International Journal of Engineering & Technology, 7 (3.33) (2018) 263-268

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

The Best Effort System to Score Subjective Answers of Tests in a Large Group

Jae-Young Lee[1]*

[1]Department of Computer Engineering and Smart Computing Lab, Hallym University
Chuncheon-si, Gangwon-do, 24252, Korea
**Corresponding author E-mail: jylee@hallym.ac.kr*

Abstract

The subjective tests can improve the quality of education by measuring the cognitive abilities, but the biggest drawback is the lack of fairness, consistency, and accuracy. To improve the drawback, we proposed the best effort system that scores the correct subjective answers based on the correct answer table made by committee members, then classifies the rest of subjective answers into groups of similar answers so that the latest automatic scoring systems and graders assign each reasonable credit to each group of similar subjective answers. In the scoring system, the groups of the similar answers are evaluated by raters and the latest automatic scoring systems, such as syntax tree comparison grading, and the syntax and semantic tree-oriented grading. All the scores for each similar answer are added and then an average for each similar is stored in the similar answer table. Finally, the system grades applicant's answers using the correct answer table and the similar answer table. This paper proposes the algorithm for the best effort scoring system to include the latest automatic scoring system in order to be as fair, consistent, and accurate as possible.

Keywords: *automatic essay scoring; automatic scoring system; content-based scoring ; Internet-based scoring system; short answer scoring; subjective-type evaluation*

1. Introduction

In an information society and the age of 100, humans can learn immediately the knowledge that we need at any time and at any place. There are educational activities, such as, learning, testing, and evaluation that are freely constructed between huge learners and teachers via the Internet in cyber education[1]. In the testing and evaluation, one of the most important things is fair evaluation for subjective questions to increase the quality of education as well as to have learning effects. There are various kinds of questions, such as true-false questions, cloze questions, multiple choice questions, subjective test, and so on in the way to evaluate the learners' abilities.

The system to generate true-false questions first search for an informative sentence from text in order to make true sentence and false sentence is made by replacing key word with antonym or inverting the meaning of the sentence using 'not'[2]. The cloze questions, known as fill-in-the-blank questions, are useful for evaluating the applicant's ability that finds the suitable word to fit a sentence with blank[3]. Multiple choice questions consist of questions, the corresponding correct answer, and some of the corresponding distractors. It is an efficient tool in order to measure applicants' achievements and it is also used worldwide both for evaluation and diagnostics of applicant's mental[4]. Although the multiple-choice questions can increase fairness and reliability, it has disadvantage to decrease the quality of education. On the other hand, the subjective tests are good to improve the quality of education by measuring the cognitive abilities, but it makes the fairness and reliability lower. However, the biggest drawback of evaluating subjective tests is the lack of fairness, consistency, and accuracy.

There were several researches for scoring the subjective tests to evaluate them with fairness and consistency. After lots of applicants take the subjective question tests through Internet, and then system informs raters of the end of the subject test through Internet, or telephone. The raters should quickly grade the applicants' subjective answers through Internet and the system notifies each result to each applicant [5]. Automatic scoring research proposed the system which grades subjective answers for the subjective-type evaluation, and the automatic scoring system was designed and implemented by using the synonym thesaurus and the system achieved a success rate of 73%. An intelligent grading system automatically grades descriptive examination papers 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[6]. In automatic scoring system for English, c-rater is the automatic scoring system that grades responses to context-based short answer questions of less than 100 words and it uses predicate argument structure, pronominal reference, morphological analysis and synonyms to assign full or partial credit to a short answer question instead of simply a string matching program and it agree with human graders about 84% of the time[7]. C-rater that is Educational Testing Service's technology grades the content of short student responses, where its major part in the scoring process is Model Building (MB) which generates a set of model answers that correspond to the rubric for each item or test question. The Model Building approaches to automating Model Building in c-rater instead of knowledge-engineered (KE), so c-rater achieves comparable accuracy on automatically built and KE models[8,9]. But the general scoring system has limitations because of the application of word and phrase-oriented grading. In the automatic scoring system to measure the similarity of vocabulary of the syntax tree-oriented re-

