Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

An Intelligent Evaluation Framework for Personalized Learning

Albert YANG[a]* & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*yang.ming.35e@st.kyoto-u.ac.jp

Abstract: Testing has been demonstrated to be an effective strategy to promote retention of knowledge and improve students' self-regulated learning skills. However, the integration of retrieval practice into an actual curriculum still remains challenges. In this paper, we propose an intelligent evaluation framework to address the tasks of question generation and test item selection to facilitate repeated testing. The results of this research can be used to motivate more instructors to involve testing in their teaching approaches.

Keywords: Repeated testing, question generation, adaptive testing, personalized learning

1. Introduction

Traditionally, testing is used to assess students' knowledge and assign grades. However, its employment to facilitate learning is an application of testing that has been largely neglected by educationalists. Empirical studies have emphasized that testing can improve learners' retention of learned (Roediger & Karpicke, 2006) and new (Pastötter & Bäuml, 2014) information and self-regulated learning (SRL) skills (Fernandez & Jamet, 2017).

Despite the strong evidence for the efficacy of retrieval practice, the integration of retrieval practice into an actual curriculum presents many challenges. One of the challenges that discourages educators from implementing retrieval practice is creating repeated tests for large numbers of learners. Recent researchers have tried to address this challenge by automatically generating questions from a given text using artificial intelligence and natural language processing techniques. The question generation (QG) has been mainly addressed with two types of methods. One is based on heuristic rules, which uses manually constructed templates to create questions and ranks the generated results. However, the rule-based method relies heavily on human effort, which makes it difficult to scale up and be generalized in various fields. Another method that is increasingly used by researchers is to use sequence-to-sequence or encoder-decoder frameworks to train end-to-end neural networks which has been shown to significantly outperformed the state-of-the-art rule-based system (Du et al. 2017). Although the performance of the existing QG model and the quality of the generated results have been greatly improved, there are still few studies on the usefulness of QG in the education field.

Another challenge that educators face is deciding what to test. Everything within a curriculum cannot be tested - especially on a repeated basis. Given this fact, tests that cover the most concepts in the materials or fit learners' knowledge state should be prioritized. One approach to address this challenge is adaptive assessment. Instead of asking the same questions to every students, adaptive assessment aims to select the next question based on their previous answers. Students can, therefore, concentrate on the knowledge they lack and not feel bored during the assessment. Most of the research, however, only focused on previous responses when selecting the next questions, while students' behaviors in each test has received less attention. Taking testing behavior into account may allow us to find questions that are worth testing again. For example, recommending questions that have been answered correctly but not tested for a period of time may help students update their memory of earlier knowledge.

Finally, educators might consider that standard course examination is enough for evaluating students' retention of knowledge. However, the standard course examinations used in most cases largely measure students' cramming ability rather than durable learning. As a result, it is important to design activities that encourage students to review the earlier elements of the curriculum for long-term retention.

2. Proposed Research Work

This study plans to resolve the challenges of integrating retrieval practice into an actual curriculum by proposing an intelligent evaluation framework built on top of an e-book reading system, BookRoll, for personalized learning. BookRoll is an e-book reading system (Flanagan & Ogata, 2017) developed by Kyoto University; instructors can upload materials, and students can use the e-book reader to read the content and interact with the text using the provided tools, such as notes and highlights. Figure 1 is the conceptual framework. The framework consist of three modules: text summarization module, question generation module, and quiz recommendation module. When instructors upload a learning material to BookRoll, the text summarization module extracts sentences from the e-book as the essential knowledge. Next, the question generation module converts the extracted sentences into questions in different formats. Then, the quiz recommendation module provides students with personalized quizzes on the basis of their performance and behaviors in previous attempts.

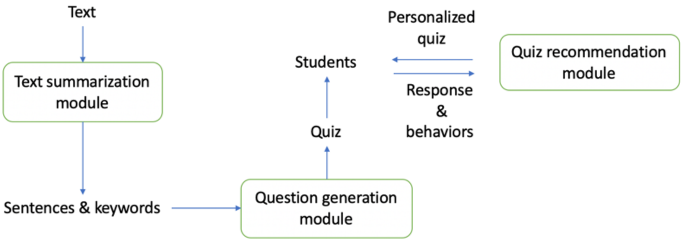
Text
Text summarization module
Sentences & keywords
Question generation module
Quiz
Students
Personalized quiz
Response & behaviors
Quiz recommendation module

Figure 1. The proposed evaluation framework

The proposed framework can be used to evaluate students' various skills. Since students' various reading actions are logged in BookRoll database, the key knowledge extracted by the text summarization can be used to assess students' text marking skills, which is an important reading skill that has been shown to be positively correlated with academic performance (Junco & Clem, 2015). In addition, by analyzing students' quiz-taking behaviors, we can evaluate their SRL skills. In this research, we aim to answer the following questions:

1. Can machines extract concepts that are approximate to the key concepts extracted by humans for marker grading?
2. Can students improve their reading skills, reading engagement, and reading comprehension with machine-generated quizzes?
3. Will different behavioral patterns in quizzes affect student learning performance?
4. Can student improve their reading skills, reading engagement, and learning performance with personalized quizzes?

3. Contribution

The proposed framework aims to resolve the current challenges of integrating retrieval practice into actual curriculums. First, our text summarization module and question generation module can extract important sentences from various types of materials and generate questions for retrieval practice, which saves the time and effort for the instructors to create the quizzes. Second, the quiz recommendation module enables students to skip questions they are already familiar with and answer questions that are worth testing again, which helps them assess their knowledge more efficiently and fully benefit from repeated testing. Finally, the proposed framework can be used as formative assessment to repeatedly assess students' knowledge. Instructors can measure students' retention of knowledge by their performance in the low- or no-stake quizzes. In sum, this research proposes a framework to motivate educationalists to involve testing in their teaching approaches and provide a more continuous approach to evaluate students' long-term retention of knowledge.

