Towards Generalized Methods for Automatic Question Generation in Educational Domains

Huy A. Nguyen(✉), Shravya Bhat, Steven Moore, Norman Bier, and John Stamper

Carnegie Mellon University, Pittsburgh, PA 15213, USA
hn1@cs.cmu.edu

Abstract. Students learn more from doing activities and practicing their skills on assessments, yet it can be challenging and time consuming to generate such practice opportunities. In our work, we examine how advances in natural language processing and question generation may help address this issue. In particular, we present a pipeline for generating and evaluating questions from text-based learning materials in an introductory data science course. The pipeline includes applying a text-to-text transformer (T5) question generation model and a concept hierarchy extraction model on the text content, then scoring the generated questions based on their relevance to the extracted key concepts. We further evaluated the question quality with three different approaches: information score, automated rating by a trained model (Google GPT-3) and manual review by human instructors. Our results showed that the generated questions were rated favorably by all three evaluation methods. We conclude with a discussion of the strengths and weaknesses of the generated questions and outline the next steps towards refining the pipeline and promoting natural language processing research in educational domains.

Keywords: Question generation · Concept extraction · Question evaluation

1 Introduction

As online education continues to expand during and after the COVID pandemic, the need for effective and scalable assessment tools emerges as a pressing issue for instructors and educators. On one hand, frequent formative assessments are crucial in reinforcing student learning in an online environment, where the learning experience may be undermined by a multitude of factors, including the lack of motivation [1] and student-teacher interaction [19]. On the other hand, summative assessments that rely on human grader evaluation, such as group projects and essays, are difficult to carry out at scale, making multiple-choice and short-answer questions, which are amenable to automatic grading, a more practical alternative. Consequently, amid many other logistical issues that arise from emergency online education [16], instructors often find themselves having to generate a large question bank to accommodate this new learning format. In turn, this challenge motivates the need for supporting instructor efforts via methods that automatically generate usable assessment questions based on the learning materials, in a way that requires minimal inputs from instructors and domain experts.

I. Hilliger et al. (Eds.): EC-TEL 2022, LNCS 13450, pp. 272–284, 2022.
https://doi.org/10.1007/978-3-031-16290-9_20

Recent advances in natural language processing (NLP), question answering (QA) and question generation (QG) offer a promising path to accomplishing this goal. While QA has been a longtime area of interest for NLP researchers, with wide applications ranging from beating the Jeopardy! challenge [14] to supporting modern intelligent assistants [12], QG has only garnered attention in recent years. Much of the interest in QG stems from the large number of BERT-based models trained on very large corpuses that demonstrate the ability to generate interesting results in open domains [49]. QG in educational domains is an even narrower focus, but holds great potential in transforming the way assessments are generated and conducted [39]. Most theories of learning emphasize repeated practice as an important mechanism for mastering low-level knowledge components, which altogether contribute to the high-level learning objectives [20]. We therefore envision that having the ability to generate questions on-demand would accommodate students' varying levels of learning needs, while allowing instructors to allocate resources to other components of the course.

Our work presents an initial step towards realizing this capability. We applied state-of-the-art Text-To-Text Transfer Transformer (T5) models [45] on conceptual reading materials from a graduate-level data science course to generate potential questions that may be used for assessment. We then evaluated these questions in three different ways. First, we conducted a separate concept hierarchy extraction process on the reading materials to extract the important concept keywords and scored each generated question based on how many such keywords it contains. Second, we applied a fine-tuned GPT-3 model to classify the questions as either pedagogically sound or not. Finally, we had two data science instructors perform this same classification task manually. Our results contribute key insights into the feasibility of applying state-of-the-art NLP models in generating meaningful questions, with a pipeline that generalizes well across learning domains.

2 Background

Recent advances in deep learning have revitalized many areas of artificial intelligence. Within the fields of NLP and QG, significant progress has been made since the introduction of neural transformer-based methods [42], particularly deep bidirectional transformers (BERT [11]), which differ from previous language models in their training approach (masked language modeling and next sentence prediction) as well as their subsequent learned representation of text from both sides (left and right) of the sentences. We summarize recent NLP improvements that are pertinent to QG below.

