Comparison of Distributed Representations and Context Vectors for Japanese Onomatopoeia Classification

Kanako Komiya[1] Minoru Sasaki[1] and Hiroyuki Shinnou[1]

Ibaraki University, 4-12-1 Nakanarusawa, Hitachi-shi, Ibaraki, 316-8511 JAPAN,
{kanako.komiya.nlp,
minoru.sasaki.01,hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

Abstract. Japanese language has thousands of onomatopoeias and they convey a subtle difference of the senses. [7] proposed to classify onomatopoeias depending on the context vectors. On the other hand, the distributed representation has been studied intensively in recent years. Therefore, we propose to classify the onomatopoeias depending on the distributed representations instead of the context vectors. We compare the results of the classifications and evaluate them based on the correct classification developed by humans in this paper.

1 Introduction

Onomatopoeias are words that mimic the sound they describe and are often used in order to expresses a feeling or a sensibility of the atmosphere they describe [7]. Japanese has thousands of onomatopoeias and they are used in daily life and have recently started to attract a lot of attention of researchers on natural language processing in Japan (Section 2). There are onomatopoeia dictionaries (cf. [16] and [1]) in Japan. [16] mentions they consist of GIONGO, i.e., words that express the sound and GITAIGO i.e., words that express the situation and are grouped into three major types: (1) words that express the sound from outside of humans vocal equipments, (2) words that express the sound or the voice from humans vocal equipments and cannot be separated into each sounds, and (3) words express the situation of things with no sound by a sensibility the sound suggests.

Some onomatopoeias are semantically or phonologically similar each other and the choice of these onomatopoeias sometimes gives a big difference among Japanese sentences. For example, both NIKONIKO にこにこ and NIYANIYA にやにや are onomatopoeias that indicate someone smiles. However NIKONIKO indicates that it is affable, and NIYANIYA indicates that it is not. On the other hand, NIKONIKO and NIKORI(TO) にこり（と）have almost the same meaning. Therefore, [7] proposed a clustering method to classify onomatopoeias and visualize the relationships of them depending on their contexts such as surrounding words, their part-of-speeches and their meanings from the perspective of word senses.

On the other hand, work using distributed representations, e.g., word2vec, is prevalent in the latest years. Distributed representations [11] are the vector representations of the word meanings. They are calculated based on contexts of each word and used for comparisons of similarities of word meanings and compositions of them. Therefore, this paper proposes to classify the onomatopoeias depending on them instead of the context vectors, such as the surrounding words and the part of speeches (Section 3). We conducted clustering for the onomatopoeias on the same corpora using the same settings as [7] (Section 4) and compared the results and evaluated them based on the correct classification developed by humans (Section 5). We discuss the results in Section 6 and conclude the paper in Section 7.

2 Related Work

As we described above, onomatopoeias have recently started to attract a lot of attention of researchers on natural language processing in Japan. [12] focused on onomatopoeias have interesting sounds and rhythms and proposed an educational system named "Onomato Planets" for children. [14] conducted a trend analysis of appearance of onomatopoeias in defect reports in English and Japanese. [8] described the way to narrow down the examples of onomatopoeias on the Web using co-occurrence and improved the quality of the examples of them. There are some works on construction of onomatopoeia dictionary. [15] reported the automatic way to construct Japanese onomatopoeia dictionary using examples on the Web. [3] reported an onomatopoeia usage dictionary they compiled based on examples.

