Machine-Generated Questions Attract Instructors When Acquainted with Learning Objectives

Machi Shimmei[1](✉), Norman Bier[2], and Noboru Matsuda[1]

[1] North Carolina State University, Raleigh, NC 27695, USA
{mshimme,Noboru.Matsuda}@ncsu.edu
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA
nbier@cmu.edu

Abstract. Answering questions is an essential learning activity on online courseware. It has been shown that merely answering questions facilitates learning. However, generating pedagogically effective questions is challenging. Although there have been studies on automated question generation, the primary research concern thus far is about if and how those question generation techniques can generate answerable questions and their anticipated effectiveness. We propose QUADL, a pragmatic method for generating questions that are aligned with specific learning objectives. We applied QUADL to an existing online course and conducted an evaluation study with in-service instructors. The results showed that questions generated by QUADL were evaluated as on-par with human-generated questions in terms of their relevance to the learning objectives. The instructors also expressed that they would be equally likely to adapt QUADL-generated questions to their course as they would human-generated questions. The results further showed that QUADL-generated questions were better than those generated by a state-of-the-art question generation model that generates questions without taking learning objectives into account.

Keywords: Question Generation · MOOCS · Learning Engineering

1 Introduction

Questions are essential components of online courseware. For students, answering questions is a necessary part of learning to attain knowledge effectively. The benefit of answering questions for learning (known as *test-enhanced learning*) has been shown in many studies [1, 2]. The literature suggests that answering retrieval questions significantly improves the acquisition of concepts when compared to just reading the didactic text [3]. Formative questions are also important for instructors. Students' answers for formative questions provide insight into their level of understanding, which, in turn, helps instructors enhance their teaching.

Despite the important role of questions on online courseware, creating large numbers of questions that effectively help students learn requires a significant amount of time and

N. Wang et al. (Eds.): AIED 2023, LNAI 13916, pp. 3–15, 2023.
https://doi.org/10.1007/978-3-031-36272-9_1

experience. To overcome this issue, researchers have been actively engaged in developing techniques for automated question generation [4]. However, in the current literature, most of the studies on question generation focus on linguistic qualities of generated questions like clarity and fluency. In other words, there has been a lack of research concern about the pedagogical value of the questions generated. It is therefore critical to develop a pragmatic technique for automatically generating questions that effectively help students learn.

The pedagogical value of questions can be discussed from multiple perspectives. In this study, we define pedagogical relevance as the degree to which a question helps students achieve learning objectives. Learning objectives specify goals that the students are expected to achieve, e.g., "*Explain the structure of the inner ear*." With this intention, the goal of the current study is to develop a machine-learning model that can generate questions that are aligned with given learning objectives. As far as we know, there has been no such question generation model reported in the current literature.

We hypothesize that if a key term or phrase related to a specific learning objective can be identified in a didactic text, then a verbatim question can be generated by converting the corresponding text into a question for which the key term or phrase becomes an answer. A verbatim question is a question for which an answer can be literally found in a related text. It is known that answering verbatim questions (even without feedback) effectively facilitates learning conceptual knowledge, arguably because doing so encourages students in the retrieval of relevant concepts [5].

Based on this hypothesis, we have developed a deep neural network model for question generation, called QUADL (**QU**iz generation with **A**pplication of **D**eep **L**earning). QUADL consists of two parts: the answer prediction model and the question conversion model. The answer prediction model predicts whether a given sentence is suitable to generate a verbatim question for a given learning objective. The output from the answer prediction model is a token index *<Is, Ie>* indicating a start and end of the target tokens, which are one or more consecutive words that represent key concepts in the given sentence. The question conversion model then converts the sentence into a question whose verbatim answer is the target tokens.

The primary research questions in this paper are as follows: Does QUADL generate questions that are pedagogically relevant to specific learning objectives? Is QUADL robust enough to apply to existing online courseware?

