International Journal of Advanced Intelligence
Volume 8, Number 1, pp.84-98, May, 2016.

AIA International Advanced Information Institute

Slang Analysis Based on Variant Information Extraction Focusing on the Time Series Topics

Kazuyuki Matsumoto[1], Minoru Yoshida[1], Seiji Tsuchiya[2], Kenji Kita[1] and Fuji Ren[1]

Faculty of Engineering, Department of Information Science and Intelligent Systems,
[1] *Tokushima University, 2-1 Minamijyousanjima-cho,Tokushima 770-8506, Japan*
matumoto;mino;kita;ren@is.tokushima-u.ac.jp
[2] *Faculty of Science and Engineering, Department of Intelligent Information Engineering and Science, Doshisha University,*
1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto-fu 610-0394, Japan
stsuchiy@mail.doshisha.ac.jp

Received (10 June 2015)
Revised (10 Sep. 2015)

Recently, with the increase in the number of users of Social Networking Sites (SNS), online communications have become more and more common, raising the possibility of using big data on SNS to analyze the diversity of language. Japanese language uses a variety of character types that are combined to create words and phrases. Therefore, it is difficult to morphologically analyze such words and phrases, even though morphological analysis is a basic process in natural language processing. Words and phrases that are not registered in morphological analysis dictionaries are usually not defined strictly, and their semantic interpretation seems to vary depending on the individual. In this study, we chronologically analyze the topics related to slang on Twitter. In this paper, as a validation experiment, we conducted a topic analysis experiment chronologically by using the sequential Tweet data and discussing the difference of topic change according to the slang types.

Keywords: Slang; Topic analysis; Time-series analysis.

1. Introduction

The languages we use daily are sometimes different in their degree of semantic recognition and are used in different ways by different people. Impressions received from the word or the sentence also vary by individual. These characteristics provide rich diversity in texts and contribute to the development of literature. On the other hand, the field of natural language processing (which treats language automatically) has focused on the universal characteristics of language. For example, semantic disambiguation or category classification of text have been issues to study in natural language processing. However, language contains exceptions or ambiguities because it is used by humans. If such ambiguous information can be appropriately handled, it would be possible to make computers more useful for humans and to deepen communication between human and machine. Social Networking Sites (SNS) are one of the most popular services made possible by the Internet. SNS such as Twitter and Facebook produce new economic value far beyond their basic function as communi-

cation tools. For example, there are studies using web data as marketing resources to investigate consumer buying patterns.

If we can analyze those media sites with high update frequency and high instancy, such as Twitter, the analysis results can be applied to better marketing strategy.

Our study focused on Twitter, which has been getting a lot of attention recently as a research resource on big data. We sought to analyze the changes in meaning of the words and phrases. In particular, we analyzed net slang in chronological order and investigated how the contexts where the slang was used changed.

2. Related Research

In the past, a few studies tried to treat slang as an engineering issue. [1] and [2] studied the net slang extracted from the web. However, there are not many studies that targeted construction of net slang dictionaries or semantic understanding (such as NetSlangWeb3 and NetSlangWeb4), partially due to ambiguous definition of net slang, its semantic polysemy or diversity. To get a sense of slang, focus should be placed on the context words co-occurring with the target slang.

If the sense of the word cannot be understood even after reference to its context words, its sense might not be fixed. When the apparent tendency of the context words is fluid, it is difficult to try to make sense of the word automatically. Therefore, invariant information should be focused on instead of focusing on semantic fluctuation.

In their language model, Suzuki et al. [5] proposed a method to estimate parts of speech of unknown words by using vocabulary divided PLSA (VD-PLSA).

Mikolov Mikolov proposed a method that can calculate the semantic relation between words by converting each word into a vector expression called a distributed representation. In this method, the context words near the target word in the corpus are trained as features. A problem of this method is that it cannot understand the majority of the meanings because it does not consider the time-series change in the corpus.

Murawaki [7] studied online unknown expression acquisition for extraction of vocabulary. Their study removed the systemic errors by using notation variations and then constructed the n-grams. One problem of this method is that it depends on the output of the morphological analyzer in preprocessing.

Some studies (e.g., [8], [9], [10]) focused on onomatopoeia, which are Japanese unknown expressions having a huge number of variations. Onomatopoeia differs from new words such as slang because onomatopoeia is patterned expression, which means that the meaning of the word or sensibility can be estimated by focusing on its notation. On the other hand, a dictionary known as mecab-ipa-neologd [a] registers new words. Many new words, such as proper nouns, can be split by using

[a] https://github.com/neologd/mecab-ipadic-neologd

this dictionary. One of its disadvantages is that the size of the dictionary enlarges each time it is updated with new words.

