Short Note

Word and Comments Evaluation Using Recursive Evaluation in Lecture Questionnaire

Asami Shiwaku*, Nobuyuki Kobayashi** and Hiromitsu Shiina***

(Received 30 November 2016 and accepted in revised form 12 June 2017)

Abstract Recently, as part of faculty development, universities employ questionnaires about lectures, and improvements are planned using investigations of student satisfaction with teachers and lectures. In this study, we perform value prediction of comments and words in free comments from a questionnaire about lectures. The comment evaluation technique involves value prediction of the entire comments via mutually recursive repetition of multiple person partial comment evaluations and value prediction of those comments and words. In addition, we evaluate the differences between individually evaluated words and comment-predicted values.

Keywords: comment evaluation, word polarity, opinion distribution, lecture evaluation

1. Introduction

Recently, universities have been required to conduct faculty development, and student questionnaires have been administered as one way to achieve class improvement. Lecture evaluations and free-answer questionnaires have been implemented; however, the obtained text data cover various evaluation points, such as opinions regarding teachers and requests regarding classes; thus, the data may be problematic in terms of objective evaluation. Furthermore, aggregating opinions from questionnaires requires significant time and effort; thus, evaluations may be limited by the available resources. In other words, evaluating all comments is difficult.

In recent years, employment of sentiment classification using word evaluation and machine learning has increased significantly. Related studies have considered evaluating the polarity of an entire document based on the word appearance ratio of evaluation expressions. Other studies have extracted evaluation information token groups to evaluate polarity. In addition, document classification studies have achieved high classification accuracy using machine learning techniques, such as Naive Bayes and support vector machines.

In this study, we focus on an open-answer midterm questionnaire and perform analysis using positive and negative polarities for word evaluation. We evaluate the answers using a set of words from the questionnaire and then evaluate teachers and classes.

Furthermore, regarding the evaluation method for free answers, because point values are not assigned to free answers, value prediction must be performed. We predict the value of an entire free answer by evaluating part of it manually. Then, word polarity values can be estimated based on the predicted value. However, the value assigned to the free answer depends on the person performing the evaluation. Consequently, since word polarity differs significantly for each evaluator, the values from multiple persons’ free answers must be evaluated. Furthermore, when using values from multiple persons, word polarities become increasingly dispersed. Thus, words that have greatly dispersed polarities can be distinguished from those that have polarities that are less dispersed, which enables us to ascertain the nature of the words. It should be noted that, before teacher and class evaluations were performed, words with greatly dispersed polarity values were removed from the target values of interest.

2. Questionnaire Data

This study uses an open question item in the five items of a lecture questionnaire conducted at Okayama University of Science in Japan. The questionnaire was
conducted in the mid-course phase (8th lecture in a course with 15 lectures) of the 2014 spring term (April to September). The number of target teachers was 15. The number of lecture courses was 41. The number of responses was 1,978. The questionnaire form was common to all the courses.

3. Free-answer Comment Analysis

3.1 Outline of the Analysis Process

The evaluation of the free-answer comments was performed as follows (Figure 1).

1. Manual evaluation of comment rank.
   1–1 Teacher data (rank) creation.
   A six-stage evaluation was manually performed on a section of the comments, rating them as “Very bad (rank 1)”, “Bad (rank 2)”, “Quite bad (rank 3)”, “Quite good (rank 4)”, “Good (rank 5)” or “Very good (rank 6)”. Here, we present evaluation from two comment and word evaluation methods. From the 1,978 free-answer comments, we manually evaluated the top 100 comments with the greatest number of words. We called these manually evaluated comments as rated comments. Then, we evaluated the remaining 1,878 unrated comments. The distribution of comment rank for 3 of 12 evaluators (E1, E2, E3) is shown in Table 1. We note that the top 100 comments included 97% of all words used in all comments.

1–2 A predictive evaluation of word rank in a comment was performed using rated comments. The distribution of word rank was performed as follows.

