Computing Evaluation Scores with An Arbitrary Aspect from Evaluation Texts in Review Sites

Satoru Hosokawa[†], Etsuko Inoue[‡], Takuya Yoshihiro[‡*], Masaru Nakagawa[‡]
[†]Graduate School of Systems Engineering, Wakayama University, Japan
[‡]Faculty of System Engineering, Wakayama University, Japan
[*]tac@sys.wakayama-u.ac.jp

Abstract - Recently, evaluation sites have become to be popular in which we can share evaluation comments over various objects including restaurants, shops, and commercial products. In such sites, users can write evaluation comments as evaluators, and also refer to the comments of others to grasp the evaluation of the object that the users are interested in. To grasp the evaluation of an object in such sites is, however, very laborious because users have to look over too many evaluation comments. In this paper, to reduce the labor to grasp the evaluation, we propose a method to compute numerical scores of objects from a set of evaluation comments with an arbitrary given aspect, which can be determined according to users' own preferences. With this method, users can refer to numerical scores of various objects with their own free aspects in order to reduce the objects to compare, so they can reduce the labor in grasping evaluation by reading evaluation comments for only high-score objects.

Keywords: Evaluation Analysis, Aspects, Evaluation Scores

1 INTRODUCTION

The Internet has grown rapidly in the several decades, and which enabled people to state their opinions or comments in public. As an example of the public statements, several review sites appeared, in which people write evaluation values or review comments for various evaluation entities such as restaurants, shops, and products for sale. This kind of web sites plays an important role to share so called word-of-mouth information among users of the Internet; Some part of people write their evaluation values and review comments into the site, and others refer them. These sites actually are useful for people to select shops or products to buy, or for companies in their marketing activities.

In review sites, however, people generally have to read so many evaluation comments as to grasp the real evaluation for each entity because the reviews are often quite different depending on individuals and further sometimes include unreliable or irresponsible comments. The problem here is that it requires considerable labor to refer and examine these review comments. Here, some people would be bored to read all review comments and quit it in the halfway, but note that we actually need to read considerable amount of review comments if we wish to grasp the real evaluation from text-based evaluation comments. One direct approach for this problem is to summarize the review comments so as to reduce the amount of comments to read. However, such a simple summarization rarely works well because in many cases the number of evaluated entities and the review comments are too large, and also currently the quality of summarizing techniques are not sufficiently high.

As another approach to reduce the labor of users, it is possible to sort the evaluated entity in the order of evaluation scores, and users only see the review comments of high-rank (e.g., top-10) entities for their selection of entities to buy or use. A history of studies to compute evaluation scores for entities is seen in the literature. As seminal work, Turney [1] proposed a method to classify review articles into two levels of polarity, *positive* and *negative*. Koppel et al. [2] extend the method to classify them into three polarity levels, *positive*, *neutral*, and *negative*. Later, they lead to methods compute finer grained numerical scores, say, rating of entities [3][4]. However, their methods are not based on particular "aspects" of evaluation, so they cannot follow variation and difference of users' viewpoints or tastes. The viewpoints or tastes in evaluating entities are usually different depending on individuals, so these methods would not meet the requirements that users would like to know the evaluation results from various practical aspects.

On the other side, there are several studies on analysis of review comments considering various aspects in evaluation. References [5] and [6] consider typical evaluation aspects, and summarize review comments with each evaluation aspect through retrieving sentences related to each evaluation aspect. Here, the typical evaluation aspects for hotels, for example, would be "cleanness," "location," "service of staffs," etc. These studies assume that such an evaluation aspect is given as a few words in advance. Their methods actually consider several aspects in evaluation, but they only treat typical evaluation aspects given by simple words. Therefore, they do not sufficiently cover the requirements of users because users' requirements have large diversity of aspects reflecting on wide variety of users' viewpoints and tastes in evaluating entities.

As for the diversity of evaluation aspects, several studies [8]–[10] try to retrieve words as "topics" that represent evaluation aspects. If we retrieve topics using these methods and summarize review comments with each topic, we may cover larger diversity of users' requirements. Also, we can compute evaluation scores instead of summarizing texts. Then, users will achieve efficient use of review sites by reducing their labor via referring evaluation comments of only high-rank entities. However, these methods do not cover all possible aspects of users, and the range of "topics" is still limited within a word.

