Evaluation Metrics for Automatically Generated Metaphorical Expressions

Akira Miyazawa†, ‡ Yusuke Miyao†, ‡
†, ‡The Graduate University for Advanced Studies / †, ‡National Institute of Informatics
{miyazawa-a, yusuke}@nii.ac.jp

Abstract

This paper proposes metrics to evaluate the quality of automatically generated metaphors not restricted to similes. The metrics are metaphoricity, novelty, comprehensibility, and overall evaluation. First, we discuss their importance and necessity. Next, we show that it is feasible to evaluate them by crowdsourcing. The targets of the evaluation are 1,360 expressions, each of which consists of a noun taken from a list of 40 nouns and a verbal phrase taken from a list of 34 verbal phrases. Then, we analyze the results to check the validity of the metaphoricity, novelty, and comprehensibility, and clarify their relationship. Finally, we argue that high-ranked expressions in the overall evaluation are considered to be "good metaphors" by showing that they actually are metaphors and are preferred by a human judge.

1 Introduction

In natural language processing (NLP), there are three main tasks that deal with metaphors: *detection*, *comprehension*, and *generation*. Generation has been studied less intensively than the others, but it has many applications. In poetry and prose, a metaphor is a tool that gives originality to works by helping writers avoid the banality of the everyday usage of the language (Leech, 2014, Ch. 2). Politicians use metaphor to make their statements more persuasive (Charteris-Black, 2011, Ch. 2). Generally, using metaphors makes language more visual or emotional (Mohammad et al., 2016). Thus, a metaphor generating system that suggests metaphorical expressions based on genres, purposes, or objects that we want to describe would be beneficial.

Most of the studies of metaphor generation have focused on similes of some fixed form such as "*T* like *S*". However, metaphor (in a broad sense) or *trope* has many other subclasses such as metaphor in a narrow sense, metonymy, or synecdoche. Thus, for example, previous works were not able to determine which is better: "his despair overflows" and "despair fills him". In this study, we propose metrics to evaluate metaphors not restricted to similes that help us find "good" metaphors from possible candidates. In the experiment, we calculated the scores of each metric by crowdsourcing. This allowed us to collect and analyze how general readers feel about expressions on a large scale. Preferably, metrics should be calculated automatically for objectivity and scalability. We expect that this is possible but do not discuss specific ways of achieving automation in this paper.

The rest of the paper is arranged as follows. Section 2 introduces works on automatic generation of similes, and their evaluation method. In Section 3, we introduce three metrics: *metaphoricity*, *comprehensibility*, *novelty*, and *overall evaluation*, which is calculated from the other three metrics. In Section 4, we check that it is feasible to evaluate expressions in terms of the metrics by conducting crowdsourcing. The target expressions in the evaluation are made by combining a noun taken from a list of 40 nouns and a verbal phrase taken from 34 verbal phrases. Section 5 shows the results of the experiment, analysis on the validity of the metrics, and their relationship. Finally some conclusions are given in Section 6.

2 Related Work

Existing works on metaphor generation have focused on similes such as "*T* like *S*" (Abe et al., 2006; Kitada and Hagiwara, 2001). Words used in the position of *S* are called the *source* or *vehicle*, and *T* is called the *target* or *topic*. Kitada and Hagiwara (2001) suggested a system that finds a word for the source that can be used in a given sentence. For example, given a sentence "The moon was red", the system outputs "The moon was as red as the setting sun". It uses some scores to select the source from the candidates and one of them is metaphoricity. This is calculated from the affective similarity and categorical dissimilarity of the target and each candidate for the source. In the evaluation, they asked volunteers who used the system to rate how good the generated sentences were as metaphors. Abe et al. (2006) proposed a model that finds suitable nouns for the source according to the properties that the target word has. For example, their model suggests "grandchild" for *S* in "a character like *S*", when given the list of properties "young, innocent and fine character". They evaluated the generated phrases in terms of *adequacy*, *ease of visualization*, *amusingness*, and *novelty*.

