Automatically Computable Metrics to Generate Metaphorical Verb Expressions

Akira Miyazawa† and Yusuke Miyao††

The automatic generation of metaphorical expressions helps us write imaginative texts such as poems or novels. This paper proposes a new metaphor generation task, evaluation metrics, and a method to solve the task. Our task is formalized as a problem of finding metaphorical paraphrases for a literal Japanese phrase consisting of a subject, an object, and a verb. We use four evaluation metrics: synonymousness, metaphoricity, novelty, and comprehensibility. Our proposed method generates metaphorical expressions by using three automatically computable scores—similarity, figurativeness, and rarity—corresponding to one of the evaluation metrics. By crowdsourcing, we show how these scores are related to those given by humans in terms of the evaluation metrics and how they are useful in finding human’s preferred expressions in pairwise comparisons.

Key Words: Metaphor, Text Generation

1 Introduction

Metaphor is an essential tool in writing imaginative works such as poems or novels. In a practical sense, for example, metaphors can make political speeches more persuasive or attractive (Charteris-Black 2011). It can be also useful in educating learners and helping them understand through analogy between new ideas and familiar things (Williams 1986). The automatic generation of metaphors would help us use effective ones that are appropriate to the situation.

Automatic metaphor generation poses an advantage in creating novel metaphors. Once metaphorical expressions become part of general usage, they lose their rhetorical effect and are, therefore, called dead metaphors (Richards 1965; Deignan 2005). Such expressions are easy to conjure or find in ordinary dictionaries or thesauri. On the other hand, novel or “living” metaphors are not, by their nature. This causes a demand for metaphor generation systems. Thus we focus on producing such “living” metaphors.

Verbs are a part of speech that is often used metaphorically (Steen, Dorst, Herrmann, Kaal, Krennmayr, and Pasma 2010). For example, we describe emotions using such verbs as “hatred
burns” or “being starved for love.” In the context of computers, we use “spawn” and “kill” when talking about processes. We can generate various metaphorical expressions even if we only care about verbal expressions. On the other hand, previous works in the field have focused on “A like B” similes where A and B are nouns (Abe, Sakamoto, and Nakagawa 2006; Terai and Nakagawa 2010). The differences in the types of target cause some difficulties. First, we need to be mindful whether generated results are actually metaphorical, as they do not necessarily have explicit metaphor signals (e.g., “like” or “as”).1 We also need to pay more attention to the comprehensibility of generated expressions. The absence of explicit signals for metaphors discourages us from optimistically expecting readers to understand the meaning of generated expressions in a nonliteral sense. For example, we have difficulty in understanding “He is a kangaroo,” but we can easily understand “He can jump high as if he were a kangaroo.”

This paper proposes a method to generate metaphorical expressions using verbs. More precisely, generation targets are phrases of a subject, object, and verb, where verbs are used metaphorically. To our knowledge, this work is the first regarding the automatic generation of metaphors other than similes. First, we define a new task for automatic metaphor generation as a paraphrasing task and propose two evaluation metrics—synonymousness and preference—in addition to metaphoricity, novelty, and comprehensibility, which were proposed by Miyazawa and Miyao (2017). Then we introduce a model to solve the task that finds paraphrasing candidates using automatically computable scores corresponding to the metrics: similarity, figurativeness, rarity, and overall score. We conduct three experiments. The first measures the performance of the metaphor classifier for computing figurativeness score. In the second and third, we use crowdsourcing. The second experiment examines how our system’s scores correlate with those provided by humans in terms of the evaluation metrics. In the last experiment, we show the usefulness of our system’s scores in finding good metaphorical expressions, that is, expressions preferred by humans in pairwise comparisons.

2 Related Works

The term “metaphor” is used for several different but closely related meanings across domains. Conceptual mapping theory (CMT) (Lakoff 1987, 1993) defines metaphors, or conceptual metaphors, as mappings from one conceptual domain to another. Expressions such as “inflam-

---

1 Strictly speaking, generation of similes also needs to verify if results are metaphorical because signals of simile are also used for exemplification, subcategorization, and many other usages. Refer to Low (2010) for how to identify metaphorical similes from simile-like expressions.
Miyazawa and Miyao

Metrics to Generate Metaphorical Verb Expressions

“imatory remarks” or “he was breathing fire” are regarded as linguistic realizations of a conceptual metaphor, from FIRE to ANGER in this case, and called linguistic metaphors. It is linguistic metaphors that we are going to generate, and we basically use the term “metaphor” in this sense hereafter. However, we do not apply CMT explicitly in the generation because fixing conceptual domains would restrict the variety or novelty of generated expressions.

Identifying metaphors in texts is an essential process in analyzing the real uses of metaphors. We also require it when collecting literal expressions for input and picking up only metaphorical expressions from output. Steen et al. (2010) suggested a guideline called MIPVU (Metaphor Identification Procedure VU University Amsterdam), which was used to construct the VU Amsterdam Metaphor Corpus. The main procedure of MIPVU is as follows. First, annotators determine the contextual meaning of the target word. Then they look up the word in a dictionary and search for a more basic meaning than the contextual meaning. Basic meanings tend to be concrete and easy to imagine, see, hear, feel, smell, or taste, or are related to bodily action. If a more basic meaning exists, the contextual meaning differs from the basic meaning, and if it can be understood in comparison, then it is judged as a metaphor-related word.

Automatic metaphor identification is a field in NLP (Natural Language Processing) that tries to automatically classify expressions into metaphorical or literal, aiming to contribute to tasks such as word sense disambiguation as well as linguistic analysis. Birke and Sarkar (2006) developed a classifier as a by-product of a dataset called the TroFi Example Base, which contains literal and nonliteral usages of 50 English verbs, and was first built through human annotation and later augmented by active learning. Tsvetkov, Boytsov, Gershman, Nyberg, and Dyer (2014) propose a method to identify adjective-noun and subject-object-verb metaphors across languages. They trained a classifier on the TroFi dataset using four features, the abstractness, imageability, supersenses, and vector space word representation of English words. These features are based on concepts rather than individual words or languages, and proved to be effective by experiments in Russian, Farsi, and Spanish as well as English. We employ their method to score how metaphorical generated Japanese expressions are.

