output some improper instances of features.

In addition to the probabilistic models, we have examined several conceivable measures.

Count-based measures: A phrase should appear in a reasonably large corpus if it is grammatical. Count-based measures assume that the more frequently a phrase appears, the more likely it is to be grammatical. The following two implementations have been evaluated.

- **HITS-News**: The number of occurrences of t in Mainichi is regarded as the score.
- **HITS-Web**: The number of Web pages that contain t is regarded as the score. The estimated value can be retrieved via Yahoo! Web search API.

Distributional similarity measures: Two versions of distributional similarity measures have also been examined. The score, which falls within $[0, 1]$, is computed using BOW, MOD, and HAR extracted from Web snippets. We assume that the larger the value is, the more likely the pair of phrases is to be paraphrases.

- Par_{Lin} (Equation (1)): Unlike Lin and Pantel (2001), which utilized a static corpus to determine feature weights, we directly used the frequency of each feature within the Web snippets retrieved by querying each phrase.
- Par_{skew} (Equation (2)): The conditional probabilities $P(f|s)$ and $P(f|t)$ of the probabilistic model were used as the probability distributions P_s and P_t. As the value of the parameter α, we examined 0.99 only, as an approximation of KL divergence (Lee 1999).

4.2 Test collection

First, existing predicate phrases were collected from Mainichi[8]. Referring to the dependency structures derived by ChaSen and CaboCha, we extracted approximately 1,000 types of the most frequent phrases for each of the following six phrase types.

[8] The corpus was also used to build language models $P(t)$ and one version of $P(f)$. However, we think it still enables us to conduct a fair experiment, because those models do not directly evaluate the source phrase s.

Table 3 Sampled source phrases s and their paraphrase candidate pairs $\langle s, t\rangle$.

Phrase type	Collected		Sampled			Generated		
	Tokens	Types	th_{freq}	s	Cov.	s	$\langle s, t\rangle$	Yld.
N:C:V	20,200,041	4,323,756	1,000	1,014	0.107	489	1,536	3.1
N_1:N_2:C:V	3,796,351	2,013,682	107	1,005	0.063	966	88,036	91.1
N:C:V_1:V_2	325,964	213,923	15	1,022	0.129	982	75,340	76.7
N:C:Adv:V	1,209,265	923,475	21	1,097	0.039	523	8,281	15.7
Adj:N:C:V	378,617	233,952	20	1,049	0.141	50	128	2.6
N:C:Adj	788,038	203,854	86	1,003	0.314	992	3,212	3.2
Total	26,698,276	7,912,642	6,190			4,002	176,533	44.1

th_{freq}: the threshold of frequency for sampling phrases.
Cov. (Coverage): the ratio of the total tokens of phrases that the sampled s cover.
Yld. (Yield): the average number of generated phrases per paraphrased source phrase.

(11) a. N**:**C**:**V **type phrase**
確認-を 急ぐ
checking-ACC to hurry
to hurry checking (something)

b. N_1**:**N_2**:**C**:**V **type phrase**
損害-賠償-を 求める
damage-compensation-ACC to require
to demand compensation for damages

c. N**:**C**:**V_1**:**V_2 **type phrase**
統計-を 取り-始める
statistics-ACC to take-to start
to start taking the statistics

d. N**:**C**:**Adv**:**V **type phrase**
検討-を さらに 進める
consideration-ACC further to advance
to make a further consideration

e. Adj**:**N**:**C**:**V **type phrase**
早い 復興-を 祈る
rapid recovery-ACC to wish
to wish the rapid recovery

f. N**:**C**:**Adj **type phrase**
のど-が 痛い
throat-NOM be painful
to have a sore throat

Assuming that these predicate phrases are grammatical, we then fed them to a paraphrase generation system proposed in (Fujita et al. 2007). Given an input phrase, the system overgenerates candidates of its phrasal variants using a catalog of handcrafted syntactic transformation patterns and dictionaries tailored for handling lexical derivations. Henceforth, we refer to those automatically generated candidates as "paraphrase candidates" and pairs of a source phrase, s, and one of its paraphrase candidates, t, as "paraphrase candidate pairs."

Table 3 summarizes the statistics of our test collection, where the "Sampled/s" column denotes the numbers of phrase types sampled as the source, while the "Generated/s" and "Generated/$\langle s, t\rangle$" columns present those of the paraphrased source phrases and paraphrase candidate pairs, respectively. At least one paraphrase candidate was generated for 65% (4,002/6,190) of the input source phrases, although the ratio and the numbers of paraphrase candidates per

Table 4 Paraphrase candidate pairs $\langle s, t\rangle$ whose appropriateness was computed.

Phrase type	HITS-Web			BOW			MOD			HITS-News		
	s	$\langle s, t\rangle$	Yld.	s	$\langle s, t\rangle$	Yld.	s	$\langle s, t\rangle$	Yld.	s	$\langle s, t\rangle$	Yld.
N:C:V	489	1,425	2.9	489	1,425	2.9	488	1,414	2.9	457	1,103	2.4
N_1:N_2:C:V	965	10,533	10.9	930	10,279	11.1	927	9,453	10.2	948	3,040	3.2
N:C:V_1:V_2	894	4,130	4.6	793	3,744	4.7	777	3,297	4.2	548	1,155	2.1
N:C:Adv:V	378	742	2.0	267	564	2.1	256	533	2.1	167	215	1.3
Adj:N:C:V	20	49	2.5	20	49	2.5	19	46	2.4	7	14	2.0
N:C:Adj	907	1,542	1.7	906	1,541	1.7	889	1,482	1.7	459	559	1.2
Total	3,653	18,421	5.0	3,405	17,602	5.2	3,356	16,225	4.8	2,586	6,086	2.4

paraphrased source phrase ("Generated/Yld." column) were remarkably different depending on the phrase type. The system generated numerous paraphrase candidates (the maximum was 186); however, most of them were not appropriate. For example, among 159 paraphrase candidates for the phrase in (11b), only 8 phrases were grammatical, and only 5 out of 8 were appropriate paraphrases.

Finally, the appropriateness of each paraphrase candidate pair as a paraphrase is computed by each measure described in Section 4.1. As those pairs include many inappropriate ones, the task can also be illustrated as filtering them out and ranking the remaining pairs. Table 4 shows the numbers of paraphrase candidate pairs whose appropriateness can be computed[9]. The numbers were diverse depending on the features being referred to. Approximately 90% of the paraphrase candidate pairs were discarded because either s or t did not appear at all. On the other hand, at least one candidate survived for 84% (3,356/4,002) of the paraphrased source phrases and 54% (3,356/6,190) of the input source phrases. With the Web, we could compute the appropriateness score at a significantly higher rate (267%; 16,225/6,086) than with the limited size of a well-controlled corpus, i.e., Mainichi, sacrificing only 352 pairs whose scores were computed only by HITS-News.