While BERT could help address the problem of handling long sequences that a traditional recurrent neural network encounters, its initial performance in QG was rather poor, as it did not consider the decoding results of previous steps while producing tokens [7]. Lopez et al. [28] solved this issue with fine-tuning techniques on a single pre-trained language model to design a QG system that generates robust questions at reduced training cost and time. Subsequent research also investigated ways to encode common sense and domain knowledge in the QG process, with Jia et al. [18] utilizing concept-relevant knowledge triples from ConceptNet, a freely available knowledge graph, and Wang et al. [43] building custom knowledge graph models to prevent the generation of irrelevant and uninformative questions. More recently, Liu [24] attempted to increase the relevance of generated questions with an attention-based, sequence-to-sequence model that incorporates target answer information into the question generation process. QG models have also been used to generate training corpora for Question Answering tasks [3].

A subset of QG research involves generating questions specifically for educational purposes, to be used as assessment materials [2]. Towards automatically generating educationally usable questions, previous work has investigated targeting certain cognitive levels of questions, including high-level ones that require synthesis and evaluation or low-level ones that focus on recall [47]. For example, recent work has used the GPT-2 model to generate mathematical word problems at varying levels of difficulty [9]. This approach was found to yield high quality questions, as judged by both automatic and human evaluation, with the capability of altering the perceived difficulty of generated questions. Related work by Liu et al. [27] also investigated automatically generating educational questions in math, with a knowledge graph as the source document for their model. The generated questions were evaluated as coherent, diverse and reflective of real-life scenarios that students may encounter. However, a recent review of question generation for educational purposes found that, while methods for producing educationally valid questions are improving, there is a greater need to properly evaluate them [22].

Question evaluation is traditionally split into two core methods, based on whether the evaluation is performed by trained machine learning models or expert human judges. Automatic assessment of questions often involves the use of evaluation metrics such as BLEU and ROGUE, which quantify how close the generated question text is to an existing human-generated text [31]. However, recent work has reported interpretability issues with these metrics, along with a lack of correlation between them and human evaluation [41]. At the same time, Sha et al. [37] found that using BERT to classify student forum posts based on the question type, post sentiment and confusion level achieved similar results as human evaluators. For human evaluation, a recent meta analysis found that over half of the reviewed research involved criteria related to grammar, fluency, topic relevance, and naturalness [9, 13, 47], most frequently on a numerical scale [4]. Previous work involving human evaluation of questions has also utilized different rubrics, such as being useful or not useful for learning [8] or being shallow or deep [34]. In line with these approaches, our work also employs both automated and expert labeling of the generated questions, so as to arrive at a holistic evaluation of their usability.

Another metric for evaluation involves how much the generated questions align with the "ground truth" data, such as reference questions created by human experts [36]. In educational QG, we expect assessment items to match the target skills of the corresponding unit and module, which raises the need to identify these skills from the learning material. A recent effort in automating this task was carried out by the researchers behind MOOCCubeX [46], an open-access, educational data repository created with the aim of supporting research on adaptive learning in massive open online courses. This framework is capable of performing data processing, weakly supervised fine-grained concept graph mining, and data curation to re-organize data in a concept-centric manner. The published toolkit also assists with the creation of new datasets for adaptive learning and concept mining. In our work, we will examine how well the generated questions match the knowledge concepts identified by MOOCCubeX.

3 Methods

3.1 Dataset

We used the learning materials from a graduate-level introductory data science course at an R1 university in the northeastern United States. The course has been offered every semester since Summer 2020, with class sizes ranging from 30–90 in general. The course content is divided into the conceptual components and the hands-on projects. Students learn from six conceptual Units, further broken down into sixteen Modules, each consisting of several data science Topics such as *Feature Engineering* and *Bias-Variance Trade-off*. Each Module consists of reading assignments, ungraded formative assessments and weekly quizzes serving as graded summative assessments. Students also get to practice with the learned concepts through seven hands-on coding projects, which are evaluated by an automatic grading system. In the scope of this work, we will focus on generating questions from the textual content of the six Units in the course, using the pipeline introduced in the following section.

3.2 Question Generation Pipeline

The overall pipeline for question generation and evaluation consists of six steps. First, we extract the learning materials from an online learning platform which hosts the course. This extracted data is in XML format, which preserves not only the text content but also its hierarchy within the course structure, i.e., which Unit, Module and Topic each paragraph of text belongs to. We scraped the text content from the XML files using the Beautiful Soup library[1]. From this point, the resulting text data was input to two separate processes, Concept Hierarchy Extraction and Question Generation.