For automatic metaphor generation, Abe et al. (2006) proposed a model that outputs “A like B” expressions from phrases with a noun and adjectives modifying it. For example, given “a young, innocent, and fine character,” the model suggests phrases such as “a character like a grandchild” or “the character like a baby.” In evaluating the generated results, Abe et al. (2006) used adequacy, ease of visualization, amusingness, and novelty, showing that some phrases,
through their system, get higher scores than those by humans.

To evaluate metaphors unrestricted to similes, Miyazawa and Miyao (2017) suggested using three metrics—metaphoricity, comprehensibility, and novelty—and showed the possibility to obtain reliable scores in terms of these metrics through crowdsourcing. They composed targets for evaluation by combining a noun mainly taken from the emotion domain and verbs from the water domain. For example, zouo 憎悪 (hatred) and X-ga afureru ○○があふれる (something spills out) constitute the expression zouo-ga afureru 憎悪があふれる (hatred spills out). They also introduce overall evaluation, defined as the arithmetic mean of the three metrics, and show that it can indicate “good metaphors” in that expressions with high overall evaluation are preferred by a judge in pairwise comparison, and that they are metaphorical in the sense of MIPVU.

In this paper, we also use metaphoricity, comprehensibility, novelty, and a modified version of overall evaluation.

3 Task

3.1 Task Definition

We introduce a new metaphor generation task: Given a literal expression, a system suggests alternative verbs that make the expression synonymous and metaphorical. We assume a practical use where writers have expressions in their mind and want to make them rhetorically appealing. In this paper, the expressions are in Japanese, each of them consisting of a subject-object-verb (SOV) triple. For example,

(i) 彼が気持ちを考える。
He-NOM feelings-ACC consideration-do
“He considers someone’s feelings.”

Given this expression, the system would output some alternative verbs such as kumitoru 汲み取る (to draw water from a well). Then it would make an expression kare-ga kimochi-wo kumitoru, which can be interpreted as having the same meaning as the input. Here the verb kumitoru is metaphorical, as it is used to mean “to consider” while it basically means the physical action of somebody or something drawing water from a well.

The task is formalized as follows. Let $\varepsilon$ be an expression of the form “s-ga o-wo v,” in Japanese “s が o を v,” where s, o, and v are the subject, object, and verb, respectively. The suffix -ga が and -wo を are the nominative case and accusative case markers, respectively. We denote it
As $\varepsilon = (s, o, v)$. Then, the task is to find an alternative verb $v'$ and make another expression $\varepsilon' = (s, o, v')$ for a given expression $\varepsilon$ so that $\varepsilon'$ achieves high scores in the five evaluation metrics introduced in Section 3.2.

We focus on verb expressions because verbs tend to be metaphorical compared to other parts of speech (Steen et al. 2010) and expect that this would make it easier for us to construct balanced datasets of metaphorical and literal expressions and generate many and various metaphorical expressions. Also, we can practically use the cross-lingual metaphor classifier by Tsvetkov et al. (2014) with small modifications, which is necessary to compute how metaphorical expressions are.

### 3.2 Evaluation Metrics

In the evaluation of results, we use the same metrics by Miyazawa and Miyao (2017)—metaphoricity, novelty, and comprehensibility—and introduce two additional metrics, synonymousness and preference. Table 1 illustrates some alternative verbs for the input expression *kare-ga kimochi-wo kouryo-suru* 彼が気持ちを考慮する (he considers someone’s feelings) and their expected classes of scores regarding the metrics except preference, which is used in comparison with another result.

**Synonymousness** measures how synonymous input and output expressions are. It is required because we define the task as a paraphrasing problem. For example, for the source expression (i), *sukuitoru* 拭い取る (to scoop up, to consider) and *suisoku-suru* 推測する (to guess) are more appropriate than *tsutaeru* 伝える (to tell).

**Metaphoricity**, the metaphorical degree of output expressions, is essential in the task. Given (i), *sukuitoru* makes more metaphorical expressions than *kouryo-suru* 考慮する (to consider), because it is also used to mean a physical action, “to scoop up” as well as “to consider,” in contemporary Japanese.

<table>
<thead>
<tr>
<th>Syn.</th>
<th>Metaphoricity</th>
<th>Novelty</th>
<th>Compr.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>sukuitoru</em> 拭い取る (to scoop up, to consider)</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td><em>suisoku-suru</em> 推測する (to guess)</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td><em>tateru</em> 建てる (to build, to construct)</td>
<td>?</td>
<td>?</td>
<td>High</td>
</tr>
<tr>
<td><em>tsutaeru</em> 伝える (to tell)</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
</tbody>
</table>

The symbol ? means that it is difficult to measure.
Novelty measures how novel output expressions are. As with the other metrics, it is measured as an expression rather than as a verb. Consequently, a common verb, tateru 建てる (to build), is evaluated as novel if it is suggested for the source expression (i) because kare-ga kimochi-wo tateru 彼が気持ち建てる (he builds someone’s feelings) sounds unnatural and is hardly used.

Comprehensibility measures the ease with which an expression is understood. For example, kare-ga kimochi-wo tsutaeru 彼が気持ちを伝える (he tells someone his feelings) is more comprehensible than kare-ga kimochi-wo tateru. In each metric’s evaluation, the other ones are not considered. However, comprehensibility affects synonymousness and metaphoricity in that low comprehensibility makes it difficult to evaluate them, as they can be assessed after evaluators understand the target expression’s meaning. This implies that we need to divide evaluation into two stages: comprehensibility in the first stage and the others in the second stage, or vice versa.