Table 5 summarizes the statistics of the extracted BOW and MOD features, revealing that fewer MOD features are obtainable than BOW features. The rightmost two columns in the table show the numbers of phrase types among the union of sets of s and t ("Phrase" column) and paraphrase candidate pairs ("$\langle s, t\rangle$" column) for which we could retrieve at least N_s Web snippets. We could retrieve 100 Web snippets for only 6,029 paraphrase candidate pairs and 1,000 Web snippets for 3,708 pairs within our test collection, respectively.

[9] The Web snippets were retrieved from September 14 to 17, 2008.

Table 5 Retrieved contextual features of phrases.

N_S	BOW			MOD			N_S snippets	
	Types	Avg.Types	Avg.Tokens	Types	Avg.Types	Avg.Tokens	Phrase	$\langle s, t\rangle$
1,000	77,642	1,668	8,019	624,784	297	650	5,463	3,708
500	74,807	1,312	5,113	479,462	209	414	6,246	4,464
200	69,606	847	2,507	314,505	118	203	6,740	4,986
100	65,074	578	1,430	223,760	73	115	7,613	6,029

Types: the total number of feature types.

Avg.Types: the average number of feature types for phrases that gained at least one snippet.

Avg.Tokens: the average number of feature tokens for phrases that gained at least one snippet.

4.3 Criteria for human judgment

The appropriateness of a pair of expressions as paraphrases can only be judged by humans. We therefore asked assessors to answer the following four questions, which reflect the criteria described in Section 1.

Q_{sc}: Is s an acceptable Japanese phrase?

Q_{tc}: Is t an acceptable Japanese phrase?

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

Q_{t2s}: Does s always hold if t holds and can s be substituted for t in some contexts?

To reduce the amount of labor, a assessor does not answer the latter two questions if he/she answers either of the former questions "No." The following examples are sampled from pairs on which the three assessors in our experiment agreed for all of the above four questions.

(12) a. *s*. *基準-を 厳しい
criterion-ACC be severe
*(not translatable)
t. 厳しい 基準-だ
be severe criterion-COP
be a severe criterion
(No, Yes, -, -)

b. *s*. 幅-が 広い
width-NOM be wide
the width is wide
t. *幅-が 広げる
width-NOM to widen
*the width widens
(Yes, No, -, -)

c. *s*. 映画-を 見-終わる
movie-ACC to see-to finish
to finish seeing the movie
t. 映画-が 終わる
movie-NOM to end
the movie ends
(Yes, Yes, Yes, No)

d. *s*. 承認-を 得る
approval to gain
to clear
t. 承認-さ-れる
to approve-PASS
to be approved
(Yes, Yes, Yes, Yes)

Incidentally, while some previous studies have attempted to collect knowledge even for plausible inferences, such as shown in (13), our criteria regard them as inappropriate.

Table 6 Paraphrase candidate pairs $\langle s, t\rangle$ for the 200 sampled source phrases.

Phrase type	All	Sampled			HITS-Web			BOW			MOD			HITS-News		
	s	s	$\langle s,t\rangle$	Yld.	s	$\langle s,t\rangle$	Yld.	s	$\langle s,t\rangle$	Yld.	s	$\langle s,t\rangle$	Yld.	s	$\langle s,t\rangle$	Yld.
N:C:V	489	18	57	3.2	18	55	3.1	18	55	3.1	18	55	3.1	18	41	2.3
N_1:N_2:C:V	966	57	4,596	80.6	57	610	10.7	57	610	10.7	57	567	9.9	57	182	3.2
N:C:V_1:V_2	982	54	4,767	88.3	54	276	5.1	54	276	5.1	54	246	4.6	33	71	2.2
N:C:Adv:V	523	16	51	3.2	16	39	2.4	16	39	2.4	16	39	2.4	13	15	1.2
Adj:N:C:V	50	2	8	4.0	2	6	3.0	2	6	3.0	2	6	3.0	1	3	3.0
N:C:Adj	992	53	173	3.3	53	91	1.7	53	91	1.7	53	86	1.6	34	44	1.3
Total	4,002	200	9,652	48.3	200	1,040	5.4	200	1,040	5.4	200	999	5.0	156	356	2.3

(13) a. X marries $Y \Rightarrow X$ dates Y (One may marry without dating) (Pantel et al. 2007)
b. X eats $Y \Rightarrow X$ likes Y (One may eat what he/she dislike) (Bhagat et al. 2007)

5 Input-wise evaluation

Paraphrase candidates for a given predicate phrase are ranked by each measure. This section describes how accurately each measure can rank an appropriate candidate first for each source phrase. To perform this evaluation, we randomly sampled 200 source phrases and extracted all of their paraphrase candidates. Table 6 shows the statistics of sampled data, where the "Sampled/Yld." column denotes that there is still a considerable diversity with regard to the numbers of paraphrase candidates per source phrase.

The rest of this section is devoted to answering the following questions.

Q$_1$: Which measure performs the task best in practice?

Q$_2$: Which contextual features are superior to the others?

- Which feature set performs the task better?
- What happens when a larger number of Web snippets are used?

Q$_3$: What the options of the probabilistic model lead to?

- Which language model is superior to the others?
- Which corpus for estimating $P(f)$ results in better performance?

5.1 Sampling and judgment

For each measure, the top-ranked paraphrase candidate pair for each of 200 source phrases

Table 7 Agreement of human judgment ($n = 469$).

	Agr.	$\mathrm{Q_{sc}} \wedge \mathrm{Q_{tc}}$	κ of $\mathrm{Q_{s2t}}$	κ of $\mathrm{Q_{t2s}}$
Judge A–Judge B	297 (0.633)	309 (0.659)	0.693	0.622
Judge A–Judge C	305 (0.650)	283 (0.603)	0.683	0.620
Judge B–Judge C	319 (0.680)	314 (0.670)	0.753	0.683

Table 8 Appropriate paraphrases among the top-ranked candidates: summary.

