Analyzing Strategies in MATHia with BERT

Abisha Thapa Magar[1 [0009−0006−7705−3735]], Stephen E. Fancsali[2 [0000−0002−4016−8145]], Vasile Rus[1 [0009−0002−4739−0440]], April Murphy[2 [0000−0002−8283−2868]], Steve Ritter[2 [0000−0003−2807−9390]], and Deepak Venugopal[1 [0000−0001−7466−5417]]

[1] University of Memphis
{thpmagar, vrus, dvngopal}@memphis.edu
[2] Carnegie Learning
{sfancsali, amurphy, sritter}@carnegielearning.com

Abstract. Understanding how students use math strategies is important to help us build tools and techniques that improve cognitive flexibility in students, i.e., select strategies that are appropriate and efficient for a problem. In this work, we focus on instructional content within MATHia, where students choose between strategies that were previously taught to them independently. Some problems favor one strategy over the other, giving us the opportunity to understand how/if students learn to pay attention to problem characteristics that suggest one strategy over the other. Using data from over 600 schools, we show that students find it hard to adapt their strategies to suit a problem. Further, we learn a BERT model to learn embeddings for strategies, develop a prediction task to distinguish between successful and unsuccessful strategies, and analyze its results to reveal deeper insights into student strategies.

Keywords: math strategies · representation learning · deep models.

1 Introduction

Problem-solving strategies are foundational to math learning [11]. However, in Intelligent Tutoring Systems (ITSs), understanding strategies is challenging since the system's interface may limit the possible strategies. In this paper, we analyze strategies in MATHia, an ITS that is part of a blended curriculum for middle-school math. In MATHia, problem-solving is broken down into sub-goals, and students perform actions to complete these sub-goals. We focus our analysis on a specific MATHia workspace (instructional module) that teaches ratio and proportion. This module follows the general pedagogical principle of teaching students multiple problem-solving strategies and then allowing them to choose between strategies. We first analyze if students use the optimal strategy based on data collected from over 600 schools. Our results seem to indicate that students have a hard time choosing efficient strategies, and even when they do, they do not execute the strategy correctly. To understand more deeply how strategies are related to student performance, we develop advanced AI models to distinguish between strategies using hidden patterns in the data. In particular, we

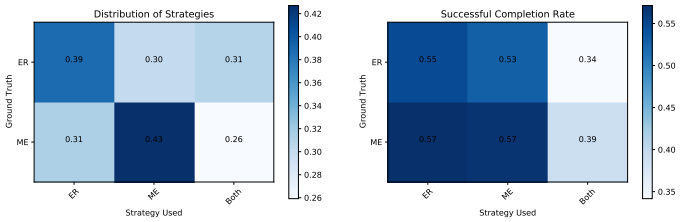
use Bidirectional Encoding Representations From Transformers (BERT) [6], a model typically used in language understanding, to separate strategies that are correctly executed from those that are not. Specifically, we *pre-train* BERT using student interaction data from 100 schools to learn representations (*embeddings*) for strategies. We then *fine-tune* the embeddings on data from the remaining schools to classify strategies. To add more context to the strategy, during fine-tuning, we augment the embedding of a strategy with prior skills inferred by Bayesian Knowledge Tracing [5], temporal features extracted from actions performed by the student and from problem text. Our model shows excellent generalizability, i.e., fine-tuning on just the first 10% of the problems worked on by a student in the workspace yields a maximum ROC-AUC score of 92% when we test the model on the remaining 90% of problems in the workspace.

2 Related Work

Several different approaches have been explored to identify strategies in virtual environments [14]. In [13], an approach was developed to augment the strategies in model tracing tutors. In [1], a machine learning model was learned using data labeled from Cognitive Tutor, and similar approaches have been used in other domains [7, 15]. Sequence modeling has often been used to identify strategies [8, 9] in virtual environments such as Betty's brain [3]. In [20], sequence pattern mining was applied to a MOOCs platform to analyze activity sequences of learners [20]. Other studies have explored Markov models and sequence mining to uncover patterns in student interactions within ITSs [2]. More recently, deep models such as LSTMs have been used to learn strategies [16]. In [18], a foundational BERT model was developed to learn embeddings for MATHia strategies, but unlike our work, context-specific strategies were not analyzed.

