An Automatic Self-explanation Sample Answer Generation with Knowledge Components in a Math Quiz

Ryosuke Nakamoto[1(✉)], Brendan Flanagan[2], Yiling Dai[2], Kyosuke Takami[2], and Hiroaki Ogata[2]

[1] Graduate School of Informatics, Kyoto University, Kyoto, Japan
s0527225@gmail.com
[2] Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan

Abstract. Little research has addressed how systems can use the learning process of self-explanation to provide scaffolding or feedback. Here, we propose a model automatically generating sample self-explanations with knowledge components required to solve a math quiz. The proposed model contains three steps: vectorization, clustering, and extraction. In an experiment using 1434 self-explanation answers from 25 quizzes, we found 72% of the quizzes generated sample answers with all necessary knowledge components. The similarity between human-created and machine-generated sentences was 0.719, with a significant correlation of R = 0.48 for the best performing generation model by BERTScore. These results suggest that our model can generate sample answers with the necessary key knowledge components and be further improved by using the BERTScore.

Keywords: Self-explanation · Rubric · Automatic summarization · NLP

1 Introduction

Self-explanation is defined as generating explanations to oneself and explaining concepts, procedures, and solutions to deepen understanding of the material [1]. It has been widely recognized for its learning effects for a long time [2]. The iSTART system is the leading research method in self-explanation evaluations, which guides learners through the exercise to support active reading and thinking [3].

In mathematics, there is a procedure for solving a quiz, and the quiz is solved according to that procedure. Therefore, we proposed a method to check whether students can describe each step in a self-explanation by comparing the similarity between the human-created sample answer and students' self-explanations [4]. It was judged that the student's knowledge was likely to be insufficient because the information and words of the unit required were included or, if not, they were missing some knowledge components. We defined "Rubric" as can-do descriptors that clearly describe all the essential knowledge components of the quiz and "Sample Answer" as model answers of self-explanations with knowledge components, which are prepared according to the step rubric number(Table 1). In this study, we propose an automatic generating sample answers model

M. M. Rodrigo et al. (Eds.): AIED 2022, LNCS 13356, pp. 254–258, 2022.
https://doi.org/10.1007/978-3-031-11647-6_46

in place of human-created sample answers. Our contributions have a wide range of implications, such as scoring self-explanations and generating self-explanation scaffold templates based on sample sentences.

Table 1. Rubrics and a sample answer of self-explanation in a quiz.

Number	Rubric	Sample Answer of Self-explanations
Step 1	Be able to find the equation of a linear function from two points	Substituting the y-coordinate of p into the equation of the line AC
Step 2	Be able to find the equation of the line that bisects the area of a triangle	Find the area of triangle ABC, and then find the area of triangle OPC
Step 3	Be able to represent a point on a straight line using letters	Since the coordinates of P are (3, 5/2), the line OP is y = 5/6, and Q are (t, 5/6)

2 Data Collections and Model Architecture

We collected the data from January 1, 2020, to December 31, 2021, using the LEAF platform [5], which consists of a digital reading system named BookRoll, and a learning analytics tool LAViEW (Fig. 1). For this experiment, we chose quizzes with at least five answers. The number of quizzes were 25, and the total number of answers were 1434. Figure 2 illustrates the proposed model, which consists of (i) Vectorizing component, (ii) Clustering component, and (iii) Extracting Component. As the vectorizing component, we adopted Sentence BERT and BERT Japanese pre-trained model to represent the sentences [6, 7]. As the clustering component, we employed an unsupervised learning model, K-means. The reason for generating meaning-intensive clusters through unsupervised learning is to reproduce the solution steps in mathematics. From an educational point of view, a problem for junior high school students would probably contain at least two steps and at most six steps of unit knowledge components and set the number of clusters in the range of 3–5 by the elbow method. As the extracting component, for each semantic cluster, the most representative sentences are extracted and sorted by multiplying them by their position in the problem, obtained from pen strokes. For extracting a representative sentence, Lexrank [8] was tested to extract the most representative sentences from each cluster. The input is all the self-explanation sentences associated with the quiz, and the output is the summarization with knowledge components for the quiz.

3 Experiments

Firstly, we set the rubrics for each quiz for evaluation (Table 2). Secondly, two authors and one assistant evaluated the machine-generated self-explanations to determine if they contained the necessary knowledge components. Though the Fleiss' kappa coefficient [9] was 0.518 initially, after discussing the differences among the three, the final coefficient

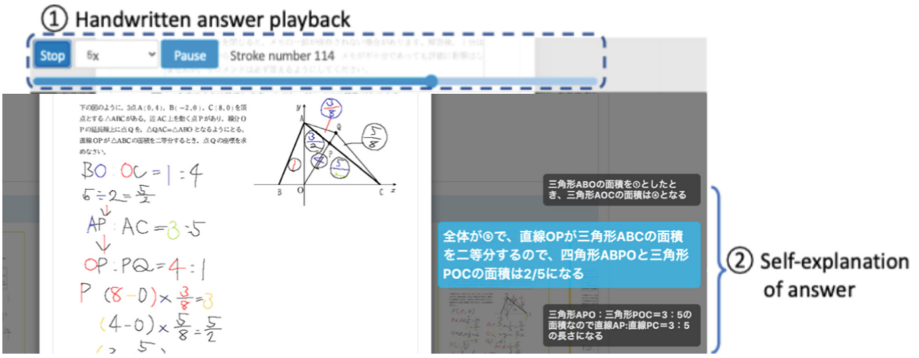

Fig. 1. The students input a sentence of explanation every time they think they have completed some step in their answers during the playback. Therefore, the self-explanations are temporally associated with the pen stroke data.

