

Classify and Examine: Using Semi-supervised Learning with Ensemble Learning to Score Short-Answer Questions in Korean

Min-Ah Cheon¹, Chang-Hyun Kim², Jae-Hoon Kim¹
Korea Maritime and Ocean University¹, Korea
Electronics and Telecommunications Research Institute², Korea
minah0218@kmou.ac.kr, chkim@etri.re.kr, jhoon@kmou.ac.kr

Eunhee Noh³, Mi-Young Song³, Jong-Im Park³, Yuhyang Kim³, Do-Gil Lee⁴
Korea Institute of Curriculum & Evaluation³, Korea University², Korea
{noro, mysong, pji0310, kyh15}@kice.re.kr, motdg@korea.ac.kr

Abstract

We propose a sentence-level Korean automated scoring tool using a semi-supervised learning with ensemble learning to ensure scoring consistency and to save labeling efforts for training a scoring classification. The tool consists of two phases: language processing and scoring processing. Through experiments, we have shown that the correlation coefficient between the results of the tool and those of human raters are 0.941 in the strong positive range. As a result, our proposed tool is practical and useful.

1 Introduction

Short-answer questions or essay questions differ from multiple-choice questions in that require students to write their correct answers. They are suitable for measuring critical thinking, problem-solving, and creativity (KSEE, 2004; Kim, 2009; Jin, 2007). However, it is difficult to guarantee the reliability and fairness of scoring results (Kim, 2009; Jin, 2007) since they are scored depending on the subjective judgment of human raters. Thus, many countries use automated scoring systems (Attali and Burstein, 2005; Shermis and Burstein, 2003; Runder et al., 2006; ETS, 2010) and most of them utilize machine learning techniques. Linguistically, Korean and English fall into different categories (Kang, 2002). Thus, an automated scoring tool developed in English cannot be applied to Korean. Therefore, it is necessary to develop the automated scoring tool for Korean.

This paper introduces a sentences-level Korean Automated Scoring Tool (KASTSL) using semi-supervised learning with ensemble learning algorithm currently under development. In addition, we demonstrate practicality of KASTSL by comparing scoring results of KASTSL and human raters. The rest of this paper is structured as follows: Section 2 describes KASTSL in detail. Performance evaluation through some experiments is described in Section 3 and Section 4 draws conclusions and discusses potential future work.

2 KASTSL: sentences-level Korean Automated Scoring Tool

Figure 1 shows the overall structure of KASTSL. KASTSL consists of a language processing phase and a scoring phase. In Figure 1, The solid arrow represents the work flow and the dashed arrow represents the flow of data. The input of the proposed tool is students’ answers and the output is the score of each answer.
The input of the language processing phase is students’ answers. In this phase, a lot of NLP modules such as text normalization, morphological analysis & part of speech tagging, negative expression recognition, chunking, paraphrasing, and dependency parsing are required. All modules in the language processing are implemented in Python 3. The implementation of each module is described in detail in (Noh et al., 2014; Noh et al., 2015; Noh et al., 2016). Once language processing is complete, completely identical answers become a group by themselves.

In the scoring phase, a human rater repeats the follows: First, human rater scores students’ answers in the top several groups. These scoring results and criteria answers are used as initial training data. Second, features are extracted from the training data to generate an ensemble classifier model consist of logistic regression, random forest, and SGD. The features are words, Eojis, and dependency obtained from the results of each NLP module in language processing phase. The generated classifier is used to assign the most appropriate points to unscored answers. After that, a human rater confirms them by density peak clustering (Rodrigues & Laio, 2014) of the automated scoring answers. The human rater confirms the answers as training data for the next scoring step.

3 Experimental and Verification

3.1 Datasets
The short-answer questions used in the evaluation were selected from the Korean, Social and Science subjects of (Korea Institute for Curriculum & Evaluation, 2014; Korea Institute for Curriculum & Evaluation, 2015); information on each test question is summarized in Section A.

3.2 Evaluation measures
The criteria score is assumed to be error-free since it has been fully verified by subject experts. We evaluated the performance of KASTSL by comparing the criteria score with KASTSL’s results. The evaluation measures are accuracy, precision, recall, and F1 score (Manning et al., 2008), and correlation coefficient (Corey et al., 1998; Fleiss et al., 2003). An additional description of each measure is given in Section A.

3.3 Results
Figure 2 briefly shows performance evaluation of KASTSL. # Sentence(s) is the number of answers for each question. A baseline system is a prototype of KASTSL (Noh et al., 2014; Noh et al., 2015; Noh et al., 2016). The results of one sentence show similar performance. However, our tool performs better than the baseline for more than two sentences.

![Figure 2: Performances evaluation](image)

For all evaluation questions Pearson’s coefficient and Kappa coefficient were in the range of [0.742, 0.95]. This means a very strong positive correlation for two sentences or less. In other words, it means that the scores scored using the KASTSL and the criteria score are almost the same.

4 Conclusion
We have presented KASTSL which is an automated scoring tool for Korean short-answer questions based on semi-supervised learning with ensemble learning. The KASTSL used several NLP modules and ensemble learning algorithms of logistic regression, random forest and SGD for automated scoring. By adding density peak clustering to the scoring phase of KASTSL, we improved user friendliness compared to the baseline system. The experimental results show that the KASTSL is not only useful for one sentence but more than two sentences for automated scoring. However, there is a limit to the existence of performance deviations depending on the characteristics of the question.