The aim of our study is to analyze how slang changes when it is used by many people. Therefore, we did not automatically extract unknown words from text data. In this paper, we targeted those slang words that can be split correctly with mecab-ipa-neologd.

3. Target Data

3.1. *Tweet Database Collection*

To analyze slang, a huge amount of Tweet data was collected from Twitter. As a collecting condition, standard words or slang words were used as queries and only the matched Tweets were registered into the corpus. The Tweets were merged according to the collected dates.

3.2. *Word Segmentation of Tweets*

The keywords were extracted from the collected Tweets. Concretely, we decided the part of speech by morphological analysis using a Japanese morphological analyzer, and extracted the basic form of a word that met the conditions of that part of speech.

4. Proposed Method

The proposed method analyzes how the usage of a word changes in time series based on the collected Tweets. However, it is difficult to obtain the changes correctly by only analyzing those changes during a short period. The meaning or usage can be estimated from word co-occurrence relation. For example, we can obtain the distributed expressions of a word from the context features around the target word by Mikolov's method. However, this cannot extract the relation between words in the same document by considering only the context words around the target word. In this study, we focused on the latent topic expressed by the document.

LSA and pLSA are famous in the research of latent semantic analysis. Latent Dirichlet Allocation (LDA) has been often used in studies on latent topics. By analyzing the latent topic, it is possible to retrieve documents with similar topics or obtain the word set used in the similar topics. If the latent topic is analyzed with time series, we will be able to observe the tendency of topic change of the word.

4.1. *Topic Analsysis*

To analyze the topic, we used LDA. LDA makes it possible to express the latent topics based on topic occurrence probability of each word. If the topic is extracted from Twitter based on LDA by focusing on a specific day, a set of keywords strongly related to the Tweets including slang words is obtained with each word's occurrence

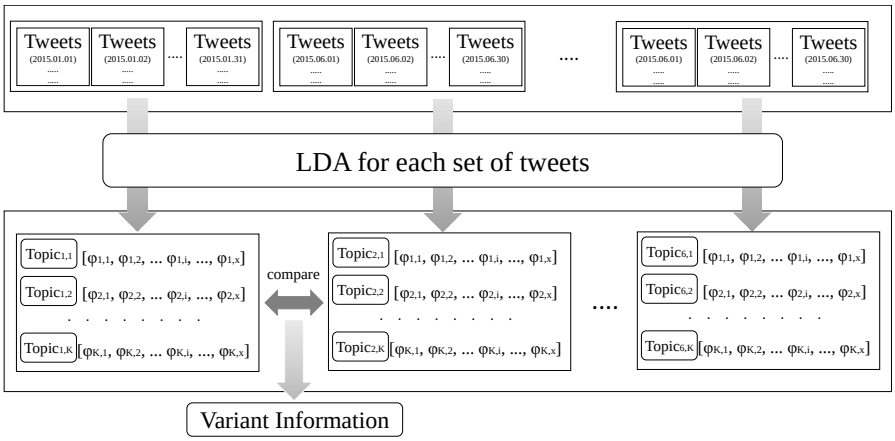


Fig. 1. Flow of comparing topics by using LDA

probability. It is thought that the keyword set variants can be obtained by repeating in each day. Word distribution in each topic can be expressed as formula 1 and 2 after sampling the topic for each word. ϕ_x indicates the occurrence probability of a word wi in the topic x. $freq_{x,i}$ indicates the number of words i belonging to topic x. $freq_x$ indicates the total number of words belonging to topic x. K is the number of the kinds of words. alpha is the hyper parameter.

$$\phi_{x,i} = \frac{freq_{x,i} + \alpha}{freq_x + K\alpha} \tag{1}$$

$$worddist_x = (\phi_{x,1}, \phi_{x,2}, ..., \phi_{x,i}, ..., \phi_{x,n}) \tag{2}$$

Because not all of the Tweets can be collected, it might not be possible to detect the changes in one day. Therefore, we analyzed the changes in one-month increments. In this study, we used a corpus that collected Tweets during seven months (from December 2014 to June 2015).

It is possible to find the most related topic to a certain word on a given day by calculating the occurrence probability of the word for each topic. We calculated the similarity degree between the word distributions on consecutive days by using the topic with the highest occurrence probability as the topic of the word on a given day.

