Knowledge Extraction from Web Reviews Using Feature Selection Based on Onomatopoeia

Fumiaki Saitoh(✉), Hikaru Aoki, and Shohei Ishizu

Department of Industrial and Systems Engineering, College of Science and Engineering, Aoyama Gakuin University, 5-10-1 Fuchinobe, Chuo-Ku, Sagamihara City, Kanagawa, Japan
saitoh@ise.aoyama.ac.jp

Abstract. In the field of Buzz marketing, it is important to extract knowledge to improve products and services from the voice of the customer represented by customer reviews. In Japanese web review sentences, words that co-occur with onomatopoeia it has been confirmed that easy to combine with use sense of product. For sensory evaluation using a products can be easily associated with the satisfaction is obvious, onomatopoeia can be expected to contribute in knowledge extraction on customer satisfaction. A knowledge model for customer satisfaction is constructed by a regression tree that co-occurrence words with onomatopoeias are used as explanatory variables. Effectiveness of the proposed method I was confirmed through the analysis for the customer review data of ramen shop in Tokyo. The knowledge model acquired by our approach contained many words associated with noodles and food, on the other hand the normal regression tree model was included many meaningless words and unrelated words.

Keywords: Text mining · Online reviews · Voices of the customer (VOC) · On-omatopoeia · Regression tree

1 Introduction

This research presents a method of extracting knowledge about customer satisfaction from customer review data. In the field of buzz marketing, it is important to extract knowledge in order to improve products and services by taking cues from the voice of the customer expressed in customer reviews. In this study, knowledge models are constructed by using the regression tree method in order to predict customer satisfaction.

However, it is difficult to understand the regression tree because certain words that have no relationship with the evaluation object are also included in the model. Prior studies on Japanese web review have confirmed that words which co-occur with onomatopoeia frequently comment on the utility of the product. For sensory evaluation, using a products can be easily associated with the satisfaction is obvious. Onomatopoeia is expected to generate evaluative feedback and support knowledge extraction on customer satisfaction.

C. Stephanidis (Ed.): HCII 2015 Posters, Part I, CCIS 528, pp. 650–655, 2015.
DOI: 10.1007/978-3-319-21380-4_110

Therefore, in this study, we focus on feature selection based on onomatopoeia for the construction of knowledge models. Co-occurrence words with onomatopoeias are used as explanatory variables in the regression tree. By using this model, factors affecting customer satisfaction (or dissatisfaction) will be determined.

Effectiveness of the proposed method was confirmed by analyzing the customer review data of a shop serving ramen (Japanese noodle soup dish) in Tokyo. We confirmed the intelligibility of knowledge acquired by using the proposed method, wherein we compared the contents of knowledge of the normal regression tree and the knowledge that was acquired through the proposed method. The knowledge model constructed through our approach contained many words associated with noodles and food. On the other hand, the normal regression tree model included many irrelevant and unrelated words. We evaluated the prediction accuracy of customer satisfaction by using cross-validation, and confirmed that the proposed method makes knowledge extraction possible without sacrificing prediction accuracy.

2 Knowledge Extraction on Customer Satisfaction

2.1 Feature Selection Based on Onomatopoeia

In this study, we focus on the onomatopoeia in feature selection of text mining. Onomatopoeia has the property that easily related to the feeling of use of the product in Japanese language (see Fig. 1). We thought words that co-occur with onomatopoeia are easy to combine with product evaluation and satisfaction. Therefore, onomatopoeias can be expected as powerful extraction toot of words that that are directly involved in product evaluation.

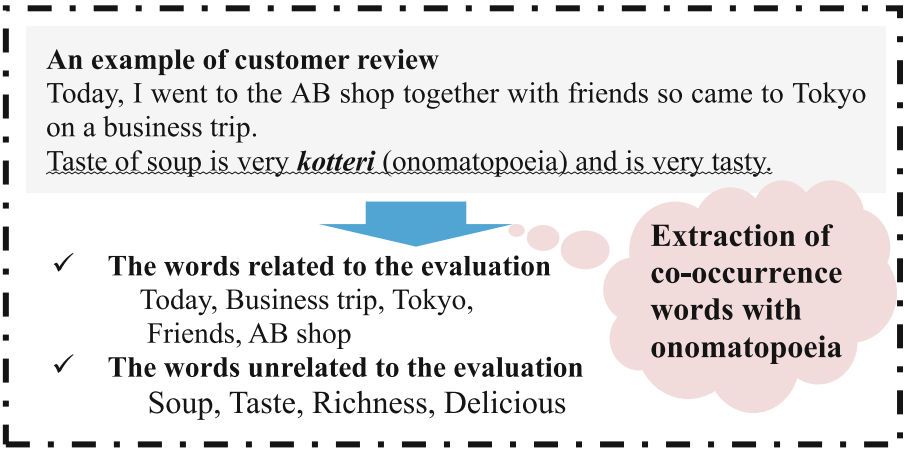


Fig. 1. Schematic diagram of the feature selection in the proposed method

2.2 Knowledge Extraction Using a Regression Tree

In our proposal, the knowledge about the factors that affect the level of satisfaction from the text data of the review sentences are extracted by the regression tree. It becomes possible to understand without reading all the review sentences by constructing a model that predicts satisfaction by the combination of conditional branches. In the prediction of satisfaction, the knowledge constructed by regression tree applied the word frequency in conditional branch, and predictive model will be available as a knowledge that can be understood.