Preference measures how good an expression is compared to others. Therefore, it is different from the other metrics in that it cannot be measured by one expression. In a practical sense, users of metaphor generation systems are likely clueless as to which metrics they should give more weight to; in such cases, preference is helpful. We quantify preference by collecting pairwise comparisons from multiple judges and aggregating them into a single score, the construction of which is described in Section 3.3. Miyazawa and Miyao (2017) also used pairwise comparisons in evaluation, but they collected these comparisons from only one judge. Thus, our preference is more objective.

3.3 Crowdsourcing

We use crowdsourcing to evaluate expressions according to the above-mentioned five metrics. Target expressions are sampled from bins that are made according to the corresponding automatically computable scores (discussed in the next section). For example, for synonymousness, we have the corresponding score called similarity. Then, to evaluate synonymousness, we divide the range of similarity into 10 bins and sample one from each bin. This process makes it easier to analyze the relation between these metrics and the system’s scores. We ask evaluators to rate the expressions on a scale of 1 to 5 as illustrated in Fig. 1. We provide short descriptions of the choices for 1 and 5 and tell evaluators that higher scores indicate higher degrees of synonymousness, metaphoricity, novelty, and comprehensibility. For each question, we collect answers \( a_1, \ldots, a_N \) from \( N \) evaluators and aggregate scores by scaling the answers to \([0, 1]\) and averaging them:

\[
\frac{1}{N} \sum_{i=1}^{N} \frac{a_i - 1}{4}. \tag{1}
\]
We collect evaluators for the experiments through Yahoo! Crowdsourcing.³ Crowdsourcing participants are not restricted by age, sex, or district of residence but are required to have enough ability to give a correct answer in a Japanese grammar quiz.

As stated in Section 3.2, we expect the evaluation process to be in two stages—comprehensibility and the others—for a strict analysis. In crowdsourcing, however, we conduct the evaluation at the same time. Evaluating comprehensibility first and removing expressions with low comprehensibility might result in imbalanced data and analysis issues. Meanwhile, the opposite process—other metrics first and comprehensibility later—has the problem of cost. Some evaluation results will be discarded if the expressions have low comprehensibility scores. Thus, we evaluate all the metrics at once.

### 3.4 Preference

We evaluate expression preference by asking evaluators to choose the superior of two expressions and aggregating the answers. An example of the style of questioning is shown in Fig. 2. To create practical systems, it would be better to ask how good expressions are in some contexts or how good they are for some specific purposes. However, we phrase the questions as in Fig. 2 to reduce evaluator load and simplify the task. The target expressions are sampled the same way as the other metrics.

³ https://crowdsourcing.yahoo.co.jp/
Fig. 2 Example of questions for preference.

Answers provided by evaluators are aggregated into real-valued scores using the Bradley–Terry model (Bradley and Terry 1952). Suppose the input expression is \((s, o, v)\) and the system suggests a set of verbs \(V\); then we have a set of expressions \(E = \{(s, o, v') \mid v' \in V\}\). We denote that \(\varepsilon \in E\) is preferred to \(\varepsilon' \in E\) by \(\varepsilon \succ \varepsilon'\). Then we model the probability that \(\varepsilon \succ \varepsilon'\) is observed as

\[
P(\varepsilon \succ \varepsilon') = \frac{\pi_\varepsilon}{\pi_\varepsilon + \pi_{\varepsilon'}},
\]

where \(\pi_\varepsilon\) is a positive parameter for each \(\varepsilon \in E\), which is supposed to represent a kind of quality of expression. Suppose we observe the \(n_{\varepsilon \varepsilon'}\) occurrences of \(\varepsilon \succ \varepsilon'\) for each \(\varepsilon, \varepsilon' \in E\). The log likelihood of this event is given by

\[
L(\pi) = \log \prod_{\substack{\varepsilon, \varepsilon' \in E \\
\varepsilon \neq \varepsilon'}} \left( \frac{\pi_\varepsilon}{\pi_\varepsilon + \pi_{\varepsilon'}} \right)^{n_{\varepsilon \varepsilon'}},
\]

where \(\pi = (\pi_\varepsilon)_{\varepsilon \in E}\). We can find a unique optimal parameter vector \(\pi^*\) that maximize \(L\) according to Ford (1957). Consequently, we use \(\pi_\varepsilon\) as the preference score of expression \(\varepsilon\).

4 Method

To find verbs that make highly evaluated expressions in terms of the above evaluation metrics, our model uses four automatically computable scores: similarity, figurativeness, rarity, and overall score. The correspondence is summarized in Table 2. Here we put commonness as the counterpart of comprehensibility, but this is a tentative solution. We do not suggest a specific score for comprehensibility, as it requires much complex word and world knowledge, and is, therefore,
Table 2  Correspondence between evaluation metrics and automatically computable scores used in finding paraphrases.

<table>
<thead>
<tr>
<th>Evaluation Metric</th>
<th>Automatic Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymousness</td>
<td>Similarity</td>
</tr>
<tr>
<td>Metaphoricity</td>
<td>Figurativeness</td>
</tr>
<tr>
<td>Novelty</td>
<td>Rarity</td>
</tr>
<tr>
<td>Comprehensibility</td>
<td>(Commonness)</td>
</tr>
<tr>
<td>Preference</td>
<td>Overall Score</td>
</tr>
</tbody>
</table>

difficult to formulate. When an automatic score for comprehensibility is required, we compute a tentative score, commonness, as a negative monotonic transformation of rarity.

This idea is based on the observation by Miyazawa and Miyao (2017) that novelty is negatively correlated with comprehensibility.

In the model, we assume that evaluation metrics are independent and modeled by a single automatic score. We check the validity of this hypothesis in Section 6.2.