The calculation of the similarity of the word distributions $ws_i(t, t+1)$ on the consecutive days $t, t+1$ of a word i is shown in formula 3. $worddist_{i,t}$ and $worddist_{i,t+1}$ express the topic with highest occurrence probability of word i at the points of t and $t+1$ as a vector of occurrence probability of each word on the topic.

$$ws_i(t, t+1) = \frac{worddist_{i,t} \cdot worddist_{i,t+1}}{|worddist_{i,t}| \times |worddist_{i,t+1}|} \tag{3}$$

4.2. *Affective Analysis*

Generally, it is considered that the latent topic expresses semantic information on the topics. Slang is thought to include a large amount of sensibility information [11],[12],[13]. In addition, it is necessary to analyze not only the semantic aspect but also the affective variant of the slang. Therefore, we used the emotion expressions included in the Japanese Appraisal Expression Dictionary [14] and Emotion Expression Dictionary [15] as the labeled sensibility information. In the latent topics analyzed by LDA, the sensibility keywords might be underestimated because analysis is exclusively based on nouns. Therefore, we expressed the sensibility vector of a slang word based on the strength of co-occurrence between that slang word and sensibility words in Tweets without using the LDA analysis result. We thought that sensitivity change of the slang word could be obtained by observing the time series change of the sensibility vector. The method to analyze the sensibility is explained as follows:

(1) Only sensitivity words and slang words are extracted from the text data that exclusively includes content words in the Tweets. The strength of co-occurrence between the slang words and the sensitivity words is calculated based on self mutual information from the co-occurrence frequency in the same Tweets. The equation of the self mutual information is shown in the formula 4. $cofreq_{x,y}$ indicates the co-appearance frequency between a slang word x and sensitivity word y in the same sentence. $freq_x$ and freqy indicate the frequencies of the slang word x and the sensitivity word y. N indicates the total frequencies of slang words and sensitivity words.

$$MI_{x,y} = log\frac{cofreq_{x,y} \times N}{freq_x \times freq_y} \tag{4}$$

(2) We analyzed sensitivity change of the slang word by calculating the summation of the positive/negative values of the co-occurring sensitivity words. The equation 5 calculates the positive/negative value of the slang word when ew_x is defined as a set of co-occurrence sensitivity words to the slang word x. $sign_i$ indicates the sign of the sensitivity word i, and $MI_{x,i}$ indicates the self mutual information.

$$pn\_value_x = \sum_{i \in ew_x} MI_{x,i} \times sign_i \tag{5}$$

Using the following process to calculate what sensitivity words the slang words are similar to, we detected the change of sensibility. The analysis result is shown in figures 2 and 3. These figures suggest that the values of positive slang words were stable while the values of some negative slang words had wider range of change. Some slang words were originally used to weaken the negative meaning of the word. It is thought that the negative slang words are sometimes used with strong impressions.

word	Jan.	Feb.	Mar.	Apr.	May	Jun.	Tendency
ネトゲー (Netoge-)	0.204	1	1	0.442	1	1	
シモネーター (Shimone-ta-)	1	1	0	0.751	1	1	
マッタリ (Mattari)	0.752	1	1	1	1	0	
根明 (Neaka)	1	0.537	1	1	1	0.227	
ゲト (Geto)	0.509	0.88	0.76	0.75	0.866	1	
WK (WK)	1	1	1	0.696	0.564	0.564	
デコ電 (Dekoden)	0.664	1	1	0.708	0.739	0.83	
妄女 (Moujo)	1	1	0	1	1	1	
キレカワ (Kirekawa)	0.714	1	1	1	0.6	1	
デコメ (Dekome)	0.556	1	1	1	0.909	0.92	
ジョイナス (Joinus)	0.398	1	1	1	1	1	
メチャ (Mecha)	0.856	1	1	1	1	1	
チリモン (Chirimon)	1	1	1	1	0.894	1	
2ch脳 (Nichannou)	1	1	1	1	1	1	
メッチャ (Meccha)	1	1	1	1	1	1	

Fig. 2. Analysis result of emotion change(Positive Emotion)

5. Experiment

5.1. *Analysis of Slang Changes by LDA*

As the experimental data, the tweet data collected during seven months were used. The details of the data are shown in Table 1.

Gibbs LDA++[b] was used as a tool to estimate LDA parameters. In the experiment, the number of topics was set as 10 because the appropriate number of topics was unknown. Other parameters were set to the default values.

