Automatic Question Generation for Evidence-based Online Courseware Engineering

Machi Shimmei[1[0000-0002-4790-8863]] and Noboru Matsuda[1[0000-0003-2344-1485]]

[1] North Carolina State University, Raleigh NC USA
<mshimme,nmatsud>@ncsu.edu

Abstract. The goal of the current study is to develop an algorithm for generating pedagogically valuable questions. We focus on verbatim questions whose answer, by definition, can be literally identified in a source text. We assume that an important keyphrase relative to a specific learning objective can be identified in a given source text. We then further hypothesize that a pedagogically valuable verbatim question can be generated by converting the source text into a question for which the keyphrase becomes an answer. We therefore propose a model that identifies a keyphrase in a given source text with a linked learning objective. The tagged source text is then converted into a question using an existing model of question generation, QG-Net. An evaluation study was conducted with existing authentic online course materials. Corresponding course instructors judged 66% of the predicted keyphrases were suitable for the given learning objective. The results also showed that 82% of the questions generated by pre-trained QG-Net were judged as pedagogically valuable.

Keywords: Question Generation, Deep Neural Network, Natural Language Processing, Learning Engineering, MOOC.

1 Introduction

Questions plays important roles on learning and teaching. On Massive Open Online Courses (MOOC), formative questions are essential component to make the courseware effective. A research demonstrated, for example, that students learn better when they practice skills by answering questions than by only watching videos or reading text [1]. In a broader context, the benefit of answering questions for learning has been shown in many studies, aka *test-enhanced learning* [2, 3]. However, creating questions that effectively help students' learning requires experience and extensive efforts.

Although there are several studies on the automation of question generation in the field of AI in education [4, 5], little has been discussed about the pedagogical value of the questions generated. To fill this gap, we propose a method for generating questions that supposedly ask about the key concepts the students need to learn to attain the learning objectives. As far as the authors are aware, there have been little study conducted

to generate questions that align with the learning objectives. We propose to develop a technique called QUADL (QUestion generation with an Application of Deep Learning) that generates *verbatim* questions from a pair of a learning objective and a sentence. The verbatim question is a question for which an answer can be literally identified in a related instructional text (i.e., *source text*).

Our central hypothesis is that pedagogically valuable verbatim questions can be generated if source texts are tagged with keyphrases relative to a given learning objective. Once a source text is tagged, then existing seq2seq technologies for question conversion can be used (e.g., [6-8]). The technological contribution of the current research is therefore to develop a deep neural-network model to identify a keyphrase given a pair of a source text and a learning objective.

Accordingly, QUADL consists of the Answer Prediction model and the Question Conversion model. The Answer Prediction model identifies a keyphrase in a given source text. The Question Conversion model generates a question by converting the source text into a question for which the keyphrase becomes the answer.

2 Related Work

The research on the automatic question generation has been growing rapidly among the AIED researchers. Most of the early studies of question generation adapted the rule-based models that relied on templates constructed by experts [9-11]. The scalability is, however, a concern for the rule-based models. They often do not work for complex sentences. The linguistic diversity in resulted questions is therefore limited.

More recent works on question generation take a data-driven approach using neural networks. Many variants of RNN-based models have been proposed and showed considerable advances in the question generation task [12-17]. For general-purpose question generation, large datasets collected from articles in Wikipedia or news media, such as SQuAD [18], NewsQA [19], and MSMARCO [20], enabled to build neural-network based models. Wang *et al*. [21] demonstrated that an LSTM-based model, called QG-Net, trained on a general question generation dataset (SQuAD) can be used for generating questions on educational contents. Questions were generated from textbooks on Biology, Sociology and History for evaluation and showed the highest BLEU score among the state-of-the-art techniques. Yet, the pedagogical value of the generated questions has not been reported.

Techniques for keyphrase extraction has been studied to suggest an answer candidate from a given paragraph text (e.g., [22]). Since our model aims to select target tokens that are aligned with a given learning objective, our proposed Answer Prediction model is essentially different from those existing keyphrase extraction models.

3 Methods

Figure 1 shows an overview of QUADL. Given a pair of a learning objective $\boldsymbol{LO}$ and a source text $\boldsymbol{S}$, $<\boldsymbol{LO}, \boldsymbol{S}>$, QUADL generates a question $\boldsymbol{Q}$ that will be suitable to achieve the learning objective $\boldsymbol{LO}$. The question $\boldsymbol{Q}$ is a verbatim question whose answer can be

literally found in source text ***S***. The following is an example of <***LO***, ***S***> and ***Q***:

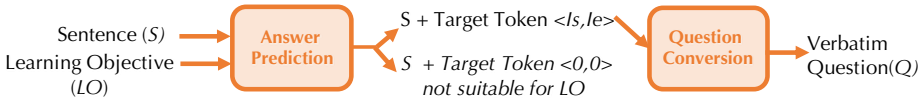

Figure 1 : The QUADL model

Learning objective (***LO***): Describe metabolic pathways as stepwise chemical transformations either requiring or releasing energy; and recognize conserved themes in these pathways.
Source Text (***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 in source text ***S*** (underlined in the example above). We call the tagged answer in the given source text ***S*** a *target token* hereafter. The target token might contain multiple words as shown in the example above.