Because these systems generate metaphors by filling templates of similes, metaphoricity has been ignored in the evaluation process. A problem here is that judging metaphoricity in a systematic way is difficult. In corpus linguistics, researchers also needed such a method to annotate words that are used metaphorically. For that purpose, Steen et al. (2010) created a detailed guideline called *MIPVU*. We followed this guideline to make the gold standard. The basic 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 are 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*.

In this study, we use metaphoricity, novelty and comprehensibility. We introduce and describe this issue in the next section. Adequacy in the work of Abe et al. (2006) is regarded as the same as comprehensibility in this paper. Because adequacy and ease of visualization are close values in the work of Abe et al. (2006), we do not use ease of visualization. Similarly, amusingness and novelty have similar scores in the work of Abe et al. (2006). Thus, we integrate amusingness into novelty. Our work is also different in that the number of target expressions in evaluation is much larger[1]. This enables us to analyze the relationship among metrics more precisely.

3 Metrics

In this study, we propose three metrics: metaphoricity, comprehensibility, and novelty. The reason we use multiple metrics is that the importance of each metric differs from application to application. For example, poems, proses, and novels need creativity. For these, novelty is important. In the case where no important feature is selected, we also propose overall evaluation after introducing the other three metrics.

3.1 Metaphoricity

Metaphoricity measures how metaphorical an expression is. As stated in the preceding section, existing studies on metaphor generation have focused on similes of forms such as "*T* like *S*". Therefore, they have not paid much attention to whether the generated expressions are metaphorical. However, this matters a great deal in the generation of general metaphors because generated expressions are not necessarily metaphorical. To solve this problem, we measure metaphoricity of expressions on a five-point scale. The reason we do not classify them into two classes, metaphorical or nonmetaphorical, is that we need to compare metaphoricity with the other metrics to investigate their relationship. We collect metaphoricity scores by asking simple questions of crowdworkers. After that, we check if the results are reliable by using MIPVU. We did not ask workers to make judgments according to MIPVU because it is for experts.

[1]Those in the work of Kitada and Hagiwara (2001) and Abe et al. (2006) are 226 and 15 respectively.

3.2 Novelty

Next, we introduce *novelty* to measure how novel an expression looks or sounds. It is hard to define "good expressions" in general. However, creativity or originality can be an important criterion and novelty is a tool to generate creative or original expressions (Leech, 2014, Ch. 2). An advantage of using novelty is that it is easier to evaluate than creativity or originality. In the evaluation of novelty, an evaluator gives the highest score if he or she has never heard of or used the target expression, and the lowest score if the expression is widely used and conventional. It may be possible to measure novelty by counting the number of occurrences of expressions. However, in this study, we use crowdsourcing to measure novelty so that we can compare it with other metrics in the same conditions.

3.3 Comprehensibility

The third metric is *comprehensibility*. This quantifies how easy it is to understand the meaning of expressions. It is necessary because we use novelty as a metric and nonsense phrases (e.g., "she drinks sleep") tend to be highly novel. Comprehensibility serves as a constraint to keep the generated result meaningful.

This metric is more important for metaphors than for similes. Metaphors are often harder to understand in that they require that the writer and readers share some kind of similarity between the source and target. For example, readers would conceive the Japanese sentence "*ano kaisya made ensyou-sita*" (the fire spread to that company) in a metaphorical sense only when they have a detailed context or already know that the concept FIRE is used in contemporary Japanese to describe the situation where many people accuse someone of misbehaving over the Internet. On the other hand, similes usually do not require such knowledge, because readers try to find some kind of similarity between the source and target when they notice that the expressions are similes by their syntactic form.

3.4 Overall Evaluation

It is expected that some users of metaphor generation systems do not know which metric is important. For them, we introduce one integrated metric called *overall evalutation*. While it can be defined in many ways, for simplicity,we calculate it by just adding the three metrics.

4 Experiment

We conducted crowdsourcing to verify the validity of the metrics and clarify the relation among them[2].