3 Strategies in MATHia

Newell and Simon [12] in their classical work provide a formal framework to understand problem-solving as a search through the problem-space. Thus, a strategy corresponds to making choices of actions to perform within this space. In the MATHia ITS, a student solves sub-goals (or steps) within the problem-space. In this work, we focus on a specific piece of instructional content where the student is supposed to make deliberate choices reflecting their strategy. Specifically, the instructional module for the ratio and proportions workspace in MATHia follows the general pedagogical strategy of teaching different strategies one at a time and then asking students to choose between strategies based on the context of a problem. The steps in this module include setting up a proportion based on the problem text and then solving the proportion for the unknown variable. For instance, given a question, *Akuro is the manager of a car wash. Last week there were 50 customers. This week there were 50% more customers than last week. What is the number of customers this week?*, we can set up the proportion for the change in amount as, $\frac{50}{100} = \frac{x}{50}$. Now, students are presented with two choices



(a) (b)

Fig. 1. Analyzing strategies in the ratio and proportions workspace. (a) shows the distribution of how students used strategies and (b) shows the Successful Completion Rate (SCR) for these strategies.

called *optional tasks*. in choice 1, they solve the unknown variable by observing that $100 \div 2 = 50$ and therefore $x = 25$. We call this approach as equivalent ratios (ER). In choice 2, called means and extremes (ME), they use cross multiplication, i.e., $50 \times 50 = 100 \times x$ and solve for x. Both these approaches are valid strategies. However, the ER approach requires the student to recognize the pattern where it can be easily applied, i.e., when one value can be scaled up/down to another value (using an integer value), the ER strategy is very efficient. Thus, we have a notion of *ground truth* in terms of which strategy is efficient for a problem. Students solve prerequisite workspaces that train them on both these approaches separately. Here, the goal is for them to make a choice depending on the problem context and thus build cognitive flexibility.

Data Analysis. To understand how students use strategies within the ratio and proportions workspace, we analyze data collected from 655 schools. This workspace is part of blended instruction in 7^{th}-grade math. Our data consists of 414K unique instances (identified by a student, problem pair), 27946 unique students and 280 word-problems. The data format is similar to the tutor interaction format available through the PSLC datashop [17]. The data includes the sequence of steps students work on within the problem, hint usage, feedback from MATHia, and the skills (knowledge components [10]) tracked for the problem. Fig. 1 compares the strategies followed by the students with the ground truth strategies. Specifically, Fig. 1 (a) shows the distribution of strategies. As seen here, only 39% correctly choose the ER strategy and 43% correctly choose the ME strategy. This indicates that in majority of the cases, students were unable to switch to the optimal strategy. Further, it can be seen that 26% use both strategies, which is a strong signal of poor meta-cognition since these students failed to correctly comprehend that both strategies are in fact alternate but equivalent approaches to solving the problem. Next, we analyzed the *successful completion rate* (SCR) for strategies. Specifically, a strategy is successfully completed if after completing the last optional task step, the student is able to answer the question without any additional help, i.e., without requiring multiple attempts or hints to answer the question. We believe this indicates that the student is able to apply

the strategy to a specific problem. The SCR for a set of strategies **S** is the ratio $\frac{s}{N}$, where s is the number of successful completions of strategies in **S** and N is the total number of times a strategy in **S** was used. Fig. 1 (b) shows the strategy-wise SCR breakdown. As seen here, using the ER strategy for problems with ER as the ground truth has a slightly higher SCR compared to the ME strategy. For problems where the ground truth is ME, the SCR remains the same for either ER or ME strategies. Since ME is more algorithmic in nature, it can be applied to all problems and therefore, students seem to have similar success applying this strategy regardless of the ground truth strategy. When both strategies are applied to a problem, as expected, the SCR is significantly lower.

3.1 BERT Model

From the analysis in the previous section, it is evident that even when the optimal strategy is used, there is only around 55-57% chance of success. Therefore, we next explore how AI methods can help distinguish between successful and unsuccessful strategies. To do this, we use BERT to learn a *low-dimensional embedding* for the problem-space. As in the classical BERT model used for language understanding, there are two key steps within our BERT framework. The first step pre-trains the model and learns general embeddings for strategies. The next step fine-tunes these embeddings for a specific prediction task. The architecture for our model is a multi-layer bidirectional Transformer encoder [19] which uses context in both directions, i.e., in our case, a transformer block scans the input from both the right-to-left and left-to-right directions. The main hyperparameters include the number of Transformer blocks (L), the size of the hidden representation (H), and the number of attention heads (A) within each block to implement the attention mechanism. For our model, we use $L = 4$, $H = 64$, and $A = 8$. Next, we describe the main components of our model.