The Answer Prediction model identifies the *target token index*, <*Is,Ie*>, where *Is* and *Ie* show the index of the start and end of a target token within a given source text ***S*** relative to the learning objective ***LO***. For the Answer Prediction model, we adopted BERT, Bidirectional Encoder Representation from Transformers [23]. In our application, the learning objective (***LO***) and the source text (***S***) were combined as a single input <***LO***, ***S***> to the model. The vector representation computed by the BERT model is given to two different classification models: the one for predicting the start index (*Is*) and the other is for the end index (*Ie*) of the target token. The models may output <*Is*=0, *Ie*=0>, indicating that the given source text is not suitable to generate a question for the given learning objective. For the rest of the paper, we call source texts that have non-zero indices (i.e., Is ≠ 0 and Ie ≠ 0) the *target source texts*, whereas others are referred to as the *non-target source texts* (i.e., has the zero token index <0, 0>). The Answer Prediction model was trained using training data that we created from existing online courses at Open Learning Initiative[†] (OLI).

The Question Conversion model generates a question for which the target token becomes the answer, given a source text with the non-zero target token index. We use QG-Net, a bidirectional-LSTM seq2seq model with attention and copy mechanisms [21]. We used an existing, pre-trained QG-Net model that was trained on the SQuAD datasets[‡]. We could train QG-Net using the OLI course data mentioned above. 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.

[†] https://oli.cmu.edu
[‡] https://rajpurkar.github.io/SQuAD-explorer/

Table 1: Examples of triplet <*LO, S<Is, Ie>, Q*> used in the AMT survey. The target token is bold and underlined in *S*.

(a) A participant judged a target token was suitable, but the question was not suitable.	
LO:	Explain how the cellular organization of fused skeletal muscle cells allows muscle tissue to contract properly.
S:	**Myofibrils** are connected to each other by intermediate, or desmin, filaments that attach to the Z disc.
Q:	What is connected to each other?

(b) A participant judged both a target token and a question were suitable.	
LO:	Identify and discuss the functions of the large intestine and its structures.
S:	The first part of the large intestine is the **cecum**, a small sac-like region that is suspended inferior to the ileocecal valve.
Q:	What is the first part of the large intestine?

4 Evaluation Study

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

To answer these research questions, we conducted a survey on Amazon Mechanical Turk (AMT). In AMT, the participants were shown triplets <***LO***, ***S***<*Is*, *Ie*>, ***Q***>. For each of the triplets, the participants were asked if they agreed or disagreed with the following two statements: (1) To create 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. The examples of triples are shown in Table 1.

Majority votes are used to consolidate the evaluation from participants. Table 2 summarizes the results for RQ1. The data showed that 49% (166/342) of the total predictions about the target token index from the Answer Prediction model were accepted by the participants. For the predictions with a non-zero target index, 88% (155/178) of the predictions were accepted including tie. As for the non-target source text predictions (i.e., the Answer Prediction model output the zero <0,0> index), only 41% (68/164) were accepted. The participants considered 55% (90/164) of the predicted non-target source texts to be target source texts. These results show that *the Answer Prediction model is rather conservative. When it outputs "positive" predictions (i.e., treating a given source text as a target source text), 70% of such predictions are appropriate. However, there is a large number of source texts that should have been predicted as a target source text but missed*. We argue that for the educational purposes, these results are accepted and pragmatic.

Table 3-a shows the results for the RQ2. The table shows that participants considered

Table 2. The evaluation of the predicted target tokens by the Answer Prediction model. There were 178 source texts that the Answer Prediction model predicted target tokens and 164 source texts that the model predicted non-target. The table shows how many of them were accepted/not accepted by the majority vote by Amazon Mechanical Turk participants.

	non-zero target index <*Is* ≠ 0, *Ie* ≠ 0>	zero-index <*Is*=0, *Ie*=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%)

Table 3. The evaluation of the questions generated by QG-Net. The table shows the acceptance of these questions with the majority votes done by the Amazon Mechanical Turk participants. (a) All source texts that the Answer Prediction model predicted a non-zero target token index (N=178) were converted into questions. (b) Only the target source texts accepted by participants in Table 2 (N=123) were converted into questions.

(a)

	Number of questions
Accepted	**80 (45%)**
Tie	50 (28%)
Not accepted	43 (24%)
Nonsensical	5 (3%)
Total	178 (100%)

(b)

	Number of questions
Accepted	**76 (62%)**
Tie	24 (20%)
Not accepted	21 (17 %)
Nonsensical	2 (2%)
Total	123 (100%)

that 73% (130/178) of the questions generated by QG-Net were appropriate for achieving the associated learning objective.

