Learning Behavior Policies for Interactive Educational Play

Samuel Spaulding
MIT Media Lab
Cambridge, MA 02143
Email:samuelsp@media.mit.edu

Cynthia Breazeal
MIT Media Lab
Cambridge, MA
Email: cynthiab@media.mit.edu

***Abstract*—Autonomous robotic systems have the potential to deliver significant benefits via social interaction. The development of such "socially assistive" robots could help address global shortfalls in caregiving resources, improving quality-of-life in areas such as nutrition, education, and autism therapy. Due to the difficulty of pre-specifying behavior across a wide range of scenarios, these robots must be able to learn social interaction behavior.**

One domain of particular interest is the development of autonomous educational robots. It is difficult to program robust models of interactive social behavior for educational co-play, in part because the interaction takes place in a high-dimensional state space with noisy state dynamics and sparse rewards. Moreover, an effective tutor must balance multiple objectives across different time scales, such as teaching new words, acquiring information about the student to model their knowledge, and keeping the interaction fun and engaging.

In this paper we describe some of the challenges of learning policies for educational co-play behavior, describe a work-in-progress game to serve as a testbed for learning algorithms, and outline a computational framework formulating educational co-play as a Multi-Obective POMDP.

I. Introduction

Developing robots capable of interactive, educational play could have profound impact on early childhood education. In this paper we describe work to develop an educational tutoring robot that can learn behavioral policies for playing an interactive second language vocabulary game with a child. Educational co-play is a challenging multi-objective problem: at each interval in the game, the robot must decide which of several distinct objectives to pursue, each of which provides rewards at different time scales.

One can model educational tutoring as a planning problem, with the robot acting on partially-observable *mental* states and the success of the task based on the mental state of the child after the interaction. As in navigation tasks, additional information about the current state can be acquired, but at a cost. In the case of the tutor, this might equal the cost of administering a test, which both takes time and is likely to negatively impact the rapport of the interaction.

Planning techniques and models, such as multi-objective reinforcement learning (MORL) or partially-observable Markov decision processes (POMDPs), have been successful applied to real-world tasks with similar computational structure, such as construction or navigation. However, "primarily interactive" tasks, such as tutoring or dialogue, differ from physical tasks (even those that require some degree of social interaction) in significant ways.

For example, in collaborative manufacturing, many planning-based solutions are cast as scheduling problems, solving for a sequence of tasks that the robot and human complete mostly independently, engaging on a single task together only when necessary. In contrast, "primarily interactive" tasks such as educational tutoring or co-play are socially interactive *per se*, the social interaction is not a by-product of the task, it *is* the task.

Such tasks place a high emphasis on legible behavior, that is, behavior that fits with, or helps the human interaction partner refine, their mental model of the robot's behavior. If the human and robot are solving different parts of the task separately, illegible behavior may have little consequence, as the human can focus on their own task without considering the robot's step-by-step behavior. In domains with a high degree of interaction throughout the task, illegible behavior may cause the complete failure of the interaction (e.g., if the human partner quits the task or decides to solve the task alone because they do not trust or cannot predict the robot's behavior).

More generally, the Markov assumption of many models may limit their success in primarily interactive tasks. Equating mental states to physical states may be a useful representation, but typical mental states are not "memoryless". The sequence of events leading to the present moment are of vital importance, especially for legible behavior. The very concept of legibility, (colloquially, the idea that "how you get there is just as important as where you end up") implies that relaxing the Markov assumption may be necessary to learn policies for legible socially interactive behavior.

II. Related Work

Despite these challenges, planning models, especially POMDPs, have been applied to derive educational tutoring behavior policies under certain conditions. Typically, this takes the form of learning a sequence of assignments or activities, intended to communicate a new concept in as few actions as possible. For example, Rafferty et al. modeled students as optimal Bayesian learners for abstract category learning and used a POMDP to guide a teacher's policy [4]. Whitehill et al. extended this work to the domain of second-language

vocabulary, using a POMDP-based model to simultaneously assess students' knowledge and learn a policy for when to *teach* a new word, *ask* about a specific word, and *test* the student, ending the learning session [6]. In both cases, the state space is the model of the student's knowledge, and does not contain any real-time social or interaction-based features.

Knox et al. learned a decision-tree based behavior policy for unstructured educational play with children from Wizard-of-Oz demonstrations of play, a technique dubbed "Learning from the Wizard" (LfW) [3]. While the learned behavior was largely successful at engaging the child and emulating human-like play, no significant educational effects were found. In this work, the state representation consisted entirely of real-time social and interaction-based features; the child's mental state was not explicitly modeled.