4.1 Target

In this study, the targets of evaluation are short expressions, each of which consists of one noun and verbal phrase. For example, the expression "*zouo ga afureru*" (hatred spills out) is made up of "*zouo*" (hatred) and "*X ga afureru*" (*X* spills out). In this paper, we use the symbol X as a placeholder for a noun. We also use the symbol Y or Z as an anonymous subject or object in English translations to make the distinction between transitive and intransitive verbs clear or make them sound more natural. For example, we translate the original Japanese phrase "*X ni hitasu*" into "*Y* dips *Z* in *X*". The nouns are chosen from a list of 40 nouns. Most of them are related to emotion such as "*ai*" (love) or "*ikari*" (anger), while some unrelated nouns, such as "*neko*" (cat), are included for contrast. Verbal phrases are chosen from a list of 34 verbal phrases, and consequently we get $40 \times 34 = 1360$ expressions in total. All of the verbs in them are often used with "*mizu*" (water) and stand for physical actions such as "*X ga nagareru*" (*X* flows) rather than cognitive actions such as "*X ni tuite kangaeru*" (*Y* thinks about *X*).

Our method follows the method of Nabeshima (2011, Ch. 6). He made 336 expressions from 12 nouns and 28 verbs, and evaluated them in terms of the acceptability to examine the productivity and

[2]The result is available at `https://github.com/pecorarista/metaphor-evaluation-result`.

structural basis of *conceptual metaphors*: EMOTION IS WATER, WORDS ARE WATER, and MONEY IS WATER[3]. Each score of acceptability is the average of the scores given by 6 undergraduate students. The advantage of this method is that it can generate diverse expressions regarding metaphoricity, conventionality, or comprehensibility.

Our lists contain all the words that Nabeshima (2011) used. We added 28 nouns to the original list because it lacks words for specific types of emotion such as "*ai*" (love) while it includes more abstract words such as "*kanjou*" (emotion). We also added 6 verbal phrases that represent actions related to water and can be used metaphorically. For example, we added the phrase "*X wo kumitoru*" (*Y* scoops up *X*). It sometimes means "understand" or "consider" as in "*kimoti wo kumitoru*" (*Y* considers *Z*'s feelings).

4.2 Metrics

We asked workers to evaluate the metaphoricity, comprehensibility, and novelty of each expression on a scale of one to five. The choices for the lowest or highest score had short descriptions as the followings.

- **Metaphoricity**
  Do you feel that the expression is metaphorical?
  5. It seems to be metaphorical.
  1. It doesn't seem to be metaphorical.
- **Comprehensibility**
  Is the following expression easy to understand?
  5. I understand it without any problem.
  1. I don't understand it at all.
- **Novelty**
  Is the expression novel?
  5. It is so novel that I have never seen or heard it before.
  1. It is conventional and widely used.

4.3 Evaluator

We used the crowdsourcing platform *Yahoo! Crowdsourcing*, which is provided by Yahoo! Japan, to collect evaluators. Because the application form is written in Japanese, applicants are considered to be able to understand Japanese. We did not put any restriction on age, sex, or district of residence. However, each applicant was asked to pass a test to prove he or she was not a spammer. The test is to choose the correct part of speech of a given word. The answers given by applicants who failed the test are excluded from the results that we analyze later.

The total number of questions was 4,080, which is the product of the number of metrics and the number of expressions. We collected 10 workers for each question. The questions were divided into several *tasks*. A task is a collection of questions and a unit that crowdworkers apply. In a task, we asked each worker to answer 21 questions including one for the test. We restricted the number of applications of a worker to a task to one so that he or she did not answer the same questions multiple times. However, the restriction could cause a delay in collecting sufficient answers. Preparing multiple tasks solved the problems of duplicate answers and delay.

5 Analysis of Result

5.1 Analysis of Individual Metrics

By conducting the crowdsourcing, we got 10 scores for each expression and metric. After that, because Nabeshima (2011) evaluated the acceptability on a scale of 0 to 4, we subtracted 1 from every score so that we could compare the scores of comprehensibility to those of acceptability directly. Then, we calculated the average score of each expression and metric. We use this result in the following analysis.