Pre-Training. We use the Masked Language Model (MLM) approach that is typically used for pre-training language models. Specifically, we encode the interaction between the student and MATHia as a sequence of tokens where each token contains the step-name, the response of MATHia to the action performed in this step (e.g., whether the step had an error), and whether a hint was required to perform the action. We then mask tokens and train the model to predict the masked tokens. To do this, the model must infer the token from its context (neighboring tokens) and therefore, it learns to represent the sequence of tokens as an embedding in a fully unsupervised manner. Our model consists of approximately 200K trainable parameters. We used training data from 100 schools that had the largest number of instances in the dataset. The number of instances we used in pre-training is 130K and 20K for validating the pre-trained model. We masked 15% of the steps similar to the approach used in the original BERT model and set the maximum length of a sequence as 128. We performed pre-training in 8 hours on two parallel Tesla GPUs.

Fine-Tuning. Fine-tuning the pre-trained model adapts it to downstream tasks without the need to retrain the full model. We initialize the BERT model with parameters learned during pre-training and then supervise the fine-tuning with

FT	FT+skills	FT+time	FT+gt	FT+gt(u)	FT+skills+time
0.67	0.80	0.83	0.67	0.67	**0.92**

Table 1. Comparing ROC-AUC scores of fine-tuned models on test data.

a small set of labeled examples. A classification layer is added on top of the pre-trained model to predict the outcome of a strategy, i.e., whether the final answer step was solved by the student correctly without additional help. During fine-tuning, we augment the embedding generated for (s, p), a student-problem pair, with additional context that is relevant to the strategy as follows.

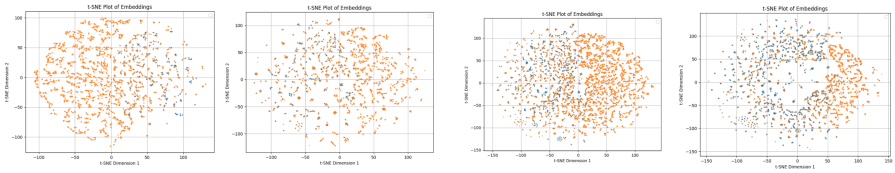
Skills. We track skills or knowledge components [10] for each student that are relevant to the workspace. The skills are updated using the standard Bayesian Knowledge Tracing (BKT) model [4]. We use the probabilities of correctness for each of the skills based on the BKT parameters computed using the prior problems that the student has attempted.

Temporal features. We track the activity time of students for each step and augment the embedding with the following temporal features. The cumulative time spent over all the steps of the problem indicates how long the student engaged with MATHia in solving the problem. We measure the time spent between completing the final step in an optional task and attempting the final answer. This is an indicator of how long the student takes to think about the strategy worked out in the optional task and relate this to answering the problem. Finally, we also measure the total time taken to complete all the optional task steps and the total time taken to answer all the other (non-optional) steps in the problem.

Ground Truth Strategy. We add the ground truth strategy label (i.e., whether the problem's optimal strategy is ER/ME) to the embedding. Further, we considered a second variant in which we computed the uncertainty in the strategy label based on the problem text. For instance, for the ER strategy, some scaling factors may be harder than others (e.g., scaling 18 to 54 is harder than scaling 10 to 100). To encode this, we generated an embedding for the problem text using a pre-trained BERT language model and trained a binary classifier to distinguish between ER and ME problems based on the text embedding. We then add the probability output of this classifier as an input feature to the fine-tuned model.

Model Evaluation. We evaluated the model on data from 555 schools (we excluded the 100 schools used in pre-training). We fine-tune the pre-trained model only with data from the first 10% of problems that the students have attempted and evaluate the accuracy of the model on the remaining 90% of future problems they worked on. The training time for each fine-tuned model was less than 30 minutes. We develop 6 models, FT is the baseline model that only uses embeddings. To this, we add skills, time (temporal features), ground truth strategy label (gt), and the probability associated with the ground truth strategy (gt(u)). The results from all our fine-tuned models[3] are summarized in Table 1. We use the Receiver Operating Characteristic Area Under the Curve

[3] https://github.com/abisha-thapa/ft_bert



(a) Promoted (b) Graduated

Fig. 2. t-SNE plots for strategy embeddings. Promoted/Graduated student strategies are shown separately. In each case, the first plot shows embeddings from the first 10% of problems and the second from the last 10% of problems. Blue colored points follow the ground truth strategy and orange colored points do not follow the ground truth.

(ROC-AUC) metric (since outputs from our model are continuous values) to compare the performance of our models. As seen from our results, skills and temporal features significantly improve the baseline model (FT). In contrast, the ground truth strategy for a problem does not improve performance over the baseline. This seems to indicate that choosing the right strategy does not imply correct execution of the strategy. One possible explanation is that the ME strategy is algorithmic and is more generally applicable. Therefore, students who learn ME do not want to switch strategies.