4.1 Similarity

By the task definition, we need to compute the semantic similarity between input and output phrases, but this is too difficult. Instead, we consider verbs only and calculate the cosine similarity between their vector representations. Let $W$ be the set of all the words in the corpus. Using the continuous bag-of-word (CBOW) algorithm (Mikolov, Chen, Corrado, and Dean 2013), we obtain a map $\gamma : W \rightarrow \mathbb{R}^n$, where $n$ is a predefined integer. Given a source expression with a verb $v$, the similarity between $v$ and its alternative $v'$ is calculated by the affine transformation of cosine similarity:

$$\text{sim}(v, v') = \frac{1}{2} \left( 1 + \frac{\gamma(v) \cdot \gamma(v')}{\|\gamma(v)\| \|\gamma(v')\|} \right).$$

Affine transformation is applied so that the score ranges from 0 to 1 and becomes compatible with the other scores.

4.2 Rarity

Let $p(*)$ and $p(*, *)$ represent the probabilities of a word and OV pair, respectively, estimated by a corpus. Let $g(o)$ be the label representing the cluster which $o$ belongs to. We compute the rarity of an expression whose object is $o$ and verb is $v$ by
rar \((o, v)\) = \(\begin{cases} \frac{\log p(g(o), v) \log p(v)}{A_{o,v} \log p(v)} + B_{o,v} & \text{if } \log p(g(o), v) > 0 \text{ and } \log p(v) > 0, \\ 1 & \text{otherwise,} \end{cases}\) \tag{5}

where \(A_{o,v}\) and \(B_{o,v}\) are constants dependent on \(o\) and a set of alternative verbs \(V\) to make the score take value in \([0, 1]\). Accurately,

\[
A_{o,v} = -\frac{1}{M_{o,v} - m_{o,v}}, \quad B_{o,v} = \frac{M}{M_{o,v} - m_{o,v}}, \tag{6}
\]

\[
M_{o,v} = \max_{v \in V} \frac{\log p(g(o), v)}{\log p(v)}, \quad m_{o,v} = \min_{v \in V} \frac{\log p(g(o), v)}{\log p(v)}. \tag{7}
\]

We estimate probabilities as follows. When the corpus contains \(L\) words, and a word \(w\) occurs \(c_w\) times, then

\[p(w) = \frac{c_w}{L}.\] \tag{8}

To obtain probabilities of OV pairs, we first extract all the pairs of a verb and its direct object. Suppose the corpus has \(L'\) pairs in total. When a pair \((o, v)\) appears \(c_{o,v}\) times in the corpus, we have

\[p(o, v) = \frac{c_{o,v}}{L'}.\] \tag{9}

Equation 5 has a problem that many possible OV pairs are not observed in the corpus, and the probabilities become 0. Consequently, we have difficulty in comparing rarity. To overcome the issue, we reduce OV pair variety by clustering nouns. First we construct clusters \(\{C_i\}_{i=1}^K\) according to vector representations of words. We obtain vector representations in the same way as Section 4.1. After that, we make \(K\) clusters by spherical \(K\)-means (Banerjee, Dhillon, Ghosh, and Sra 2005), a variant of \(K\)-means clustering algorithm where the nearness between a point and each cluster representative is computed by cosine similarity. After that, we replace nouns in the text with the corresponding labels cluster_1, ... , cluster_K. This is \(g\) in Equation 5. Probabilities tend to take large values for frequent words and almost zero for most words as illustrated in Fig. 3. Thus, we take the logarithm of probabilities and obtain less skewed values representing degree of novelty.

4.3 Figurativeness

To automatically score metaphoricity, we build a classifier that categorizes an expression into the metaphorical class (\(\text{Met}\)) or the literal class (\(\text{Lit}\)). The score can be binary but can be continuous when we adopt a discriminative classifier that models the posterior probability of the
class $\text{Met}$ when the feature of the expression is given. Let $\phi$ be a function that assigns a feature vector to SOV triples. Then we define the \textit{figurativeness} score of an expression consisting of a subject $s$, object $o$, and verb $v$ by

$$\text{fig}(s, o, v) = p(\text{Met} \mid \phi(s, o, v)),$$

where $p$ is the probability estimated by a discriminative classifier.

We construct feature vectors the same way as in Tsvetkov et al. (2014). We calculate \textit{abstractness and imageability}, \textit{supersenses}, and \textit{vector space word representations} by translating Japanese words into English. In addition, as in Tsvetkov, Mukomel, and Gershman (2013), we use \textit{named entities} as an additional feature, which is directly calculated from Japanese. The mapping $\phi$ concatenates each feature of $s$, $o$, and $v$ and then concatenates the three vectors again.

\textit{Abstractness and imageability} are real-valued scores, which human evaluators give to each word according to how abstract and imageable it is. These scores are good indicators of metaphors.
For example, a verb afureru あふれる (to overflow) usually becomes metaphorical when its subject is an abstract word such as kanjou 感情 (emotion) but becomes literal when its subject is a concrete word such as mizu 水 (water). While abstractness and imageability are similar concepts, they are not redundant. Some abstract words, such as “torture,” call up visual images (Tsvetkov et al. 2014). Supersenses are coarse semantic categories in WordNet (Miller 1995), which consists of 47 categories, including 27 for nouns (noun.act, noun.quantity, etc.) and 15 for verbs (verb.motion, verb.cognition, etc.). Vector space representations are the same as those used in computing similarity and rarity. Named entities are used to specify subject and object types: human, organization, or location. Such distinction is effective in finding some metaphorical usages of verbs. For example, Seto-naikai-wo idaku 澱戸内海を抱く (to embrace Seto Inland Sea) is metaphorical because idaku 抱く is interpreted as “to be located surrounding something” rather than “embrace” or “hold someone tight.” Section 5.5 discusses which language resources are used to retrieve features and how they are preprocessed.