[2] proposed an onomatopoeia dictionary. They proposed an online onomatopoeia example-based dictionary named ONOMATOPEDIA, that comprises extensive example sentences collected from the Web, for learners of Japanese. These papers propose a clustering algorithm for onomatopoeias, but they focused on the onomatopoeias that have several meanings depending on the contexts. On the other hand, this paper proposes a clustering method of onomatopoeias, instead of each instances of them. [13] classified the onomatopoeias that express sensibilities using SOM. [18] proposed and developed an onomatopoeia thesaurus map. They labeled objects with onomatopoetic words on the thesaurus map in order to visualize the similarity relationships. However they focused on the phonological features of onomatopoeias and not the context features of them. [5] also automatically classified onomatopoeias. They collected examples from the Web and used co-occurrence and phonological features of onomatopoeias. However they did not use context features such as surrounding words again. [9] classified Japanese psychomimes, a kind of onomatopoeia, using pLSA and SMO. They used verbs which located immediately after the psychomimes as a feature. [19] analyzed the onomatopoeias from the two perspective, i.e., using physical image and Web corpora, in order to develop a Japanese onomatopoeia dictionary. [6] used Japanese onomatopoeias for sentiment analysis. [4],

The closest work to ours is [7]. They proposed a clustering method of onomatopoeias based on contexts such as surrounding words, their part-of-speeches and so on with the aim of classifying onomatopoeias and visualizing the relationships of them. They classified the onomatopoeias using the method of word sense disambiguation to analyze them from the perspective of the word senses. On the other hand, the distributed representations are also effective for word sense disambiguation as [17] reported. Therefore, this paper proposes to classify the onomatopoeias depending on the distributed representations instead of the context vectors, compare the results of the classifications, and evaluate them based on the correct classification developed by humans.

3 Classification of Onomatopoeia

The clustering algorithm of the onomatopoeias in this paper consists of the following three steps.

1. Generate the distributed representations of each onomatopoeia from the corpus.
2. Calculate the distances among the distributed representations of the onomatopoeias.
3. Perform clustering based on the distances among onomatopoeias.

We used the Balanced Corpus of Contemporary Japanese (BCCWJ) [10] and Mecab [1] as a morphological analyzer [2]. The distributed representations were generated using Word2Vec[3]. The window size of each target word was set to five and the size of dimensionality of the distributed representation was set to 50. We used default settings for the other parameters. In addition, the cosine similarities among the distributed representations were used to measure the distance among the onomatopoeias. On the other hand, [7] conducted clustering using the Jensen-Shannon divergence of the context vector sets whose features were bag-of-words, part-of-speeches, abstracted semantic classes, and a syntactic feature[4]. The window size of each target word was set to two.

We used single-link, bottom-up, and hierarchical clustering, followed [7].

Single-link clustering is a clustering method where the similarity between two clusters is the similarity of two most similar members of each cluster. In other words, if clusters c_u and c_v are merged into $c_w = c_u \cup c_v$, then the similarity of c_w and another cluster c_k is the maximum of the two individual similarities:

$$sim(c_w, c_k) = max(sim(c_u, c_k), sim(c_v, c_k)) \quad (1)$$

[1] http://mecab.googlecode.com/svn/trunk/mecab/

[2] [7] used the same corpus but used Chasen as a morphological analyzer.

[3] http://word2vec.googlecode.com/svn/trunk/

[4] We cannot combine the dense continuous representation and the sparse representation because the former uses a vector and the latter uses a set of vectors to represent a word type.

Bottom-up clustering starts with a separate cluster for each onomatopoeia and the most similar clusters are merged into a new cluster in each step of the clustering.

4 Experiment

We extracted the following four types of onomatopoeias which was chosen from [16] and classified them, followed [7]. We used the ones which appeared 10 times or more in the corpus.

- Onomatopoeias about the sunshine
- Onomatopoeias about the coldness
- Onomatopoeias about the rain, the snow, and the ice
- Onomatopoeias to express that someone gets upset

Accordingly, we used 12 onomatopoeias about the sunshine, which can be classified into four groups:

1. Sunny, bright: ぎらぎら GIRAGIRA, てかてか TEKATEKA, じりじり JIRIJIRI, ぽかぽか POKAPOKA, かっと KATTO, かんかん KANKAN
2. Sunshine: おっとり OTTORI, ぽっと POTTO,
3. Fair weather: からっと KARATTO, すかっと SUKATTO, からり KARARI,
4. Sundown: とっぷり TOPPURI.

