Sensibility Estimation Method for Youth Slang by Using Sensibility Co-occurrence Feature Vector Obtained from Microblog

Kazuyuki Matsumoto
Faculty of Engineering, Tokushima University
Tokushima city, Japan
matumoto@is.tokushima-u.ac.jp

Minoru Yoshida and Kenji Kita
Faculty of Engineering, Tokushima University
Tokushima city, Japan
{mino;kita}@is.tokushima-u.ac.jp

***Abstract*—Social networking sites such as Twitter provide more opportunities to express what people think or intend in short text. In short text, abbreviations such as "ASAP" or "joinus" and emoticons are often used. Because these expressions are not registered into the existing dictionaries, these are analyzed as unknown expressions. That can be a bottleneck for improving accuracy of reputation analysis in text mining. To use context for unknown word clustering is a major method, however, it usually requires word segmentation process and it has weakness for split errors of unknown expressions such as youth slang. In this paper, we proposed a method to obtain the appropriate context even though unknown expressions cause split errors and estimate sensibility expressed in the text. Because the dimensions of the obtained context vector were enormous, we also proposed a method to create a feature vector based on the co-occurrence of the sensibility words as simple expression with low dimension. As an evaluation experiment, the proposed method showed certain accuracy even with the small training data.**

Keywords-youth slang; microblog; Twitter; affective computing; sensibility classification

I. INTRODUCTION

The opportunity to express our thoughts has widely spread on social networking sites such as Twitter. In these sites, it is important to express the information simply. Therefore, the texts tend to include many ungrammatical expressions or tend to describe in spoken language. These seem to suggest that a mechanism to analyze unknown expression or spoken language adequately will be necessary for analyzing tweet data that is increasing rapidly.

In the existing studies, the semantic analysis of the unknown words had been widely studied. Most of the analysis target of these studies is limited the context such as noun expressing proper thing [1-7]. Some studies aimed to process unknown onomatopoeia [8-9]. Onomatopoeia is an expression that can be patterned with character combinations or orders. For this reason, to obtain the pattern, it is not necessary to use context for processing onomatopoeia.

Some youth slang words can be estimated their impressions by referring to the characters consisting of the words as clue. However, a lot of youth slang words form their impressions from information associated with the words in addition to the superficial information.

The aim of this paper is not to estimate the meanings of youth slang but to estimate sensibility of youth slang. The aim of this paper is not to estimate the meanings of youth slang but to estimate sensibility of youth slang. Matsumoto et al. [10] proposed a method of estimating emotion of youth slang by using emotion vector. However, the expressions such as net slang or youth slang are expressing various impressions. It seems insufficient to annotate the existing emotion vector with two dimensions.

In this paper, to register new youth slang into the affective dictionary, we proposed a sensibility estimation method based on the result of the sensibility questionnaire about youth slangs.

Youth slang is subject to change in fashion. New words are daily created whereas many of them quickly go out of use. So, it is necessary to update the information registered in the affective dictionary each time. Our proposed method can estimate sensibility considering new semantic background in real time by obtaining co-occurrence words based on Twitter.

II. YOUTH SLANG QUESTIONNAIRE

To obtain the data that is necessary to make training data for youth slang sensibility estimation, we conducted a questionnaire about youth slang. The questionnaire was targeted for five university students in their twenties.

The questionnaire asked about their impressions to each youth slang word. The impressions were answered from the sixteen pairs of impressions. The user interface of the tool to answer the questionnaire is shown in Fig. 1. A part of the questionnaire results is shown in Fig. 2.

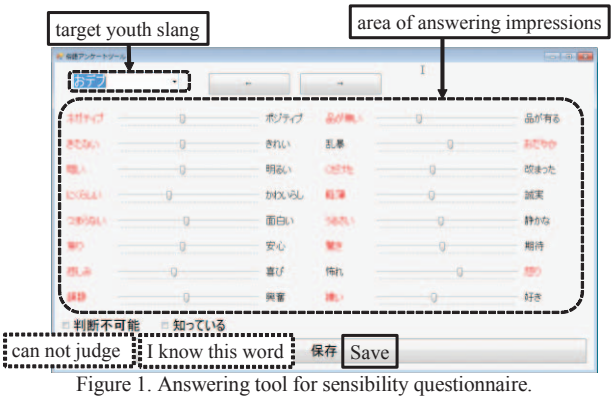

Figure 1. Answering tool for sensibility questionnaire.