search for scoring essays, e-rater automatic scoring system expresses the results of natural language analysis in a set of structured features, and then it grades the essay in a way that compares standardized samples which are made by vectors[10,11]. Question Answering archives are considered as a very useful resource for instant access to comprehensive information in response to user queries. Access to this resource through Short Message Service (SMS) requires that a high precision automatic similar question matching system be built in order to decrease the search time. To decrease the number of SMS exchanges, the system models the problem as one of combinatorial search and it uses syntactic tree matching to improve the ranking scheme[12]. Because the result to score can have great influences on applicants' life, such as, admissions in the schools and promotions in the companies, the automatic scoring system requires more accuracy. On the other hand, subjective tests and the answers which are written in pencil for exams are scanned to grade the pencil-and-paper test in a large group. Internet-based scoring system that two or three raters grade the subjective answers using the scanned paper instead of the pencil-and-paper test was studied to increase the reliability. The pencil-and-paper scoring system has the drawback that is graded unfairly and inconsistently with subjective judgments and requires lots of manpower and processing time. Automatic scoring systems also have the drawback that is not perfect although it grades fairly and consistently and does not requires lots of manpower and processing time.

To score subjective answers as accurate as possible and to solve the unfair and inconsistent problems, we have proposed the best effort scoring system that first scores correct answers by using the correct answer table and scores similar answers by using the new similar answer table which is updated by average of new scores for the similar answers. The new scores of similar answers are evaluated by the latest automatic scoring systems and raters. The reason to include the latest automatic scoring systems is to improve as fair and consistent as possible. There are three passes in the best effort system to score subjective answers. The pass 1 is to score the applicant's answers which exist in the correct answer table. If it does not exist in the correct answer table, raters including the latest automatic scoring systems assign each credit to each group of the similar answers and return new credits to the best effort scoring system in the pass 2. The pass 2 updates the similar answer table with the average of the new credits. Finally, the pass 3 is to score similar answers based on the updated similar answer table.

2. Scoring of subjective answers

The subjective scoring systems are classified as two categories. One is paper and pencil scoring and the other is automatic scoring system. The most important criteria in the scoring subjective questions can be considered as accuracy, fairness, consistency, processing time, and human resource. In paper and pencil scoring, the accuracy is better than the accuracy in the automatic scoring system**,** so this paper and pencil scoring have been used for evaluating a small group of high-quality applicants, although fairness and consistency are imperfect as well as it requires more time, many man powers, and high cost. In the automatic scoring system, the accuracy is not perfect, although fairness and consistency are perfect as well as process is very fast.

2.1. Paper and pencil scoring

The subjective question is very suitable for the measuring applicants' higher-order thinking skills but it is very complicated and difficult to score the subjective questions. And it is also likely to be a mistake in scoring. Furthermore, the mistakes in scoring tend to have a great impact on an applicant's life. In the paper and pencil scoring, raters evaluate each answer of each subjective question one by one and then they write each score by hand.

For example of the secondary teacher certification test at domestic, there are complex steps which set the subject test and grade the subjective answers by hand. There are six steps in the procedure for both setting the subject test and scoring the answers by hand in the major field on the secondary teacher certification test at domestic as follows. The first step sets subjective questions. The second one makes answer sheets and the criteria of scoring. The third one simulates the procedure to score 3 times. The fourth one determines both the final answer sheets and the final criteria of scoring. The fifth one scores subjective answers. The last one transfers the results.

The paper and pencil scoring is useful for evaluating subject answers in a small group, especially, but it is not suitable for applicants in large group because it is difficult to keep fairness for all the subject answers.

2.2. Automatic scoring system using syntactic-semantic analysis

In the researches related to the syntax tree comparison grading, there are three types of research areas, such as word-oriented one and syntax tree-oriented one, syntax and semantic tree- oriented one.

In the automatic scoring system to use the similarity of vocabulary of the word-oriented research, when the system uses a large corpus to construct a semantic kernel and to score answers, it automatically scores the subjective answers using the meaning kernel and the Korean wordnet. The scoring system utilizes semantic information using Korean wordnet, but it extracts words through morpheme interpretation and uses the meanings of words extracted. Because the syntactic relationship between words is not used, the subjective scoring system has limitation for scoring.

In the automatic system to score subjective answers at the word and phrase level, the system designs the correct answer templates for words and phrases, and then it grades correct answers based on which the scoring of whether the answer matches the correct answer and the answer matches the concept was carried out.

