AUGMENTING CLASSROOMS WITH AI FOR PERSONALIZED EDUCATION

Ravi Kokku, Sharad Sundararajan, Prasenjit Dey, Renuka Sindhgatta, Satya Nitta, Bikram Sengupta

IBM Research

ABSTRACT

Intelligent tutoring systems (ITS) have been a topic of great interest for about five decades. Over the years, ITS research has leveraged AI advancements, and has also helped push the boundaries of AI capabilities with grounded usage scenarios. Using ITSs along with classroom instruction to augment traditional teaching is a canonical example of how humans and machines can work together to solve problems that are otherwise overwhelming and non-scalable individually. The experiences of personalized learning created by (1) seamless orchestration of human decision-making at few critical points with (2) scalability of cognitive capabilities using AI systems can drive increased student engagement leading to improved learning outcomes. By considering two particular use-cases of early childhood learning and higher education, we discuss the challenges involved in designing these complex human-centric systems. These systems integrate technologies involving interactivity, dialog, automated question generation, and learning analytics.

***Index Terms*—** Intelligent tutoring systems, ITS, AI in Education, Dialog-based tutoring, Assessments

1. INTRODUCTION

Technology-driven intelligent tutoring systems (ITS) provide a way for computing systems to autonomously teach learners by giving them immediate and personalized feedback. Intelligent tutoring systems have been envisioned since the dawn of AI but it is only over the last two decades that significant progress has been made in this field. ITS systems have been envisioned as aids for learners both inside and outside classrooms, primarily as supplemental learning aids. In classrooms, ITS systems act as scalable augmentation aids to traditional multi-student settings for automating a number of tasks in the teaching and learning process. This in turn helps teachers focus their effort on critical tasks that humans are inherently good at (ranging from interventions with empathy to nurturing creativity), which machines cannot necessarily emulate effectively. Recent advances in AI are driving stronger human-machine collaboration in the learning process, especially making aspects of Intelligent Tutoring scale up to larger masses of students. This paper focuses on the use of ITS systems within classrooms.

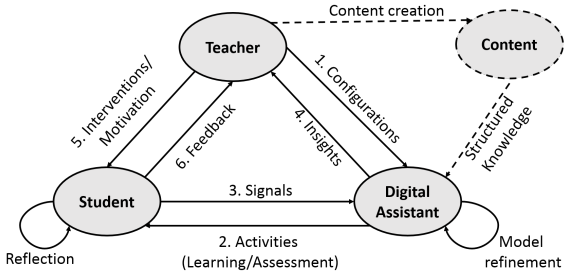

Fig. 1. Interactions between a teacher, a digital assistant and a student.

In one envisioned scenario (Figure 1), the traditional teacher-student instruction model can be augmented by including a digital assistant such as an ITS which can be configured by a teacher with specific high level learning objectives. The ITS then exposes various personalized learning and assessment activities to each student, observes various performance and behavioral signals from the student, and provides insights to the teacher. The teacher in-turn can use the insights specific to each student or in aggregate as a class to customize interventions, and provide informed motivation and remediation to the student for continued involvement in the learning process.

This collaborative process is significantly facilitated by two important recent advancements. The first is the digital presentation of learning and assessment activities. Assessment activities in particular are benefiting from significantly higher levels of machine automation and processing than was possible previously. The second advance is the increased access students have to smart devices that have a multitude of sensors and that can provide data for better understanding of the student's context and usage patterns. More importantly, these advancements have resulted in a deluge of data that can be mined using AI algorithms to gain significant insights into behavioral processes during learning. However, in typical usage situations, the data can also have a high degree of variability and variety. To be able to exploit this data for better learning outcomes, we require a combination of (a) deriving actionable intelligence from the large amount of data, and (b) abstract decision making on a continuum of knowledge. This is where the teacher and digital-assistant symbiosis can pro-

vide a richer experience to the students.

Our efforts to build tutoring systems for both early childhood and higher education have enabled us to explore: 1.The design space of automation of the learning process, 2: Understand the different components of the process that can be automated with state-of-the-art technologies, and 3: Identify the components that are better done by human teachers. We discuss the design space we explored, our experience in implementing our solutions, and several challenges we encountered, some of which still require additional research.

