A Text Mining Approach for Automatic Modeling of Kansei Evaluation from Review Texts

Atsuhiro Yamada[1], Sho Hashimoto[2] and
Noriko Nagata[3]

[1,2,3] Kwansei Gakuin University, 2-1 Gakuen Sanda Hyogo, Japan
[3] nagata@kwansei.ac.jp

Abstract. In the field of product design, it is important to meet the specific user's affective needs in addition to function and price. For that purpose, a Kansei evaluation model expressing the relationships between a low-level impression related to physical characteristics and a high-level impression (affective needs) is being constructed in the field of Kansei engineering. However, the conventional modeling method involves a lot of time and effort, because it depends on experiments involving humans and on subsequent analyses. To solve this problem, we propose a method that automatically constructs a Kansei evaluation model for product design using review texts on the Web. The method consists of three steps. First, we collect and select evaluation words with word classes and Japanese dictionaries of evaluation expressions. Second, we estimate the impression axes particular to a domain with the topic model that uses only the evaluation words. Finally, we score each product for each evaluation axis, using the frequencies of the appearance of evaluation words and term-scores. The results of the application of the method to reviews of wristwatches and the subjective evaluation experiment with topics for visual impressions show the utility and validity of the method.

Keywords: Affect, Auto scoring, Kansei evaluation model, Text mining.

1 Introduction

In the field of product design, it is important to meet the specific user's needs or feelings [1]. Particularly, with the development of technology, in mature industries where it is difficult to differentiate products provided by each company, it is important to satisfy affective needs in addition to needs such as function and price. The Kansei engineering approach is considered to be the most reliable and useful methodology to handle users' affective needs [2]; therefore, it has been successfully applied in various design domains. The impression involves a low-level impression related to physical characteristics and a high-level impression (affective needs) related to emotion. In the field of Kansei engineering, construction of a Kansei evaluation model that expresses the relationships between these impressions is one of the main subjects of study [3, 4].

In the conventional method of construction of the Kansei evaluation model, the impression axes of the domain and the scores on the axes for each product are estimated from the subjective evaluation experiment using the semantic differential (SD) method. For the preparation of this experiment, the collection of the evaluation words suitable for evaluating products, structuralization of the evaluation word space, and selection of evaluation words to be used for the experiment are carried out. Stimulation is similarly collected, structured, and selected. The experiments and analyses are required at each stage; hence, it involves a lot of time and effort.

In order to solve this problem, in the field of Kansei engineering, many studies are already being performed to automatically construct the model. Qu [1] attempted to optimize the appearance of the Web from the user's preference based on the approaches of the neural network and the evolutionary genetic algorithm. The preferences of the users are acquired through a questionnaire that evaluates using a word pair for a web image (stimulation). In addition, Lokman, Haron, Abidin, and Khalid [5] automated the clustering of evaluation words using Natphoric algorithm for the purpose of supporting the analysis process such as factor analysis. However, these methods are mainly aimed at automating the analysis process, and the processes of these systems include evaluation experiments performed by humans, so it is not complete automation without evaluation experiments. Therefore, in this research, we will consider constructing a Kansei evaluation model automatically using a review corpus instead of an evaluation experiment.

As an analysis example of a review corpus, many studies have been conducted to estimate the functions and features of products using natural language processing (NLP) techniques. Htay and Lynn [6] proposed an idea to extract opinion words or phrases for each feature by getting the patterns from the review text through some specific Part of Speech (POS) to generate a summary that can support a customer's product selection. In addition, Santosh, Babu, Prasad, and Vivekananda [7] proposed a method to construct a domain independent Feature Ontology Tree (FOT) by applying a topic model to the review in order to identify appropriate product features.

The purpose of this study is to automatically construct a Kansei evaluation model from the product reviews on the Web for product design by applying NLP methods to impressions, not to the functions and features of products. We explain a flow of the proposed method of automatic construction of the Kansei evaluation model. First, in order to collect and select evaluation words, we use POS data and Japanese dictionaries of evaluation expressions. Second, the impression axes are estimated by the Hierarchical Dirichlet Process LDA (HDP-LDA) [8], which uses only the evaluation words. Finally, the scores on the impression axes are estimated using the frequency of the appearance of evaluation words and a term-score [9] from the result of HDP-LDA. The next chapter will give a detailed description of this method.

