A Scoring System for Short Answers on the Test in a Large Group

Jae-Young Lee

Department of Computer Engineering, Hallym University, Gangwon, Korea 200-702
jylee@hallym.ac.kr

Abstract. In this paper, we have developed the scoring system that scores short answers based on the score table updated by average scores of non-existing answers. The accuracy and the consistency are very important, because the score influences a life. Automatic mark systems have consistency but need more accuracy. In the paper and pencil, a consistency is difficult to maintain. To achieve accuracy and consistency, the scoring system consists of three passes. The first pass is to score the applicant's answers based on the ready-made score table if it is in the table. If not, the second pass updates the table with the average of credits for which committee members evaluate the non-existing answer. Finally, the third pass is to score non-existing answer based on the updated table.

Keywords: remote education, subjective-type evaluation, automatic scoring system, Internet-based scoring system.

1 Introduction

Learners in an information society can learn immediately the knowledge and technologies they need at any time and at any place and educational activities, such as, learning, testing, and evaluation, are done freely between huge learners and teachers in cyber education via the Internet [1] and [2].

The important thing in this education is fair evaluation for subjective questions to increase the quality of education. How to evaluate the learners' abilities can normally be classified as multiple choices or subjective tests. The multiple-choices can increase fairness and reliability but decrease the quality of education. On the other hand, the subjective tests can improve the quality of education by measuring the cognitive abilities, but lower the fairness and reliability. The biggest drawback of evaluating subjective tests is the lack of fairness.

There were several researches to solve these problems in evaluation of the subjective test.

After applicants solve the subjective questions through Internet, raters are informed the finish of the test by Internet, or telephone. Then, the raters should quickly score the answers through Internet, and the system notifies each result to applicant [3].

In automatic scoring, Park and Kang [4] proposed the model which grades for the subjective-type evaluation, and designs and implements the evaluation system using

J.J. Park et al. (Eds.): NPC 2012, LNCS 7513, pp. 42–47, 2012.

the synonym thesaurus and the system results the 73% success rate. Kim et. Al. [5] had developed an intelligent grading system, which scores descriptive examination papers automatically, based on Probabilistic Latent Semantic Analysis(PLSA) and it can acquire about 74% accuracy of a manual grading, 7% higher than that from the Simple Vector Space Model. Kang [6] designed and implemented a subjective-type evaluation system using syntactic and case-role information and the system results the 75% success rate. Scores have a great influence on applicants' life, such as, admissions and promotions, the automatic scoring system needs more accuracy. On the other hand, subjective tests and the answers written in pencil for large group exam are scanned to grade the pencil-and-paper test. Internet-based scoring system that two or three raters score the scanned paper instead of the pencil-and-paper test to increase the reliability was studied [7]. It also has the drawback that raters can score unfairly with subjective judgments.

To solve the problems, we have proposed the scoring system that particularly scores non-existing answers based on the new score table updated by average of new scores. The new scores are evaluated non-existing answers by members of committee. This system has three passes to score fairly. The first pass is to score the applicant's answers based on the ready-made score table if it is found in the table. If not, committee members evaluate the non-existing answer and return new credits to the system. The second pass updates the table with the average of the new credits. Finally, the third pass is to score non-existing answers based on the updated score table.

2 Paper and Pencil Scoring

In the paper and pencil scoring, raters evaluate each item of the subjective questions by writing score by hand. It is used for evaluating a small group, specifically, the group of high-quality human resources, but is not suitable for large group because of fairness. In the major field on the secondary teacher certification test at domestic, there are six steps to process between setting the subject test and scoring the answers by hand as follows [7]:

[Step 1] Setting questions:
-To set questions to majors for the measurement of higher-order thinking skills.
[Step 2] Making answer sheets and the criteria of scoring:
- At first, every member of committee scores each item of question.
- And then every member should systematically check the validity and relevance of the contents through group discussion.
[Step 3] Simulation to score 3 times:
- Every rater or committee member scores 3 times each question of every majors and a group updates answer sheets and the criteria of scoring after analyzing the results.
[Step 4] Determining the final answer sheets and the final criteria of scoring:
- Committee members and raters should determine final answer sheets and the criteria of scores after checking the validity and relevance of the contents and confirming the purpose of the questions through group discussion.

[Step 5] Scoring:
- Each question is independently scored three raters.
- Final score of each can be calculated as the average of 3 scores.

[Step 6] Transfer of Score results:
- Education office in-local or city will take over the score results and test papers from raters.

