An Automated Scoring Tool for Korean Short-Answer Questions Based on Semi-Supervised Learning

Min-Ah Cheon
Korea Maritime and Ocean University
minah014@outlook.com

Hyeong-Won Seo
Korea Maritime and Ocean University
wonn24@gmail.com

Jae-Hoon Kim
Korea Maritime and Ocean University
jhoon@kmou.ac.kr

Eun-Hee Noh
Korea Institute for Curriculum and Evaluation
noro@kice.re.kr

Kyung-Hee Sung
Korea Institute for Curriculum and Evaluation
Kelly9147@kice.re.kr

EunYong Lim
Korea Institute for Curriculum and Evaluation
elim@kice.re.kr

Abstract

Scoring short-answer questions has disadvantages that may take long time to grade and may be an issue on consistency in scoring. To alleviate the disadvantages, automated scoring systems are widely used in America or Europe, but, in Korea, there has been researches regarding the automated scoring. In this paper, we propose an automated scoring tool for Korean short-answer questions using a semi-supervised learning method. The answers of students are analyzed and processed through natural language processing and unmarked-answers are automatically scored by machine learning methods. Then scored answers with high reliability are added in the training corpus iteratively and incrementally. Through the pilot experiment, the proposed system is evaluated for Korean and social subjects in Programme for National Student Assessment. We have showed that the processing time and the consistency of grades are promisingly improved. Using the proposed tool, various assessment methods have got to be development before applying to school test fields.

1. Introduction

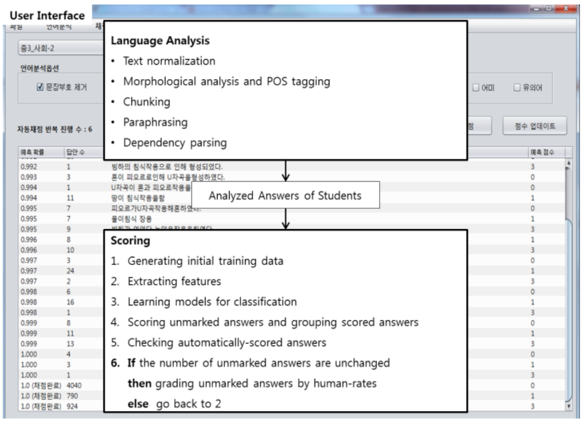
Multiple choice items can be more efficient and reliably scored than short-answer questions (Case and Swason, 2002). For this reason, questions of large-scale testing generally are multiple choice questions such as College Scholastic Ability Test (CSAT). Multiple choice questions, however, have a serious disadvantage that the limited types of knowledge, so that Korea Institute of Curriculum and Evaluation (KICE) should provide short-answer questions. The short-answer questions are difficult to score in an economical, efficient, and reliable scoring (Latifi et al., 2013). One of possible solution for such problems is using the machine learning technology of automated essay scoring (AES), e.g. Project Essay Grader (PEG) , Intelligent Essay Assessor (IEA), e-Rater and Bayesian Essay Test Scoring sYstem (BESTY) (Attali and Burstein, 2006, Shermis and Burstein, 2003).

The goal of the paper is to propose an automated scoring tool for Korean short-answer questions using semi-supervised learning. The tool consists of three components: User interface, Language analysis, Scoring. The user interface component allows users human raters interact with other components and controls them. The language analysis component analyzes and processes the answers of students through natural language processing modules like spacing normalizers, morphological analyzers, and parsers. Finally, the scoring component first grades unmarked-answers by machine learning methods and then iteratively and incrementally adds the scored answers with high reliability in the training corpus. Through the pilot experiment, the proposed system is evaluated for Korean and social subjects in Programme for National Student Assessment. We have showed that the processing time and the consistency of grades are

Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 59–63,
Beijing, China, July 31, 2015. ©2015 Association for Computational Linguistics and Asian Federation of Natural Language Processing

promisingly improved. The rest of the paper is structured as follows: Section 2 describes the proposed tool. The experiments carried out with the proposed system are discussed in Section 3. Finally, Section 4 draws conclusions and discusses future works.