1–2–1 Nouns, verbs, and adjectives were extracted from the rated comment, and then, a comment rank was assigned to the words in the comment.

1–2–2 Step (1–2–1) was performed for all rated comments and the words contained in the comments to determine the distribution of word rank based on the frequency of rank for each word.

1–2–3 From the distribution of word rank, the ranking of the most frequent rank was selected as the word rank. We note the lower rank was selected if the word ranking frequency of word rank was the same.

2. A predictive evaluation of unrated comment ranks was performed using the distribution of word rank. The distribution of word rank produced from the rated comments was used directly, and the comment rank obtained was reused for word rank.

3. A predictive evaluation of all comment ranks was performed using the rated comments, the word rank estimated from the rated comments and the unrated comments. The prediction method repeated steps (3–1) and (3–2) until the values of all comments could not be further improved.

3–1 The word rank was evaluated based on ranks of all comments.

3–2 The unrated comment rank was evaluated based on estimated word rank.

After completing the repeated predictive evaluation of the word rank for all words in comments, we selected the final prediction of rank for comments and words.

3.2 Recursive Evaluation of Comments and Words

In the previous section, we indicated that comments were classified as rated or unrated, and comment evaluation was repeated until the unrated comment ranks became fixed. In this process, word evaluation was based on the comment ranks, and comment evaluation was based on the word rank, i.e., evaluation was performed recursively. The proposed method sums frequency of rank for word evaluation based on comment rank and comment evaluation based on word rank (Figure 2). The predictive evaluation of word and comment ranks using the sum of the frequency of rank done
(1) A predictive evaluation of word rank was made.

(1–1) To perform word rank evaluation based on comment evaluations, we created a word rank distribution by converting the frequency distribution of the rank so that each frequency of the rank was divided by total frequency of the rank of six stages, i.e., we normalized each word by rank. To multiply this in (1–2), we added one to the distribution of word rank of each rank.

(1–2) From each distribution of word rank, we took the maximum rank as the word rank.

(2) A predictive evaluation of comment rank was made.

(2–1) To carry out comment evaluation based on word evaluation, we produced a comment rank frequency by totaling the word rank frequencies of the words within the comment.

(2–2) Among the words in a comment, we excluded words whose evaluation differed among evaluators and whose standard deviation was greater than 1.5 (Figure 3). This rank was determined by a one-sided Smirnov–Grubbs test (5% significance).

(2–3) Comments were decomposed into morphemes using the ChaSen (9), nouns, verbs, and adjectives were extracted. Then these were multiplied by the evaluation value for each rank of each word.

(2–4) From each distribution of comment ranks, we took the maximum rank as comment rank.

The comment rank evaluation for “Voice of explanation is small” is shown in Figure 4. In Figure 4, comments were decomposed into morphemes, and “voice,” “explanation,” and “small” were extracted. These were multiplied by evaluations at each rank for each word, and the maximum rank was taken as the comment rank evaluation.

As evaluation example, we show the results given by Evaluator 1. Table 2 shows the comment evaluations at each iterations of the recursive evaluation for the comments evaluated by Evaluator 1. We note that the recursive evaluation terminated at the second iteration. The estimation results for unrated comments are shown later in Table 4 and discussed there.

To terminate repetition, we used the difference between the reevaluation ranks of the rated comments and their manually evaluated original ranks. If the Euclidean distance was used to calculate the difference between reevaluation ranks and their manually evaluated original ranks, the average of the Euclidean distance was 2.6 after the first comment rank evaluation repetition.