4. Methodology

4.1 Apply Text Summarization Techniques to Extract Key Concepts

In order to generate questions that best represent the knowledge in the learning materials, we need to find a model that can effectively and accurately obtain the knowledge in the text. To achieve this, we compare three text summarization models for automatically extracting key concepts from learning materials, namely TextRank (Mihalcea & Tarau, 2004), RAKE (Rose et al., 2010), and BERT (Devlin et al., 2018). We use the learning materials on BookRoll as the input text for the models. Figure 2 shows the text summarization process. We convert the PDF learning materials into plain text files and perform preprocessing techniques, such as removing special characters and converting text to lowercase. Then, the preprocessed text is passed to the models as input, and a sentence list will be generated as the essential knowledge in the text. Finally, we use the markers provided by the humans as the reference answer to evaluate the quality of summaries by the machine using BLEU 1, BLEU 2, BLEU 3, BLEU 4 (Papineni et al., 2002), and METEOR (Denkowski & Lavie, 2014) scores.

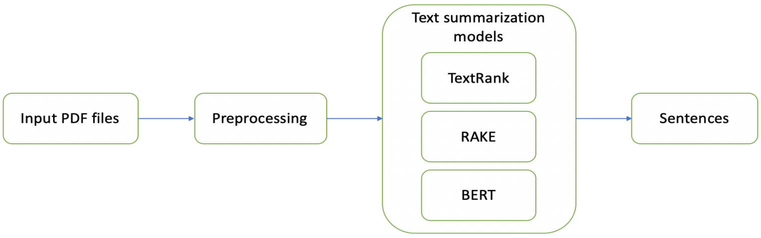

Figure 2. Text summarization process

4.2 Automatic Question Generation for Repeated Testing

Next, with the ability to extract key knowledge from text, we propose a QG module which automatically generates two types of questions for testing the extracted knowledge. Figure 3 demonstrates the question generation process. We use sentences extracted by the text summarization module as candidate for question generation. Then, we perform syntax analysis to filter out incomplete sentences. For example, a sentence composed of subject and verb is a complete sentence, otherwise it is incomplete. Next, two models are used to generate cloze items and short answer question. We apply TextRank to select keywords from each sentence and mask them to generate the cloze items, and use GPT-2 (Radford et al., 2019) to convert the factual sentences to interrogative sentences to generate short answer questions.

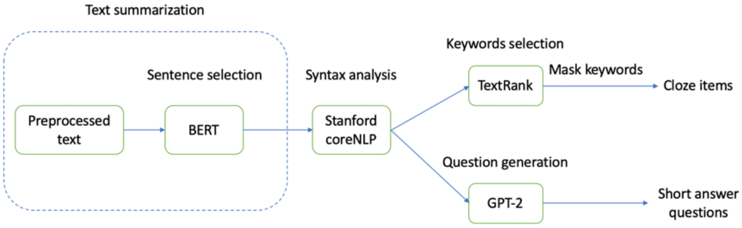

Figure 3. Question generation process

4.3 Quiz Recommendation for Personalized Learning

We propose a module to adaptively recommend quiz to students. Figure 4 shows the adaptive testing process. When students complete a quiz, their responses and the testing behaviors will be logged in the database. Then, the recommendation module recommends the next quiz to students according to their past performance and behaviors. The module first uses adaptive testing approaches (such as item response theory or cognitive diagnosis models) to evaluate students' knowledge state to select questions that suit their current knowledge state. Next, the module considers students' testing behaviors, such as response time, number of attempts for each quiz, and the interval between each quiz to select questions that are worth testing again. Finally, when students complete the recommended quiz, the module will update their knowledge state and testing behaviors, and recommend the next quiz.

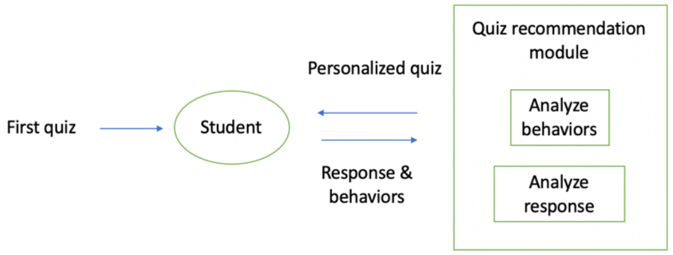

Figure 4. The adaptive testing process

Acknowledgments

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) 20H01722, JSPS Grant-in-Aid for Scientific Research (S) 16H06304 and NEDO JPNP20006 and JPNP18013.

References

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.*

Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the ninth workshop on statistical machine translation (pp. 376-380). Baltimore, Maryland USA: Association for Computational Linguistics.

Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. Retrieved from https://arxiv.org/abs/1705.00106

Fernandez, J., & Jamet, E. (2017). Extending the testing effect to self-regulated learning. Metacognition and Learning, 12(2), 131-156.

Flanagan, B., & Ogata, H. (2017, November). Integration of learning analytics research and production systems while protecting privacy. In The 25th International Conference on Computers in Education, (pp. 333-338). Christchurch, New Zealand: ResearchGate.

Junco, R., & Clem, C. (2015). Predicting course outcomes with digital textbook usage data. The Internet and Higher Education, 27, 54-63.

Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404-411). Barcelona, Spain: Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318). Philadelphia, USA: Association for Computational Linguistics.

Pastötter, B., & Bäuml, K. H. T. (2014). Retrieval practice enhances new learning: the forward effect of testing. Frontiers in psychology, 5, 286.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological science, 17(3), 249-255.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory, 1,* 1-20.