Distributed Representation of Misconceptions

Zachary A. Pardos, University of California, Berkeley, zp@berkeley.edu
Scott Farrar, Khan Academy, scottfarrar@gmail.com
John Kolb, University of California, Berkeley, jkolb@berkeley.edu
Gao Xian Peh, University of California, Berkeley, pehgaoxian@berkeley.edu
Jong Ha Lee, University of California, Berkeley, jonghalee@berkeley.edu

Abstract: Tutoring systems deployed at scale present an opportunity to reinvigorate the study of how misconceptions or partial understanding develops in a wide range of STEM domains by connecting to critical pedagogical theories from the learning sciences by way of a distributed representation of the learner. Using answer sequence data from three Khan Academy exercises, we generate high-dimensional vector representations of incorrect student answers using a model of distributed representation more commonly applied to natural language. After clustering wrong answers in the learned vector space, we use these clusters as the basis for analysis of student misconceptions with a quantitative comparison to manual coding and a deeper qualitative discussion based on a constructivist framework. The result is a demonstration of how big data from conventional tutoring systems can act as a bridge to more critical pedagogies from the learning sciences via a distributed, connectionist model of student concept formation.

Keywords: misconception analysis, distributed representation, big data, tutoring systems

Introduction

The computational cognitive sciences and learning sciences have come from different pedagogical points of view on the framing and intervening on misconceptions (Anderson, 1993; Piaget, 1952; Smith, DiSessa, & Roschelle, 1994). Big data from tutoring systems paired with rich models of representation can serve as a bridge between the two, demonstrating the utility of scale while being more amenable to representing theories from the learning sciences and empirically testing them. In this paper, we present an approach to algorithmically cull together common wrong answers that share a misconception across instances of variablized problems. The intuition is that a misconception, exhibited by a particular answer on a problem, can distinguish itself by the answers given by a student to problems immediately before and after that particular answer. Those adjacent answers may themselves belong to their own misconception groups. This distribution of groups possibly comprises the signature of a misconception. A skip-gram model, which creates vector representations that capture a contextual distribution, is used to position problem instantiations into a vector space in which we find that common wrong answers of the same misconception are clustered together with considerable fidelity. We present an algorithm aided misconception analysis of three fraction arithmetic exercises in Khan Academy (www.khanacademy.org) and provide a discussion of the results and designs for potential future interventions.

The manual process of diagnosing the misconception associated with a common wrong answer is a laborious one and can be made more difficult if several hypotheses exist for how an answer was generated. Adding to the challenge of misconception analysis in tutoring systems is the use of templatized or variablized problems. This variablization, where a single problem template can create hundreds of instances with different numbers filled in, is often used to (1) allow students to continue to practice a particular skill without exhausting the available item pool and (2) reduce the possibility for cheating to occur in certain contexts. The challenge this poses to misconception analysis is that instead of having all students contribute to a single distribution of common wrong answers for a problem, that distribution is split across the hundreds of instantiations of the now variablized problem, potentially negating the benefit of the scale of data afforded by a widely used tutoring platform. Additionally, the differences in the numbers used in each instantiation changes the distribution of common wrong answers, such that the proportion of each misconception is not equal across instances. Having the ability to cluster together common wrong answers of a shared misconception across different instantiations of a problem would address these challenges. Furthermore, it would allow a tutoring platform to adapt instruction to each cluster as opposed to each individual answer. Without this ability, a platform is limited to addressing the misconceptions, and their respective common wrong answers, that teachers and instructional designers can anticipate and proceduralize help for.

Related work

Misconception research has roots in the cognitive development theories of Piaget (1952, 1962), which argue all

persons' continual adaptations to new knowledge can produce "systematic errors". The person may assimilate and accommodate new information that develops a conception sufficient for their current needs yet is inconsistent with more advanced knowledge from the mathematical community. Erlwanger's (1973) case study on a student developing a partially consistent-- yet incorrect-- set of mathematical knowledge highlighted the fact that misconceptions are durable even in the face of repeated negative reinforcement from an individualized and automatic instructional system.