In this paper, we propose a method to compute evaluation scores of entities to reduce labor of users to grasp evaluations

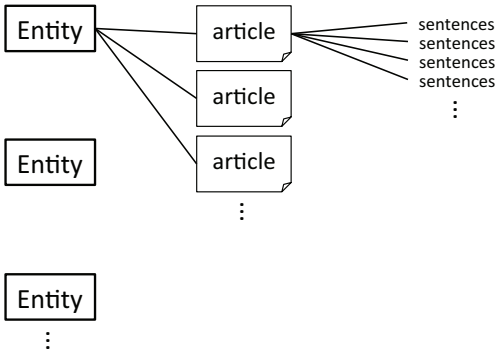

Figure 1: The Structure of Review-site Data

in review sites, while covering all possible evaluation aspect of users. In our study, we assume that an aspect for evaluation is given as a form of text description, and we compute the evaluation score with the given aspect. For example, in case of restaurants, "Good restaurant for family with reasonable cost" can be a typical practical aspect description that is useful for many people. With our method, users can obtain an ordered list of evaluation entities with respect to the computed evaluation scores, and so they can focus on high-rank entities based on their own evaluation aspect, which enable them efficient use of review sites. To the best of our knowledge, this is the first method to compute scores based on a text-style aspect description.

This paper is organized as follows: In Section 2, we describe the proposed method that computes evaluation scores with respect to the given aspect description. In Section 3, we give an evaluation results for the proposed method, and show that the method computes the evaluation scores that fit the feeling of the users of review sites. Finally, we conclude the work in Section 4.

2 COMPUTING EVALUATION SCORES WITH AN ARBITRARY ASPECT DESCRIPTION

2.1 Overview of the Proposed Method

In this paper, we compute a numerical score for each entity based on the given text description of an evaluation aspect. The structure of the review-site data is shown in Fig. 1; For each entity to be evaluated, we have text evaluation articles that consist of many sentences, which forms a tree structure. The proposed method, which computes a score for each entity from this form of data, consists of three folds:

(a) Learning a dictionary of evaluation words.

(b) Computing the evaluation score for each sentences included in each review article.

(c) Computing the evaluation score for each entity using the scores of the sentences computed in step (2).

Figure 2 illustrates the overview of the proposed method. First, (a) we learn a dictionary of evaluation words from a

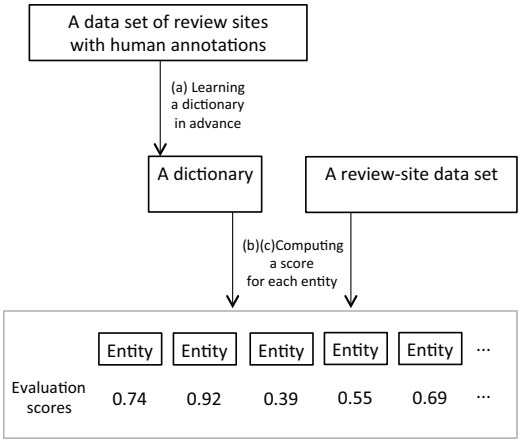

Figure 2: The Process of The Proposed Method

small data set of review sites with human annotations, and then (b)(c) compute the evaluation score for each entity. Here, the data set to learn the dictionary could be different from the full data-set of the review-sites, could be rather small data set, but the sort of entity to be evaluated should be the same as the review-data with which we compute the evaluation score. (Namely, if you want to evaluate restaurants, then the data set to learn the dictionary should include the evaluation articles for restaurants.) Further, note that the data set to learn the dictionary should include human annotations; for each sentence in the evaluation articles, a reliable person should perform the following.

(1) We judge whether the sentence is surely related to the evaluation with the given evaluation aspect or not. If yes, the sentence is called an *evaluation sentence* under an aspect a.

(2) For each *evaluation sentence* under a, we further add the polarity, i.e., *positive*, *neutral*, or *negative*, to each sentence.

From this manipulated data set with human annotations, the dictionary is constructed. The method to construct the dictionary is described in Section 2.2.

Facing on (b)(c) computing evaluation scores, our basic strategy is to first compute the score for each sentence (not for each article) based on the dictionary, and collect them to compute the score for each entity. We adopt this strategy because we expect the averaging effect; It is important to collect a sufficient number of units for evaluation to perform statistical computation, so we choose a "sentence" as a unit of evaluation because a relatively large number of sentences are included in each evaluation article, while in many cases each sentence includes sufficient number of words to judge their polarity roughly. Specifically, in our method, we first compute a polarity value, i.e., *positive*, *neutral*, or *negative* for each *evaluation sentence* using the dictionary, and merge them to compute finer-grained rating score for each entity according to the ratio of *positive* and *negative* sentences.