2. Korean Automated Scoring Tool



User Interface

Language Analysis

- Text normalization
- Morphological analysis and POS tagging
- Chunking
- Paraphrasing
- Dependency parsing

Analyzed Answers of Students

Scoring

1. Generating initial training data
2. Extracting features
3. Learning models for classification
4. Scoring unmarked answers and grouping scored answers
5. Checking automatically-scored answers
6. **If** the number of unmarked answers are unchanged
   then grading unmarked answers by human-rates
   else go back to 2

Figure 1. The overall architecture of the proposed tool

The overall architecture of Korean automated scoring tool is given in Figure 1. The tool consists of three components: User interface, Language analysis, Scoring. The user interface component allows users as human raters interact with other components and controls them and we do not describe more details of this component because it is not important for readers to understand it. The language component and the scoring component will be described in sequent subsection in more detail.

2.1. Language analysis

As mentioned before, the language analysis component analyzes and processes the answers of students through natural language processing modules: Text normalization, Morphological analysis and POS tagging, Chunking, Paraphrasing, Dependency parsing as you can see in Figure 2. All modules in the language analysis component is implemented in Python 3.

Text normalization is composed of spacing normalization and spelling correction. Like English, Korean language uses white spaces as separators of words called Eojeol, which is a sequence of characters and represent an inflected word. Students as well as educated persons can often make spacing errors because the regulation is so flexible. The spacing normalization is performed using maximum entropy model (Berger et al., 1996). The spelling correction is implemented using Levenshtein distance algorithm. The morphological analyzer is implemented using the modified CYK algorithm (Kim, 1983) and the pre-analyzed data. The POS tagging is to find the longest path on the weighted network (Kim, 1998). The weighted network is made of a lattice structure constructed by using the morphological analysis results, contextual probability, and lexical probability. The chunker is based on the maximum entropy model and a chunking dictionary. The paraphrasing replaces consecutive words or phrases with representative words or phrases. We perform a small scale of paraphrasing, for example, synonyms, endings, and particles. The purpose of the paraphrasing is twofold. First, it helps to alleviate data sparseness of dependency parsing. Second, it increases the accuracy of automated scoring. The dependency parsing finds direct syntactic relationships between words by connecting head-modifier pair into a tree structure and is implemented by the MaltParser (Niver, 2008). Actually we use just dependency relations as one of features, described in the next subsection, but not the tree structure.

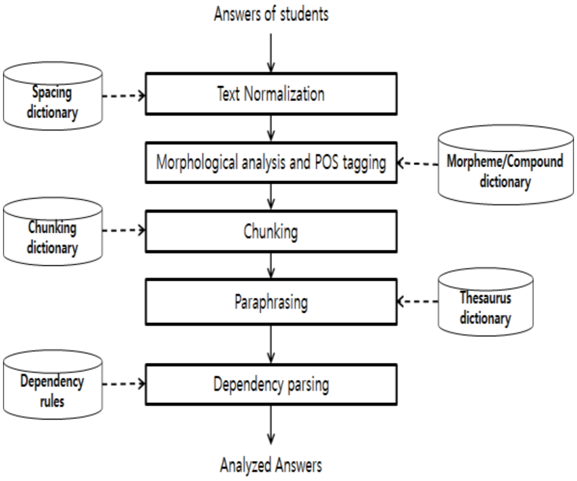
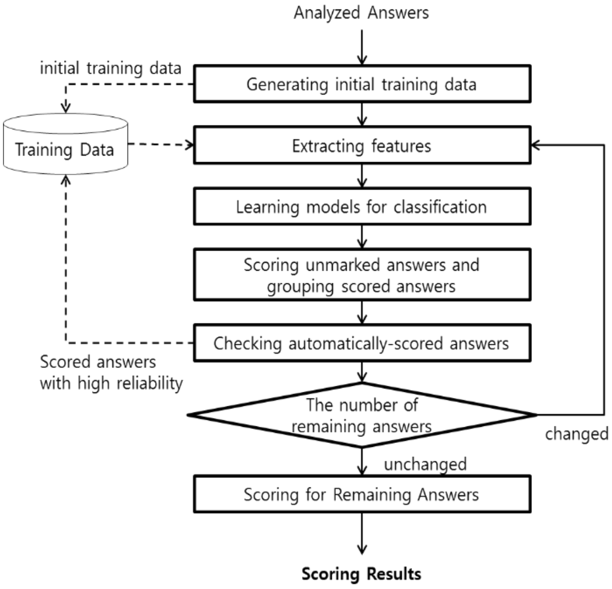


Figure 2. The processing order in the language analysis component

2.2. Scoring

The scoring component first grades unmarked-answers by machine learning methods and then iteratively and incrementally adds the scored answers with high reliability in the training corpus. The process order in the scoring component is shown in Figure 3.