From the disciplinary angle of the computational cognitive sciences, misconceptions have been cast as *buggy rules*, and have been foundational to modern Intelligent Tutoring Systems (Anderson, Corbett, Koedinger, & Pelletier, 1995; Zinn, 2006). In this field buggy rules represent incorrect variations of the correct reasoning processes in the ideal model (Brown & Burton, 1978; Sleeman & Brown, 1982). Previous studies of buggy rules largely revolved around student learning processes in programming (Putnam, Sleeman, Baxter, & Kuspa, 1986; Reiser, Anderson, & Farrell, 1985) and mathematics (Jurkovic, 2001; Milson, Lewis, & Anderson, 1990). In programming problems, a buggy rule may represent various student misconceptions such as semantic confusions between functions (VanLehn, 1990). In a similar fashion, buggy rules have also been studied in mathematical problems (Star, 2005) such as signed subtraction (Tatsuoka, 1985) and basic algebra (Milson et al., 1990).

Tutoring systems assume a set of ideal rules that produce correct and consistent answers, and efforts have been made towards cataloguing collections of buggy rules that could explain incorrect answers. These buggy rules could represent misconceptions that students often have during the learning process (Brown & VanLehn, 1980). This large collection of buggy rules has been referred to as a bug catalogue (Johnson & Soloway, 1984). The bug catalogue enables tutoring systems, such as ASSISTments (Razzaq et al., 2009), to model a student's path through a problem, which enables it to respond to common wrong answers that a student may express with their answer. In the spirit of exploring methodologies for improving this catalogue and its association with answers and with skills (Birenbaum, Kelly, & Tatsuoka, 1992) or knowledge components (Barnes, 2005) through the analysis of big data, Liu, Patel & Koedinger (2016) explored how the incorporation of buggy rules could improve model fit to dichotomous response data in an algebra tutor.

Smith, DiSessa, & Roschelle (1994), in contrast, made an explicit call to shift attention away from only cataloguing misconceptions, as education theorists noted conflicts between the learning theory of constructivism and the assumptions of misconceptions as thoughts to be replaced (Sfard & Cobb, 2014). Smith et al. argue that misconceptions should not be seen as knowledge to be confronted and replaced but rather as knowledge to be appreciated and developed in the student. "Persistent misconceptions, if studied in an evenhanded way, can be seen as novices' efforts to extend their existing useful conceptions to instructional contexts in which they turn out to be inadequate." For example, learners may *productively* order integers ($125 > 99$) in such a way that their strategy produces a correct answer. However, the same strategy may produce an incorrect answer in another setting ($1.25 > 9.9$), leaving the learner confused about keeping or discarding their conception of ordering. Instead of a *confrontation* between keeping and discarding a conception, Smith et al. (1994) pressed for the importance of *developing* and refining the student's conceptions through reflection and discussion. Teachers have increasingly incorporated misconception-based strategies in their classrooms from the 1980s to present day (Franke et al., 2015; Sfard & Cobb, 2014; Watkins, Hammer, Radoff, Jaber, & Phillips, 2017).

An assessment-driven approach in most tutoring systems have struggled to represent partial understanding to substantial effect (Ostrow, Donnelly, Adjei, & Heffernan, 2015). With the advent of big data and reemergence of the application of connectionist models (Pardos, 2017), progress has been made towards a more relational and structural representation of students' cognitive (Piech et al., 2015) and behavioral (Tang, Peterson, & Pardos, 2017) states during the learning process. Berland, Baker, & Blikstein (2014) allude to a bridge, asserting that assessment has much to gain from interdisciplinary cooperation.

Methodology

Selecting exercises from the Khan Academy dataset

We develop our representation learning approach to misconceptions using anonymized data granted from Khan Academy. Subjects on Khan Academy are broken into groups of *exercises*, each of which contain questions on a specific concept for students to master, loosely analogous to knowledge components (Piech et al., 2015).

