KASS: Korean Automatic Scoring System for Short-answer Questions

Eun-Seo Jang[1], Seung-Shik Kang[1], Eun-Hee Noh[2],
Myung-Hwa Kim[2], Kyung-Hee Sung[2] and Tae-Je Seong[2]
[1]*School of Computer Science, Kookmin University, Seoul, Republic of Korea*
[2]*Korea Institute for Curriculum and Evaluation, Seoul, Republic of Korea*

Keywords: Automatic Scoring, Short-answer Questions, Token-based Answer Template, Morphological Analysis, Natural Language Processing, JSON Format.

Abstract: The scoring of short-answer questions in a national-wide achievement test to public school students needs a lot of human efforts and financial expenses. Since we know that natural language processing technology can be applied to replace the manual scoring process by automatic scoring software, many researchers tried to build an automatic scoring system like c-rater and e-rater in English. In this paper, we explored a Korean automatic scoring system for short and free-text responses. NLP techniques like morphological analysis are used to build a token-based scoring template for increasing the coverage of the automatic scoring process. We performed an experiment to measure the efficiency of the automatic scoring system and it covered about 90 to 95% of the student responses with an agreement rate 95% to the manual scoring.

1 INTRODUCTION

Educational achievement test is moving from the offline handwritten platform to an online computer-based assessment (Page, 1966; Valenti, 2003). It causes an automatic scoring method being investigated to save the overall cost of the assessment (Elliot, 2003; Shermis, 2003; Dikli, 2006). Now, automated scoring is one of the interesting research topics in the field of natural language processing in which scoring process includes NLP techniques such as morphological analysis (Kang, 1994). Especially in the case of short-answer questions, lots of different answers to the same question are possible. Therefore, the resources like time, human, finance that is required for assessment can be reduced by the automated system that can handle those questions properly (Sung, 2010; Noh, 2012).

Large-scale exams like CSAT(College Scholastic Ability Test) and NAEA(National Assessment of Educational Achievement) are annual assessment tests in Korea. Automatic scoring for CSAT or NAEA should be fast and accurate. CSAT is very critical to Korean students and the result of the exam determines the entrance of a college or university. Another test TOPIK (Test of Proficiency in Korean) is a Korean language test that is offered four times a year to foreigners.

In this paper, we propose a concept-based scoring method that can evaluate complex questions. Also, we developed a Korean automated scoring system that is suitable for assessing large-scale exams. We performed an experiment on some questions with student answer sets and the result will be discussed in Chapter 4.

2 RELATED WORKS

C-rater is an automatic scoring system for English short-answer questions. The system uses predicate argument structure, pronominal reference, morphological analysis and synonyms to improve scoring accuracy. C-rater system is applied to two studies and results show that about 84% of the responses are assessed identically by the system and human graders (Sukkarieh, 2009; Leacock, 2003). But as mentioned earlier, c-rater is applicable only to English questions.

On the other hand, several methods have been tried for scoring Korean questions automatically. Chung (2009) used cosine similarity between each student response and given model answer. The system assesses responses as correct if critical keywords of the question match (Chung, 2009). Park

Jang E., Kang S., Noh E., Kim M., Sung K. and Seong T..
KASS: Korean Automatic Scoring System for Short-answer Questions.
DOI: 10.5220/0004864302260230
In *Proceedings of the 6th International Conference on Computer Supported Education* (CSEDU-2014), pages 226-230
ISBN: 978-989-758-021-5

(2003) classified open-ended questions to 4 categories; short answer, fill-blank answer, a sentence answer, many sentences answer. The scoring system uses classifying questions, exact matching, applying partial credits, critical keywords and heuristic similarity (Park, 2003).

Kang (2011) classified questions to 6 categories. The scoring system classifies questions then processes them with the proper methods for specific category. Kang's system uses morphological analysis, similarity with heuristics and score calculator (Kang, 2011). Oh (2005) and Cho (2006) used techniques from information retrieval for the scoring system. The system uses semantic kernels, vector space model and latent semantic analysis. Experiments were performed with actual exam papers and the result shows that accuracy of the system was about 80% (Oh, 2005; Cho, 2006).