Concept Hierarchy Extraction. This process was carried out by the MOOCCubeX pipeline [46], which performs weakly supervised fine-grained concept extraction on a given corpus without relying on expert input. As an example, given a paragraph that explains Regression, some of the extracted concepts include *least-squared error*, *regularization*, and *conditional expectation*; these could be viewed as the key concepts which students are expected to understand after reading the materials. A researcher in the team reviewed the generated concepts and manually removed those which were deemed invalid, including prepositions (e.g., *'around'*), generic verbs (e.g., *'classifying'*) and numbers (e.g., *'45'* – this is part of a numeric example in the text, rather than an important constant to memorize).

Question Generation. For this process, we applied Google's T5 [45], which is a transformer-based encoder-decoder model. Since its pre-training involves a multi-task structure of supervised and unsupervised learning, T5 works well on a variety of natural

[1] https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

language tasks by merely changing the structure of the input passed to it. For our use case, we collect all the text within each Topic (which typically consists of 3–6 paragraphs), prepend this text by a header name (which is the name of either the Topic itself, or the corresponding Module, or the corresponding Unit), and input the resulting corpus to T5 (see an example in Fig. 1). In this way, we generate three questions for each Topic in the course. Our rationale for including the header name in the T5 input text is to inform the model of the high-level concept which the generated questions should center around. We had previously tried extracting answers from the text content using a custom rule-based approach with a dependency parse tree, but found that this resulted in the creation of more nonsensical than sensible questions; in comparison, incorporating the headers led to higher quality questions. Before applying the model to our dataset, we also fine-tuned it on SQuAD 1.1 [32], a well known reading comprehension dataset of questions curated by crowd workers on Wikipedia articles and a common benchmark for question-answering models.

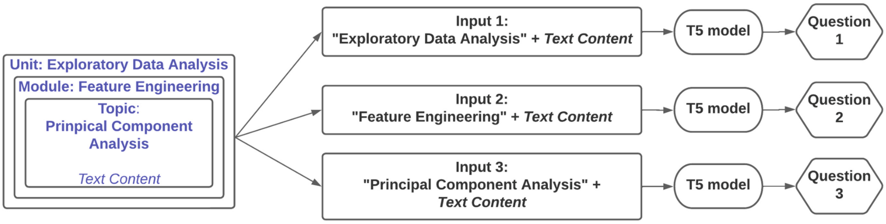
Unit: Exploratory Data Analysis
Module: Feature Engineering
Topic:
Prinpical Component Analysis
Text Content
Input 1:
"Exploratory Data Analysis" + Text Content
Input 2:
"Feature Engineering" + Text Content
Input 3:
"Principal Component Analysis" + Text Content
T5 model
T5 model
T5 model
Question 1
Question 2
Question 3

Fig. 1. Example question generation process for the text content in one Topic.

3.3 Evaluation

We evaluated the generated questions with three different methods as follows.

Information Score. This is a set of custom metrics that denote how relevant each question is to the key concepts identified in the Concept Hierarchy Extraction step. We denote this set of key concepts as C. For every generated question q, we further denote $T(q)$ as the set of tokens in it and compute the *information score* as the number of tokens in q that coincide with an extracted concept,

$$IS(q) = \frac{1}{|T(q)|}\sum\nolimits_{t \in T(q)} 1(t \in C). \tag{1}$$

where the division by q's length is to normalize the metric so that longer questions are not inherently favored. With this formulation, higher scores indicate better questions that touch on more of the key learning concepts.

GPT-3 Classification. We used a GPT-3 classification model [6], as it has been a popular choice for text classification tasks such as detecting hate speech [10] and text sentiment

[48]. Our classification task involves rating each question as either *pedagogically sound* or *not*. A pedagogically sound question is one that pertains to the course content and is intended to assess the domain knowledge of the student. An example of a question classified as pedagogically sound in a Physics course is "*Why can't voltage-gated channels be placed on the surface of Myelin?*". A question is classified as not sound if it is vague, unclear, or not about assessing domain knowledge. For example, the question "*What programming language do I need to learn before I start learning algorithms?*" is a valid question, but it is classified as not sound, as it pertains to a course prerequisite rather than assessing domain knowledge.

