香港中文大學 The Chinese University of Hong Kong

國立臺灣大學 National Taiwan University

Microsoft

Adversarial Advantage Actor-Critic Model for Task-Completion Dialogue Policy Learning

*Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, **Yun-Nung (Vivian) Chen**, Kam-Fai Wong*

Summary

➢ **Motivation**

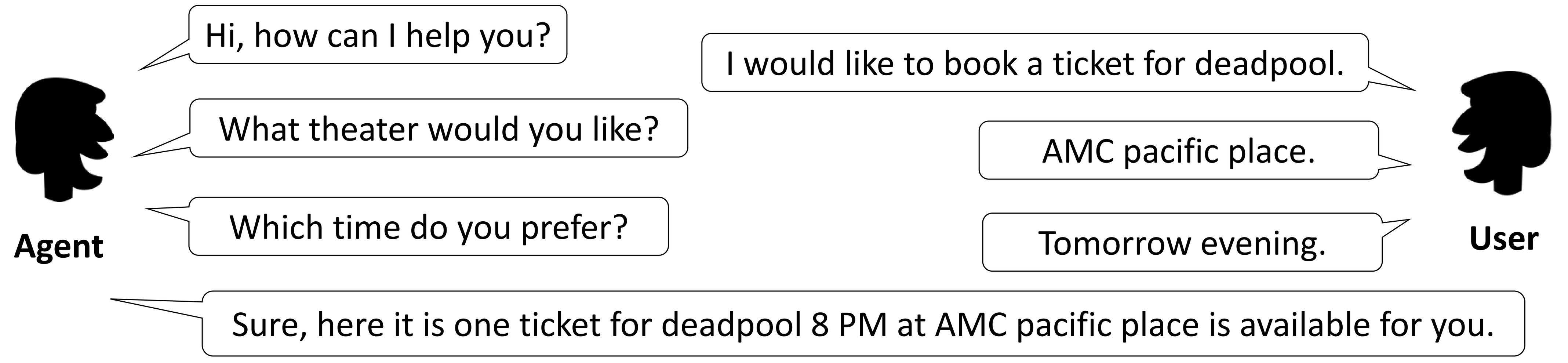
- Exploiting reinforcement learning for dialogue policy learning
- Exploration in the large state-action space is challenging
- Reward is delayed and sparse with a long trajectory

➢ **Approach**

- Propose an **Adversarial Advantage Actor-Critic** algorithm
- Leverage expert-generated dialogues as priors
- Use a discriminator to differentiate responses from an agent or human experts
- The output of discriminator as intrinsic reward to explore state-action regions similar to what human experts do

➢ **Results**

- Significant improvement of efficiency and performance on a movie-ticket booking domain



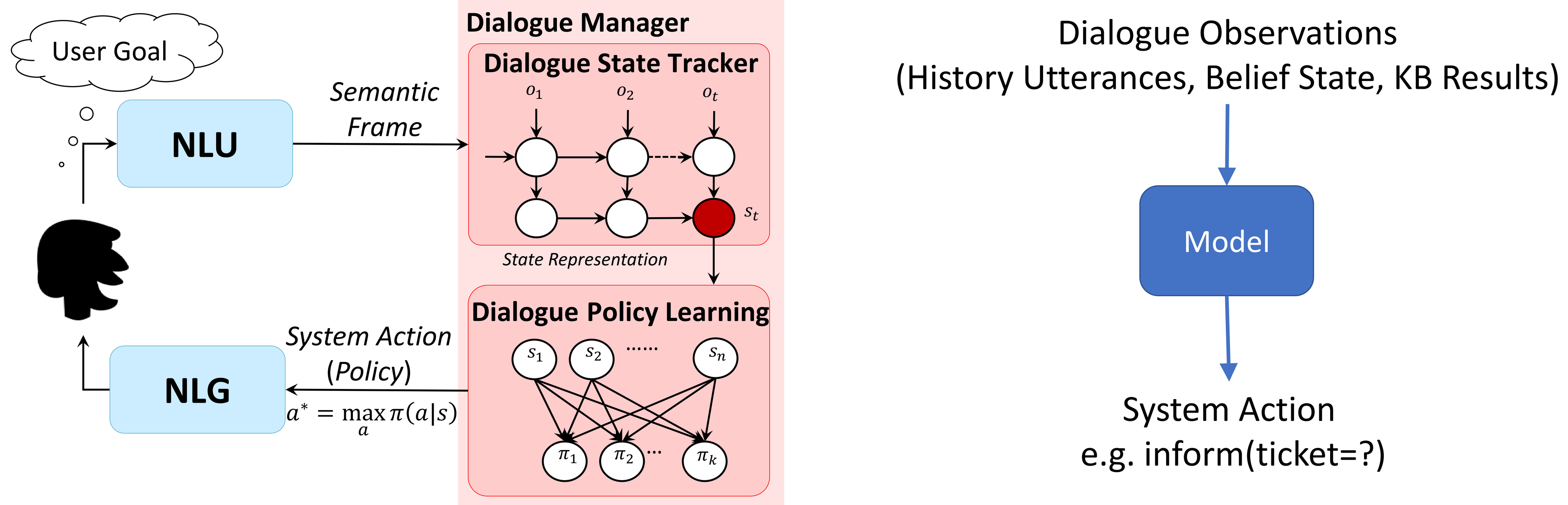
1. Task Definition

➢ ***Natural Language Understanding (NLU)*** turns natural language into intents and slot-values