The scoring component is based on a semi-supervised learning (Chapelle et al., 2006), which is halfway between supervised learning and unsupervised learning. It uses a small amount of labeled data and a large amount of unlabeled data. Actually, a grade in scoring can be considered a label in automated scoring. In other words, automated scoring classifies grades as labels from students' answers. The scoring component comprises six steps described in the follows.



Analyzed Answers
initial training data
Generating initial training data
Training Data
Extracting features
Learning models for classification
Scoring unmarked answers and grouping scored answers
Checking automatically-scored answers
Scored answers with high reliability
The number of remaining answers
changed
unchanged
Scoring for Remaining Answers
Scoring Results

Figure 3. The processing order in the scoring component

The first step is to generate initial training data by the human raters who grade high frequency answers as many as they want. The graded answers will be the initial training data.

The second is to extract features for machine learning. We use word features, syntactic features, and dependency relation features. A word feature is a content word, a syntactic feature comprises a content word and a syntactic relation like Subj and Obj. A dependency relation feature is composed of a triplet of a dependent, a governor of features consists of TF-IDF which is widely used in information retrieval.

The third step is to generate learning model for classification. We use two classification models: Logistic regression model and k-NN (k-Nearest Neighbors) model. The logistic regression model is used to classify answers as well as to get the probability of classification. The k-NN model is used to increase the reliability of classification by comparing the result with that of logistic regression classifier.

The fourth step is to grade unmarked answers and to group the scored answers. We classify grades of unmarked answers using the two learning models. If the two results are same and if the predictive probability as the regression probability is greater than a threshold, the scored answers are considered as correct scoring results which are candidates added in the training corpus. The threshold is arbitrarily set by human-raters (default is 0.99) through the user interface and is automatically decreased by 0.03 during iteration. The interval value can also be determined through the user interface. Each group of scored answers has the same probability and is showed as one row on the user interface in order that it is easy to check whether the scored answer is correct.

The fifth step is check whether the automatically-scored answers are correct. The Human-raters have to confirm the results. If there is some wrong results, the human raters should correct them or put back them into unmarked answers. After that, the confirmed results are added to the training data. The system repeats the second step to the fifth step until the number of unmarked answers is unchanged. Repeating this process can increase the amount of training data, thus both reliability and accuracy of automated scoring are increased.

Finally the sixth step is to manually grade still-unmarked answers by human-raters.

3. Pilot Experiments

3.1. Experimental setting

We have evaluated the proposed tool on the short-answer questions which are selected from "Programme for National Student Assessment (KICE, 2013)". The eleven items are from subjects such as Korean and social. The number of students' answers in each item is 1000. All the answers are composed only one sentence.

The correct answers as gold standards are graded by experts throughout three rounds. The round defines as grading the same problem by two experts in subjects. If scored results of the two experts are different, other experts perform the round again. The round is repeated by three times.

We use Pearson's correlation coefficient (Corey, 1998), Cohen's Kappa coefficient (Carletta, 1996; Fleiss, 2003) and an accuracy which generally used from information retrieval. For example, interpreting any kappa value can be considered as follows: $\kappa < 0.4$ (poor), $0.4 \leq \kappa < 0.75$ (fair to good), and $0.75 \leq \kappa$ (excellent).

Table 1. Results of Evaluation

-		**Pearson's correlation coefficient (*r*)**		**Kappa correlation coefficient (κ)**		**Accuracy (%)**	
subject	**Item no.**	**H-G**	**S-G**	**H-G**	**S-G**	**H-G**	**S-G**
Korean: Middle school	2-(1)	0.96	0.82	0.90	*0.80*	98.6	97.3
	2-(2)	0.97	0.93	0.91	0.87	97.5	96.1
	4-(2)	0.97	0.93	0.93	0.81	96.9	92.0
Korean: High school	2-(1)	0.99	1.00	0.99	1.00	99.5	100.0
	2-(2)	0.98	0.87	0.98	0.87	99.5	96.3
	4-(1)	0.99	0.88	0.97	0.83	98.6	91.5
	5-(2)	0.99	0.93	0.99	0.88	99.1	92.3
	6-(1)	0.98	0.94	0.98	0.94	98.9	97.2
	6-(2)	1.00	0.90	0.98	0.84	98.9	92.4
Social: Middle school	4-(3)	0.86	0.95	0.85	0.95	96.8	99.0
	8	1.00	0.92	0.99	0.93	99.8	97.8
Average (standard derivation)		**0.97 (0.04)**	**0.92 (0.05)**	**0.95 (0.05)**	**0.88 (0.06)**	**98.6 (1.04)**	**95.6 (3.05)**

As another example, interpreting r can be considered as follows: r ≤ 0.2 (very small), 0.2 < r ≤ 0.4 (small), 0.4 < r ≤ 0.6 (medium), and 0.6 < r ≤ 0.8 (large), r ≤ 0.8 (very large).