[3]"*T* IS/ARE *S*" denotes a conceptual metaphor that maps a concept *S* to another concept *T*. See Lakoff (1993) for the details of conceptual metaphors.

Rank	Noun (X)	Verbal phrase	Score
1	*kotoba* (word)	*X ga huttou-suru* (*X* boils)	3.9
2	*kanjou* (emotion)	*X ni oboreru* (*Y* almost drowns in *X*)	3.8
3	*zetubou* (despair)	*X ga ahureru* (*X* overflows)	3.7
4	*oto* (sound)	*X ga simiru* (*X* soaks into *Y*)	3.6
5	*zetubou* (despair)	*X ga koboreru* (*X* spills out)	3.5
		⋮	
1356	*koe* (voice)	*X wo kakeru* (*Y* sprays *X*)	0.0
1356	*mizu* (water)	*X wo susuru* (*Y* sips *X*)	0.0
1356	*mizu* (water)	*X ga huttou-suru* (*X* boils)	0.0
1356	*mizu* (water)	*X ga nagareru* (*X* flows)	0.0
1356	*mizu* (water)	*X wo nomu* (*Y* drinks *X*)	0.0

Table 1: High-ranked and low-ranked expressions in terms of metaphoricity.

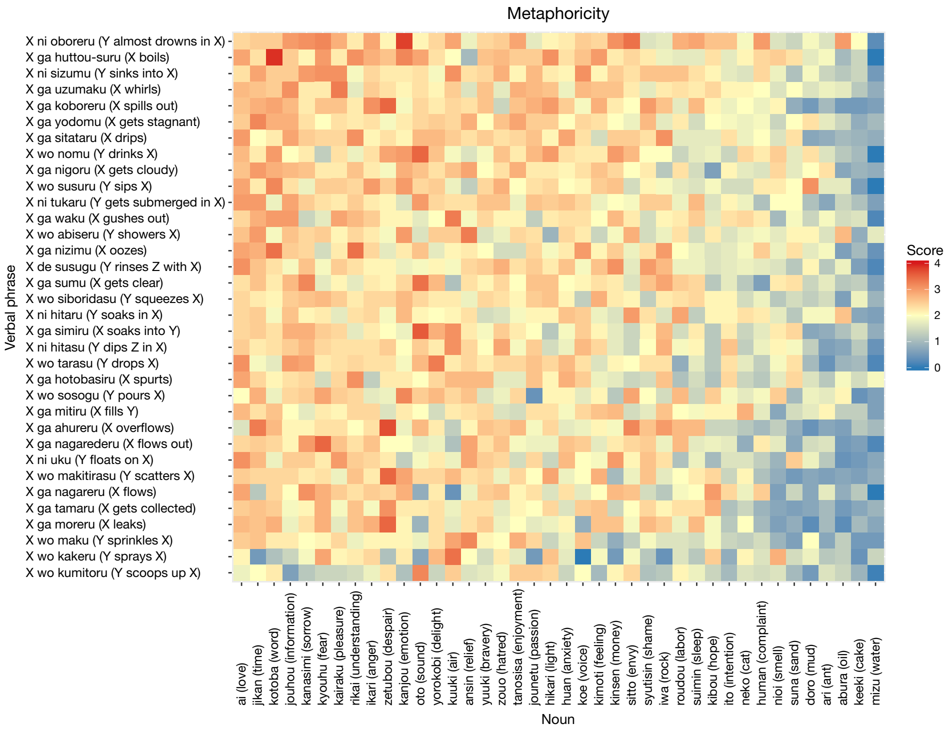

Figure 1: Scores of metaphoricity visualized in heat map. Nouns and verbal phrases are arranged in descending order of column- and row-wise sums of scores respectively.