Measure	Grammatical t			Appropriate $\langle s, t\rangle$		
	1 judge	2 judges	3 judges	1 judge	2 judges	3 judges
MDS, NewsCP, N_S = 1,000, HAR	182	159	$115^{\star\star}$	$122^{\star\star}$	$87^{\star\star}$	$61^{\star\star}$
MDS, WebCP, N_S = 1,000, HAR	$181^{\star}$	159	$120^{\star}$	$117^{\star\star}$	$86^{\star\star}$	$63^{\star\star}$
PWDS, NewsCP, N_S = 1,000, HAR	182	159	$114^{\star\star}$	$123^{\star\star}$	$87^{\star\star}$	$62^{\star\star}$
PWDS, WebCP, N_S = 1,000, HAR	$181^{\star}$	159	$118^{\star\star}$	$121^{\star\star}$	$88^{\star\star}$	$65^{\star\star}$
HITS-News	$150^{\star\star}$	$140^{\star\star}$	$115^{\star\star}$	$103^{\star\star}$	$83^{\star\star}$	$64^{\star\star}$
HITS-Web	**187**	**166**	**131**	$123^{\star\star}$	$91^{\star\star}$	$69^{\star}$
Par_{Lin}, N_S = 1,000, HAR	185	164	$123^{\star}$	135	102	76
Par_{skew}, N_S = 1,000, HAR	184	164	$122^{\star\star}$	**138**	**105**	**78**

was first selected. Although we had 14,756 pairs for 74 measures in total[10], the union of resultant sets consisted of only 469 paraphrase candidate pairs. Then, three human assessors separately judged each paraphrase candidate pair by answering the four questions shown in Section 4.3.

The inter-assessor agreement of each pair of assessors is summarized in Table 7, where the "Agr." column denotes the number of paraphrase candidate pairs for which two assessors agreed on all of the four questions. On the other hand, Kappa values were calculated on the basis of paraphrase candidate pairs that passed both $\mathrm{Q_{sc}}$ and $\mathrm{Q_{tc}}$. As shown in the table, the assessors agreed on all of the four questions for 63 to 68% of paraphrase candidate pairs. We also obtained substantial κ-values: 0.68 to 0.75 and 0.62 to 0.68 for judging $\mathrm{Q_{s2t}}$ and $\mathrm{Q_{t2s}}$, respectively.

As the appropriateness of a paraphrase candidate pair is estimated with the assumption that the source phrase s is given, the rest of this section does not refer to the answers for $\mathrm{Q_{t2s}}$.

5.2 Overview of the results

Although we have evaluated all 74 measures, we show the results of selected combinations of parameters in Table 8 as a brief summary of the performance. "$\star\star$" and "$\star$" in Table 8 denote the significance levels $p < 0.01$ and $p < 0.05$ of McNemar's test between the best measure and

[10] HITS-News did not select any paraphrase candidates for 44 source phrases, because none of their paraphrase candidates appeared in the corpus (see Table 6).

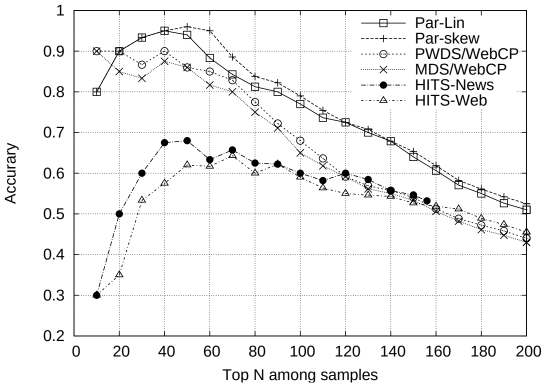

Fig. 4 Comparison of measures (2 judges).

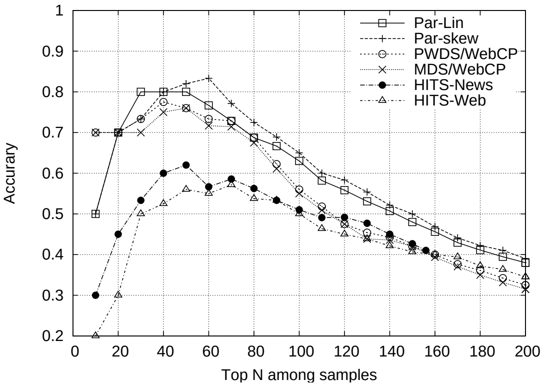

Fig. 5 Comparison of measures (3 judges).

each measure, respectively. HITS-Web selected the grammatical paraphrase candidate best, while Par_{skew} with $N_S = 1{,}000$ and HAR performed the best for the entire task.

As paraphrase candidates were generated from a given source phrase using a set of handcrafted transformation patterns, their constituent words were similar to those of their source phrase to some degree. Thus, the results of the count-based measures were tolerably high, considering that they ranked paraphrase candidates without referring to their source phrases. HITS-Web had a better coverage than HITS-News as a result of harnessing the Web as a corpus.

Par_{Lin} and Par_{skew} do not explicitly evaluate the grammaticality but only compare the feature sets of two phrases; nevertheless, contrary to our expectation, they performed the best among all the measures. As a result of taking into account the similarity of contextual features in addition to that of constituent words, these measures performed significantly better than the count-based measures. Distributional similarity measures were also superior to the probabilistic model for assessing the grammaticality of paraphrase candidates. This implies that the grammaticality of a paraphrase candidate could be assessed as a side effect of querying the Web to extract its contextual features. We expect that this technique will work well as long as we deal with relatively short phrases.

The aim in introducing the probabilistic model is to explicitly assess the grammaticality and to combine it with the similarity measurement. No combination of parameters, however, outperformed the existing measures, Par_{Lin} and Par_{skew}, which only assessed similarity, nor even one of the count-based measures, HITS-Web. Furthermore, the accuracies of all versions of the probabilistic model were significantly worse than that of the best measure.

Figures 4 and 5 give a closer look at the correlation between the score and accuracy. The

Table 9 Appropriate paraphrases among the top-ranked candidates: distributional similarity measures.