The reason for these complex procedures to score the subjective questions is to increase the reliability in Paper and pencil scoring. However, it increases processing time and cost because it needs complex procedures to reduce problems occurred by the difference of individuals among members, to maintain the consistency of the scoring, and to inform scoring results to raters.

3 A Scoring System for Short Answers

In this paper, the scoring system for short answers on the test accepts questions and score tables from the committee and then saves them in the database. The table consists of pair of correct answer and its credit, or another pair of similar answer and its credit. The system should not only show questions to the applicants but also store the answer received from the applicants into the database. There are three passes to score the answers. The first pass is to score automatically the applicant's answers based the score table if it is the same as the correct answer, or, similar answer. If not, raters evaluate the non-existing answer and return new credits, and the second pass updates the table with the average of new credits. Finally, the third pass is to score non-existing answers based on the updated score table. Such a clients and server system including database is shown in Fig. 1.

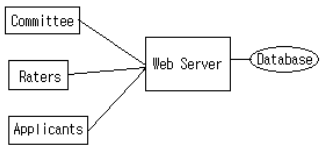

Fig. 1. A scoring system for short answer question

4 A Scoring Algorithm for Short Answers

In the process of scoring short answer questions, it's said that the feature of the short answers tends to classify an answer, so the feature will make answers easier on the test in a huge group. On the other hand, the feature of the essay question is difficult to classify an answer, so it is not useful to score on the test in a huge group.

The scoring system for short answers on the test accepts questions and score tables from the committee. For scoring, the system score an answer based on the ready-made table. But for the non-existing answers that are not in the table, non-existing answers are grouped to make score easier. Rater gives a score every group. The system updates the table with the average of various new credits which raters give and score

again the non-existing answers based on the updated score table. Such an algorithm consisting of 3 passes is as follows:

Pass 1: The procedure to score answers in the score table

[Step 1] If there is any applicant's answer, then read the answer and go to [Step 2]. Otherwise, go to Pass 2.

[Step 2] Look up the answer in the score table.

[Step 3] If the answer and credit exist in the score table, then the answer will be scored based on the score table and go to [Step 1].

[Step 4] Otherwise, the answer will be appended in the table and go to [Step 1].

Pass 2: The procedure to update the table with the average of new credits for non-existing answers

[Step 1] Send the score table including non-existing answers to committee members.

[Step 2] Average the scores evaluated by members of committee.

[Step 3] Update the score table with the average and go to Pass 3.

Pass 3: The procedure to score non-existing answers

[Step 1] If there is non-existing answer, read the non-existing answer and go to [Step 2], else stop.

[Step 2] Look up the non-existing answer in the updated table.

[Step 3] If the non-existing answer and new credit exist in the updated table, then the answer will be scored based on the updated table and go to [Step 1].

[Step 4] Otherwise, go to [Step 1].

The score table consists of number, question, a pair of an answer and a credit, as shown in Table 1. The pair can be one of four types: a pair of correct answer and credit, or a pair of similar answer and credit for Pass 1, or a pair of non-existing answer and credit for Pass 2 and Pass3.

Table 1. Score table consisting of questions, answer, and credit

NO	Question	Answer	Credit
1			
2			
3			

5 Comparative Analysis

Performance of scoring is able to meet a certain number of criteria. The most important things of these are accuracy, fairness, consistency, processing time, and human resource.

Accuracy A(x) is the probability of scoring answers correctly. Fairness F(x) is the probability of scoring answers with objectivity justly. Consistency C(x) is the probability that the same score is given to the same answer from first to the end. Processing time T(x) is the time that scores from first to the end. Human resource H(x) is the number of humane needed to score from first to the end.

The most important thing to evaluate subjective questions is accuracy, fairness, and consistency, because the evaluation results have a significant impact on the lives. The criteria of accuracy, fairness, and consistency are more important than time.

In these respects, there are comparisons of three types of scoring: the automatic scoring system, the paper and pencil scoring, and the scoring system for short answers.

First, the performance of the automatic scoring system is as follows:

$$Pa(x) = Aa(x) + Fa(x) + Ca(x) + 1/\,Ta(x) + Ha(x) = Aa(x) + 1/\,Ta(x). \quad (1)$$

Where Aa(x), Fa(x), Ca(x), Ta(x), and Ha(x) mean accuracy, fairness, consistency, processing time, and human resources for the automatic scoring system, respectively. The accuracy Aa(x) of this system falls, so this system is not suitable. Particularly, a wrong result is a fatal influence on a person's life, although fairness and consistency are perfect and process is very fast. The answers are scored by computer instead of human resources.

