Dialogue-based tutoring at scale: Design and Challenges

Maria Chang[1], Matthew Ventura[2], Jae-wook Ahn[1], Peter Foltz[2], Tengfei Ma[1], Tejas I. Dhamecha[1], Smit Marvaniya[1], Patrick Watson[1], Cassius D'helon[1], Amy Wetzel[2], Andy Packard Haas[2], Kaitlyn Banaszynski[2], John Behrens[2], Gailene Nelson[2], Sharad C. Sundararajan[1], Ravi Tejwani[1], Shazia Afzal[1], Nirmal Mukhi[1]

[1] IBM Research

[2] Pearson Education

Email: maria.chang@ibm.com, matthew.ventura@pearson.com, jaewook.ahn@us.ibm.com, peter.foltz@pearson.com, Tengfei.Ma1@ibm.com, tidhamecha@in.ibm.com, smarvani@in.ibm.com, pwatson@us.ibm.com, cdhelon@us.ibm.com, amy.wetzel@pearson.com, andy.packard-haas@pearson.com, kaitlyn.banaszynski@pearson.com, john.behrens@pearson.com, gailene.nelson@pearson.com, sharads@us.ibm.com, rtejwan@us.ibm.com, shaafzal@in.ibm.com, nmukhi@us.ibm.com

Abstract: In 2016, IBM and Pearson announced a partnership to deliver a next generation learning service in the form of a dialogue-based tutor. Dialogue-based tutoring systems have demonstrated efficacy, but they are difficult to design and scale across domains. We have developed a framework for enabling digital courseware with a dialogue-based tutoring experience that can be applied to new domains with additional domain-specific content, but without re-design of the conversation flow or use case. The framework uses a content model that's consistent across domains, which enables a general dialogue-based tutoring strategy. We identify several challenges to this approach, as well as recommendations for future work.

The Pearson-IBM partnership

In 2016, IBM and Pearson announced a partnership to deliver next generation learning services in the form of a digital tutoring system. The aim of this partnership was to create learning experiences powered by Pearson's high-quality content and IBM's Watson technologies. While there have been many successful intelligent tutoring systems (ITSs), our challenge was scaling the process across multiple disciplines and titles. This paper is organized into 4 parts. First, we briefly review the value of dialogue-based tutoring systems. Second, we describe what we call the Watson dialogue-based tutor. Third, we will describe some of our approaches to scaling and evaluating the Watson dialogue-based tutor. Finally, describe the limitations of our approach and recommendations for future work.

Dialogue-based tutoring systems

Dialogue-based tutoring systems (DBTs) are an approach to ITSs that create a learning experience driven by natural language dialogue and classification of student natural language responses (e.g., Graesser, 2011). DBT conversations can be described as Socratic because the tutor guides the student through concepts via dialogue moves, which can include questions, hints, and other prompts. DBTs have been claimed to support a variety of learning principles and strategies, like encouraging constructive behaviors and self-explanations (M. T. H. Chi, 2009), deep reasoning questions (e.g., Graesser & Person, 1994), and conceptual understanding through scaffolding (e.g., VanLehn, 2011). DBTs require students to construct natural language responses, which can have a positive impact on memory and comprehension of source text (e.g., McNamara, 1992). DBTs can also provide immediate feedback to facilitate learning (Shute, 2008).

For example, AutoTutor is a DBT, i.e., an ITS that initiates discourse with a student. The discourse patterns of the earliest AutoTutor were inspired by analyses of approximately 100 hours of non-expert human tutoring interactions (Graesser, 2011), which showed that students in need of tutoring are not active, self-regulated learners, and are not aware of their knowledge deficits. This affects how students converse: they do not effectively take command of the tutorial agenda and typically ask only 6-8 genuine information-seeking questions per hour. In contrast, tutors set 100% of the agenda, introduced 93% of the topics, presented 82% of examples, and asked 80% of the questions. Tutors do this by invoking a curriculum script of topics, problems, questions, and examples to drive a Socratic tutoring dialogue with students. Based on this tutor analysis, AutoTutor was designed to control the conversation through an expectation-misconception discourse model of tutoring (Graesser, 2011). This consists of a set of anticipated correct ideal answers (expectations) and a set of invalid answers frequently expressed by students (misconceptions). AutoTutor follows this design in a five-step tutoring framework: (1)

tutor poses a question/problem, (2) student attempts to answer, (3) tutor provides brief evaluation as feedback, (4) collaborative interaction to improve the answer, (5) tutor checks if student understands.

