February 2022

Training Intelligent Tutors on User Simulators Using Reinforcement Learning

Paul BRICMAN, Matthia SABATELLI
University of Groningen

The vast majority of conversational assistants and dialogue systems being researched and deployed today are focused on information retrieval (IR) or task-based intent [10]. Dialogue systems designed for facilitating IR aim to understand what piece of information the user is searching for, before accessing it and relaying it to the user [7]. Similarly, task-based dialogue systems focus on understanding the user's intent with respect to a discrete action to be taken on their behalf. In contrast, relatively few resources have been invested in developing conversational assistants which help users reason to conclusions and identify appropriate actions *themselves*, instead of merely delegating the decision-making process, missing rich opportunities for promoting user agency.

However, this user-centered approach has been intimately adopted by a different research field, namely intelligent tutoring systems (ITS) [8, 15]. Developments in ITS aim to create systems which help actively cultivate both theoretical and practical skills in students through adaptive exercises and timely feedback. The goal of those systems is not merely to outsource the problem-solving skills of their users, but instead to nurture those very skills in their users. Adapting the difficulty of exercises to the current user skill level and providing hints at strategic points in the curriculum are only a handful of the techniques which appear to promote the effectiveness of ITS [17].

Unfortunately, developing an ITS requires both a solid understanding of the target domain and comprehensive knowledge about domain-specific obstacles which students might face. This is reflected in the niched nature of the ITS literature, consisting largely in analyses of the effectiveness and design choices of highly specialized systems [8]. In contrast, dialogue systems often strive for generality, being able to conversationally interface users to large bodies of factual knowledge and vast sets of actions to be taken on their behalf [10, 18, 14, 7, 11]. To overcome the difficulty in implementing domain-agnostic ITS, insights from sample-efficient dialogue system development might be leveraged.

One way in which dialogue systems overcome the low availability of user data across various domains is by using user simulators [12, 11, 7]. Essentially, the conversational assistants are trained and evaluated against simulators which exhibit procedurally-generated intents [10]. This can be seen as a specific instance of the more general sim2real paradigm from Reinforcement Learning (RL), where agents are trained in a virtual environment due to low interaction costs, before transferring their abilities to real-world environments [9]. While most sim2real developments are seen in interaction with 3D environments mimicking real-world physics, there has been limited interest in using dialogue as a conversational and textual environment to be navigated.

Moreover, despite the user simulators being trained on limited data to learn how to mimic user behavior in new situations, the training data is often limited to previous dialogue datasets [16]. This fails to take advantage of the rich world models internalized as latent representations by models trained to autoregressively predict text in a general setting, unbound by conversational contraints [3]. Language models trained on large corpora are likely to have encoded comprehensive knowledge about the world as an instrumental goal in reducing perplexity in text generation. This makes them prime candidates for being employed as plug-and-play world simulators in a broader sense, while still qualifying as user simulators. The language model essentially becomes the RL agent's environment, changing its state as the result of the agent's actions, and supplying the agent with rewards accordingly.

In this paper, we investigate the reliability of using pretrained language models as user simulators in developing conversational tutoring systems. We address this by using Q-learning to train a tutor agent which is rewarded for helping the user simulator reach seemingly pertinent and diverse answers to an overarching question selected to guide the specific dialogue episode (e.g. "What are the most important books ever written?"), in what we refer to as the Question Answering Assistance (QAA) task. Both state and action spaces are discretized for simplicity, where states are quantized regions of the latent space populated by user simulator replies, and actions are selected from a finite set of approximately 200 intervention prompts (e.g. "How could you approach this differently?"). Following 20,000 10-turn dialogues with the user simulator, the trained tutor agent is then tested for significance against an untrained tutor agent at helping human users answer sampled questions, to examine transfer performance.

Whenever the tutor agent performs an action (i.e. replies using an intervention prompt), the sandbox environment which coordinates the training procedure appends the reply to a growing dialogue data structure. The environment's dynamics are based on prompting the user simulator to generate a user reply to the tutor agent's intervention. The simulated user reply is similarly appended to the data structure representing the present dialogue. Besides observing a new environment state, the tutor agent also receives a numerical reward following each of its interventions, based on estimated question-answer relatedness and semantic similarity to previous replies. The agent is based on Q-learning, with the discrete state space entirely described by the cluster index to which the user reply is assigned to, and the action space consisting of the set of tutor interventions.

Following an extensive training process consisting of a tutor agent interacting with the user simulator for 20,000 10-turn dialogues, the tutor agent has then been evaluated in terms of performance in transfer to a human user by means of comparing it to an untrained (i.e. random) tutor agent. Both the trained and untrained tutor agents guided five 5-turn dialogues with a human user instead of the user simulator, while the reward has been tracked in the same way as during training, except for the human user replacing the user simulator's role in conversation. While the trained agent obtained a higher mean reward than the untrained agent during the human trials, an unpaired one-sided t-test revealed the difference not to be significant ($p = 0.08, t = 1.53$). Moreover, the moving average of the reward history recorded during training exhibited limited signs of reaching a plateau, continuing its increase even towards the end of the training procedure (see Fig. 1).