Words	Fitting Level	Polarity
Flavor	50.3	−0.02
Taste	38.1	0.7
Like	34.8	0.33
Meat	30.5	0.22
Normal	23.2	−0.01

Figure 3: An Example of Evaluation Dictionary

2.2 (a) Learning A Dictionary of Evaluation Words

2.2.1 Structure of The Dictionary

The evaluation dictionary is a dictionary that is used to compute the polarity of each sentence in review articles, and is a set of tuples $(w, F_{w,a}, P_{w,a})$, where w is a word for evaluation, $F_{w,a}$ is the *fitness level* of the word w with aspect a, and $P_{w,a}$ is the *polarity level* of w with aspect a. The *fitness level* $F_{w,a}$ represents the degree how much w is important in evaluating sentences w.r.t. an aspect a, and takes a higher value if the importance is higher. The *polarity level* $P_{w,a}$ represents the degree of *positive* or *negative* feeling of the word in evaluation w.r.t. a, which takes a value in $[1, -1]$ such that $P_{w,a}$ is nearer to 1 when w gives more positive evaluation, and nearer to -1 in case of more negative evaluation.

2.2.2 Retrieving Words for Evaluation

To construct the dictionary, we first retrieve the evaluation words, which are the words that we use in evaluation, from the data set for learning. We construct the dictionary with the words retrieved as evaluation words from the data set, while other words in the data set are just ignored. To retrieve evaluation words, we apply the morphological analysis to the data set and as evaluation words we choose nouns, verbs, adjectives, adverbs, interjections, and symbols.

In our method, to judge polarity correctly, a small preprocessing of words is required. Specifically, the negate words such as “not” and “never” in English would reverse the polarity of evaluation words. Thus, if we find these negate words with a verb or a adjective, we treat the verb or the adjective as a new word that includes negative meaning. Namely, one verb or adjective word may be included in the dictionary as two different words, i.e., with and without negative meaning.

2.2.3 Computing Fitness Levels of Words

As described above, the *fitness level* $F_{w,a}$ is the real value that represents how important a word w is in evaluation w.r.t. an aspect a. We designed a formula to compute the *fitness level* based on the well-known *tf-idf* index. The *tf-idf* is an index value that takes high value for words peculiar to a given document; For a given document included in a document set, *td-idf* is the product of *tf* and *idf*, where *tf* is the *term frequency* that represents the frequency of the term (word) in the document, and *idf* is the *inverse document frequency* that represents how common the term appears in all documents in the document set. Namely, the *tf-idf* index takes higher value for the words peculiar to the document, while it takes lower value for the words that commonly appears in all documents.

In the proposed method, as the value corresponding to *tf*, we use the frequency of a word w in the *evaluation sentences* under an aspect a i.e., the number of the sentences that includes w among *evaluation sentences* under aspect a in the learning data set. On the other side, as the value corresponding to *idf*, we use the ratio of sentences including w among all the sentences in the learning data set. Thus, *idf* takes larger value when the number of sentences including w is small.

Now we give a formal description of the *fitness level*. Let S be the set of all sentences in the learning data set, S_a be the *evaluation sentence*, i.e., the set of sentences judged to be related to the aspect a, and $S_{\bar{a}}$ be those judged not to be related to a. Naturally, $S_a \cap S_{\bar{a}} = \emptyset$ and $S = S_a \cup S_{\bar{a}}$ hold. Also, let $n_{w,a}$ and $n_{w,\bar{a}}$ be the frequency of w in S_a and $S_{\bar{a}}$, respectively. Let $|\{s|w \in s \text{ and } s \in S\}|$ be the number of sentences that include w in S. Then, the definition of $F_{w,a}$ for a given word w and an aspect a is written as follows:

$$F_{w,a} = \mathrm{tf}_{w,a} \cdot \mathrm{idf}_w,$$

where

$$\mathrm{tf}_{w,a} = \frac{n_{w,a}}{n_{w,a} + n_{w,\bar{a}}},$$

$$\mathrm{idf}_w = \log \frac{|S|}{|\{s|w \in s \text{ and } s \in S\}|}.$$

2.2.4 Computing Polarity Levels of Words

Polarity level $P_{w,a}$ is the real value in range $[1, -1]$ that represents the degree of positive or negative that a word w is used to evaluate entities under an aspect a. $P_{w,a}$ takes a value near 1 when w contributes to positive evaluation, and near -1 when negative.

The polarity level of a word w with an aspect a is computed based on the ratio of positive and negative sentences among all *evaluation sentences* that includes w. We designed the formula to compute $P_{w,a}$ where the polarity takes 1 when w appears in only positive sentences, and takes -1 when w appears in only negative ones.

