Learning Association Between Learning Objectives and Key Concepts to Generate Pedagogically Valuable Questions

Machi Shimmei(✉) and Noboru Matsuda

North Carolina State University, Raleigh, NC 27695, USA
{mshimme,noboru.matsuda}@ncsu.edu

Abstract. It has been shown that answering questions contributes to students learning effectively. However, generating questions is an expensive task and requires a lot of effort. Although there has been research reported on the automation of question generation in the literature of Natural Language Processing, these technologies do not necessarily generate questions that are useful for educational purposes. To fill this gap, we propose QUADL, a method for generating questions that are aligned with a given learning objective. The learning objective reflects the skill or concept that students need to learn. The QUADL method first identifies a key concept, if any, in a given sentence that has a strong connection with the given learning objective. It then converts the given sentence into a question for which the predicted key concept becomes the answer. The results from the survey using Amazon Mechanical Turk suggest that the QUADL method can be a step towards generating questions that effectively contribute to students' learning.

Keywords: Question generation · MOOCS · Learning engineering

1 Introduction

Creating high-quality questions is important for instructors as valid questions provide insight into their students' learning status, which in turn helps instructors enhance their teaching methods. Answering questions is also an essential part of learning. The benefit of answering questions for learning has been shown in many studies, aka *test-enhanced learning* [1, 2]. On Massive Open Online Course (MOOC), questions are also an influential component that determines the effectiveness of the course. It is reported that students learn better when they practice skills by answering questions than when only watching videos or reading text [3]. However, creating questions that effectively help students' learning requires experience and extensive efforts.

When the question is generated for educational use in particular, with the focus on test-enhanced learning, machine-generated questions should have a pedagogical value in addition to general features such as clarity and fluency. Although there are a number of studies on question generation in the field of AI in education [4, 5], little has been studied about the pedagogical value of the generated questions. To fill this gap, we propose a

I. Roll et al. (Eds.): AIED 2021, LNAI 12749, pp. 320–324, 2021.
https://doi.org/10.1007/978-3-030-78270-2_57

method for generating questions that supposedly ask about the key concepts the students need to learn to attain the learning objectives. There have been no studies that aim to generate questions that align with the learning objectives.

2 Related Work

Recent works on question generation take a data-driven approach using neural networks. Large datasets such as SQuAD [6], NewsQA [7], MSMARCO [8] enabled training a recurrent neural network (RNN) for question generation. The number of studies with the aim of generating questions specifically for educational purposes has been also increasing. The limited number of relevant datasets available is among the primary challenges in educational question generation. Although there are datasets such as SciQ [9], which contains questions from science textbooks, the size of the data is considerably small. Therefore, some studies utilize general question generation datasets to train a model. Wang *et al.* [10] demonstrated that an LSTM-based model, called QG-Net, trained on a SQuAD can be used for generating questions on educational contents.

Another challenge for question generation is how to identify an answer candidate. QG-Net and other models [11–13] require that an input sentence is tagged with a candidate of an answer for the generated question. There are also some models that can find an answer candidate in a given text. For example, Willis *et al.* [14] proposed a key phrase extraction model that outputs an answer candidate from a given paragraph text. QUADL also has the Answer Prediction model that finds an answer candidate (i.e., a target token index). The key difference of our Answer Prediction model from the existing models is that *our proposed Answer Prediction model aims to select target tokens that are aligned with a given learning objective.*

3 Methods

Figure 1 shows an overview of QUADL. Given a pair of a learning objective *LO* and a sentence *S*, $<LO, S>$, QUADL generates a question *Q* that will be suitable to achieve the learning objective *LO*. The question *Q* is a verbatim question, which means that the answer can be literally found in the given sentence *S*. The following is an example of $<LO, S>$ and *Q*:

Learning objective (*LO*): Describe metabolic pathways as stepwise chemical transformations either requiring or releasing energy; and recognize conserved themes in these pathways.

Sentence (*S*): Among the main pathways of the cell are photosynthesis and cellular respiration, although there are a variety of alternative pathways such as fermentation.

Question (*Q*): Along with photosynthesis, what are the main pathways of the cell?

Answer: Cellular respiration.

Notice that the answer is tagged (underlined in *S*) in the sentence *S*. For the sake of explanation, we call the tagged token(s) in the given sentence S as a *target token* hereafter.

QUADL consists of two components: (1) the Answer Prediction model and (2) the Question Conversion model. The Answer Prediction model identifies <*Is,Ie*>, called *token index*, where *Is* and *Ie* show the index of the start and end of a target token within a given sentence *S* relative to the learning objective *LO*. We adopted BERT, Bidirectional Encoder Representation from Transformers [15] for this Answer Prediction model. The learning objective and sentence were combined as a single input <*LO, S*> to the model. The final hidden state of the BERT model was fed to the single layer classification model that outputs logit for the start index (*Is*) and another single layer classification model that outputs logit for the end index (*Ie*) for each token in the sentence *S*. The final score was calculated by taking the softmax of the sum of the start logit and end logit for every possible span (*Is* < *Ie*) in the sentence. The score was also calculated for < *Is* = *0, Ie* = *0* > indicating that the sentence is not suitable to generate a question for the learning objective. The index <*Is*, *Ie*> with the largest score became the final prediction. For the rest of the paper, we call sentences that have non-zero indices (i.e., *Is* ≠ 0 *and Ie* ≠ 0) the *target sentences*, whereas others are referred to as the *non-target sentences* (i.e., has the zero token index <0, 0>). We created training data for the Answer Prediction model using the text data from existing online courses at Open Learning Initiative[1] (OLI).