3 KASS: KOREAN AUTOMATIC SCORING SYSTEM

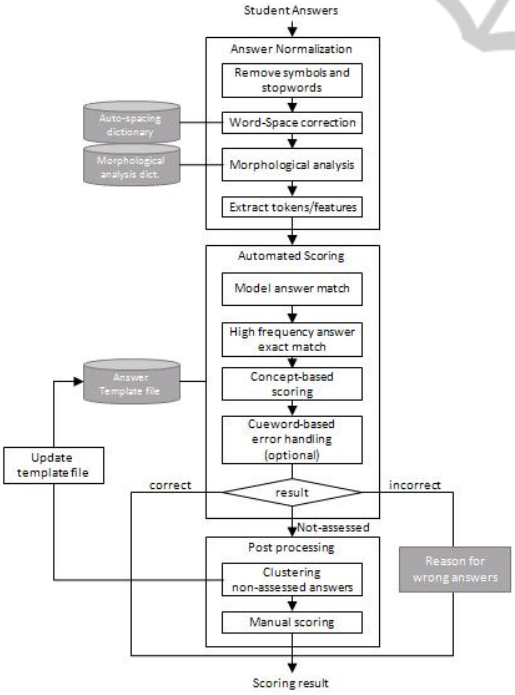

Figure 1: Structure of the proposed system.

The scoring system requires two inputs; student answers to apply automated scoring and model answers with a specific scoring guideline. Model answers are described in an answer template that is constructed by human grader and contains scoring in-formation about model answers, high-frequency student answers, concepts, cuewords and scoring options.

In order to handle automated scoring process, the system should normalize student answers with given options that are defined in an answer template. Then the system proceeds to a scoring step, which also uses an answer template. After a scoring step is finished, the system performs post processing with student answers that are not assessed in the scoring step. Figure 1 shows the structure of the automatic scoring system.

3.1 Answer Templates

Automatic scoring system does not work standalone, but it needs scoring information about a question (Park, 2012). An answer template adopted JSON (JavaScript Object Notation) format which contains detailed scoring guidelines about the question that needs to be assessed. An answer template consists of five parts and the internal structure looks like Figure 2.

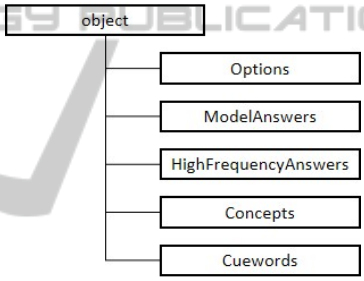

Figure 2: Structure of answer template..

Four options are given for normalize student answers; spelling correction, word-spacing correction, unnecessary symbol remove, and cueword applying option. Options can be changed and stored in the answer template file to be used at the actual scoring stage.

After the options are set, student answers are sorted in the order of frequency and a human gives the score to the high-frequency answers. Scoring information that is generated in this step is also stored in the answer template file. Apart from the "high frequency answers", human graders mark some of the student responses that is to be constructed as initial concepts. Each marked response will be generated automatically as a concept. Finally, human graders can insert some cue words of the question for treating incorrect responses, if the cueword option is set.

3.2 Student Answer Normalization

Although model answers are given for each question, student responses significantly have so many variants of the correct or incorrect answers. So, identifying the characteristics of each question is also important.

Table 1: Example of student responses and result of the normalization.

Answer types	Original text	Normalized text
Model answer	마음을치유 Heal the mind	마음을치유 Heal the mind
response 1	A: 마음을치유 A: heal the mind	마음을치유 Heal the mind
response 2	마음을치유 Healthemind	마음을치유 Heal the mind
response 3	(상처받은)마음을치유 Heal the (hurted) mind	상처받은마음을치유 Heal the hurted mind

For instance, human graders should consider spelling errors about several questions of the Korean language courses. The system should normalize every student responses including model answers with certain considerations. Three of four options are used for normalization stage, except the cueword option. Table 1 is an example of student responses and result of the normalization with all options enabled; spelling error correction, word-spacing error correction and eliminating unnecessary characters or symbols.