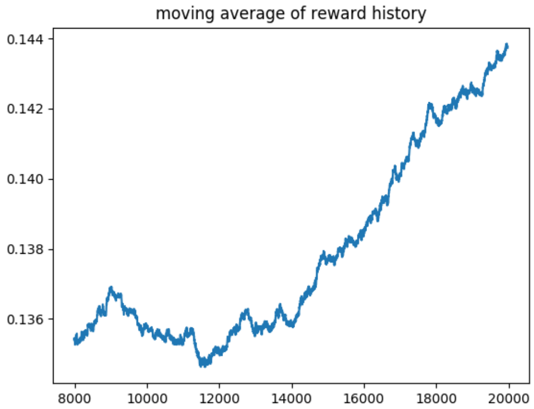

Figure 1. Smoothed reward history recorded during the tutor agent's training.

References

[1] cross-encoder/qnli-distilroberta-base · hugging face, . URL https://huggingface.co/cross-encoder/qnli-distilroberta-base.

[2] First quora dataset release: Question pairs, . URL https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs.

[3] GPT-neo, . URL https://www.eleuther.ai/projects/gpt-neo/.

[4] nreimers/MiniLM-l6-h384-uncased · hugging face, . URL https://huggingface.co/nreimers/MiniLM-L6-H384-uncased.

[5] P. Bricman. K-probes. URL https://github.com/paulbricman/k-probes/blob/3e98b87bd2bc4be748d0115beb608b97f41de8fb/prompts.json. original-date: 2020-11-25T18:45:30Z.

[6] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu. Plug and play language models: A simple approach to controlled text generation. URL http://arxiv.org/abs/1912.02164.

[7] P. Erbacher, L. Soulier, and L. Denoyer. State of the art of user simulation approaches for conversational information retrieval. URL http://arxiv.org/abs/2201.03435.

[8] A. C. Graesser, X. Hu, B. D. Nye, K. VanLehn, R. Kumar, C. Heffernan, N. Heffernan, B. Woolf, A. M. Olney, V. Rus, F. Andrasik, P. Pavlik, Z. Cai, J. Wetzel, B. Morgan, A. J. Hampton, A. M. Lippert, L. Wang, Q. Cheng, J. E. Vinson, C. N. Kelly, C. McGlown, C. A. Majmudar, B. Morshed, and W. Baer. ElectronixTutor: an intelligent tutoring system with multiple learning resources for electronics. 5(1): 15. ISSN 2196-7822. . URL https://stemeducationjournal.springeropen.com/articles/10.1186/s40594-018-0110-y.

[9] A. Kadian, J. Truong, A. Gokaslan, A. Clegg, E. Wijmans, S. Lee, M. Savva, S. Chernova, and D. Batra. Sim2real predictivity: Does evaluation in simulation predict real-world performance? 5(4):6670–6677. ISSN 2377-3766, 2377-3774. . URL http://arxiv.org/abs/1912.06321.

[10] X. Li, W. Wu, L. Qin, and Q. Yin. How to evaluate your dialogue models: A review of approaches. URL `http://arxiv.org/abs/2108.01369`.

[11] H.-c. Lin, N. Lubis, and S. Hu. Domain-independent user simulation with transformers for task-oriented dialogue systems. page 12.

[12] B. Peng, X. Li, J. Gao, J. Liu, and K.-F. Wong. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2182–2192. Association for Computational Linguistics. . URL `http://aclweb.org/anthology/P18-1203`.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. URL `http://arxiv.org/abs/2103.00020`.

[14] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, and J. Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325. Association for Computational Linguistics. . URL `https://aclanthology.org/2021.eacl-main.24`.

[15] A. C. Sales and J. F. Pane. The role of mastery learning in intelligent tutoring systems: Principal stratification on a latent variable. URL `http://arxiv.org/abs/1707.09308`.

[16] W. Wang, Y. WU, Y. Zhang, Z. Lu, K. Mo, and Q. Yang. Integrating user and agent models: A deep task-oriented dialogue system. URL `http://arxiv.org/abs/1711.03697`.

[17] J. H. Wong, S. S. Kirschenbaum, and S. Peters. Developing a cognitive model of expert performance for ship navigation maneuvers in an intelligent tutoring system. page 8.

[18] R. Zhou, S. Deshmukh, J. Greer, and C. Lee. NaRLE: Natural language models using reinforcement learning with emotion feedback. URL `http://arxiv.org/abs/2110.02148`.