Efficacy of DBTs

Steenbergen-Hu and Cooper (2014) conducted a meta-analysis of 39 studies evaluating the use of ITSs (including DBTs) in higher education. The researchers found an overall, moderate, positive effect (g = .35) favoring the use of ITS over other instructional conditions. When compared specifically to alternatives that were either "self-reliant learning activities" or no-treatment conditions, the use of ITSs appeared to offer a large advantage (g =.86). AutoTutor in particular has shown significant learning gains over non-interactive learning materials in a variety of math and science domains: computer literacy, physics, biology, and critical thinking (Graesser, 2011). Typically, higher gains were found for more complex questions, such as "how" and "why" questions, versus shallow questions, such as "who" or "what" questions (Nye, Graesser and Hu, 2014).

Scalability of DBTs

While DBTs have demonstrated a wide range of possible behaviors and pedagogical strategies, building a DBT for a new domain, course, or textbook is a non-trivial task. Even when the use case and learning goals are clearly defined, creating the necessary domain models for these tutoring systems can be very challenging for domain experts. This is a general problem for ITSs, which is why the researchers behind the most widely adopted tutoring systems have also developed authoring tools (e.g., Aleven, McLaren, Sewall, & Koedinger, 2009).

Watson DBT

Watson dialogue-based tutor (WDBT) follows the design of AutoTutor in many respects, but with a content creation and iterative design cycle to support application to new domains. WDBT begins with deep reasoning questions and then provides hints to assist students to give a response that matches a set of assertions or knowledge components. An example transcript from an interaction with WDBT is shown in Table 1. WDBT is made up of 6 components to achieve this functionality: (1) Domain Model, (2) Dialogue Content, (3) Natural Language Response Classification, (4) Question Answering, (5) Learner Modeling, and (6) Dialogue Management.

Domain Model

The domain model defines the knowledge and skills we want students to learn. We create a domain model for a specific title (i.e. textbook) that breaks down the knowledge into *educational objectives* consisting of *learning objectives* and *enabling objectives*. Learning objectives are broad learning goal statements e.g., "Analyze physical changes that occur in middle adulthood." Enabling objectives are more granular learning goal statements that support the learning objective e.g., "Identify the physical benchmarks of change in middle adulthood." Educational objectives serve as the foundation for creating content and assessment in Pearson, so the same framework was used to enable WDBT. The domain model also contains misconception statements that are aligned to educational objectives. Domain experts create the domain models, choosing learning and enabling objectives that are particularly difficult for students or that would benefit from a conversational learning experience. For each title, domain models typically contain 24 learning objectives and 110 enabling objectives (about 5 enabling objectives per learning objective). Because the learning objectives follow the general organization of the title, the domain model acts as an extension of the core content in the title.

Dialogue Content

The dialogue content for WDBT consists of a mix of content created manually by domain experts, and automatically extracted content. Subject matter experts (SMEs) are responsible for authoring:

- *Main question and answer pairs* aligned to both learning and enabling objectives. The questions are broad, high-level conceptual questions, to assess deep comprehension of source text. WDBT delivers these questions to students. The corresponding answers represent ideal responses and are used to evaluate student answer attempts.
- *Hint questions and assertions* for every main question and answer pair. Assertions are independent knowledge components that make up a main answer. Hints are intended to elicit an assertion from a student. WDBT delivers these hint questions to students.
- *Misconception statements and corresponding feedback messages* for every applicable learning and enabling objective. In the event that response classification detects that a student is expressing a misconception, WDBT delivers the corresponding feedback message.

Other forms of dialogue content are automatically or semi-automatically extracted:

- *Question-answer pairs* based on key terms and other sources within the textbook.
- *Concept groupings* based on key concepts found in the unstructured text. These groupings are automatically extracted (1) but validated by a subject matter expert.
- *Fill in the blank prompts* that are dynamically created within the dialogue using an assertion and response classification gap analysis.

This approach to content organization is applied across titles to facilitate faster adaptation to new titles, but it is not without its limitations. There is still a considerable amount of human effort required for this process, although it is much less than what would be required if new conversational scripts (and associated content) were created for each domain. This points to the inherent trade-off between human effort savings and dialogue richness that can be addressed by better automatic content extraction techniques.