In teacher appointment tests, the paper and pencil scoring is still carried out because of the accuracy problem. The performance of this scoring is as follows:

$$Pp(x) = Ap(x) + Fp(x) + Cp(x) + 1/\,Tp(x) + Hp(x). \quad (2)$$

Where Ap(x), Fp(x), Cp(x), Tp(x), and Hp(x) mean accuracy, fairness, consistency, processing time, and human resources for the paper and pencil scoring, respectively. The accuracy Ap(x) is better than Aa(x), so this system have been used for evaluating a small group of high-quality human resources, although fairness and consistency are not perfect and it needs a few days, many raters, and high cost.

Before analyzing the scoring system for short answers, let us compare pass 2 in the scoring algorithm and [Step 4] in the paper and pencil scoring.

To search similar answers of all possible cases before scoring in the paper and pencil scoring, raters should make score 3 times for every question and then check both answer sheets and the criteria of scoring, as shown in [Step 3]. In the [Step 4], the committee members and raters should determine final answer sheets and the criteria of scoring after checking the validity and relevance of the contents and purpose of the questions through group discussion. To do these, many human resources, lots of processing time, and high costs are needed.

In the scoring algorithm for short answers, on the other hand, [Step 3] and [Step 4] in the paper and pencil scoring is simply replaced by the pass 2. The key point of the credit in the non-existing answers is to use the average of credits evaluated by every rater, instead of the criteria after their discussion. The discussion is the important factor to increase time, human resource and cost.

Thus, the performance of the algorithm is as follows:

$$Ps(x) = As(x) + Fs(x) + Cs(x) + 1/\,Ts(x) + Hs(x) = As(x) + 1/\,Ts(x) + Hs(x). \quad (3)$$

Where As(x), Fs(x), Cs(x), Ts(x), and Hs(x) mean accuracy, fairness, consistency, processing time, and human resources for the scoring system for short answers, respectively. In the scoring system, the accuracy As(x) is better than Ap(x). The fairness and consistency are perfect and process is fast. The raters is to score only non-existing answers, and this scoring requires much less labor than the paper and pencil scoring, so it is suitable for scoring short answers in huge group. Therefore, the algorithm has the advantage to reduce human resources, processing time, and costs.

6 Conclusions

The most important thing to evaluate subjective questions is an accuracy and consistency, because the evaluation results have a significant impact on the lives of applicants. In this respect, automatic mark systems have consistency but need more accuracy. In the paper and pencil, a consistency is difficult to maintain.

To achieve accuracy and consistency, the scoring system scores the applicant's answers based on the ready-made score table, and it then updates the table with the average of various new credits for the non-existing answer. Finally, it scores non-existing answer based on the updated table.

In this paper, the algorithm for short answers has more accuracy than automatic mark system and less costs than the paper and pencil scoring.

Acknowledgments. This research was supported by Hallym University Research Fund, 2012(HRF-2012-06-001).

References

1. Reiser, R.A., Kegelmann, H.W.: Evaluating instructional Software: A review and critique of current method. Education Technology Research and Development 42(3), 63–69 (1994)
2. Jang, S.P., Lee, Y.M.: Alternative Formative Evaluation in Web Based Learning System. Journal of KACE 3, 43 (2000)
3. Bang, H., Kang, T.H., Kim, W.J., Won, D.H., Lee, J.Y.: A Web-based Grading System for Classifying Answers of Subjective Test. Proceeding of KIPS 28, 673–675 (2001)
4. Park, H.J., Kang, W.: Design and Implementation of a Subjective-type Evaluation System Using Natural Language Processing Technique. Journal of KACE 6, 207–216 (2003)
5. Kim, Y.S., Oh, J.S., Lee, J.Y., Chang, J.H.: An intelligent grading system for descriptive examination paper based on probabilistic latent semantic analysis, pp. 1141–1146. Springer, Heidelberg (2004)
6. Kang, W.S.: Design and Implementation of a Subjective-type Evaluation System Using Syntactic and Case-Role Information. Journal of KACE 10, 61–69 (2007)
7. Cho, J.M., Kim, K.H.: A Study on Design of the Internet-based Scoring System for Constructed Responses. Journal of KACE 10, 89–100 (2007)