978-1-4673-8126-0 /15/$31.00 ©2015 IEEE

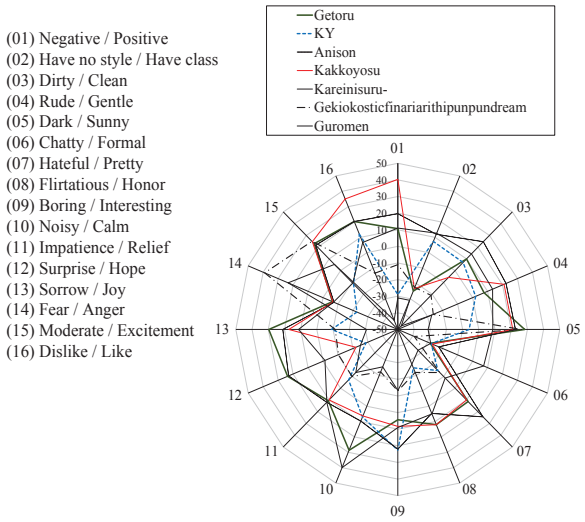

Figure 2. A part of the questionnaire answering result.

Averages of the results were calculated. The answers varied in some degree depending on the examinee. We thought that it would be necessary to regulate the varieties by normalizing the answers for each examinee.

From the obtained result, the sensibility similar words were listed up by calculating cosine similarity to confirm the sensibility similarity of each youth slang word. The result showed that rather many youth slang words with different meanings listed up in higher ranks because many of them originally did not have same meanings. However, it was very rare to have sensitivity similar words with opposite negative/positive emotion polarity. So, we thought that it would be acceptable to use the data as correct answers for the study aiming at evaluation analysis. We decided to start the sensitivity estimation based on the questionnaire result.

III. Proposed Method

A. Sensibility Co-occurrence Feature Vector

The sensibility co-occurrence feature vector defined in this paper is a feature vector indicating the characteristics related to sensibility. The feature vector is calculated based on the co-occurrence relation with the words expressing sensibility. In this study, we constructed a sensibility dictionary based on Japanese Appraisal Evaluation Expression Dictionary (JD) [11].

By using the dictionaries and thesaurus with annotations of emotional polarity, the words having the same meanings and emotion polarity with the words in the JD were added to the sensibility dictionary. We used four kinds of thesauruses; Japanese Lexicon [12], Bunrui Goihyo [13], EDR Concept Dictionary [14], and Japanese WordNet [15]. Each dictionary consists of the Japanese words with annotation of concept number describing semantic class.

To judge emotion polarity (positive/negative) of the words, the emotion expression dictionary [16], modern adjective usage dictionary [17], modern adverb usage dictionary [18], modern onomatopoeia usage dictionary [19] and the emotion polarity correspondence table [20] were used.

The method of creating the sensibility dictionary is shown in Fig. 3. The sensibility dictionary includes the words and the sensibility co-occurrence feature vector corresponding to the words. The word cluster decided by clustering based on the context and the affiliation degree to the cluster were registered into the dictionary.

The method to generate the sensibility co-occurrence feature vector to the expressions included in the dictionaries is described by using (1) – (4).

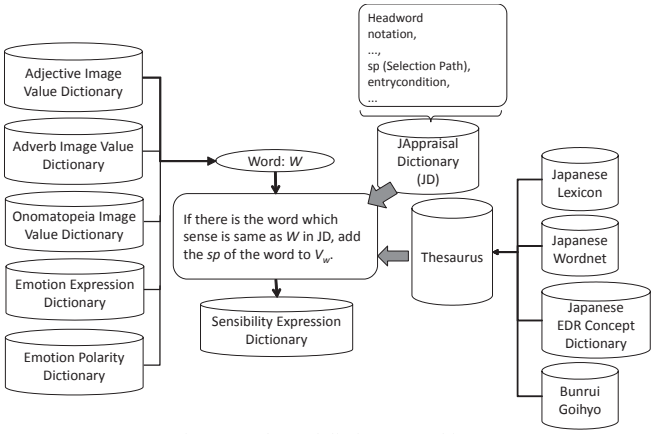

Figure 3. Flow of dictionary making.