The formal definition of $P_{w,a}$ is given in the following. Let S_a^p and S_a^n be the sets of *evaluation sentences* in S_a that are annotated as *positive* and *negative*, respectively. Also, let $f_{w,a}^p$ and $f_{w,a}^n$ be the frequency of w appearing in the sentences in S_a^p and S_a^n, respectively. Then, the polarity level $P_{w,a}$ for a word w and an aspect a is defined as follows:

$$P_{w,a} = \frac{f_{w,a}^p}{f_{w,a}^p + f_{w,a}^n} - \frac{f_{w,a}^n}{f_{w,a}^p + f_{w,a}^n}$$

2.3 (b) Computing Polarities for Sentences

2.3.1 Retrieving Evaluation Sentences under Aspect a

For each sentence in the evaluation articles in the review site, we first judge whether the sentence should be used to compute the score of the entity, i.e., whether each sentence in review articles is *evaluation sentence* or not. The *evaluation sentence*

should surely evaluate the entity under the aspect a. Thus, in this process, we judge this point using the fitting levels of the words included in the sentence.

The basic strategy is as follows. From a sentence s to be judged, we first retrieve words whose fitting level is sufficiently high, which are the words that have ability to evaluate entities. We next compute the average of the fitting levels, and if the average is sufficiently high, the sentence s has ability to evaluate entities, so judged to be *evaluation sentence*.

We present the formal description of this process. Let s be the sentence to be judged. Let F_{min} be the threshold of fitting level for *evaluation words*. With threshold F_{min}, we define the set of words that has sufficiently high fitting levels as $W_f^s = \{w | w \in s \text{ and } F_{w,a} \geq F_{min}\}$. Thus, the average of the fitting levels of *evaluation words* in s is written as

$$F_s = \frac{1}{|W_f^s|} \Sigma_{w \in W_f^s} F_{w,a}.$$

If F_s is equal to or larger than threshold T_s, i.e., $F_s \geq T_s$, then the sentence s is judged to be *evaluation sentence*, which is used in evaluating entities.

Figure 4 illustrates an example of the process to choose the *evaluation sentences* for entities. In this figure, we judge whether the sentence is an *evaluation sentence* or not under an aspect a. Here, the fitting levels of all (four) words used for evaluation are larger than threshold F_{min}, we compute the average of the fitting levels among them. Because the average value is larger than threshold $T_s = 10$, this sentence is selected as an *evaluation sentence*.

2.3.2 Computing Polarities

For the each *evaluation sentences* s, we further compute the polarity of the sentence s. Since a single sentence usually includes not many words, we choose to use the 3-graded polarity value, i.e., *positive*, *neutral*, and *negative*, rather than finer-grained polarity such as real values.

The basic strategy to compute the polarity of sentence s is to examine the total polarity of the evaluation words included in s. We first retrieve the words that has sufficiently strong polarity, and examine the total balance of the polarity of them.

Specifically, let P_{min} be the threshold to select words of strong polarity. With P_{min}, we define the set of words that has strong polarity as $W_p^s = \{w | w \in s \text{ and } |P_{w,a}| \geq P_{min}\}$. Using this set of words, we define the polarity of sentence s as follows:

$$P_s = \begin{cases} \textit{positive}, & (\text{if } \ T_p < \frac{1}{|W_p^s|} \Sigma_{w \in W_p^s} P_{w,a}), \\ \textit{neutral}, & (\text{if } \ -T_p \leq \frac{1}{|W_p^s|} \Sigma_{w \in W_p^s} P_{w,a} \leq T_p), \\ \textit{negative}, & (\text{if } \ \frac{1}{|W_p^s|} \Sigma_{w \in W_p^s} P_{w,a} < T_p). \end{cases}$$

Figure 5 illustrates an example of polarity computation of a sentence. Since the sentence is determined as *evaluation sentence* in Fig. 4, we next compute the polarity of this sentence. We first retrieve the words whose polarity values are equal to or larger than P_{min} in absolute value, and compute the average of the polarity of the selected words. Because the average is larger than threshold $T_p = 0.35$, the polarity of this sentence is determined to be *positive*.

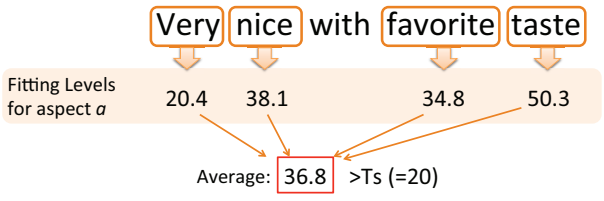

Figure 4: Judging Aspect for An Evaluation Sentence

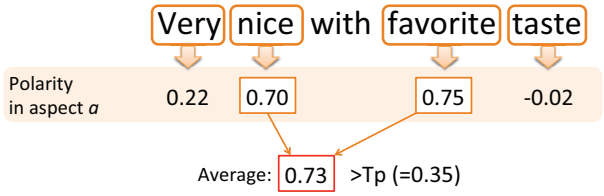

Figure 5: Judging Polarity for An Evaluation Sentence

2.4 (c) Computing Evaluation Scores for Entities

Finally, we compute the evaluation score of an entity i using the *evaluation sentences* selected under aspect a. The evaluation score of i, which is referred as $Score(i)$, is computed based on the ratio of *positive* and *negative* evaluation sentences in the review articles of i. Formally, the evaluation score is computed as