2 Method

2.1 Preprocessing

Morphological Analysis. We conduct a morphological analysis of the sentences using MeCab [10] (a Japanese morphological analyzer) and extract nouns, verbs, adverbs, adjectives, adnominal words, and prefixes. In addition, all words are converted into base form. Furthermore, if the word class is a noun (*nai*_adjective base), it is concatenated with the previous word. If the word class is a prefix and the next word class is a noun, it is connected to that word; otherwise, it is removed.

Orthographic Disambiguation. Because there is concern about the accuracy deterioration of the model construction due to the orthographic variants, all words are automatically corrected after the morphological analysis. The words in dictionaries are also corrected similarly. The process of orthographic disambiguation is described below. The *daihyo-hyoki* of the JUMAN version 7.0 [11] (a Japanese morphological analyzer) is used for this processing. The *daihyo-hyoki* is one of the types of information on the words included in JUMAN and is a representative notation within the group of orthographic variants of that word. First, each word divided by MeCab is morphologically analyzed using JUMAN. Next, when the word is divided into two or more as a result of JUMAN, it is not corrected. When it is not divided and when representative notation is included in the semantic information, the word is corrected to the "representation notation." As stated above, one can then solve the problem of the orthographic variant.

2.2 The Collection and Selection of Evaluation Words

In order to extract only the evaluation words related to the impression of the product, the evaluation words are collected from all the words, and thereafter, the evaluation words related to the individual's feelings and behaviors are removed.

First, in order to extract evaluation words related to the impression, the words whose POS is a main adjective or noun (adjective base) are collected from all words after morphological analysis. Next, evaluation words of other POS are collected using the following three Japanese dictionaries: (1) EVALDIC ver. 1.0.1 [12, 13], (2) Japanese Sentiment Dictionary (Volume of Verbs and Adjectives) ver. 1.0 [12], and (3) Japanese Sentiment Dictionary (Volume of Nouns) ver. 1.0 [14]. The conditions of the collection are as follows: (1) included in EVALDIC; (2) included in the words given the "evaluation" tag in the Japanese Sentiment Dictionary (Volume of Verbs and Adjectives); and (3) included in the words given the "evaluation" tag in the Japanese Sentiment Dictionary (Volume of Nouns).

In a previous study [15], evaluation expressions included (a) expressions that represent evaluators' feelings, such as "enjoy" and "get upset," or behaviors that represent feelings, such as "laugh" and "cry," and (b) features of evaluation objects such as "effective" and "bad." The (a) category is called internalized (*nai-hyoka*), and (b) is called externalized (*gai-hyoka*). Because internalized expressions of an individual feelings and behaviors are not impressions of products, they are inappropriate for the estimation of the relationships between impressions. However, these two expressions are mixed in the collected evaluation words. Therefore, evaluation words categorized as internalized are removed from the collected words and evaluation words related to impressions are selected using the Japanese Appraisal Dictionary-attitudinal evaluation ver. 1.0. [16], which categorizes several evaluation words as internalized or externalized.

2.3 Estimation of Impression Axes

As used in the previous research introduced in Chapter 2, unsupervised topic models are often used for product review analysis. Accordingly, the HDP-LDA is used for estimating impression axes. LDA (Latent Dirichlet Allocation) is one of the most popular classification methods for texts. The HDP-LDA is a nonparametric Bayes extension of LDA, which can recover a proper number of topics. The HDP-LDA is used after all reviews that have evaluated that the same products are concatenated. The words used for learning are evaluation words of "value" and "impression," that is, evaluation words not included in "affect." Furthermore, evaluation words included in more than 50% or less than 0.1% of documents are eliminated. A collapsed Gibbs sampling (CGS) method for HDP-LDA is used, and models are run for 500 iterations. For hyper-parameters, we set alpha=2.0, beta=0.3, and gamma=0.01.

2.4 Estimation of Scores on Impression Axes

We assume that reviews with a high score for certain topics contain a lot of evaluation words with high importance relative to that topic. Thus, in the estimation of the scores on impression axes, the scores are given to each product using the appearance frequency of the evaluation words in the review and the importance of the word in each topic.

In this method, we use a term-score as the importance of each evaluation word in each topic. This is an application of the TF-IDF weight, which is used to judge the importance of words in a document relative to the topic model. It is indicated by the following formula:

$$\text{term-score}_{k,v} = \widehat{\beta_{k,v}} \log\left(\frac{\widehat{\beta_{k,v}}}{\left(\sum_{j=1}^{K}\widehat{\beta_{k,v}}\right)^{\frac{1}{K}}}\right) \quad (1)$$

where β, k, and v are the means of each topic term's probability, topic, and word, respectively.