To answer those research questions, QUADL was applied to an existing online course (OLI[1] Anatomy & Physiology course). Then, in-service instructors were asked to evaluate the pedagogical relevance of generated questions through a survey (Sect. 4.2). In the survey, questions generated by QUADL were blindly compared to both those generated by Info-HCVAE [6] (a state-of-the-art question-generation system that does not take learning objective into account) and those generated by human experts.

The results show that QUADL questions are evaluated as on-par with human-generated questions, and remarkably better than the state-of-the-art question generation model in terms of the relevance to the learning objectives and the likelihood of adoption.

[1] Open learning Initiative (https://oli.cmu.edu).

From a broader perspective, QUADL has been developed as part of our effort to develop evidence-based learning engineering technologies that we call PASTEL (Pragmatic methods to develop Adaptive and Scalable Technologies for next-generation E-Learning) [7]. The primary goal for PASTEL is to assist instructional developers to build adaptive online courseware.

The major contributions of the current paper are as follows: (1) We developed and open sourced[2] a pragmatic question-generation model for online learning, QUADL, that can generate questions that are aligned with specific learning objectives, which is the first attempt in the current literature. (2) We demonstrated that instructors rated the pedagogical relevance of questions generated by QUADL as on-par with human-generated questions and higher than questions generated by a state-of-the-art model.

2 Related Work

There are two types of question generation models: answer-unaware models and answer-aware models. In *answer-unaware question generation models* (also known as answer-agnostic models), knowledge about the answer is not directly involved in the question generation pipeline either as an input or output (e.g., [8–10]). Given a source context (e.g., a sentence, a paragraph, or an article), the answer-unaware model generates a question(s) asking about a concept that appeared in the source context. However, the corresponding answer is not explicitly shown. Answer-unaware models are not suitable for QUADL because students' responses cannot be evaluated automatically without knowing correct answers.

In *answer-aware question generation models*, on the other hand, knowledge about the answer is explicitly involved in the question generation pipeline. In the current literature of question generation, the most actively studied models are answer-aware models for which the answer data are manually provided (e.g., [11–16]). There are also other answer-aware models that identify keywords (that become answers) by themselves and generate questions accordingly, called question-answer pair generation (QAG) models [6, 17–20]. There have been many QAG models specifically proposed for educational use [21–23]. QUADL is one such QAG model.

Some question generation models utilize extra knowledge in addition to the source context and/or an answer (though these models are not very common). For example, a model proposed by Pyatkin *et al.* [10] is given a predicate (e.g., "arrive") with a context, and produces a set of questions asking about all possible semantic roles of the given predicate (e.g., "start point", "entity in motion", etc.). Wang *et al.* [15] proposed a model whose input is a pair of a target and a paragraph. The target is a word or a phrase that specifies a topic of the question such as "size" or "sleep deprivation," but it is not an exact answer to the question. The model generates questions that ask about concepts relevant to the target in the paragraph.

QUADL also belongs to this type of question generation model. It takes a learning objective as the extra knowledge and generates questions that are suitable for the given

[2] The code and data used for the current study is available at https://github.com/IEClab-NCSU/QUADL.

learning objective. *As far as we are aware, no studies have been reported in the current literature that generate questions that align with specific learning objectives.*

3 Overview of *QuADL*

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 is assumed to be suitable to achieve the learning objective *LO*. Examples of < *LO, S*> and *Q* are shown in Table 1 in Sect. 5.3. Notice that a target token is underlined in the sentence *S* and becomes the answer for the question.