Our work involves analysis of three exercises chosen based on the criteria that (1) answers to questions in these exercises could be reliably parsed from the data logs and (2) that the exercises represented topics in which misconceptions had been studied by past learning sciences work. The exercises were, "*Adding and Subtracting Fractions with Unlike Denominators*", "*Understanding Multiplying Fractions with Whole Numbers*", and "*Multiplying Unit Fractions and Whole Numbers*" (see Table 1 for descriptive stats). Within each exercise, instructional designers create several problem templates, dubbed *problem types*. These are questions with generic

parameters that are instantiated based on a random *seed*. Khan Academy randomly selects problems from different problem types and seeds within an exercise to present to students as they interact with the system.

Table 1: Khan Academy Dataset Statistics

-	**Adding and Subtracting Fractions with Unlike Denominators**	**Multiplying Unit Fractions and Whole Numbers**	**Understanding Multiplying Fractions and Whole Numbers**
# Student Answers	103,873	78,369	134,590
# Users	24,411	21,923	36,968
# Problem Types	4	3	5
# Seeds	23	19	18

Skip-gram model

The skip-gram (more popularly known as "word2vec") was initially developed to model language (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), vectorizing words and learning syntactic and semantic relationships among them automatically from a large corpus of text. In this work, we apply the technique not to textual sequences of words but rather to chronological sequences of student answers, reasoning that regularities in student answer patterns can surface insights about each answer, much like a traditional application of skip-grams to natural language surfaces insights about words and their relationships to one another. A skip-gram is a simple neural network, or "connectionist" model (Hinton, 1990), taking an element in a sequence as an input and predicting which elements occurred adjacent to that element in the sequence (or within a specified context window length). It is similar in topology to a multinomial logistic regression (or softmax regression), but with a single hidden layer (the vector space) serving to featurize the input and outputs. In our case, this input is a wrong answer and we posit that the featurization may encode misconceptions shared among several wrong answer inputs. Ideally, two student wrong answers stemming from the same misconception but different problem seeds would be near to one another in the vector space, similar to how synonymous words cluster together in a word vector space as a result of the similar contexts in which they are used, linguistically.

We represent each answer as the concatenation of the problem type, the randomly generated seed that was used to instantiate the question, and the frequency rank of the student's answer within that seed. For example, a student submits the answer "3/4" to a question seed "ef2" that was instantiated from the problem type "adding." The answer "3/4" is the third most common answer so we encode this within the skip-gram model as the token "adding_ef2_3". We assemble a chronologically ordered sequence of answers submitted to Khan Academy for each unique student represented in each exercise's dataset.

Once a model is trained, we iterate through each student answer as a "main answer" and find its ten nearest neighbors (i.e. the ten most similar answers based on Euclidean distance within the vector space). We sort a list of these groupings by the average distance of the neighbors to the main, then proceed to select the groupings with the smallest distance for subsequent manual analysis. We enforce a certain degree of diversity in the groupings by not passing on any grouping that has greater than 50% overlap with a previously selected grouping. This style of grouping assumes that common misconceptions are clustered near one another, as opposed to manifesting as a common vector offset. The skip-gram model involves several hyperparameters: vector size (dimensionality of the vector space), window size (number of adjacent student answers in a sequence to consider when processing a given student answer), a minimum occurrence count for a token to be included in the model, and the number of training iterations. We chose a vector size of 35, a window size of 5, a minimum occurrence threshold of 25, and 40 training iterations for all our models based on the magnitude of our data and the parameters of a past study whereby a skip-gram was used to associate problems with skills based on the order of items answered (Pardos & Dadu, 2017). We leave the optimization of these hyperparameters to this context for future work.

Manual misconception labeling

After producing vectors for student answers through the skip-gram approach and identifying the ten most similar answers based on Euclidean distance to create an answer grouping, we manually verified the extent to which each of these groupings corresponds to an underlying misconception. Ideally, a single misconception label would explain all of the answers contained in the grouping. Due to the large number of answers, we focused our manual labeling on the ten groupings with the lowest average distance between the original answer and its ten neighbors. Overall, we inspected ten groupings containing 11 student answers each. Hence, we labeled 110 answers per exercise, yielding 330 total answers across the three exercises.