$$Score(i) = \frac{|\{s | P_s = \textit{positive} \text{ and } s \in E_i\}|}{|\{s | P_s = \{\textit{positive} \text{ or } \textit{negative}\} \text{ and } s \in E_i\}|}$$

where E_i represents the set of *evaluation sentences* that evaluate i selected with the process shown in Sec. 2.3.1, and s represents a sentence.

3 EVALUATION

3.1 The Viewpoints

In this paper, we propose a method to compute the evaluation scores for each entity with respect to an arbitrary text description of evaluation aspects. In other words, this method intends to predict the human evaluation scores for each entity that readers of the evaluation articles would make. Therefore, in our evaluation, we requested several persons to read evaluation articles and make a 10-grade score for each entity. We evaluated the difference between the human scores and the computed scores to measure the precision of the proposed method.

Note that, however, human scores in general vary depending on individuals, especially in the average or the standard deviation of the scores. (Imagine that some person may make relatively low scores in average, while other person may prefer high rating.) To take this diversity into account, we standardized the human scores for each person (namely, the average and the standard deviation of the scores made by a person are adjusted to be the same), and made a ranking of entities with their average scores. If the two rankings based on human scores and computed scores are similar to each other, it implies that the performance of the proposed method to predict

Figure 6: Human and Computed Scores for Restaurants under Aspect "Taste" (Experiment 1)

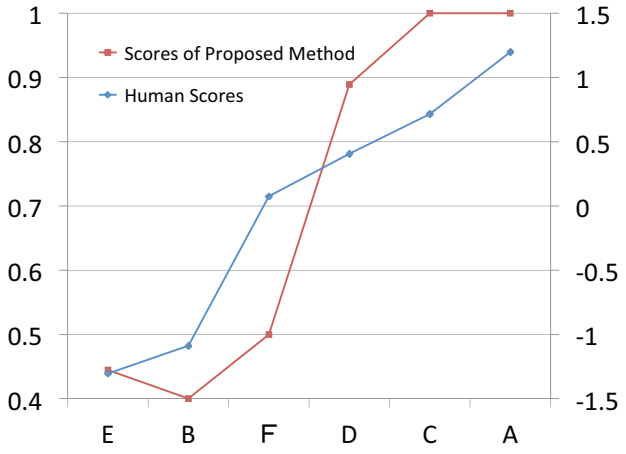

Figure 7: Human and Computed Scores for Restaurants under Aspect "Price" (Experiment 1)

Figure 8: Human and Computed Scores for Ra-men Restaurants under Aspect "Quality of Noodles" (Experiment 2)

Figure 9: Human and Computed Scores for Ra-men Restaurants under Aspect "Taste of Soup" (Experiment 2)

human scores is good. Thus, we used Spearman's rank correlation coefficient between human and computed rankings as evaluation criterion to measure the precision of the proposed method.

We conducted two evaluations using different aspects. We supposed the following two different cases in determining the aspects.

Experiment 1: In case of general evaluation aspects.

Experiment 2: In case of specific evaluation aspects that reflects on personal viewpoints of individuals.

In Experiment 1, we used general evaluation aspects that we often see in review sites. We selected "restaurants" as evaluation entities, and used two evaluation aspects "taste" and "price." In Experiment 2, we used a little specific evaluation aspects that requires several words to describe. We selected "Ra-men restaurants" as evaluation entities, and used two evaluation aspects "quality of noodles" and "taste of soup." Note that these two aspects would be still so simple and would not be as complicated as usual practical descriptions. However, in this paper, we use these two aspects to perform a first-step investigation to clarify the basic property of the proposed method.

3.2 Evaluation Methods

For Experiment 1, we selected 6 restaurants as the reviewed entities from a popular Japanese review site called "Tabelog" [11]. To guarantee fair evaluation, these 6 restaurants are chosen from high-rated restaurants in Tabelog, placed in Tokyo, where users' ratings are the same as a whole. We selected 3 review articles for each restaurants mainly under the criteria that (1) the length is almost the same as 50-60 sentences, (2) review date is not old, (3) they do not include any direct description of numerical scores, and (4) sentences are relatively tidy. As written above, the evaluation aspects are "taste" and "price," and we told the participants of our experiments (i.e., subjects in our experiments) that "taste" means how good the taste of dishes is, and "price" means how reasonable the price of dishes is.