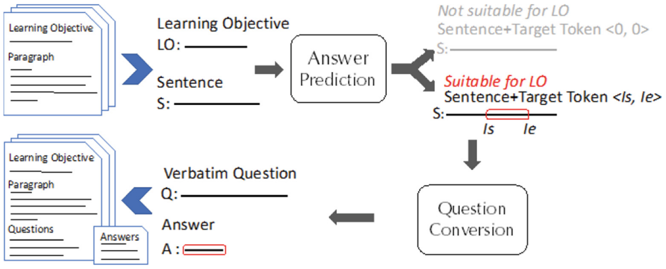

Fig. 1. An overview of QuADL. The answer prediction model identifies the start/end index <*Is, Ie*> of the target token (i.e., key term) in *S*. When *S* is not suitable for *LO*, it outputs <0,0>. The question conversion model converts *S* with target token to a verbatim question.

The input of the answer prediction model is a single sentence (or a *source sentence* for the sake of clarity) and a learning objective. In our application, each sentence *S* in a paragraph is paired with the learning objective *LO* as a single input <*LO, S*> to the model. The final output from the answer prediction model is a *target token index*, <*Is, Ie*>, where *Is* and *Ie* show the index of the start and end of the target token within the source sentence *S* relative to the learning objective *LO*. The models may output <$Is = 0$, $Ie = 0$>, indicating that the source sentence is not suitable to generate a question for the given learning objective. For the rest of the paper, we refer to source sentences that have non-zero indices (i.e., $Is \neq 0$ and $Ie \neq 0$) as *target sentences*, whereas the other source sentences that have the zero token index <0, 0> *non-target sentences*.

For the answer prediction model, we adopted Bidirectional Encoder Representation from Transformers (BERT) [24]. The final hidden state of the BERT model is fed to two single-layer classifiers. One of them outputs a vector of probabilities, $P_s(Is = i)$, indicating the probability of the *i-th* index in the sentence being beginning of the target token. Another model outputs a vector of probabilities for the end index, $P_e(Ie = j)$. To compute the final probability for being the target token, a normalized sum of $P_s(i)$ and $P_e(j)$ is first calculated as the joint probability $P(Is = i, Ie = j)$ for every possible span ($Is < Ie$) in the sentence. The probability $P(Is = 0, Ie = 0)$ is also computed, indicating a likelihood that the sentence is *not* suitable to generate a question for the learning objective. The pair <$Is = i$, $Ie = j$> with the largest joint probability becomes the final prediction.

As for the question conversion model, we hypothesize that if a target token is identified in a source sentence, a pedagogically valuable question can be generated by converting that source sentence into a verbatim question using a sequence-to-sequence model that can generate fluent and relevant questions. In the current implementation, we used the state-of-the-art technology, called ProphetNet [13], as a question conversion model. ProphetNet is an encoder-decoder pre-training model that is optimized by future n-gram prediction while predicting n-tokens simultaneously.

These two models in QUADL are domain and courseware independent. This section only describes their structures. The next section describes how those models were trained with the data from an existing online course for an evaluation study.

4 Evaluation Study

To evaluate the anticipated pedagogical value of the questions generated by QUADL, we conducted a human-evaluation study with in-service instructors who use an existing online course hosted by Open Learning Initiative (OLI) at Carnegie Mellon University.

4.1 Model Implementation

Answer Prediction Model. The answer prediction model was fine-tuned with courseware content data taken from the Anatomy and Physiology (A&P) OLI course. The A&P course consists of 490 pages and has 317 learning objectives. To create training data for the answer prediction model, in-service instructors who actively teach the A&P course manually tagged the didactic text as follows. The instructors were asked to tag each sentence *S* in the didactic text to indicate the target tokens relevant to specific learning objective *LO*. A total of 8 instructors generated 350 pairs of <*Lo, S*> for monetary compensation. Those 350 pairs of token index data were used to fine-tune the answer prediction model.

Since only a very small amount of data was available for fine-tuning, the resulting model was severely overfitted. We therefore developed a unique ensemble technique that we argue is an innovative solution for training deep neural models with extremely small data. A comprehensive description of the ensemble technique can be found in Shimmei *et al.* [25] Due to the space constraint, this paper briefly shows how the ensemble technique was applied to the answer prediction model. Note that the ensemble technique is not necessary if a sufficient amount of data are available.