Measure	N_S	1 judge			2 judges			3 judges		
		BOW	MOD	HAR	BOW	MOD	HAR	BOW	MOD	HAR
Par_{Lin}	1,000	$126^{\star\star}$	136	$\underline{135}$	$90^{\star\star}$	$101^{\star}$	102	$66^{\star\star}$	75	76
	500	$127^{\star}$	138	**139**	$90^{\star\star}$	$101^{\star}$	102	$65^{\star\star}$	$74^{\star}$	75
	200	$125^{\star\star}$	135	137	$89^{\star\star}$	101	102	$63^{\star\star}$	$71^{\star}$	$72^{\star}$
	100	$127^{\star}$	$127^{\star}$	$128^{\star}$	$91^{\star\star}$	$95^{\star\star}$	$\underline{94}^{\star\star}$	$63^{\star\star}$	$67^{\star\star}$	$\underline{66}^{\star\star}$
Par_{skew}	1,000	$128^{\star\star}$	137	138	$97^{\star\star}$	104	**105**	$70^{\star\star}$	**78**	**78**
	500	$126^{\star\star}$	137	138	$96^{\star\star}$	103	104	$69^{\star\star}$	75	75
	200	$125^{\star\star}$	136	137	$92^{\star\star}$	101	102	$65^{\star\star}$	73	73
	100	$125^{\star\star}$	$128^{\star}$	$130^{\star}$	$91^{\star\star}$	$94^{\star\star}$	$94^{\star\star}$	$64^{\star\star}$	$67^{\star\star}$	$67^{\star\star}$

horizontal and vertical axes denote the number of paraphrase candidate pairs with the highest scores among 200 samples and the cumulative accuracy of those pairs, respectively. These graphs imply that if a top-ranked paraphrase candidate for a given phrase is assigned a high enough score, it is more likely to be appropriate. However, some inappropriate paraphrase candidate pairs can accidentally have high scores. Typical errors will be exemplified in Section 7.

5.3 Investigation into contextual features

Before a detailed evaluation of the probabilistic model, we investigate the use of contextual features retrieved from the Web snippets. Table 9 summarizes the results of distributional similarity measures, where "$\star\star$" and "$\star$" denote the significance levels of McNemar's test on the performance compared to that of the best measure, i.e., Par_{skew} with $N_S = 1{,}000$ and HAR.

5.3.1 The type of contextual features

In Section 3.3, we mentioned that BOW and MOD gauge the similarity of phrases from different viewpoints: semantic equivalence and substitutability. Yet, we expect that MOD also contributes to quantifying the semantic equivalence, and thus it is superior to BOW. The combination of BOW and MOD into HAR is introduced to further enhance the performance.

The underlined numbers in Table 9 indicate the results of MOD performed worse than BOW, and those of HAR that did so for either BOW or MOD. These results confirmed that MOD and HAR were superior to BOW for Par_{Lin} and Par_{skew}. However, the impact of introducing HAR seems small. Nevertheless, we consider that HAR must be practically important in terms of robustness, because MOD features are relatively sparser than BOW features as we discussed referring to Table 5. This will be discussed in more detail in Section 5.4.2.

5.3.2 The number of Web snippets

We observed that a larger number of Web snippets led to better results. The extent of the improvement, however, became smaller when more than 200 Web snippets were used. This is because general predicate phrases were in fact less frequent, and thus we can retrieve only a small number of Web snippets and sparse contextual features for each phrase.

As it is time-consuming to obtain a large number of Web snippets, the trade-off between the number of Web snippets and the performance should be investigated further. However, as Kilgarriff (2007) has pointed out, the quality of Web snippets and what appears at the top of search results that commercial search engines return will vary according to several factors other than linguistic ones[11]. To examine the proposed features and measures further in an unbiased situation, TSUBAKI[12], a search engine developed for NLP research, may be useful, because it statically archives a huge number of Web pages written in Japanese, and allows us to obtain all the Web pages in the archive that correspond to the query.

5.4 Characteristics of the probabilistic model

Table 10 summarizes the experimental results of all versions of the probabilistic model. The underlined numbers in the table again indicate the results of MOD performed worse than BOW, and those of HAR that did so for either BOW or MOD. The marks "○" and "●" indicate the winners of comparisons of each pair of measures that share all the parameters except the model of $P(t)$ and the corpus for $P(f)$, respectively. As described in Section 5.2, no combination of parameters could achieve comparable performance to the existing measures (cf. Table 9). To clarify the reason, we analyze the characteristics of the probabilistic model by evaluating the utility of each option and their combinations in turn.

5.4.1 Grammaticality factor

Contrary to our expectation, comparisons between MDS and PWDS revealed that PWDS do not always produce better results than MDS: "○" is marked 30 times on PWDS and 16 times on MDS. To find the reason, we examined how accurately each language model was able to select grammatical phrases. We first extracted the paraphrase candidates that were given the highest probability $P(t)$ among candidates for each of 200 source phrases, and then we asked the assessors to judge only their grammaticality, $\mathrm{Q_{tc}}$. Table 11 shows the numbers of grammatical

[11] We found that the rankings of paraphrase candidates changed slightly depending on when the Web snippets were retrieved.

[12] http://tsubaki.ixnlp.nii.ac.jp/se/index.cgi

Table 10 Appropriate paraphrases among the top-ranked candidates: probabilistic model.