First, we look at metaphoricity. As shown in Table 1, high-ranked expressions tend to contain a noun related to an emotion such as "*zetubou*" (despair). The complete ranking is available at the repository. It shows that 4 of the 10 expressions use "despair". On the other hand, most of the low-ranked expressions (9 of the bottom 10 expressions) consist of the noun "*mizu*" (water). Since each verb stands for a physical action related to water, they are not metaphorical. To check that high-ranked expressions were actually metaphorical, an author judged the top 10 expressions for their metaphoricity. The criterion is that an expression is metaphorical if its verb is a metaphor-related word in the sense of MIPVU. In the judging process, he used *Shin Meikai Kokugo Jiten* (a Japanese dictionary) to find basic and other meanings. As a result, 8 of the 10 expressions were metaphorical. He could not make judgments on the two expressions, "*kotoba ga huttou-suru*" (words boil)[4] and "*zetubou ga koboreru*" (despair spills out), because it was hard to understand their meanings.

Next, we examine the result of comprehensibility. Contrary to the case of metaphoricity, many high-ranked expressions consist of "*mizu*" (water) such as "*mizu de susugu*" (*Y* rinses *Z* with water) as partially

[4] We sometimes add an article "a(n)" or "the", or suffix "-(e)s" to the noun in translation to make it sound more natural; the number of the noun is usually not expressed explicitly in Japanese.

Rank	Noun (*X*)	Verbal phrase	Score
1	*ai* (love)	*X ni oboreru* (*Y* almost drowns in *X*)	4.0
1	*kanjou* (emotion)	*X wo kumitoru* (*Y* scoops up *X*)	4.0
1	*mizu* (water)	*X de susugu* (*Y* rinses *Z* with *X*)	4.0
1	*mizu* (water)	*X wo kakeru* (*Y* sprays *X*)	4.0
1	*mizu* (water)	*X wo susuru* (*Y* sips *X*)	4.0
		⋮	
1356	*ari* (ant)	*X ga simiru* (*X* soaks into *Y*)	0.0
1356	*keeki* (cake)	*X wo sosogu* (*Y* pours *X*)	0.0
1356	*iwa* (rock)	*X wo susuru* (*Y* sips *X*)	0.0
1356	*ikari* (anger)	*X wo susuru* (*Y* sips *X*)	0.0
1356	*zouo* (hatred)	*X de susugu* (*Y* rinses *Z* with *X*)	0.0

Table 2: High-ranked and low-ranked expressions in terms of comprehensibility.