Notice that the result shown above is influenced by the performance of Answer Prediction model. To investigate the capability of QG-Net separately from the performance of the Answer Prediction model, we analyzed the performance of QG-Net given only the "appropriate" inputs (according to the survey participants). Table 3-b shows the evaluation of the questions when QG-Net was given only those source texts that the Answer Prediction model output an "appropriate" target token index according to the survey participants. There, 123 source texts satisfied this condition, which means that 82% (100/123) of questions generated from "appropriate" source texts were considered to be suitable for achieving the associated learning objective. *This indicates that the pre-trained QG-Net can generate a fair number of suitable questions for domains other than the one it was originally trained. Using QG-Net as a building block for QUADL is therefore an acceptable design option.*

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 current study showed that when the Answer Prediction model output a non-zero index for the target token, 88% of such predictions were accepted as good predictions by the study participants. Though we admit that the performance should be improved, this is an encouraging result showing the potential of the proposed model. The data also showed that the majority of the participants believed that 55% of the target source texts that the Answer Prediction model identified as not being useful for the learning objective were actually useful for creating questions. Lowering the amount of "false negative" predictions is certainly a crucial next step.

One of the challenges of the current study was a cost for creating the training data. To train the Answer Prediction model, each target source text paired with a learning objective has to be annotated to indicate the target token. For the current study, we used the existing courseware contents taken from OLI. When the training data were created, target source texts were tagged using answers (extracted from assessment questions) by exact match—i.e., a non-zero token index was assigned only when the target answer appeared literally in the source text. Those source texts that included only a part of the answer or contained synonymous words that were equally plausible as the original answer were not tagged with appropriate token indices. 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.

Acknowledgement

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

References

1. 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.
2. Rivers, M.L., *Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning.* Educational Psychology Review, 2020.
3. Pan, S.C. and T.C. Rickard, *Transfer of test-enhanced learning: Meta-analytic review and synthesis.* Psychological Bulletin, 2018. **144**(7): p. 710-756.
4. Kurdi, G., et al., *A Systematic Review of Automatic Question Generation for Educational Purposes.* International Journal of Artificial Intelligence in Education, 2020. **30**(1): p. 121-204.
5. Pan, L., et al., *Recent advances in neural question generation.* arXiv preprint arXiv:1905.08949, 2019.

6. Kim, Y., et al. *Improving neural question generation using answer separation*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019.
7. Nema, P., et al., *Let's Ask Again: Refine Network for Automatic Question Generation.* arXiv preprint arXiv:1909.05355, 2019.
8. Yuan, X., et al., *Machine comprehension by text-to-text neural question generation.* arXiv preprint arXiv:1705.02012, 2017.
9. Mazidi, K. and P. Tarau, *Automatic question generation: from NLU to NLG.* International Conference on Intelligent Tutoring Systems, 2016: p. pp.23-33.
10. Mitkov, R. *Computer-aided generation of multiple-choice tests*. in *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing*. 2003.
11. Heilman, M. and N.A. Smith, *Question generation via overgenerating transformations and ranking*. 2009, Carnegie-Mellon Univ Pittsburgh pa language technologies insT.
12. Zhao, Y., et al., *Paragraph-level neural question generation with maxout pointer and gated self-attention networks*, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Blanco and W. Lu, Editors. 2018. p. 3901-3910.
13. Wang, S., et al., *A multi-agent communication framework for question-worthy phrase extraction and question generation*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, P. Stone, P.V. Hentenryck, and Z.-H. Zhou, Editors. 2019. p. 7168-7175.
14. Song, L., et al., *Leveraging context information for natural question generation*, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, F. Liu and T. Solorio, Editors. 2018. p. 569-574.
15. Ma, X., et al., *Improving question generation with sentence-level semantic matching and answer position inferring*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, F. Rossi, V. Conitzer, and F. Sha, Editors. 2020. p. 8464-8471.
16. Kim, Y., et al., *Improving neural question generation using answer separation*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, P. Stone, P.V. Hentenryck, and Z.-H. Zhou, Editors. 2019. p. 6602-6609.
17. Tang, D., et al., *Learning to collaborate for question answering and asking*, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, T.S. Fei Liu, Editor. 2018. p. 1564-1574.
18. Rajpurkar, P., R. Jia, and P. Liang, *Know what you don't know: Unanswerable questions for SQuAD.* arXiv preprint arXiv:1806.03822, 2018.
19. Trischler, A., et al., *Newsqa: A machine comprehension dataset.* arXiv preprint arXiv:1611.09830, 2016.
20. Bajaj, P., et al., *Ms marco: A human generated machine reading comprehension dataset.* arXiv preprint arXiv:1611.09268, 2016.
21. 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.

22. 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.
23. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.