Variational Reward Estimator Bottleneck: Learning Robust Reward Estimator for Multi-Domain Task-Oriented Dialog

Anonymous COLING submission

Abstract

Despite its notable success in adversarial learning approaches to multi-domain task-oriented dialog system, training the dialog policy via adversarial inverse reinforcement learning often fails to balance the performance of the policy generator and reward estimator. During optimization, the reward estimator often overwhelms the policy generator and produces excessively uninformative gradients. We propose the Variational Reward estimator Bottleneck (VRB), which is an effective regularization method that aims to constrain unproductive information flows between inputs and the reward estimator. The VRB focuses on capturing discriminative features, by exploiting information bottleneck on mutual information. Empirical results on a multi-domain task-oriented dialog dataset demonstrate that the VRB significantly outperforms previous methods.

1 Introduction

While deep reinforcement learning (RL) has emerged as a promising solution for complex and high-dimensional decision-making problems, the determination of an effective reward function remains a challenge, especially in multi-domain task-oriented dialog systems. Many recent works have struggled on sparse-reward environments and employed a handcrafted reward function as a breakthrough (Zhao and Eskenazi, 2016; Dhingra et al., 2017; Shi and Yu, 2018; Shah et al., 2018). However, such approaches are often unable to guide the dialog policy through user goals. For instance, as illustrated in Figure 1, the user can't reach the goal because the system (S1) that exploits the handcrafted rewards completes the dialog session too early. Moreover, the user goal usually varies as the dialog proceeds.

Inverse Reinforcement Learning (IRL) (Russell, 1998; Ng and Russell, 2000) and MaxEnt-IRL (Ziebart et al., 2008) tackles the problem of recovering reward function and using this reward function to generate optimal behavior. Although Generative adversarial imitation learning (GAIL) (Ho and Ermon, 2016), which exploits the GANs framework (Goodfellow et al., 2014), has proven that the discriminator can be defined as a reward function, GAIL fails to generalize and recover the reward function. Adversarial inverse reinforcement learning (AIRL) (Fu et al., 2018) enables GAIL to take advantage of disentangled rewards. Guided dialog policy learning (GDPL) (Takanobu et al., 2019) uses AIRL framework to construct the reward estimator for multi-domain task-oriented dialogs. However, these methods often encounter difficulties in balancing the performance of the policy generator and reward estimator, and produce excessively uninformative gradients.

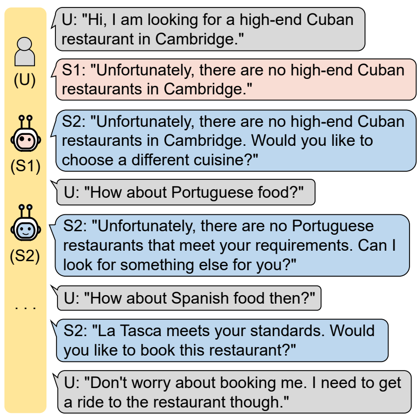

Figure 1: The system (S2) that uses well-specified rewards can guide the user through the goal while S1 can't.

In this paper, we propose the Variational Reward Estimator Bottleneck (VRB), an effective regularization algorithm. The VRB uses information bottleneck (Tishby et al., 1999; Alemi et al., 2016; Peng et al., 2019) to constrain unproductive information flows between dialog state-action pairs and internal representations of the reward estimator, thereby ensuring highly informative gradients and robustness. The experiments demonstrate that the VRB achieves the state-of-the-art performances on a multi-domain task-oriented dataset.