[b]http://gibbslda.sourceforge.net/

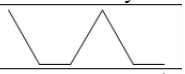
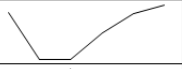
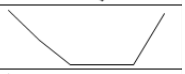
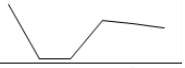
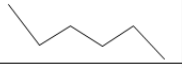
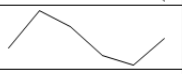
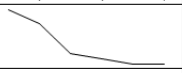
word	Jan.	Feb.	Mar.	Apr.	May	Jun.	Tendency
ノンパラ (Non-Para)	-0.52	-1	-1	-0.52	-1	-1	
汚超腐人 (Ocho-Fujin)	-0.563	-1	-1	-0.753	-0.573	-0.497	
エロカッコイイ (Erokakkoii)	-0.241	-1	0	-1	-0.867	-1	
チョバチョブ (Chobachobu)	-1	0	0	-1	-1	-1	
チョベリブ (Choberibu)	-1	0	0	-1	-1	-1	
ネチズン (Nechizun)	-0.393	-0.812	-0.747	-1	-0.576	-0.307	
キショイ (Kishoi)	0.076	-0.515	-1	-1	-1	0	
当り屋 (Atariya)	-0.093	-1	-1	-0.361	-0.401	-0.479	
ガクガクブルブル (Gakugakuburuburu)	-0.359	-1	-0.442	-0.398	-0.127	-1	
ネトウヨ (Netouyo)	-0.506	-0.568	-0.538	-0.57	-0.539	-0.59	
離活 (Rikatsu)	-0.667	0.037	-0.276	-0.818	-1	-0.5	
スケ番 (Sukeban)	0	-1	0	-1	-0.162	-1	
パワハラ (Pawahara)	-0.423	-0.456	-0.532	-0.545	-0.558	-0.559	
チャケバ (Chakeba)	0	0	0	-1	-1	-1	

Fig. 3. Analysis result of emotion change(Negative Emotion)

Table 1. Summary of Tweet Data

Number of sentences	62,727,275
Number of nouns	517,212,723
Noun per a sentence	8.25

The Tweet data collected during seven months were used as the experimental data. The details of the data are shown in Table 1. GibbsLDA++ [c] was used as a tool to estimate LDA parameters. In the experiment, the number of topics was set

[c] http://gibbslda.sourceforge.net/

as 10 because the appropriate number of topics was unknown. Other parameters were set to the default values.

The similarity degree was calculated between the word distributions obtained at the point of t and its neighboring point of $t+1$, and the average value of similarity and the standard variation of similarity were calculated.

When the standard variations of similarity were higher, the similarity values varied, showing that the topic did not change a certain amount. On the other hand, when the average of the similarity degrees was lower, the word distributions were different each time, showing that the topic was not stable. Therefore, the value obtained from dividing the standard variations by the average value was defined as the variation score. The slang words with higher variation scores were defined as variation slang and the slang words with lower variation scores were defined as invariance slang.

The equation 6 calculates the average of the topic similarity values $ws_i(t, t+1)$ of a word i during consecutive N days from t to $t+1$. The equation 7 calculates the standard variation of the topic similarity. The variation score is calculated by the equation 8.

$$avg(ws_i) = \frac{1}{N-1} \times \sum_{t=1}^{N-1} ws_i(t, t+1) \tag{6}$$

$$sd(ws_i) = \sqrt{\frac{1}{N-1} \times \sum_{t=1}^{N-1} (ws_i(t, t+1) - avg(ws_i))^2} \tag{7}$$

$$vd(ts_i) = \frac{sd(ws_i) + 1}{avg(ws_i) + 1} \tag{8}$$

Table 2 shows the example of slang words that had high/low variation degree scores. 'pn' indicates the emotional polarity value of each slang word based on the co-occurrence with sensibility words.

As seen in the slang words that had high variation degree scores such as "RT," "*Otsu*" or "*Maji*," these words are versatile and already have been used widely. Because usages of these slang words are fixed and simple, they easily can be used in various contexts. We can also point out that many of them seem to be polysemous words, or to have enlarged the range of their senses while spreading widely.

On the other hand, the slang words that had low variation degree scores seemed to be expressions that were not yet used widely. It is estimated that there are few users of these words. They tend to be used in fixed context and have been gradually falling into disuse to become what we call "dead words."