To make an ensemble prediction, 400 answer prediction models were trained independently using the same training data, but each with a different parameter initialization. Using all 400 answer prediction models, an ensemble model prediction was made as shown below.

Recall that for each answer prediction model k ($k = 1, \ldots, 400$), two vectors of probabilities, the start index ${P_s}^k(i)$ and the end index ${P_e}^k(j)$, are output. Those probabilities were averaged across all models to obtain the ensemble predictions $P_s^*(i)$ and $P_e^*(j)$ for the start and end indices, respectively. The final target token prediction $P(Is = i, Ie = j)$ was then computed using P_s^* and P_e^* (see Sect. 3).

Subsequently, we applied rejection method. This method is based on the hypothesis that a reliable prediction has stable and high probabilities across models, while an unreliable prediction has diverse probabilities across models, which results in smaller values when averaged. Therefore, if either of $P_s^*(i)$ or $P_e^*(j)$ were below the pre-defined threshold $R \in (0,1)$, the model discarded the prediction $< Is = i, Ie = j >$. When all predictions were below the threshold, the model did not make any prediction for the sentence, i.e., the prediction is void.

Rejection increases the risk of missing target sentences but decreases the risk of creating questions from non-target sentences. For the sake of pedagogy, question quality is more important than the quantity. Rejecting target token prediction with a low certainty (i.e., below the threshold) ensures that the resulting questions are likely to be relevant to the learning objective. In the current study, we used a threshold of 0.4 for rejection. We determined the number of models and the threshold using the performance on the human-annotated test dataset.

Question Conversion Model. We used an existing instance of ProphetNet that was trained on SQuAD1.1 [26], one of the most commonly used datasets for question generation tasks. SQuAD1.1 consists of question-answer pairs retrieved from Wikipedia. We could train ProphetNet using the OLI course data. However, the courseware data 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 therefore not suitable to generate a training dataset for ProphetNet.

4.2 Survey Study

Five instructors (the "participants" hereafter) were recruited for a survey study. The total of 100 questions used in the survey were generated by three origins: QUADL, Info-HCVAE, and human experts, as described below.

QUADL questions: 34 questions out of 2191 questions generated by QUADL using the method described above were randomly selected for the survey.

Info-HCVAE questions: 33 questions were generated by Info-HCVAE [6], a state-of-the-art question generation model that generates questions without taking the learning objective into account. Info-HCVAE extracts key concepts from a given paragraph and generates questions for them. We trained an Info-HCVAE model on SQuAD1.1. Info-HCVAE has a hyperparameter K that determines the number of questions to be generated from a paragraph. We chose K=5 because that was the average number of questions per paragraph in the A&P course. The Info-HCVAE model was applied to a total of 420 paragraphs taken from the courseware, and 2100 questions were generated. Questions with answers longer than 10 words or related to multiple sentences were excluded. Consequently, 1609 questions were left, among which 33 questions were randomly selected for the survey.

Human questions: 33 formative questions among 3578 currently used in the OLI A&P course were randomly collected. Most of the A&P questions are placed immediately after a didactic text paragraph. Only questions whose answers can be literally found in the didactic text on the same page where the question appeared were used because questions generated by QUADL are short-answer, verbatim questions. The course contains various types of formative questions such as fill-in-the-blank, multiple-choice, and short-answer questions. Fill-in-the-blank questions were converted into interrogative sentences. For example, “The presence of surfactant at the gas-liquid interphase lowers the ____ of the water molecules.” was changed to “The presence of surfactant at the gas-liquid interphase lowers what of the water molecules?”. The multiple-choice questions were also converted into short-answer questions by hiding choices from participants.