3.3 Automated Scoring

Once an answer template file and normalized student responses are prepared, the system is now ready to proceed to the actual scoring phase. Automated scoring process consists of four sequential steps; model answers matching, high frequency answers exact matching, concept-based assessment and cueword-based handling for incorrect responses. A cueword-based task can be skipped by the option. Each scoring step processes every student responses, and then passes the unprocessed responses to the next step.

Model answers and high frequency answers are already evaluated by human graders and normalized by the system. Since the information which are stored in the template file and student responses are already normalized, scoring process can be done with exact matching method.

After exact matching, the system proceeds to the concept-based scoring step. A concept consists of one or more tokens. There are two token types of lexical and grammatical morpheme. A Korean word is split into two tokens through morphological analysis. The system produces concepts from unprocessed student responses then tries to find a concept from the answer template in a conceptualized student response. Concept-based method can handle responses even if part of the answer matches with a concept. Figure 3 shows how the system handles student responses with concepts. The fourth element in the concept has been modified to an asterisk '*' so that any token is matched to this token.

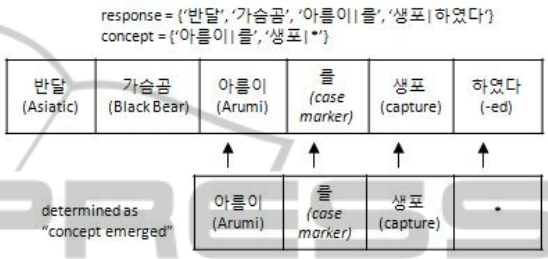

Figure 3: Example of concept matching.

Last step handles incorrect responses with cueword-based methods. Cue word is an essential keyword of the question for writing a correct response. In contrast to the previous step, each response will be treated as incorrect if the response does not include any of the cue words. As mentioned earlier, this step can be skipped by the option. If cueword option is disabled, unprocessed responses after concept-based step will be treated as non-assessed. Non-assessed responses are passed to a post processing module.

3.4 Post Processing

After automated scoring, unprocessed student responses may exist. The purpose of post processing is merging student responses. As a result, a list of merged concepts will be produced. Human graders can examine the produced list and select concepts to create.

Post processing will proceed in the following order; conceptualizing and sorting student responses then merging into concepts. First of all, every student responses are conceptualized. Then post processor sorts student responses by size because merging concept requires same sized responses. If the size of target concepts are same, the system sorts them alphabetically.

Next step is merging the concepts. The system investigates every responses of the same size and picks targets for merging. With the picked responses, the processor performs a comparison,

token by token. If there is an error of less than one or two, the different token is inserted to same position of the other response. Figure 4 shows how concept merging works, where 'Areumi' is a name of a bear. The new concepts that are generated as a result of merging the concepts will be updated in the answer template file.

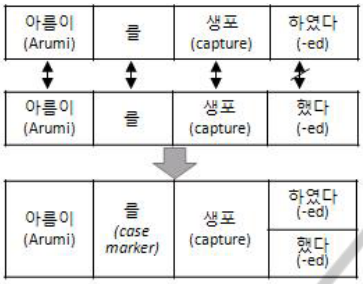

Figure 4: Example of concept merging

4 EXPERIMENTS

We performed experiments to evaluate an efficiency of the scoring system. For the experiments we used about 21,000 student responses from 2012 NAEA. We measured automatic scoring rate per each step. Results are summarized in table 2. The result in table 2 shows that about 80% of the responses are processed in model answers step. 10~15% of the rest is assessed in high frequency answers step, only 5~10% of the responses are passed to concept-based or cueword-based step.

Table 2: Result of the experiments without using cueword option.

ID_Q	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7
# of answers	3018	3013	3017	3017	3012	3023	3035
Model answers	2603	2611	2600	2771	2004	2249	1774
HighFreq answers	296	318	242	121	858	722	1229
Concept-based	66	33	93	54	30	21	23
Non-scored	53	51	82	71	120	31	9

We also compared the results between human grader and the automatic scoring system. The results are in table 3 and the average matching rate was about 95%. The result in table 3 shows that error rates of the system is very low. Note that errors might occur in experiments because the system operator is not an expert of the question.

