Towards Automated Generation and Evaluation of Questions in Educational Domains

Shravya Bhat
Carnegie Mellon University
shravyab@andrew.cmu.edu

Huy A. Nguyen
Carnegie Mellon University
hn1@cs.cmu.edu

Steven Moore
Carnegie Mellon University
stevenmo@andrew.cmu.edu

John Stamper
Carnegie Mellon University
jstamper@cs.cmu.edu

Majd Sakr
Carnegie Mellon University
msakr@cs.cmu.edu

Eric Nyberg
Carnegie Mellon University
ehn@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 this work, 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 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 classified the generated questions as either useful to learning or not with two different approaches: automated labeling by a trained GPT-3 model and manual review by expert human judges. 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 NLP research in educational domains.

1. INTRODUCTION

As education across grade levels continues to transition towards online platforms in response to the COVID-19 pandemic, the need for effective and scalable assessment tools emerges as a pressing issue for instructors and educators. Amid many other logistical issues that arise from emergency online education [5], 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. Recent advances in natural language processing (NLP), question answering and question generation (QG) offer a promising path to accomplishing this goal. 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 [7]. 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 Text-To-Text Transfer Transformer (T5) models [15] 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 useful to learning or not. Finally, we had two data science instructors perform this same classification task manually. Our results contribute 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. METHODS

2.1 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 corresponding to a data science topic 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

S. Bhat, H. Nguyen, S. Moore, J. Stamper, M. Sakr, and E. Nyberg. Towards automated generation and evaluation of questions in educational domains. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 701–704, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.6853085

questions from the textual content of the sixteen modules in the course, using the following pipeline.

2.2 Question Generation Pipeline

First, we extracted 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 module and unit each paragraph belongs to). We scraped the text content from the XML files using the BeautifulSoup[1] library and cleaned the content to remove leading questions, such as *"What does this accomplish"* and *"Why would this make sense?"*. These questions were included to help students navigate the reading more effectively but do not contain meaningful information on their own. From this point, the resulting text data was input to two separate processes as follows.

Concept Hierarchy Extraction. This process was carried out by the MOOCCubeX pipeline [16], 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 [15], 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 language tasks by merely changing the structure of the input passed to it. For our use case, the input data is the cleaned text content prepended by a header of the text. Our rationale for including the header 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. There were three hierarchical levels of header that were used in our input: Unit, Module and Title, where the former encompasses the latters. For example, the unit *Exploratory Data Analysis* includes the module *Feature Engineering*, which has a section titled *Principal Component Analysis*, among others. Before applying the model to our dataset, we also fine-tuned it on SQuAD 1.1, a well known reading comprehension dataset and a common benchmark for question-answering models [11].

2.3 Evaluation

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

Information Score. This is a custom metric that denotes 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,

$$IS(q) = \frac{1}{|T(q)|} \sum_{t \in T(q)} 1(t \in C), \tag{1}$$

where the division by $|T(q)|$ is for normalization. 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 model as it has been a popular choice for text classification tasks such as detecting hate speech [3] and text sentiment [17]. Our classification task involves rating each generated question as either *useful for learning* or *not useful*. A useful-for-learning question is one that pertains to the course content and is intended to assess the domain knowledge of the student. On the other hand, a question is classified as not useful 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 useful for learning because it pertains to a course prerequisite rather than domain knowledge assessment. To perform this classification, we first fine-tuned the GPT-3 model with default hyperparameters on the LearningQ dataset [2], which 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. Next, we passed in the T5-generated questions as the GPT-3 model's input, obtaining the output as a set of binary labels indicating if it rated each question as useful for learning or not.

Expert Evaluation. To further validate the question quality, 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 [2], the two raters indicated if each question was useful for learning 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 [9]. The remaining discordant questions were discussed between the two raters until they reached a consensus on their classification, resulting in all of the generated questions being classified by both human judges and the GPT-3 model.

3. RESULTS

Following the above pipeline, we generated a total of 203 questions across the three header levels - Module, Unit, and Title. The Appendix 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 useful for learning by the GPT-3 model. To compare this classification with the human raters' consensus, we constructed a confusion matrix as shown in Table 1. We observed that the model agreed with human raters in 135 (66.50%) instances; in cases where they disagreed, most of

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

the mismatches (52 out of 68) were due to the GPT-3 model overestimating the questions' usefulness.

	Expert: 0	Expert: 1
GPT-3: 0	36	16
GPT-3: 1	52	99

Table 1: Confusion matrix for comparing GPT-3 and expert ratings on the generated questions. 0 denotes Not Useful and 1 denotes Useful rating.

We followed up with a qualitative review of the questions rated as not useful by human experts to better understand (1) what separated them from the questions rated as useful, and (2) why the GPT-3 model might still rate them as useful. For (1), we identified two important requirements that a question generally needs to meet to be rated as useful 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 useful. We further note that useful 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 useful 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 useful 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 useful, due to *"analyze business data"* being too broad. The GPT-3 model typically fails to recognize this specificity criterion – many of the questions rated as useful by GPT-3, but not by human raters, are similar to ones such as *"What are two types of data science tasks?,"* which could be useful if *"data science tasks"* was replaced with a more targeted concept.