The rest of the paper is organized as follows. Section 2 discusses the history of AI in education and the state of the art in human-machine collaboration in the domain of education. Section 3 discusses our efforts in building tutoring solutions and the challenges involved. Section 4 concludes.

2. BACKGROUND

The primary goal of intelligent tutoring systems in formal education since the mid 1950s (when it was termed Computer Assisted Instruction) has remained the same, which is to create efficient learning environments, accomplished mostly by enabling instructors to do what they do best. Initially the systems were expensive, rule based and limited to easier tasks such as linear presentation of information, record keeping, progress tracking, or drill and practice. Modern day ITSs are cheaper, more accessible and use sophisticated machine learning to solve harder problems such as enabling more natural interactions like textual and spoken dialog, evaluating student answers, contextual information retrieval, or automatically generating assessments [1] [2].

More recently, in [3], the authors highlight key sources of learning gains namely frequent error repair, self-explanation, breaking the interaction plateau, rich natural language understanding, explorable explanations and tutor personas and call out for a renewed focus on making tutoring systems more engaging. The case studies that we discuss in this paper are an extension of the effort in [3], except for the difference that the original work was piloted for K-12 and our work is in early childhood and higher education.

ITSs have their roots in Artificial Intelligence (AI) [1] [4] [5]. Just as in AI, the underlying core problems in ITS revolve around knowledge representation and reasoning. How do we model the domain and the student? How do we induce knowledge from the content? How do we reason about the content and data to derive insights? In [5], Self questions the theoretical foundations of ITS and suggests that AI researchers might have retreated from the ITS arena when they realized the need for more fundamental work on mental models, language understanding, knowledge representation etc. Fast forward over a quarter century, how are we doing now? We attempt to address the above question in this paper. In (Figure 2) we depict the different capabilities that are best handled by machines and humans separately and some that can be done by both.

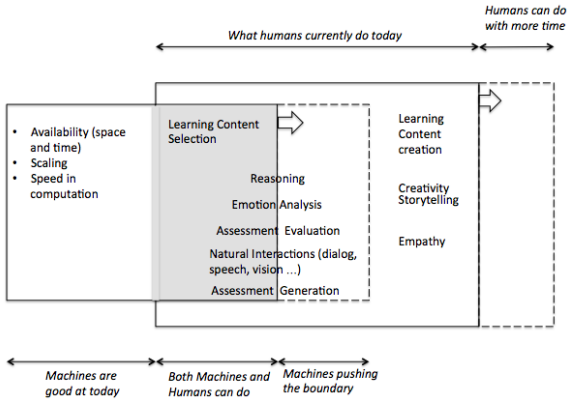

Fig. 2. Human and Machine capabilities (not exhaustive).

The general paradigm of Human-machine collaboration that can solve hard problems is not new: it has been used in several instances that would otherwise have been difficult for humans or machines individually. In some instances, machines perform intelligent tasks and humans intervene only where machine has low confidence in what it has done. In other instances, machines require human involvement at specific points in a work-flow, where only humans can take intelligent higher order decisions. For instance, the protein folding problem that was unsolved for ten years was eventually solved in three weeks by a seamless orchestration of human and machine capabilities [6]. Similarly the DARPA shredder challenge demonstrated how a difficult task such as re-assembling a shredded document can be converted into a human-in-the-loop problem and solved with accuracy as well as speed [7].

Several human-in-the-loop systems have also been explored specifically in the education domain using reinforcement learning [8, 9]. The hope is that teachers can support higher order tasks such as taking abstract or affect-aware decisions for intervention, whereas the AI system can find the most effective learning paths from numerous possibilities. While it is easy to appreciate the benefits of the collaboration between teachers and digital assistants in aiding traditional instruction, the key challenges involve pushing the capabilities of machines systematically (Figure 2), which can make the collaboration fruitful and effective. Our efforts in building multi-modal learning technologies in the early childhood and higher education domains target this exact problem of pushing the boundaries.