➢ ***Natural Language Generation (NLG)*** turns system actions into natural language

➢ ***Dialogue Manager (DM)***

- tracks dialogue states and updates state accordingly
- interacts with the database
- takes state as input to output system action → **Dialogue Policy Learning**



2. Methodology

➢ **Advantage Actor-Critic for Dialogue Policy Learning**

- Find a policy π that maximizes the expected reward $R = \sum_t^{T-1} \gamma^t r_t$
- π is a parameterized probabilistic mapping function: $\pi_\theta(a \mid s) = P(A_t = a \mid s_t = s; \theta)$
  - Update θ with following gradients $\nabla_\theta J(\theta) = \mathbb{E}[\nabla_\theta \log \pi_\theta(a \mid s) Q^{\pi_\theta}(s,a)]$
  - Baseline function for reducing variance $\mathbb{E}[\nabla_\theta \log \pi_\theta(a \mid s) A^{\pi_\theta}(s,a)], A^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)$
  - TD error as an unbiased estimation $\mathbb{E}[\nabla_\theta \log \pi_\theta(a \mid s) \delta^{\pi_\theta}], \delta^{\pi_\theta} = r + \gamma V^{\pi_\theta}(s') - V^{\pi_\theta}(s)$

➢ **Adversarial Training**

- Actor π_θ as a ***generator*** G
- A ***discriminator*** D identifies state-action pair (s, a) from experts or G
- D can be viewed as a reward function extracted from experts' trajectories
- D is to maximize the probability of classifying each pair correctly

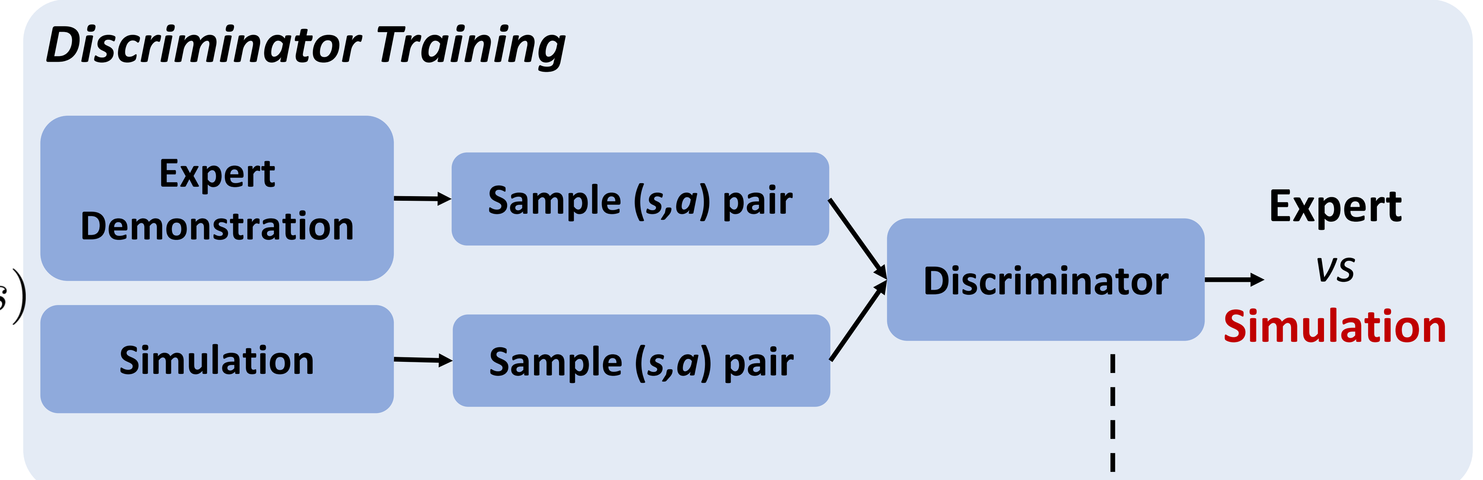
$$\min_{\theta_D} \mathcal{L}_D = -\mathbb{E}_{(s,a)\sim Simu} \log D(s,a;\theta_D) - \mathbb{E}_{(s,a)\sim Demo} \log(1 - D(s,a;\theta_D))$$

