Developing a Large-scale Feedback System to Train Dialogue-based Tutors using Student Annotations

Cassius D'Helon, Vinay Kasireddy, Nirmal Mukhi and Jae-Wook Ahn

IBM Watson Education, Yorktown Heights, NY, USA

cdhelon@us.ibm.com

Abstract. We're developing a feedback system that uses student annotations embedded in the dialogue to improve the natural language understanding and the content of dialogue-based tutors in multiple knowledge domains. The annotations are collected in the context of each dialogue turn, and are reviewed by other students and instructors. If approved by reviewers, the changes resulting from the feedback are automatically implemented by retraining natural language classifiers, or by updating domain-specific dialogue content. The number of students using our Watson dialogue-based tutors is projected to grow from hundreds in current pilot trials to tens of thousands in the commercialization phase, generating a large-scale feedback effect. We expect that the number of retraining events will peak initially, and then drop off as the tutor training becomes more comprehensive for each knowledge domain. This paper discusses our progress and identifies the challenges that we'll be addressing in future work.

Keywords: Dialogue based tutor, Natural language, Student annotation, Large-scale feedback.

1 Dialogue-based Tutors

1.1 General Approach

Dialogue-based tutoring systems (DBTs) use natural language to engage students in a Socratic conversation about a topic of interest. In general, the tutor's intent is to facilitate learning by guiding students through relevant concepts using questions, hints, pumps, etc., as well as giving feedback. Students engaging with DBTs also benefit from improvements in their memory and comprehension of the source text [7], since they are required to provide natural language responses.

Steenbergen-Hu and Cooper [8] conducted a meta-analysis of 39 studies evaluating the use of intelligent tutoring systems (including DBTs) in higher education. Overall, a large advantage (g=0.86) was observed for intelligent tutoring systems compared to self-reliant learning activities or alternatives with no tutoring.

The learning principles and strategies employed by DBTs include encouraging constructive behaviors and self-explanations [2], deep reasoning questions [5], and conceptual understanding through scaffolding [9]. For example, AutoTutor [4] is a well-

Back to Table of Contents

known DBT, which is based on an expectation-misconception discourse model, and has demonstrated significant learning gains over non-interactive learning materials in various math and science domains.

The scalability of DBTs is limited in general, since the effort required to adapt tutors for a new domain or textbook is non-trivial, even if the use case and learning goals are clearly defined. In particular, it's difficult to scale up the understanding of student responses across different knowledge domains.

1.2 Watson DBT

IBM and Pearson have partnered to develop dialogue-based intelligent tutoring systems at large scale. Currently, we are integrating Watson DBTs into a Pearson courseware product called REVEL, which provides online learning support for higher education courses. In the commercialization phase starting later this year, Watson DBTs will cover ten undergraduate textbooks in the domains of developmental psychology, physiological psychology, abnormal psychology, sociology, government and communication.

The overall design of Watson DBT is detailed in [10]. Our framework enables Watson DBT to be scaled up to a large number of textbooks and domains. Additional content is required for each new domain, but the same conversation flow and the same type of domain model are used across all domains. The domain model contains learning objectives, enabling objectives that support the learning objectives, and prerequisite relations among the learning objectives.

Currently, the dialogue content consists of questions, hints, reference answers, misconceptions and fill in the blanks. There are a total of around 600 questions (including main and hint questions) per textbook. Much of this content is created manually by domain experts from existing textbooks, and the rest is extracted automatically. We're also planning to introduce multimodal activities that are useful in a conversation, such as concept maps [1], gestures/actions [3], and speech/audio [6].

We've developed Watson DBT using a conversational strategy similar to the fivestep tutoring approach in AutoTutor: (i) tutor poses a question, (ii) student attempts to answer, (iii) tutor provides brief evaluation as feedback, (iv) collaborative interaction to improve the student answer, (v) tutor checks that student understands. A dialogue manager controls the flow of the conversation for each learning objective. Watson DBT begins by asking the student a main question, and continues to respond to students with appropriate feedback and a set of hint questions that will improve their understanding.