To make these classifications, we first fine-tuned the GPT-3 model with default parameters on the LearningQ dataset [8]. This dataset is publicly available and contains 5600 student-generated questions from Khan Academy. Each question contains a label to indicate if it is useful for learning or not, as annotated by two expert instructors. No preprocessing was performed on the questions used to fine–tune the model; they were used as-is from the publicly available dataset along with their corresponding binary labels. Fine-tuning the model with default hyperparameters[2] took approximately 10 min and incurred a cost of \$0.21. Next, we passed in the T5-generated questions as the GPT-3 model's input, obtaining the output as a set of binary rating labels.

Expert Evaluation. To further validate the quality of the questions, as well as that of the classification model, we had two expert raters with 5 + years of teaching experience in the domain of data science rate each question. Following the same classification process as in previous work [8], the two raters indicated if each question was pedagogically sound or not. We measured the Inter-Rater Reliability (IRR) between the two raters and found they achieved a Cohen's kappa of $\kappa = 0.425$, with similarity in 75.59% of the question ratings, indicating a moderate level of agreement [23]. The remaining discordant questions were discussed between the two raters until they reached a consensus on their classification.

4 Results

Following the pipeline introduced in Sect. 2, we generated a total of 219 questions across the three header levels - Topic, Module and Unit. 16 questions were removed due to being duplicates[3], leading to a final set of 203 unique questions. Table 1 shows a number of example generated questions, along with their information scores and GPT-3 model evaluation. Among the 203 questions, 151 (74.38%) were classified as pedagogically sound by the GPT-3 model. To compare this classification with the human raters' consensus, which rated 115 (56.7%) questions as pedagogically sound, we constructed a confusion matrix as shown in Table 2. We observed that the model agreed with human raters in 135 (66.50%) instances; in cases where they disagreed, most of the mismatches (52 out of 68) were due to the GPT-3 model overestimating the questions' soundness.

[2] We used the hyperparameter set suggested in https://beta.openai.com/docs/guides/fine-tuning.

[3] With our question generation routine (Fig. 1), the text content in each Topic was used as input three times, which could lead to duplicate questions, even if the accompanying header names were different.

Table 1. Example generated questions across different header levels and soundness ratings.

Generated question	Header level	IS	GPT-3 rating	Expert rating
What is the process of using domain knowledge to extract features from raw data?	Module	0.5	Sound	Sound
What are two types of decision trees?	Topic	0.57	Sound	Sound
What is the tradeoff between bias and variance?	Unit	0.375	Sound	Sound
What is used to evaluate clustering when labeled data is not present?	Module	0.33	Sound	Sound
What are two methods that can be used to improve a regression model?	Unit	0.53	Sound	Sound
What is the term for PCA?	Topic	0.16	Sound	Not sound
What is the main topic of the Data Wrangling module?	Topic	0.2	Not sound	Not sound
What is one of the easiest techniques to implement?	Topic	0.22	Not sound	Not sound
What is the title of the Information Design Unit?	Topic	0	Not sound	Not sound
What is the name of the pattern that is used in the module on regression?	Module	0.2	Not sound	Not sound

Table 2. Confusion matrix for comparing GPT-3 and expert evaluations.

	Expert: Not sound	Expert: Sound
GPT-3: Not sound	36	16
GPT-3: Sound	52	99

We followed up with a qualitative review of the questions rated as not sound by human experts to better understand (1) what separated them from the questions rated as sound, and (2) why the GPT-3 model might still rate them as sound. For (1), we identified two important requirements that a question generally needs to meet to be considered sound by human experts. First, it has to thoroughly set up the *context* (e.g., what is the scenario, how many responses are expected) from which an answer could be reasonably derived. An example question that satisfies this category is "*What are* ***two*** *types of visions that a data science team will* ***work with a client*** *to develop?,*" where the bolded terms are important contextual factors which make the question sound. Without these terms, the question would become "*what are the types of vision that a data science team will develop?,*" which is too ambiguous. We further note that sound questions with thorough contexts tend to be longer, because they necessarily include more information to describe such contexts. At the same time, short questions may still be considered sound by expert raters if they target a sufficiently *specific* concept. For example, "*what is a way to improve a* ***decision tree's performance****?*" is considered sound because the bolded term is very specific. On the other hand, a similar-looking question such as "*what is a way to analyze business data*" is not sound, due to "*analyze business data*" being too broad. It is this second requirement of *specificity* that the GPT-3 model fails to recognize. Many of the questions rated as sound by GPT-3, but not by human raters, are similar to ones such as "*What are two types of data science tasks?,*" which could not be used as a stand-alone assessment question due to a lack of specificity.