Strategy Embeddings Fig. 2 shows the t-SNE plots for embeddings over time for graduated and promoted students (graduated students demonstrate mastery over all skills related to the workspace). As seen here, the embeddings for initial strategies for graduated students shows a fuzzy separation between orange and blue colored strategies. However, over time, it seems like the model obtains a better separation between them, showing that their strategies evolve. In contrast, for promoted students, the variation in embeddings is minimal.

4 Conclusion

In this work, we analyzed instructional content in MATHia, where students can choose between two different strategies to solve ratio and proportion problems. In this workspace, based on data collected from 655 schools, we discovered that students find it hard to choose and execute the strategies that are efficient for a specific problem. Further, we developed a BERT-based model to learn strategy embeddings and, using this, we discovered that students who attain mastery evolve their strategies over time. In the future, we plan to use insights from our model to improve outcomes such as time-to-mastery.

Acknowledgements

This research was supported by an award from the Bill and Melinda Gates Foundation and NSF award #2008812. The opinions, findings, and results are solely the authors' and do not reflect those of the funding agencies.

References

1. Baker, R., de Carvalho, A.: Labeling student behavior faster and more precisely with text replays. In: Educational Data Mining 2008 (2008)
2. Baker, R.S., Yacef, K.: The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining **1**(1), 3–17 (Oct 2009)
3. Biswas, G., Segedy, J.R., Bunchongchit, K.: From design to implementation to practice a learning by teaching system: Betty's brain. International Journal of Artificial Intelligence in Education **26**(1), 350–364 (2016)
4. Corbett, A.T.: Cognitive computer tutors: Solving the two-sigma problem. In: Proceedings of the 8th International Conference on User Modeling 2001. p. 137–147 (2001)
5. Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction **4**, 253–278 (1994)
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
7. DiCerbo, K.E., Kidwai, K.: Detecting player goals from game log files. In: Proceedings of the 6th International Conference on Educational Data Mining, Memphis, Tennessee, USA, July 6-9, 2013. pp. 314–315 (2013)
8. Kinnebrew, J.S., Biswas, G.: Identifying learning behaviors by contextualizing differential sequence mining with action features and performance evolution. In: Proceedings of the 5th International Conference on Educational Data Mining, Chania, Greece, June 19-21, 2012. pp. 57–64 (2012)
9. Kinnebrew, J.S., Loretz, K.M., Biswas, G.: A contextualized, differential sequence mining method to derive students' learning behavior patterns. Journal of Educational Data Mining **5**(1), 190–219 (2013)
10. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. Cogn. Sci. **36**, 757–798 (2012)
11. Molnár, G., Greiff, S., Csapó, B.: Inductive reasoning, domain specific and complex problem solving: Relations and development. Thinking Skills and Creativity **9**, 35–45 (2013)
12. Newell, A., Simon, H.: Human Problem Solving. Prentice Hall (1972)
13. Ritter, S.: Communication, cooperation and competition among multiple tutor agents. In: Artificial Intelligence in Education: Knowledge and media in learning systems. pp. 31–38 (1997)
14. Ritter, S., Baker, R., Rus, V., Biswas, G.: Identifying strategies in student problem solving. Design Recommendations for Intelligent Tutoring Systems **7**, 59–70 (2019)
15. Rowe, E., Baker, R.S., Asbell-Clarke, J., Kasman, E., Hawkins, W.J.: Building automated detectors of gameplay strategies to measure implicit science learning. In: Proceedings of the 7th International Conference on Educational Data Mining, EDM 2014, London, UK, July 4-7, 2014. pp. 337–338 (2014)
16. Shakya, A., Rus, V., Venugopal, D.: Student strategy prediction using a neuro-symbolic approach. In: Proceedings of the 14th International Educational Data Mining Conference (EDM 21) (2021)
17. Stamper, J.C., Koedinger, K.R., de Baker, R.S.J., Skogsholm, A., Leber, B., Demi, S., Yu, S., Spencer, D.: Datashop: A data repository and analysis service for the learning science community. In: AIED. vol. 6738, p. 628 (2011)

18. Thapa Magar, A., Shakya, A., Fancsali, S.E., Rus, V., Murphy, A., Ritter, S., Venugopal, D.: "can a language model represent math strategies?": Learning math strategies from big data using bert. In: Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK 2025). pp. 655–666 (2025)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017)
20. Wong, J., Khalil, M., Baars, M., Koning, B.D., Paas, F.: Exploring sequences of learner activities in relation to self-regulated learning in a massive open online course. Computers & Education (2019)