For Experiment 2, as the reviewed entities, we selected 10 Japanese Ra-men restaurants placed in Wakayama city also from Tabelog. Note that they are all high-rated Ra-men restau-

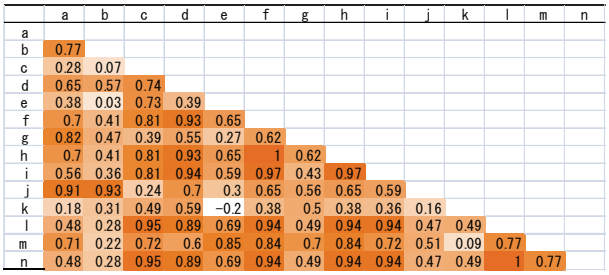

Figure 10: Rank Correlation Coefficients between Users. (Aspect "Taste" in Experiment 1)

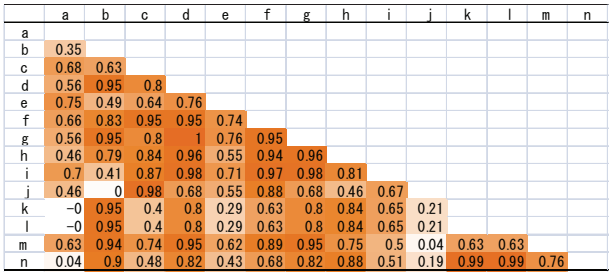

Figure 11: Rank Correlation Coefficients between Users. (Aspect "Price" in Experiment 1)

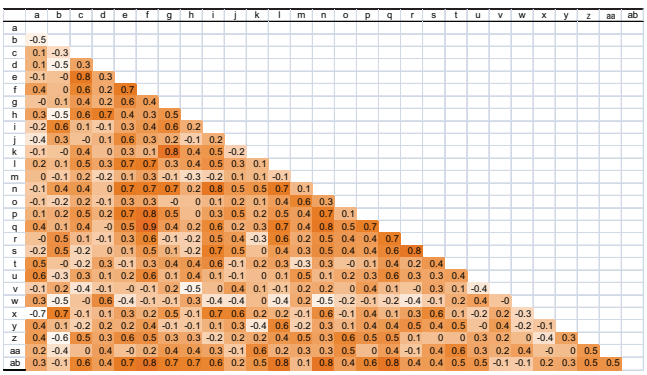

Figure 12: Rank Correlation Coefficients between Users. (Aspect "Quality of Noodles" in Experiment 2)

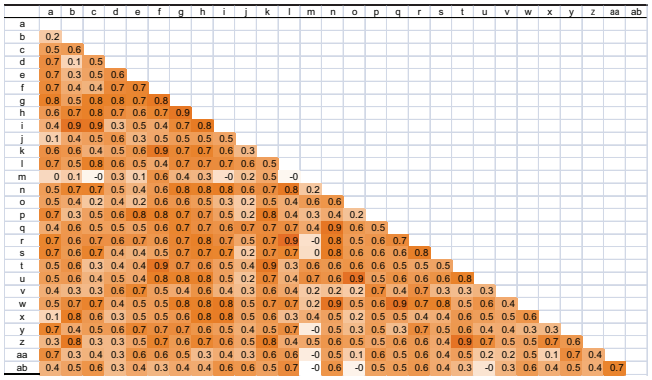

Figure 13: Rank Correlation Coefficients between Users. (Aspect "Taste of Soup" in Experiment 2)

rants in Wakayama city. We used 7 review articles for each restaurant, where each article consists of about 30 sentences. The criteria to select those review articles are the same as the case of Experiment 1. The evaluation aspects are "quality of noodles" and "taste of soup."

In advance of the experiments, we constructed the dictionary under the given four aspects. As a set of learning data, we collected review articles from Tabelog. For Experiment 1, we collected 1,500 review articles for restaurants that include about 6,000 sentences. As the result of our annotation, we obtained 4,600 evaluation words for the aspect "taste," and 1,500 words for "price." For Experiment 2, we collected 1,200 review articles for Ra-men restaurants that include about 3,000 sentences. As the result of annotation, we obtained 800 evaluation words for "quality of noodles," and 1,800 words for "taste of soup."

As the process of the experiments, we asked all the participants to read all the review articles and to make a 10-grade score for each entity, where 10 is the best, and 1 is the worst grade of scores. 14 and 28 persons (subjects) participated to the Experiment 1 and 2, respectively, where all the participants were at the age of 20's.