$P(t)$	$P(f)$	N_S	1 judge			2 judges			3 judges		
			BOW	MOD	HAR	BOW	MOD	HAR	BOW	MOD	HAR
MDS	NewsCP	1,000	111	$122^{\bullet}_{\circ}$	$122^{\bullet}$	81	84	$87^{\bullet}$	$56_{\circ}$	59	61
		500	$116^{\bullet}$	$121^{\bullet}$	$\mathbf{124}^{\bullet}_{\circ}$	$85^{\bullet}_{\circ}$	$86^{\bullet}$	$\mathbf{89}^{\bullet}_{\circ}$	58	60	63
		200	$115^{\bullet}$	$121^{\bullet}_{\circ}$	$121^{\bullet}$	$85^{\bullet}_{\circ}$	$87^{\bullet}$	$\mathbf{89}^{\bullet}$	59	59	63
		100	$113_{\circ}$	$122^{\bullet}_{\circ}$	$122^{\bullet}$	$83_{\circ}$	$86^{\bullet}_{\circ}$	$87^{\bullet}$	$61^{\bullet}_{\circ}$	$\underline{60}$	61
	WebCP	1,000	$112^{\bullet}$	117	117	$82^{\bullet}$	$85^{\bullet}$	86	$58^{\bullet}$	$65^{\bullet}$	$\underline{63}^{\bullet}$
		500	114	117	119	84	$85_{\circ}$	87	$59^{\bullet}$	$65^{\bullet}$	$\underline{64}^{\bullet}$
		200	114	115	117	84	$\underline{81}$	85	59	$62^{\bullet}$	$64^{\bullet}$
		100	$114^{\bullet}$	$\underline{113}_{\circ}$	117	$84^{\bullet}$	$\underline{81}_{\circ}$	86	$59_{\circ}$	$61^{\bullet}$	$64^{\bullet}$
PWDS	NewsCP	1,000	$112_{\circ}$	$121^{\bullet}$	$123^{\bullet}_{\circ}$	81	84	87	55	59	$62_{\circ}$
		500	116	$121^{\bullet}$	$123^{\bullet}$	84	$87^{\bullet}_{\circ}$	88	58	$61_{\circ}$	63
		200	115	$120^{\bullet}$	$121^{\bullet}$	84	$87^{\bullet}$	$\mathbf{89}^{\bullet}$	59	$60_{\circ}$	$64_{\circ}$
		100	111	$119^{\bullet}$	$122^{\bullet}$	82	$85^{\bullet}$	87	$60^{\bullet}$	60	$62_{\circ}$
	WebCP	1,000	$115^{\bullet}_{\circ}$	$118_{\circ}$	$121_{\circ}$	$84^{\bullet}_{\circ}$	$86^{\bullet}_{\circ}$	$88^{\bullet}_{\circ}$	$58^{\bullet}$	$\mathbf{66}^{\bullet}_{\circ}$	$\underline{65}^{\bullet}_{\circ}$
		500	$116_{\circ}$	117	$121_{\circ}$	$85^{\bullet}_{\circ}$	$\underline{84}$	$88_{\circ}$	$59^{\bullet}$	$65^{\bullet}$	$65^{\bullet}_{\circ}$
		200	$116^{\bullet}_{\circ}$	$\underline{115}$	$120_{\circ}$	$85^{\bullet}_{\circ}$	$\underline{82}_{\circ}$	$88_{\circ}$	59	$62^{\bullet}$	$65^{\bullet}_{\circ}$
		100	$116^{\bullet}_{\circ}$	$\underline{112}$	$119_{\circ}$	$84^{\bullet}$	$\underline{80}$	$87_{\circ}$	58	$61^{\bullet}$	$64^{\bullet}$

Table 11 Grammatical phrases among the top-ranked candidates.

Measure	1 judge	2 judges	3 judges
MDS only	$130^{\star\star}$	$106^{\star\star}$	$79^{\star\star}$
PWDS only	$141^{\star\star}$	$122^{\star\star}$	$87^{\star\star}$
HITS-News	$150^{\star\star}$	$140^{\star\star}$	$115^{\star\star}$
HITS-Web	**187**	**166**	**131**
Par_{Lin}, $N_S =$ 1,000, HAR	185	164	$123^{\star}$
Par_{skew}, $N_S =$ 1,000, HAR	184	164	$122^{\star\star}$

phrases among those each measure selected. From the results, we confirmed that PWDS is superior to MDS. However, we have not clarified why the advantage of PWDS is diminished when two factors are combined.

Unfortunately, both MDS and PWDS were significantly worse than the other measures ($p < 0.01$). The results confirm that querying a phrase to the Web contributes to assessing the grammaticality of the phrase. This may suggest that it is not necessary to assess the grammaticality explicitly. Another avenue for future work is to examine a more sophisticated language model to enhance the probabilistic model. Although the order of sibling nodes in the dependency structure is disassembled by MDS and PWDS, it reflects some preferences; for example, it sounds

Table 12 Appropriate paraphrases that were selected only on the basis of similarity factor.

$P(f)$	N_S	1 judge			2 judges			3 judges		
		BOW	MOD	HAR	BOW	MOD	HAR	BOW	MOD	HAR
NewsCP	1,000	93	107$^\bullet$	118	64	72$^\bullet$	81	40$^\bullet$	49$^\bullet$	58
	500	101$^\bullet$	107$^\bullet$	120	69	73$^\bullet$	85	42$^\bullet$	51$^\bullet$	60
	200	101$^\bullet$	112$^\bullet$	115	70	81$^\bullet$	83	45$^\bullet$	57$^\bullet$	59
	100	99	114$^\bullet$	114	70	82$^\bullet$	84	46$^\bullet$	58$^\bullet$	62$^\bullet$
WebCP	1,000	95$^\bullet$	99	122$^\bullet$	64	64	87$^\bullet$	36	45	64$^\bullet$
	500	98	101	**124**$^\bullet$	69	$\underline{66}$	**91**$^\bullet$	41	47	**67**$^\bullet$
	200	100	105	119$^\bullet$	70	70	86$^\bullet$	43	51	62$^\bullet$
	100	99	104	115$^\bullet$	71$^\bullet$	72	85$^\bullet$	45	51	61

strange to a native speaker if "左手で (with the left hand)" precedes "いつも (always)" in sentence in (9). In addition to the depth of dependencies and syntactic conditions (Uchimoto et al. 2000), discourse elements also exert a degree of influence upon the decision, e.g., highlighting old and new information. Obtaining a suitable granularity of nodes is another issue. One conceivable way is to introduce latent classes, such as the Semi-Markov class model (Okanohara and Tsujii 2007) and the hierarchical Pitman-Yor language model (Mochihashi and Sumita 2007). The existence of many orthographic variants of both content and function morphemes may prevent us from accurately estimating their grammaticality. To normalize these variations, methods and resources, such as (Ohtake et al. 2004; Matsuyoshi and Sato 2008), must be incorporated.

5.4.2 Similarity factor

We also conducted a component analysis for the similarity factor. Table 12 shows the numbers of paraphrase candidate pairs that were selected only on the basis of the similarity factor score and judged correct for all of the questions $\mathrm{Q_{sc}}$, $\mathrm{Q_{tc}}$, and $\mathrm{Q_{s2t}}$.

Basically, MOD outperformed BOW, and HAR did so for both BOW and MOD as Par_{Lin} and Par_{skew}, although we sometimes observed converse results (see underlines in Table 10). The superiority of MOD and HAR was also observed when the similarity factor was used alone (see Table 12). The impact of combining BOW and MOD into HAR was more significant than the entire probabilistic model. This supports the importance of HAR, implying, in contrast, that the single use of BOW and MOD in this probabilistic framework is not useful.

We could not find any notable tendencies regarding the number of Web snippets, N_s, from the results of the entire probabilistic model. The results were still unexplainable when the similarity factor was used alone: MOD somehow performed better when N_s was small, while the performance of HAR always peaked at $N_s = 500$.

