Disambiguation of Japanese Onomatopoeias using Nouns and Verbs

Hironori Fukushima[1], Kenji Araki[1], and Yuzu Uchida[2]

[1] Hokkaido University, Sapporo, Japan
{hukupoyo, araki}@media.eng.hokudai.ac.jp
[2] Hokkai Gakuen University, Sapporo, Japan yuzu@hgu.jp

Abstract. Japanese onomatopoeias are very difficult for machines to recognize and translate into other languages due to their uniqueness. In particular, onomatopoeias that convey several meanings are very confusing for machine translation systems to distinguish and translate correctly. In this paper, we discuss what features are helpful in order to automatically disambiguate the meaning of onomatopoeias that have two different meanings. We used nouns, adjectives, and verbs extracted from sentences as features, then carried out a machine learning classification analysis and compared the accuracy of how well these features differentiate two meanings of ambiguous onomatopoeias. As a result, we discovered that employing a combination of machine learning with nouns and verbs as a feature achieved accuracy of above 80 points. In addition, we were able to improve the accuracy by excluding pronouns and proper nouns and also by limiting verbs to those that are modified by onomatopoeias. In future, we plan to concentrate on dependency between verbs that are modified by onomatopoeia and nouns, as we believe that this approach will help machine translation to translate Japanese onomatopoeias correctly.

1 Introduction

In 2020, the Tokyo Olympics [1] will be held and many non-Japanese speakers will visit Japan. Not only these visitors, but many other non-Japanese will face a necessity to understand Japanese during their stay due to the lack of translated information in other languages. However, Japanese is one of the most difficult languages to understand because of unique expressions such as onomatopoeia [2]. There are even some onomatopoeias with multiple meanings. In addition, although these words must not be ignored in order to understand the meaning of Japanese more clearly and precisely, such onomatopoeias are very difficult to translate into other languages and recognize using machines because the meaning differs according to the context [3]. Therefore, it is very important to distinguish the meanings of onomatopoeias. In our study, we conducted a consideration of which features need to be extracted from a sentence in order to distinguish the meanings of onomatopoeias with two meanings. Furutake [4] et al. proposed a method for an automatic acquisition system of alternative expressions for onomatopoeia. Their method paraphrases an onomatopoeia as an adjective typically modified by a verb which is typically modified by the onomatopoeia. The meaning of the onomatopoeia greatly depends on the verb which

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 141–149, 2014.
This is preprint prepared by Proceedings editor for Springer International Publishing Switzerland.

it modifies, as demonstrated by the fact that the accuracy of this method achieved 80.6 points. Moreover, Uchida et al. [5] showed that the most often used parts of speech which the onomatopoeia modifies are commonly verbs, nouns, and adjectives, in descending order. Therefore, these parts of speech are very useful for distinguishing the meanings of onomatopoeias.

With the recent proliferation of social media, there are many short sentences and colloquial expressions on the Web. As a result, it can be assumed that there will be cases of Japanese sentences that have no verbs, or where it is difficult to find the part of speech that the onomatopoeia modifies. An example Tweet in Japanese, "*(Yokatta-, yuki sarari(*´□ '*.)*"[3] , meaning "I'm glad the snow is flowing" has no verbs according to the result of morphological analysis by MeCab [7], although the meaning "touch, feel" can be understood due to the existence of "*sarari*" with the noun "*yuki*", meaning "snow". Consequently, words which onomatopoeia do not modify can be useful as features for extraction, and their effectiveness should be considered. In our study, we extract verbs, nouns, and adjectives from sentences which include an onomatopoeia with two meanings. We extract these words as features and conduct machine learning using SVM-Light [8] to measure the accuracy of semantic disambiguation. Moreover, we subsequently changed the range of extracted nouns and used verbs modified by onomatopoeias as features to improve accuracy.