Fig. 1 shows the procedure of the estimation of the scores. The score of each topic of a product is the average of the scores computed by all reviews of the product. The score of each topic and each review is calculated by the appearance frequency of the evaluation word and term-score. First, the evaluation words in the target review are extracted. Next, we sum the term-scores of the target topic of these words. This calculated value is the score of each topic and each review. If two or more of the same evaluation words are included in the same review, the term-score of the word is added to the score each time.

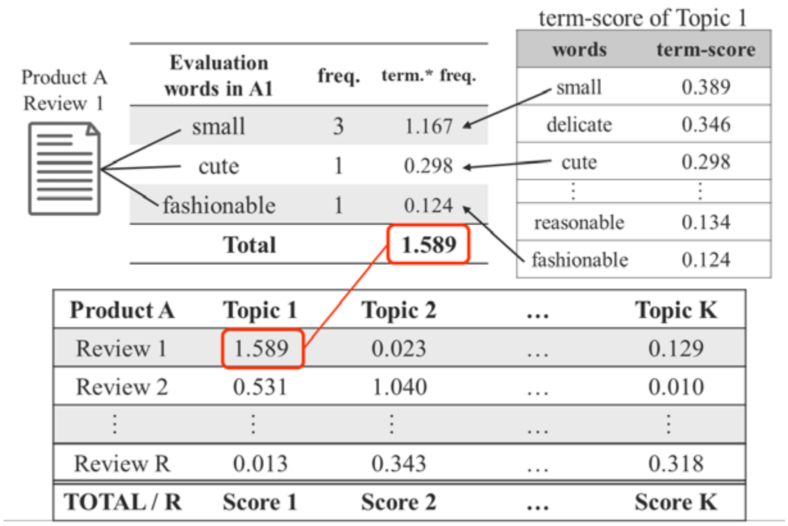


Fig. 1. Procedure of Estimation of Scores on Impression Axes.

3 Automatic Construction Experiment

3.1 Data Set

We applied the method to reviews of wristwatches in the Rakuten Data Set (https://www.nii.ac.jp/dsc/idr/en/rakuten/rakuten.html). We excluded reviews made by non-purchasers and reviews of products that had fewer than 10 reviews. We concatenated a review and its title as one review. As a result, the number of reviews was 203,086, and the number of products was 4,557. The total number of words was 4,514,325, and the total vocabulary size was 19,689.

3.2 Results

The number of collected evaluation words was 3,097. Of these, 493 words were classified as internalized, and the remaining 2,604 words were selected as externalized evaluation words. In the results for the top 10 words by appearance frequency, the internalized expressions included “do,” “think,” “favorite,” “buy,” and “want.” The externalized expressions included “good,” “satisfaction,” “reasonable,” “no,” and "few." From this result, it can be seen that the evaluation "non-affect" represents the products' characteristics. Furthermore, results show that internalized words are correctly selected because these words represent the appraiser's own behaviors or emotions, such as "do," "buy," "want," or "pleased with."

The results of the estimation of impression axes and the top 10 product images of five visual impression topics are shown in Table 1. The table indicates the top four words and the score of topics for visual impressions. After eliminating the evaluation word, the number of products with more than 10 reviews including the evaluation word was 3,644 out of 4,557. As a result of HDP-LDA, 15 topics were selected, and the perplexity was 243.5. Because topics are composed of words of similar meaning, it is easy to interpret each topic. The results indicate that the estimation of impression axes is promising.

Table 1. Impression Axes and Top 3 Images of Visual Impression Topics.

Topic	English	in Japanese	Top Images
1	high-class, heavy, massive, texture	高級だ, 重い, 重厚だ, 質感	
4	light, clear, thin, easy	軽い,見易い, 薄い,簡単だ	
6	child, tough, durable, quick	子供,タフだ, 丈夫だ,迅速な	
7	sophisticated, gorgeous, tiny, delicate	上品だ,素敵だ, 小さい,華奢	
11	interesting, fashionable, unique, rare	面白い,おしゃれだ, 変わる,珍しい	