In the automatic scoring system, the practical system uses the significant words between words, but it did not use syntactic analysis. Based on the previous automatic scoring system instead of the correct answer templates for words and phrases, the correct answer template was constructed from several student answers to improve scoring. But the system has limitations because of the application of word and phrase-oriented grading.

In the automatic scoring system to measure the similarity of vocabulary of the syntax tree-oriented research for scoring essays, e-rater automatic scoring system expresses the results of natural language analysis in a set of structured features, and then it grades the essay in a way that compares standardized samples which are made by vectors. In the automatic scoring system to use the syntax and semantic tree-oriented research in order to grade quality subjective problem, the system can analyze the meaningful relationship between the phrases in the answer, and then it grades subjective answers in a way that compares the meaning tree of the syntax that implies the structural meaning relationship between words and words using the knowledge of words.

3. The best effort system to score subjective answers

The best effort system is defined as the system that does its best to process tasks done. Similarly, the best effort system to score subjective answers is the system grades the subjective answers as fare and consistent as possible, because it is impossible to grade subjective answers in a large group particularly. The best way to score lots of subjective answer is to score the latest automatic scoring technologies, such as syntax tree comparison grading, and the syntax and semantic tree-oriented grading, in order to make scoring of subjective answer more fare and consistent.

3.1. A system to score subjective answers

In order to grade short subjective answers, the best effort scoring system reads short subjective questions, corresponding correct answers, and their full scores from the committee. And then the system makes question test sets and question table out of the questions in order to evaluate applicants and it simultaneously makes correct answer table out of both correct answers and their full scores in order to grade applicant answers. It saves them into the database. The system should not only send questions to the applicants but also store the answer from the applicants into the database.

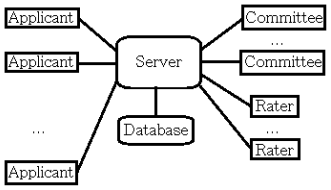

Fig. 1: A system to score subjective answers

In the scoring system, the similar answers which are not the correct answers are classified into groups of similar answers and stored the groups into the similar answer table. The groups of the similar answers are evaluated by the latest automatic scoring systems, such as syntax tree comparison grading, and the syntax and semantic tree-oriented grading. And finally raters check whether the evaluated credits by the latest automatic scoring systems is reasonable and assign each new credit to each unevaluated similar answer. All the credits for each similar answer are added and an average for each similar answer is calculated by each sum. And then the average value is stored in the similar answer table. Finally, the system grades applicant's answers using the correct answer table and the similar answer table. Such a clients and server system including database is shown in Figure 1.

3.2. Algorithm for the best effort system to score subjective answers

In the characteristics of scoring lots of subjective answers, it is easy to classify similar answers in the type of short subjective answers and the characteristics make it easy to score similar answers in a huge group. On the other hand, it is difficult to classify the subjective answers in the essay form, so it is not useful to score the similar answers on the test in a huge group. The best effort scoring system limits to grade the type of short subjective answers in this paper.

In the process to score short subjective answer, the best effort system to score subjective answers on the test first loads the subjective question table from database which is saved by the committee in advance. The subjective question table consists of subjective questions, the types of answers, corresponding scores, and the pointers of similar answers, as shown in Table 1.

Table 1: A subjective question table including an applicant's scores

No	Question	Type of Answer	Score	Pointer of Similar Answers
1	Question 1	0	5	1
2	Question 2	3	2	5
3	Question 3	1	4	7
	...	...	...	...
n	Question n	0	7	Pn(SA)

The first field represents the number of question, in which the total number of question means n subjective questions for each applicant. The second field means the subjective question, and the third type field indicates whether the answer is the location of a correct answer at Table 2 or the location of a similar answer at Table 3. Type ***0*** means that the answer is correct answer, while the other type means that the answer is similar answer and the number indicates the relative displacement of the similar answer at Table 3. The fourth field indicates each score of each subjective answer for an applicant's answer. For example, the score on question ***1*** is ***5*** points which is full score for the first subjective answer. On the other hand, the score on question ***2*** is ***2*** points which is assigned by the latest automatic scoring technologies, such as syntax tree comparison grading, and the syntax and semantic tree-oriented grading, raters. The fifth field indicates whether the answer is the pointer of the correct answer of the corresponding question at Table 2 or the pointer of the first similar answer of the corresponding question at Table 3. For example, the first pointer ***1*** is the location of the correct answer of the first question in the correct answer table, as shown in Table 2 and the second pointer ***5*** is the location of the first similar answer of the second question in the similar answer table, as shown in Table 3.