On the corpus used for $P(f)$, we found that the huge Web corpus did not always produce a better result than the relatively small controlled corpus: NewsCP was regarded as better by a single judge, while WebCP was regarded as better by three judges (see "•" of the columns MOD and HAR in Table 10). Component analysis, on the other hand, showed that NewsCP tended to perform better than WebCP when BOW or MOD was used alone (see Table 12). We speculate that the morphological analyzer and dependency parser produce errors when features are extracted from the Web corpus, because those tools are tuned to newspaper articles. Likewise, $P(f|s)$ and $P(f|t)$ are expected to involve noise even though they are estimated using relatively clean parts of Web pages retrieved by querying phrases. Surprisingly, HAR compensated for the inferiority of WebCP, achieving accuracies comparable to the entire probabilistic model.

5.5 Summary of the input-wise evaluation

The probabilistic model was derived straightforwardly from the conditional probability $P(t|s)$. However, as shown above, no combination of parameters could achieve the desired result. The disappointing results might be due to the separate estimation of each factor. In other words, if the entire probabilistic model is optimized according to their implementation, the performance may be improved. The component analyses, however, revealed the lack of utility of each factor, underlining the difficulty of this complex approach.

The experiment produced two major findings. The first is the utility of contextual similarity in combination with the constituent similarity underlying morpho-syntactic paraphrases of predicate phrases. While a variety of measures have so far been proposed and evaluated basically for lexical paraphrases of words and word sequences for the paraphrase acquisition manner, we have empirically confirmed their applicability to computing the semantic similarity and substitutability of automatically generated paraphrase candidate pairs. The second finding is the versatility of the Web for representing the characteristics of predicate phrases. We could use up to only 1,000 Web snippets for each predicate phrase; nevertheless, contextual features extracted from the Web snippets enabled us to compute the appropriateness of paraphrase candidate pairs as paraphrases at a tolerable accuracy. The lack of a sophisticated weighting function suggests that we can improve the measures further.

Table 13 Agreement of human judgment ($n = 627$).

	Agr.	$\mathrm{Q_{sc}} \wedge \mathrm{Q_{tc}}$	κ of $\mathrm{Q_{s2t}}$	κ of $\mathrm{Q_{t2s}}$
Judge A–Judge B	447 (0.713)	545 (0.869)	0.697	0.605
Judge A–Judge C	490 (0.781)	554 (0.884)	0.746	0.684
Judge B–Judge C	488 (0.778)	572 (0.912)	0.813	0.588

6 Score-based evaluation

Previous work has acquired lexical paraphrases accurately by skimming the most reliable portion of a huge number of paraphrase candidates. Our second evaluation attempts from this viewpoint, i.e., how accurately a method gives higher scores to more appropriate paraphrase candidate pairs.

6.1 Sampling and judgment

In this evaluation, we investigate the performance of only the baseline measures, i.e., 2 count-based and 24 distributional similarity measures, because one of the assumptions that we have introduced to formalize the probabilistic model, i.e., s is grammatical, does not allow us to use the model for this purpose. A paraphrase candidate for an ungrammatical phrase may be improperly found to have a high conditional probability $P(t|s)$.

We first extracted the 200 best paraphrase candidate pairs for each measure. Three assessors were then asked to separately judge all of the 627 paraphrase candidate pairs that cover the 26 sample sets. The inter-assessor agreement of each pair of assessors is summarized in Table 13. The columns in the table correspond to those of Table 7. As a result of extracting the most reliable pairs, the agreement ratio of the four questions and the proportion of grammatical phrase pairs were significantly higher than those of the experiment in Section 5 ($p < 0.01$ of the 2-sample test for equality of proportions). κ-values for judging $\mathrm{Q_{s2t}}$ were also remarkably high, while those for judging $\mathrm{Q_{t2s}}$ were at the same range as the input-wise evaluation.

6.2 Results

Table 14 summarizes the numbers of the paraphrase candidate pairs to which questions $\mathrm{Q_{sc}}$, $\mathrm{Q_{tc}}$, and $\mathrm{Q_{s2t}}$ were answered "Yes." The measures performed remarkably better in this traditional evaluation method based on N-best samples than the input-wise evaluation (cf. Table 9). Unlike the input-wise evaluation in Section 5, two sample sets derived by different measures do not necessarily contain samples for the same sets of source phrases. We therefore applied the 2-

Table 14 Appropriate paraphrases among 200 best candidates.

Measure	N_S	1 judge			2 judges			3 judges		
		BOW	MOD	HAR	BOW	MOD	HAR	BOW	MOD	HAR
Par_{Lin}	1,000	191*	197	197	182	184	184	155	**158**	**158**
	500	196	**198**	**198**	186	$\underline{185}$	$\underline{185}$	153	157	$\underline{156}$
	200	**198**	**198**	**198**	**187**	$\underline{185}$	$\underline{185}$	154	$\underline{153}$	$\underline{153}$
	100	197	197	197	186	$\underline{184}$	$\underline{184}$	155	$\underline{153}$	$\underline{151}$
Par_{skew}	1,000	197	197	197	185	$\underline{183}$	$\underline{184}$	155	155	$\underline{154}$
	500	197	197	197	185	$\underline{181}$	$\underline{183}$	155	$\underline{149}$	$\underline{151}$
	200	**198**	$\underline{197}$	**198**	184	$\underline{181}$	$\underline{182}$	154	$\underline{149}$	$\underline{150}$
	100	197	$\underline{196}$	**198**	183	$\underline{181}$	183	153	$\underline{150}$	$\underline{152}$
HITS-News		113**			85**			71**		
HITS-Web		120**			88**			68**		

sample test for equality of proportions to compare accuracies of two arbitrary measures. "$\star\star$" and "$\star$" denote the significance levels $p < 0.01$ and $p < 0.05$ between the best measure and each measure, respectively.

The results show the significant advantages of distributional similarity measures over the count-based measures. On the distributional similarity measures, the lower number of "$\star$" implies that the performance of these measures reached a ceiling. The underlined numbers in Table 14 indicate the results of MOD performed worse than BOW, and those of HAR that did so for either BOW or MOD. Fortunately, BOW performed the task as well as MOD and HAR. This is a favorable result because BOW features can be extracted much more quickly and accurately than MOD features.

7 Error analysis

The accuracies based on the assessments of the three judges were approximately 40% in the input-wise evaluation and 80% in the 200-best evaluation. This section describes typical errors, i.e., inappropriate paraphrase candidate pairs whose appropriateness was improperly estimated to be high.

Most of the common errors in our experiment were generated when the following knowledge was applied to the given N_1:N_2:C:V type phrase together (Fujita et al. 2007).