5.2. *Variation Analysis Based on Word2Vec*

To investigate whether the meaning of the slang word changes or not, we focused on the context instead of the topic. In the experiment, the semantic vector of the

Table 2. Example of slangs with low/high variant degree

Low vd slangs	vd	pn	High vd slangs	vd	pn
ダッサイ (*Dassai*)	0.09	0.378	RT	18.8	0.313
キショ(*Kisho*)	0.17	0	乙 (*Otsu*)	17.3	-0.123
オソロ (*Osoro*)	0.18	0.379	神 (*Kami*)	16.7	0.259
くれない族 (*Kurenaizoku*)	0.21	-0.264	ネット (*Net*)	16.1	-0.115
ぱくり (*Pakuri*)	0.21	0.309	草 (*Kusa*)	15.7	0.060
メタポン (*Metapon*)	0.22	0	マジ (*Maji*)	15.6	-0.034
自宅難民 (*Jitakunanmin*)	0.22	0	激 (*Geki*)	15.3	0.214
ノンパラ (*Nonpara*)	0.23	-0.840	ググれカス (*Gugurekasu*)	15.2	0.254
バメン (*Bamen*)	0.24	0.167	微妙 (*Bimyo*)	14.9	0.042
ドクハラ (*Dokuhara*)	0.25	-0.450	姫 (*Hime*)	14.9	0.169
同中 (*Onachu*)	0.32	0.055	リアル (*Real*)	14.7	0.018
こそアド (*Kosoado*)	0.33	0	天然 (*Tennen*)	14.6	0.451
お祈りメール (*Oinorimail*)	0.34	-0.237	ツボ (*Tsubo*)	14.5	0.054
生存フラグ (*Seizonflag*)	0.36	0.723	ガチ (*Gachi*)	14.2	-0.102
IK	0.36	0.334	イケメン (*Ikemen*)	14.1	0.306

slang word was created by using word2vec [d]. Only the content words were used for learning to create the semantic vector of the slang word. As the learning parameters, we set the window size as 5 and the vector dimension as 200. We used skip-gram in order to consider semantic similarity.

We investigated what changes occur during one month by focusing on the top 100 similar words for each slang word. The cosine similarity was used as similarity value. When the variation degree score was calculated based on the monthly average value of similarity degrees and the standard deviations, the result was obtained as shown in table 3.

For the emotional expressions and the words that are often used recently, such as "*Gekiokopunpunmaru*," variation degree scores were high because such words could be used in various contexts.

On the other hand, there were many unfamiliar words among the words that had low variation degree scores. It was an unexpected result that the expressions such as "'*Arege*" had low variation degrees even though they can be used in various contexts.

Table 4 shows the top 10 synonym candidates for "*Arege*" obtained by using the training model trained by word2vec on all Tweet corpora. The synonyms of "*Arege*" rarely appeared and there were few words having the same meanings. The slang word "*Arege*" means "something useless although it looks useful."