Table 2 shows the table that consists of correct answers, the types of answers, their scores, and the statics of correct answers. The first field represents the number of each correct answer, in which ***n*** means the total number of correct answers. The second field is each correct answer for each question and the third field is ***0*** that means full score and Type ***0*** means that the answer is correct. In the fourth field, the score of each answer indicates full score ***5***. The fifth statistics field is the number of applicants who write correct answers for the question.

Table 2: A correct answers table including full scores of correct answers

No	Correct Answer	Type of Answer	Score	Statistics of Correct Answers
1	Correct Answer 1	0	5	S1(CA)
2	Correct Answer 2	0	5	S2(CA)
	...	...	...	...
	...	...	...	...
n	Correct Answer n	0	5	Sn(CA)

Table 3 shows the table that consists of classified similar answers, the types of answers, their scores, the statics of similar answers, and links. The first field represents the number of each similar answer, in which ***m*** means the total number of similar answers for all the questions. For example, the first question has 4 similar answers and the second question has 2 similar answers. The second field indicates the relative displacement of the similar answer for each question. The fourth score field indicates the average score of corresponding question which is assigned by the latest automatic scoring technologies, such as syntax tree comparison grading, and the syntax and semantic tree-oriented grading, raters. The fifth statistics field is the number of applicants who write each similar answer for every question. The last field indicates the next similar answer for the same question. The link is used for finding the similar answer, until an applicant's similar answer matches the similar answer at this table.

Table 3: A similar answers table including scores of similar answers

No	Similar Answers	Types of Answer	Score	Statistics of Similar Answers	Link
1	Similar Answer 11	1	4	S11(SA)	L2
2	Similar Answer 12	2	3	S12(SA)	L3
3	Similar Answer 13	3	2	S13(SA)	L4
4	Similar Answer 14	4	1	S14(SA)	L1
5	Similar Answer 21	1	4	S21(SA)	L6
6	Similar Answer 22	2	2	S21(SA)	L5
7	Similar Answer 31	1	4	S31(SA)	L7
	...	...	...	...	...
m	Similar Answer nm	1	5	Snm(SA)	Lm

Consider the procedures for the best effort scoring an applicant's answers using the above tables, such as the subjective question tables, the correct answer table, and the similar answer table. After lots of applicants solve the subjective question, the best effort system reads each applicant's answers from database. If the applicant's answer is in the correct answer table, the system loads the type ***0*** for the correct answer and the full score ***5*** from the correct answer table and then it saves the type and the full score in the

type field and the score field of the subjective question table, respectively. Otherwise, it classifies the applicant's answer by the type of the similar answer and save the corresponding similar answer in the similar answer table. Additionally, frequencies of the correct answer and similar answer are updated at statistics fields of the correct answer table and the similar answer table, respectively.

After it processes all the answers, the best effort system offers evaluators the information that is made by three tables. The evaluators, such as new automatic scoring technologies and raters, assign a reasonable point to every group of similar answers after evaluating the similar answers. The system saves the average score in the score field of the similar table for scoring the applicants' answers. There are 3 passes in the algorithm that grade subjective answers. The first pass is to classify similar answers to grade easily. The second one is that evaluators assign a point to each similar answer and saves the point in the score field of similar answer table. The last one is to grade every applicant's answer using the value of the score field of the correct answer table or the similar table.

Pass 1: The procedure that classifies similar answers and saves them in the similar table

Step 1: Read an applicant's answer and if there are not an applicant's answers, then go to **Pass 2**.

Step 2: Search for the new applicant's answer ***nA*** in the correct answer table ***CT***.

Step 3: If the new applicant's answer ***nA*** does not exist in the correct answer table ***CT***, then go to **Step 5** in order to process it as a similar answer.

Step 4: Procedure for correct answers:

Load a value of type ***tc*** and a full score ***sc*** of the correct answer from the correct answer table ***CT***, save the type and the score into the type field ***tq*** and the score field **sq** of the question table ***QT***, and increase the statistics field ***stc*** of the correct answer table ***CT***. Go to **Step 1**.