GIRAGIRA and TEKATEKA represent the situation that something (cf. the sun) shines very brightly. JIRIJIRI means that the sun shines too brightly and sometimes it is too hard for people. POKAPOKA represents the warmth. KATTO indicates that something burns up. KANKAN is the onomatopoeia that represents the situation that the sun shines very brightly. The intensity of fire was focused on as for this onomatopoeia. OTTORI represents calmness, typically used for nature of person or climate. POTTO means that something (cf. a light) burst into light. KARATTO and SUKATTO are very similar each other and they both indicate that the clear sky and the sunshine. KARARI represents the clear sky. TOPPURI indicates the sunset.

We used six onomatopoeias about the coldness, which can be classified into three groups:

1. Chill: ぞくぞく ZOKUZOKU,
2. Coldness: じわじわ JIWAJIWA, しんと SHINTO, りんと RINTO,
3. Iciness: ひんやり HINYARI, ひやひや HIYAHIYA.

ZOKUZOKU represents the situation that someone feels a chill because of fever, fear or cold air. ZIWAZIWA suggests that it get cooler and cooler. SHINTO can be used to express the coldness but it is mainly used to express the silence. RINTO is similar case to SHINTO, but it is mainly used to express the bravery. HINYARI indicates the situation that someone feels cold. HIYAHIYA indicates that someone feels cold on one's skin.

We used 13 onomatopoeias about the rain, the snow, and the ice, which can be classified into three groups:

1. Rain: じめじめ JIMEJIME, どんより DONYORI, ぽつぽつ POTSUPOTSU, ぱらぱら PARAPARA, ばらばら BARABARA, しょぼしょぼ SHOBOSHOBO, しとしと SHITOSHITO, びしょびしょ BISHOBISHO, さっと SATTO, ざっと ZATTO,
2. Snow: ちらほら CHIRAHORA, はらはら HARAHARA,
3. Thunder: ごろごろ GOROGORO.

JIMAJIME and DONYORI mean the cloudy sky but it may not rain yet. Other onomatopoeias except GOROGORO indicate that it rains now. POTSUPOTSU and PARAPARA describe the sound of the rain and suggest a light rain. BARABARA describes the sound of large drops of rain. SHOBOSHIOBO indicates the continuous rain. SHITOSHITO represents the situation that the rain falls silently. BISHOBISHO is an onomatopoeia that suggests something/someone is wet. SATTO indicates the rain falls in a short time and ZATTO means the rain falls heavily in a short time. CHIRAHORA represents that the situation the light snow falls. HARAHRA represents the situation that a little snow or rain falls silently. GOROGORO describes the sound of the thunder.

Finally, we used 19 onomatopoeias to express that someone gets upset, which can be classified into five groups:

1. Anger: むっと MUTTO, むしゃくしゃ MUSHAKUSHA, かっと KATTO, ぐらぐら GURAGURA, むかむか MUKAMUKA, ぷりぷり PURIPURI, かんかん KANKAN, がみがみ GAMIGAMI, ぷんと PUNTO, ぷんぷん PUNPUN, ぞっと ZOTTO
2. Frustration: ぐつぐつ GUTSUGUTSU, かちん KACHIN
3. Bad mood: ぶすっと BUSUTTO, つんと TSUNTO, つんけん TSUNKEN,
4. Unfriendliness: むっつり MUTTSURI, つっけんどん TUKKENDON
5. Hardening of attitudes: きっと KITTO.

MUTTO suggests that someone gets upset with the other person's speech or behavior. MUSHAKUSHA suggests that someone loses one's composure because of anger, and KATTO indicates that something burns up and it can be used for person like "That burns me up". GURAGURA indicates the rage and this onomatopoeia can be used to describe the sound of a boiling. MUKAMUKA suggests the sudden anger. PURIPURI represents someone gets ratty and KANKAN represents something burns intensely and it can be used for people as well. GAMIGAMI is an onomatopoeia that describes someone berates. PUNTO represents the situation where someone gets sulky. PUNPUN means that someone is in a fume. GUTSUGUTSU represents that someone gets a flash of anger about something. KACHIN suggests that someone gets upset with what they are said. BUSUTTO describes the sulky face, TSUNTO represents the bad temper, and TUKKENDON represents the unfriendly attitude. MUTTSURI suggests the glum mood. TSUNTSUN indicates that someone is cranky. KITTO represents the grim look.