Next, we examined whether our IS metric, which calculates the number of important concepts that a question encapsulates, aligns with its pedagogical soundness. Figure 2 (left) shows the distribution of information scores for the questions in each class (pedagogically sound or not), within each type of header level. A one-way ANOVA showed that, among the questions generated with the Topic header names, there was a significant difference in IS between questions rated as pedagogically sound and those rated as not sound by human experts, $F(1, 68) = 8.60, p < .01$. In this case, the pedagogically sound questions ($M = 0.39, SD = 0.14$) had higher IS values than their counterparts ($M = 0.30$, $SD = 0.12$). However, the difference in IS between these two groups was not significant among the questions generated by the Unit header names, $F(1, 66) = 0.07, p = .79$, or the Module header names, $F(1, 63) = 0.41, p = .53$. Figure 2 (right) shows the same distribution based on the GPT-3 model's ratings; in this case, however, the IS between sound and non-sound questions were similar across all three header levels.

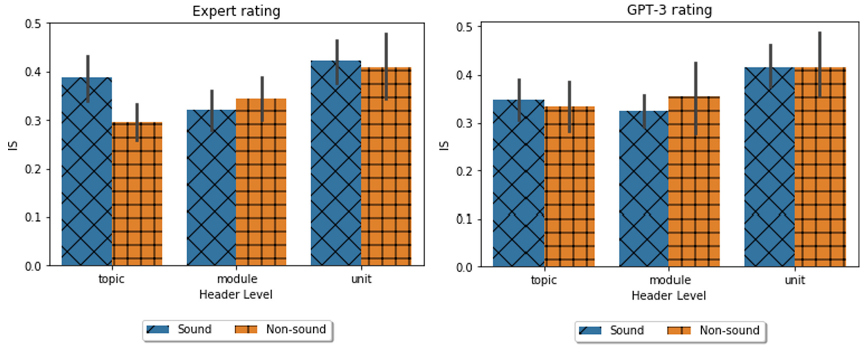

Fig. 2. Distribution of information score at each header level, partitioned by expert ratings (left) and GPT-3 ratings (right).

Finally, we examined which level of header tended to yield the most pedagogically sound questions, based on human ratings. We observed that the number of sound and non-sound questions were respectively 35 and 35 at the Topic level, 37 and 28 at the Module level, and 43 and 25 at the Unit level. Among the sound questions, those generated with the Unit headers were the most common, while those generated with the Topic headers were the least. Conversely, among not-sound questions, those generated with the Topic header were the most common. These distributions suggest that the Topic levels were not suitable for question generation.

5 Discussion

In this work, we propose and evaluate a domain-independent pipeline for generating assessment questions based on instructional materials in an introductory data science course. Our research is motivated by the general lack of practice opportunities in online higher education, as well as the high labor cost in manual question generation, which was reported to be approximately 200 h for one course [35]. Furthermore, the ability to generate questions on-demand can greatly assist adaptive and personalized learning technologies, especially in the context of mastery learning where students are prompted to continue practicing until they reach mastery [33]. To this end, our work makes use of state-of-the-art language models for question generation, concept extraction and question evaluation, in addition to custom scoring metrics and expert labeling as additional validation measures. In general, we found a moderate level of agreement between the three evaluation methods – information score, GPT-3 classification and human judgment – which all rate a high percentage of the generated questions as capturing important concepts or being pedagogically sound. We discuss the features of the generated questions and the possibilities of extending the proposed pipeline as follows.

We saw that the GPT-3 model, fine-tuned on the LearningQ dataset [8], was able to replicate 66.50% of the two expert raters' consensus, which is well above chance. The model appeared to learn that long questions are likely sound, which is a reasonable assumption as these questions might contain more relevant contextual information. However, it also classified a number of short questions as sound, despite the lack of specificity which human evaluators could easily recognize. As the LearningQ dataset did not contain data science questions, it is no surprise that our model was not particularly good at distinguishing between specific data science concepts (e.g., "*decision tree's performance*") and ambiguous ones (e.g., "*business data*"). Additional fine-tuning of the GPT-3 model on a new dataset with questions and expert-generated labels that are closer to our learning domain would therefore be a promising next step.