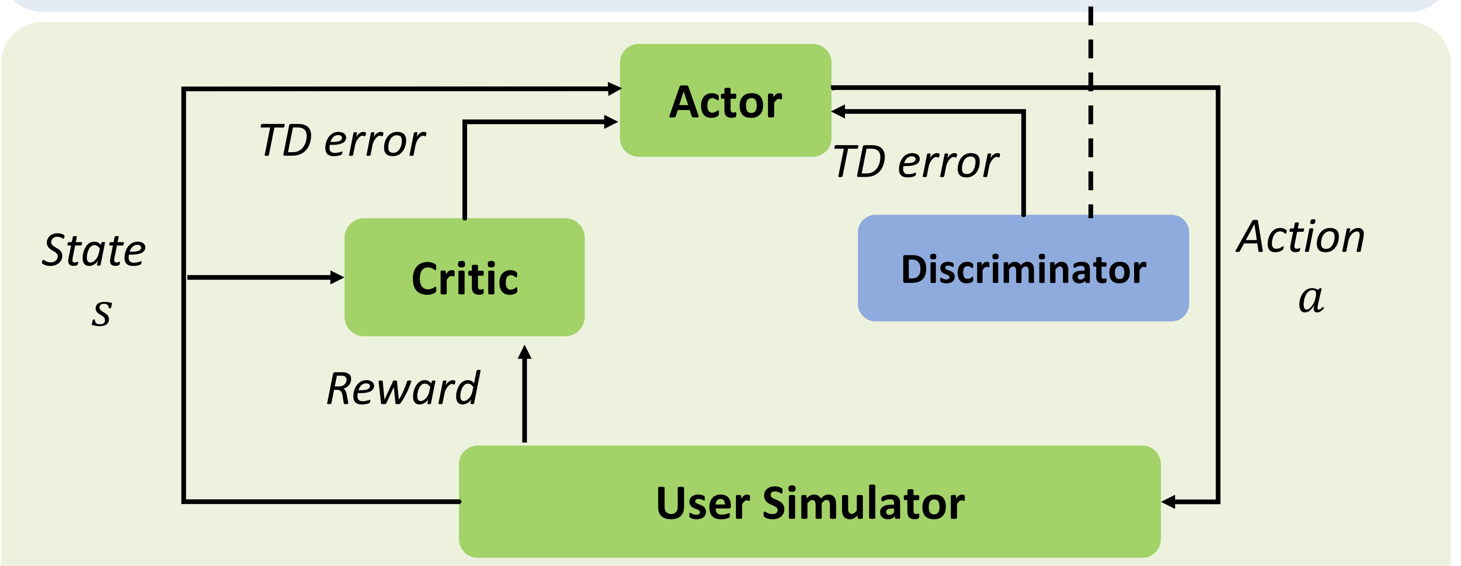
- **Actor π_θ (G) can be improved with $-log(1 - D(s,a))$ as the reward function**

$$\nabla_\theta J(\theta) = \mathbb{E}[\nabla_\theta \log \pi_\theta(a \mid s) \delta^{\pi_\theta}_{\text{GAN}}], \delta^{\pi_\theta}_{\text{GAN}} = r_{\text{GAN}} + \gamma V^{\pi_\theta}_{\text{GAN}}(s') - V^{\pi_\theta}_{\text{GAN}}(s)$$

➢ *Combine A2C with a reward function learned from experts' demonstrations with adversarial training.*

➢ *The discriminator D guides actor to explore state action regions where human experts will explore.*





Adversarial Advantage Actor-Critic

3. Experiments & Results

➢ **Dataset:** human-human conversations in the movie-ticket booking scenario

- collected via AMT and annotated by human experts
- *280 labeled dialogue with 11 average turns*
- *11 dialogue acts, 29 slots*
- *Informable (narrow down search), requestable (ask info from agent)*
- *Use a publicly available user simulator*

➢ **Evaluation**

- Success rate
- 10 run averaged learning curve
- 2000 dialogues for testing

➢ **Baselines**

- *Rule Agent*: handcrafted rule-based policy that in- forms and requests a hand-picked subset of necessary slots.
- *A2C*: trained with a pre-defined reward function and a standard advantage actor-critic algorithm
- *BBQN-Map Agent (AAAI'18)*: the best agent among a set of BBQN variants that has great efficiency for policy exploration in dialogue systems

➢ *Adversarial A2C learns faster and more stable with better exploration.*

Agent	Succ.	Turn	Reward
Rule	41.34	16.00	0.26
A2C	81.24	15.43	5.08
BBQN-MAP	81.56	18.75	5.00
Adversarial A2C	**87.52**	**13.52**	**5.93**

4. Conclusion

- We propose an adversarial advantage actor-critic model with ***efficient exploration***.
- The discriminator severs as an ***additional critic*** to guide policy exploration towards human-like one. It also has connection with inverse reinforcement learning that ***learns reward function***.
- Our experiments in a movie-ticket booking domain show its superiority.