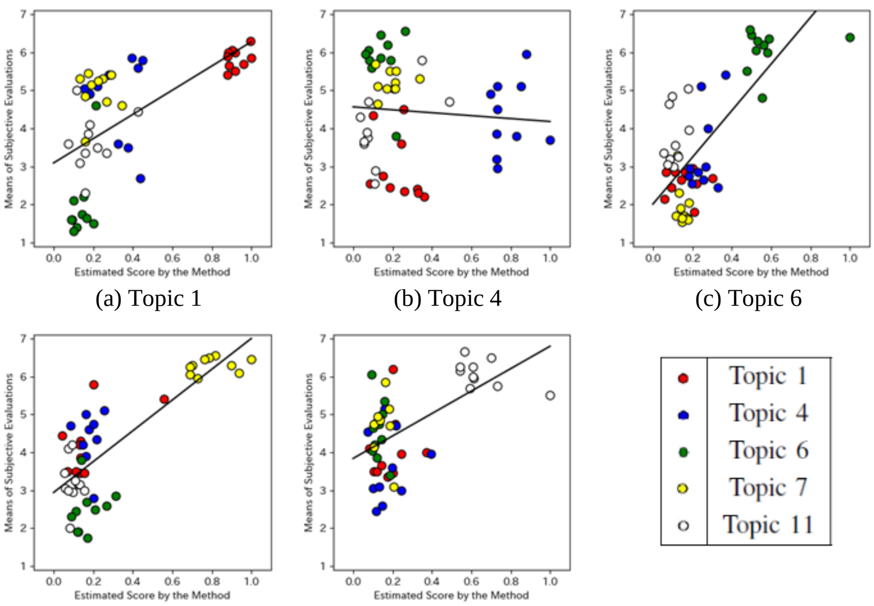
3.3 Subjective Evaluation Experiment

To confirm the consistency between the estimated score of the wristwatches and people's evaluations, we conducted a subjective evaluation experiment of the top 10 product images of the above five topics for visual impressions using the evaluation words.

Method. Twenty participants, 9 women and 11 men, whose ages were between 20 and 24, participated in this experiment. In this experiment, 50 wristwatch images from the top 10 of the five visual impression topics were presented. These images were collected from the website based on product names and product numbers. The image size was unified to 400 × 400 pixels, and the height of the wristwatch body was resized to 320 pixels. Moreover, the background colors were unified to the same gray (RGB 128, 128, 128). The participants evaluated the images using the seven-point

Likert-type scale of the SD method. The seven-point scale ranged from (1) completely disagree (*hizyoni-atehamaranai*), to (7) completely agree (*hizyoni-atehamaru*). The number of evaluation items was five, and each item consisted of the top four words for a topic relative to visual impression. In addition, some words were transformed into a form suitable for the experiment. All images were presented in random order.

Results and Discussion. First, in order to verify the validity of the proposed method, we confirmed the correlation between the estimated score given to each wristwatch and the experimental score. The experimental score of each wristwatch was determined by the mean value of participants. The scores for topic 6 and topic 7 were strongly related to the experimental scores, r = .743 and .787, respectively. In addition, the scores for topic 1 and topic 11 were moderately related to the experimental scores, r = .622 and .576, respectively. However, there was no correlation between the score for topic 4 and the experimental scores, r = -0.081. We plotted each wristwatch in Fig. 2. The abscissa represents the means of subjective evaluation score, and the ordinate represents the score estimated using this method. The estimated scores were normalized to a range between 0 and 1. According to Fig. 2, most watches other than watches with high estimated scores gather in the low range of 0.0 to 0.3 on the axis of the estimated score. Conversely, the watches with high estimated scores have a high evaluation in the corresponding evaluation item, except for topic 4. It can be considered that because the proposed estimated score of the wristwatch with the middle experimental scores is close to 0, the correlation was not as strong.