$$flg_{x,i} = f_{sem}(ew_x, jw_i) \times f_{pn}(ew_x, jw_i) \tag{1}$$

$$f_{sem}(ew_x, jw_i) = \begin{cases} 1, & if \quad sem_x = sem_i \\ 0, & otherwise. \end{cases} \tag{2}$$

$$f_{pn}(ew_x, jw_i) = \begin{cases} 1, & if \quad pn_x = pn_i \\ 0, & otherwise. \end{cases} \tag{3}$$

$$S'_x = flg_{x,1} \times S_1 + \cdots + flg_{x,i} \times S_i + \cdots flg_{x,m} \times S_m \tag{4}$$

S'_x in (4) indicates the sensibility co-occurrence feature vector annotated to the word ew_x in a set of words included in the Japanese appraisal dictionary. S_1, S_2, ..., S_i, ..., S_m indicate the sensibility co-occurrence feature vector of the word jw_i included in the Japanese appraisal dictionary. $flg_{x,i}$ takes the value of 1 when the meanings of ew_x and jw_i (sem_x, sem_i) are same and the emotion polarities of ew_x and jw_i (pn_x, pn_i) are same. Because the value of S_x' in each dimension can become too large, the vector is actually normalized with the vector norm.

B. Co-occurrence Word Obtaining Method

Some methods for obtaining co-occurrence words, such as collocation analysis or n-gram analysis method, try to obtain the words that appear near the target word.

In this study, we used the method that obtains the words nearby the target word as co-occurrence word. However, if the utterance sentence is short, it will be difficult to obtain many co-occurrence words from a sentence, therefore, the distance from the target word was limited to 10 words and weighting by the distance was not considered.

At the time of obtaining the co-occurrence words, the target youth slang word was morphologically analyzed and their character strings were full text retrieved from the morphologically analyzed text corpus collected from Twitter to obtain the nearby words. Fig. 4 shows the flow of creating of the sensibility co-occurrence feature vector from the inputted youth slang.

The co-occurrence frequency vector CV_x is created as in (5) based on the set of co-occurrence words obtained from the full text retrieval for the inputted youth slang ys_x,

$$CV_x = \left[f_{t_1}, f_{t_2}, \cdots, f_{t_N}\right] \tag{5}$$

Then, which word cluster each word t_i of CV_x belongs to is judged. The obtained cluster ID c_{ti} and the affiliation degree a_{ti} are acquired and the matrix CM_x with each row indicating the data of co-occurrence word is created.

$$CM_x = \begin{matrix} f_{t_1} & c_{t_1} & a_{t_1} \\ \vdots & \vdots & \vdots \\ f_{t_N} & c_{t_N} & a_{t_N} \end{matrix} \tag{6}$$

After that, the sensibility expression e_j with the same cluster number c_{ti} is obtained for each word t_i in CM_x by referring to the sensibility dictionary. By using the absolute value of the difference between the affiliation degree a_{ej} of e_j to c_{tj} and the affiliation degree c_{ti}, we weighted the sensibility co-occurrence feature vector $S_j = \left[s_{j,1}, s_{j,2}, \cdots, s_{j,16}\right]$ of the affective expression of e_j. The $S_{x,i}$ calculated with (7) indicates the sensibility co-occurrence feature vector obtained from the co-occurrence word t_i.

$$S_{x,i} = \frac{2}{1.0 + \left|a_{t_i} - a_{e_j}\right|} \times f_{t_i} \times \left[s_{j,1}, s_{j,2}, \cdots, s_{j,16}\right] \tag{7}$$

Finally, we calculated the linear combination of the sensibility co-occurrence feature vector $S_{x,1}$, ..., $S_{x,N}$ and regarded it as the sensibility co-occurrence feature vector of youth slang ys_x (8).