Figure 5 shows the changes of the average of the Euclidean distance estimated after second evaluation for
WORD AND COMMENTS EVALUATION USING RECURSIVE EVALUATION IN LECTURE QUESTIONNAIRE

The teacher focuses on teaching what students do not understand. 5 5 5 4
It is hard to prepare handouts every time I attend the lecture. 2 1 2 3
This is a challenging class because I can study with senior students and there are additional assignments. 6 1 5 5

Table 2. Comment Evaluation at Each Iteration for Manually Evaluated Comments (Evaluator 1).

| Comments                                                                 | Manual evaluation | First iteration | Second iteration | Third iteration |
|--------------------------------------------------------------------------|-------------------|-----------------|------------------|-----------------|
| The teacher focuses on teaching what students do not understand.         | 5                 | 5               | 5                | 4               |
| It is hard to prepare handouts every time I attend the lecture.          | 2                 | 1               | 2                | 3               |
| This is a challenging class because I can study with senior students and | 6                 | 1               | 5                | 5               |
| there are additional assignments.                                        |                   |                 |                  |                 |

Figure 5. Number of Iterations.

Table 3. Estimated Word Rank for Each Evaluators (E1-E3).

| Word      | Estimated rank | Average | Variance |
|-----------|----------------|---------|----------|
| Allocation| 2 2 2 2        | 2.08    | 0.08     |
| Enjoyable | 5 5 4          | 4.83    | 0.14     |
| Design    | 5 5 4          | 4.83    | 0.14     |
| Create    | 5 5 4          | 4.83    | 0.14     |
| Wasteful  | 1 1 1          | 1.17    | 0.14     |
| Specialty | 5 5 5          | 5.00    | 0.17     |
| Easily viewable | 5 5 4 | 4.75 | 0.19 |
| Thankful  | 5 5 4          | 4.75    | 0.19     |
| Learning  | 5 5 4          | 4.75    | 0.19     |
| Java      | 5 5 4          | 4.75    | 0.19     |
| Emotion   | 1 5 4          | 3.58    | 2.24     |
| Homework  | 1 5 4          | 3.83    | 2.31     |
| Annoying  | 1 5 4          | 3.25    | 2.35     |
| University| 1 2 4          | 3.58    | 2.41     |
| Think     | 1 5 4          | 3.67    | 2.56     |
| Test      | 1 5 4          | 3.67    | 2.56     |
| Friendly  | 5 5 4          | 3.58    | 2.91     |
| Utmost effort | 6 6 2 | 4.01 | 3.74 |

4. Evaluation Examples for Word and Comment Ranks

On the basis of the comment ranks given in Table 1, we evaluated words and comments by totaling the frequencies of ranks. The word ranks are shown in Table 3. A list of words used to generate comments was used in the word evaluation, and rather than using comment ranks to evaluate words, word ranks were used to evaluate the comments in Table 4.

The word and comment evaluations were performed as follows.

(1) Among the word evaluations, there were those with a stable numerical rating that was common among the raters and those with variable ratings depending on the rater. The word “enjoyable” had an average rating of 4.83 with a low variance of 0.139; thus, it had a good and stable valuation. Similarly, “to create” had a low variance and high average rating; thus, it had a good and stable valuation. Conversely, “wasteful” had a low variance and low average rating; thus, it could be considered a term with poor valuation. On the other hand, the words “think” and “utmost effort” had a high variance; thus, their evaluations varied depending on the evaluator.

(2) Regarding the evaluations of the comments, comments with a low variance in the numerical ratings could be considered to have a common valuation among raters. Therefore, it was necessary to consider comments with low variance and low average ratings. We can be considered that “verifying answers to problems” and “time allocation” required consideration. On the other hand, for comments with a low variance and high average rating,
we considered that these concepts were what students desire from learning support for each assessed category, such as “practical training is possible” and “preparation and review are possible.”

5. Evaluation of Teachers and Lectures Using Recursive Evaluation

With the recursive evaluation method, it is possible to obtain rating distribution for comments about teachers and lectures. The rating distribution for comments about teachers is given in Table 5, and the rating distribution for comments about lectures is in Table 6. In addition, plots for the average ratings and rating variances are shown in Figure 6.

For the teacher evaluations, the variance was not small. Thus, in this section, we evaluate the relative variance. For teachers B and G (Table 5), we considered that they had high comment ratings without a significant variance. In contrast, teacher D had a low comment rating with a moderate variance.

