select the responses that satisfy the student best. The reward function ρ_i is defined to measure the satisfaction, in terms of the number of times the student explicitly rejects the responses, or asks about some "prerequisite" concepts.

The policy of the student agent $\pi_j(b)$ models the behavior of the student. It can be used to predict the most probable action, e.g. question, by the student in a given b. The prediction is based on the student's current study state, as well as information about the teacher agent, like observation history, preference, and rationality. When interpreting an ambiguous student input, the agent may use the prediction to help resolve the ambiguity. π_j is created and repeatedly improved for best match between predicted and actual student actions. Reward function ρ_j is defined in terms of the similarity between predicted and actual actions.

5. Summary

We propose an I-POMDP architecture for building a dialogue tutoring system, in which two agents model the teacher (system) and the human student, addressing the two challenging tasks of disambiguating student input and selecting appropriate responses. The major advantage of the architecture is that, in a state space where the states are not fully observable, the two agents can make decisions based on their beliefs about their states and each other's preferences and rationality. This fits well the nature of tutoring between a teacher and a student.

6. References

[1] K. Forbes-Riley, D. Litman, A. Huettner, and A. Ward, "Dialogue-learning correlations in spoken dialogue tutoring". In *Proc. of the 2005 conf. on AI in Education*, 2005.

[2] M. Frampton and O. Lemon, "Recent research advances in reinforcement learning in spoken dialogue systems", *The Knowledge Engineering Review*, Vol. 24(4), pp. 375-408, 2009.

[3] P. Gmytrasiewicz and P. Doshi, "A framework for sequential planning in multi-agent settings", *Journal of Artificial Intelligence Research*, Vol. 24, pp. 49-79, 2005.

[4] F. Jurcicek, B. Thomson, S. Keizer, M. Gasic, F. Mairesse, K. Yu, and S. Young, "Natural Belief-Critic: a reinforcement algorithm for parameter estimation in statistical spoken dialogue systems" *Proc. s of Interspeech10*, pp 90-93, 2010.

[5] L. Kaelbling, M. Littman, and A. Cassandra, "Planning and acting in partially observable stochastic domains", *Artificial Intelligence*, Vol. 101, pp. 84-98, 1998.

[6] D. Litman, and S. Silliman, "Itspoke: an intelligent tutoring spoken dialogue system". In *Proc. of Human Language Technology Conf. 2004*.

[7] S. Png and J. Pineau, "Bayesian Reinforcement Learning for POMDP-based dialogue systems", *Proc. Of Conf. on Acoustics, Speech and Signal Processing*, pp. 2156-2159, 2011.

[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: The MIT Press, 2005.

[9] J. Williams and S. Young, "Partially observable Markov decision processes for spoken dialog systems", *Elsevier Computer Speech and Language*, Vol 21, pp. 393-422, 2007.

[10] M. Woodword and R. Wood, "Learning from Humans as an I-POMDP", http://www.eecs.harvard.edu/~woodword/papers/woodword_2010_ipomdp_position_paper.pdf, 2010.