3.3 Evaluation Results

In Figs. 6-9, we show the average of human scores and the scores computed by the proposed method. The horizontal axis represents the restaurants in the order of scores. The left vertical axis represents the score of the proposed method, and the right one represents the average of human scores. The rank correlation coefficient in Experiment 1 is 0.94 for the aspect "taste" and 0.92 for "price," which show that the proposed method predicts the human score with high precision in case of general aspects. In Experiment 2, the rank correlation coefficient is 0.74 for "quality of noodles" and 0.72 for "taste of soup," which is not so high as Experiment 1, but relatively high value.

Figures 10-13 shows the rank correlation coefficients between participants for each aspects in our experiments. Each alphabet represents a participant (14 and 28 persons participated in our Experiments 1 and 2, respectively), and for every pairs of the participants, we compute the rank correlation coefficient of the two rankings. In Experiment 1, the correlation coefficients takes high values as a whole, where the average value is 0.60 for "taste" and 0.67 for "price," meaning that the ranking of participants are relatively similar to each other. On the other hand, in Experiment 2, they take low values where the average value is 0.22 for "quality of noodles" and 0.52 for "taste of soup," meaning that the ranking differs significantly according to individuals.

3.4 Discussion

We obtained the result that the rank correlation coefficients take relatively high values in case of general aspects, while they take low values in case of specific aspects. In this section, we discuss the reason of this point.

First of all, we focus on the rank correlation coefficient

between participants shown in Figs. 10-13, where they take quite low values in case of specific aspects. Especially, with the aspect "quality of noodles," it takes very low value 0.22. It means that the rankings of participants are basically similar to each other for the general aspects in Experiment 1, whereas they are quite different for the specific aspects in Experiment 2. The reason of this is quite simple; It is due to likes and dislikes among people. In fact, from the hearing from participants after the Experiment 2, it is clarified that several participants were strongly affected by the expression words such as "hard" or "soft" on noodles, or "rich" or "plain" on soup. It would be natural that someone likes "hard" noodle or "rich" soup, while others would like "soft" or "plain" ones. On this point, we also examined the polarities of those words in the dictionary and found that the polarities of them are mostly neutral (i.e., near 0). It is considered that the person who annotated to the learning data set seemed to select *neutral* if a sentence includes the words that depends on like and dislikes of people. As a result, the proposed method also gave neutral polarities for this kind of words.

As another reason on this point, the precision of the dictionary possibly affects the performance in Experiment 2. Note that the number of words in Experiment 1 is 4,600 words for "taste" and 1,500 words for "price," while that in Experiment 2 is 800 words for "quality of noodle" and 1,800 words for "taste of soup." The number of words is smaller in Experiment 2, which may affects the performance. (Note that the performance for "price" is good although the number of words is relatively small. This may be because most of the words that evaluates "price" is clear to understand; the polarity of words "expensive" or "cheap" would be clear for most of people.)

Therefore, to examine the effects of the number of words in the dictionary, we conducted another experiment. Using the four dictionaries constructed for each evaluation aspects as used in Experiments 1 and 2, we evaluated the precision of evaluation-sentence judgments and polarity judgments for sentences described in Sections 2.3.1 and 2.3.2, respectively. For evaluation-sentence evaluation, we prepared 1,200 sentences that are related with the aspect and another 1,200 sentences not related with the aspect, and examined the precision of the proposed judging algorithm described in Section 2.3.1. For polarity evaluation, we prepared 1,200 sentences for each of *positive*, *neutral*, and *negative* polarities, and examined the precision of the proposed method described in Section 2.3.2. Results are shown in Figs. 14 and 15, respectively. Both results show that the proposed method marks about 90% of precisions regardless of aspects, which indicates that the precision of the dictionary is not related to the number of words in the dictionary. Thus, the cause that the rank correlation coefficient is relatively low in Experiment 2 would not be the number of words in the dictionary, but would be the effect of like and dislike of people for several specific evaluation words.

4 CONCLUSION

In this paper, we proposed a method that computes the evaluation scores for entities in review sites with respect to a given aspect of arbitrary text description. As a result of our evaluation, the proposed method computes evaluation scores that have high rank correlation coefficients with human scores. That is to say, the proposed method predicts the human scores with high precision. Also, from the evaluation result, we found there are aspects that include likes and dislikes of people, and that the correlation coefficients degrade for such aspects. The main reason of this degradation is the existence of the evaluation words for which people may give wide-range of polarities depending on persons.

One of the most important future task on the proposed method is to cope with the problem described in Section 3.4, i.e., the problem that likes and dislikes exist in several evaluation words. Although by nature this problem occurs inevitably, we have several choices to avoid the inconvenience that comes from the problem. The easiest solution is to detect the words that include likes and dislikes of people and exclude them from the computation of evaluation score. Another choice, which would be a more challenging solution, would be to classify people into two groups, say, 'like' group and 'dislike' group, and show the scores of both groups to the users who use our method. To propose and evaluate the methods to achieve either of them would be challenging work toward practical use of our method.