2 Experiment Preparation

2.1 Defining target onomatopoeia

First, we extracted onomatopoeias and mimetic words which have multiple meanings from 'Basic Word Usage Dictionary for Foreigners' [9]. These can be classified into two types depending on whether or not the parts of speech are the same when the onomatopoeia is used in each meaning. When the onomatopoeia has meanings that differ according to the part of speech, we need only distinguish the part of speech of the onomatopoeia in order to disambiguate the meanings. The meanings of such onomatopoeias are distinguished easily. Therefore, in our study, we define onomatopoeias with meanings that are difficult to distinguish as being "onomatopoeias that are used as the same part of speech in two different meanings". An example of this is "*garari*". When we use "*garari*" as an adverb, it has the following two meanings: (1) Opening the door vigorously. (e.g., "*Genkan no to wo "garari" to akeru*", He open the front door "*garari*".), (2) Changing an attitude suddenly. (e.g., " *Hanashi no tochuu de kareno taido ga "garari" to kawatta.*", His attitude changed "*garari to*" in the middle of talking.)

We found 21 onomatopoeias with two meanings which match the above definition in 'Basic Word Usage Dictionary for Foreigners'. This research is based on the concept that onomatopoeias that are used widely and regularly should be studied. Therefore, we targeted the two onomatopoeias that rank highest in terms of number of hits in a Google search [10]. As a result, "*sarari to*" and "*gatan to*" were selected. Both of these consist of the particle "*to*" and an onomatopoeia. However, it was found that onomatopoeias

[3] Extracted from Twitter [6]

often appear in blogs accompanied by the particle "*to*" [11]. In order to reduce errors in morphological analysis we retained the particle "*to*" in this research, using these onomatopoeias in the forms of "*gatan to*"and "*sarari to*".

2.2 Definition of meanings of onomatopoeias

It is necessary to provide original definitions of the meanings of "*sarari to*" and "*gatan to*", because there are some sentences that are impossible to classify using only the meanings described in 'Basic Word Usage Dictionary for Foreigners'.

The meanings of "*sarari to*" and "*gatan to*" that we defined are shown in Table 1 and 2. we define "Meaning 1 " as "m1" and "Meaning 2 " as "m2".

Table 1. The meanings of "*sarari to*" that we defined

Meaning 1 (m1)	**A touch, a feel, a flavor.** Wind blowing "*sarari to.*"(English) *"sarari to" shita kaze ga huku.*(Japanese)
Meaning 2 (m2)	**With good grace, good decisiveness.** He said difficult things "*sarari to*".(English) *Kare ha muzukashii koto wo "sarari to" itta.*(Japanese)

Table 2. The meanings of "*gatan to*" that we defined

Meaning 1 (m1)	**Appearance situations or sounds of heavy things faling or crushing heavy things, crush heavy things.** Things fall "*gatan to*" from a shelf.(English) *tana kara monoga "gatan to" ochiru.*(Japanese)
Meaning 2 (m2)	**A sudden drop in price, performance etc.** Sales fell "*gatan to*".(English) *uriage ga "gatan to" ochiru.*(Japanese)

2.3 Collection of sentences that include onomatopoeia

We collected sentences from Ameba blog [12] and Twitter [6] due to the assumption that such blogs contain a high number of onomatopoeia that are difficult to distinguish, and have many colloquial and everyday expressions. Any noisy sentences were removed manually. Example sentences which were removed are shown in Table 3.

Example 1 contains sentences consisting only of onomatopoeias. Example 2 is a sentence which can not be generally understood. Example 3 contains an onomatopoeia used as a proper noun in the sentence. In Example 4, features of the string which construct the onomatopoeia are repeated directly before or after the onomatopoeia.

