Adaptive Natural-Language Targeting for Student Feedback

Y. Alex Kolchinski*, Sherry Ruan*, Dan Schwartz, Emma Brunskill
Stanford University, Stanford, CA, USA
{yakolch, ssruan, daniel.schwartz}@stanford.edu, ebrun@cs.stanford.edu

ABSTRACT

In tutoring software, targeting feedback to students' natural-language inputs is a promising avenue for making the software more effective. As a case study, we built such a system using Natural Language Processing (NLP) to provide adaptive feedback to students in an online learning task. We found that the NLP targeting mechanism, relative to more traditional multiple-choice targeting, was able to provide optimal feedback from fewer student interactions and generalize to previously unseen prompts.

Author Keywords

Intelligent tutoring systems; natural language processing; adaptive feedback.

INTRODUCTION

Tutoring has long been held to be an effective practice in education, and for good reason: in a number of studies, tutors have been shown to raise students' performance levels by a standard deviation or more [11]. Software tutors have been proposed as a way to expand access to tutoring, and have in certain contexts approached the performance of human tutors. However, the differences in the capabilities of human and software tutors are still quite large. One of the most important contributors to these differences is granular feedback, or the ability to target useful and frequent responses to students [11]. Developing more effective feedback-targeting systems for tutoring software is therefore an important avenue for improving their performance.

An important difference between human and software tutors is their ability to use natural language. While human tutors are able to both explain concepts and get a gauge on what students understand through natural language, software is generally better at presenting natural-language material to students than interpreting the natural-language responses of a student. This presents a limitation in the ability of tutoring software to effectively target feedback to students. However, should this change, the software could come one step closer to the performance of human tutors.

It is with this in mind that we developed a feedback targeting method based on natural-language inputs from students. Our learning task taught students to tell poison ivy from other plants (see details in Experiment section). The contributions of this paper are as follows:

- Our system models the interaction between a student's natural language responses and the tutoring system as a contextual bandit [10] which considers a student's responses to exercises as states, and candidates for subsequent feedback as available actions. The contextual bandit framework allows the optimization of these actions for a reward signal, which is in this case pre-test to post-test gain. This framing is sufficiently general to be applicable to most, if not all, uses of tutoring software.
- We showed that the natural-language based targeting policies were able to choose optimal feedback actions from fewer exercises (i.e. sparser state) than multiple-choice based policies, and that the natural-language policies were effective even when tested on previously unseen interactions. This indicates significant promise for further applications in tutoring software.

RELATED WORK

Shute [9] categorizes granular targeted feedback in tutoring software as "micro adaptive" functionality, alongside scaffolding and other locally optimized interactions. Such functionality can be seen in a number of tutoring systems, including the well-known Cognitive Tutor family of software [3]. These approaches give feedback in a number of contexts, including correcting mistakes, providing hints, or even metacognitive feedback meant to encourage good habits [8].

However, the predominant ways in which these systems target their feedback to users are based on task-specific inputs such as multiple-choice responses, computer code, and math equations written by the student [3]. By using less expressive inputs than natural language, these systems lose out on a potentially powerful source of signal for targeting feedback to students.

Approaches to providing tutorial feedback targeted to natural-language student input exist. Aleven et al [1] demonstrated a system for adding natural-language inputs to the Cognitive Tutor approach to support self-explanatory behavior in a student. Graesser et al [6] created a chat-style system for student interactions in which the software prompts the student with questions and provides feedback to the student's responses.

*The two authors contributed equally to this work.

L@S 2018, June 26–28, 2018, London, United Kingdom

ISBN 978-1-4503-5886-6/18/06…$15.00

DOI: https://doi.org/10.1145/3231644.3231684

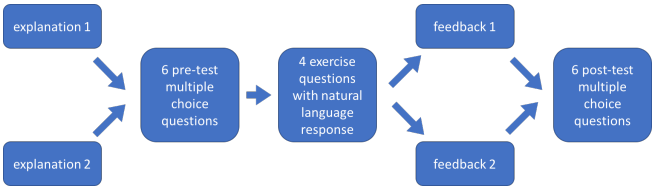

Figure 1. Survey Flow

Figure 2. (a) Left: Six questions used in pre-test and post-test sections. (b) Right: Four questions used in exercise section.