Each survey item consists of a paragraph, a learning objective, a question, and an answer. Participants were asked to rate the prospective pedagogical value of proposed questions using four evaluation metrics that we adopted from the current literature on question generation [15, 21, 27]: answerability, correctness, appropriateness, and adoptability. Answerability refers to whether the question can be answered from the information shown in the proposed paragraph. Correctness is asking whether the proposed answer adequately addresses the question. Appropriateness is asking whether the question is appropriate for helping students achieve the corresponding learning objective. Adoptability is asking how likely the participants would adapt the proposed question to their class. Each metric was evaluated on a 5-point Likert scale.

Every individual participant rated all 100 questions mentioned above (i.e., 100 survey items). Five responses per question were collected, which is notably richer than any other human-rated study for question generation in the current literature that often involve only two coders.

5 Results

5.1 Instructor Survey

Inter-Rater Reliability: We first computed Krippendorff’s alpha to estimate inter-rater reliability among study participants. There was moderate agreement for answerability, correctness, and appropriateness (0.56, 0.49, and 0.47 respectively). The agreement for adoptability was weak (0.34), indicating that there were diverse factors among participants that determine the adoptability of the proposed questions.

Overall Ratings: Figure 2 shows the mean ratings per origin (QUADL vs. Infor-HCVAE vs. Human) across participants for each metric. The plot shows that QUADL- and human-generated questions are indistinguishable on all four metrics, whereas questions generated by Info-HCVAE are clearly different.

Statistical tests confirmed the above observations. One-way ANOVA tests (when applied separately to each metric) revealed that origin (QUADL vs. Infor-HCVAE vs. Human) is a main effect for ratings on all four metrics; F(2, 97) = 36.38, 24.15, 26.11, and 25.03, for answerability, correctness, appropriateness, and adoptability, respectively, with $p < 0.05$ for each of them. A post hoc analysis using Tukey’s test showed that there was a statistically significant difference between QUADL and Info-HCVAE at $p < .05$, but no significant difference between QUADL and human-generated questions for each of the four metrics.

Overall, the results suggest that in-service instructors acknowledged that questions generated by QUADL and humans had equal prospective pedagogical values when asked blindly. It is striking to see that *the in-service instructors suggested that they would be equally likely to adapt QUADL- and human-generated questions to their courses.*

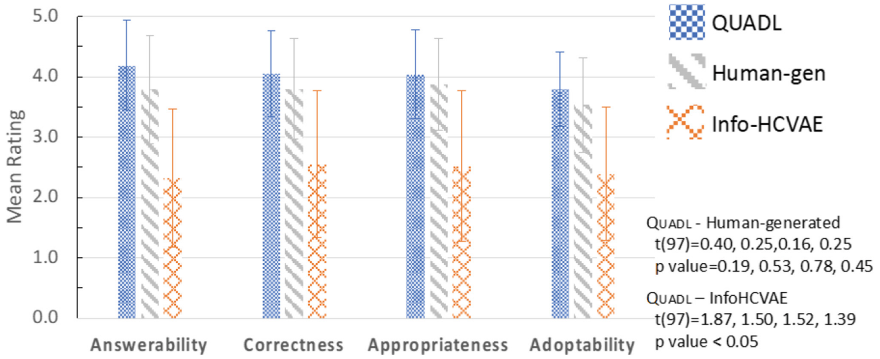

Fig. 2. Mean of the ratings for each metrics. The standard deviation is shown as an error bar. The value ranges from 1 as strongly disagree to 5 as strongly agree.

As for the comparison between question generation technologies with and without taking learning objectives into account, the results suggested that learning objective-specific questions (QUADL) were rated higher than learning objective-unaware questions (Info-HCVAE) on all four metrics. Since Info-HCVAE does not aim to generate questions for particular learning objectives, appropriateness might not be a fair metric. Yet, instructors showed clear hesitation to adapt questions generated by Info-HCVAE to their courses. The poor scores on answerability and correctness for Info-HCVAE questions might be mostly due to the poor performance of the question conversion model used (see Sect. 6 for the further discussion).