Table 3. Examples of manually removed sentences

Ex1	*"Gatanto."*, *"Sarari sarari to."*
Ex2	*Nama de mireru nante sasu "ga tanto"n san.* - I'm proud to see Tanton in real life (English)
Ex3	*"Sarari to" shita umeshu* - *Sarari to* plum liquor. (Product name)
Ex4	*Otonari no akishitsu kara "gatan gatan" to to ga aku oto.* - The sound of the door opening in the vacant room next to mine.(English)

2.4 Constructing correct data

We constructed a corrected version of the data to input in SVM-Light. We asked ten Japanese participants to judge which meaning is correct for each onomatopoeia in each sentence. We determined a meaning to be correct if eight or more people gave the same answer. Sentences with "*sarari to*" were checked by ten participants, consisting of five women and five men. Sentences with "*gatan to*" were checked by another ten participants, consisting of six women and four men. We collected approximately 200 sentences for each onomatopoeia. Of the "*sarari to*" sentences, we collected 78 sentences with meanings of m1 (defined in **2.2**), and 124 sentences with meanings of m2. Of the "*gatan to*" sentences, we collected 100 sentences each of m1 and m2. The number of sentences given correct meanings for "*sarari to*" is 70 sentences as m1, and 119 as m2. For "*gatan to*" 98 sentences were given correct meanings as m1, and 94 as m2. There were 13 sentences which more than three people judged to have different meanings among 202 sentences featuring "*sarari to*". The rate of occurrence is 6.4 percent. On the other hand, there were eight such sentences among 200 sentences containing "*gatan to*", an occurrence of 4.0 percent. As a result, it was demonstrated that there are cases of different interpretations of the meanings of Japanese onomatopoeia even among native Japanese speakers.

3 Meanings Distinction Experiment

3.1 Onomatopoeia semantic disambiguation method

Firstly, we conducted morphological analysis on the sentences using MeCab [7] and extracted verbs, nouns, and adjectives. Secondly, we constructed the data in sets of words (verbs, nouns and adjectives) and the number of occurrences. Finally, we classified the meanings of the onomatopoeias using SVM-Light. We compared the accuracy using 10-fold cross-validation, as the amount of data is small. There are seven patterns of features in SVM-Light : only verbs; only nouns; only adjectives; nouns and adjectives; verbs and adjectives; nouns and verbs; and nouns, verbs and adjectives. Our definition of accuracy is as the equation (1). The meaning given is the correct meaning.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

TP = The meaning given “m1” and its output is “m1”
FP = The meaning given “m2” and its output is “m1”
TN = The meaning given “m2” and its output is “m2”
FN = The meaning given “m1” and its output is “m2”
(“m1” and “m2” are defined in **2.2**)

3.2 Result of experiment

We performed 10-fold cross-validation twelve times and calculated the average after omitting the highest and lowest values for each onomatopoeia. The calculated categorie are accuracy(defined in **3.1** (1)). These results are shown in Table 4.
n : noun, v : verb, a : adjective.

Table 4. Accuracy for “*sarari to*” and “*gatan to*”

POS	*“sarari to”*[%]	*“gatan to”*[%]
n	75.66	75.70
a	67.06	52.60
v	79.81	72.59
n and a	76.07	74.10
n and v	83.50	80.50
a and v	78.09	73.32
n,a and v	83.66	80.77

3.3 Consideration of results

For both “*sarari to*” and “*gatan to*”, the feature pattern of “nouns, adjectives and verbs” gave the highest accuracy. The feature pattern of “nouns and verbs” gave the second highest accuracy. The difference in accuracy between “nouns, adjectives and verbs” and “nouns and verbs” is very small. Therefore, we demonstrated that adjectives have poor potential for improving accuracy. On the other hand, we found some cases where adjectives did make a contribution to improving accuracy. For example, “nouns and adjectives” has higher accuracy than “nouns” in the results for “*sarari to*”. Thus, adjectives cannot be said to be completely noisy. This fact is related to the frequency of co-occurrence with the onomatopoeia. Furthermore, when we compare the features “verbs”, “nouns” and “adjectives”, “verbs” has the highest accuracy for “*sarari to*” , whereas “nouns” has the highest accuracy for “*gatan to*”. Therefore, it was demonstrated that in the semantic disambiguation of onomatopoeia, the most significant parts of speech are dependent on the type of onomatopoeia. The conclusion of our consideration is that “nouns” and “verbs” are highly significant for distinguishing meanings of onomatopoeia, and combining them as features makes a major contribution to semantic disambiguation.