4.4 Overall Score

We define overall score in two ways: through the arithmetic mean (AM) of similarity, rarity, and figurativeness, and through the harmonic mean (HM) of the same three scores. Precisely, when the source expression is $\varepsilon = (s, o, v)$, and an alternative verb is $v'$, the overall scores are calculated as follows.

$$os_{AM}(s, o, v, v') = \frac{1}{3} \left( \text{sim}(v, v') + \text{rar}(o, v') + \text{fig}(s, o, v') \right),$$

(11)

$$os_{HM}(s, o, v, v') = 3 \left( \frac{1}{\text{sim}(v, v')} + \frac{1}{\text{rar}(o, v')} + \frac{1}{\text{fig}(s, o, v')} \right)^{-1}.$$

(12)

We do not use them together, but we compare their performances in experiments. Miyazawa and Miyao (2017) defined overall evaluation as the arithmetic mean of the scores for comprehensibility, metaphoricity, and novelty. However, we exclude the score for comprehensibility, that is, commonness, from Equation 12, as we have no automatic score for comprehensibility, and using commonness leads to the mere cancellation of rarity. On the other hand, we include similarity in Equation 12, which was not used by Miyazawa and Miyao (2017) because their task was not paraphrasing. We use harmonic mean along with the arithmetic mean in expecting that the score becomes robust to outliers.
5 Dataset

In this section, we describe the resources used in the experiments and how we preprocessed them.

5.1 Input and Metaphor Classifier Datasets

We needed two datasets of Japanese SOV triples: one was inputted to the system as input, and the other was used to train a metaphor classifier for computing figurativeness scores. For these purposes, each verb required a label of metaphoricity, \textit{Met} or \textit{Lit}. We used a dataset by Miyazawa, Yoshida, and Miyao (2016), which was constructed on part of \textit{Kyoto University Text Corpus Version 4.0} (KU Corpus hereafter) (Kawahara, Kurohashi, and Hasida 2002). The KU Corpus has lexical and syntactic annotations on text taken from Japanese news articles. In addition, some sections of the corpus have semantic annotations such as referent and implicit subjects. This helped us construct SOV triples from Japanese sentences, as they often lack explicit subjects. We used this dataset rather than TroFi or other datasets for two reasons: First, metaphor criteria were clearer, which made it easy to analyze classification results. Second, the dataset contained more diverse expressions; TroFi focuses on literal and nonliteral uses of a limited number of verbs and only contains expressions related to them. As we explain below, Miyazawa et al. (2016) annotated based on MIPVU and referred to dictionary explanations. As a result, words were often labeled \textit{Met} even if they were so conventional that people do not take them as metaphorical. Such usage of words usually make for dull expressions, and they are not our main targets to generate. However, we had other automatic metrics, such as rarity, to filter out dull or inappropriate expressions from other perspectives, so we did not expect classifiers to filter out many expressions in advance. In MIPVU and Miyazawa et al. (2016), metaphoricity assessments did not explicitly depend on the conventionality of each usage, which met the policy.

Miyazawa et al. (2016) aimed to create a guideline for a Japanese metaphor corpus based on MIPVU. To check its validity, they annotated 764 OV pairs from the first part of fully-annotated sections of the KU Corpus. They assigned labels to both objects and verbs. In this paper, we just needed verbs annotation but also briefly review results including objects. After extracting 764 OV pairs, two annotators made judgments on metaphoricity according to a guideline based on a slightly modified version of MIPVU. One difference between MIPVU and Miyazawa et al. (2016) is that the latter used binary labels while the former allows annotators to assign a special label in ambiguous cases. Another difference was the dictionary: to refer to meanings of words, they used an electronic version of \textit{Shin Meikai Kokugo Jiten} (Yamada, Shibata, Sakai, Kuramochi, Yamada,
In the dictionary, metaphorical meanings are often marked as _hiyu-teki-ni_ 比喻的に (in a metaphorical sense). According to their guideline, the annotators labeled the word as _Met_ when the dictionary marked the contextual meaning as metaphorical.

To check the quality of the guideline, they calculated inter-annotator agreement (Fleiss’ $\kappa$), which is shown in Table 3. They had $\kappa$ of 0.477 and 0.587 for objects and verbs respectively. These were not high compared to the results of Steen et al. (2010), but they were high enough to recognize moderate agreement and thus we used this result in the later analysis. When the two annotators differed in judgment, we adopted the labels given by Annotator 1.

For our experiments, we removed some expressions from the dataset of Miyazawa et al. (2016) and obtained 202 metaphorical expressions (whose verbs are labeled _Met_) and 326 literal expressions (whose verbs are labeled _Lit_). Expressions that have labels on _koto_ こと (fact) and _mono_ もの (thing) were removed. This is because we cannot assign proper vector representations them. Our method assumes that subjects and objects are single nouns, but _koto_ and _mono_ are often preceded by a verb phrase and create such noun phrases as _eiga-wo miru koto_ 映画を見る (to watch a movie), for which it is difficult to acquire good representations. We also excluded objects containing numerals (e.g., “20%”) because sometimes it was challenging to link them to the quantifying nouns and vector representations. Finally, we removed expressions with labels on formal verbs: _suru_ する (to do), _motsu_ 有 (to have), and _tsukuru_ 作る (to make). However, _suru_ was not excluded when linked to _verbal nouns_, which are nouns that function as verbs by taking _suru_ at the end, such as _kouryo-suru_ 考慮する (to consider).

The dataset was divided into a paraphrase dataset and training dataset. The former consisted of nine expressions listed in Table 4, and the latter consisted of the rest. In Section 6.1, we compared the performance of two metaphor classifiers. The performance of classifier JA was measured by 10-fold cross validation on this training dataset while that of classifier EN was measured on the _TroFi Example Base_ (Birke and Sarkar 2006, 2007), which was used by Tsvetkov.