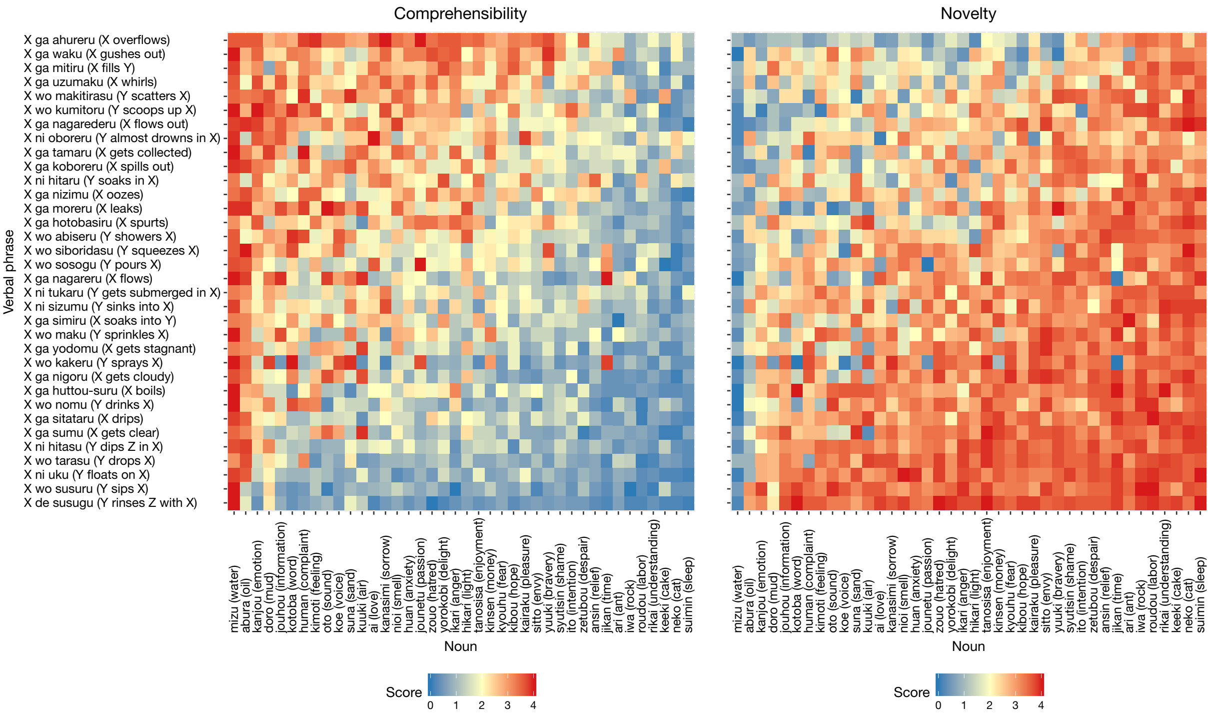

Figure 2: Scores of comprehensibility and novelty visualized in heat maps. Nouns and verbal phrases are arranged in descending order of column- and row-wise sums of comprehensibility.

shown in Table 2. On the other hand, low-ranked expressions tend to contain concrete objects (e.g., "ant" or "cake"). To confirm that our method actually captures how easy it is to understand the expression, we compare it to the acceptability reported in Nabeshima (2011). It is expected that the scores are close and show a similar tendency. First, when dealing with the common 336 expressions, we calculate the average of absolute differences of corresponding scores. The result is 0.64. It is less than one step of the scale of evaluation. Thus, it is considered to be minor. Next, we calculate the correlation coefficient between comprehensibility and acceptability. It is as high as 0.81 and means that they have a high positive correlation. As a result, we conclude that our evaluation on comprehensibility is as confident as that of Nabeshima (2011). The left graph of Figure 2 illustrates the scores of the expressions. Though nouns and verbal phrases were arranged by column- and row-wise sum, they form a dapple pattern. This implies that synonymous expressions have different grades of comprehensibility. For example, while "*kyouhu wo makitirasu*" (*Y* (disorderly) scatters fear) obtained 2.9, "*kyouhu wo maku*" (*Y* sprinkles fear) obtained 1.6. Analyzing this kind of discord is essential to examine the systematicity of metaphor and we succeeded in collecting the data for that purpose in a scalable and controllable way. We do not analyze such patterns in this paper, but will analyze them with the results in other domains.

Rank	Noun (*X*)	Verbal phrase	Score
1	*ari* (ant)	*X ga simiru* (*X* soaks into *Y*)	4.0
1	*neko* (cat)	*X wo siboridasu* (*Y* squeezes *X*)	4.0
1	*neko* (cat)	*X ga nagarederu* (*X* flows out)	4.0
1	*nioi* (smell)	*X ni uku* (*Y* floats on *X*)	4.0
1	*roudou* (labor)	*X wo susuru* (*Y* sips *X*)	4.0
		⋮	
1353	*mizu* (water)	*X ga nagareru* (*X* flows)	0.0
1353	*mizu* (water)	*X ga waku* (*X* gushes out)	0.0
1353	*mizu* (water)	*X wo nomu* (*Y* drinks *X*)	0.0
1353	*jikan* (time)	*X wo kakeru* (*Y* sprays *X*)	0.0
1353	*kotoba* (word)	*X wo kakeru* (*Y* sprays *X*)	0.0

Table 3: High-ranked and low-ranked expressions in novelty.