To determine how well each grouping identified by the skip-gram process reflects a misconception shared among its constituent student answers, we manually attribute each student answer to a misconception. Table 2 shows an example of answers assigned to misconceptions. The first row is the *main answer* generating the cluster, and the subsequent ten rows are its ten nearest neighbors in increasing order by Euclidean distance.

Table 2. Answer grouping with its main answer and two nearest neighbors labeled with misconceptions.

Generating ("Main") Seed	Student Answer	Seed	Similarity Rank	Misconception Index
xe29d6ce8acbcc688	2	xe29d6ce8acbcc688	0	2
xe29d6ce8acbcc688	3	x0ac68cd06ed51fc6	1	2
…	…	…	…	…
xe29d6ce8acbcc688	2/6	x3f10fca965656d09	10	3

The Main Seed column value shared across all rows represents the answer that the group was based around. The Seed and Student Answer column values differ because they pertain to individual neighboring answers. The explanations for Similarity Rank and Misconception Index column values are outlined in the next section.

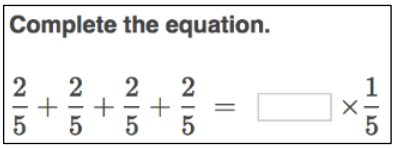

Figure 1. The main seed from Table 2.

In labeling the misconceptions for student answers in Table 2, we access each problem through a Khan Academy problem preview link (https://www.khanacademy.org/preview/content/items/SEED). Figure 1 shows the table's first problem. Here, the correct answer is 8, but the student's answer was 2. We hypothesize that the student misunderstood the problem as simply multiplying a numeric factor, in this case: 2, to match the unit fraction 2/5 apparent in the problem. We create and add this misconception label to an indexed catalogue and label the student answer as misconception 2. We extend this process to all the neighboring answers within this group, and to all the groupings within an exercise, and lastly to all three exercises we analyzed. We generated separate misconception labels for the three exercises.

Results

We evaluated the degree to which our hypothesis held true; that answers sharing the same misconception would cluster together in a vector space produced by the model. We used homogeneity of answer misconception labels in a cluster as an evaluation metric, as well as agreement between the main answer's label with the most common label of its cluster. If the method successfully clustered answers with shared misconceptions together, then the homogeneity and agreement measures should be high. Our manual labeling process was performed on the ten main answers with the smallest average distances to their own ten nearest neighbors. This was repeated for each of the three exercises for a total of 30 main answers and 300 nearest neighbor answers. Our analysis produced nine misconception labels for "*Adding and Subtracting Fractions with Unlike Denominators*", nine for "*Understanding Multiplying Fractions with Whole Numbers*", and fourteen for "*Multiplying Unit Fractions and Whole Numbers.*"

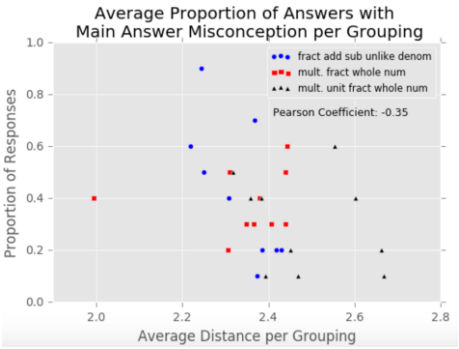

Figure 2. Proportion of answers with the main answer's misconception vs. average distance of the grouping.

With the 330 answers manually labeled, we calculated metrics for each of the 30 clusters of answers. The most common misconception of a cluster was shared, on average, by 46% of the answers in that cluster. Further, in 63.33% of the clusters, the most common misconception matched the misconception of the main answer. Finally, we found a weak negative correlation (p = -0.35) between the average Euclidean distance of answers in a cluster to their main answer and their misconception agreement. As seen in Figure 2, besides the outlier, there is a weak inverse relationship between the proportion of neighboring answers with the main answer's misconception and the average distance between the main answer and the other answers in the cluster.

Discussion

In this paper, we combine modern machine learning techniques with manual misconception labeling, providing an illustration of how the communities of learning analytics and learning sciences may benefit from combined efforts. While our experiment showed promise in the ability of algorithms to aid manual analysis, we became further interested in how methods like ours may allow conceptions to be studied in a distributed representation framework. In this discussion we give a brief evaluation of this paper's results followed by a deeper analysis of the clusters with potential future applications inspired by ideas from the learning sciences.