### Table 3

<table>
<thead>
<tr>
<th></th>
<th>Object</th>
<th>Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td># Met + # Lit</td>
<td>764</td>
<td>725</td>
</tr>
<tr>
<td># Met by Annotator 1</td>
<td>60 (7.9%)</td>
<td>214 (29.5%)</td>
</tr>
<tr>
<td># Met by Annotator 2</td>
<td>75 (9.8%)</td>
<td>205 (28.3%)</td>
</tr>
<tr>
<td>Fleiss’ $\kappa$</td>
<td>0.477</td>
<td>0.587</td>
</tr>
</tbody>
</table>

4 For example, they report $\kappa$ was 0.77 in the News register in the first trial.
Miyazawa and Miyao

Metrics to Generate Metaphorical Verb Expressions

Table 4  Source expressions in the paraphrase dataset.

<table>
<thead>
<tr>
<th>Expression</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 kai-ga giin-wo settoku-suru</td>
<td>(the group persuades a member of the Diet)</td>
</tr>
<tr>
<td>2 seifu-ga gidai-wo kentou-suru</td>
<td>(the government considers a subject for discussion)</td>
</tr>
<tr>
<td>3 kazoku-ga kyauka-wo tanoshimu</td>
<td>(the family enjoys holidays)</td>
</tr>
<tr>
<td>4 X-san-ga sakuhin-wo erabu</td>
<td>(someone chooses a piece of work)</td>
</tr>
<tr>
<td>5 kaisha-ga seihin-wo shukka-suru</td>
<td>(the company ships its products)</td>
</tr>
<tr>
<td>6 X-san-ga ronbun-wo keisai-suru</td>
<td>(someone makes his/her article carried (in a journal))</td>
</tr>
<tr>
<td>7 X-san-ga keikaku-wo shiru</td>
<td>(someone gets to know the project)</td>
</tr>
<tr>
<td>8 X-san-ga chousen-wo hajimeru</td>
<td>(someone starts a challenge)</td>
</tr>
<tr>
<td>9 X-san-ga bunka-wo mamoru</td>
<td>(someone saves a culture)</td>
</tr>
</tbody>
</table>

et al. (2014) as training data. The classifier for computing figurativeness scores for human evaluation (Section 6.2, Section 6.3) was trained on the entire training dataset.

5.2 Candidates for Alternative Verbs in the Generation

Alternative verbs in the generation were chosen from a verb set collected from the Balanced Corpus of Contemporary Written Japanese (BCCWJ). It is a large Japanese corpus with about 100 million words taken from a wide variety of registers including books, magazines, and online Q&A forums, for example. We also used it to obtain word vectors for similarity and rarity. By using the lexical annotation of the corpus, we collected all the verbs in the corpus, including combinations of verbal nouns and suru.

Since our targets are SOV triples, verbs needed to be transitive. We extracted them from the verbs in BCCWJ using a case frame dictionary, Kyoto University Case Frames (KU Case Frames hereafter) (Kawahara and Kurohashi 2006a, 2006b), which is a collection of predicate-argument examples taken from 1.6 billion sentences of online text. The top entries of the dictionary are indexed by verb, with each having examples of cases and complement nouns as subentries. We assumed that transitive verbs have at least one example of the accusative case, which is indicated by -wo. Collecting such examples, we obtain 10,743 transitive verbs including verbal nouns.

5.3 Similarity

We obtained the word vectors used in the calculation of similarity scores by training the

---

5 http://pj.ninjal.ac.jp/corpus_center/bccwj/en/
CBOV model\(^6\) on the sentences in BCCWJ. The size of vectors was set to 100, and the window size was set to 8.

### 5.4 Rarity

To calculate rarity, first, we clustered the BCCWJ words by applying spherical \(K\)-means\(^7\) to their vector representations constructed for similarity. The number of clusters \(K\) was 1,000. Before computing the score, we removed particles such as case markers from the text to put verbs adjacent to their respective objects. After that, we replaced each noun with the corresponding cluster tag, and finally we had \(p(v)\) and \(p(t, v)\) for each verb \(v\) and tag \(t\).

### 5.5 Figurativeness

The classifier EN and JA share some language resources but with some differences. They are summarized in Table 5.

We retrieved abstractness and imageability scores from the MRC Psycholinguistic Database\(^8\) (Wilson 1988). The classifier JA also used the dataset, but we translated words into Japanese using Japanese WordNet (Isahara, Bond, Uchimoto, Utiyama, and Kanzaki 2008). Japanese WordNet has links between words’ synsets (set of synonyms) and corresponding English words, so we can get English translations of Japanese words via these links.

We retrieved and encoded supersenses from WordNet the same way as in Tsvetkov et al. (2014) for both EN and JA.

To retrieve information on named entities, we did not use translation because the annotation of named entities was available in the KU Corpus. For the English dataset, we used the parsing result of spaCy\(^9\) with the \texttt{en\_core\_web\_sm} model.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Classifier EN</th>
<th>Classifier JA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstractness &amp; Imageability</td>
<td>MRC</td>
<td>MRC &amp; trans. by Japanese WordNet</td>
</tr>
<tr>
<td>Vector Representation</td>
<td>Huang et al. (2012)</td>
<td>CBOV applied to BCCWJ</td>
</tr>
<tr>
<td>Named Entity</td>
<td>spaCy (\texttt{en_core_web_sm})</td>
<td>KU Corpus</td>
</tr>
</tbody>
</table>

\(^6\) We used an implementation provided in the gensim library (https://radimrehurek.com/gensim/models/word2vec.html).
\(^7\) We used the spherecluster library (https://github.com/clara-labs/spherecluster).
\(^8\) http://ota.oucs.ox.ac.uk/headers/1054.xml
\(^9\) https://spacy.io/
Miyazawa and Miyao

As vector representations for EN, we used Huang, Socher, Manning, and Ng (2012) as Tsvetkov et al. (2013) did. Classifier JA used the vector representation constructed for computing similarity.