Both of these approaches showed promising results, but shared the key limitation of feedback targeting mechanisms which depend on some combination of hand-tuned knowledge engineering and semantic matching. With such models, every feedback condition must be carefully tailored and matched to possible inputs and/or targeted to students using some notion of semantic similarity. The former is extremely effort-intensive: hand-creating a database of feedback matched to all possible responses is not feasible at scale. The latter is also not ideal: for example, a student who fixates on one aspect of a concept might benefit from feedback that delves more into other sides of the issue, but the semantic-matching strategy would be likely to give the student feedback closely associated to the concepts where the student was already devoting too much attention.

A more effective mechanism would be to target feedback not by semantic similarity but by predicted results on the actual end goal of knowledge gain, which is what our method accomplishes. Moreover, it requires the hand-creation only of a set of feedback texts useful for the task at hand, a significant reduction in labor cost relative to hand-tailored approaches.

EXPERIMENT

In order to evaluate the feasibility of natural-language feedback targeting, we conducted a proof-of-concept experiment to demonstrate the capabilities of the technique in a simple setting. Our proposed learning activity took inspiration from the educational principle of contrasting cases, where students learn and demonstrate comprehension of the key similarities and differences between concepts by observing and discriminating between contrasting pairs [4]. In our activity the contrasts were between poison ivy and other plants. Using an unfamiliar subject enables us to design a tutoring system that is agnostic to users' prior knowledge. In other words, the learning gain gauged in our experiment is a valid measure of the effectiveness of the tutoring system. The task is simple but demonstrates a proof-of-concept of tutoring system that can leverage the student's natural-language responses to provide effective feedback.

Survey Design

As shown in Figure 1, the survey flow contains five stages.

Initial explanation: At the start of the activity, we gave all participants one of two explanations at random. Half of participants saw paragraph 1 and half saw paragraph 2. Each paragraph was designed to communicate one key feature of poison ivy: the first for leaves, the second for thorns and colors. By giving the participants one of the two explanations, we induced variation in "prior knowledge" going into the rest of the activity to simulate initial variation between students in real learning tasks.

Pre-test: All participants answered six multiple-choice questions in random order. Each question consisted of a photo (see Figure 2a) and three answer choices: "Poison ivy", "Not poison ivy", and "I don't know", which was included to discourage guessing.

Exercises: After the pre-test, participants answered four exercise questions (Figure 2b), which were similar in format to the pre-test questions but did not include an "I don't know" option. We also required participants to complete an open-text response with a minimum of 25 characters to explain their reasoning.

Feedback: After the exercises, participants were given one of two feedback paragraphs, each a paraphrase of one of the two initial explanations. Both the initial explanation and the feedback were uniformly randomized, which allowed us after the fact to precisely evaluate the effectiveness of any possible policy using the technique of importance sampling.

Post-test: After reading the feedback, participants took a post-test, considering of the same questions as in the pre-test but in shuffled order. Scores were computed as number of questions answered correctly. We measured the learning gain by subtracting the pre-test score from the post-test score.

Participants

We recruited 949 UK residents from Prolific to participate in our study. Since poison ivy does not grow in the UK, 93.7% of participants indicated no or little knowledge of poison ivy. The rest indicated only "a moderate amount." The study took 7 minutes to finish on average, and participants were paid $1 each.

POLICIES

Feedback selection can be naturally modeled as a contextual bandit problem [10]. In our study, state came from participants' responses to the exercises, and the actions were the two possible feedback paragraphs. The reward, which we sought to maximize, was the learning gain. It was calculated as post-test score minus pre-test score.

We gathered data using a random policy, trained various policies on this dataset, and used importance sampling [10] for policy evaluation. In this way, we were able to gather one large dataset and use it to evaluate any policy which we wished to test. No matter which of the two possible actions a policy would select for a given student, there would be a 50% chance that the randomly-selected feedback would match that selected

by the policy - in this manner, any policy would have approximately 475 unbiased examples against which to evaluate its performance.

We implemented and tested four polices: oracle, multiple choice targeted policy, and two NLP targeted policies (word vector sums and bag of words). All policies were trained and tested on all four exercise responses and one exercise response at a time. NLP policies were also tested *out of sample*, meaning they were trained on three responses and tested on the fourth.

Oracle: This policy "peeked" at the initial explanation a participant saw (which other policies could not do) and chose the feedback paragraph which provided the explanation a student had not already seen. While it is possible that in some cases this would not be the optimal choice (e.g. a student had an unusual amount of trouble absorbing the explanation about the thorns), we found that students who had been provided both explanations outperformed students who had only seen one, repeated twice, by a substantial amount. Thus, we used the oracle policy as a pedagogically near-optimal point of comparison against which to evaluate the other policies.