On the other hand, regarding the lecture assessments, lecture A2 had extremely high comment ratings with a low variance, and lecture F3 had high comment ratings with a rather low variance. However, lecture F1 had low comment ratings despite the fact that it did not receive many comments. The correlation coefficient between the average and variance of the comment ratings was $-0.613$ for teachers and $-0.645$ for lectures. A high negative correlation was observed for both. The correlation coefficient between the average comment rating and comment count was $-0.00854$ for teachers and $-0.495$ for lectures. For teachers, the results indicated a lack of correlation, and we considered that the relationship between the rating and comment count was not significant. For lectures, comment assessments were performed for each lecture although we showed only selected lectures; thus, comment ratings tended to be slanted. Although it was only our impression, we thought that good comments evaluated by the proposed method often mentioned the fact that they were devising a lecture method.

6. Conclusion and Future Work

In the recursive evaluation performed in this study, predictive evaluations of comment and word were con-
DUCTED BASED ON THE HIGHEST EVALUATION FREQUENCIES IN DISTRIBUTION OF RANK. IN PARTICULAR, THE DIFFERENCE BETWEEN THE EVALUATION AND ESTIMATED RANKS FOR EVALUATED COMMENTS WAS MINIMIZED. HOWEVER, WE DID NOT EVALUATE THE ESTIMATES FOR UNRATED COMMENTS.

IN TERMS OF RELIABILITY, EVALUATION OF THIS PART WILL BE THE FOCUS OF FUTURE WORK.

IT IS NECESSARY TO PERFORM TEACHER/LECTURE EVALUATIONS VIA COMMENT EVALUATIONS IN CONSIDERATION OF POLARITY REVERSAL Owing TO THE NEGATIVE WORDS, SUCH AS THE JAPANESE AUXILIARY VERB “NAI.” EXPRESSIONS THAT ARE LIKELY TO CAUSE POLARITY INVERSION IN COMMENTS OFTEN CONTAIN THE AUXILIARY VERB “NAI,” thus, IT IS NECESSARY TO REMOVE SUCH EXPRESSIONS.

THE PROPOSED METHOD IS REQUIRED TO RATED COMMENTS TO A CERTAIN EXTENT. IN OTHER WORDS, IT IS NECESSARY TO EVALUATE SUBSTANTIAL WORDS. ONE FUTURE ISSUE IS REDUCING COMMENT EVALUATIONS USING THE PROPAGATION OF WORD EVALUATIONS.

REFERENCES

(1) Turban, J. W.: “Students Prefer Audience Response System for Lecture Evaluation”, Int’l J. of Emerging Technologies in Learning, Vol. 6, No. 4, pp. 52–55 (2011).

(2) Jarahi, L. and Najaf, M.: “Evaluation of Teaching through Lecture with New Methods of Student-Centered Teaching in Medical Students”, Future of Medical Education J., Vol. 3, No. 4, pp. 6–9 (2013).

(3) Nigam, K., McCallum, A., Thrun, S. et al.: “Text Classification from Labeled and Unlabeled Documents Using EM”, Machine Learning. Vol. 39, No. 2, pp. 103–134 (2000).

(4) Turney, P. D.: “Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews”, Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424 (2002).

(5) Takamura, H., Inui, T. and Okumura, M.: “Extracting Semantic Orientations Using Spin Model”, Proc. of ACL 2005, pp. 133–140 (2005).

(6) Rish, I.: “An Empirical Study of the Naive Bayes Classifier”, IJCA 2001 Workshop on Empirical Methods in Artificial Intelligence, Vol. 3, No. 22, pp. 41–46 (2001).

(7) Vapnik, V. N.: Statistical Learning Theory, Wiley (1998).

(8) Shawe-Taylor, J. and Cristianini, N.: Kernel Methods for Pattern Analysis, Cambridge University Press (2004).

(9) Asahara, M. and Matsumoto, Y.: “Extended Models and Tools for High-Performance Part-of-Speech Tagger”, Proc. of COLING 2000, pp. 21–27 (2000).