Using Machine Learning to Detect Student Learning Levels along a Learning Progression

Abstract: We introduced a novel method of using Natural Language Processing and Machine Learning to detect student learning levels using student short responses to a question measuring a learning progression of functions. Accuracy rate of our model was at 86%. Our precision, recall, and F1 scores ranged from .77 to .92.

Introduction

Learning progressions (LP) (Corcoran et al., 2008) can serve as a foundation to develop personalized assessments to support student learning (Heritage, 2008). One key challenge in working with LPs is to figure out how to identify student learning levels (Wilson, 2012). Several psychometric methods have been proposed to identify student learning levels along a LP (e.g., Author et al., 2021; Wilson et al., ; Shin et al., 2017). However, classification into levels by different psychometric methods using student item scores can be inconsistent (e.g., Author, 2019). In this study, we offer a new approach using natural languague processing (NLP) and machine learning (ML) to detect student learning levels from similated student responses to an open-ended question developed to measure a LP of function in middle-school mathematics (Author et al., 2021).

Method

Data generation

Within the LP of function, there are two consecutive learning levels which are (i) lower level: students believe that all functions are linear and thus their graph is a straight line, (ii) higher level: students understand that the slope of a function can change, thus graph of such function can be curvy. In a previous study, a set of one multiple-choice item (item 1) and an open-ended question were developed to measure the LP of function as shown in Figure 1 below.

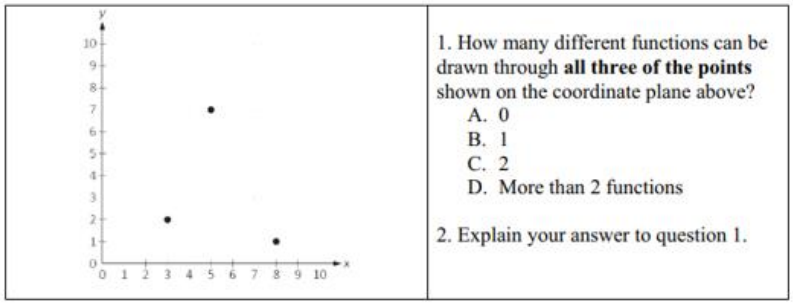

Figure 1: Items to Measure the Two Learning Levels

In this study, we used ChatGPT (OpenAI, 2023) to generate short student responses to question 2 conditioned on their selection for the first item. We adopted the following script to ask ChatGPT to generate responses:

> Question: “How many different functions can be drawn through all three of the points that are not on the same line shown on a cordinate plane?”
>
> Answer: “n” (n=0,1,2, more than 2 functions)

"Give 20 different explanation of you are a 5th grade student."

To mimic the number of upper elementary students in a typical elementary school, we ran three rounds of data generation for each answer. All together, we generated 240 responses. Since there were four duplicated responses, we filtered them out and kept 236 responses in our analyses. Due to the data generation procedure, we should expect there are around 75% of the explanations were incorrect, and 25% of them received a full credit.

Learning Level Annotation

To prepare the training data, we annotated the generated responses using the following label identifiers:

- 0 or lower learning level: students answered the question incorrectly or offered incorrect explanation to their correct answer
- 1 or higher learning level: students answered the question correctly and offer corrrect explanation

One of the authors had a master's degree in mathematics read through the 236 responses and assigned these labels to each of them. The data was then used to build model as follows.

Model Building and Evaluation

We use the BERT transformer in Python to generate NLP features which were then fed into TextClassification to predict the label of 0 or 1 described above (Devlin et al., 2018). Due to the quite small sample size, we used 5-fold cross-validation to evaluate the classification (Bishop & Nasrabadi, 2006). We relied on common metrics of accuracy, precision, recall, and F1 score (Cui, 2021) to evaluate our model.

Results

Data Annotation

Out of the 236 responses, 169 accounting for 71.6% of them were assigned a label of 0. The remaining 67 responses acounting for 28.4% of the responses received a label of 1.

Human-Machine Interater Reliability

Table 1 below shows our confusion matrix. Based on the results, our accuracy rate was at 86%.

Table 1. Confusion Matrix for Human and Machine Scores

	Human Score of 0	**Human Score of 1**	**Row Total**
Machine Score of 0	151	18	169
Machine Score of 1	14	53	67
Column Total	165	71	

Table 2 contains the precision, recall, and F1 scores for each point.

Table 2. Precision, Recall, and F1 Scores by Score Point

Score Point	Precision	Recall	F1
0 (lower level)	.92	.89	.90
1 (higher level)	.77	.79	.77

Conclusions and Discussion

Our overall accuracy of 86% indicated that existing tool such as the BERT transformer and TextClassification can help us identify effectively student learning levels using their short constructed responses. The precision, recall, and F1 scores for the score point of 0 were around 90% which is higher than these of the score point of 1. The result suggested that the model we proposed can correctly detect more than 90% of students with lower level of understanding. Compared to results reported in the literature, our accuracy, precision, recall, and F1 scores outperformed many studies on short answer scoring and meaning detection (e.g., Burstein et al., 2013; Cui, 2021; Foltz et al., 2013).

References

Authors (2019). [Title omitted for blind review]. *[Journal Name Omitted For Blind Review]*.

Authors (2021). [Title omitted for blind review]. *[Journal Name Omitted For Blind Review]*.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Briggs, D. C., & Peck, F. A. (2015). Using learning progressions to design vertical scales that support coherent inferences about student growth. *Measurement: Interdisciplinary Research and Perspectives*, *13*(2), 75-99.

Burstein, J., Tetreault, J., Chodorow, M., Blanchard, D., & Andreyev, S. (2013). 16 Automated Evaluation of Discourse Coherence Quality in Essay Writing. *Handbook of automated essay evaluation: Current applications and new directions*, 267-280.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (Report No. 63). New York, NY: Center on Continuous Instructional Improvement, Teachers College - Columbia University.

Cui, Z, (2021). Machine learning and small data. *Educational Measurement: Issues and Practices*, *40*(4),8–12. https://doi.org/10.1111/emip.12472

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the intelligent essay assessor. *Handbook of automated essay evaluation*: *Current applications and new directions*, 68-88.

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: Chief Council of State School Officers (CCSSO).

OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat

Shin, H. J., Wilson, M., & Choi, I. H. (2017). Structured constructs models based on change-point analysis. *Journal of Educational Measurement*, 54(*3*), 306–332. doi:10.1111/jedm.12146

Wilson, M. R. (2012). Responding to a challenge that learning progressions pose to measurement practice. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 317–343). Rotterdam, The Netherlands: Sense Publishers.