Rank	Noun (*X*)	Verbal phrase	Score
1	*kuuki* (air)	*X ni sizumu* (*Y* sinks into *X*)	8.9
2	*kimoti* (feeling)	*X ga huttou-suru* (*X* boils)	8.8
3	*kyouhu* (fear)	*X ga nagarederu* (*X* flows out)	8.7
4	*kanjou* (emotion)	*X ga huttou-suru* (*X* boils)	8.6
4	*syuutisin* (shame)	*X ga koboreru* (*X* spills out)	8.6
		⋮	
1356	*keeki* (cake)	*X wo sosogu* (*Y* pours *X*)	4.0
1356	*mizu* (water)	*X ga huttou-suru* (*X* boils)	4.0
1356	*mizu* (water)	*X ga nagareru* (*X* flows)	4.0
1356	*mizu* (water)	*X wo nomu* (*Y* drinks *X*)	4.0
1360	*koe* (voice)	*X wo kakeru* (*Y* sprays *X*)	3.8

Table 4: High-ranked and low-ranked expressions in overall evaluation.

Finally, we look at novelty. Table 3 shows the opposite tendency. That is, many high-ranked expressions contain concrete objects, while many low-ranked expressions contain "*mizu*" (water). This is shown visually in Figure 2.

5.2 Relationships

To analyze the relationships among the three metrics, we calculate the correlation coefficients. The result is shown in Table 5. It reveals a strong negative correlation between comprehensibility and novelty. This is natural because we expect that comprehensible expressions get used more often and are more conventional. Thus, it may be possible to integrate one into the other. However, there are some expressions that achieve high scores in both comprehensibility and novelty such as "*human wo nomu*" (*Y* suppresses complaints; literally, *Y* drinks or swallows complaints) (comprehensibility: 3.3; novelty: 2.6) or "*syuutisin ga waku*" (shame gushes out) (comprehensibility: 3.1; novelty: 2.4). These are preferable when we want to generate expressions that are creative as well as comprehensible. Consequently, both comprehensibility and novelty are needed to retrieve such expressions.

5.3 Overall Evaluation

The high-ranked and low-ranked expressions in the overall evaluation are shown in Table 6. In high-ranked expressions, there are several expressions that are not common but are understandable. For example, we can understand "*kuuki ni sizumu*" (*Y* sinks into the air) by interpreting "the air" as "an atmosphere" or "an ambience", and "*Y* sinks into" as "*Y* gets depressed in". Similarly, we can comprehend the meaning of the expression "*kanjou ga huttou-suru*" (emotion boils) if we assume that "emotion" stands for a specific type of emotion such as "anger", "hatred", or "excitement". In Japanese, there are many idioms that describe such kinds of emotion by using words related to fire.

To examine the validity of the overall evaluation, we check if high-ranked expressions are actually "good metaphors". We define "good metaphors" as expressions that are metaphorical and "good", and

	Metaphoricity	Comprehensibility	Novelty
Metaphoricity	1.0	−0.19	0.28
Comprehensibility	−0.19	1.0	−0.92
Novelty	0.28	−0.92	1.0

Table 5: Correlation coefficients between metrics.

High-ranked expression [rank in overall evaluation]	Low-ranked expression [rank in overall evaluation]	Match
human wo nomu (*Y* drinks complaints) [23]	*abura wo kumitoru* (*Y* scoops up oil) [1087]	✓
ikari ga koboreru (anger spills out) [6]	*iwa ni oboreru* (*Y* almost drowns in a rock) [1117]	✓
syuutisin ga tamaru (shame gets collected) [44]	*syuutisin wo sosogu* (*Y* pours shame) [856]	✓
jouhou ga nigoru (information gets cloudy) [106]	*kuuki X wo makitirasu* (*Y* scatters the air) [212]	✓
kanasimi ga simiru (sorrow soaks into *Y*) [32]	*rikai ga nagareru* (understanding flows) [721]	✓
tanosisa ga uzumaku (enjoyment whirls) [81]	*human ni tukaru* (*Y* gets submerged in complaints) [1241]	–
kotoba ga nizimu (words ooze) [14]	*kyouhu ga nagareru* (fear flows) [307]	–
kanjou wo sosogu (*Y* pours emotion) [44]	*ito X ni tukaru* (*Y* gets submerged in intention) [654]	✓
huan ga nagarederu (anxiety flows out) [44]	*jounetu wo kumitoru* (*Y* scoops up passion) [165]	✓
jouhou ni oboreru (*Y* almost drowns in information) [23]	*abura ga tamaru* (oil gets collected) [1241]	✓

