Generation of Student Questions for Inquiry-based Learning

Kevin Ros, Maxwell Jong, Chak Ho Chan, ChengXiang Zhai

Department of Computer Science, The Grainger College of Engineering, University of Illinois at Urbana-Champaign

Introduction

Asking questions during an in-person, synchronous lecture is a central part of a traditional classroom setting.

Such a process helps both students and instructors:

1. Instructors can get implicit feedback on the difficult aspects of their lecture content, and they are able to update the content accordingly.
2. Students can test their understanding of the material by hearing other students' questions.

This is possible, in part, due to the coupling between asked questions and lecture content, as everyone shares the same context.

However, in online, asynchronous lectures, students generally engage with the course material alone.

Questions are often asked independently via search engines. As a consequence, other students and instructors cannot use the questions to their implicit benefit.

Discussion forums help mitigate this, but the barriers to posting are often higher than to asking a question in a live lecture, and the coupling is not consistent, resulting in a possible lack of context.

How can we recover the benefit of in-person question-asking for online classrooms?

Aim and Contributions

Automatically generating plausible student questions given a lecture transcript would help recover some of the benefit of in-person question-asking. The aim of this work is to provide an initial exploration of student question generation.

To this end, we:

1. Release a new data set (the Lecture-Question data set) from an online, asynchronous classroom where real student-asked questions are coupled to the lecture windows when they were asked.
2. Explore the feasibility of generating student questions using these identified lecture windows.

In general, we hope that this work:

1. Encourages online instructors to collect questions in this manner for future benefit and analysis.
2. Encourages research on the questions that are asked about the content, not on the questions that are answerable by the content.

The Lecture-Question Data Set

Students from an online, asynchronous class were asked, if they had a question, to submit the question, along with a timestamp window corresponding to the point in the lecture where they had the question.

Questions were submitted in the following format:

<Lecture name, start time, end time, question>

The class used two MOOCs as the primary lectures:

1. Text Retrieval and Search Engines
2. Text Mining and Analytics

The data collection process resulted in a total of 536 questions over 90 lectures, which we release on GitHub.

In contrast with many existing question data sets, our questions are *about* the lecture content, rather than *answerable* by the content.

Examples of Student Questions
What is the point of compression? Will the access times really be that impactful to the overall indexing?
Are the doc-ids sorted with the term-ids in the "local" sort?
Can we get more examples of using gamma-code?
I'm still very confused how integer compression actually reduces size of storage since some of the examples make it seem like you're using more bits than before on some inputs

Question Generation

Goal: Given a start and end time of a lecture window, generate the question asked during that window.

Due to the low amount of data, we explored two low-data techniques with the T5 language model - a pre-trained generative language model based on the Transformer [1].

Research Question #1: How does pre-training on search engine query generation affect student question generation performance?

To answer this, we compared t5-base to docTTTTTquery [2], a version of T5 fine-tuned to generate search engine queries given the ground truth passage.

Research Question #2: How does continuous prefix tuning affect student question generation performance?

To answer this, we implemented prefix tuning on top of T5 and docTTTTTquery. Prefix tuning is a method for fine-tuning a generative language model by freezing the original parameters and adding a learnable prefix that is prepended to every input [3].

Results

The table below presents the recall, precision, and F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L measurements for the question generation approaches on the test set (53 questions). "FT" refers to traditional fine-tuning and "Prefix" refers to continuous prefix tuning. Each model was selected based on the best performance across the cross validation folds.

Model	Recall	Precision	F1
	ROUGE-1 (%)		
t5-base FT	20.06	14.47	14.82
t5-base Prefix	20.13	21.56	18.63
docTTTTTquery	14.41	25.17	16.83
docTTTTTquery FT	15.70	23.34	17.45
docTTTTTquery Prefix	17.19	24.00	18.74
	ROUGE-2 (%)		
t5-base FT	1.679	1.656	1.502
t5-base Prefix	3.267	3.391	3.043
docTTTTTquery	3.237	4.596	3.358
docTTTTTquery FT	4.011	4.730	3.903
docTTTTTquery Prefix	4.790	6.247	5.010
	ROUGE-L (%)		
t5-base FT	15.82	15.57	11.77
t5-base Prefix	16.89	17.65	15.47
docTTTTTquery	13.17	22.32	15.18
docTTTTTquery FT	14.34	20.64	15.76
docTTTTTquery Prefix	15.47	21.00	16.73

Some examples of generated questions are as follows:

Model	Question
Ground Truth	Does is the delta-code use gamma-code twice recursively?
t5-base FT	What is the difference between delta coding and delta coding? Is it possible to use delta coding for inverted index distribution?
t5-base Prefix	What is the difference between delta and gamma?
docTTTTTquery	what is gamma coding
docTTTTTquery FT	what is the difference between delta and delta coding?
docTTTTTquery Prefix	what is the difference between delta and gamma coding?

Research Question #1: Pre-training on search engine query generation appears to offer clear benefit in increasing the precision, though the benefit appears to be more for traditional fine-tuning.

Research Question #2: There seems to be marginal benefit for using continuous prefix tuning in a low-data setting to generate student questions.

We can only draw preliminary conclusions from the findings due to the small data set size and due to some runs having insignificant differences.

Overall, student question generation from lecture context appears to be a difficult task, but existing generative models show promise towards generating meaningful questions from limited data.

Conclusions

In this work, we explored the feasibility of generating student questions from selected lecture windows using explicitly marked locations in lectures. We also released a new data set for this exploration, and we plan to expand it in the future.

We found the generation task to be promising but difficult, and we can only draw preliminary conclusions on model differences due to the small data set size.

Future Work

Some general areas for future exploration include:

- Testing more controlled or few-shot generation approaches.
- Evaluating question generation effectiveness by direct student evaluation.
- Examining how question utility differs by context, student, or instructor.
- Expand the contextual information provided to the generation models (e.g. from different lectures).
- General practical applications, such as integrating generated questions directly into classrooms.

Corresponding author: Kevin Ros, kjros2@illinois.edu

References

[1] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

[2] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019a. From doc2query to docTTTTTquery. Online preprint.

[3] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190.

Acknowledgements

This work is supported in part by the National Science Foundation under Grant No. 1801652. We would also like to thank Dr. Heng Ji for her insightful comments.

I ILLINOIS