Step 5: Procedure for similar answers:

Search for the new applicant's answer in the similar answer table ***ST***.

Step 6: If the new answer exists in the similar answer table ***ST***, then go to **Step 8**.

Step 7: Procedure for classifing the new similar answers:

Classify the applicant's answer by group of similar answers and save the new applicant's answer ***nA*** into a similar answer field ***sA***. And update new value of type for the similar answer and save it in both the type field ***ts*** of the similar answer table and the type field ***tq*** of the question table. Save the location of the similar answer into the pointer field of similar answer ***pq*** of the question table. Increase the statistics field ***sts*** and update the link of the previous similar answer ***ls*** at the similar answer table ***ST***. Go to **Step 1**.

Step 8: Procedure for updating information of the existing similar answers:

Load a value of type ***ts*** of the similar answer table ***ST*** and save the value into the type ***tq*** of the question table ***QT***. Increase the statistics field ***sts*** of the corresponding group of similar answers in the similar answer table ***ST*** and go to **Step 1**.

Pass 2: The procedure to update score field with the average of new scores for similar answers

Step 1: For each question, send raters a set of information, such as a correct answer, its full score, its statistics of the correct answer table ***CT***. And send evaluators, such as new automatic scoring technologies and raters, the set of information, such as similar answers, their statistics of the similar answer table ***ST***.

Step 2: Average the scores evaluated by the raters.

Step 3: Update the score field of the similar table with the average of score.

Step 4: If the question is not final, then go to **Step 1**, else go to **Pass 3**.

Pass 3: The procedure to grade each applicant's answer using the correct answer table and the similar answer table

Step 1: Read each subjective answer from each applicant.

Step 2: Search for the correct the answer at the correct answer table ***CT***.

Step 3: If the applicant's answer is in the correct answer table ***CT***, the best effort system loads the type and the full score of the correct answer from the correct answer table ***CT*** and then it saves the type and the full score in the type field and the score field ***sq*** of the subjective question table ***QT***, respectively. Save the location of the correct answer into the pointer field ***pq*** at ***QT.*** Additionally, frequencies of the correct answer are updated at statistics fields of the correct answer table. Go to **Step 5**.

Step 4: If the applicant's answer is in the similar answer table ***ST***, the best effort system loads the type and the score of the similar answer from the similar answer table ***ST*** and then it saves the type and the full score in the type field and the score field ***sq*** of the subjective question table ***QT***, respectively. Save the location of the similar answer into the pointer field ***pq*** at ***QT.*** Additionally, frequencies of the similar answer are updated at statistics fields of the similar answer table.

Step 5: If the set of answers is not processed, then go to **Step 1** in order to grade the next answer, else if the answers of all student is not processed then go to **Step 1** in order to grade the next applicant's answer. Otherwise go to stop.

4. Implementations and discussions

In the best effort system to score subjective answers of tests in a large group, it accepts applicants' answers, a correct answer table, and then it scores the subjective answers by the correct answer table. The answers which are not in the table are classified and they are saved in a similar table without scores and then the similar answers in the similar answer table are graded by raters through the Internet. Finally, the system scores the applicant's answers by both the correct answer table and the similar answer table with scores. The system is implemented on environment of JSP, apache server, file system and mysql.

For example, the system give applicants two subjective questions, such as "what are five components of data communications?", and "What is TCP?". The applicants write various answers including correct answers through Internet. In the first subjective question, the correct answer with a perfect score 6 is "They are sender, receiver, message, transmission line, and protocol." and the other similar answers without scores are "They are sender, receiver, message, and transmission line.", "They are sender, receiver, message, and protocol.", "They are sender, receiver, and protocol.", and so on, as shown in Figure 2.

In the second subjective question, the correct answer with a perfect score 4 is "TCP is a connection-oriented and reliable protocol." and the other similar answers without scores are "TCP is a connection-oriented protocol.", "TCP is a reliable protocol.", "TCP is a connection-oriented and unreliable protocol.", "TCP is a connectionless and reliable protocol.", "TCP is a connectionless and unreliable protocol.".