There are two levels of natural language classification operating in Watson DBT for understanding student responses. The first level is an intent classifier to detect utterances that remain consistent across many different domains. The domain-specific training for this classifier is seeded from the dialogue content. The second level is a response analyzer that assesses student answers against reference answers and drives the feedback given to students. This uses a supervised machine learning model, which requires less than ~35 scored student answers per question.

Back to Table of Contents

Watson DBT also has a learner model, which estimates a student's mastery as they progress through each topic. The mastery is estimated for both learning and enabling objectives, to give high-level and more detailed views of student understanding.

Our plans for preliminary evaluations of Watson DBTs are described in [11]. The learning efficacy of Watson DBT is still being investigated in our current pilot trials, with formal evaluations scheduled to begin later this year. We will publish those results later in another paper.

2 Feedback system

The feedback system is designed to improve the natural language understanding and the dialogue content of Watson DBT. A tutor with a comprehensive understanding of student responses is able to cover a broad set of topics within a domain, and to recognize an array of variations in the language style.

Annotations are collected from students during their chat with the tutor, and they are used to drive the feedback process. In our feedback system, an annotation is a predefined, structured comment that students can select quickly and intuitively, whenever they have an issue with the tutor response. For example, if students disagree with the tutor assessment of their answer, they may select "My answer is entirely correct" as their annotation.

A flowchart of the feedback system is shown in Fig. 1. Student annotations are stored in a transcript database, along with the entire conversation. Each annotation is then sent to a number of active reviewers, who can see the annotation category, the student response and other context about the dialogue turn. If the reviews cross an approval threshold, the relevant API is triggered to automatically retrain the natural language classifiers, or to update the dialogue content.

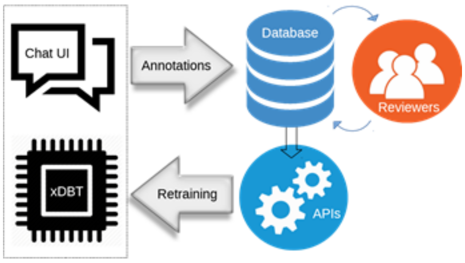


Fig. 1. Feedback system for Watson DBT using student annotations.

2.1 Collection of Annotations

When students provide an answer to a tutor question, they are given feedback by the tutor on their answer. Students have the ability to annotate that dialogue turn by selecting a predefined, structured comment from a list shown by the tutor. The annotation categories can be grouped by tutor functionality, here is an example list:

Back to Table of Contents

a) Natural language understanding

“The tutor didn't understand my synonym”

“My answer is partly correct”

“My answer is entirely correct”

“I asked a question” or *“That was my answer”*

b) Dialogue content

“The fill-in word isn't a key concept”

“The last question is too confusing”

c) Graphical interface

“The tutor is not responding”

“I couldn't type in the text box”

These annotations were chosen for Watson DBT by analyzing their relative frequency in student feedback from pilot trials. Only a subset of the annotations are visible in each dialogue turn, depending on the context. For example, students can choose “The fill-in word isn't a key concept” only after answering a fill-in-the-blank type question. In addition, students may leave general comments, and new categories can be introduced to reflect other issues.

Our current pilot trials indicate that students welcome the opportunity to give feedback when they perceive the tutor response isn't quite right. When we reach the commercialization phase, we expect that tens of thousands of students (or more) will start using Watson DBT, generating a large number of annotations.

2.2 Review Process

The annotations are stored in the transcript database of Watson DBT, and are presented to students or instructors who use the same tutor. As an example, active reviewers may see the feedback annotation: “My answer is entirely correct” plus (i) the tutor question (ii) the reference answer and (iii) the student answer. Reviewers can either agree or disagree with the annotation, and when the sum of their responses crosses an approval threshold, the decision is used to trigger a retraining event for the tutor.