Evaluation of results

Though we examined only a small subset of the entire Khan Academy dataset, our analysis of student answers helped identify the most frequent misconceptions and how they differed across exercises, even when two exercises may have had the same mathematical concept behind them. We found that only a handful of misconceptions were present in each exercise – leading us to evaluate quantitatively how the skip-gram clustered student answers associated with these misconceptions. Grouping the skip-gram produced wrong answer vectors by nearest neighbors showed mixed results. Generally, the process did cluster student answers sharing a common misconception within each main student answer cluster. However, the most common misconception in an answer cluster was not necessarily the misconception evidenced in the main answer that the cluster was based on. We did verify that the more similar neighboring-answers were to the main answer (determined by Euclidean distance), the higher proportion of them shared the main answer's misconception. This observation implies that our model did cluster neighboring answers based on common misconceptions, though some may not be due to the exact same manual label. Future research can work to compare clusters with more robust misconception frameworks, and may better test validity by blinding the manual analysis. Overall, the combination of the qualitative manual misconception analysis and the quantitative evaluation of our model showed potential in clustering and identifying student answers based on shared misconceptions.

Ideas from Learning Sciences

Templated exercises aim to enable variety by parameterizing elements of the question, while remaining true to the nature of the problem. However, parameterization can produce different misconceptions across instances. Models like the one in this paper, may support a latent space of intertwined conceptions, with templated (or not) exercises shining light on traces of evidence supporting a larger distributed theory of knowledge. This dovetails with research interests from the learning sciences, providing an opportunity to draw upon teacher knowledge and education theory.

Latent conceptions and productive strategies

The answer groupings given by this model may help reveal latent student misconceptions that are more apparent when viewed across instances of a template. The specific parameters of a problem may prime differing solution strategies. Thus, one instance may conceal a misconception that another instance reveals.

If a student answers ¼ of 8 as "4", (Figure 3b) they may have performed 8-4 (trying a familiar operation), or they may have confused fourths with the more familiar halves, due to the 8/4 interaction with doubling/halving concepts. Our model clustered the response of 4 with the group centered around a different instance of the problem template (Figure 3a). These possible misconceptions may be the type expected by our analysis: a description of a student's mathematical reasoning. However, this main answer and its nearest neighbors inform new hypotheses on the students' thinking:

We offer a new hypothesis that the students are using the pictorial clues. For question (a), "15" is the entire set, for question (c), "2" is the size of the partitions portrayed of 10. In this context, the answer "4" to the question (b) may be giving the number of partitions displayed of "8". The misconception displayed by students giving these answers may be less specific than our labeling process performed in this paper. Instead, the answer grouping generated may represent a more general student solution strategy of picking out properties of a graphical hint as answer candidates. The result is that answers identifying different properties of the image (total, group size, count of groups) are grouped by the skip-gram model. A diagnosis of the single answer ¼ of 8 = "4" (above) could be augmented to consider that 4 represents one instance of a heuristic for this kind of question.

Further research will need to establish the degree to which such algorithmically generated groupings are meaningful; but if they are, tutoring systems may have a new lens on student conceptions. Consider an answer grouping from the *Multiplying Unit Fractions with Whole Numbers* exercise (Figure 4). Most (six) of the group's answers are extremely similar to the main answer (a): each of their inputs is the numerator of the repeated addends. The remaining grouped answers appear different: Figure 4 (b, c) do not match the numerator pattern, Figure 4 (d, e) even come from a different prompt structure in which the correct input is a fraction rather than a whole number. It may be that the later elements of the group are not significantly linked to the main, (a), but if they are linked,

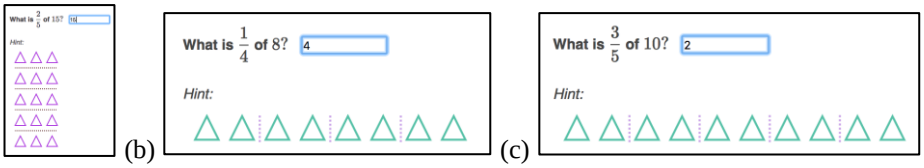

Figure 3: Selected elements (b,c) of the answer grouping around the main answer (a): 2/5 of 15 = 15. Despite the word "hint," these clues are given as part of the problem statement for this Khan Academy exercise.