[d] https://code.google.com/p/word2vec/

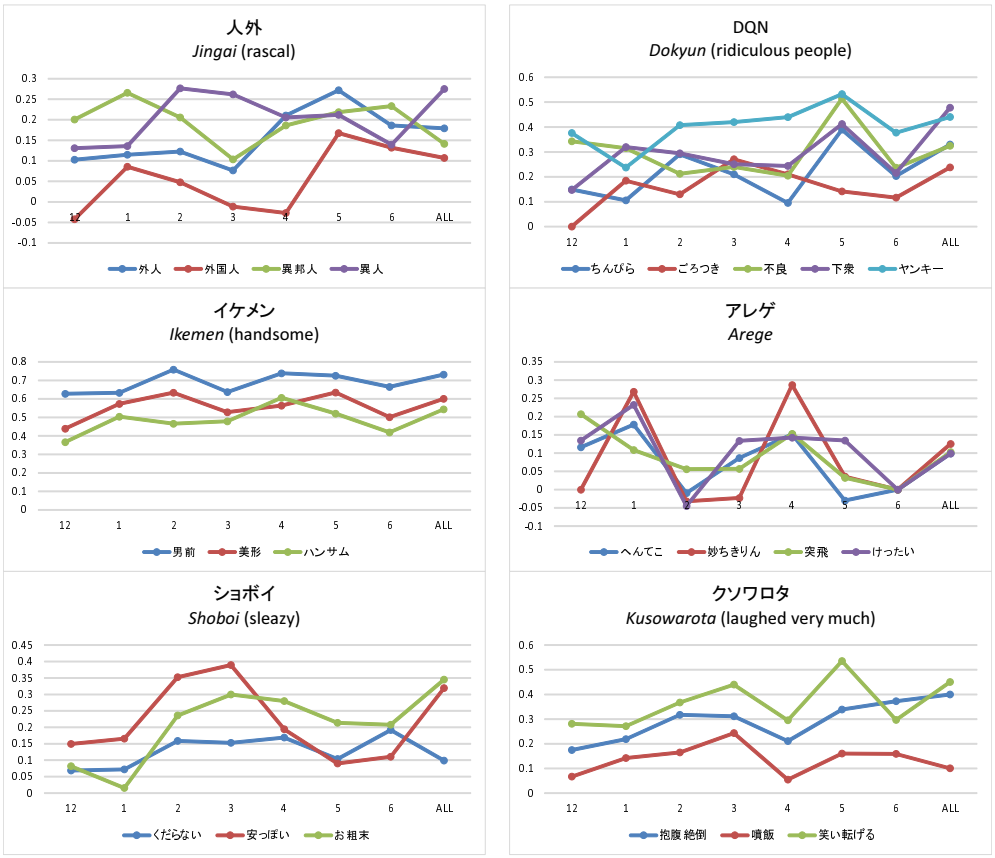


Fig. 4. Similarity change between the slang and the corresponding standard words

The expression tends to be used with a more subtle nuance in the topics favored by people called *Otaku* or geeks. Because the expression does not have clear meanings or substitute words, it seems not to appear in the synonyms.

6. Discussions

In the experiment, the variation of slang was analyzed with two methods. One was the method based on the topic and the other was the method based on the context-based semantic vector. We discussed the results in detail. First, we would like to think about the problem of judging the word as a dead word when its variation degree score was low. A dead word is a word that used to be used often, but only temporarily for a certain period of time, and which is rarely being used now. Low variation degree score was likely due to the low frequency of use. Therefore, because

Table 3. The slangs with low/high variant degree based on the context similarity

Low vd slangs	vd	pn	High vd slangs	vd	pn
パギャル (*Pagyaru*)	0.001	0.056	キティラー (*Kitira-*)	0.22	0.484
ハラッサー (*Harassa-*)	0.001	-0.141	鮫 (*Same*)	0.20	0.027
テクハラ (*Tekuhara*)	0.002	0	激おこぷんぷん丸 (*Gekiokopunpunmaru*)	0.21	-0.108
マゴギャル (*Magogyaru*)	0.002	0.333	スレ (*Sure*)	0.21	-0.140
ピンアポ (*Pinapo*)	0.002	0	般若 (*Hannya*)	0.21	-0.073
CKY	0.002	0.3808	VIP	0.20	0.04
ユニクラー (*Yunikura-*)	0.003	0.502	安全牌 (*Anzenpai*)	0.19	0.122
ブヒアゲ (*Buhiage*)	0.003	-0.167	鯖 (*Saba*)	0.19	-0.048
空気嫁 (*Kuukiyome*)	0.003	-0.400	トゥー (*Tuu*)	0.19	0.101
ニカニカ (*Nikanika*)	0.004	0.667	アニドル (*Anidoru*)	0.19	0.691
タカピー (*Takapi-*)	0.004	0.437	バリチュー (*Barichuu*)	0.18	0.112
アレゲ (*Arege*)	0.004	0.133	確信犯 (*Kakushinhan*)	0.18	-0.148
フケ専 (*Fukesen*)	0.004	0.167	ギタフリ (*Gitafuri*)	0.18	-0.075
根明 (*Neaka*)	0.004	0.794	ミーハー (*Mi-ha-*)	0.18	0.065
モロバレ (*Morobare*)	0.004	0.333	希ガス (*Kigasu*)	0.18	0.308
MKY	0.004	0.321	デュクシ (*Dyukushi*)	0.17	0.138
イタメシ (*Itameshi*)	0.004	0.393	除草 (*Josou*)	0.17	0.159
CIK	0.005	0.333	ゼニガメ (*Zenigame*)	0.17	-0.125
ブルセラ (*Burusera*)	0.005	0.047	ゆるきゃら (*Yurukyara*)	0.17	-0.098

Table 4. The context similar words of "*Arege*"

Similar Words	Similarity
OASYS	0.46
D 言語 (D Language)	0.46
スマートフォーン (Smart Phone)	0.45
yurinekofuran	0.45
高周波回路 (high-frequency circuit)	0.44
プロセスルール (process rule)	0.44
タイプフェイス (type face)	0.44
pisponpan	0.43
MS ゴシック (MS Gothic)	0.43
階差機関 (difference engine)	0.43

these words can become trend words later, it is not possible to conclude that they are dead words.