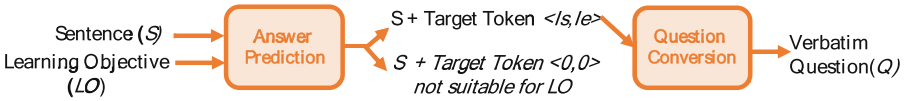

Fig. 1. The QUADL model

Given a sentence with the non-zero target token index, the Question Conversion model generates a question for which the target token becomes the answer. We use an existing bidirectional-LSTM seq2seq model with attention and copy mechanisms, QG-Net [10], for the Question Conversion model. We used an existing *pre-trained* QG-Net model that was trained using SQuAD datasets[2]. We could train the QG-Net using the OLI course data. However, the OLI courses we used for the current study do not contain a sufficient number of verbatim questions—many of the questions are fill-in-the-blank and multiple-choice questions hence not suitable to generate training data for QG-Net.

4 Evaluation

We have the following research questions: **RQ1**:How well does the Answer Prediction model identify target tokens (including zero token indices) in a given sentence relative to a given learning objective? **RQ2**:How well does the pre-trained QG-Net generate questions for a given sentence tagged with the target tokens? To answer the questions,

[1] https://oli.cmu.edu.

[2] https://rajpurkar.github.io/SQuAD-explorer/.

we conducted a survey on Amazon Mechanical Turk (AMT). In AMT, for each triplet $<LO, S < Is, Ie>, Q>$ shown, the participants were asked if they agreed or disagreed with the following two statements: (1) To get a question that helps attain the learning objective *LO,* it is adequate to convert the sentence *S* into a question whose answer is the token $<Is, Ie>$ highlighted. (2) The question *Q* is suitable for attaining the learning objectives *LO*. Each statement corresponds to each research question.

Table 1 summarizes the results for RQ1. The table shows that, for the predictions with a non-zero target index, 70% (123/178) of the predictions *were accepted*. As for the non-target sentence predictions (i.e., the Answer Prediction model output the zero <0,0> index), only 26% (43/164) were accepted. That is, 55% (90/164) of the predicted non-target sentences were considered to be target sentences by participants.

Table 1. The evaluation of the predicted target tokens by the Answer Prediction model. There were 178 sentences that the Answer Prediction model predicted target tokens (non-zero index) and 164 sentences that the model predicted non-target (zero index $<0, 0>$). The table shows how many of them were accepted/not accepted by the majority vote by Amazon Mechanical Turk (AMT) participants.

AMT	Model prediction			
		Non-zero target index $<Is \neq 0, Ie \neq 0>$	Zero-index $<0, 0>$	Total
	Accepted	**123 (70%)**	**43 (26%)**	166 (49%)
	Tie	32 (18%)	25(15%)	57 (17%)
	Not accepted	22 (12%)	90 (55%)	112 (33%)
	Nonsensical	1	6 (4%)	7 (2%)
	Total	178 (100%)	164 (100%)	342 (100%)

As for the RQ2, the results showed that the participants considered that 45% (80/178) of the questions generated by QG-Net (used in QUADL) were appropriate for achieving the associated learning objective. Notice that the result is influenced by the performance of the Answer Prediction model because questions are generated from sentences that the Answer Prediction model predicted target tokens.

5 Conclusion

We proposed QUADL for generating questions that are aligned with the given learning objective. As far as we are aware, there have been no studies that aim to generate questions that are suitable for attaining the learning objectives. The evaluation through Amazon Mechanical Turk revealed that the 70% of the predicted target tokens were considered to be appropriate. The result also showed there is a need for improvement to reduce the false negatives—incorrectly predicting that a given sentence is not suitable to attain the learning objective. The current study utilized a survey on Amazon Mechanical Turk.

Evaluating the effectiveness of generated questions with real students in an authentic context is an important next step to be conducted.

Acknowledgements. The research reported here was supported by National Science Foundation Grant No. 2016966 and No. 1623702 to North Carolina State University.

References

1. Rivers, M.L.: Metacognition about practice testing: a review of learners' beliefs, monitoring, and control of test-enhanced learning. Educ. Psychol. Rev. (2020)
2. Pan, S.C., Rickard, T.C.: Transfer of test-enhanced learning: meta-analytic review and synthesis. Psychol. Bull. **144**(7), 710–756 (2018)
3. Koedinger, K.R., et al.: Learning is not a spectator sport: doing is better than watching for learning from a MOOC. In: Proceedings of the Second (2015) ACM Conference on Learning@ Scale (2015)
4. Kurdi, G., et al.: A systematic review of automatic question generation for educational purposes. Int. J. Artif. Intell. Educ. **30**(1), 121–204 (2020)
5. Pan, L., et al.: Recent advances in neural question generation. arXiv preprint arXiv:1905.08949 (2019)
6. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822 (2018)
7. Trischler, A., et al.: Newsqa: A machine comprehension dataset. arXiv preprint arXiv:1611.09830 (2016)
8. Bajaj, P., et al.: Ms marco: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268 (2016)
9. Welbl, J., Liu, N.F., Gardner, M.: Crowdsourcing multiple choice science questions. arXiv preprint arXiv:1707.06209 (2017)
10. Wang, Z., et al.: QG-net: a data-driven question generation model for educational content. In: Proceedings of the Fifth Annual ACM Conference on Learning at Scale (2018)
11. Kim, Y., et al.: Improving neural question generation using answer separation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2019)
12. Nema, P., et al.: Let's Ask Again: Refine Network for Automatic Question Generation. arXiv preprint arXiv:1909.05355 (2019)
13. Yuan, X., et al.: Machine comprehension by text-to-text neural question generation. arXiv preprint arXiv:1705.02012 (2017)
14. Willis, A., et al.: Key phrase extraction for generating educational question-answer pairs. In: Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale (2019)
15. Devlin, J., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)