$$FV_x = S_{x,1} + S_{x,2} + \cdots + S_{x,N} \tag{8}$$

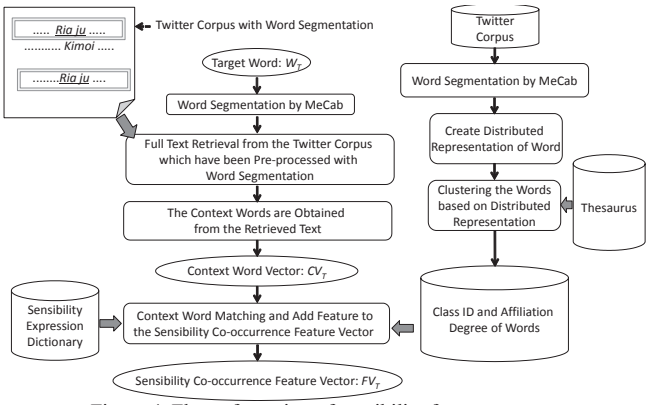

Figure 4. Flow of creation of sensibility feature vector.

C. Sensibility Estimation Method

We used machine learning method for classification in sensibility estimation. The dimension number of the sensibility co-occurrence feature vector was 224 and the dimension number of the sensibility vector to be estimated was 16 as the same with the number of the impression pairs used in the questionnaire.

If we use the multi value classification method, sensibility estimation is not expected to work well due to the amount of the training data. Therefore, we trained the binary classifier for each impression pair. This classifier was defined as Sensibility Binary Classifier (SBC). Fig. 5 shows a flow of sensibility estimation. Each element of sensibility vector has 1 or -1 value.

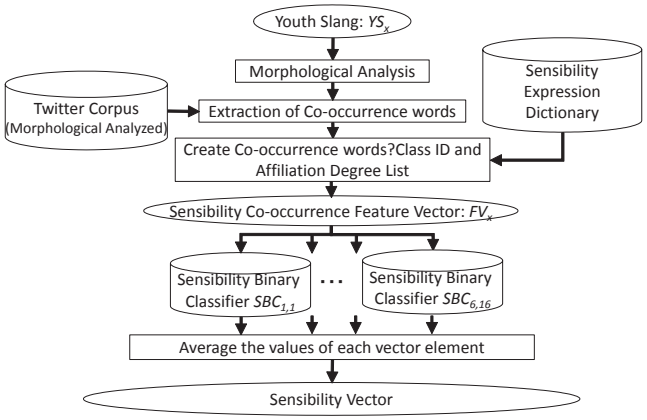

Figure 5. Flow of sensibility vector estimation.

As the binary classification algorithms, we used the six kinds of algorithms: (i) Linear Support Vector Classifier, (ii) Adaboost Classifier, (iii) Gaussian Naïve Bayes Classifier, (iv) Extra Trees Classifier, (v) Gradient Boosting Classifier, and (vi) Random Forest Classifier. The classifiers constructed for each impression pair are expressed as $SBC_{1,1}$,..., $SBC_{1,16}$, ..., $SBC_{6,1}$,..., $SBC_{6,16}$.

The average of the outputted result of each classification algorithm was calculated, and considered the value as the classification result. The value obtained for each impression

pair was defined as the element value of the sensibility vector, then we evaluated it.

IV. EXPERIMENT

A. Data

As training data, we selected the 185 words, which were able to be extracted sensibility co-occurrence feature vector, out of the 671 youth slang words prepared for the questionnaire. The sensibility classifiers were created to classify binarily each impression pair in each questionnaire element. We evaluate this sensibility classifier by comparing the classification result and the questionnaire result with cross validation.

As Twitter corpus, we used approximately thirty million sentences (32,234,154) that were collected for one month by using the words in the thesaurus as search queries and removed duplicate Tweets.

B. Evaluation Method

Because the classification results based on the cross validation test were outputted as binary values of 16 impression pairs, it is difficult to compare them with the questionnaire result. Therefore, we binarized the questionnaire result for each impression pair for evaluation. We decided the value obtained by dividing the matched number for each impression pair by 16. The baseline method (Baseline-1) for comparison is described as the follows.

We calculated TF*IDF value from the frequency of the co-occurrence words of the youth slang words then made context vector. The sensibility binary classifies of the 16 kinds of impressions were generated by performing machine learning. We used "Truncated Gradient Hinge" as machine learning algorithm.