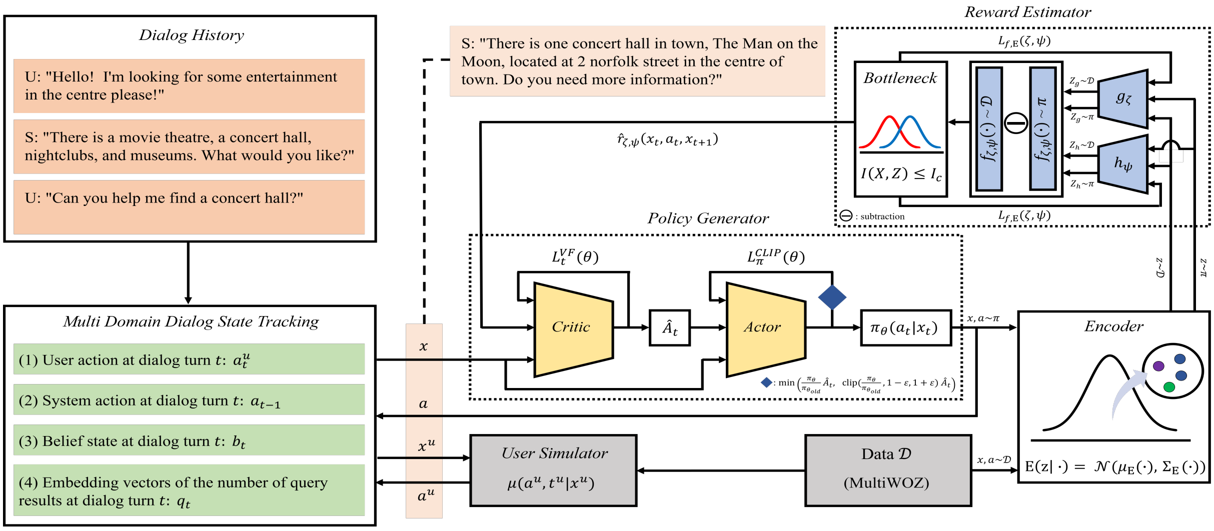

Figure 2: Schematic depiction of the Variational Reward Estimator Bottleneck.

2 The Proposed Method

2.1 Background: Reward Estimator

The reward estimator (Takanobu et al., 2019), which is a core component in multi-domain task-oriented dialog systems, evaluates dialog state-action pairs at dialog turn t and estimates the reward that is used for guiding the dialog policy through the user goal[1]. Based on MaxEnt-IRL (Ziebart et al., 2008), each dialog session τ in a set of human dialog sessions $\mathcal{D} = \{\tau_1, \tau_2, ..., \tau_H\}$ can be modeled as a Boltzmann distribution that does not exhibit additional preferences for any dialog sessions: $f_\zeta(\tau) = \log\left(\frac{\exp(\mathcal{R}_\zeta)}{Z}\right)$ where $\mathcal{R}_\zeta = \sum_{t=0}^{T} \gamma^t r_\zeta(x_t, a_t)$, Z is a partition function, ζ is a parameter of reward function, and $\mathcal{R}_\zeta$ denotes a discounted cumulative reward. The reward estimator is trained using gradient-based optimization[2]:

$$L_f(\zeta, \psi) = \mathbb{E}_{\tau\sim\mathcal{D}}[f_{\zeta,\psi}(x_t, a_t, x_{t+1})] - \mathbb{E}_{\tau\sim\pi}[f_{\zeta,\psi}(x_t, a_t, x_{t+1})] \tag{1}$$

2.2 Variational Reward Estimator Bottleneck

The Variational information bottleneck (Tishby et al., 1999; Alemi et al., 2016; Peng et al., 2019) is an information-theoretic approach that restricts unproductive information flow between inputs and the discriminator. Inspired by this concept, we propose a regularized objective that constrains the mutual information between encoded state-action pairs and original inputs, thereby ensuring highly informative internal representations and robust adversarial model. Our proposed method learns an encoder that is maximally informative regarding human dialogs. To this end, we employ a stochastic encoder and an upper bound constraint on the mutual information between the dialog states X and latent variables $\mathbf{Z}$:

$$\begin{aligned} L_{f,\mathbf{E}}(\zeta,\psi) = \mathbb{E}_{x,a\sim\mathcal{D}}[\mathbb{E}_{\mathbf{z}\sim\mathbf{E}(\mathbf{z}|x_t,x_{t+1})}[f_{\zeta,\psi}(\mathbf{z}_g, \mathbf{z'}_h, \mathbf{z}_h)]] - \mathbb{E}_{x,a\sim\pi}[\mathbb{E}_{\mathbf{z}\sim\mathbf{E}(\mathbf{z}|x_t,x_{t+1})}[f_{\zeta,\psi}(\mathbf{z}_g, \mathbf{z'}_h, \mathbf{z}_h)]] \\ \text{s.t.} \quad I(Z,X) \le I_c \end{aligned} \tag{2}$$