Question No	Questions		
1	What are five components of data communications?		
Kinds of Answers	Applicant's Answer	Statistics	Score
Correct Answer	They are sender, receiver, message, transmission line, and protocol.	110	6
Similar Answer 1	They are sender, receiver, message, and transmission line.	87	
Similar Answer 2	They are sender, receiver, message, and protocol.	73	
Similar Answer 3	They are sender, receiver, and protocol.	68	
Similar Answer 4	They are message, transmission line, and protocol.	52	
Similar Answer 5	They are sender, receiver, and message.	46	
Similar Answer 6	They are sender, and message.	37	
Similar Answer 7	They are sender and receiver.	21	
Similar Answer 8	It is sender	19	

Fig. 2: The screen that a rater can grade the similar answers

Question No	Questions		
2	What is TCP		
Kinds of Answers	Applicant's Answer	Statistics	Score
Correct Answer	TCP is a connection-oriented and reliable protocol.	127	4
Similar Answer 1	TCP is a connection-oriented protocol.	97	2
Similar Answer 2	TCP is a reliable protocol	85	2
Similar Answer 3	TCP is a connection-oriented and unreliable protocol.	76	1.5
Similar Answer 4	TCP is a connectionless and reliable protocol.	61	1.5
Similar Answer 5	TCP is a connectionless and unreliable protocol.	24	0

Fig. 3: The screen to show the similar answers with scores

Figure 3 shows a correct answer with a perfect score 4 as well as the five similar answers with scores 2, 2, 1.5, 1.5, and 0 which are graded by raters after the raters check and discuss with other members in the committee for scoring subjective questions through Internet, where the statistics which represent the number of applicants to write the corresponding similar answer are chosen arbitrary numbers for understanding it easily.

There are some criteria in the performance for the system to score subjective question. The most important criteria in the scoring subjective questions can be considered as accuracy, fairness, consistency, processing time, and human resource in three types of scoring, such as the automatic scoring system, the paper and pencil scoring, and the best effort scoring system.

Let accuracy $\boldsymbol{P_a(x)}$ and Fairness $\boldsymbol{P_f(x)}$ be the probability to score subjective answers correctly and the probability to score subjective answers fairly, respectively. The fairness is the ratio of grading the same score on the same subjective answer. Let consistency $\boldsymbol{P_c(x)}$ and processing time $\boldsymbol{T(x)}$ be the probability that gives the same answer the same score and the time that requires for scoring them in the same way as ever, respectively. Human resource $\boldsymbol{M(x)}$ is the number of humane needed to score them during the process of evaluating subjective answers. In the evaluation of subjective questions, the criteria of accuracy, fairness, and consistency are more important than those of processing time and human resource, because the evaluation results have a significant impact on the applicants' lives.

Consider comparisons of performances in three types of scoring: the automatic scoring system, the paper and pencil scoring, and the best effort scoring system.

In the automatic scoring system, let $\boldsymbol{P_{a1}(x)}$, $\boldsymbol{P_{f1}(x)}$, $\boldsymbol{P_{c1}(x)}$, $\boldsymbol{T_1(x)}$, and $M_1(x)$ be accuracy, fairness, consistency, processing time, and human resources, respectively. The accuracy $P_{a1}(x)$ of this system is not perfect. The system has the disadvantage that a mistaking in grading can affect a person's life, although fairness $\boldsymbol{P_{f1}(x)}$ and consistency $\boldsymbol{P_{c1}(x)}$ are perfect as well as process is very fast.

The paper and pencil scoring is still carried out in teacher appointment tests. Let $\boldsymbol{P_{a2}(x)}$, $\boldsymbol{P_{f2}(x)}$, $\boldsymbol{P_{c2}(x)}$, $\boldsymbol{T_2(x)}$, and $\boldsymbol{M_2(x)}$ be accuracy, fairness, consistency, processing time, and human resources in the paper and pencil scoring, respectively. The accuracy $\boldsymbol{P_{a2}(x)}$ in the paper and pencil scoring is better than the accuracy $\boldsymbol{P_{a1}(x)}$ in the automatic scoring system, so this paper and pencil scoring have been used for evaluating a small group of high-quality applicants, although fairness $\boldsymbol{P_{f2}(x)}$ and consistency $\boldsymbol{P_{c2}(x)}$ are imperfect as well as it requires more time, many man powers, and high cost.