We finally note that someone would consider that the cost to create a dictionary for each aspect could be a problem for the practical use of the proposed method. Although the proposed method actually requires considerable human labor, we would note that the laborious task would be performed easily with low cost if we use a tool called crowd sourcing. With this useful tool, the proposed method would be one of realistic methods that works in practice.

Aspects	Precision	Recall	F-value
"taste"	0.904	0.947	0.925
"price"	0.988	0.898	0.941

(a) Experiment 1

Aspects	Precision	Recall	F-value
"quality of nodles"	0.873	0.870	0.871
"taste of soup"	0.905	0.859	0.881

(b) Experiment 2

Figure 14: Accuracy of Aspect Judgments

Polarity	Precision	Recall	F-value
Positive	0.835	0.820	0.827
Neutral	0.725	0.788	0.755
Negative	0.898	0.833	0.864

(a) "taste"

Polarity	Precision	Recall	F-value
Positive	0.877	0.844	0.860
Neutral	0.741	0.845	0.789
Negative	0.939	0.840	0.887

(b) "price"

Polarity	Precision	Recall	F-value
Positive	0.856	0.780	0.816
Neutral	0.701	0.829	0.760
Negative	0.868	0.760	0.810

(c) "quality of noodles"

Polarity	Precision	Recall	F-value
Positive	0.822	0.845	0.833
Neutral	0.716	0.809	0.760
Negative	0.961	0.791	0.868

(d) "taste of soup"

Figure 15: Accuracy of Polarity Judgments

REFERENCES

[1] P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), pp. 417-424 (2002).

[2] M. Koppel and J. Schler, "The Importance of Neutral Examples in Learning Sentiment." Computational Intelligence, Vol.22, No.2, pp.100-109 (2006).

[3] B. Pang, and L. Lee, "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respace to Rating Scales," In Proceeding of the 43rd Meeting of the Association for Computational Linguistics (ACL), pp. 115-124 (2005).

[4] D. Okanohara and J. Tsujii, "Assigning Polarity Scores to Reviews Using Machine Learning Techniques," Natural Language Processing, Vol.14, No.3 (2007).

[5] G. Carenini, R. Ng, and A. Pauls, "Multi-Document Summarization of Evaluative Text," In Prof. of the conference of the Eutopean chapter of the association for computational linguistics (2006).

[6] M. Hu and B. Liu, "Mining Opinion Features in Customer Reviews," In Proc. of Nineteenth National Conference on Artificial Intelligence (2004).

[7] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," In Proc. of the 29th annual international ACM conference on Research and development in information retrieval (SIGIR), pp.244-251 (2006).

[8] G. Carenini, R. Ng, and E. Zwart, "Extracting Knowledge from Evaluative Text," In Proc. of the 3rd international conference on knowledge capture, pp.11-18 (2005).

[9] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," In Proc. of the 2004 ACM international conference on knowledge discovery and data mining (SIGKDD), pp.168–177 (2004).

[10] I. Titov, R. McDonald, "Modeling Online Reviews with Multi-grain Topic Models," In Proc. of the 17th international conference on World Wide Web (WWW), pp.111-120 (2008).

[11] Tabelog, http://tabelog.com/ (In Japanese, referred to in 2015).

(Received Nobember 10, 2014)

Satoru Hosokawa received the B.E. and M.E. degrees from Wakayama University in 2011 and 2013, respectively. He is currently working with Yahoo Japan Corporation.

Etsuko Inoue received the B.E., M.E. and Ph.D. degrees from Wakayama University in 2002, 2004 and 2007, respectively. She is an Assistant Professor in Wakayama University from 2007. She is interested in database systems, web applications, data visualization, and so on. She is a member of IPSJ.

Takuya Yoshihiro received his B.E., M.I. and Ph.D. degrees from Kyoto University in 1998, 2000 and 2003, respectively. He was an assistant professor in Wakayama University from 2003 to 2009. He has been an associate professor in Wakayama University from 2009. He is currently interested in the graph theory, dis- tributed algorithms, computer networks, medial applications, and bioinformatics, and so on. He is a member of IEEE, IEICE, and IPSJ.

Masaru Nakagawa received the B.E, M.E, and Ph.D. degrees from Osaka University in 1970, 1972, and 1990, respectively. He was a researcher in NTT laboratory from 1972 to 1994, and from 1994 he was a Professor in Kinki University. He is a Professor in Wakayama University from 1994. He is interested in Database Design and Computer System Engineering. He is a member of IPSJ, JSAI and JSIK.