where $f_{\zeta,\psi}(\mathbf{z}_g, \mathbf{z}'_h, \mathbf{z}_h) = D_g(\mathbf{z}_g) + \gamma D_h(\mathbf{z'}_h) + D_h(\mathbf{z}_h)$ and D is modeled with nonlinear function. Note that $f_{\zeta,\psi}(\mathbf{z}_g, \mathbf{z}'_h, \mathbf{z}_h)$ is divided into the three terms $D_g(\mathbf{z}_g)$, $\gamma D_h(\mathbf{z'}_h)$, and $D_h(\mathbf{z}_h)$, based on GANs (Goodfellow et al., 2014), GAN-GCL (Finn et al., 2016), and AIRL (Fu et al., 2018). D_g represents the encoded disentangled reward approximator with the parameter ζ, and D_h is the encoded shaping term with the parameter ψ. Stochastic encoder $\mathbf{E}(\mathbf{z}|x_t, x_{t+1})$ can be defined as $\mathbf{E}(\mathbf{z}|x_t, x_{t+1}) = \mathbf{E}_g(\mathbf{z}_g|x_t) \cdot$

[1]For background on the dialog state tracker, the user simulator, and the policy generator, see Appendix A

[2]For the complete derivation, see Appendix B.2

Algorithm 1 Variational Reward Estimator Bottleneck

Initialize dialog policy generator π_θ and reward estimator $f_{\zeta,\psi}$
for $i \leftarrow 0$ ***to*** N **do**
  Obtain random samples from human dialog corpus $\mathcal{D}$
  Gather dialog sessions using user simulator $\mu(a^u, t^u|x^u)$ and policy generator $\pi_\theta(a|x)$
  Encode dialog sessions using stochastic encoder $\mathbf{E}(\mathbf{z}|\cdot) = \mathcal{N}(\mu_{\mathbf{E}}(\cdot), \Sigma_{\mathbf{E}}(\cdot))$
  Compute information bottleneck $\mathbb{E}_{x,a\sim\pi}[\mathrm{KL}[\mathbf{E}(\mathbf{z}|x)||r(\mathbf{z})]]$
  Update reward estimator $f_{\zeta,\psi}$ by optimizing $L_{f,\mathbf{E}}(\zeta,\psi)$ (Equation 3)
  Estimate reward function $\hat{r}_{\zeta,\psi}$ for each state-action pair
  Update state-value function $V(\mathcal{X})$ and dialog policy π_θ given the reward $\hat{r}_{\zeta,\psi}$ (Equation 4)
end

$\mathbf{E}_h(\mathbf{z}_h|x_t) \cdot \mathbf{E}_h(\mathbf{z'}_h|x_{t+1})$ which maps states to a latent distribution $\mathbf{z}$: $\mathbf{E}(\mathbf{z}|x_t) = \mathcal{N}(\mu_{\mathbf{E}}(x_t), \Sigma_{\mathbf{E}}(x_t))$. $r(\mathbf{z}) = \mathcal{N}(0, I)$ is standard gaussian and I_c stands for an enforced upper bound on mutual information. To optimize $L_{f,\mathbf{E}}(\zeta,\psi)$, VRB introduces a Lagrange multiplier φ:

$$\begin{aligned} L_{f,\mathbf{E}}(\zeta,\psi) = \mathbb{E}_{x,a\sim\mathcal{D}}[\,\mathbb{E}_{\mathbf{z}\sim\mathbf{E}(\mathbf{z}|x_t,x_{t+1})}[\,f_{\zeta,\psi}(\mathbf{z}_g, \mathbf{z'}_h, \mathbf{z}_h)]\,] - \mathbb{E}_{x,a\sim\pi}[\,\mathbb{E}_{\mathbf{z}\sim\mathbf{E}(\mathbf{z}|x_t,x_{t+1})}[\,f_{\zeta,\psi}(\mathbf{z}_g, \mathbf{z'}_h, \mathbf{z}_h)]\,] \\ + \varphi\left(\mathbb{E}_{x,a\sim\pi}\left[\,\mathrm{KL}[\mathbf{E}(\mathbf{z}|x_t, x_{t+1})]\,||\,r(\mathbf{z})\,\right] - I_c\right) \end{aligned} \tag{3}$$

where the mutual information between dialog states X and latent variable Z is

$$\begin{aligned} I(Z,X) &= \mathrm{KL}[p(\mathbf{z},x)||p(\mathbf{z})p(x)] \\ &= \int d\mathbf{z}\,dx\,p(\mathbf{z},x)\log\frac{p(\mathbf{z},x)}{p(\mathbf{z})p(x)} = \int d\mathbf{z}\,dx\,p(x)\mathbf{E}(\mathbf{z}|x)\log\frac{\mathbf{E}(\mathbf{z}|x)}{p(\mathbf{z})} \\ &\le I_c = \int d\mathbf{z}\,dx\,\pi_\theta(x)\mathbf{E}(\mathbf{z}|x)\log\frac{\mathbf{E}(\mathbf{z}|x)}{r(\mathbf{z})} = \mathbb{E}_{x,a\sim\pi}[\mathrm{KL}[\mathbf{E}(\mathbf{z}|x)||r(\mathbf{z})]] \end{aligned}$$