When treating the expert classification of question soundness as the ground truth labels, we were able to draw a number of comparisons. First, we found that the sound questions generally had higher information score values than those rated as not sound (Fig. 2), suggesting that our rationale for the formulation of these metrics (i.e., that higher scores reflect more concepts captured and therefore higher quality) was justified. Our qualitative review further showed that pedagogically sound questions differ from non-sound questions primarily in their context and specificity. While the current information score metric doesn't capture how specific the terms used in each question are, this task has been explored in previous work [17] and could be incorporated in the next iteration of the question evaluation process in our pipeline. Critically, this evaluation method, which combines concept extraction with information score computation, could be applied in many other learning domains, as it represents a general strategy of identifying high quality and pedagogically sound questions. Second, we found that combining the instructional content with a summary of this content (e.g., the header names) could lead to better question generation with T5. In our case, the header names at the Module and Unit levels were shown to result in more sound questions than those at the Topic level.

At the same time, there are ample opportunities to further promote the adoption of our pipeline across different learning domains. First, more research is needed to investigate question generation when the learning contents are not entirely textual, but may include multimedia components, such as math formulas and images. Recent advances in the area of *document intelligence* [5, 15, 30], combining NLP techniques with computer vision, might be particularly helpful in this direction. Second, there remains the need to diversify the generated questions, so as to meet a wider range of assessment goals. In particular, most of our current questions start with "what" (e.g., those in Table 1), which are primarily geared towards *remembering* information. Incorporating other question types in the generation pipeline could elicit more cognitive processes in Bloom's taxonomy [21] – for example, "how" questions can promote *understanding* and "why" questions are designed for *analyzing* – which in turn contribute to better student learning. This diversifying direction is also an area of active research in the NLP and QG community [40, 44].

In addition, the proposed pipeline is generalizable yet also customizable to individual domains, so as to enable higher quality questions. As previously mentioned, the fine-tuning steps for both T5 and GPT-3 could be carried out on datasets that are closely related to the learning contents and with cross-validated hyperparameter tuning to better fit the dataset. Similarly, the concept extraction process could be enhanced with a combination of machine-generated and human-evaluated skill mappings, which have been shown to yield more accurate knowledge models across several works [25, 26, 38]. Finally, the question evaluation criteria may also benefit from subject matter experts' inputs to closely reflect the distinct nature of the learning domain; for example, chemistry assessments could potentially include both conceptual questions (e.g., "*what is the chemical formula of phenol?*") and scenario-based questions (e.g., "*describe the phenomenon that results from mixing sodium metal and chlorine gas?*").

Finally, we should note the limitations that may influence the interpretation of our results. First, the text input to our T5 model was the content of an entire Topic, consisting of 3–6 paragraphs. Constructing more fine-grained inputs at the paragraph or sentence level could potentially yield more targeted questions, although at the cost of a larger number of questions for human experts to evaluate. This direction could be viable once the evaluation metrics have been refined to more closely replicate expert judgments, allowing them to be applied at scale on large question corpuses. Second, while the human raters' pedagogical soundness ratings provide preliminary evidence of the generated questions' usability, there remains the need to empirically validate their impacts on student learning. To this end, we plan to deploy the pedagogically sound questions identified in this work to formative assignments in the next iteration of the data science course. As shown in prior research [29], the low-stake formats, such as optional quizzes, can still yield crucial insights on student performance while not impacting the overall grades. In this way, they are highly useful for experimenting with new assessment items, especially those not generated by instructors and domain experts, such as in the present study.

6 Conclusion

Our work raises attention to the potential of applying state-of-the-art NLP models in automated question generation. Through applying this process on learning materials from a data science course, we highlight a number of ideas that merit additional investigation in future works. First, we propose an initial method of scoring the quality of automatically generated questions, which provide instructors with the ability to recognize pedagogically sound questions and give the field a baseline to derive comparable methods. Second, we identified the potential of incorporating summary data in the input to QG models, such as Google's T5, to improve the quality of the generated questions. Third, we demonstrated the use of a fine-tuned GPT-3 model in classifying question quality, which in turn serves as a potential feature to add to future models of question quality.