Next, we would like to think about the effectiveness of the similarity calculation between word distributions in obtaining the semantic variation. The distribution of the slang word was decided based on the topic that showed the highest occurrence probability of the slang word. However, when there are other topics with the same degree of occurrence probability, it means that the analysis was insufficient. This problem might often occur when the slang word has multiple meanings.

Researchers should focus on the part of speech of the slang word when the slang word includes affective variation. When the slang word is used as an adjective, the affective variation might be small; when the slang word is used as a noun, the affective variation might be large. Therefore, it might be possible to improve the emotion estimation accuracy of sentences including slang words by judging the parts of speech of those words.

Here, we would like to discuss whether it is possible to estimate emotion of a slang word from the topic of that word or a set of similar words. We assumed that the sensibility expressed by the slang word could be understood by searching the tendency of the topic or context of the slang. In this paper, we conducted sensibility analysis based on the co-occurrence with sensibility words; however, we could only analyze positive/negative change. One of the reasons is complexity of the sensibility expressions.

For example, in the sentence "*Tencho ni okorarete furubokko ni atta. Kanashisu*" (I got chewed out and beaten up by a store manager. I am too sad.), considering that "*furubokko*" and "*Kanashisu*" are co-occurring with the sensibility word "*Okorare*," the sensibility of each slang word could be estimated. However, in fact, because the event mentioned in the sentence indicates the relation of cause and result (the subject became sad because he/she was scolded), it is not possible to associate these words by co-occurrence alone. In particular, because the slang "*Kanashisu*" means "too sad," it should not be associated with the sensibility word "*Okorare*," which is associated with "Angry."

When word2vec is used to judge similarity of words based on the context, an expression such as "*Kuyashisu*" can be obtained as a candidate that does not co-occur with "*Kanashisu*" but has similar usage. However, because the expression is semantically different, we should not judge the word in the same way as the sensibility word. Therefore, it seems insufficient to judge sensibility of the word by only using similar words.

In the future, to construct a slang sensibility dictionary, it would be necessary to consider co-occurrence relation with the words that have strong relation with the sensibility words, and not just use direct co-occurrence with the sensibility words.

7. Conclusions

In this research, we collected sets of Tweets from the microblog Twitter, which offers an advantage for getting real-time posts. Then, we analyzed the variant of the slang word by analyzing the topic keyword vector in time-series. We obtained

the topic keywords by calculating the latent topic based on LDA and obtaining the word's occurrence probability. Then, we proposed two methods for extracting the variant information of the slang word:

(i) the method to detect the change strength of the variation
(ii) the method to analyze the sensibility variant by calculating the similarities between topic keyword vectors.

In the experiment, the similarities were calculated by using topic or context information that was extracted for each period. We analyzed the word variants by using a new index variant degree score. As a result, we found it difficult to identify the period during which semantic change occurred by using only the variant degree score. It is necessary to compare the word to the expressions that appeared with similar frequency.

In the case of judging dead words, we propose the effective corpus collection method to rapidly recognize the sense of words that were recently made. This may be the best method to judge whether the unknown expression is slang or not, by using the difference of the context distribution between the slang word and the other words as feature. It should be analyzed by considering the difference for each situation such as user or scene.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 15H01712, 15K16077, 15K00425.

References

1. K. Manuel, K. V. Indukuri, and P. R. Krishna. *Analyzing Internet Slang for Sentiment Mining*, Proceedings of the 2nd Vaagdevi International Conference on Information Technology for Real World Problems (VCON), pp. 9–11, 2010.
2. W. Muliady, H. Widiputra. *Generating Indonesian slang lexicons from twitter*, Proceedings of the 2nd International Conference on Uncertainty Reasoning and Knowledge Engineering (URKE), pp. 123–126, 2012.
3. M. Arora, and V. Kansal. *A framework for informal language: Opinion mining*, Proceedings of the International Conference on Computing, Communication & Automation (ICCCA), 2015.
4. K. Matsumoto and F. Ren. *Construction of Wakamono Kotoba Emotion Dictionary and Its Application*, Computational Linguistics and Intelligent Text Processing, Vol.LNCS6608, pp.405–416, Feb. 2011.
5. M. Suzuki, N. Kuriyama, A. Ito and S. Makino. *Automatic clustering of part-of-speech for vocabulary divided PLSA language model*, Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, 2008. NLP-KE '08, , pp. 1–7, 2008.
6. T. Mikolov, K. Chen, G. Corrado, and J. Dean. *Efficient Estimation of Word Representations in Vector Space*, In Proceedings of Workshop at ICLR, 2013.
7. Y. Murawaki and S. Kurohashi. *Online Japanese Unknown Morpheme Detection using Orthographic Variation*, In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pp. 832–839, Valletta, Malta, 2010.5.
8. R. Sasano, S. Kurohashi, and M. Okumura. *A Simple Approach to Unknown Word Processing in Japanese Morphological Analysis*, Journal of Natural Language Processing **21**(6), pp. 1183–1205, 2014.