As another baseline method (Baseline-2), we used the method that converted the co-occurrence words into the word cluster ID and trained the vector whose elements were obtained by multiplying the word TF*IDF value and the cluster affiliate degree value.

To segment the Twitter corpus into morphemes, we used the Japanese morphological analysis tool MeCab [21]. We also trained the distributed representations of the words by using word2vec [22]. We used the k-means method for clustering the words. The number of the clusters was optimized by using the evaluation function. We weighted the sensibility co-occurrence features by calculating the affiliation degrees to the clusters of the words.

C. Experimental Result

Fig. 6 shows the comparison of the experimental results by the proposed method and by the baseline method for each impression pair. The vertical axis indicates the average of the accuracies, and the horizontal axis indicates the sensibility IDs. The correspondence table of the sensibility IDs and the impression pairs is in Table I.

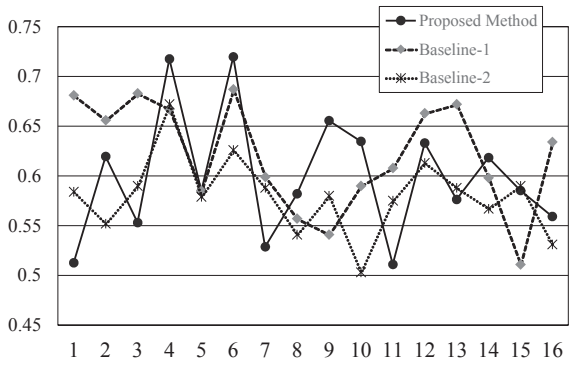

Figure 6. Accuracy of each sensibility.

TABLE I. SENSIBILITY IDS AND SENSIBILITY PAIRS

Sensibility ID	Sensibility Pair (+ / -)
1	Positive / Negative
2	Have Class / Have No Style
3	Clean / Dirty
4	Gentle / Rude
5	Sunny / Dark
6	Formal / Chatty
7	Pretty / Hateful
8	Honor / Flirtatious
9	Interesting / Boring
10	Calm / Noisy
11	Relief / Impatience
12	Hope / Surprise
13	Joy / Sorrow
14	Anger / Fear
15	Excitement / Moderate
16	Like / Dislike

As in this figure, the Baseline-1 obtained the stable accuracies on average although in the some impression pairs the accuracies of the baseline method were better than those of the proposed methods. The accuracies of the Baseline-2 were low overall. In the proposed method, the impression pairs of Positive/Negative or Relief/Impatience had low accuracies of appox. 0.5. Both of these pairs were often used as emotion categories. When the youth slang word itself is expressing emotion, it is not so strange if the nearby co-occurrence words are expressing opposite emotions.

The comparison of the accuracies of the proposed method and the baseline methods are shown for each youth slang word in Fig. 7. Although there were several words that could not be estimated sensibility at all by the baseline method, the proposed method could estimate sensibilities of the all words. It suggested the effectiveness of the sensibility co-occurrence feature vector.

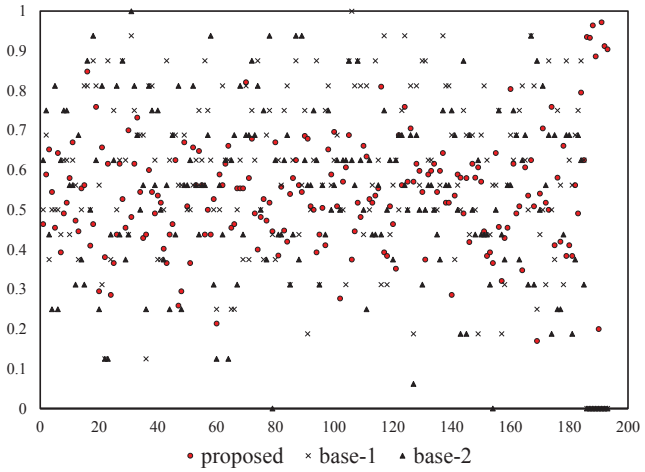

Figure 7. Comparison of the accuracies for each youth slang.