4 Improvement of Features

4.1 Method of improvement

Firstly, we conducted a consideration of verbs. From related research, we know that the verbs that the onomatopoeia modifies are very important in order to distinguish meanings. Therefore, we extracted only verbs that the onomatopoeia modifies. We used CaboCha [13] to determine dependency relations. Furthermore, we extracted only verbs that belong to independent words using MeCab with the exception of suffixes. A suffix includes the “*ori*” of “ *orimasu*” or “*i*” of “ *imasu*” (English: both forms of “to be”) . Secondly, we conducted a consideration of nouns. There are proper nouns, numerals, and pronouns in sentences. For example, names such as “*Ogawa*” and “*Darvish*”, and the pronoun “*watashi*” (“I”) . These belong to the category of nouns, but they do not affect semantic disambiguation. Therefore, we removed them. Moreover, nouns which belong to the category of non-independent words were also removed. For example, “*suru*” of “*~suru koto ni natta*” (English: “(I) ended up doing ”) . Therefore, as a method of improvement, we extracted nouns belonging to the categories of “general”, “sahen-connection”, “nai-adjectivestem” and “quoted string”. In this way, we improved the features selection method, and performed 10-fold cross-validation again.

4.2 Results and consideration after improving the features

As the purpose of this study is improvement of accuracy(defined in **3.1** (1)), we compared the results of experiments after improving the features selection method with the previous results. These results and a comparison shown in Table 5, Table 6, Table 7 and Table 8. n : features are nouns. v : features are verbs. n’: features are only nouns, with the exception of proper nouns and pronouns, etc. v’: features are only verbs modified by the onomatopoeia.

Table 5. Comparison of “nouns”

POS	“*sarari to*”[%]	“*gatan to*” [%]
n	75.66	75.70
n’	72.70	80.73
Change	-2.96	5.03

Table 6. Comparison of “verbs”

POS	“*sarari to*”[%]	“*gatan to*”[%]
v	79.81	72.59
v’	85.23	81.67
Change	5.42	9.08

Table 7. Results after improving features

POS	“*sarari to*”[%]	“*gatan to*”[%]
n’	72.70	80.73
v’	85.23	81.67
n’ and v’	83.87	82.52

Table 8. Combination of nouns and verbs

POS	“*sararito*”[%]	“*gatanto*”[%]
n and v	83.50	80.50
n’ and v’	83.87	82.52
Change	0.37	2.02

Firstly, we will discuss nouns. For "*gatan to*", accuracy increased by 5.03 points. On the other hand, for "*sarari to*" accuracy decreased by 2.96 points. When we change the types of extracted nouns, there are 19 sentences that have no nouns for "*sarari to*", and nine sentences without nouns for "*gatan to*". These sentences were removed from the data. Therefore, the target sentences for "*sarari to*" were reduced to 170 sentences, and "*gatan to*" to 183 sentences. The number of distinct nouns that appeared was decreased to 327 from 456 for "*sarari to*", and to 341 from 500 for "*gatan to*". It can be considered that the frequency of occurrence of proper nouns and pronouns was an influential factor on the increased rate of accuracy. For example, the pronoun "*watashi*" ("I") appeared in six sentences of 183 sentences for "*sarari to*", whereas it appeared in only three sentences of 170 sentences for "*gatan to*". This difference is about 1.5 points. This affected the increased rate of accuracy. Moreover, failures of morphological analysis affected the decreasing rate of accuracy. For example, "*Darvish ga watashi no me no mae wo "sarari to" tootte kandou* " (I was impressed when Darvish (a baseball player) "*sarari to*" passed right in front of me). The extracted words from this sentence are "*Darvish*", "*me*", "*kayou*" and "*kandou*". The point of interest here is "*Darvish*". The word "*Darvish*" belongs to the category of "proper nouns" , but it was analyzed by MeCab in this sentence as belonging to the category of "general nouns". As a result , the word can not be removed as a proper noun. This failure causes a reduction in accuracy.