3. MULTI-MODAL LEARNING

3.1. Higher Education

In higher education, mixed-initiative multi-modal conversations have been shown to help students construct knowledge [10]. A dialog based tutoring system answers students' questions, assesses student understanding by analyzing their nat-

ural language responses to tutor initiated questions, personalizes the next activity suitable for each learner by modeling their engagement and mastery. Dialog based tutors offer a rich opportunity to elicit self-explanation from students [11] and extract signals of confusion, misconception, frustration and knowledge gaps all of which can be relayed back to the instructors via a dashboard.

We have built a conversational tutoring system that orchestrates multiple learning activities such as natural language exchanges, visual concept grouping, worked examples, fill-in-the-blanks to name a few. The rationale behind that is that not all questions, particularly those belonging to different Bloom's levels lend themselves to the same kind of experience. We discuss a few of those activities below.

3.1.1. Assessment Evaluation

For tutor initiated questions, students respond in short sentences. Existing dialog based tutors use different techniques to analyze student responses [12]. In our tutoring system, we use an ensemble of machine learning models to analyze and classify the student responses. Here, we train the models and compare the ability of the system to mimic a human tutor. The responses are graded both by humans and the tutor. Figure 3 shows the distribution of F1-scores [13] between two humans and a human and the system. F1-score is the harmonic mean of precision and recall, and is a measure of a test's accuracy. The distribution is similar with the humans having higher agreement (>0.8) for a larger percentage of responses. However, the distribution indicates that for a significant percentage of responses, the system can mimic humans. We have also found that student responses and our analyses of them can directly inform content creation, particularly the formulation of questions and reference answers. Short answer analysis remains a challenge because of the myriad ways in which ideas can be expressed in natural language and the need to compare student responses not just against a reference answer but against a knowledge base in order to provide meaningful feedback.

3.1.2. Assessment Generation

During the course of the dialog, if a student provides a response that is partially correct, the tutoring system generates a fill-in-blank assessment questions automatically, that varies for each student depending on the knowledge gap. The knowledge gap is identified by comparing the gap in the student response and the reference answer(s). Variations in the questions can be automatically generated to provide a different experience to a student, each time they converse with the system. Domain experts still need to initially validate the assessments so the challenge is to scale the validation step. With sufficient training data, we are confident that the system can learn the most important concepts.

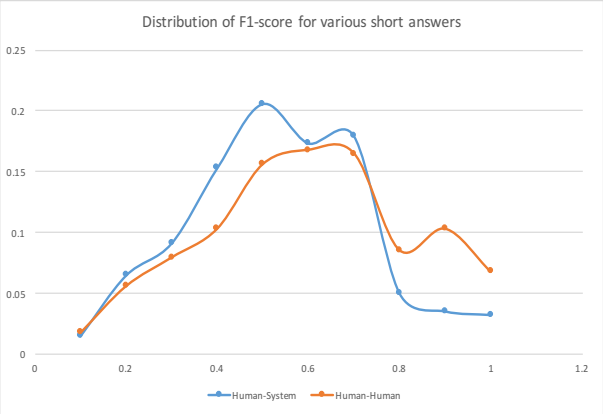

Fig. 3. Distribution of F1-scores for student responses comparing two humans and a human and the system.

3.1.3. Knowledge Extraction and Induction

Extracting examples and concepts: When students struggle to answer questions even with hints, pumps or prompts, we present contextual examples to help them understand the concepts better. We automatically extract some examples based on metadata linking the examples with learning objectives. But there are a few challenges to fully automating example extraction in some STEM domains as they demand a very good understanding of the examples to ensure they are relevant. Additionally we use the Watson Natural Language Understanding (NLU) service [1] to extract concepts and keywords and apply clustering approaches to create concept groupings. We then surface these as alternative assessments. We require a human-in-the-loop to help validate the groupings and train the system initially but are improving our algorithms to create richer concept clusters automatically.

Extracting question and answers: Answering student questions requires the system to identify frequently asked questions (FAQ). These questions can then be used as input to train a question answering system such as, the Watson Conversation service [2]. There are several valuable sources like forums and learning management systems that can be used to extract FAQ [14]. Similarly, definitions from the content can be extracted to generate questions and answers. We are currently not focusing on deeply conceptual questions.

