2021 International Conference on Advanced Learning Technologies (ICALT) | 978-1-6654-4106-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICALT52272.2021.00108

Automatic Question Generation for Repeated Testing to Improve Student Learning Outcome

Danny C.L. Tsai
Department of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
dan860202@gmail.com

Anna Y.Q. Huang
Department of Computer Science
and Information Engineering
National Central University
Taoyuan, Taiwan
anna.yuqing@gmail.com

Owen H.T. Lu
College of Computer Science

National Pingtung University
Pingtung, Taiwan
owen.lu.academic@gmail.com

Stephen J.H. Yang
Department of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
stephen.yang.ac@gmail.com

***Abstract*—In recent years, educational resources have gradually been digitized, and digital education platforms have gradually become popular. We use AI to accurately assist people in performing daily tasks through a machine learning process. In education, we can use AI in many situations, such as predicting student's learning outcome and discovering student's learning strategies. However, most solutions have not yet utilized modern AI capabilities, such as natural language processing. This research aims to help teachers use machines to automatically generate short answer questions to reduce the time for teachers to write exam questions. In addition, the main reason we focus on short answers is that many studies prove that short answer exercises can enhance student's long-term memory, thereby improving their learning performance. We propose an automatic question generation (AQG) system that combines syntax-base and semantics-base, in order to prove that the system is highly available and improve student's learning performance, we conducted experiments with 41 students. The experimental results show that student's learning performance has been significantly improved, which means that by repeatedly testing the machine question generation system, students can deepen their long-term memory of course knowledge.**

Keywords-automatic question generation; learning outcome; repeated testing

I. Introduction

In programming courses, proficiency is usually achieved through homework, tests and a lot of practice. Unfortunately, the process of creating exams and quizzes is usually time-consuming and error-prone. Therefore, we use short answer AQG technology to solve creating a large number of exercises and questions, these questions can be used for assignments or quizzes in programming courses. The main purpose of question generation is to use it as an assessment tool to help teachers quickly and effectively understand the learning situation of students. Zavala and Mendoza (2018) [1] report usage in the classroom environment, where the generator was used to generate quizzes for multiple courses and generate assignments for students. As mentioned in [2, 3], To reduce the expenses associated with manual construction of questions and to satisfy the need for a continuous supply of new questions, AQG techniques were introduced. Finally, summarize the current trends and advances in AQG, highlight the changes that the area has undergone in the recent years, and suggest areas for improvement and future opportunities for AQG.

In the field of education, using short answer quizzes can gain various benefits: (1) let students construct knowledge through short answer questions, (2) identify incorrect concepts through learner feedback, (3) repeated testing important concepts to improve memory, (4) let teachers understand the learning situation of each student [2]. Funk and Dickson (2011) [4] found that in laboratory research, students usually spend more time studying for short-answer exams than for multiple-choice exams. In addition, if students expect to recall information to answer questions, then their scores are better than those that rely on cognition to answer questions.

As mentioned above, we know that AI can bring many benefits to the education field, and in recent years, AI has developed significantly in the field of natural language processing. Therefore, this research uses machine to automatically generate short-answer questions and provides students with conceptual knowledge review after class. Finally, the research questions of the study are defined as follows:

- RQ1: How to measure the quality of automatic short answer question generation?
- RQ2: Have student's engagement of repeated testing effect on learning outcomes?

II. Literature reviews

A. Automatic question generation

Yao, Bouma and Zhang (2012) [5] divide the methods of question generation into three types: (1) syntax-based, (2) semantic-based and (3) template-based. The traditional AQG method is mainly to extract parts of speech or parse trees in sentences, and then use rules to convert the sentence from a declarative sentence to an interrogative sentence. However, the main disadvantage of this method is that it lacks flexibility and relies heavily on pre-established rules, but it is still an important basis for question generation. With the development of deep learning, there are many neural network models for question generation. In the study of [6], use attention mechanism seq-2-seq architecture to achieve question generation, and [7] use the transformer architecture for question generation. On the other hand, pre-training models are also useful tools for question generation, such as BERT and T5 language models, or GPT-3 language generation models, these models can produce fairly fluent natural language. Therefore, in this research, we propose a method that combines syntax-based and semantic-based methods to automatically question generation.

978-1-6654-4106-3/21/$31.00 ©2021 IEEE
DOI 10.1109/ICALT52272.2021.00108

B. How to benchmarking the automatic question generation

In the field of natural language processing, most studies use standard data sets and automatic evaluation indicators to measure the quality of the question [2], the most popular data set is SquAD (Sandford question answering dataset), and the most popular evaluation indicators are BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation). We take the questions in the SQuAD dataset as basic facts, and use BLEU and ROUGE to calculate the n-gram similarity between the reference sentence and the generated sentence.