5.2 Accuracy of the Answer Prediction Model

To evaluate the performance of the answer prediction model, we operationalize the accuracy of the model in the target token identification task using two metrics: token precision and token recall. *Token precision* is the number of correctly predicted tokens divided by the number of tokens in the prediction. *Token recall* is the number of correctly predicted tokens divided by the number of ground truth tokens. For example, suppose a sentence “*The target tissues of the nervous system are muscles and glands*” has the ground truth tokens as “*muscles and glands*”. When the predicted token is “*glands*”, the token precision is 1.0 and token recall is 0.33.

We computed the average of token precision and token recall across the predictions for a test dataset that contains 25 target and 25 non-target sentences. The averages of token precision and token recall were 0.62 and 0.21, respectively. Even with the very limited amount of training data (350), the proposed ensemble-learning technique (Sect. 4.1) achieved 0.62 token precision with the cost of recall. The recall score is notably low because predictions with low certainty were rejected. For our primary purpose of making instructional questions, we value quality over quantity. Therefore, we consider the currently reported performance of the answer prediction model to be adequate in particular when considering the extremely low amount of data available for training. Of course, improving token recall is important future research.

5.3 Qualitative Analysis of Questions Generated by *QUADL*

There are questions generated by QUADL that were rated low by the instructors. How could QUADL be further improved to generate better questions? To answer this question, we sampled a few QUADL- and human-generated questions from the same paragraphs for the same learning objectives. Table 1 compares these questions, with their appropriateness scores as rated by participants.

We found that the QUADL questions had low appropriateness scores (<2) when they were not particularly specific and did not clearly require students to review the concepts mentioned in the corresponding learning objective. For example, Q2 in Table 1 is ambiguous because it does not mention that it is about the circulation of blood in the urinary tract. The human-generated question (Q1) has enough context and is clear enough to effectively encourage students to review important concepts in the learning objective. For Q2, the answer prediction model predicted a target sentence "Once filtered, the blood exits through the *renal vein.*" with "*renal vein*" as a target token. However, in the original paragraph just before the predicted target sentence, there were other relevant sentences: "The process begins with waste carrying blood entering each of the two kidneys through the renal artery. Urine is produced by the nephrons in the kidney." The question would be more appropriate if it included this context. Q4 better provides the information students need to answer the question and appropriately asks about the concept in the learning objective. It is rated as high as the human-generated question Q3.

Table 1. Example of generated questions. Human- and QUADL-generated questions from the same paragraph and learning objective are compared. ***S*** shows a source sentence of a question, and an answer is written in *italics*. For human-generated questions, the sentences that should be referred to in order to answer the question were retrieved from the paragraph.

Learning Objective: Identify gross and microscopic anatomy of the urinary tract.
Q1 (Human): When enough urine has been produced in the nephrons, it leaves through the ureters and urine is then transported to what? [Appropriateness: 4.4] S: Each ureter transports the urine via peristalsis to the urinary *bladder*.
Q2 (QUADL): Through what part of the body does the blood exit? [Appropriateness: 2.2] ***S*****:** Once filtered, the blood exits through the *renal vein.*
Learning Objective: Describe how the structure of these macromolecules allow the structures c the respiratory system to perform their functions.
Q3 (Human): The presence of surfactant at the gas-liquid interphase lowers what of the water molecules? [Appropriateness: 4.2] ***S***: At the gas-liquid interface of the alveoli cell membranes, surfactants found in the liquid surface layer lower *surface tension***.** Surface tension arises when water molecules hydrogen bond with each other.
Q4 (QUADL): What helps humidify and buffer the cells in direct contact with air? [Appropriateness: 4.2] ***S***: Throughout the entire respiratory system, *mucus* helps humidify and buffer the cells that are in direct contact with air.