9. 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 **19**(5), pp. 367–379, 2012-12-14.
10. Y. Shimizu, R. Doizaki, and M. Sakamoto. *A System to Estimate an Impression Conveyed by Onomatopoeia*, Transactions of the Japanese Society for Artificial Intelligence **29**(1), pp. 41–52, 2014.
11. K. Matsumoto, K. Kita, and F. Ren. *Emotion Estimation of Wakamono Kotoba Based on Distance of Word Emotional Vector*, Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp. 214–220, Tokushima, Nov. 2011.
12. K. Matsumoto, K. Kita, and Fuji Ren. *Emotion Estimation from Sentence Using Relation between Japanese Slangs and Emotion Expressions*, Proceedings of 26th Pacific Asia Conference on Language,Information and Computation, pp. 377–384, Nov. 2012.
13. F. Ren and K. Matsumoto. *Semi-automatic Creation of Youth Slang Corpus and Its Application to Affective Computing*, IEEE Transactions on Affective Computing, Issue 99, 2015.
14. M. Sano. *Classification of evaluative expressions in Japanese : an Appraisal perspective*, IEICE technical report. Natural language understanding and models of communication 110(400), pp. 19–24, 2011.
15. A. Nakamura. *Emotion Expression Dictionary*, Tokyodo publisher, 1993.

Kazuyuki Matsumoto

He received the Ph.D degree in 2008 from Faculty of Engineering, the University of Tokushima. He is currently an Assitant Professor at the university of Tokushima. His research interests include affective computing, Emotion Recognition and Natural Language Processing. He is a member of IPSJ, IEICE, IEEJ.

Minoru Yoshida

He is a lecturer at the Department of Information Science and Intelligent Systems, University of Tokushima. After receiving his BSc, MSc, and PhD degrees from the University of Tokyo in 1998, 2000, and 2003, respectively, he worked as an assistant professor at the Information Technology Center, University of Tokyo. His current research interests include Web Document Analysis and Text Mining for the Documents on the WWW.

Seiji Tsuchiya

He received the B.E., M.E., and Ph.D. degrees from Doshisha University, Kyoto, Japan, in 2000, 2002, and 2007, respectively. Since 2002, he worked with Sanyo Electric Co., Ltd. In 2007 he became an Assistant Professor in the Institute of Technology and Science, the University of Tokushima. Since 2009, he worked with Faculty of Science and Engineering, Doshisha University, Kyoto, Japan, as an Assistant Professor, and in 2011, he became an Associate Professor. He has been engaged in natural language processing, intelligent information processing. He is a member of the Japanese Society for Artificial Intelligence, the Association for Natural Language Processing, the Institute of Electronics, Information and Communication Engineers, the Information Processing Society of Japan and Japanese Cognitive Science Society.

Kenji Kita

He received the B.S. degree in mathematics and the PhD degree in electrical engineering, both from Waseda University, Tokyo, Japan, in 1981 and 1992, respectively. From 1983 to 1987, he worked for the Oki Electric Industry Co. Ltd., Tokyo, Japan. From 1987 to 1992, he was a researcher at ATR Interpreting Telephony Research Laboratories, Kyoto, Japan. Since 1992, he has been with Tokushima University, Tokushima, Japan, where he is currently a Professor at Faculty of Engineering. His current research interests include multimedia information retrieval, natural language processing, and speech recognition.

Fuji Ren

He received his Ph.D. degree in 1991 from Hokkaido University, Japan. From 1991, he worked at CSK, Japan, where he was a chief researcher of NLP. From 1994 to 2000, he was an associate professor in the Faculty of Information Sciences, Hiroshima City University. He became a professor in the Faculty of Engineering of the University of Tokushima in 2001. His research interests include natural language processing, articial intelligence, language understanding and communication, and affective computing. He is a member of IEICE, CAAI, IEEJ, IPSJ, JSAI, and AAMT, and a senior member of IEEE. He is a fellow of the Japan Federation of Engineering Societies. He is the president of the International Advanced Information Institute. Faculty of Engineering, University of Tokushima, 2-1, Minamijyousanjima-cho, Tokushima 770-8506 Japan