Multiple choice targeted linear regression: We used the participant's multiple-choice answers to the four exercise questions as input and their corresponding learning gain as an output. We trained two linear regression models, one for each of the two feedback conditions, to predict a student's learning gain based on how they had answered the exercises and which feedback they were shown. The policy selected the feedback with higher predicted learning gain.

Word vector sum targeted linear regression: We trained two linear regression models like in the previous policy, but here the inputs were the natural-language responses. We conducted basic preprocessing (e.g., removing punctuation and stop words) and converted words to 100-dimensional Glove embeddings [7], which we summed for each participant.

Bag of five words: We found the most common five words used by participants in the system (more than five was not found to improve performance), and represented each student's responses as a five-dimensional vector of how many times the participant had used each of those five words. The sum model aggregated this over all exercise responses, while the concatenation model treated them separately.

RESULTS

Participants' average learning gains by different policies presented above are shown in Figure 3. We used a 20% held-out test set to ensure there was no overfitting. The oracle showed the best gains: 1.84 on the validation set. Random policies provide a baseline performance. Always Feedback 1 ($p = 1$), Always Feedback 2 ($p = 0$), and Coin Flip ($p = 0.5$), gave 1.35, 0.78, and 1.12 gains on the validation set. All other policies which were trained and tested on all four exercise responses showed statistically indistinguishable performance from the oracle policy, demonstrating the ability of both multiple-choice and NLP feedback targeting to pick the right feedback condition essentially all of the time given four exercise responses worth of input data.

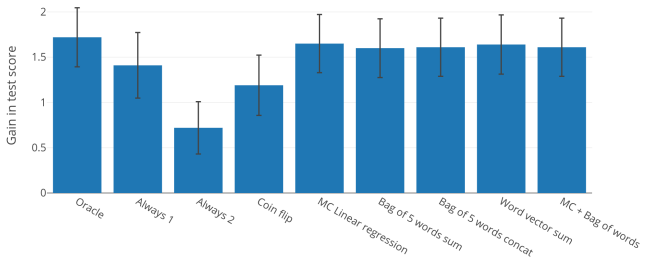

Figure 3. Gains for policies trained on all 4 exercise responses. Error bars represent 95 % confidence intervals.

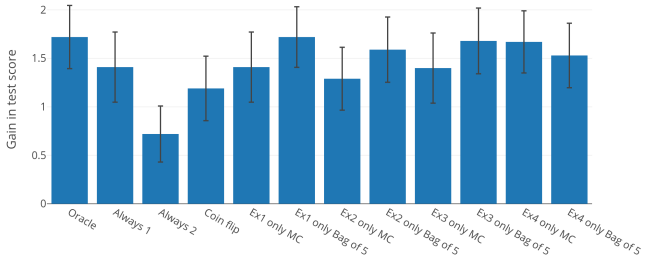

Figure 4. Participants' learning gains by different one-question policies. Error bars represent 95 % confidence intervals.

Participants' learning gains by four different one-question policies compared against baselines and the oracle are illustrated in Figure 4. As is shown, policies targeted through NLP outperformed multiple choice targeted ones on all of the exercise responses except the last one. The fact that the NLP-targeted policies consistently performed at or near oracle levels even when trained and tested on a single exercise response is a powerful testament to the greater signal density provided by natural-language student inputs as compared to multiple-choice responses. While the multiple-choice-targeted policy was able to achieve competitive performance when trained on the fourth exercise response alone, this highlights the fact that questions must be very carefully calibrated to provide good signal from the multiple-choice responses alone. The NLP policies were much more robust to variation between questions.

Finally, we show results of out-of-sample policies versus baselines and the oracle in Figure 5. It is impossible to use any multiple choice targeted policies in this setting because of the change of the input domain. However, NLP-targeted policies achieved 1.69, 1.62, 1.61, and 1.5 average gains for exercise responses 1, 2, 3, and 4 out-of-sample 5-bag-of-words poli-

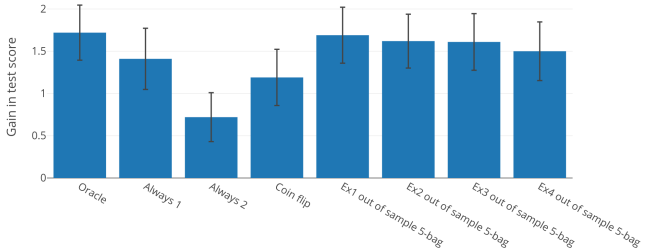

Figure 5. Participants' learning gains by different out-of-sample policies. Error bars represent 95 % confidence intervals.

cies respectively. This means that unlike the multiple-choice models, the NLP models can generalize to unseen questions and interactions by learning a relatively general notion of how students respond to prompts depending on what they know.