Table 3: Comparison between KASS and human grader's scoring result.

ID_Q	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7
# of mismatch	4	3	11	4	7	1	12
Matching rate(%)	98.1	98.2	97.1	97.6	95.8	99.3	99.3

5 CONCLUSIONS

We designed and implemented an automatic scoring system for Korean short-answer questions that requires free-text answers. To evaluate the efficiency and accuracy of the system, we performed experiments with 7 questions. The results of experiments show that the system can be applied to a large-scale test like CSAT or NAEA. As a future work, we will focus on some part of the system which is relevant with concepts. Building an answer template semi-automatically will be a part of the work.

ACKNOWLEDGEMENTS

This research has been conducted through R&D program for national assessments of educational achievement funded by Korea Institute for Curriculum and Evaluation (KICE).

REFERENCES

E. B. Page, 1966. The imminence of grading essays by computer, Phi Delta Kappa, pp.238-243.

S. Valenti, F. Neri, and A. Cucchiarelli, 2003. An overview of current research on automated essay grading, *Journal of Information Technology Education*, pp.319-330.

S. Dikli, 2006. An overview of automated scoring of essays, *The journal of technology, learning, and assessment*, pp.1-35.

S. Elliot, 2003. Intellimetric: From here to validity, Automated essay scoring: A cross-disciplinary perspective, pp.71-86.

M. D. Shermis, J. C. Burstein, 2003. Automated essay scoring: A cross-disciplinary perspective, Lawrence Erlbaum Associates, Inc.

C. Leacock and M. Chodorow, 2003. C-rater: automated scoring of short-answer questions, Computers and the Humanities, vol.37, pp.389-405.

J. Sukkarieh and J. Blackmore, 2009. C-rater: automatic content scoring for short constructed response,

Proceedings of the Twenty-Second International FLAIRS Conference, pp.290-295.

T. J. Sung, K. S. Yang, T. H. Kang, and E. Y. Chung, 2010. A survey of computer-based scoring method of free-text or constructed answers in study-advancement evaluation, Korea Institute for Curriculum and Evaluation, Research Report RRE 2010-1.(in Korean).

E. H. Noh, J. H. Sim, M. H. Kim, and J. H. Kim, 2012. A prospect of the developing an automatic scoring system for Korean free-text answers in a large-scale evaluation, Korea Institute for Curriculum and Evaluation, Research Report ORM 2012-92.(in Korean).

E. M. Chung, M. S. Choi, and J. C. Shim, 2009. Design and implementation of automatic marking system for a subjectivity problem of the program, *Journal of Korea Multimedia Society*, vol.12, no.5, pp.767-776.(in Korean).

H. J. Park and W. S. Kang, 2003. Design and implementation of a subjective-type evaluation system using natural language processing technique, *The Journal of Korean association of computer education*, vol.6, no.3, pp.207-216.(in Korean).

W. S. Kang, 2011. Automatic grading system for subjective questions through analyzing question type, *The Journal of the Korea Contents Association*, vol.11, no.2, pp.13-21. (in Korean).

J. S. Oh, W. J. Cho, Y. S. Kim, and J. Y. Lee, 2005. A descriptive question marking system based on semantic kernels, *The Journal of Korean institute of information technology*, vol.3, no.4, pp.95-104. (in Korean).

W. J. Cho, 2006. An Intelligent Marking System based on Semantic Kernel and Korean WordNet, Master Thesis, Hallym University.(in Korean).

I. N. Park, E. H. Noh, J. H. Sim, M. H. Kim, and S. S. Kang, 2012. Answer template description for automatic scoring of Korean free-text or constructed answers, *Proceedings of the 24th Hangul and Korean Language Processing*, pp.138-141.(in Korean).

I. N. Park, E. H. Noh, J. H. Sim, M. H. Kim, and S. S. Kang, 2012. Concept-based automatic scoring system for Korean free-text or constructed answers, *Proceedings of the 24th Hangul and Korean Language Processing*, pp.69-72. (in Korean).

S. S. Kang and Y. T. Kim, 1994. Syllable-based model for the Korean morphology, *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, vol.1, Kyoto, Japan, pp.221-226.