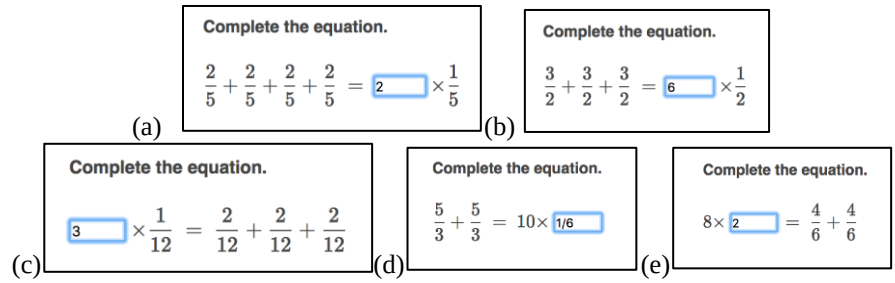

Figure 4: An answer grouping from Multiplying Unit Fractions with Whole Numbers.

we offer a hypothesis that these inputs are the partial results of student thinking, perhaps impatiently entered. In the process of doing each problem a student may arrive at statements such as: (a) 2/5 = **2** * 1/5, (b) 3/2 + 3/2 = 6/2 = **6** * 1/2, (c) 2/12 + 2/12 + 2/12 = **3** * 2/12, (d) 5/3 + 5/3 = 10/6 = 10 * **1/6**, (e) 4/6 + 4/6 = **2** * 4/6. In each case, the student may consider the expression similar enough to the prompt to try to input it. Note, all but statement (d) are true statements, similar but not equivalent to the question being asked. Statement (d) contains a hypothesized false statement from a common error. We note the connection of this analysis to that of Figure 3. Students may have picked out true properties or statements from the given information, and then used that initial idea to input their answer. The search for mathematical misconceptions may be concealed by the student's conceptions of how to succeed in the tutoring system. With online tutoring often given instant feedback, the students may feel they can work quickly and productively via this method, overcoming negative consequences for wrong answers if the strategy produces enough correct answers along the way. Future work may investigate the degree to which mathematical misconceptions are hidden behind the student's mental model of the software and its incentives.

Developing conceptions with positioning

If a system knows more about student thinking, the system is posed to give better feedback. The modern expert teacher carefully interprets representations of student thought to supply that student with challenges appropriate to the student's internal model (Franke et al., 2015). One method involves prompting students to reflect upon their thinking. Another, *positioning*, places a student in a situation set up for them to make productive engagement with their own work or the work of another learner. In Watkins et al. (2017), the teacher encourages student voicing of uncertainty as a mechanism to provide peer feedback to address incomplete understanding in a way that fosters co-construction. Our system of representation of misconceptions could aid in the orchestration of this positioning online, matching students up with peers with complimentary constructions.

We may avoid directly confronting a student's misconception with a "wrong" message. The instant feedback may encourage gaming the system, but more importantly the feedback is imprecise: leaving the student in a difficult position of not knowing which ideas to prune and which to keep. Instead, we may more gently give feedback by leveraging a machine learned similar question-answer to better respond to the student's actual input. In a normal interaction on Khan Academy, the student answering in the way shown in Figure 5(a) would be alerted that their response is wrong as soon as they submitted.