Both classifiers used gradient boosting decision tree as the classification algorithm. We used an implementation provided by the LightGBM library\(^{10}\) with default hyperparameters.

6 Experiment

We conducted three experiments. Experiment 1 measured the performance of the metaphor classifier by applying it to the Japanese test dataset. Experiments 2 and 3 used crowdsourcing. In Experiment 2, we examined whether the system’s scores actually correlated with scores provided by human evaluators. Experiment 3 was conducted to see how the system’s scores reflect human preferences.

6.1 Experiment 1

Experiment 1 measured the performance of the two metaphor classifiers, EN and JA. Performance was measured by 10-fold cross-validation. The results are shown in Table 6. Though the scores were not so high, JA achieved an accuracy greater than 0.5, implying that they work as metaphor classifiers. Thus, we used JA to compute figurativeness scores in the later experiments.

It is not directly comparative, but the performance was worse than that in Tsvetkov et al. (2014). They reported that their classifier achieved F-measures of 0.85 and 0.86 in English and Russian, respectively. These results are attributed to the difference in annotating schemes in the English and Japanese datasets.

6.2 Experiment 2

Experiment 2 was conducted to analyze the relation between human scores in the evaluation

<table>
<thead>
<tr>
<th></th>
<th># Lit</th>
<th># Met</th>
<th>A</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>1,166</td>
<td>1,716</td>
<td>0.44</td>
<td>0.39</td>
<td>0.84</td>
<td>0.53</td>
</tr>
<tr>
<td>JA</td>
<td>489</td>
<td>175</td>
<td>0.83</td>
<td>0.78</td>
<td>0.50</td>
<td>0.61</td>
</tr>
</tbody>
</table>

The columns # Lit and # Met show the numbers of instances of Lit and Met contained in each dataset, and A, P, R, and F stand for accuracy, precision, recall, and F-measure respectively.

\(^{10}\) https://github.com/Microsoft/LightGBM

293
metrics and system scores computed in the generation. We fed nine literal expressions to our system, and generating 10,743 expressions for each. We sampled the target expressions as described in Section 3.3. Thus, we had 10 expressions for each source expression and each score. We also evaluated the source expressions. Consequently, we had $9 \times 11 \times 4 = 396$ questions. For each question, we collected 10 evaluators. This means $N$ in Equation 1 was set to 10.

The results are shown in Fig. 4 and Table 7. The Underlines in Table 7 indicate that the

---

**Source Expression**

- kai-ga giin-wo settoku-suru (the group persuades a member of the Diet)
- sesfu-ga gidai-wo kentou-suru (the government considers a subject for discussion)
- kazoku-ga kyuuka-wo tanoshimu (the family enjoys holidays)
- X-san-ga saukin-wo erabu (someone chooses a piece of work)
- kusha-ga seihin-wo shukka-suru (the company ships its products)
- X-san-ga ronbun-wo keissai-suru (someone makes his/her article carried (in a journal))
- X-san-ga kessaku-wo shiru (someone gets to know the project)
- X-san-ga chouen-wo hajimeru (someone starts a challenge)
- X-san-ga bunka-wo mamaru (someone saves a culture)

---

*Fig. 4 Relationship between scores given by human and our system.*
automatic score of the row was introduced to model the evaluation metric in the column. Regarding synonymousness, metaphoricity, and novelty, the corresponding automatic scores, similarity, figurativeness, and rarity, have positive correlations and are the highest among the scores. This means that they work as models of the corresponding evaluation metrics. The results for comprehensibility were different from what we expected. The automatic score with the highest correlation coefficient was similarity (0.44) rather than negative rarity (0.37). However, this is natural because the original phrases are highly comprehensible, and similarity was measured in the sense that they often appear in similar contexts. For example, \textit{X-san-ga ronbun-wo kakiorosu} (someone newly writes an article) was rated 1.0 in the comprehensibility evaluation, while similarity and commonness were 0.75 and 0.91, respectively. This result suggests that we need to consider dependencies and a way to mix the scores to model each evaluation metric.

We observed that human judges tended to give high metaphoricity scores when expressions were incomprehensible.\footnote{Here we regard expressions “incomprehensible” based on our intuition rather than in terms of the comprehensibility scores because the target expressions are different in each metrics and expressions rated in terms metaphoricity do not have comprehensibility scores.} For example, \textit{kaisha-ga seihin-wo aogu} (the company looks up at its product) was rated 0.66 in metaphoricty but 0.10 in figurativeness. Opposite cases were also observed. \textit{Kazoku-ga kyuuka-wo furikaeru} (the family looks back on holidays) was rated 0.19 by human evaluators, but its figurativeness score was 0.89. The verb in the expression is metaphorical according to MIPVU, but it is so conventional that people usually do not consider it a metaphor.

As for novelty, we hypothesized that there were two reasons for the weak correlation. First, alternative verb candidates in the generation were sometimes incompatible with the object. For example, \textit{deru} (to get out), which appeared in \textit{X-san-ga giin-wo deru} (the group gets out of a member of the Diet) \textit{O○さんが議員を出る} (novelty: 0.78, rarity: 0.21), sounds novel or...
unnatural because it can be transitive only when its object stands for a place. The second reason is that the set of alternative verb candidates wrongly contains some verbs that are always intransitive, such as mieru 見える (to come into sight) in *X-san-ga bunka-wo mieru *○○さんが文化を見る. We regarded a verb as transitive when it had at least one example of -wo. However, we needed to adopt a more elaborate method because the KU Case Frames sometimes contains unnatural examples; it was constructed automatically, and its examples were collected from the Web. We expect to have more reliable results if we can remove such ungrammatical cases.