3.2. Experiment Results

Table 1 shows performance evaluation results of the proposed tool. In the Table 1, H-G stands for human-rater and gold standard and S-G for our system and gold standard.

The average of Pearson's correlation coefficient between results of our system and gold standards (S-G) is 0.92. It means a strong positive linear relationship between the automated scores as results of our system and the gold standard scores, therefore it can be mostly similar to our automatic grading and gold standards. The average of Kappa correlation coefficient is 0.88, so results of our system are broadly same like standard scores. The accuracy of the answer that contains negative expressions and the inversion of word order is relatively low as compared to other answers. According to report of KICE (Noh et al., 2014), this system can save significant time and cost in comparison with scoring methods of human raters.

4. Conclusion

We have presented an automated scoring tool for Korean short-answer questions based on semi-supervised learning. The tools use several NLP technologies for analyzing answers of students, and some machine learning methods of logistic regression and k-NN algorithm for automated scoring. The scoring process is iterative and incremental under the semi-supervised learning. The experimental results show that the proposed automated scoring tool is very promising in automated scoring for the short-answer questions.

In future work, we will be going to study a method for increasing the accuracy of our automated tool and to find a way to minimize the intervention of the human-raters.

Acknowledgements

This work was partly supported by KICE (Korea Institute of Curriculum & Evaluation) and the IT R&D program of MSIP/KEIT. [10041807, Development of Original Software Technology for Automatic Speech Translation with Performance 90% for Tour/International Event focused on Multilingual Expansibility and based on Knowledge Learning]

Reference

Y. Attali and J. Burstein. 2006. Automated Essay Scoring with E-rater. *The Journal of Technology, Learning, and Assessment*, 4(3):12-15.

A. Berger, S. A. Della Pietra, and V. J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39-71.

J. Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249-254.

S. M. Case and D. B. Swason. 2002. Constructing Written Test Questions for the Basic and Clinical Sciences. *National board of Medical Examiners*.

O. Chapelle, B. Schölkopf,and A. Zien. 2006. *Semi-Supervised Learning*. The MIT Press Cambridge, Massachusetts London, England, pages 1-3.

D. M. Corey, W. P. Dunlap and M. J. Burke. 1998. Averaging Correlations: Expected Values and Bias in Combined Pearson rs and Fisher's z Transformations. *J. Gen. Psychol.*, 125:245-261.

J. L. Fleiss, B. Levin, M.C. Paik (Eds.). 2003. Statistical methods for rates and propositions 3rd. *John Wiley & Sons, Inc.*, pages 598-626.

KICE. 2013. *Programme for National Student Assessment*. Korean Institute of Curriculum & Evaluation.

J. Kim. 1998. Korean Part-of-Speech Tagging using a Weighted Network. *Journal of the Korea Information Science Society (B): Software and Applications*, 25(6):951-959.

S. Kim. 1987. *A Morphological Analyzer for Korean Language with Tabular Parsing Method and Connectivity Information*. MS Thesis, Department of Computer Science, Korea Advanced Institute of Science and Technology, pages 21-37.

S. M. F. Latifi, Q. Guo, M. J. Gierl, A. Mousavi, K. Fung. 2013. Towards Automated Scoring using Open-source Technologies. In *Proceedings of the 2013 Annual Meeting of the Canadian Social for the Study of Education Victoria, British Columbia*, pages 1-27.

J. Nivre, 2008. Algorithms for Deterministic Incremental Dependency Parsing. *Computational Linguistics*, 34(4):513-553.

E. Noh, S. Lee, E. Lim, K. Sung and S. Park, 2014. Development of Automatic Scoring System for Korean Short-answer question and Verification of Practicality. Korean Institute of Curriculum & Evaluation, page 87-120.

M. D. Shermis and J. Burstein. 2003. *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum Associates, Cambridge, England, MIT Press.