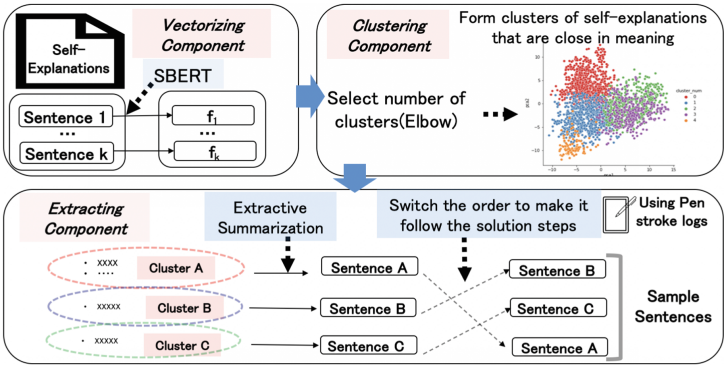

Fig. 2. Overall model architecture.

was 0.870. Table 2 shows the human evaluation results in 72% of the quizzes, it could generate all of the maximum five knowledge components.

Next, we evaluated the similarity of human-created and machine-generated sentences from several metrics: BERTScore, BLEU [10, 11]. In addition, we conducted a Spearman correlation analysis to investigate the correlations between the summary index and human evaluation. The Human Evaluation Score (HES) was scored according to how well machine-generated answers met the knowledge components against rubrics in the following form.

Table 3 presents the F1 Metrics scores. The highest similarity metric was BERTScore with an average of 0.719. Table 4 shows the correlations and RMSE between HES and metrics. As for correlation, it was 0.48 for BERTScore, showing a moderate correlation. As for RMSE, the BERTScore with the minor error was 0.273, while the other metrics were over 0.5, a significant difference.

Table 2. Missing knowledge components of each quiz by Human evaluation

Missing knowledge components	0	1	>=2
Num of quizzes	18	4	3
Probability density	0.72	0.16	0.12

Table 3. The similarity evaluation(F1)

BERTScore		BLEU	
M	SD	M	SD
0.719	0.032	0.300	0.093

Table 4. RMSE and Correlations between HES and metrics.

	BERTScore	BLEU
Correlations	0.48**	0.46**
RMSE	0.273	0.582

Note. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

4 Conclusion

This study attempted to generate sample self-explanation sentences from collected data. The collected 1434 self-explanations from 25 quizzes were fed into a model and the results showed that 72% of the quizzes could generate all of the maximum five knowledge components. The similarity between human-created and machine-generated sentences was 0.715, with a significant correlation of R = 0.48(BERTScore). Results suggest it is possible to generate sample answers using the proposed model to extract the necessary knowledge components and improving the BERTScore accuracy correlates with extracting essential knowledge components.

Acknowledgments. This work was partly supported by JSPS Grant-in-Aid for Scientific Research 20H01722, 21K19824, and NEDO JPNP20006, JPNP18013.

References

1. Rittle-Johnson, B.: Promoting transfer: effects of self-explanation and direct instruction. Child Dev. **77**(1), 1–15 (2006)
2. Bisra, K., Liu, Q., Nesbit, J.C., Salimi, F., Winne, P.H.: Inducing self-explanation: a meta-analysis. Educ. Psychol. Rev. **30**(3), 703–725 (2018). https://doi.org/10.1007/s10648-018-9434-x

3. McNamara, D.S., Levinstein, I.B., Boonthum, C.: iSTART: interactive strategy training for active reading and thinking. Beh. Res. Methods, Inst. Comput. **36**(2), 222–233 (2004)
4. Nakamoto, R., Flanagan, B., Takam K., Dai Y., Ogata, H.: Identifying students' stuck points using self-explanations and pen stroke data in a mathematics quiz. In: ICCE 2021, 2021.11.22–26 (2021)
5. Flanagan, B., Ogata, H.: Learning analytics platform in higher education in Japan. Knowl. Manage. E-Learn. (KM&EL) **10**(4), 469–484 (2018)
6. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-Networks, arXiv preprint arXiv:1908.10084 (2019)
7. Suzuki, M.: Pretrained Japanese BERT models, GitHub repository. https://github.com/cl-tohoku/bert-japanese. Accessed 10 Aug 2020
8. Erkan, G., Radev, D.: LexRank: graph-based lexical centrality as salience in text summarization. arXiv:1109.2128 (2004)
9. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**, 378–382 (1971)
10. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
11. Papineni, K., Roukos, S., Ward, T. & Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002). Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135(2002)