Gordon & Breazeal used a Bayesian active learning algorithm to model the child's mental state and select new words to introduce to a child in a vocabulary learning task. However, outside of the choice of words introduced, the robot's game behavior was scripted [1].

III. A Multiobjective Policy Model for Educational Co-play

In this section we describe a proposed game as an test domain for educational gameplay algorithms as well as a sketch of two models for teacher behavior and student knowledge.

A. Game Domain Description

A robot and child sit across from each other with a tablet in between. Each side of the tablet has a button, which the players tap to 'ring in' (the robot can press the button digitally) and answer when an image of a vocabulary word appears in the center of the tablet screen. The first player to tap their button is given an opportunity to say the word. If word is pronounced correctly (assessed by a speech recognition system), that player 'wins' the round. This game is largely about assessing children's productive vocabulary on nouns, though there are opportunities to assess the child's receptive vocabulary on the same words, e.g., by having the robot "win" a round by correctly pronouncing a word, then seeing if the child is able to do the same a few rounds later. Because the robot has immediate knowledge of when the picture will appear, we assume the robot can always ring in first to answer, if the policy calls for it. Thus, the behavioral policy of the robot largely determines how each round plays out.

B. Teacher Model

For each question, the robot can choose to ring in and win and say word correctly (*teaching* the child how to pronounce the word), not ring in and let the child demonstrate their ability to pronounce the word, or ring in and win and say word *incorrectly*.

These actions correspond to three different objectives. The first objective is to teach the child new words, which the robot achieves by demonstrating how to correctly pronounce a word. The second objective is to gain information about the child's state of knowledge, by allowing the student to pronounce words and evaluating their pronunciation. The robot chooses which words to teach based on the student model, and refining the student model can also help achieve the teaching objective. The third objective is to maintain the social fluency of the game, keeping the student engaged and enjoying the interaction. As gameplay goes on, the words taught and the robot's knowledge about the child increase monotonically. To be an effective tutor, the robot must keep the social and affective dimensions of the task balanced, allowing the interaction to continue, even if this objective may run counter to the other two goals in the short term.

To balance these three objectives, we propose to model the teacher as a multi-objective POMDP (MOPOMDP) [5], with both student model and real-time interaction features (e.g., facial expression analysis, game state, time since last action) represented in the state space, allowing the robot to draw on a rich feature set relevant to each objective.

C. Student Model

We propose to model student vocabulary learning as a Gaussian process, paired with active learning to decide which word to introduce at the next time step. Gaussian processes have a number of nice properties: at each point in the domain they produce Gaussian outputs, so the robot can model both how likely it is the child knows the word, as well as the uncertainty about that estimate and use this information during the active learning process to select the most appropriate word for the current objective. Kapoor et al. combined Gaussian processes with active learning for image classification, demonstrating that GPs can be a computationally efficient way to classify data in a large concept space [2]. For vocabulary words, the kernel values could be derived from an pedagogically informed concept-distance heuristic, or a phonetic similarity metric [1].

IV. Personalization

In addition, the most effective tutors are those that can personalize their pedagogical behavior to the specific child. The benefits of personalized tutoring are wide-ranging and well known, and the challenge of developing *personalized* policies for educational co-play could be addressed through transfer learning or other forms of domain adaptation.

V. Conclusion

In this paper we have described a number of challenges related to deriving autonomous social interaction behavior for educational co-play. We have also presented a work-in-progress game environment, and Teacher and Student models for learning social interaction behavior for educational tutoring: a MOPOMDP model that balances competing objectives in an educational game and a Gaussian process model of student learning that allows for efficient assessment of the student's knowledge via active learning.

REFERENCES

[1] Goren Gordon and Cynthia Breazeal. Bayesian active learning-based robot tutor for children?s word-reading skills. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI-15)*, 2015.

[2] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Gaussian processes for object categorization. *International journal of computer vision*, 88(2):169–188, 2010.

[3] W Bradley Knox, Samuel Spaulding, and Cynthia Breazeal. Learning from the wizard: Programming social interaction through teleoperated demonstrations. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 1309–1310. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[4] Anna N Rafferty, Emma Brunskill, Thomas L Griffiths, and Patrick Shafto. Faster teaching via pomdp planning. *Cognitive Science*, 2015.

[5] Diederik Marijn Roijers, Shimon Whiteson, and Frans A Oliehoek. Point-based planning for multi-objective pomdps. In *IJCAI*, pages 1666–1672, 2015.

[6] Jacob Whitehill and Javier Movellan. Approximately optimal teaching of approximately optimal learners. *IEEE Transactions on Learning Technologies*, 2017.