6.3 Experiment 3

Experiment 3 compared human preference and the overall score. We used the same input expressions as those in the previous experiment and recruited evaluators the same way. We collected binary comparison answers and constructed aggregate preference orders as described in Section 3.4. The evaluation was done using Kendall’s τ. It takes a value in [−1, 1], and a higher value means two orderings are more similar.

Table 8 shows τ calculated between orderings made by preference and the automatically computable scores including the overall scores, AM and HM. We cannot conclude that overall scores are consistent with human evaluation, but it helps us find preferable expressions in some cases. For example, X-san-ga chousen-wo erabu 〇〇さんが挑戦を埋める (someone buries a challenge) was ranked at the top by the system and human judges (similarity: 0.47, figurativeness:

<table>
<thead>
<tr>
<th>Source Expression</th>
<th>Sim.</th>
<th>Fig.</th>
<th>Rar.</th>
<th>AM</th>
<th>HM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  kai-ga giin-wo settoku-suru</td>
<td>0.24</td>
<td>−0.20</td>
<td>−0.24</td>
<td>−0.16</td>
<td>−0.24</td>
</tr>
<tr>
<td>2  seifu-ga gidai-wo kentou-suru</td>
<td>−0.02</td>
<td>−0.11</td>
<td>0.20</td>
<td>−0.11</td>
<td>0.51*</td>
</tr>
<tr>
<td>3  kazoku-ga kyuuka-wo tanoshimu</td>
<td>0.24</td>
<td>0.07</td>
<td>0.24</td>
<td>0.02</td>
<td>0.24</td>
</tr>
<tr>
<td>4  X-san-ga sakuhin-wo erabu</td>
<td>−0.42</td>
<td>−0.11</td>
<td>−0.16</td>
<td>−0.11</td>
<td>−0.02</td>
</tr>
<tr>
<td>5  kaisha-ga seihin-wo shukka-suru</td>
<td>−0.07</td>
<td>0.07</td>
<td>0.02</td>
<td>0.07</td>
<td>−0.11</td>
</tr>
<tr>
<td>6  X-san-ga ronbun-wo keisai-suru</td>
<td>−0.29</td>
<td>0.29</td>
<td>−0.07</td>
<td>0.24</td>
<td>−0.02</td>
</tr>
<tr>
<td>7  X-san-ga keikaku-wo shiru</td>
<td>0.20</td>
<td>−0.11</td>
<td>0.16</td>
<td>−0.02</td>
<td>0.20</td>
</tr>
<tr>
<td>8  X-san-ga chousen-wo hajimeru</td>
<td>−0.24</td>
<td>0.11</td>
<td>−0.16</td>
<td>0.11</td>
<td>0.07</td>
</tr>
<tr>
<td>9  X-san-ga bunka-wo mamoru</td>
<td>0.16</td>
<td>0.07</td>
<td>−0.07</td>
<td>0.20</td>
<td>0.11</td>
</tr>
</tbody>
</table>

The star (*) indicates that τ is significantly different from 0 at the 5% level.

12 The expression makes sense if we use a transitive verb miru 見る (to see) instead of mieru 見える (to come into sight) as X-san-ga bunka-wo miru 〇〇さんが文化を見る (someone sees the culture).
When we looked at mismatching examples, **kai-ga giini-wo haki-suru** 会が議員を破棄する (the group destroys a member of the Diet) was the most preferred by human evaluators, but it was ranked at the bottom by oSAM among the results from the same input (similarity: 0.57, figurativeness: 0.00, rarity: 0.73). The expression ranked at the top by oSAM was **X-san-ga chousen-wo kaku 〇〇さんが挑戦を書く** (someone writes a challenge) (similarity: 0.67, figurativeness: 0.34, rarity: 0.73). This inconsistency was due to significant differences in figurativeness. We expected that the harmonic mean would weaken the effect, but it did not work well. This observation reveals that the need to revise the definition of figurativeness so that they become more robust.

## 7 Conclusion and Future Work

We defined a new task for automatic metaphor generation as a paraphrasing task. To evaluate the results, we proposed two metrics—synonymousness and preference along with metaphoricity, novelty, and comprehensibility, which were proposed by Miyazawa and Miyao (2017). We introduced a model that finds alternative verbs for input expressions by using automatically computable scores corresponding to the metrics. By applying the method to a Japanese dataset, we showed that automatically computable scores were weakly correlated with human scores, and in some cases the overall score was effective in finding preferable expressions. On the other hand, it revealed that the definitions of the automatically computable scores need to be revised. Specifically, we need to solve two problems. First, we assumed that we can model each evaluation metric using a single automatic score, but this was proven to be invalid. For example, similarity correlated highly with both synonymousness and comprehensibility. Second, overall scores were not consistent with human scores. We simply used arithmetic and harmonic means, but we need to try many patterns to weight the scores. Making figurativeness more robust may help us improve overall scores.

We tested the task and method in a Japanese dataset, but it is also applicable to other languages with slight modifications, as metaphors are conceptual rather than lexical. For the same reason, our method would work in syntactic structures other than SOV such as a noun and its modifier. We will examine whether these are true by conducting experiments in other datasets. To make our system practical, we are going to evaluate how good generated results are with contexts or in specific purposes such as speeches or poems.
Reference


Miyazawa and Miyao

Metrics to Generate Metaphorical Verb Expressions

Linguistics.

Akira Miyazawa: He received his B.A. degree from Waseda University in 2012. He is currently a Ph.D. candidate at the Graduate University for Advanced Studies. His research interests include natural language processing, in particular, computational approaches to metaphor.

Yusuke Miyao: He received his Ph.D. from the University of Tokyo in 2006. He has been Assistant Professor at the University of Tokyo from 2001 to 2010, Associate Professor at National Institute of Informatics since 2010, and Professor at the University of Tokyo from 2018. He has been engaged in research on natural language processing, in particular on syntactic parsing and its applications.

(Received October 31, 2018)
(Revised February 18, 2019)
(Accepted March 2, 2019)