In order to test our conversational system, we built a Wizard of Oz (WoZ) application [15] but with a twist. When a student provides a response to a tutor-initiated question, the response would first go to the backend (Watson Tutor in this case) for analysis. The system returns feedback, which now a human wizard can choose to accept and pass on to the student or reject and substitute. This allowed us to debug our dialog flows. We also added the capability to annotate every turn

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

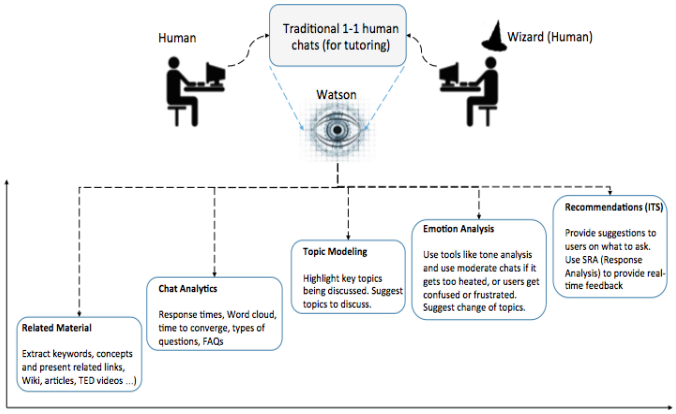

Fig. 4. Possible capabilities for a machine in the loop.

with metadata that would later inform the short answer scoring as well as sentiment analysis. We plan to extend the WoZ application to handle other capabilities as shown in Figure 4.

3.2. Early Childhood

In building early childhood solutions, we have been exploring automated assessment generation, assessment evaluation through gamification [16], and contextual visual recognition [17] techniques that enable in-class tools for testing the understanding of children and providing learning reinforcement activities as a part of the digital exposure. The dashboards of insights derived out of click-stream and performance data collected through gamified learning and assessment activities are used by teachers in turn to select specific topics of instruction. For instance, in a vocabulary application we built for classrooms [16], the dashboards accessible to teachers highlight the words that most students are struggling to understand, and hence create an evidence-based prioritized list of words that a teacher can take up on any particular day. In what follows, we describe specific solution components to drive such applications, and highlight some challenges for future work.

3.2.1. Assessment Generation

Assessments play an important role in the overall learning process, since they provide a way to continuously measure the learner's level of understanding, and personalize their learning. For the vocabulary application we built, we automatically generate Multiple Choice Questions (MCQ) that are multimodal (the questions were presented using text-to-speech and the choices of correct answers and distractors were images to be chosen from by the child).

In order to relieve the teacher from worrying about assessments, a machine should be able to generate questions of various levels of conceptual difficulty. For multiple choice question (MCQ) generation, this boils down to generation of different kinds of distractors (wrong answer options). Generating MCQs with appropriate level of distractors for each child, based on his prior knowledge is an interest problem for future research. For instance, to test the understanding of a word *insect*, an *arachnid* (e.g. a spider) is a harder distractor than a *lion*. Secondly, MCQ generation needs to ensure that the images themselves that are being presented as correct answers and distractors are of the right level of visual complexity. Depending on prior knowledge, an image can have varied level of visual complexity for a child. We are currently working on algorithms that combine the visual complexity of an image along with its conceptual complexity to come up with very high quality MCQ assessments.

3.2.2. Multimodal Interactions

Since most children at an early age (preK to KG) cannot read or write, their primary communication modalities are speech and visual interfaces. Hence a digital assistant needs to use multimodal interaction to engage with the learners. There are several challenges, however, since visual and speech recognition are still inherently hard for machines, especially under challenging situations of child speech, noisy environments, poorly lit environments, child's improper use of camera devices etc. We are attempting to build systems that work with these imperfections of speech and visual recognition and augment them with contextual information from the interaction the child has with the system. However, speech recognition for children (aged 3-5) is a long-standing missing feature, and significant advancements, including speech data collection for training speech models are necessary. Similarly, state of the art visual recognition solutions provide a list of possible labels for each image provided, along with a confidence measure. For better learning experience, the accuracy of such solutions need to be significantly improved and contextualized to the learning setting for a child, several challenges of which we address in [17]. Overall, significant potential of improvement exists for each of the AI technologies to enable ITSes to effectively augment traditional classroom instruction.