To overcome this shortcoming, one of the solutions would be to change the unit-of-analysis for QUADL from sentence-level to paragraph-level. In the current study, the target sentence is always a single sentence. Capturing context more properly by taking an entire paragraph (or some relevant sentences) into account might reduce the ambiguity of the question and effectively encourage the retrieval of relevant concepts in the learning objective. Testing this hypothesis is an important future study.

6 Discussion

As far as we are aware, QUADL is the first model that generates pedagogical questions while taking learning objectives into account. The current study is also the first in the literature that evaluates the appropriateness of machine-generated questions relative to the learning objectives.

Through the current study, we encountered many obstacles related to education research. First, collecting datasets of adequate quality and size in order to train a deep neural network model is challenging. For the current study, even though some datasets for question generation are available [28–30], they did not necessarily meet our need—there are various domains, types of questions (e.g., multiple-choice vs. fill-in-the-blank), and difficulty levels. The current study demonstrated that the proposed ensemble method with rejection is a powerful solution for the issue of low data regime.

Second, with the lack of ground-truth data, comparing the machine- and human-generated questions is challenging. In the current study, to obtain human-generated questions, questions from an existing online course were retrieved using an automated process (as mentioned in Sect. 4). However, we noticed that some questions on the online course had extra text associated that provided information about the context. When those questions were converted into survey items, the information about the context was removed, which resulted in ambiguous questions that allowed different answers or were too general. We speculate that this technological complication might be a reason why some human-generated questions in the current survey study received low answerability and correctness scores by study participants.

Third, evaluating the validity and quality (or the "utility") of machine-generated questions is challenging. Strictly speaking, it requires a close-the-loop experiment where students are assigned to courseware with machine-generated questions, and their subsequent learning outcomes are measured. However, due to the cost of conducting a rigorously controlled study in an authentic learning environment, subjective evaluation by human experts is a common technique in the current literature [31]. Nonetheless, with the current promising results, we have been preparing a close-the-loop evaluation study with college students as an important next step towards the wide dissemination of the QUADL technology.

The results in Fig. 2 show that Info-HCVAE received notably low scores on answerability and correctness. This might be merely due to the inferior performance of the question conversion model that it uses. The current implementation of QUADL utilized an existing state-of-the-art question conversion module, ProphetNet. Due to this confounding factor, the current results do not allow us to draw a rigorous conclusion on how QUADL's feature of learning-objective awareness contributed to the better evaluation of

QUADL than Info-HCVAE. Technically speaking, the question of which of the models in QUADL—the answer prediction model or the question conversion model—lead to satisfactory performance has yet to be investigated. To achieve this goal, we plan to modify Info-HCVAE by replacing its question conversion model with ProphetNet. We acknowledge that a lack of conducting this rigorous comparison is a limitation of the current study and constitutes important future research.

The current study demonstrated the significant potential of QUADL as a pragmatic technology for creating next-generation evidence-based online courseware. As stated in the Introduction section, QUADL is part of PASTEL, a suite of evidence-based online learning engineering methods. The promising results reported in this paper encourage us to consider further extension of PASTEL. For example, it may be possible for a machine to learn concepts from texts on the internet and automatically generate didactic text for online courseware.

7 Conclusion

We found that a deep neural network model designed to generate questions that are aligned with the given learning objective, QUADL, performs on-par with human experts for the task of generating pedagogically valuable questions. QUADL is the first model in the literature that aims to generate questions that are suitable for attaining learning objectives. The results encourage us to further conduct an in-class evaluation study to measure students' learning with the machine-generated questions.