We propose an alternative that delays communicating right/wrong for the student's *own* response. Instead, we position their thinking next to a similar answer, Figure 5(b), pulled as a nearest neighbor from 5(a)'s answer grouping. The system asks the student to find the error in this other student's work. **"Another student answered a question like this but made a mistake. See if you can find the mistake and fix their answer for them."** Upon submitting this answer, the student could be shown their own question-answer again and could be asked if they'd like to keep or change their answer. Upon submitting, the normal right/wrong flow resumes. This intervention may need only happen sparingly and may also act on correct answers. We notice 4/3 is the sum 5/6+1/2 from the first exercise item. This could be naïve adding of the two fractions, or a confusion of balancing the equation. But in either case, 4/3 may indicate a partial understanding, not to be harshly rejected. The student is asked to interpret the answer of a different question, but because this answer is selected via the algorithm it has

a high chance of containing similar thinking. We notice the same procedure 5/2+1/3 yielding the answer 17/6. The goal of the intervention is to direct the student to analyze their own thinking. We note also that this intervention asks a new task of the student (analysis) while staying focused on the same skill.

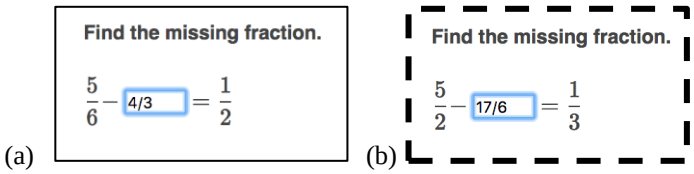

Figure 5: a student answer (a) and a similar question-answer (b) found via algorithm.

Limitations

The methods utilized to produce our results involved a selection of exercises and a limited focus on answer groupings, both for ease of manual analysis. In addition, the "long tail" of potentially interesting student responses with low frequency were lumped into a miscellaneous category and we did not have the skip-gram model create groupings across problem types both to allow the model to focus on common answers within one problem type. Consequently, we acknowledge that the results were produced in a very controlled setting which may not reflect the entirety of Khan Academy exercises. Finally, the manual analysis was not performed blindly, rather the answers were analyzed after being grouped by the model, thus researcher bias may exist. Despite the limitations we believe we offer a novel prototypical approach to identifying conceptions through a distributed, algorithmic perspective.

Conclusion

In this work, we conducted an analysis of misconceptions among the common wrong answers in three exercises chosen from Khan Academy. Through a novel application of a representation learning model, we observed that answers exhibiting common misconceptions tended to group together in the vector space, combining traces of a common misconception across answers to different numeric instantiations, or templates, of a problem. This paper contributes to theory in the parallels it draws between representation learning and the theories and best practices arrived at in the learning sciences.

There are several steps we can take as future work, starting with an improvement to the accuracy of the model by tuning its various hyper parameters. Next, we can connect to the STEM teaching community to assist in the characterization of student wrong answers and ultimately in the validation of our model representations. Finally, we can deploy pilot online interventions based on peer co-construction of understanding based on inferences about the misconceptions and partial understanding inferred to be held by each student. The learning sciences have had a strong pedagogical voice over the past decades and it is our belief that student models based on distributed representation using big data are an appropriate vehicle to empirically test and scale its impact.

References

Anderson, J. R. (1993) *Rules of the mind*. Psychology Press.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995) Cognitive tutors: lessons learned. *The Journal of the Learning Sciences*, *4*(2), 167–207.

Barnes, T. (2005) The q-matrix method: mining student response data for knowledge. In *American Association for Artificial Intelligence 2005 Educational Data Mining Workshop* (pp. 1–8).

Berland, M., Baker, R. S., & Blikstein, P. (2014) Educational Data Mining and Learning Analytics: Applications to Constructionist Research. *Technology, Knowledge and Learning*, *19*(1–2), 205–220.

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1992) Diagnosing knowledge states in algebra using the rule space model. *ETS Research Report Series*, *1992*(2).

Brown, J. S., & Burton, R. R. (1978) Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, *2*(2), 155–192.

Brown, J. S., & VanLehn, K. (1980) Repair theory: a generative theory of bugs in procedural skills. *Cognitive Science*, *4*(4), 379–426.

Erlwanger, S. H. (1973) Benny's conception of rules and answers in IPI mathematics. *Journal of Children's Mathematical Behavior*, *1*(2), 7–26.