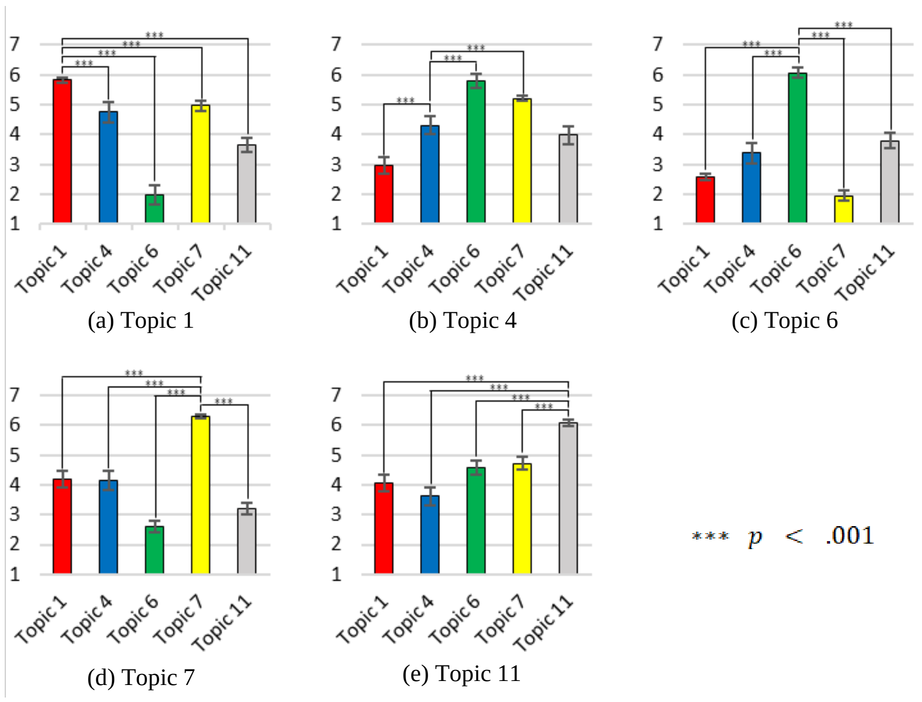
(a) Topic 1 (b) Topic 4 (c) Topic 6

(d) Topic 7 (e) Topic 11

Fig. 2. Scatterplot of Each Wristwatch.

Next, we verified whether the mean value of the watches with the high estimated score of the topic corresponding to each evaluation item was statistically higher than the mean value of the wristwatches of other topics. In order to verify the interaction between evaluation items and topics, we used a four-way analysis of variance (ANOVA). The factors were participants, topics, evaluation items, and wristwatch. If an interaction between the topic and the evaluation item was confirmed, we checked that the simple main effect of the topic was statistically significant in each evaluation item. If it was statistically significant, we performed a TukeyKramer multiple-comparison test and confirmed topics with high scores in the evaluation items.

The results of four-way ANOVA were as follows: Topics*Evaluation items: $F(16,4911)=202.870$, $p<0.001$; Participants: $F(19,4911)=7.191$, $p<0.001$; Evaluation items: $F(4,4911)=80.114$, $p<0.001$; Wristwatch: $F(45,4911)=4.146$, $p<0.001$. The interaction between the topics and the evaluation items was statistically significant at the 0.1% level. Next, the results of the simple main effects of the topics in each evaluation item were as follows: Topic 1: $F(4,995)=191.549$, $p<0.001$; Topic 4: $F(4,995)=51.207$, $p<0.001$; Topic 6: $F(4,995)=35.801$, $p<0.001$; Topic 7: $F(4,995)=185.109$, $p<0.001$; Topic 11: $F(4,995)=81.516$, $p<0.001$. Lastly, Fig. 3 shows a graph of the mean value for each topic of each evaluation item based on the result of the multiple-comparison. The mean value of the corresponding evaluation items of all topics except topic 4 was statistically significantly higher at the 0.1% level compared to the mean value of the other topics.



(a) Topic 1 (b) Topic 4 (c) Topic 6

(d) Topic 7 (e) Topic 11

Fig. 3. The Mean Values of Each Topic in Each Evaluation Item.

Based on these results, the proposed method for products with high estimated scores was proven to be beneficial. In addition, after examining the findings, it seems that topic 4 does not involve many visual impressions. For instance, the experimental score of the top watch for topic 4 was not as high in the evaluation items of topic 4; however, it is rated “light” in most product reviews. The material of the dial of this watch is titanium, which is lighter than stainless steel; therefore, it was assumed that it would be evaluated as “light,” contrary to its appearance. As mentioned above, the material is difficult to judge by appearance; hence, it is difficult to confirm the validity of the result of topic 4 using visuals alone. For further study, it is necessary to give appropriate scores to products that were given the middle impression score in the subjective evaluation experiment. Many of the products that were evaluated as a middle impression in topic 1 were products of topic 7, and vice versa; therefore, to solve this problem, a method that uses similarity between topics or evaluation words should be considered.