Table 1: An excerpt of an interaction between a learner and Watson Dialogue-Based Tutor (WDBT). Dialogue moves and responses from WDBT are left aligned. Learner responses are right aligned.

What is the nature of gender differences?	
	its is common for gender to be defined by sex
Looks like I have something different… How do sociologists define sex?	
	biological differneces
Well said. Sex refers to the biological differences between men and women. There are 2 key facts left in this review. What is gender?	
	idk
No problem. I'll help you get this! Gender refers to socially created differences between men and women. How are gender differences primarily constructed?	
	socially
Very good! Gender differences are primarily socially constructed, not biologically, constructed.	

Natural Language Response Classification

WDBT has two levels of natural language classification. The first level is a student intent classifier to detect classes of student utterances that are expected to be relatively consistent for different titles. The second level is a domain-specific response analyzer that is used to drive the expectation-misconception dialogue.

Student Intent Classification

The student intent classifier identifies different types of student utterances, such as answers to a hint question, indications that they don't know something (or would like a hint), expressions of frustration, and requests to give feedback. The student intent classifier also detects whether the student is asking an on-topic question. This intent classifier is trained in WDBT using the Watson Conversation service (2).

Student Response Analyzer

When a student responds to a tutor question, our student response analyzer (SRA) classifies the student response into three categories (correct, incorrect, and partially correct). It also analyzes the gap between the expected answer and the student response, which can be used to dynamically generate fill in the blank prompts. As part of feature extraction, SRA identifies important parts of the assertions using syntactic and semantic processing of sentences. A set of nonlinear classifiers are learned from labeled training data and domain-general corpora. Since different features and classification models capture various aspects of semantics, the final classification is obtained as an ensemble.

Many general natural language understanding challenges are relevant here, including co-reference resolution, anaphora, and dealing with differences in vocabulary and presentation of the knowledge. Even though the expectation-misconception tailored framework simplifies response classification, there are still challenges to handling mismatches in language specificity and breadth of content (e.g., when a student utterance does not match the most contextually relevant expectation but is not necessarily incorrect). These issues are mitigated with high quality training data.

Question Answering

WDBT can answer student questions and recommend further questions by using the question-answer pairs that are included in the domain model, as well as question-answer pairs that are automatically extracted from unstructured text. If the student's question matches one of the questions in the domain model (other than the one currently asked by the tutor), or it matches one of the automatically extracted questions, then WDBT is able to provide an answer.

Learner Modeling

The goal of the learner model is to estimate a student's degree of mastery of each domain model topic (based on student performance on assessments) to identify the optimal zone of proximal development and recommend mastery appropriate content. However, modeling a learner's developing knowledge in a domain general ITS framework presented a cold-start challenge. To overcome this, we adapted a technique used to estimate player ranks in competitive games (Glickman 2013, Pelánek et al. 2017). Each assessment was assigned its own mastery rating, as if it were a player in a competitive game. We labeled students correctly answering an assertion a "win" against that assertion, incorrect answers a "loss," and partial answers a "draw," updating both the student and the question's mastery rating accordingly. This enabled the model to converge quickly to reasonable estimates of learner skill and question difficulty without requiring large datasets.

Dialogue Management

The dialogue management system controls the rules for how to deliver the conversation with a student. In many cases these rules depend on inputs from the other 5 components. Each conversation with a student is centered around a learning objective (and the main question, answer, and enabling objectives associated with it). The dialogue begins with the tutor asking the main question. When the student responds to this question, WDBT must analyze the student response with respect to the expected correct answer and known misconceptions. Based on the results of response classification, the tutor provides qualitative feedback (e.g. positive, negative, neutral, comforting) and if required, one of the following scaffolds to elicit more information from the student:

- Hint question
- Misconception feedback (if the student expressed a misconception)
- Dynamically generated fill in the blank prompt
- Asking for clarification/disambiguation
- Asking another question (i.e. a main question for an associated enabling objective)

This choice depends on the student's progress towards the conversation goal and the tutor's estimates of student mastery from the learner model. For instance, if the tutor detects that the student is struggling, it may choose to focus on a different enabling objective before returning to the main learning objective question.

Because this is an open natural language dialogue, the student may respond in many different ways, including utterances that are off-topic, expressions of frustration or confusion, or questions that the student wants answered. These alternative utterances from the student may require the tutor to temporarily shift away from the current conversation flow and respond appropriately to the student. For example, if the student asks a known question, the tutor can provide a succinct answer, followed by a transition back to tutoring around the main question (e.g. "Let's get back to our original question...").