These types are manually classified according to the description in Dictionary of Japanese Onomatopoeias [16]. In addition, we followed [7] and used the onomatopoeias with the suffix TO like KATTO and KARATTO for the distributed

Type of onomatopoeia	Dictionary	Final exp.
Sunshine	17	12
Coldness	17	6
Rain, snow, ice	28	13
To get upset	40	19

Table 1. The number of onomatopoeias in the dictionary and in the final experiment

Type of onomatopoeia	Min	Max	Avg.
Sunshine	17	228	80.33
Coldness	34	640	235.50
Rain, snow, ice	12	1,154	225.62
To get upset	15	5,812	384.05

Table 2. The minimum, maximum, and average number of instances of onomatopoeias in the final experiment using distributed representation

representations except for RINTO. We used RIN, without the suffix TO, because the new morphological analyzer did not extract RINTO but extracted RIN.

Table 1 shows the number of the types of onomatopoeias in the dictionary and in the final experiment. The maximum, minimum, and average number of the instances of onomatopoeias in the final experiment is summarized in Table 2. These numbers are slightly different from those in [7] (shown in Table 3) because we used a new morphological analyzer.

5 Results

Figures 1 and 2 in appendix show the results of the hierarchical clustering of onomatopoeias about the sunshine based on the distributed representations and the context vectors, respectively. Figures 3 and 4 in appendix show the results of those about the coldness. Figures 5 and 6 in appendix show the results of those about the rain, the snow, and the ice. Figures 7 and 8 in appendix show the results of those to express that someone gets upset. We evaluated the entropy

Type of onomatopoeia	Min	Max	Ave
Sunshine	12	228	79.00
Coldness	24	243	139.83
Rain, snow, ice	13	1,175	230.46
To get upset	15	5,894	390.58

Table 3. The minimum, maximum, and average number of instances of onomatopoeias in the final experiment using context vectors

Onomatopoeia	Entoropy		Purity	
	Distributed rep.	Context vec.	Distributed rep.	Context vec.
Sunshine	0.62	**0.54**	**0.67**	**0.67**
Coldness	**0.63**	0.71	**0.67**	0.50
Rain, snow, ice	0.56	**0.54**	**0.79**	**0.79**
To get upset	**0.47**	0.50	**0.74**	0.68

Table 4. Entropies and Purities of Clusters by Distributed Representations and Context Vectors

and purity based on the manually labeled correct answers, which are the groups in Section 4. Table 4 shows the results. The better results are written in bold. The cluster numbers used for the evaluation of the entropy and purity are also shown in the figures in the appendix.

6 Discussion

According to Figures 1 and 2, these two results of clustering bear little resemblance to each other. However, the most similar onomatopoeias are the same in Figures 3 and 4 (ZOKUZOKU and HINYARI), in Figures 5 and 6 (JIMEJIME and DONYORI), and in Figures 7 and 8 (ZOTTO and MUTTO). In addition, when Figures 3 and 4 are compared, they are almost the same but one exception: HIYAHIYA is grouped with SHINTO first or it is integrated into the group of ZOKUZOKU, HINYARI, and JIWAJIWA first. Moreover, the only difference between Figures 5 and 6 is that ZATTO, SATTO, BARABARA, and BISHOBISHO are grouped into the first and the largest group or not, and the difference between Figures 7 and 8 is KACHIN, KANKAN, and GUSTUGSTU. The part-of-speeches, abstracted semantic classes, and a syntactic feature were not used for the distributed representations, and the window size of each target word for the context vectors was set to two whereas that for the distributed representations was set to five. We think that the common feature, i.e., bag-of-words adjacent to the onomatopoeias greatly contributed the results of clustering, because the results are sometimes similar each other despite these differences.