Next, we examined whether our score metric, which calculates the normalized number of important concepts that a question encapsulates, aligns with the expert classification of question usefulness for learning. We observed from Figure 1 that, across the three header levels, questions rated as useful tended to have similar or higher information scores than their counterparts.

4. DISCUSSION AND CONCLUSION

In this work, we propose and evaluate a domain-independent pipeline for generating assessment questions based on reading materials in a data science course. Our results showed that the GPT-3 model, fine tuned on the LearningQ dataset [2], was able to reach an acceptable level of agreement (on 66.50% of the questions) with the consensus of two expert raters. The model appeared to learn that long questions are likely useful, which is a reasonable assumption as these questions might contain more relevant contextual information. However, it also classified some short questions as useful, 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 differentiating 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 labeled dataset closer to our learning domain would therefore be a promising next step.

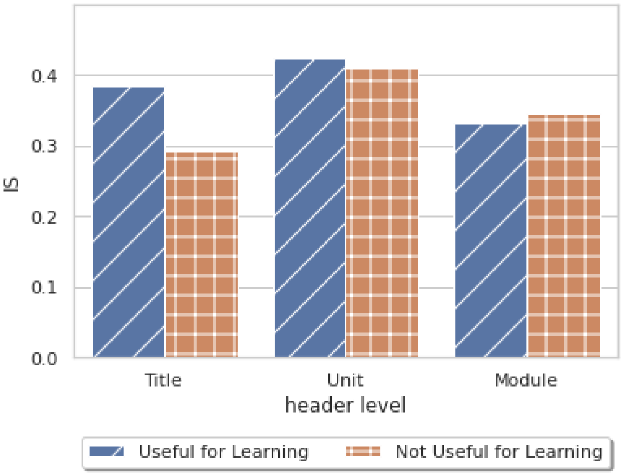

Figure 1: Distribution of information score, partitioned by expert raters' evaluation, at each header level.

When treating the expert rating of question usefulness as the ground truth, we found that the useful questions generally had higher information scores than those not rated as useful, 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. At the same time, several questions had relatively low information scores but were still rated as useful by experts (e.g., *"What are two types of decision trees?"*) because they target a sufficiently *specific* concept. To detect these questions, it would be beneficial to incorporate measures of the generated questions' level of specificity [6] into the existing information score metric.

The above results have been obtained without the need for any human-labeled domain encoding, which makes our question generation pipeline highly domain-agnostic and generalizable. At the same time, there are ample opportunities to further promote its adoption 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. Recent advances in the area of document intelligence [1,4], combining NLP techniques with computer vision, could be helpful in this direction. Second, there remains the need to diversify the generated questions to meet a wider range of assessment goals. In particular, most of our current questions start with "what" (e.g., those in Table **??**), which are primarily geared towards recalling information. Incorporating other question types in the generation pipeline could elicit more cognitive processes in Bloom's taxonomy [8] – for example, "how" questions can promote understanding and "why" questions are designed for analyzing – which in turn contribute to better learning overall. This diversifying direction is also an area of active research in the NLP and

QG community [13,14].

We further note that the proposed pipeline is also customizable to individual domains, so as to enable higher quality questions. First, hyperparameter tuning on a dataset relevant to the learning domain would likely improve the performance of the T5 and GPT-3 models. Second, the concept extraction process could be enhanced with a combination of machine-generated and human-evaluated skill mappings, which have been shown to result in more accurate knowledge models [10,12]. 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?"*).

5. REFERENCES

[1] D. Baviskar, S. Ahirrao, V. Potdar, and K. Kotecha. Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions. *IEEE Access*, 2021.

[2] G. Chen, J. Yang, C. Hauff, and G.-J. Houben. Learningq: a large-scale dataset for educational question generation. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[3] K.-L. Chiu and R. Alexander. Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*, 2021.

[4] B. Han, D. Burdick, D. Lewis, Y. Lu, H. Motahari, and S. Tata. Di-2021: The second document intelligence workshop. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4127–4128, 2021.

[5] C. B. Hodges, S. Moore, B. B. Lockee, T. Trust, and M. A. Bond. The difference between emergency remote teaching and online learning. 2020.

[6] H. Huang, T. Kajiwara, and Y. Arase. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, 2021.

[7] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798, 2012.

[8] D. R. Krathwohl. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

[9] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[10] R. Liu and K. R. Koedinger. Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains. *Journal of Educational Data Mining*, 9(1):25–41, 2017.

[11] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[12] J. C. Stamper and K. R. Koedinger. Human-machine student model discovery and improvement using datashop. In *International Conference on Artificial Intelligence in Education*, pages 353–360. Springer, 2011.

[13] M. A. Sultan, S. Chandel, R. F. Astudillo, and V. Castelli. On the importance of diversity in question generation for qa. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, 2020.

[14] S. Wang, Z. Wei, Z. Fan, Z. Huang, W. Sun, Q. Zhang, and X.-J. Huang. Pathqg: Neural question generation from facts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9066–9075, 2020.

[15] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

[16] J. Yu, Y. Wang, Q. Zhong, G. Luo, Y. Mao, K. Sun, W. Feng, W. Xu, S. Cao, K. Zeng, 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*, pages 4643–4652, 2021.

[17] R. Zhong, K. Lee, Z. Zhang, and D. Klein. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. *arXiv preprint arXiv:2104.04670*, 2021.