Lastly, let $\boldsymbol{P_{a3}(x)}$, $\boldsymbol{P_{f3}(x)}$, $\boldsymbol{P_{c3}(x)}$, $\boldsymbol{T_3(x)}$, and $\boldsymbol{M_3(x)}$ be accuracy, fairness, consistency, processing time, and human resources in the best effort scoring system, respectively. The accuracy $\boldsymbol{P_{a3}(x)}$ in the algorithm for the best effort scoring system is better than the accuracy $\boldsymbol{P_{a2}(x)}$ for the paper and pencil scoring. The fairness $\boldsymbol{P_{f3}(x)}$ and consistency $\boldsymbol{P_{c3}(x)}$ are perfect and processing time of the best effort scoring system is faster than that of the paper and pencil scoring but slower than that of the automatic scoring system. Because the latest automatic scoring technologies as well as raters scores only similar answers classified by the best effort scoring system to obtain better results. The best effort scoring system also requires much less labor than the paper and pencil scoring, so it is suitable for scoring subjective answers in huge group.

5. Conclusion

In the evaluation using the subjective questions, the most important thing is an accuracy, fairness and consistency of scoring the subjective answers, because a mistaking in evaluating can affect an applicant's life. Automatic scoring systems have consistency but its accuracy is not perfect. In the paper and pencil scoring, it is difficult to maintain its consistency and it increases processing time, human resource, and cost. To improve accuracy, fairness and consistency, the best effort scoring system grades the applicant's answers based on the correct answer table, and it then updates the similar answer table with the average of points for the similar answers. Before raters grade similar answers, the latest upgraded automatic scoring systems grade the same answers as perfect as possible. Finally, it grades applicant's answers using these tables. In this paper, the algorithm for the best effort scoring system has more accuracy than automatic scoring system and it also has less cost than the paper and pencil scoring.

Acknowledgement

The research was supported by Hallym University Research Fund, 2012(HRF-201206-001).

References

[1] Reiser RA & Kegelmann HW (1994), Evaluating instructional Software: A review and critique of current method. *Education Technology Research and Development*, 1994;Vol.42, No.3, 63-69.

[2] Lee JY (2017), Dynamic Relocation of True-False Questions Using Ready-made Arrays with Random Numbers. *International Journal of Software Engineering and Its Applications*, Vo.10, No.8, 91-100.

[3] Correia R, Baptista J, Eskenazi M & Mamede N (2012), Automatic Generation of Cloze Question Stems. *In Computational Proceeding of the Portuguese Language, Springer-Verlag Berlin Heidelberg*, 168-178.

[4] Majumder M & Saha SK (2015), A System Multiple Choice Questions: With a Novel Approach for Sentence Selection. *Proceedings of the 2nd Workshop on Natural Language Proceeding*, 64-72.

[5] Fairon C (1999), A Web-based System for Automatic Language Skill Assessment: EVALING. *Proceedings of Computer Mediated Language Assessment and Evaluation in Natural Language Processing Workshop*, 62-67.

[6] Kim YS, Oh JS, Lee JY & Chang JH (2004), An intelligent grading system for descriptive examination paper based on probabilistic la-

tent semantic analysis. *Springer-Verlag Berlin Heidelberg*, 1141-1146.
[7] Leacock C & Chodorow M(2003), C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities*, Vol.37, No.4, 389-405.
[8] Sukkarieh JZ & Stoyanchev S (2009), Automating Model Building in C-rater. *Proceeding of the 2009 Workshop on Applied Textual Inference*, 61-69.
[9] Sukkarieh JZ & Blackmore J (2009), C-rater: Automatic Content Scoring for Short Constructed Response. *Proceedings of the Twenty-Second International FLAIRS Conference*, 290-295.
[10] Attali Y & Burstein J (2006), Automated Essay Scoring With e-rater® V.2. *Journal of Technology, Learning, and Assessment*, Vol.4, No.3, 1-31.
[11] Dikli S (2006), An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment*, Vol.5, No.1, 1-36.
[12] Langer A, Banga R & Mittal A (2010), Subramanian LV. Variant Search and Syntactic Tree Similarity Based Approach to Retrieve Matching Questions for SMS Queries. *'10 Proceedings of the Fourth Workshop on Analytic for Noisy Unstructured Text Data*, 67-72.