In addition, Table 4 shows that the entropies of the clusters developed using the context vectors slightly outperformed those using the distributed representations when the onomatopoeias about the sunshine or those about the rain, the snow, and the ice are compared, but those using the distributed representations slightly outperformed those using the context vectors when the onomatopoeias about the coldness or those to express that someone gets upset are compared. Moreover, Table 4 shows that the purity of the clusters developed using the distributed representations slightly outperformed those using the context vectors when the onomatopoeias about the coldness or those to express that someone gets upset are compared, and there was no difference between the context vectors and the distributed representations when the onomatopoeias about the sunshine or those about the rain, the snow, and the ice are compared. Therefore, we can

see that the clustering based on the distributed representations classified the onomatopoeias more similarly to manually classified onomatopoeias in general.

In addition, some onomatopoeias have more than one sense. For example, KATTO and KANKAN are included in both onomatopoeias about the sunshine and onomatopoeias to express that someone gets upset. We could not use the word sense tagged data of onomatopoeia and did not consider them like [7]. Therefore, the word sense disambiguation of the onomatopoeias is our future work.

7 Conclusion

This paper proposed to conduct clustering for Japanese onomatopoeias based on the distributed representations and compared the results to the method where the clustering was conducted based on the context vectors, which was proposed in [7]. The single link and hierarchal clustering was employed as the clustering method and the cosine similarities were used to measure the distance between the onomatopoeias. The following four types of onomatopoeias were classified: (1) Onomatopoeias about the sunshine, (2) onomatopoeias about the coldness, (3) and onomatopoeias about the rain, the snow, and the ice, and (4) onomatopoeias to express that someone gets upset. Although the features for clustering were different, the most similar onomatopoeias were the same when the distributed representation were used and when the context vectors were used, for the three type of the onomatopoeias, which are those about the coldness, those about the rain, the snow, and the ice, and those to express that someone gets upset. In addition, the experiments revealed that the distributed representation was slightly better than the context vectors in general when the entropies and the purities of the clusters were evaluated based on the manually labeled data. The problems to analyze the onomatopoeias, e.g., the suffix of them, remain because it is difficult for morphological analyzer to correctly delimit and extract the onomatopoeias. The word sense disambiguation of the onomatopoeias should also be considered in the future.

Acknowledgement

This work was partially supported by some grants, which will be discribed later.

References

1. Amanuma, Y.: Giongo, Gitaigo Jiten. Tokyodo Shuppan publisher (In Japanese) (1993)
2. Asaga, C., Mukarramah, Y., Watanabe, C.: Clustering technique with consideratiaon of modification relation for classifying sentences by meaning of onomatopoeia on online onomatopoeia example-based dictionary onomatopedia. In: Proceedings of DEWS2008 (In Japanese). pp. A3–4 (2008)