In addition to these contributions, we are also making our full pipeline and results available[4] in hopes of providing a baseline for the community to use and improve on the proposed methods. We believe that achieving generalized, usable methods of automatic question generation will likely require multiple techniques in an ensemble approach to produce content at a sufficiently high quality. Our long term goal is to create generalized QG methods in a widely available open format that use an ensemble of scoring metrics, with the expectation that different metrics will produce better results in different domains. The field of QG for specific educational domains needs a baseline for measuring improvement and we envision this research as a starting point.

References

1. Aguilera-Hermida, A.P.: College students' use and acceptance of emergency online learning due to COVID-19. Int. J. Educ. Res. Open. **1**, 100011 (2020)
2. Ai, R., Krause, S., Kasper, W., Xu, F., Uszkoreit, H.: Semi-automatic generation of multiple-choice tests from mentions of semantic relations. In: Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pp. 26–33 (2015)
3. Alberti, C., Andor, D., Pitler, E., Devlin, J., Collins, M.: Synthetic QA corpora generation with roundtrip consistency. arXiv preprint arXiv:1906.05416 (2019)
4. Amidei, J., Piwek, P., Willis, A.: Evaluation methodologies in automatic question generation 2013–2018 (2018)
5. Baviskar, D., Ahirrao, S., Potdar, V., Kotecha, K.: Efficient automated processing of the unstructured documents using artificial intelligence: a systematic literature review and future directions. IEEE Access **9**, 72894–72936 (2021)
6. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
7. 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, pp. 154–162 (2019)
8. Chen, G., Yang, J., Hauff, C., Houben, G.-J.: LearningQ: a large-scale dataset for educational question generation. In: Twelfth International AAAI Conference on Web and Social Media (2018)

[4] https://github.com/MCDS-Foundations/data-science-question-generation.

9. Cheng, Y., et al.: Guiding the growth: difficulty-controllable question generation through step-by-step rewriting. arxiv preprint arXiv:2105.11698 (2021)
10. Chiu, K.-L., Alexander, R.: Detecting hate speech with gpt-3. arXiv preprint arXiv:2103.12407 (2021)
11. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
12. Dimitrakis, E., Sgontzos, K., Tzitzikas, Y.: A survey on question answering systems over linked data and documents. J. Intell. Inf. Syst. **55**(2), 233–259 (2019). https://doi.org/10.1007/s10844-019-00584-7
13. Du, X., Shao, J., Cardie, C.: Learning to ask: neural question generation for reading comprehension. arXiv preprint arXiv:1705.00106 (2017)
14. Ferrucci, D., et al.: Building watson: an overview of the DeepQA project. AI Mag. **31**(3), 59–79 (2010)
15. Han, B., Burdick, D., Lewis, D., Lu, Y., Motahari, H., Tata, S.: DI-2021: the second document intelligence workshop. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 4127–4128 (2021)
16. Hodges, C.B., Moore, S., Lockee, B.B., Trust, T., Bond, M.A.: The difference between emergency remote teaching and online learning (2020)
17. Huang, H., Kajiwara, T., Arase, Y.: Definition modelling for appropriate specificity. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 2499–2509 (2021)
18. Jia, X., Wang, H., Yin, D., Wu, Y.: Enhancing question generation with commonsense knowledge. In: China National Conference on Chinese Computational Linguistics, pp. 145–160. Springer (2021) https://doi.org/10.1007/978-3-030-84186-7_10
19. Kalman, R., Macias Esparza, M., Weston, C.: Student views of the online learning process during the COVID-19 pandemic: a comparison of upper-level and entry-level undergraduate perspectives. J. Chem. Educ. **97**(9), 3353–3357 (2020)
20. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The knowledge-learning-instruction framework: bridging the science-practice chasm to enhance robust student learning. Cogn. Sci. **36**(5), 757–798 (2012)
21. Krathwohl, D.R.: A revision of Bloom's taxonomy: an overview. Theor. Pract. **41**(4), 212–218 (2002)
22. Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S.: A systematic review of automatic question generation for educational purposes. Int. J. Artif. Intell. Educ. **30**(1), 121–204 (2020). https://doi.org/10.1007/s40593-019-00186-y
23. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
24. Liu, B.: Neural question generation based on Seq2Seq. In: Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence, pp. 119–123 (2020)
25. Liu, R., Koedinger, K.R.: Closing the loop: automated data-driven cognitive model discoveries lead to improved instruction and learning gains. J. Educ. Data Min. **9**(1), 25–41 (2017)
26. Liu, R., McLaughlin, E.A., Koedinger, K.R.: Interpreting model discovery and testing generalization to a new dataset. In: Educational Data Mining 2014. Citeseer (2014)
27. Liu, T., Fang, Q., Ding, W., Li, H., Wu, Z., Liu, Z.: Mathematical word problem generation from commonsense knowledge graph and equations. arXiv preprint arXiv:2010.06196 (2020)
28. Lopez, L.E., Cruz, D.K., Cruz, J.C.B., Cheng, C.: Transformer-based end-to-end question generation. arXiv preprint arXiv:2005.01107, vol. 4 (2020)
29. Moore, S., Nguyen, H.A., Stamper, J.: Examining the effects of student participation and performance on the quality of learnersourcing multiple-choice questions. In: Proceedings of the Eighth ACM Conference on Learning@ Scale, pp. 209–220 (2021)