In Equation 3, the VRB minimizes the mutual information with dialog states to focus on discriminative features. The VRB also minimizes the KL-divergence with the human dialogs, while maximizing the KL-divergence with the generated dialogs, thereby distinguishing effectively between samples from human dialogs and dialog policy. Our proposed model is summarized in Algorithm 1.

3 Experiments

3.1 Dataset

We evaluate our method on Multi-domain wizard-of-oz (Budzianowski et al., 2018) (MultiWOZ), which contains approximately 10,000 of large-scale, multi-domain, and multi-turn conversational dialog corpora. MultiWOZ consists of seven distinct task-oriented domains, 24 slots, and 4,510 slot values.

3.2 Baselines And Evaluation Metrics

We compare the proposed method with the following existing methods: GP-MBCM (Gašić et al., 2015), ACER (Wang et al., 2017), PPO (Schulman et al., 2017), ALDM (Liu and Lane, 2018), and GDPL (Takanobu et al., 2019). Moreover, we evaluate our proposed model using four metrics: (i) *Turns*: we record the average number of dialog turns between the dialog agent and user simulator. (ii) *Match rate*: we conduct *match rate* experiments to analyze whether the booked entities are matched with the corresponding constraints in the multi-domain environment. For instance, in Figure 2, *entertainment* should be matched with *concert hall in the centre*. The match rate ranges from 0 to 1, and scores 0 if an agent fails to book the entity. (iii) *Inform F1*: we test the ability of the model to inform all of the requested slot values. For example, in Figure 1, the price range, food type, and area should be informed if the user wishes to visit a *high-end Cuban restaurant* in *Cambridge*. (iv) *Success rate*: in the *success rate* experiment, a dialog session scores 0 or 1. We obtain 1 if all required information is presented and every entity is booked successfully. Experimental setup is available in Appendix C.1.

Model	Agenda				VHUS			
	Turns	Match	Inform	Success	Turns	Match	Inform	Success
GP-MBCM (Gašić et al., 2015)	2.99	44.29	19.04	28.9	-	-	-	-
ACER (Wang et al., 2017)	10.49	62.83	77.98	50.8	22.35	33.08	55.13	18.6
PPO (Schulman et al., 2017)	9.83	69.09	83.34	59.1	**19.23**	33.08	56.31	18.3
ALDM (Liu and Lane, 2018)	12.47	62.60	81.20	61.2	26.90	24.15	54.37	16.4
GDPL (Takanobu et al., 2019)	7.64	83.90	**94.97**	86.5	22.43	36.21	52.58	19.7
VRB (Ours)	**7.59**	**90.87**	90.97	**90.4**	20.96	**44.93**	**56.93**	**20.1**
Human	7.37	95.29	66.89	75.0	-	-	-	-

Table 1: Results on Agenda-based and VHUS-based user simulators.

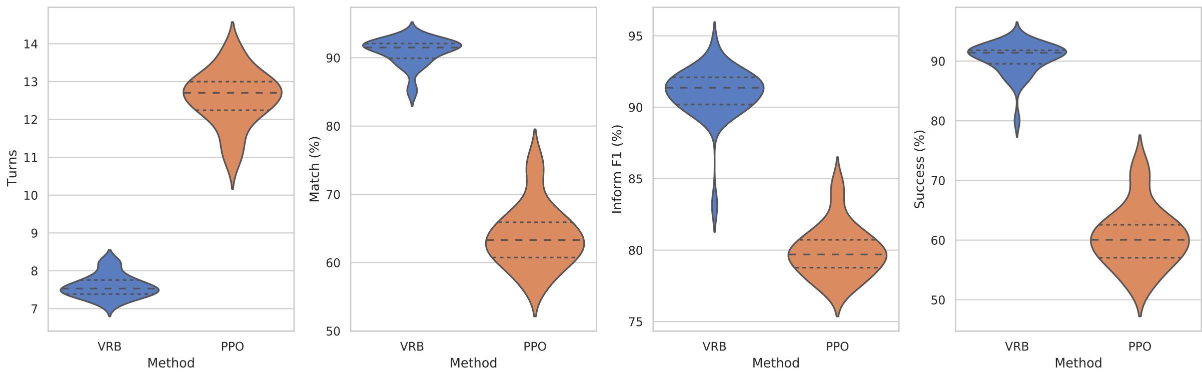

Figure 3: Performance on the MultiWOZ and the Agenda-based user simulator. Higher is better except *Turns*. Quartiles marked with dashed lines.