Table 6: Result of human evaluation. Column "Match" is checked if high-ranked expression is preferred by judge.

we define "good" as "making us more inclined to use". The process of examination was as follows. First, we divided all the expressions into two groups: the top 10% and bottom 90% in the overall evaluation. Then, we randomly picked one expression from each group and made 10 pairs. For evaluation, we asked a volunteer to choose the one from each pair that he prefers without caring about whether it is metaphorical. This volunteer is a graduate student in NLP, native speaker of Japanese, and has a basic knowledge of linguistics. In addition, we changed the border to 20%, 30%, 40%, and 50% to find the effective range.

The result of 10% is shown in Table 6. The volunteer preferred 8 high-ranked expressions in 10 pairs. We regard the overall evaluation as valid in that high-ranked ones are preferred in most cases. Changing the boundary to 20%, 30%, 40%, and 50% made only small changes in the number: 6, 6, 6, and 7. In the cases where the low-ranked ones are preferred, four cases had the nouns related to emotion only in the low-ranked ones. Three cases contained the high-ranked expressions that are hard to understand; the noun "rock" is used with incompatible verbal phrases. To make the preferred ones highly ranked in these cases, taking the abstractness or concreteness of the nouns into the overall evaluation would be effective.

Finally, an author judged if high-ranked expressions are actually metaphorical. Table 6 shows high- and low-ranked expressions divided by 10%. He followed MIPVU in making judgments in the same way as he did in checking the validity of metaphoricity. As a result, all but six expressions are metaphorical. Three of them are nonmetaphorical: "*nioi wo kakeru*" (*Y* sprays smell), "*abura ni sizumu*" (*Y* sinks into oil), and "*iwa wo nomu*" (*Y* drinks oil). The rest of the expressions are nonsense: "?*iwa ga sitataru*" (a rock drips), "?*iwa ga nizimu*" (a rock oozes), and "?*suna ga sitataru*" (sand drips). All the expressions that are not metaphorical use concrete nouns. Thus, the situation will be improved by taking concreteness/abstractness into consideration in the overall evaluation.

In total, 33 of the 50 high-ranked expressions in Table 6 are preferred. Moreover, 44 of the 50 high-ranked expressions are metaphorical. Consequently, we conclude that the overall evaluation is valid in finding "good metaphors".

6 Conclusion

In this study, we proposed metrics to evaluate automatically generated metaphors not restricted to similes. Then, we actually evaluate expressions by crowdsourcing. The analysis of the result revealed the validity of the metrics and their relationship. Finally, we confirmed that high-ranked expressions in the overall evaluation are good metaphors. In future, we will apply the evaluation to an automatic metaphor generation system to help writers.

References

Abe, K., K. Sakamoto, and M. Nakagawa (2006). A computational model of metaphor generation process. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, pp. 937–942.

Charteris-Black, J. (2011). *Politicians and rhetoric: The persuasive power of metaphor*. Springer.

Kitada, J. and M. Hagiwara (2001). Figurative composition support system using electronic dictionaries (in Japanese). *Transactions of Information Processing Society of Japan 42*(5), 1232–1241.

Lakoff, G. (1993). The contemporary theory of metaphor.

Leech, G. N. (2014). *A linguistic guide to English poetry*. Routledge.

Mohammad, S. M., E. Shutova, and P. D. Turney (2016). Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*Sem)*, Berlin, Germany.

Nabeshima, K. (2011). *Nihongo no metafā (in Japanese)*. Kurosio Publishers.

Steen, G. J., A. G. Dorst, J. B. Herrmann, A. Kaal, T. Krennmayr, and T. Pasma (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins Publishing.