3. Furutake, Y., Sato, S.: Yorei ni motodsuku onomatope youhou jiten no hensan. In: Proceedings of NLP 2010 (In Japanese). pp. 994–997 (2010)
4. Hashimoto, K., Takeuchi, K.: A method to estimate subjective context-sensitive kansei polarities of japanese onomatopoeic expressions. In: Proceedings of the 25th Annual Conference of the Japanese Society for Artificial Intelligence (In Japanese). pp. 1C2–OS4b–6 (2011)
5. Ichioka, K., Fukumoto, F.: Graph-based clustering for semantic classification of onomatopoetic words. In: Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing. pp. 33–40 (2008)
6. Igarashi, T., Sasano, R., Takamura, H., , Okumur, M.: Use of sound symbolism in sentiment classification. Journal of Natural Language Processing Vol.20(No.2), 183–200 (2013)
7. Komiya, K., Kotani, Y.: Classification of japanese onomatopoeias using hierarchical clustering depending on contexts. In: Proceedings of the 2011 English International Joint Conference on Computer Science and Software Engineering. pp. 108–1013 (2011)
8. Kurosawa, Y., Mera, K., Takezawa, T.: Ziko shoshikika mappu som niyoru shinzyo wo arawasu onomatope bunrui no saikentou. In: Proceedings of NLP 2010 (In Japanese). pp. 1058–1061 (2010)
9. Kurosawa, Y., Takezawa, T.: Psychomime classification by using self-organizing map and probabilistic latent semantic analysis (in japanese). In: Proceedings of the 25th Annual Conference of the Japanese Society for Artificial Intelligence. pp. 1C2–OS4b–7 (2011)
10. Maekawa, K.: Balanced corpus of contemporary written japanese. In: Proceedings of the 6th Workshop on Asian Language Resources (ALR). pp. 101–102 (2008)
11. Mikolov, T., tau Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of NAACL 2013. pp. 746–751 (2013)
12. Miyazaki, A., Tomimatsu, K.: Onomato planets: Physical computing of japanese onomatopoeia. In: Proceedings of the 3rd International Conference on Tangible and Embedded Interaction. pp. 301–304 (2009)
13. Morita, K., Suzuki, Y.: Onomato planets: Physical computing of japanese onomatopoeia. In: Proceedings of NLP 2010 (In Japanese). pp. 924–927 (2010)
14. Nasukawa, T., Unno, Y., Murakami, A.: Kikino fuguai wo kijutsu shita niongo to eigo no kopasu ni okeru onomatope. In: Proceedings of NLP 2010 (In Japanese). pp. 154–158 (2010)
15. Okumura, M., Okumura, A., Saito, S.: Automatic construction of a japanese onomatopoeic dictionary using text data on the www. In: Proceedings of the 11th International Conference on Applications of Natural Language to Information Systems. pp. 209–215 (2006)
16. Ono, M.: Nihongo Onomatope Jiten (Dictionary of Japanese Onomatopoeias). Shogakukan publisher (In Japanese) (2007)
17. Sugawara, H., Takamura, H., Sasano, R., Okumura, M.: Context representation with word embeddings for wsd. In: Proceedings of PACLING 2015 (2015)
18. Tomoto, Y., Nakamura, T., Kanoh, M., Komatsu, T.: Visualization of similarity relationships by onomatopoeia thesaurus map. In: Proceedings of the IEEE World Congress on Computational Intelligence. pp. 3304–3309 (2010)
19. Uno, R., Kaji, N., Ogai, Y., Ikegami, T., Kitsuregawa, M.: Semantic stability and subjcectivity in mimetics. In: Proceedings of the 25th Annual Conference of the Japanese Society for Artificial Intelligence (In Japanese). pp. 1C2–OS4b–8 (2011)

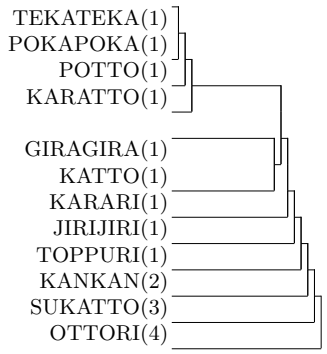

Fig. 1. Hierarchical clustering of onomatopoeias about the sunshine using distributed representations

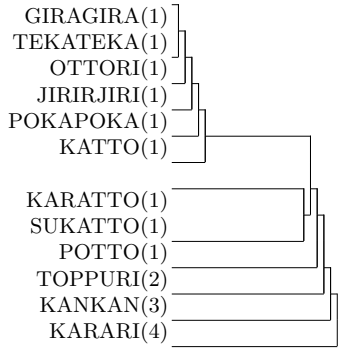

Fig. 2. Hierarchical clustering of onomatopoeias about the sunshine using context vectors

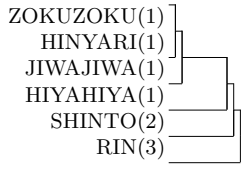

Fig. 3. Hierarchical clustering of onomatopoeias about the coldness using distributed representations