(14) a. Transformation pattern: N_1:N_2:C:$V \Rightarrow np(N_1, N_2)$:$C$:$V$

b. Generation function: $np(N_1, N_2) \Rightarrow N_2$

Dropping a nominal element N_1 of the given nominal compound N_1:N_2 by (14b) normally generalizes the meaning that the compound conveys, and thus generates appropriate (directional) paraphrases such as those shown in (15).

(15) *s.* 損害-賠償-を 求める
damage-compensation-ACC to require
to demand compensation for damages

t. 賠償-を 求める
compensation-ACC to require
to demand compensation

However, it caused errors in some cases; for example, dropping N_1 of the source sentence in (16) generates an anomaly, because it was the semantic head of the sentence.

(16) *s.* 出血-多量-で 死亡する
bleeding-much-because to die
to die due to heavy blood loss

t. *多量-で 死亡する
much-because to die
*to die due to plenty

In our experiment, source phrases were assumed to be grammatical. However, as exhibited by the examples (12a) and (17*s*), some of extracted sub-parses were ungrammatical.

(17) *s.* *気圧-配置-が 強まる
pressure-pattern-NOM to be strengthened
*pressure pattern is strengthened

t. *配置-が 強まる
layout-NOM to be strengthened
*layout is strengthened

For 11 source phrases among the 200 samples in the first evaluation, Q_{sc} was answered "No," i.e., ungrammatical, by at least 1 judge. Thus, for automatic generation of paraphrase knowledge, the task of capturing the appropriate boundary of a given phrase should be addressed.

The other notable observation is that the appropriateness of paraphrase candidates for predicate phrases containing an adjective was poorly computed. The primal source of the errors for *Adj*:*N*:*C*:*V* type phrases was the subtle change of nuance by switching syntactic heads as illustrated in (18).

(18) *s.* 良い 仕事-を する
be good work-ACC to do
to do a good job

t_1. ≠良く 仕事-する
much to work
≠to work hard

t_2. ≠仕事-を 良く-する
work-ACC be good-to make
≠to improve the work

Most errors in paraphrasing *N*:*C*:*Adj* type phrases, on the other hand, were caused due to the difference of the aspectual property and agentivity between adjectives and verbs. For example, (19*s*) can describe not only things whose quality has been improved as inferred by (19*t*), but also those that were originally of high quality (Fujita et al. 2006).

(19) *s.* 質-が 高い
quality-NOM be high
(Its) quality is high

t. $^{\neq}$質-が 高まる
quality-NOM to rise
$^{\neq}$(Its) quality rises

$\mathrm{Q_{s2t}}$ for (19) was thus judged "No." To realize paraphrases that involve the change of syntactic categories, we need to enhance the lexical derivation dictionary by capturing the subtle difference between the original and derived words.

8 Conclusion

A pair of expressions qualifies as paraphrases if and only if they are semantically equivalent, substitutable in some contexts, and grammatical. When paraphrase knowledge is represented with general transformation patterns to attain high coverage of paraphrases, we should assess not only the first and second criteria, but also the third criterion. On the basis of this recognition, in this paper, we address the task of measuring the appropriateness of the given pair of phrases as paraphrases, as a post-generation assessment of automatically generated candidates of phrasal variants. We examined several measures including a novel probabilistic model, which consists of two components: (i) a structured N-gram language model that ensures grammaticality and (ii) a distributional similarity measure for estimating semantic equivalence and substitutability between two phrases.

Through the experiment, we empirically evaluated the performance of the measures, analyzed the characteristics, and found the following.

- Contextual similarity in combination with the constituent similarity of morpho-syntactic paraphrases is effective for measuring the appropriateness of the given pair of predicate phrases as paraphrases. Among several measures, two existing distributional similarity measures achieved a tolerable level of performance. They performed the task significantly better than not only the count-based measures, which only assess the grammaticality of paraphrase candidates, but also the probabilistic model. We also showed that combining two feature sets was beneficial.
- The Web is versatile for representing the characteristics of predicate phrases. Contextual features extracted from the Web snippets contributed to the task. Our first aim in harnessing Web snippets was to overcome the data sparseness problem; however, additionally, issuing a phrase as a query to a commercial search engine also contributes to assessing the grammaticality of the phrase to some degree.

Two different evaluations also revealed the difference between the difficulties of the recognition and generation tasks.

We are developing a two-step plan for our future work. In the first step, we will attempt to enhance the measures by incorporating a more sophisticated language model for the probabilistic model, and further examining distributional similarity measures: the three elements described in Section 2.2. Once the performance reaches a reasonable level, we will generate a huge paraphrase knowledge base consisting of millions of phrasal variant pairs and provide it to the research community.

Reference

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

Barzilay, R. and McKeown, K. R. (2001). “Extracting paraphrases from a parallel corpus.” In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 50–57.

Barzilay, R. and Lee, L. (2003). “Learning to paraphrase: an unsupervised approach using multiple-sequence alignment.” In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp. 16–23.

Bhagat, R., Pantel, P., and Hovy, E. (2007). “LEDIR: an unsupervised algorithm for learning directionality of inference rules.” In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (*EMNLP-CoNLL*), pp. 161–170.

Dolan, B., Quirk, C., and Brockett, C. (2004). “Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources.” In *Proceedings of the 20th International Conference on Computational Linguistics* (*COLING*), pp. 350–356.

Dras, M. (1999). *Tree adjoining grammar and the reluctant paraphrasing of text.* Ph.D. thesis, Division of Information and Communication Science, Macquarie University.

Fujita, A., Inui, K., and Matsumoto, Y. (2004). “Detection of incorrect case assignments in automatically generated paraphrases of Japanese sentences.” In *Proceedings of the 1st International Joint Conference on Natural Language Processing* (*IJCNLP*), pp. 14–21.

Fujita, A. and Inui, K. (2006). “Building a paraphrase corpus based on class-oriented candidate generation.” *Journal of Natural Language Processing*, **13** (3), pp. 133–150. (in Japanese).

Fujita, A., Masuno, N., Sato, S., and Utsuro, T. (2006). "Adjective-to-verb paraphrasing in Japanese based on lexical constraints of verbs." In *Proceedings of the 4th International Natural Language Generation Conference* (*INLG*), pp. 41–43.

Fujita, A., Kato, S., Kato, N., and Sato, S. (2007). "A compositional approach toward dynamic phrasal thesaurus." In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (*WTEP*), pp. 151–158.