Secondly, we will discuss verbs. In sentences containing "*sarari to*", accuracy increased by 5.42 points, and for "*gatan to*", accuracy increased by 9.08 points ; a major increase for both features. This shows that the verbs modified by onomatopoeias are very important for distinguishing the meanings. However, the number of features is insufficient, because a Japanese onomatopoeia basically only modifies one verb. Therefore, there are few features to distinguish the meaning. There are a few sentences in which the meanings are different even though the features are the same. Such sentences are very difficult to disambiguate clearly. The combination of nouns and verbs increased accuracy by 2.02 points for sentences containing "*gatan to*", and 0.37 points for "*sarari to*". When "v' and n' " were combined and used as a feature, the highest accuracy was achieved for "*gatan to* ", although this was not the combination of nouns and verbs increased accuracy by 2.02 points for sentences containing "gatan to", and 0.37 points for "sarari to". When "v' and n' " were combined and used as a feature, the highest accuracy was achieved for "*gatan to* ", although this was not the case for "*sarari to*". The reason why the accuracy for "*gatan to*" is high is that the accuracy of the feature "n' " is high. The accuracy of "*gatan to*" is 8 points higher than '*sarari to*" with the feature "n' ". This shows that "n' " has a positive influence on improvement of accuracy. On the other hand, for "*sarari to*", the feature "n' and v' " is 1.36 points lower than "v' " , because the accuracy of "n' " was low. Below, we analyze examples of failed semantic disambiguation. These examples are classified into three patterns. These patterns and an example show in Table 9, Table 10, and Table 11. We conducted analysis for each pattern.

Example of pattern (1), there are three forms of the verb "*suru*" ("to do") in this sentence. The other nouns extracted from this sentence nouns do not exist in any other sentences in our data. Therefore, the only feature for disambiguation is "*suru*". Furthermore, there are 75 sentences which include "*suru*", of which 45 sentences were judged as m1(defined in **2.2**), and 30 sentences as m2. The proportion of sentences judged as m1 is larger than those judged as m2. Therefore, there is a trend to wards output as m1 although correct meaning is m2.

Table 9. Pattern (1) : In case that semantic disambiguation depends on non-important verbs

Example sentence	*Watashi mo gochisou shitari purezento shitari suru no ga suki dakedo kono "sarari to" to iu no ga otokomae.* (I like treating people to meals and giving presents too, but the way he does it so smoothly (*"sarari to"*) is really manly.)
(Words : frequency)	(gochisou : 1), (purezento : 1), (otokomae:1), (suru : 3)

Example of pattern (2), the nouns and verbs in this sentence do not occur in other sentences. Therefore, there are insufficient features for disambiguation. The output meaning of this sentence is m2 though the correct meaning is m1.

Table 10. Pattern (2) : In case that there is a lack of information to distinguish meanings

Example sentence	*Hontou, futshuu no ro-ru ke-ki nan desu ga, suponji ha fuwafuwa de, kuri-mu ha tottemo nameraka de, shita no ue de "sarari to" tokemasu.* (It is a secret that I freaked out a bit the cream is so smooth it melts *"sarari to"* on the tongue.)
(Words : frequency)	(shita : 1), (suponji : 1), (hontou : 1), (kuri-mu : 1), (tokeru : 1)

Example of pattern (3), the output meaning of this sentence is m2(defined in **2.2**) though the correct meaning is m1. Both "*yuka*" and "*ochiru*" exist in other sentences in our data: two sentences include "*yuka*" and 57 sentences include "*ochiru*". Therefore, if we consider the proportions, "*yuka*" has very poor potential to distinguish meanings even though there are tendencies for "*yuka*" to be an important word for disambiguation and to exist disproportionately in either meaning.

