VitalSource

Toward Effective Courseware at Scale: Investigating Automatically Generated Questions as Formative Practice

This is an overview of the research investigating the use of artificial intelligence to generate Learn-by-Doing more affordably in courseware by the VitalSource Research and Development team as presented at the L@Scale '21 Virtual Event from June 22-25, 2021. You can read the entire paper here.

INTRODUCTION

If you follow trends in online education, you already know that online courseware can, when used well, provide great learning gains for students. Full courseware provides students with content to learn from, but also opportunities to check and apply their knowledge as they learn with formative practice. Unfortunately, courseware can be costly to create. Each learning opportunity must be crafted individually.

Recent advances in artificial intelligence (AI), however, now afford the ability to create these learning opportunities automatically and at scale. There are many ways in which AI can be used to tackle this problem, and it is therefore important that we analyse the quality of these practice questions. Many existing studies show the effectiveness of practice, but mostly that research involves human created practice for obvious reasons. We wanted to know if automatic question generation (AQG) can do the work of providing students practice opportunities as well as those written by content authors. If they do, effective courseware suddenly becomes much more broadly accessible to students who otherwise might have only a text-based resource.

As AI driven question generation is a relatively recent field of study, most existing research happens in a lab setting with analysis done on a small number of questions. What we wanted to know was how well the automatically generated (AG) questions perform with students in a natural learning environment, when compared to human-authored (HA) questions. Our research question was, "Are student interactions with AG questions equivalent to HA questions with respect to engagement, difficulty, and persistence metrics?" In order to answer that, we looked at data from 109 students who worked through courseware which contained both AG questions and HA questions. Students were not aware any of the questions were generated; they were simply learning material as they needed for their regular class.

KEY TERMINOLOGY

Automatic question generation (AQG): a method of creating formative practice questions at scale within courseware using artificial intelligence to minimise investments in human time and cost

Automatically generated (AG) questions: formative practice questions created by natural language processing and artificial intelligence using the course's textbook as source material

Human-authored (HA) questions: formative practice questions created manually by an individual and taken from the textbook's ancillary materials or written by subject matter experts

Recall question types: questions that require students to fill in a missing word rather than select from a fixed group of choices (in this study: AG or HA fill-in-the-blank questions)

Recognition question types: questions that require students to evaluate provided terms or concepts and select an answer (in this study: AG or HA matching; HA multiple choice, multiple choice multi-select, multiple choice grid, drag and drop, and pulldown)

DATA

109 students in a university Communication course	**263** automatically generated questions	**390** human-authored questions	**20,990** observations

WHAT DID WE FIND?

The data we have shows that ***our first generation of generated questions performed just as well as the questions written by authors.***

Of course, not every question is the same. Even in carefully crafted courseware some questions turn out to be "tricky" or too hard for students and only putting those questions out into the field and collecting data on them will tell you that. Similarly, multiple choice questions where a student is picking an answer from a list (a recognition question) is typically easier than one where the student needs to come up with the answer on their own (a recall question). Although there are always exceptions to the rule, we see the same thing in our data. Whether or not a question was generated or authored, recognition questions are similarly less difficult than recall questions. But the nature of how the question was created – by AI or by a person – does not change the results.

GETTING INTO THE DETAILS

Here is an excerpt from our research that summarises what we found:

- Levels of engagement, difficulty, and persistence with AG questions and HA questions in the same courseware used by the same students were found to be largely equivalent.
- While there were differences among results for individual question types, there was no evidence that students preferred HA over AG questions.
- The format of a question (recognition vs. recall) had the greatest impact on initial engagement, and that difficulty had an impact on persistence, a metric which has been shown to have a strong impact on learning.

Q: Will students engage with generated questions in the same way they do with authored questions?

A: Yes, they will! In no way did students seem to engage in the questions differently based on their source.

Q: Are generated questions too easy or too hard as compared to authored questions?

A: No! Within our ability to measure difficulty based on student data, you could not make any generalisations about either source of questions being too easy or hard for students.

Q: Are students more likely to keep working until they reach the right answer if the question was authored?

A: Great news! Students generally want to reach the right answer no matter of the source of the question. In our data they persisted to reach the right answer well over 90% of the time no matter what.

Van Campenhout, R., Brown, N., Jerome, B., Dittel, J. S., & Johnson, B. G. (2021). Toward Effective Courseware at Scale: Investigating Automatically Generated Questions as Formative Practice. Learning at Scale. pp. 295–298. https://doi.org/10.1145/3430895.3460162