Franke, M. L., Turrou, A. C., Webb, N. M., Ing, M., Wong, J., Shin, N., & Fernandez, C. (2015) Student engagement with others' mathematical ideas: the role of teacher invitation and support moves. *The Elementary School Journal*, *116*(1), 126–148. https://doi.org/10.1086/683174

Hinton, G. E. (1990). Connectionist learning procedures. In *Machine Learning, Volume III* (pp. 555-610).
Johnson, W. L., & Soloway, E. (1984) Intention-based diagnosis of programming errors. In *Proceedings of the 5th national conference on artificial intelligence, austin, tx* (pp. 162–168).
Jurkovic, N. (2001) Diagnosing and correcting student's misconceptions in an educational computer algebra system. In *Proceedings of the 2001 international symposium on Symbolic and algebraic computation* (pp. 195–200). ACM.
Liu, R., Patel, R., & Koedinger, K. R. (2016) Modeling common misconceptions in learning process data. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 369–377). ACM.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013) Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).
Milson, R., Lewis, M. W., & Anderson, J. R. (1990) The teacher's apprentice project: building an algebra tutor. *Artificial Intelligence and the Future of Testing*, 53–71.
Ostrow, K., Donnelly, C., Adjei, S., & Heffernan, N. (2015) Improving student modeling through partial credit and problem difficulty. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 11–20). ACM.
Pardos, Z. A. (2017) Big data in education and the models that love them. *Current Opinion in Behavioral Sciences*, *18*, 107–113.
Pardos, Z. A. & Dadu, A. (2017) Imputing KCs with Representations of Problem Content and Context. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (UMAP). Bratislava, Slovakia. ACM. Pages 148-155.
Piaget, J. (1952) *The child's concept of number*. New York.
Piaget, J. (1962) Commentary on Vygotsky's criticisms of Language and thought of the child and judgment and reasoning in the child. *Lev Vygotsky, Critical Assessments*, *1*, 241–260.
Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015) Deep knowledge tracing. In *Advances in Neural Information Processing Systems* (pp. 505–513).
Putnam, R. T., Sleeman, D., Baxter, J. A., & Kuspa, L. K. (1986).A summary of misconceptions of high school Basic programmers. *Journal of Educational Computing Research*, *2*(4), 459–472.
Razzaq, L., Patvarczki, J., Almeida, S. F., Vartak, M., Feng, M., Heffernan, N. T., & Koedinger, K. R. (2009) The ASSISTment Builder: supporting the life cycle of tutoring system content creation. *IEEE Transactions on Learning Technologies*, *2*(2), 157–166.
Reiser, B. J., Anderson, J. R., & Farrell, R. G. (1985) Dynamic student modelling in an intelligent tutor for LISP programming. In *IJCAI* (Vol. 85, pp. 8–14).
Sfard, A., & Cobb, P. (2014) Research in mathematics education: what can it teach us about human learning. *The Cambridge Handbook of the Learning Sciences*, 545–63.
Sleeman, D., & Brown, J. S. (1982) *Intelligent tutoring systems*. London: Academic Press.
Smith, J. P., DiSessa, A. A., & Roschelle, J. (1994) Misconceptions reconceived: a constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, *3*(2), 115–163.
Star, J. R. (2005) Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, 404–411.
Tang, S., Peterson, J., & Pardos, Z. (2017) Predictive modelling of student behaviour using granular large-scale action data. *The Handbook of Learning Analytics*, 223–233.
Tatsuoka, K. K. (1985) A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, *10*(1), 55–73.
VanLehn, K. (1990) *Mind bugs: The origins of procedural misconceptions*. MIT press.
Watkins, J., Hammer, D., Radoff, J., Jaber, L. Z., & Phillips, A. M. (2017). Positioning as not-understanding: The value of showing uncertainty for engaging in science. *Journal of Research in Science Teaching*, n/a-n/a.
Zinn, C. (2006) Supporting tutorial feedback to student help requests and errors in symbolic differentiation. In *Intelligent Tutoring Systems* (pp. 349–359). Springer.

Acknowledgements

We thank Khan Academy for sharing their anonymized exercise data and Alan Schoenfeld for his assistance in identifying exercise topics in which misconceptions have been well studied. This work was supported, in part, by a grant from the National Science Foundation (#1547055).