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. **33**(3), 823–862 (2021)
2. Pan, S.C., Rickard, T.C.: Transfer of test-enhanced learning: meta-analytic review and synthesis. Psychol. Bull. **144**(7), 710 (2018)
3. Smith, M.A., Karpicke, J.D.: Retrieval practice with short-answer, multiple-choice, and hybrid tests. Memory **22**(7), 784–802 (2014)
4. Pan, L., et al.: Recent advances in neural question generation. arXiv preprint arXiv:1905.08949 (2019)
5. Roediger, H.L., Karpicke, J.D.: The power of testing memory: basic research and implications for educational practice. Perspect. Psychol. Sci. **1**(3), 181–210 (2006)
6. Lee, D.B., et al.: Generating Diverse and Consistent QA pairs from Contexts with Information-Maximizing Hierarchical Conditional VAEs. Association for Computational Linguistics (2020)
7. Matsuda, N., et al.: PASTEL: Evidence-based learning engineering methods to facilitate creation of adaptive online courseware. In: Ouyang, F., et al. (eds.) Artificial Intelligence in STEM Education: The Paradigmatic Shifts in Research, Education, and Technology, pp. 1–16. CSC Press, New York, NY (in press)

8. Lewis, M., et al.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)
9. Du, X., et al.: Learning to ask: neural question generation for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (2017)
10. Pyatkin, V., et al.: Asking it all: generating contextualized questions for any semantic role. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2021)
11. Bao, H., et al.: Unilmv2: pseudo-masked language models for unified language model pre-training. In: International Conference on Machine Learning. PMLR (2020)
12. Chan, Y.-H., Fan, Y.-C.: A recurrent BERT-based model for question generation. In: Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Association for Computational Linguistics (2019)
13. Qi, W., et al.: ProphetNet: Predicting future N-gram for sequence-to-sequence pre-training. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics (2020)
14. 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)
15. Wang, Z., Valdez, J., Mallick, D.B., Baraniuk, R.G.: Towards human-like educational question generation with large language models. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part I, pp. 153–166. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-11644-5_13
16. Xiao, D., et al.: ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In: IJCAI (2020)
17. Du, X., Cardie, C.: Harvesting Paragraph-level Question-Answer Pairs from Wikipedia. Association for Computational Linguistics (2018)
18. Back, S., et al.: Learning to generate questions by learning to recover answer-containing sentences. In: Findings of the Association for Computational Linguistics (2021)
19. Subramanian, S., et al.: Neural Models for Key Phrase Extraction and Question Generation. Association for Computational Linguistics (2018)
20. Wang, B., et al.: Neural question generation with answer pivot. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)
21. Steuer, T., Filighera, A., Rensing, C.: Remember the facts? investigating answer-aware neural question generation for text comprehension. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12163, pp. 512–523. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52237-7_41
22. Willis, A., et al.: Key phrase extraction for generating educational question-answer pairs. In: Proceedings of the Sixth ACM Conference on Learning@ Scale (2019)
23. Qu, F., et al.: Asking questions like educational experts: Automatically generating question-answer pairs on real-world examination data. arXiv preprint arXiv:2109.05179 (2021)
24. Devlin, J., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
25. Shimmei, M., Matsuda, N.: Can't inflate data? let the models unite and vote: data-agnostic method to avoid overfit with small data. In: 14th Inernational Conference on Educatinal Data Mining (to appear)
26. Rajpurkar, P., et al.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
27. Tamang, L.J., Banjade, R., Chapagain, J., Rus, V.: Automatic question generation for scaffolding self-explanations for code comprehension. In: Rodrigo, M.M., Matsuda, N., Cristea,

A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part I, pp. 743–748. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-11644-5_77

28. Chen, G., et al.: LearningQ: a large-scale dataset for educational question generation. In: Twelfth International AAAI Conference on Web and Social Media (2018)
29. Lai, G., et al.: RACE: Large-scale ReAding Comprehension Dataset From Examinations. Association for Computational Linguistics (2017)
30. Welbl, J., et al.: Crowdsourcing multiple choice science questions. arXiv preprint arXiv:1707.06209 (2017)
31. Kurdi, G., et al.: A systematic review of automatic question generation for educational purposes. Int. J. Artif. Intell. Educ. **30**(1), 121–204 (2020)