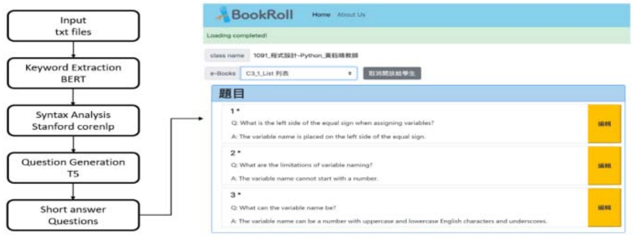
III. Methodology

A. Participants and data sets

This research conducted an experiment to explore whether a system that automatically generates questions by repeatedly practicing short-answer questions can improve student learning outcomes, the experiment started in September 2020 and lasted for one semester. The subjects were students in the first-year programming course of the university. A total of 41 students participated in the experiment. The course was mainly to teach the basic Python programming language.

B. Automatic question generation system

This research proposes a method that combines syntax-based and semantic-based methods to automatically question generation, Fig. 1 shows the system architecture and user interface designed in this study. This interface allows teachers to view and modify machine generated questions, and allows students to answer questions. The main purpose of semantic analysis is to enable machines to understand the content of textbooks and extract important keywords from its. We use BERT to extract keywords from textbooks, and use a large amount of unlabeled data to train the model through unsupervised learning and transfer learning methods, using transfer learning to train a language model can bring many benefits: (1) no need to mark training data, (2) the language model can understand the grammatical structure, interpret the semantics, and even use the attention mechanism for coreference resolution, researchers only need to fine-tune the model to complete many downstream NLP tasks, thereby improving training efficiency. In addition to semantic analysis, the purpose of syntax analysis is to extract sentences that contain keywords. We use the Stanford CoreNLP tool developed by Stanford University to perform syntactic analysis on the extracted sentences. The result will be displayed in the parse tree to find the complete sentence containing the subject, verb and target.



Loading completed!

1 *

Q: What is the left side of the equal sign when assigning variables?

A: The variable name is placed on the left side of the equal sign.

2 *

Q: What are the limitations of variable naming?

A: The variable name cannot start with a number.

3 *

Q: What can the variable name be?

A: The variable name can be a number with uppercase and lowercase English characters and underscores.

Figure 1. Data flow and screenshot of the AQG system

Since syntax analysis can only output declarative sentences containing keywords, we use T5 to convert declarative sentences into interrogative sentences, T5 is a text-to-text framework that combines encoder and decoder transformer architecture. It treats each NLP question as "text-to-text", which means using text as input and generating new text as output. T5 model uses the Colossal Clean Crawled Corpus (C4) data set for pre-training. C4 data set is a new open source pre-training data set created by Google, which collects about 750GB of data. In addition, the number of parameters of T5 is much higher than that of BERT and GPT-2, with up to 11 billion parameters, in contrast, although BERT has 340 million parameters and GPT-2 has 1.5 billion parameters, it is still much less than T5. Finally, we input descriptive sentences containing keywords into the T5 model, and use the T5 model to transform the descriptive sentences into interrogative sentences beginning with WH questions.

C. Evaluation the quality of machine-question

Many previous studies have used automatic evaluation indicators to evaluate the performance of the problem generation model [2], for comparison, we also use Bleu and ROUGE-L to evaluate question generation methods. BLEU is an evaluation indicator based on accuracy, usually used as an evaluation tool for machine translation or question generation, BLEU uses N-gram matching rules to calculate the similarity between the generated sentence and the reference sentence. The score is between 0 to 1. If the score is closer to 1, indicates that the question generation quality is better, the calculation of BLEU as follows:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp(\textstyle\sum_{n=1}^{N} w_n \log P_n) \tag{1}$$

ROUGE is an evaluation indicator proposed in the context of text summarization. Four methods are proposed in [12]: (1) ROUGE-N, (2) ROUGE-L, (3) ROUGE-W and (4) ROUGE-S. Generally, most studies use ROUGE-L, L represents the longest common subsequence (LCS), it is based on the F-score of the LCS between the generated sentence and the reference sentence, a common subsequence means that a group of words appear in the same order in two sequences, the difference with n-gram is that the common subsequence can be a combination of discrete words. For example, the generated sentence is: "He went to school", the reference sentence is: "He will go to school", and the longest common subsequence is: "He to school". The calculation of ROUGE-L as follows:

$$R_{lcs} = \frac{LCS(X,Y)}{m} \tag{2}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \tag{3}$$

$$F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{4}$$

Among them, LCS(X, Y) represents the length of the longest common subsequence between the generated sentence and the reference sentence, m represents the length of the reference sentence, n represents the length of the generated sentence, β is a user-defined parameter. If β is larger, it means that F score will focus more on R_{lcs} instead of P_{lcs}.