4. CONCLUSION

Augmenting traditional instruction with intelligent tutoring systems can relieve teachers from doing activities that are more effectively done by machines, and lets the teachers focus on what humans are good at. We are instantiating the use of AI technologies in early childhood and higher education scenarios. We encountered several challenges in the process of this integration into traditional instruction and are addressing these challenges in our future research. Our exercise highlights the need for continued research in speech, vision and conversation modalities of interaction.

5. REFERENCES

[1] Hyacinth S. Nwana, "Intelligent tutoring systems: an overview," *Artif. Intell. Rev.*, vol. 4, no. 4, pp. 251–277, 1990.

[2] Beverly Park Woolf, *Building Intelligent Interactive Tutors: Student-centered Strategies for Revolutionizing e-Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007.

[3] S. C. Sundararajan and S. V. Nitta, "Designing engaging intelligent tutoring systems in an age of cognitive computing," *IBM Journal of Research and Development*, vol. 59, no. 6, pp. 10:1–10:9, Nov 2015.

[4] J. R. Carbonell, "Ai in cai: An artificial-intelligence approach to computer-assisted instruction," *IEEE Transactions on Man-Machine Systems*, vol. 11, no. 4, pp. 190–202, Dec 1970.

[5] John Self, "Theoretical foundations for intelligent tutoring systems," 1990.

[6] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popovi, M. Jaskolski, and D. Baker, "Crystal structure of a monomeric retroviral protease solved by protein folding game players," *Nat. Struct. Mol. Biol.*, vol. 18, no. 10, pp. 1175–1177, Sep 2011.

[7] Tom Geller, "Darpa shredder challenge solved," *Commun. ACM*, vol. 55, no. 8, pp. 16–17, Aug. 2012.

[8] Min Hyung Lee, Joe Runde, Warfa Jibril, Zhuoying Wang, and Emma Brunskill, "Learning the features used to decide how to teach," in *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, New York, NY, USA, 2015, L@S '15, pp. 421–424, ACM.

[9] Andrea L. Thomaz and Cynthia Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, no. 6, pp. 716 – 737, 2008.

[10] Arthur C. Graesser, Kurt VanLehn, Carolyn P. Ros, Pamela W. Jordan, and Derek Harter, "Intelligent tutoring systems with conversational dialogue," *AI Magazine*, vol. 22, no. 4, pp. 39–51, 12 2001.

[11] O. Popescu C. Torrey K. Koedinger V. Aleven, A. Ogan, "Evaluating the effectiveness of a tutorial dialog system for self-explanation," in *Proceedings of 8th International Conference on Intelligent Tutoring Systems*, 2004, pp. 443–454.

[12] Nabin Maharjan, Rajendra Banjade, and Vasile Rus, "Automated assessment of open-ended student answers in tutorial dialogues using gaussian mixture models," in *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017.*, 2017, pp. 98–103.

[13] "F1 Score," https://en.wikipedia.org/wiki/F1_score.

[14] Renuka Sindhgatta, Smitkumar Marvaniya, Tejas Dhamecha, and Bikram Sengupta, "Inferring frequently asked questions from student question answering forums," in *Proc. of the 10th International Conf. on Educational Data Mining*, 2017, EDM 2017, pp. 256–261.

[15] Jae-wook Ahn, Patrick Watson, Maria Chang, Sharad Sundararajan, Tengfei Ma, Nirmal Mukhi, and Srijith Prabhu, *Wizard's Apprentice: Cognitive Suggestion Support for Wizard-of-Oz Question Answering*, pp. 630–635, Springer International Publishing, Cham, 2017.

[16] Miles Ludwig and Satya Nitta, "Cognitive learning goes to school with ibm and sesame street," https://www.ibm.com/blogs/think/2017/06/sesame-street/.

[17] Vijay Ekambaram, Ruhi Sharma Mittal, Prasenjit Dey, Ravi Kokku, Aditya K Sinha, and Satya V Nitta, "Tell me more: Digital eyes to the physical world for early childhood learning," *Proceedings of the 10th International Conference on Educational Data Mining*, 2017.