AN EMPIRICAL ANALYSIS OF BERT EMBEDDING FOR AUTOMATED OPEN ENDED RESPONSE MATHEMATICS QUESTIONS SCORING

HueyMin Wu[1], YuChing Lin[2], & HungSheng Lin[3]

[1]Graduate Institute of Educational Information and Measurement,
National Taichung University of Education (Taiwan)
[3]YuLin County Qishan Elementary School (Taiwan)

Abstract

With the development of technology, the number of users of online tests is increasing. The online test platform's multiple choice and closed question types have automatic scoring functions. The automatic scoring of online tests allows students to get instant feedback on their answers and take advantage of information technology. However, in the open-ended response questions, most of the platform questions have not yet had an automatic scoring mechanism. The scoring still has to rely on experts in the professional field for manual grading, and the correction process is time-consuming. Currently, the research on automatic scoring is mainly in the essay, and automatic scoring is performed through Latent Semantic Analysis (LSA). Since the answers to open-ended response questions in mathematics often contain specific mathematical symbols in addition to text, semantic analysis is difficult to achieve good results. The contribution of this paper is to establish the automatic scoring of mathematical open-ended response questions through deep learning Bidirectional Encoder Representation from Transformers (BERT) and solve the dilemma of mathematical open-ended response questions.

The data set is taken from Taiwan's digital learning platform, mainly used by elementary school students and junior middle school students, including four-theme mathematical construction response questions: algebra (1755 datasets), space and shape (412 datasets), data and uncertainty (1518 datasets), and number and quantity (1435 datasets). All open-ended response questions were scored by human raters using holistic 3-point scoring rubrics. This research uses Colaboratory as the development environment of the automatic scoring system. It is a development platform for virtual servers provided by Google. It is an editing and execution software running on the cloud, which allows software developers to quickly edit and execute Python code directly through the browser. In order to explore the effectiveness of BERT applied to the automatic scoring of open-ended response questions in mathematics, this study chose the application of the common automatic scoring method, LSA, as the benchmark for comparison, and to explore the application of LSA and BERT to automatic scoring of open-ended questions in mathematics. The consistency between the automatic scoring results and human raters' scoring will be presented in this research. In order to determine the accuracy of the model, 5-fold cross-validation is used to divide the datasets into the training set and testing set for model training and testing.

This study uses Exact Accuracy Rate and F1-score as evaluation indicators. The exact accuracy rate refers to the score results of each level of automatic scoring and must be completely consistent with the scoring results of each level of human rater scoring in the total number of all test questions. The F1 score is calculated based on precision rates and recall rates. The F1-score ranges between 0 and 1. If the F1-score is close to 1, the model is better. The results are presented in the table below. The results show that BERT was better than LSA in both the accuracy and F1 score performance of automatic scoring of open-ended questions in mathematics. In practice, BERT can be used to automatically score open-ended mathematics questions and provide immediate feedback.

theme	Exact Accuracy Rate		F1- score	
	LSA	BERT	LSA	BERT
algebra	73.41%	89.77%	.6326	.8436
space and shape	84.00%	90.99%	.6786	.8521
data and uncertainty	76.18%	89.37%	.6700	.8246
number and quantity	75.56%	91.39%	.7427	.8987

Keywords: *Automated Scoring (AS), Latent Semantic Analysis (LSA), Mathematics, Natural language processing (NLP), Open-constructed response.*