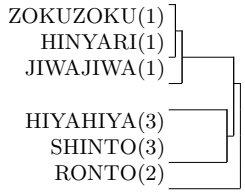

Fig. 4. Hierarchical clustering of onomatopoeias about the coldness using context vectors

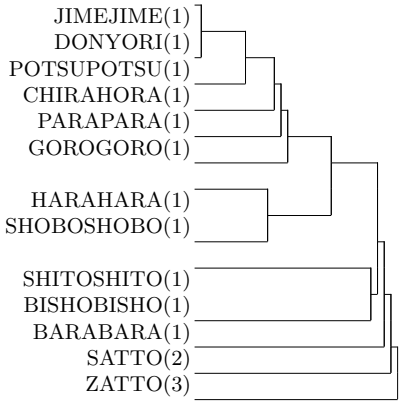

Fig. 5. Hierarchical clustering of onomatopoeias about the rain, the snow, and the ice using distributed representations

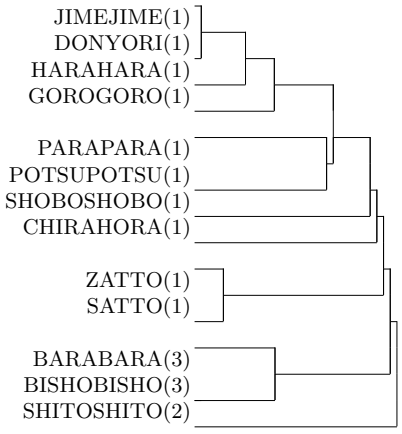

Fig. 6. Hierarchical clustering of onomatopoeias about the rain, the snow, and the ice using context vectors

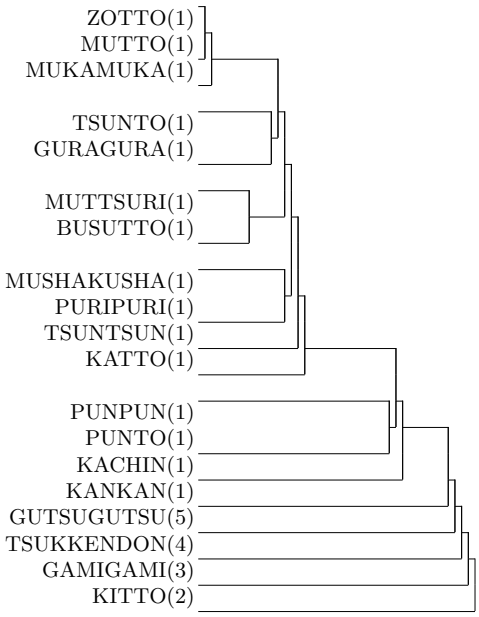

Fig. 7. Hierarchical clustering of onomatopoeias to express that someone gets upset using distributed representations

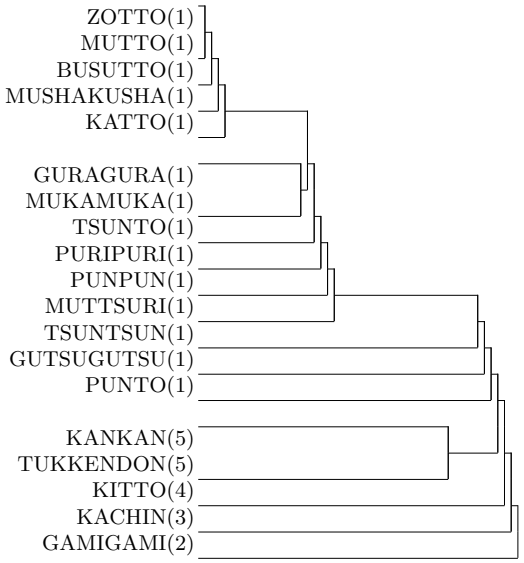

Fig. 8. Hierarchical clustering of onomatopoeias to express that someone gets upset using context vectors