30. Motahari, H., Duffy, N., Bennett, P., Bedrax-Weiss, T.: A report on the first workshop on document intelligence (DI) at NeurIPS 2019. ACM SIGKDD Explor. Newsl. **22**(2), 8–11 (2021)
31. Novikova, J., Dušek, O., Curry, A.C., Rieser, V.: Why we need new evaluation metrics for NLG. arXiv preprint arXiv:1707.06875 (2017)
32. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
33. Ritter, S., Yudelson, M., Fancsali, S.E., Berman, S.R.: How mastery learning works at scale. In: Proceedings of the Third (2016) ACM Conference on Learning@ Scale, pp. 71–79 (2016)
34. Ruseti, S., et al.: Predicting question quality using recurrent neural networks. In: Penstein Rosé, C., et al. (eds.) artificial intelligence in education. LNCS (LNAI), vol. 10947, pp. 491–502. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93843-1_36
35. Rushkin, I., et al.: Adaptive assessment experiment in a HarvardX MOOC. In: EDM (2017)
36. Sai, A.B., Mohankumar, A.K., Khapra, M.M.: A survey of evaluation metrics used for NLG systems. arXiv preprint arXiv:2008.12009 (2020)
37. Sha, L., et al.: Which hammer should i use? A systematic evaluation of approaches for classifying educational forum posts. Int. Educ. Data Min. Soc. (2021)
38. Stamper, J.C., Koedinger, K.R.: Human-machine student model discovery and improvement using DataShop. In: Biswas, G., Bull, S., Kay, J., Mitrovic, A. (eds.) International Conference on Artificial Intelligence in Education, pp. 353–360. Springer (2011). https://doi.org/10.1007/978-3-642-21869-9_46
39. Steuer, T., Bongard, L., Uhlig, J., Zimmer, G.: On the linguistic and pedagogical quality of automatic question generation via neural machine translation. In: European Conference on Technology Enhanced Learning, pp. 289–294. Springer (2021) https://doi.org/10.1007/978-3-030-86436-1_22
40. Sultan, M.A., Chandel, S., Astudillo, R.F., Castelli, V.: On the importance of diversity in question generation for QA. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5651–5656 (2020)
41. Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., Krahmer, E.: Best practices for the human evaluation of automatically generated text. In: Proceedings of the 12th International Conference on Natural Language Generation, pp. 355–368 (2019)
42. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
43. Wang, S., et al.: PathQG: neural question generation from facts. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9066–9075 (2020)
44. Wang, Z., Rao, S., Zhang, J., Qin, Z., Tian, G., Wang, J.: Diversify question generation with continuous content selectors and question type modeling. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2134–2143 (2020)
45. Xue, L., et al.: mT5: a massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934 (2020)
46. Yu, J., et al.: MOOCCubeX: a large knowledge-centered repository for adaptive learning in MOOCs. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 4643–4652 (2021)
47. Zhang, R., Guo, J., Chen, L., Fan, Y., Cheng, X.: A review on question generation from natural language text. ACM Trans. Inf. Syst. (TOIS) **40**(1), 1–43 (2021)
48. Zhong, R., Lee, K., Zhang, Z., Klein, D.: Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. arXiv preprint arXiv:2104.04670 (2021)
49. Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., Chua, T.-S.: Retrieving and reading: a comprehensive survey on open-domain question answering. arXiv preprint arXiv:2101.00774 (2021)