In future work, we will study how to improve the performance of KASTSL while minimizing its performance variation of according to the question characteristics.
Acknowledgments

This work was partly supported by the ICT R&D program of MSIP/ITIP [R7119-16-1001, Core technology development of the real-time simultaneous speech translation based on knowledge enhancement] and [Advanced Refinements and Application of Automated Scoring System for Korean Large-Scale Assessment].

References


A Supplementary Material

Subsection 3.1 Datasets

The short-answer questions used in the evaluation were selected from the Korean, Social and Science subjects of (Korea Institute for Curriculum & Evaluation, 2014; Korea Institute for Curriculum & Evaluation, 2015); information on each test question is summarized in Table A. The initial training answers were generated by scoring one blank answer group and 10 high-frequency answer groups. Answer unit is the number of sentences in the criteria answer. Test year is the year in which the exam was taken. Subject and Question number are the question information of the middle school and high school subjects. Criteria answer is the scoring criteria. Score category is the point information including the partial score of questions. # Of students’ answers is the student’s number of answers.

Subsection 3.2 Evaluation measures

The criteria score is assumed to be error-free since it has been fully verified by subject experts. We evaluated the performance of KASTSL by comparing the standard score with KASTSL results. The evaluation criteria are accuracy, precision, recall, F1 score and correlation coefficients. Accuracy, precision, recall, and F1 score are often used as the performance measures of the general classification system (Manning et al., 2008). However, these methods are not suitable for KASTSL performance evaluation, because the KASTSL includes the initial training data and manual scoring. Therefore, we use precision, recall and accuracy, as defined in (Noh et al., 2014; Noh et al., 2015; Noh et al., 2016). The definitions of precision, recall and accuracy used in this experiment are shown in Table B. \(a\) is the number of correctly answered answers. \(b\) is the number of incorrectly scored answers. \(c\) is the number of answers correctly hand-scored. \(d\) is the number of errors with hand-scored answers. \(e\) is the number of initial training answers correctly scored. \(f\) is the number of incorrectly scored answers in the initial training data. If there is no error in the initial training answer, \(f = 0\).

<table>
<thead>
<tr>
<th>Automated scoring target</th>
<th>Automated scoring</th>
<th>Manual scoring</th>
<th>Manual scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a)</td>
<td>(c)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(b)</td>
<td>(d)</td>
<td></td>
</tr>
</tbody>
</table>

Table B: Classification of answers to evaluate the performance evaluation of KASTSL

\[
\text{precision} = \frac{a}{a + b} \\
\text{recall} = \frac{a}{a + b + c + d} \\
\text{accuracy} = \frac{a + c + e}{a + b + c + d + e + f} \\
F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

<table>
<thead>
<tr>
<th>Category</th>
<th>Test year</th>
<th>Subject</th>
<th>Question number</th>
<th>Criteria answer</th>
<th>Score category</th>
<th># of students’ answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>One sentence</td>
<td>2014</td>
<td>Korean (High)</td>
<td>6(1)</td>
<td>우리는 자연에 손을대야 한다고 생각한다.</td>
<td>0, 1</td>
<td>7,965</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Korean (Middle)</td>
<td>6(2)</td>
<td>야생 동물과 친구가 되려면 무엇을 할 수 있을지를 생각해본다.</td>
<td>0, 1</td>
<td>7,453</td>
</tr>
<tr>
<td></td>
<td>2015</td>
<td>Korean (High)</td>
<td>3</td>
<td>독서 역량과 수준의 갖는 끈을 선택해야겠다.</td>
<td>0, 1, 2</td>
<td>7,501</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Social (Middle)</td>
<td>9</td>
<td>근본 원칙은 사수하고 근본 원칙들은 감소합니다.</td>
<td>0, 1, 2</td>
<td>7,430</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Science (Middle)</td>
<td>6(2)</td>
<td>꿈의 원을 이루지 않는 동물의 깊이는 적절하다.</td>
<td>0, 1, 2</td>
<td>7,428</td>
</tr>
<tr>
<td>Two sentences</td>
<td></td>
<td>Korean (High)</td>
<td>4(2)</td>
<td>(서 영감을) 예절을 그만두라고 한다. (화 시발로) 예절이 좋아지고 싶다.</td>
<td>0, 1, 2</td>
<td>7,501</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Korean (Middle)</td>
<td>6</td>
<td>현재 방과과학 기본 순서입니다 (수업 내역의 수업에 대한 연속도가 낮습니다. (그러므로) 수업별로 다른 내용 개설해 주세요. (그러면) 방과과학 기본 순서와 수업 참여도가 높아질 것입니다.</td>
<td>0, 1, 2, 3</td>
<td>7,432</td>
</tr>
<tr>
<td>Three sentences</td>
<td></td>
<td>Korean (Middle)</td>
<td>6</td>
<td>현재 방과과학 기본 순서입니다 (수업 내역의 수업에 대한 연속도가 낮습니다. (그러므로) 수업별로 다른 내용 개설해 주세요. (그러면) 방과과학 기본 순서와 수업 참여도가 높아질 것입니다.</td>
<td>0, 1, 2, 3</td>
<td>7,432</td>
</tr>
</tbody>
</table>

Table A: Summary information on the short-answer questions used in this experiment.