4 Conclusion

In this paper, we proposed a method to automatically construct a Kansei evaluation model for product design from review texts using text mining techniques. First, evaluation words are collected and selected using Japanese dictionaries of evaluation expressions and POS data. Second, the impression axes are estimated using a topic model that uses only evaluation words. Lastly, the scores on impression axes are estimated using frequencies of appearance of evaluation words and term-scores. We applied the method to the product reviews of Internet shopping sites and constructed the kansei evaluation model of the wristwatch. Moreover, we conducted a subjective evaluation experiment on the estimation results on the impression axes. Based on these results, the study found that the estimated scores are related to scores evaluated by subjective experiment and that the method is useful for products that are highly appreciated on evaluation axes existing in relation to people. This indicates that the proposed method is very promising in performing its task. We believe that the demand for the automatic construction of a kansei evaluation model will increase, because affective needs are increasingly regarded as important in many fields.

Acknowledgment

This research is partially supported by the Center of Innovation Program from Japan Science and Technology Agency, JST.

References

1. Qu, Q.-X.: Kansei knowledge extraction based on evolutionary genetic algorithm: an application to e-commerce web appearance design. Theoretical Issues in Ergonomics Science 16(3), 299–313 (2015).

2. Chen, C. H., Khoo, L. P., Chen, K., Pang, J. H., & Huang, Y.: Consumer-oriented product form creation via Kansei engineering. Proceedings of the international symposium for emotion and sensibility e emotion research in practice, 184–191 (2008).
3. Jing, Y., Ting, J., & Jianbo, X.: A Study on Interaction Design of Elderly Mobile Terminals Based on Kansei Engineering. In: Proceedings of the 2016 International Conference on Robots & Intelligent System (ICRIS), pp. 59–62. IEEE, Zhangjiajie (2016).
4. Aiba, E., Takamatsu, N., Numata, K., Yanagita, S., Suzuki, S., Sato, T., ... Takada, K.: Differences in ‘Kansei’Space between Age Groups. Transactions of Japan Society of Kansei Engineering 15(7), 677–685 (2016).
5. Lokman, A. M., Haron, M. B. C., Abidin, S. Z. Z., & Khalid, N. E. A.: Automating ex-pert role to determine design concept in Kansei Engineering. In: Aip conference proceedings (Vol. 1705), p. 20003. AIP Publishing, Bali (2016).
6. Htay, S. S., & Lynn, K. T.: Extracting product features and opinion words using pat-tern knowledge in customer reviews. The Scientific World Journal 2013 (2013).
7. Santosh, D. T., Babu, K. S., Prasad, S. D. V., & Vivekananda, A.: Opinion Mining of Online Product Reviews from Traditional LDA Topic Clusters using Feature Ontology Tree and Sentiwordnet. International Journal of Education and Management Engineering (IJEME) 6(6), 34–44 (2016).
8. Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M.: Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. In: Advances in Neural Information Processing Systems 17, pp. 1385-1392, MIT Press (2005).
9. Blei, D. M., & Lafferty, J. D.: Topic models. Text mining: classification, clustering, and applications 10(71), 34 (2009).
10. Kudo, T., Yamamoto, K., & Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis. In: Lin, D., Wu, D. (eds.) Proceedings of EMNLP 2004, pp. 230–237. Association for Computational Linguistics, Barcelona (2004).
11. Kurohashi, S.: Improvements of Japanese morphological analyzer JUMAN. In: Proceedings of the workshop on sharable natural language resources, pp. 22–28. Nara (1994).
12. Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., & Fukushima, T.: Collecting Evaluative Expressions for Opinion Extraction. Journal of Natural Language Processing 12(3), 203–222 (2005).
13. Kobayashi, N., Inui, K., & Matsumoto, Y.: Designing the task of opinion extraction and structurization. IPSJ SIG Technical Report, NL171-18, 111-118 (2006).
14. Higashiyama, M., Inui, K., & Matsumoto, Y.: Learning sentiment of nouns from selectional preferences of verbs and adjectives. In: Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing, pp. 584–587. The Association for Natural Language Processing, Tokyo (2008). (In Japanese)
15. Sano, M.: The construction of “Japanese dictionary of appraisal -attitude-” ～ Development of language resource to capture diversity of evaluation. In: Proceedings of the 17th Annual Meeting of the Association for Natural Language Processing, pp. 115‐118. The Association for Natural Language Processing, Toyohashi (2010). (In Japanese)
16. Sano, M.: Japanese Dictionary of Appraisal —attitude— (JAppraisal Dictionary ver1.0). Gengo Shigen Kyokai (2010). (In Japanese)