The processing flow of the proposed method is as follows.

Step.1:	Words that co-occur with onomatopoeia are extracted, and they are treated as a word set *O*.
Step.2:	All the words contained in the data set to be analyzed are treated as a word set A.
Step.3:	A words set A∩O that has common elements to A and O is extracted, in this step.
Step.4:	To learn a regression tree, term frequencies of $A \cap O$ are used as explanatory variables, and customer satisfaction are used as objective variables.

3 Experimental

3.1 Experimental Settings

Effectiveness of the proposed method was confirmed through the analysis for the customer review data of ramen (Japanese noodle soup dish) shop in Tokyo. Parts of speech that used for analysis are verbs and adjectives. We adopted these frequencies as explanatory variables, and we utilized customer satisfaction, which is represented by five stars as objective variable. Before (or after) 5 words of onomatopoeia are co-occurrence range.

In order to confirm the prediction accuracy of the knowledge that was constructed by regression tree, we evaluated the model by Leave-one-out Cross-validation (LOOCV). We applied CART algorithm to data sets A∩O and A respectively, and compared the learning result of regression tree.

3.2 Experimental Results

This section describes the experimental results. Table 1 shows the part of examples of Japanese onomatopoeia that frequently appear in our data set. Table 2 shows prediction accuracy of each model that calculated by mean of residual and its standard deviation.

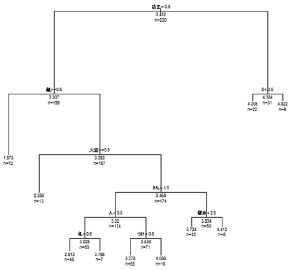
Figure 2 shows regression tree that is applied to the normal data and Fig. 3 shows regression tree that has been acquired by the proposed method.

Table 1. Examples of Japanese onomatopoeia

Japanese onomatopoeia	Alphabetical notation	Meanings in English
あっさり	assari	*ASSARI* describes (1) a light, plain, simple flavor, (2) To appear simple, plain, or light yet delicate.
こってり	kotteri	*KOTTERI* describes rich or heavy taste of a food.
べたべた	beta-beta	*BETA-BETA* describes (1) something sticky (2) someone clinging to another.
つるつる	tsuru-tsuru	*TSURU-TSURU* describes (1) slipping on a smooth surface (2) the sound made by someone slurping noodles.
がつがつ	gatsu-gatsu	*GATSU-GATSU* describes someone eating greedily.

Table 2. prediction accuracy

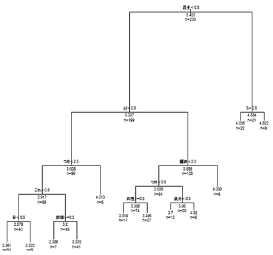
Prediction Accuracy	Word set A	Word set $A \cap O$
Mean	0.575	0.598
SD	0.554	0.483



The words used for conditional branches

Shopkeeper, Soy, Mr., 8, man, time
RYU (It is representing the Dragon in Japanese),
etc.

Fig. 2. A knowledge model acquired by the proposed method



The words used for conditional branches

Shopkeeper, Minute,
Dipping noodle, Soy sauce,
Seafood (Soup stock), Woman customer
etc.

Fig. 3. A knowledge model acquired by regression tree using a normal data

3.3 Discussion

In the comparison of two models, a clear difference in the prediction accuracy was not observed (see Table 2). On the other hand, clear difference between these models appeared on the readability and understanding. The proposed method has been included words related to evaluation objects more than the conventional method. From the above, it can be said that understandable knowledge is acquired without compromising of the prediction accuracy of satisfaction by the proposed method.

4 Conclusion

In this study, we improved the quality of the knowledge of customer satisfaction extracted from web review sentences, through feature selection based on co-occurrence of onomatopoeia. By focusing on onomatopoeia, we achieved an easy extraction of word sets tend to be associated with the evaluation. Furthermore, intelligibility and readability of the acquired regression tree (knowledge) were improved. We have confirmed the effectiveness of the proposed method through the analysis of Web Review for Tokyo ramen shop.

References

1. Sakamoto, M., Ueda, Y., Doizaki, R., Shimizu, Y.: Communication Support System Between Japanese Patients and Foreign Doctors Using Onomatopoeia to Express Pain Symptoms. J. Adv. Comput. Intell. Intell. Inf. **18**(6), 1020–1025 (2014)
2. Komatsu, T.: Choreographing Robot Behaviors by Means of Japanese Onomatopoeias. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, pp.23–24 (2015)
3. Lertsumruaypun, K., Watanabe, C., Nakamura, S.: Onomatoperori: Recipe Recommendation System Using Onomatopeic Words, IPSJ SIG Technical report vol.73, no.6, pp.1–7 (2009) (in Japanese)
4. Fukushima, H., Araki, K., Uchida, Y.: Disambiguation of Japanese Onomatopoeias using Nouns and Verbs. In: Proceedings of 17th International Conference Text, Speech and Dialogue, pp.141–149 (2014)

5. Kato, A., Fukazawa, Y., Sanada, H., Mori, T.: Extraction of food-related onomatopoeia from food reviews and its application to restaurant search. J. Adv. Comput. Intell. Intell. Inf. **18**(3), 418–428 (2014)
6. "The JADED NETWORK," http://thejadednetwork.com/sfx/ (Accessed on 18 March 2015)