Geffet, M. and Dagan, I. (2004). "Feature vector quality and distributional similarity." In *Proceedings of the 20th International Conference on Computational Linguistics* (*COLING*), pp. 247–253.

Geffet, M. and Dagan, I. (2005). "The distributional inclusion hypotheses and lexical entailment." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 107–116.

Habash, N. (2004). "The use of a structural N-gram language model in generation-heavy hybrid machine translation." In *Proceedings of the 3rd International Natural Language Generation Conference* (*INLG*), pp. 61–69.

Hagiwara, M., Ogawa, Y., and Toyama, K. (2008a). "Effective use of indirect dependency for distributional similarity." *Journal of Natural Language Processing*, **15** (4), pp. 19–42.

Hagiwara, M., Ogawa, Y., and Toyama, K. (2008b). "A comparative study on effective context selection for distributional similarity." *Journal of Natural Language Processing*, **15** (5), pp. 119–150.

Harris, Z. (1957). "Co-occurrence and transformation in linguistic structure." *Language*, **33** (3), pp. 283–340.

Harris, Z. (1968). *Mathematical structures of language.* John Wiley & Sons.

Ibrahim, A., Katz, B., and Lin, J. (2003). "Extracting structural paraphrases from aligned monolingual corpora." In *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications* (*IWP*), pp. 57–64.

Inui, K. and Fujita, A. (2004). "A survey on paraphrase generation and recognition." *Journal of Natural Language Processing*, **11** (5), pp. 151–198. (in Japanese).

Iordanskaja, L., Kim, M., Kittredge, R., Lavoie, B., and Polguère, A. (1992). "Generation of extended bilingual statistical reports." In *Proceedings of the 14th International Conference on Computational Linguistics* (*COLING*), pp. 1019–1023.

Jacquemin, C. (1999). "Syntagmatic and Paradigmatic Representations of Term Variation." In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 341–348.

Kawahara, D. and Kurohashi, S. (2006). "Case frame compilation from the Web using high-performance computing." In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*.

Kilgarriff, A. (2007). "Googleology is bad science." *Computational Linguistics*, **33** (1), pp. 147–151.

Lee, L. (1999). "Measures of distributional similarity." In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 25–32.

Lin, D. (1998). "Automatic retrieval and clustering of similar words." In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL)*, pp. 768–774.

Lin, D. and Pantel, P. (2001). "Discovery of inference rules for question answering." *Natural Language Engineering*, **7** (4), pp. 343–360.

Matsuyoshi, S. and Sato, S. (2008). "Automatic paraphrasing of Japanese functional expressions under style and readability specifications." *Journal of Natural Language Processing*, **15** (2), pp. 75–99. (in Japanese).

Mel'čuk, I. and Polguère, A. (1987). "A formal lexicon in meaning-text theory (or how to do lexica with words)." *Computational Linguistics*, **13** (3-4), pp. 261–275.

Mochihashi, D. and Sumita, E. (2007). "The Infinite Markov Model." In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS)*.

Ohtake, K., Sekiguchi, Y., and Yamamoto, K. (2004). "Detecting transliterated orthographic variants via two similarity metrics." In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pp. 709–715.

Okanohara, D. and Tsujii, J. (2007). "A discriminative language model with pseudo-negative samples." In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 73–80.

Pang, B., Knight, K., and Marcu, D. (2003). "Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences." In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 102–109.

Pantel, P., Bhagat, R., Coppola, B., Chklovski, T., and Hovy, E. (2007). "ISP: Learning Inferential Selectional Preferences." In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 564–571.

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

Sekine, S. (2005). "Automatic paraphrase discovery based on context and keywords between NE pairs." In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP)*, pp. 80–87.

Shinyama, Y., Sekine, S., Sudo, K., and Grishman, R. (2002). "Automatic paraphrase acquisition from news articles." In *Proceedings of the 2002 Human Language Technology Conference (HLT)*.

Szpektor, I., Tanev, H., Dagan, I., and Coppola, B. (2004). "Scaling Web-based acquisition of entailment relations." In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 41–48.

Szpektor, I., Dagan, I., Bar-Haim, R., and Goldberger, J. (2008). "Contextual preferences." In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 683–691.

Takahashi, T., Iwakura, T., Iida, R., Fujita, A., and Inui, K. (2001). "Kura: a transfer-based lexico-structural paraphrasing engine." In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS) Workshop on Automatic Paraphrasing: Theories and Applications*, pp. 37–46.

Torisawa, K. (2002). "An unsupervised learning method for associative relationships between verb phrases." In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pp. 1009–1015.

Torisawa, K. (2006). "Acquiring inference rules with temporal constraints by using Japanese coordinated sentences and noun-verb co-occurrences." In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 57–64.

Uchimoto, K., Murata, M., Ma, Q., Sekine, S., and Isahara, H. (2000). "Word order acquisition from corpora." In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pp. 871–877.

Weeds, J. (2003). *Measures and applications of lexical distributional similarity.* Ph.D. thesis, University of Sussex.

Weeds, J., Weir, D., and Keller, B. (2005). "The distributional similarity of sub-parses." In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pp. 7–12.

Wu, H. and Zhou, M. (2003). "Synonymous collocation extraction using translation information." In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*

(*ACL*), pp. 120–127.

Yamamoto, K. (2002). "Acquisition of lexical paraphrases from texts." In *Proceedings of the 2nd International Workshop on Computational Terminology* (*CompuTerm*), pp. 22–28.

Yoshida, M., Nakagawa, H., and Terada, A. (2008). "Gram-free synonym extraction via suffix arrays." In *Proceedings of the 4th Asia Information Retrieval Symposium* (*AIRS*), pp. 282–291.

Atsushi Fujita: Atsushi Fujita is an associate professor at the School of Systems Information Science, Future University-Hakodate. His research interests include natural language processing and computational linguistics, especially automatic paraphrasing and lexical semantics. He received B.E. and M.E. from Kyushu Institute of Technology in 2000 and 2002, respectively. He received Doctor of Engineering from Nara Institute of Science and Technology in 2005.

Satoshi Sato: Satoshi Sato is a professor of Department of Electrical Engineering and Computer Science in Nagoya University. His recent work in natural language processing focuses on automatic paraphrasing, controlled language, and automatic lexicon compilation. He received B.E., M.E., and Doctor of Engineering from Kyoto University in 1983, 1985, and 1992, respectively.

(Received July 2, 2009)
(Revised August 30, 2009)
(Accepted September 30, 2009)