Table II shows the list of youth slang words obtained relatively high accuracies. From this result, we found that the word expressing sensibility strongly with adjectival usage had relatively high accuracy. On the other hand, some proper nouns were included in the list. One of the reasons seemed that a certain value was generally provided to the things that the proper nouns indicated. On the other hand, "*Yankii*," "*Nekama*," "*Rougai*," "*Joinus*", etc. were among the words resulted with lower accuracies. Although these words were already pervasive words, the meanings of them changed according to the times.

Depending on whether they were recognized as ironic meanings or literal meanings, the impression evaluation would vary. In the impression evaluation questionnaire, the number of the examinees was small. Therefore, it might be possible that the answers with large values became noise for training data.

TABLE II. EXAMPLES OF YOUTH SLANG THAT OBTAINED HIGH ACCURACY

Word	Accuracy	Word	Accuracy
wktk	0.848	*Korabo*	0.688
Makudo	0.821	*Yuuai*	0.688
Kakkoyosu	0.804	*Inaire*	0.686
Kawayui	0.795	*Kyaradeza*	0.679
Gurabo	0.759	*Gitafuri*	0.679
Meruado	0.759	*Rori*	0.670
Gochi	0.732	*Icharabu*	0.670
Moteki	0.705	*Moro*	0.670
Niiso	0.705	*Anison*	0.661
Meniakku	0.70	*Sebuire*	0.661
Warosu	0.696	*Gakugakuburuburu*	0.661

V. DISCUSSIONS

The experiment showed that in the binary classification task for the 16 impression pairs, the efficiency of classifying sensibility of youth slang words was better in the proposed method than in the baseline method based on the co-occurrence word cluster. However, the averaged accuracy of the proposed method was lower than that of the baseline method based on the simple co-occurrence word's TF*IDF (Baseline-1). The proposed method replaced the co-occurrence words into the sensibility information by using the co-occurrence frequency of the words and the sensibility expressions, therefore, the information of the words without direct relation with sensibility was removed. This seems to be a factor of the problem.

However, because the baseline method 1 (Baseline-1) treated the information of the words without changing them, it might cause over-training. As in this experiment, when a small scaled training data is used, the influence of the over-training should be considered.

In this experiment, we did not evaluate the youth slang words that were not targeted in the questionnaire. It was because there is no established method to evaluate the result of sensibility estimation of unknown youth slang even though the estimation could be possible. By applying the estimated sensibility vector to text mining task, the practical utility should be confirmed.

It is a problem when the proposed method sometimes failed to extract the co-occurrence words when it generated the sensibility co-occurrence feature vector. When co-occurrence words cannot be extracted or when the number of the extracted co-occurrence words is small, classification becomes difficult.

In case of the words consisting of Hiragana characters, segmentation will be sometimes different depending on the context from the case of analyzing the words by themselves. In that case, we cannot extract all of the co-occurrence words by using our proposed method.

VI. CONCLUSIONS

In this paper, we focused on the youth slang words as unknown sensibility expressions and proposed a method to estimate sensibility of the youth slang words. Because it is difficult to process youth slang by using only the existing language knowledge, we created the word cluster by using the Twitter corpus including a lot of expressions such as youth slang as language resource.

By creating the sensibility co-occurrence feature vector of the existing sensibility expressions based on the Japanese Appraisal Evaluation Expression Dictionary (JD), the dimensions of the feature vector used for sensibility estimation was reduced to 224 dimensions. As the experimental result, it was found that the proposed method could classify with higher accuracy compared to the baseline method (binary classifier) that trained the word vectors using TF*IDF of the co-occurrence word.

The biggest factor was that the size of the training data used for the experiment was small. The data requiring human resources such as a questionnaire evaluation cannot be always available. Our method could obtain a certain level of classification performance even with a small sized training data.

Our proposed method extracts the words near the target word to obtain the co-occurrence information from the training data. This method targeted the co-occurrence words in one tweet. However, we often describe our feelings in several tweets (not in single tweet). To be able to apply for such case, it would be necessary to obtain the co-occurrence information from more than one tweet near the target tweet by using chronological information.

