Crowdsourcing Explanations for Improving Assessment Content and Identifying Knowledge Components

Steven Moore, Carnegie Mellon University, StevenJamesMoore@gmail.com
Huy Nguyen, Carnegie Mellon University, hn1@andrew.cmu.edu
John Stamper, Carnegie Mellon University, john@stamper.org

Abstract: Refining assessment items to improve their clarity and identifying the intended knowledge components required to solve them is a time-consuming task. In this study, we present the results of crowdsourcing insights into the underlying concepts of problems in mathematics and English writing, as a step towards leveraging the crowd to expedite the task. This work demonstrates a method to use the crowd's knowledge that can lead to knowledge component identification and improved assessments.

Introduction

Intelligent tutoring systems and other adaptive courseware often employ knowledge component modeling, which treats student knowledge as a set of interrelated knowledge components (KCs), where each KC is "an acquired unit of cognitive function or structure that can be inferred from performance on a set of related tasks" (Koedinger et al., 2012). Operationally, a KC model is defined as a mapping between each question item and a hypothesized set of associated KCs that represent the skills or knowledge needed to solve that item. This mapping is intended to capture the student's underlying cognitive process and is vital to many core functionalities of an intelligent educational software, enabling features such as adaptive feedback and hints (Moore & Stamper, 2019). Once student data has been collected, the initial mapping can then be improved as poorly associated KCs come to light (Corbett & Anderson, 1994). As the next step, instructional designers are often leveraged to revise the KC model; however, this is often a time consuming task, making this continuous iteration challenging. While machine learning methodologies have been developed to assist in the automatic identification of new KCs, prior research has shown that human judgment remains critical in the interpretation of the improved model and acquisition of actionable insights (Liu & Koedinger, 2017; Nguyen et al., 2019).

An emerging area that has the potential to provide the human resources needed for scaling KC modeling is crowdsourcing, although the challenge with this approach is that the population of crowdworkers is highly varied in their education level and domain knowledge proficiency. Therefore, as a first step towards examining and promoting the feasibility of crowdsourced KC modeling, we studied how crowdworkers can provide insights into identifying KCs in a set of word problems. Using a crowdsourcing platform, we gathered participants with no background in pedagogical training or learning sciences and varying levels of math and English writing expertise. We then asked them to provide explanations of what makes a problem challenging, particularly for questions involving geometry and prose writing style. Based on their responses, our research questions are as follows:

RQ1: Are the explanations provided by crowdworkers indicative of any KCs that the problems require?
RQ2: Do the explanations provide insights into how the presented assessment items may be improved?

Methods

Our study consists of two experiments with the same procedure but involve different domain knowledge. The first domain is mathematics, with a focus on the area of shapes; the second is English writing, with a focus on prose style involving agents and clause topics. In both domains, we deployed an experiment using AMT. Forty crowd workers on AMT, known as "turkers," completed the math experiment, and thirty turkers completed the writing experiment, for a total of 70 participants. In each domain, the tasks took roughly five minutes to complete. Participants were compensated $0.75 upon completion, providing a mean hourly wage of $9.

Participants completed a series of demographic questions about gender, education level, and occupational field. They were then asked two questions regarding their expertise in the given domain, either English writing or math. Following this, they moved onto the main task of the experiment, where they were presented with two word problems positioned next to one another. In the math experiment, both of these problems involved finding the area of two different shapes, composed of squares and triangles. In the writing experiment, both word problems involved identifying the agents and actions of different sentences. Across both experiments, participants were asked to compare the two side-by-side questions and provide three explanations as to why students might find one problem more challenging than the other.

The participant provided explanations were then coded, using a unique codebook for each experiment, by two researchers following the process of DeCuir-Gunby et al. (2011). In total 11 codes were created and applied to the 120 explanations from the math experiment and 10 different codes were created and applied to the 90 writing experiment explanations. Due to space limitations, the codebook can not be shared here, but the results extrapolate on the codes that were found to be the most meaningful.

Results

From the coded explanations in the math experiment, six of the eleven codes were applied to explanations that were indicative of a KC that was fitting to the problems. These codes involved the number of steps a problem took, the layout of the depicted shapes, and the need to calculate different areas. Additionally, two other codes from the math experiment were tagged to explanations that indicated an area of question improvement. These often involved the wording being confusing or the shapes used in the problems being hard to decipher. In total 90/120 (75%) of the participant explanations were suggestive of at least one KC and 15/120 (12.5%) of the explanations suggested an area of the problem(s) that could be improved.

The writing experiment yielded slightly lower results, as only four of the ten codes applied to participant explanations were ones indicative of a KC that was fitting the problems. These four codes covered explanations that detailed the grammatical rules of a sentence and the need for technical jargon to identify the different parts of the word problem. Three other codes of the ten were tagged to explanations that indicated an area of question improvement, akin to the math experiment ones. These often detailed the format of the question, being open ended or multiple choice, and the wording of one of the questions being not as understandable. In total, 20/90 (22.22%) of the participant explanations were suggestive of at least one KC in the writing experiment and 42/90 (46.67%) of them suggested an area of the problem(s) that could be improved.

Discussion

The results indicate that many of the provided explanations were relevant to the problems more often than not, either indicating a KC required to solve the problem or suggesting an area of improvement. Understandably the math experiment achieved a greater number of explanations indicative of a KC than the writing one, likely due to the domain being more familiar to participants. These explanations that suggested a KC required to solve the problems could be leveraged to help create a KC model or as a starting point for experts to begin developing such a knowledge mapping. Across both experiments, participants were also able to suggest more surface level features of the problems, such as confusing wording or confusing images, that could lead students to have difficulty in solving them. These explanations could be leveraged by course designers to indicate where their efforts should be spent in correcting assessments items. They also may lead to clarification as to why students are struggling with a particular set of problems, not due to the content, but due to the question semantics.

References

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278.

DeCuir-Gunby, J.T., Marshall, P.L. and McCulloch, A.W. 2011. Developing and using a codebook for the analysis of interview data: *An example from a professional development research project*. Field methods. 23, 2 (2011), 136–155.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, *36*(5), 757–798.

Liu, R. and Koedinger, K.R. 2017. Closing the Loop: Automated Data-Driven Cognitive Model Discoveries Lead to Improved Instruction and Learning Gains. Journal of Educational Data Mining. 9, 1 (2017), 25–41.

Moore, S., & Stamper, J. (2019). Decision Support for an Adversarial Game Environment Using Automatic Hint Generation. *International Conference on Intelligent Tutoring Systems*, 82–88. Springer.

Nguyen, H., Wang, Y., Stamper, J., & McLaren, B. M. (2019). Using Knowledge Component Modeling to Increase Domain Understanding in a Digital Learning Game. *Proceedings of the 12th International Conference on Educational Data Mining*, 139–148.