Table 11. Pattern (3) : In case that the difference in frequency causes a failure

Example sentence	*Tenjo ya yuka ga "gatan to" ochite karuku tenpatta no ha naisho* (I like treating people to meals and giving presents too, but because the ceiling and floor came crashing ("*gatan to*") down)
(Words : frequency)	(yuka : 1),(tenjo : 1),(naisho : 1),(ochiru : 1)

5 Conclusions and Future Works

Our research evaluated a selection method for features used to semantically disambiguate onomatopoeias with two meanings. As a result, using nouns and verbs as features achieved a stable accuracy of over 80 points. Moreover, using verbs that are modified by onomatopoeias as a feature improved the accuracy without depending on the type of onomatopoeia. In addition, according to types of onomatopoeia, the combination of verbs modified by onomatopoeias and nouns further improved the accuracy. Therefore, these features highly effective for semantic disambiguation. However, there remains a problem in the handling of nouns. The frequency of word occurrence has a major influence on semantic disambiguation. Consequently, nouns whose frequency of occurrence is low are not emphasized in disambiguation, even when the noun is important in order to distinguish the meaning. Therefore, we need to apply weighting to words according to frequency of co-occurrence of "nouns and onomatopoeias" or "verbs modified by onomatopoeias and nouns". In future work, we will incorporate this co-occurrence as a feature and aim for further improvements in accuracy.

References

1. Tokyo Olympics, `http://tokyo2020.jp/`
2. Yuichiro Shimizu, Ryuichi Doizaki and Maki Sakamoto, A System to Estimate an Impression Conveyed by Onomatopoeia, Transactions of the Japanese Society for Artificial Intelligence, vol. 29, no. 1, pp. 41–52, 2014. (in Japanese)
3. Chisato Asaga, Mukarramah Yusuf, Chiemi Watanabe, ONOMATOPEDIA: Onomatopoeia Online Example Dictionary System Extracted from Data on the Web, Proceeding of the 10th Asia Pacific Web Conference (APweb), 8b-1, 2008.
4. Yasuki Furutake, Satoshi Sato and Kazunori Komatani, *Onomatope wo iikaeru hyougen no jidoushushu* [Automatic Collection of Expression in Other Words of Onomatopoeias], Proceeding of the 17th Annual Meeting of the Association for Natural Language Processing, pp. 904–907, 2011. (in Japanese)
5. Yuzu Uchida, Kenji Araki and Jun Yoneyama, Semantic Ambiguity of Onomatopoeia Extracted from Blog Entries, Proceeding of the 27th Fuzzy System Symposium, pp. 853–856, 2011. (in Japanese)
6. Twitter, `https://twitter.com/`
7. Taku Kudo, et al. MeCab, Yet Another Part-of-Speech and Morphological Analyzer, `http://mecab.sourceforge.net/`.
8. Thorsten Joachims, SVM-Light, `http://svmlight.joachims.org/`.
9. Agency for Cultural Affairs, *gaikokujin no tame no kihongoyourei jitenn* (Basic Word Usage Dictionary for Foreigners), National Printing Bureau, 1990.
10. Google: `https://www.google.co.jp/`
11. Yuzu Uchida, Kenji Araki and Jun Yoneyama, Affect Analysis of Onomatopoeia Sentences Extracted from Blog Entries Proceeding of the 10th Forum on Information Technology, pp. 274–279, 2011. (in Japanese)
12. Ameba, `http://www.ameba.jp/`.
13. CaboCha, Yet Another Japanese Dependency Structure Aalyser, `http://code.google.com/p/cabocha/`