Content Collection and Iterative Design Lifecycle

We use a content collection and iterative design lifecycle that enables us to conduct experiments with incomplete and/or early-stage models. The lessons learned from these activities can be fed into the initial authoring and training stages of our pipeline, to improve the overall system or test alternative models and strategies.

Content collection tools for training SRA

We developed a student answer collection tool that (1) collects student answers to our SME-created hint questions, and (2) enables SMEs to score the answers as being correct, partially correct, or incorrect. We collected roughly 35 student answers for every WDBT hint question to ensure robust accuracy around SRA. A title typically has around 1000 hint questions resulting in the collection of over 35,000 student answers per title. This training improved our SRA model by allowing it to adapt to specific titles. Note that determining ground truth is not trivial, since even SMEs may disagree in their ratings of student answers.

Automated Content Curation

To scale to multiple domains, we must adopt a process that semi-automatically takes unstructured domain content, such as textbooks or webpages, and produces structured data that may be used to drive an educational dialogue experience. Such data can take many forms, including question and answer pairs, concept maps/graphs, annotated content chunks, anecdotes, and/or examples. Our current implementation of WDBT uses a subset of the content elements that may be used in a dialogue. A critical feature of our framework is that domain experts continue to be responsible for manual parts of content curation and for the validation of automatically extracted data. A continual shift away from creation towards validation and automated extraction will lead to greater scalability.

Iterative Design

The iterative design phase of WDBT is intended to be an ongoing source of feedback on the design of the system and training data for underlying models. Over the past year we have been conducting micro experiments with small groups of individuals (often, but not necessarily students) to provide us with qualitative data on user experience. We have also conducted experiments using a hybrid Wizard of Oz (WoZ) design (Ahn et al., 2017), where the Wizard is assisted by the prototype. These pre-pilot activities have enabled us to improve our response classifiers, refine our proposed dialogue flows, improve our UI design, and make recommendations for how to author clearer dialogue content (e.g. avoiding vague or potentially subjective questions).

Evaluation

The Watson dialogue-based tutor is currently being used by students through a Pearson-developed interactive learning environment, called REVEL. As of this writing, about 430 students have used WDBT across 10 higher education institutions. We are collecting feedback in the form of written statements from learners, dialogue usage metrics, and manually annotated transcripts. Usage statistics show that our dialogue strategy produces different experiences for different students, with conversations averaging about 15 turns per session, and sometimes running over 100 turns per session. As with our pre-pilot activities, these sources of feedback enable us to improve our response analysis models and fine tune our dialogue strategies, but on a larger scale.

We are in the process of preparing more formal studies to evaluate the impact of WDBT on learning. We have two types of studies we are conducting across 10 titles: in-class pilots and learning experiments. In-class pilots involve using WDBT directly with real students in real courses. Learning experiments involve highly controlled lab-based studies comparing students using WDBT with a control group.

Progress on Journey to Scale

Through the Pearson-IBM partnership, over approximately 12 months, we were able to create domain models and dialogue content for 8 titles and build complete WDBTs for two titles. Moving forward we believe we can scale our approach to create WDBTs for approximately 30 titles per year. For intelligent tutoring systems in general, it is estimated that up to 300 hours of development are required to enable just 1 hour of student-tutor interaction, with sophisticated authoring environments like Cognitive Tutor Authoring Tools (CTAT) reducing that range to 50-100 hours (Aleven et al. 2009). Based on our internal estimates, we can enable 1 hour of student-tutor interaction with approximately 40-45 hours of development from subject matter experts. However, this is not exactly a level comparison because of the adaptive nature of WDBT: not all students view all content and some students may have a shallower (and less time consuming) experience than others. As we collect more usage data from students, we can further understand exactly how WDBT's ratio between subject matter expert hours and students' time on task compares with other DBTs. Furthermore, as we progress to new domains, these numbers may vary as we discover new challenges, opportunities for efficiency, and improvements to overall tutoring strategies.