D. *Student's engagement of AQG system*

To improve student's learning outcome, the instructor ask students to review learning concepts after class through AQG system. There are five learning concepts of C_1, C_2, C_3, C_4, C_5 in this course. To measure students engagement of review behavior after class, we used the equation (5) to calculate the student's practice situation in AQG system. For student S_i, the symbol $S_{RT}(\mathrm{i})$ indicates the engagement score of using AQG system. The symbol N_k represents the number of automatically generated questions for the learning concept C_k. The symbol N_k^i indicates the number of questions with correctly answered of learning concept C_k for students S_i.

$$S_{RT}(\mathrm{i}) = \sum_{k=1}^{5} \frac{C_k^i}{N_k} \tag{5}$$

E. *Student's learning outcome*

This study applied four scores to comprehensively measure the learning performance of students in all learning concepts. The S_{t1} and S_{t3} scores focus on measure student's learning performance through short answer questions and program coding, respectively. The S_{t2} and S_{t4} are the score of final exam and semester scores respectively.

IV. RESULT AND DISCUSSION

A. *Quality of automatic question generation*

For RQ1, we take the questions in the SQuAD dataset as basic facts, and use automatic evaluation indicators to calculate the n-gram similarity between the reference sentence and the generated sentence, according to the results, the score of BLEU is 0.567, which is not satisfactory, mainly due to the following three reasons: (1) in the process of machine question generation, the subject in the question will be converted into a pronoun form, while in the SQuAD data set, it is represented by the subject, (2) the quality of the generation questions may be reduced due to the combination of singular and plural nouns, verb tenses and prepositions, (3) in the process of machine question generation, if the key sentence of the textbook is too short, the model will not be able to understand the core point of the key sentence when generating the question, thereby reducing the quality of the generated question. On the other hand, the score of ROUGE-L is 0.613, which means that the machine-generated questions have a certain degree of similarity to the questions of the SQuAD dataset, it also means that our AQG system can generate questions of a certain quality. Finally, edit the generated questions through the user interface in Figure 1 to improve the above problems and improve the quality of the generated questions.

B. *Correlation analysis*

For RQ2, we performed Pearson correlation analysis, as shown at TABLE I. . The results show that by repeated testing the AQG system positively related to S_{RT} in S_{t1}, S_{t2}, S_{t3} and S_{t4}. students can enhance their understanding of classroom knowledge concepts. We also found that more serious students who use the AQG system will greatly improve their semester grades, which is also significantly related to their learning outcome.

TABLE I. STATISTICS RESULTS AND PEARSON CORRELATION OF FOUR TEST

test	N	Mean	S.D.	Pearson correlation	
				coefficient	p-value
S_{t1}	41	89.34	10.90	.435	.005**
S_{t2}	41	73.80	19.98	.403	.009**
S_{t3}	41	70.05	21.76	.320	.041*
S_{t4}	41	90.22	11.45	.840	.000**

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

V. CONCLUSION

This study focuses on natural language processing, aims to implement AI applications for the education purpose, and look forward to the benefits of emerging AI technologies that can bring into education. After a semester of experiment, we verified two expectations through data analysis, (1) using syntax-based and semantic-based question generation methods, it has good performance in terms of the quality of the generated questions, (2) repeated testing are beneficial to student's long-term memory, even if the exercises are generated by machines. Finally, there are two limitations to this study, (1) the quality of the teaching materials and the format setting will affect the machine generated question's quality, (2) in generating question types, we only use WH questions. In the future, we will try other types of questions, such as cloze or multiple-choice questions, and we will also improve the problems that cause the BLEU score to drop.

ACKNOWLEDGMENT

This work is supported by Ministry of Science and Technology, Taiwan under grants MOST-109-2511-H-008-007-MY3, MOST-108-2511-H-008-009-MY3, MOST-110-2511-H-153-001, and Ministry of Education, Taiwan.

REFERENCES

[1] L. Zavala and B. Mendoza, "On the use of semantic-based aig to automatically generate programming exercises," in Proceedings of the 49th ACM Technical Symposium on Computer Science Education, 2018, pp. 14-19.

[2] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," International Journal of Artificial Intelligence in Education, vol. 30, no. 1, pp. 121-204, 2020.

[3] S. J. Yang, H. Ogata, T. Matsui, and N.-S. Chen, "Human-centered artificial intelligence in education: Seeing the invisible through the visible," Computers and Education: Artificial Intelligence, vol. 2, p. 100008, 2021.

[4] S. C. Funk and K. L. Dickson, "Multiple-choice and short-answer exam performance in a college classroom," Teaching of Psychology, vol. 38, no. 4, pp. 273-277, 2011.

[5] X. Yao, G. Bouma, and Y. Zhang, "Semantics-based question generation and implementation," Dialogue & Discourse, vol. 3, no. 2, pp. 11-42, 2012.

[6] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," arXiv preprint arXiv:1705.00106, 2017.

[7] K. Kriangchaivech and A. Wangperawong, "Question generation by transformers," arXiv preprint arXiv:1909.05017, 2019.