The approval threshold in our pilot trials is set at 80%, i.e., at least 4 out of 5 reviewers are required to validate a student annotation. Other approaches are still being investigated, as we're seeking to balance quality and speed when implementing changes in the tutor training. We will publish those approaches later in another paper.

In general, approval thresholds can vary for different types of annotations, and may also depend on the level of expertise of each reviewer. For example, an annotation that's validated by an expert instructor will most likely lead to retraining of the tutor.

As expected, the review process is quicker and more effective if a large number of reviewers are available. We're considering all options to motivate high mastery students (and instructors) to become reviewers, e.g., (i) enable students to unlock UI customizations after a certain number of contributions, (ii) a rewards scheme awarding points or levels that can be traded by students for access to other textbooks, (iii) extra credit, or

Back to Table of Contents

access to hidden topics with the same tutor, and (iv) monetary compensation for instructors.

2.3 Automation of Changes

Approved annotations are used to automatically trigger retraining events by invoking the relevant API for each annotation category. For example, if a student claims that their answer is entirely correct and the reviewers agree, then an API for the response analyzer will retrain the corresponding reference answer.

The types of retraining events are specific to the natural language classifiers and the dialogue content of Watson DBT:

- (i) new synonyms are added to the domain synonym dictionary,
- (ii) the response analyzer is retrained using additional student answers,
- (iii) the intent classifier is retrained to handle misclassifications (e.g., to distinguish between student answers and student questions),
- (iv)domain keywords are updated,
- (v) the hint questions asked by the tutor are updated.

The training changes are implemented on an incremental basis, though it may be more efficient to run a batch of changes for some annotation categories.

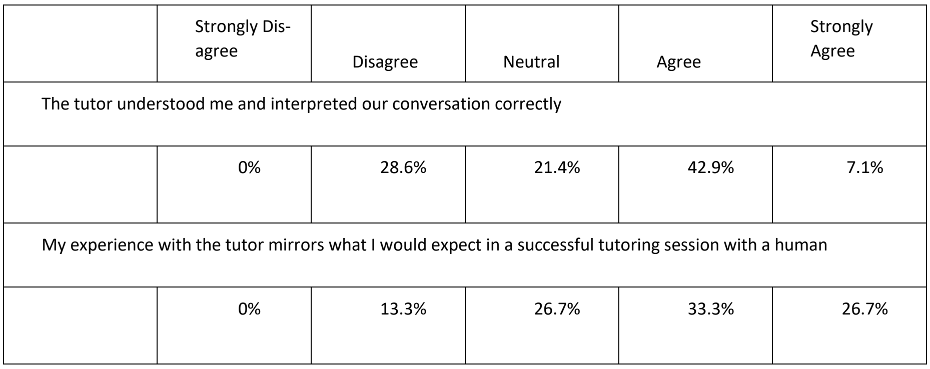
2.4 Effect of Student Feedback

The effect of training changes on tutor performance is monitored using a set of metrics for the domain-specific training and the dialogue quality. We use: (i) the accuracy of the intent classifier, (ii) F1 scores for the response analyzer, and the proportions of those classifications that have a strong adverse effect on the student experience, e.g., correct answers classified as incorrect, and (iii) the level of student satisfaction with the dialogue quality.

The metrics in (i) and (ii) can be measured automatically by using test data, as well as manually by analyzing a sample of conversation transcripts. It's important to monitor if feedback changes are having the intended effect, either at regular intervals, or after a set number of events for each type of retraining, so that changes can be rolled back if a significant negative impact is observed.

The base performance of the tutor is determined using only the training seeded from the initial dialogue content. Baselines in our pilot trials measured 95% accuracy for the intent classifier, and F1 scores ranging between 0.45 - 0.55 for the response analyzer. Fig. 2 shows the type of survey results collected about the dialogue quality from students who participated in the trials.

The number of training changes is expected to peak in the initial period of use, and then it should drop as the tutor gains a more comprehensive understanding. However, if changes persist at a high rate in the longer term, then the dialogue content and/or training will have to be inspected manually, at least for some annotation categories.