Conclusion & Future Work

We have developed an approach for scalable dialogue-based tutoring systems that uses a content creation, deployment, and feedback cycle. We believe that this reduces the amount of effort required to create a DBT for a new domain because the content creation and DBT development are closely connected and interdependent. Several challenges remain in the areas of natural language understanding and learner modeling. However, when it comes to quickly scaling to new domains, the two most immediate challenges involve support for richer interactions and more automated content extraction techniques. A wider range of interactions could include support for multimodal activities such as drawings (Ainsworth 2011), gestures/actions (Goldin-Meadow, 2000), speech and/or other audio (Litman et al. 2006). An extended interaction model might also support richer forms of explanation, such as analogies (Thagard 1992), qualitative modeling (Bredeweg & Forbus 2003), argumentation (Erduran & Jiménez-Aleixandre, 2008), and peer learning (Topping, 2005). In reality, rich one-on-one tutoring involves a variety of pedagogical strategies that engage multiple senses and types of reasoning. We believe that conversational systems that support any of these behaviors in a scalable way would represent a major advance in intelligent learning environments. However, in order to achieve these capabilities at scale, we need to advance the state of the art in automated content extraction as well. Currently, the types of content used in our WDBT represent a small subset of what is potentially available from the titles. As tools for ingesting structured content improve, the potential use of new dialogue moves (including those that use examples, annotated diagrams, images, analogies, and causal models) becomes tractable. We believe that our framework for content organization, dialogue management, and user feedback provides a baseline for reducing the effort required of applying DBT to new domains, both enabling teachers to focus on complex dialog moves, and supporting future technical development.

Endnotes

(1) https://www.ibm.com/watson/services/natural-language-understanding/
(2) https://www.ibm.com/watson/services/conversation/

References

Ahn, J. W., Watson, P., Chang, M., Sundararajan, S., Ma, T., Mukhi, N., & Prabhu, S. (2017, June). Wizard's Apprentice: Cognitive Suggestion Support for Wizard-of-Oz Question Answering. In International Conference on Artificial Intelligence in Education (pp. 630-635).

Ainsworth, S., Prain, V., & Tytler, R. (2011). Drawing to learn in science. Science, 333(6046), 1096-1097.

Aleven, V., McLaren, B.M., Sewall, J., & Koedinger, K.R. (2009). A New Paradigm for Intelligent Tutoring Systems: Example-Tracing Tutors. *International Journal of Artificial Intelligence in Education, 19*(2), 105-154.

Bredeweg, B., & Forbus, K. D. (2003). Qualitative modeling in education. AI magazine, 24(4), 35.

Chi, M. T. H. (2009). Active-Constructive-Interactive: A conceptual framework for differentiating learning activities. Topics in Cognitive Science, 1, 73–105.

Erduran, S., & Jiménez-Aleixandre, M. P. (2008). Argumentation in science education. Perspectives from classroom-Based Research. Dordre-cht: Springer.

Glickman, M. E. (2012). Example of the Glicko-2 system. Boston University.

Goldin-Meadow, S. (2000). Beyond words: The importance of gesture to researchers and learners. Child development, 71(1), 231-239.

Graesser, A. C. (2011). Learning, thinking, and emoting with discourse technologies. *The American Psychologist, 66*(8), 743-757.

Graesser, A. C. & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31*, 104 –137.

Litman, D. J., Rosé, C. P., Forbes-Riley, K., VanLehn, K., Bhembe, D., & Silliman, S. (2006). Spoken versus typed human and computer dialogue tutoring. International Journal of Artificial Intelligence in Education, 16(2), 145-170.

McNamara, D. S. (1992). *The Generation Effect: A Detailed Analysis of the Role of Semantic Processing*. Citeseer.

Nye, B. D., Graesser, A. C., & Hu, X. (2014) AutoTutor and Family: A review of 17 years of science and math tutoring. *International Journal of Artificial Intelligence in Education, 4*(24), 427–469.

Pelánek, R., Papoušek, J., Řihák, J., Stanislav, V., & Nižnan, J. (2017). Elo-based learner modeling for the adaptive practice of facts. User Modeling and User-Adapted Interaction, 27(1), 89-118.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153-189.

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, *106*(2), 331-347.

Thagard, P. (1992). Analogy, explanation, and education. Journal of research in science teaching, 29(6), 537-544.

Topping, K. J. (2005). Trends in peer learning. Educational psychology, 25(6), 631-645.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197-221.

Acknowledgments

This work was heavily supported by the broader teams at Pearson Education, IBM Research (T.J. Watson and India Research Labs), and IBM Watson Education. We also thank the students and participants of our micro pilots for their invaluable feedback.