In future work, we would like to propose a mechanism to obtain the timelines prior to and posterior to the target tweet by using the posted time information, and create the sensibility co-occurrence feature vector to evaluate the effectiveness.

ACKNOWLEDGMENT

This research was partially supported by JSPS KAKENHI Grant Numbers 15K00425, 15K16077.

REFERENCES

[1] J. Techo, C. Nattee and T. Theeramunkong, "A Corpus-based approach for automatic thai unknown word recognition using boosting techniques," IEICE Trans. INF. & SYST., vol. E92-D, no.12, Dec. 2009.

[2] M. Asahara and Y. Matsumoto, "Japanese Unknown Word Identification by Character-based Chunking," Proc. the 20th International Conference on Computational Linguistics, 2004, Article No. 459.

[3] K.-J. Chen and W.-Y. Ma, "Unknown Word Extraction for Chinese Documents," Proc. the 19th International Conference on Computational Linguistics, 2002, vol. 1, pp. 1-7.

[4] W.-Y. Ma and K.-J. Chen, "A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction," Proc. the Second SIGHAN Workshop on Chinese Language Processing, vol.17, 2003, pp. 31-38.

[5] R. Sasano, S. Kurohashi and M. Okumura, "A Simple Approach to Unknown Word Processing in Japanese Morphological Analysis," Proc. International Joint Conference on Natural Language Processing, 2013, pp.162-170.

[6] J. Miyake, S. Takeuchi, H. Kawanami, H. Saruwatari and K. Shikano, "Automatic reading annotation of trendy keywords by web text mining focused on parentheses expression," IEICE Technical Report, pp. 1-6, 2009.

[7] K. Uchimoto, S. Sekine and H. Isahara, "The Unknown Word Problem: a Morphological Analysis of Japanese Using Maximum Entropy Aided by a Dictionary," Conference on Empirical Methods in Natural Language Processing, 2001.

[8] S. Tsuchiya, M. Suzuki, F. Ren, and H. Watabe, "A novel estimation method of onomatopoeic word's feeling based on mora sequence patterns and feeling vectors," Journal of Natural Language Processing, vol.19, no. 5, 2012, pp. 367-379.

[9] R. Sasano, S. Kurohashi, and M. Okumura, "A simple approach to unknown word processing in Japanese morphological analysis," Journal of Natural Language Processing, vol. 21, no. 6, 2014, pp. 1183-1205.

[10] K. Matsumoto, K. Kita, and F. Ren, "Emotional vector distance based sentiment analysis of Wakamono Kotoba," China Communications, vol. 9, no. 3, 2012, pp.87-98.

[11] M. Sano, "Classification of evaluative expressions in Japanese: An appraisal perspective," IEICE technical report. Natural language understanding and models of communication, vol. 110, no. 400, 2011, pp. 19-24.

[12] S. Ikehara et al. "GoiTaikei : A Japanese Lexicon (CD-ROM)," NTT Communication Science Laboratories, 1999.

[13] Bunrui Goihyo (Word List by Semantic Principles, Revised and Enlarged Edition), National Institute for Japanese Language and Linguistics, 2004.

[14] EDR Electronic Dictionary, Japan Electronic Dictionary Research Institute, LTD, 1996.

[15] F. Bond, T. Baldwin, R. Fothergill and K. Uchimoto, "Japanese SemCor: A Sense-tagged Corpus of Japanese," the 6th International Conference of the Global WordNet Association (GWC-2012), Matsue, 2012.

[16] A. Nakamura "Emotion Expression Dictionary", Tokyodo, 1993.

[17] Y. Hida, "Gendai Keiyoushi Yoho Jiten," Tokyodo, 1991.

[18] Y. Hida, "Gendai Fukushi Yoho Jiten," Tokyodo, 1994.

[19] Yoshifumi Hida, "Gendai Giongo Gitaigo Yoho Jiten," Tokyodo, 2002.

[20] H. Takamura, T. Inui, and M. Okumura, "Extracting Semantic Orientations of Words using Spin Model," Proc. the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005), 2005, pp.133-140.

[21] MeCab: Yet Another Part-of-Speech and Morphological Analyzer: http://taku910.github.io/mecab/ .

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Proc. Workshop at ICLR, 2013.