	Strongly Dis-agree	Disagree	Neutral	Agree	Strongly Agree
The tutor understood me and interpreted our conversation correctly					
	0%	28.6%	21.4%	42.9%	7.1%
My experience with the tutor mirrors what I would expect in a successful tutoring session with a human					
	0%	13.3%	26.7%	33.3%	26.7%

Fig. 2. Students were asked to rate how they felt about these statements after completing a tutoring session using Watson DBT.

3 Conclusion and Future Work

3.1 Summary

We've developed a feedback system for DBTs that uses valid student annotations to improve the understanding of natural language student responses and the dialogue content. We ex(i)pect that our approach to student feedback will work in general for all types of dialogue-based tutors, and will scale tutors more efficiently across multiple domains. Later this year, when Watson DBTs start interacting with a large number of students, we will monitor the effect of the feedback to verify that changes are having a positive effect on the tutor performance.

3.2 Challenges

There are still a number of challenges ahead for the complete implementation of our student feedback system. We'll address these questions in future work, as Watson DBTs enter the commercialization phase:

- (i) how can we motivate reviewers to participate in the long term?
- (ii) what is a good balance of quality and speed for the approval threshold, so that student annotations can be processed quickly and effectively?
- (iii) if necessary, can we roll back training changes from events before the last iteration without negative side effects?
- (iv) can we extend the scope of the feedback system to other types of training, aswe add new tutor features such as multimodal activities?

Acknowledgments

This work was supported by the IBM Watson Education team. We also thank all of the students who participated in the pilot trials for their invaluable participation.

Back to Table of Contents

References

1. Ainsworth, S., Prain, V., Tytler, R.: Drawing to learn in science. Science 333(6046), 10961097 (2011).
2. Chi, M. T. H.: Active-Constructive-Interactive: A conceptual framework for differentiating learning activities. Topics in Cognitive Science 1, 73–105 (2009).
3. Goldin-Meadow, S.: Beyond words: The importance of gesture to researchers and learners. Child development 71(1), 231-239 (2000).
4. Graesser, A. C.: Learning, thinking, and emoting with discourse technologies. The American Psychologist 66(8), 743-757 (2011).
5. Graesser, A. C., Person, N. K.,: Question asking during tutoring. American Educational Research Journal 31, 104 –137 (1994).
6. Litman, D. J., Rosé, C. P., Forbes-Riley, K., VanLehn, K., Bhembe, D., Silliman, S.: Spoken versus typed human and computer dialogue tutoring. International Journal of Artificial Intelligence in Education 16(2), 145-170 (2006).
7. McNamara, D. S.: The generation effect: A detailed analysis of the role of semantic processing. ICS Technical Report 92-2. Boulder, CO (1992).
8. Steenbergen-Hu, S., Cooper, H.: A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. Journal of Educational Psychology 106(2), 331-347 (2014).
9. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist 46(4), 197-221 (2011).
10. Chang, M., Ventura, M., Ahn, J., Foltz, P., Ma, T., Dhamecha, T. I., Marvaniya, S., Watson, P., D'helon, C., Wetzel, A., Packard Haas, A., Banaszynski, K., Behrens, J., Nelson, G., Sundararajan, S. C., Tejwani, R., Afzal, S. Mukhi, N.: Dialogue-based tutoring at scale: Design and Challenges. To be published in Proceedings of International Conference of the Learning Sciences (ICLS), London, UK (2018).
11. Ventura, M., Chang, M., Foltz, P., Mukhi, N., Behrens, J., Ahn, J., Ma, T., Dhamecha, T. I., Marvaniya, S., Watson, P., D'helon, C., Tejwani, R., Afzal, S.: Preliminary Evaluations of a Dialogue-based Tutor. To be published in Proceedings of Artificial Intelligence in Education (AIED), London, UK (2018).

Back to Table of Contents