DISCUSSION

Our experiment has shown several promising aspects of the use of natural-language inputs from students for feedback selection. One is that NLP-powered policies are just as capable of inferring students' knowledge as those that use carefully constructed multiple-choice questions. This means that with NLP-powered feedback targeting, there can be less reliance on extensive hand-tuning of exercises for tutoring software, allowing for more expansive curricula and broader reach to more topics and contexts with the same time spent on development.

In addition, while a single multiple-choice response was not enough in most cases, a single sentence proved to be enough in our setting to effectively target feedback. A sentence-long explanation is not an undue burden to a student. In fact, asking students to self-explain in a tutor has been shown to be a helpful instructional tactic in tutoring software [2]. Thus, incorporating NLP feedback targeting in a tutoring system can allow for more frequent and relevant feedback.

Finally, NLP-powered feedback targeting has the powerful advantage of generality. Feedback-selection policies that use multiple-choice answers as input must be trained to those specific questions, and much of existing tutoring software that selects feedback based on student's natural-language inputs rests upon either extensive special-purpose knowledge engineering or semantic matching of feedback to student inputs. On the contrary, our approach only requires a consistent knowledge-gain metric and a set of feedback conditions, and can then effectively infer students' knowledge states and select feedback to maximize students' knowledge gains based on students' responses, even to prompts unseen in the training data. This opens many avenues to making curricula more flexible and reducing the labor load of developing tutoring software.

FUTURE WORK

One direction for future work would be to apply these techniques to more complex domains. In various domains, experts and novices describe problems in different ways depending on how sophisticated their mental representations are [5]. This represents a promising source of signal for models such as those described in this paper, which could be used in fields containing well-defined component concepts, such as physics and medicine. For example, a system that gives medical student automatic feedback on their diagnoses could be built based on the strategies we proposed.

Although in this study we observed substantial learning gains from a system based on relatively simple NLP models, it will be important to assess the performance of NLP-based systems on more complex tasks than poison ivy identification, where identifying discrete features like thorns and leaves is not sufficient for good performance.

Additional avenues for future work include incorporating NLP feedback-targeting mechanisms into existing educational software. This would provide multi-step interactions with students - a more complex use case than evaluated in this paper - and real-world differentiation in prior knowledge. Ultimately, a real-world test bed is the true test of how useful these techniques will prove to be.

CONCLUSION

We demonstrated that targeting feedback in educational software to natural-language inputs is a more effective and context-independent method than multiple-choice targeting. Such techniques present promise for applying NLP to tutoring software.

REFERENCES

1. Vincent Aleven, Octav Popescu, and Kenneth R Koedinger. 2001. Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In *Proceedings of Artificial Intelligence in Education*. Citeseer, 246–255.
2. Vincent AWMM Aleven and Kenneth R Koedinger. 2002. An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive science* 26, 2 (2002), 147–179.
3. John R. Anderson, Albert T. Corbett, Kenneth R. Koedinger, and Ray. Pelletier. 1995. Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences* 4, 2 (1995), 167–207.
4. John Bransford and Nancy S. McCarrell. 2018. A sketch of a cognitive approach to comprehension: Some thoughts about understanding what it means to comprehend. (03 2018).
5. Michelene TH Chi, Paul J Feltovich, and Robert Glaser. 1981. Categorization and representation of physics problems by experts and novices. *Cognitive science* 5, 2 (1981), 121–152.
6. Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M Louwerse. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers* 36, 2 (2004), 180–192.
7. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.
8. Ido Roll, Vincent Aleven, Bruce Mclaren, and Kenneth Koedinger. 2011. Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. 21 (04 2011), 267–280.
9. Valerie Shute. 1993. A macroadaptive approach to tutoring. 4 (01 1993), 61–93.
10. Richard S. Sutton and Andrew G. Barto. 1998. *Introduction to Reinforcement Learning* (1st ed.). MIT Press, Cambridge, MA, USA.
11. Kurt VanLehn. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist* 46, 4 (2011), 197–221.