3.3 Results

Table 1 presents the empirical results on both simulators and MultiWOZ. In the agenda-based setting, we observe that our proposed method achieves a new state-of-the-art performance. Note that an outstanding model should obtain high scores in every metric, not just a single one, because to regard a dialog as having ended successfully, every request should be informed precisely, thereby guiding a dialog through the user goal. Although GDPL achieves the highest score in Inform F1, our proposed model acts more human-like with respect to *Turns*, which is closed to human evaluation score: 7.37, and provides more accurate slot values and matched-entities than the other methods. In VHUS setting, on the other hand, though PPO behaves more human-like in *Turns*, PPO exhibits greater difficulty in providing accurate information, while our model doesn't because our method constrains unproductive information flows.

In Figure 3, to evaluate the robustness of the models, we conduct experiments over 30 times for each model and visualize the results using a violin plot. Note that our proposed method outperforms PPO in every metric despite some negative outliers and has much lower standard deviation than PPO[3]. Both results in Table 1 and in Figure 3 demonstrate that our proposed model outperforms existing models, providing more definitive information than the other methods.

4 Conclusions

In this paper, we develop a novel and effective regularization method known as the Variational reward estimator bottleneck (VRB) for multi-domain task-oriented dialog systems. The VRB contains a stochastic encoder which enables the reward estimator to be maximally informative, as well as provides information bottleneck regularization, which constrains unproductive information flows between the inputs and reward estimator. The empirical results demonstrate that VRB achieves a new state-of-the-art performances on two different user simulators and a multi-turn and multi-domain task-oriented dialog dataset.

[3]An example of dialog session comparison between VRB and PPO is available in Table 3

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. cite arxiv:1612.00410Comment: 19 pages, 8 figures, Accepted to ICLR17.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October-November. Association for Computational Linguistics.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Chelsea Finn, Sergey Levine, and Pieter Abbeel. 2016. Guided cost learning: Deep inverse optimal control via policy optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 49–58. JMLR.org.

Justin Fu, Katie Luo, and Sergey Levine. 2018. Learning robust rewards with adverserial inverse reinforcement learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

M. Gašić, N. Mrkšić, P. Su, D. Vandyke, T. Wen, and S. Young. 2015. Policy committee for adaptation in multi-domain spoken dialogue systems. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 806–812.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Izzeddin Gür, Dilek Hakkani-Tür, Gokhan Tür, and Pararth Shah. 2018. User modeling for task oriented dialogues. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 900–906. IEEE.

Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4565–4573. Curran Associates, Inc.

Bing Liu and Ian Lane. 2018. Adversarial learning of task-oriented neural dialog models. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 350–359, Melbourne, Australia, July. Association for Computational Linguistics.

Andrew Ng and Stuart Russell. 2000. Algorithms for inverse reinforcement learning. *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*, 05.

Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. 2019. Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow. In *International Conference on Learning Representations*.

Stuart Russell. 1998. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 101–103.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152, Rochester, New York, April. Association for Computational Linguistics.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1889–1897, Lille, France, 07–09 Jul. PMLR.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2016. High-dimensional continuous control using generalized advantage estimation. *International Conference on Learning Representations*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana, June. Association for Computational Linguistics.

Weiyan Shi and Zhou Yu. 2018. Sentiment adaptive end-to-end dialog systems. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Ryuichi Takanobu, Hanlin Zhu, and Minlie Huang. 2019. Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 100–110, Hong Kong, China, November. Association for Computational Linguistics.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. 2017. Sample efficient actor-critic with experience replay. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–10, Los Angeles, September. Association for Computational Linguistics.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. 2008. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA.

Appendix

A Background

A.1 Dialog State Tracker

The dialog state tracker (DST) (Wu et al., 2019), which takes dialog action a and dialog history as input, updates the dialog state x and belief state b for each slot (For background and notations on MDP, see Appendix B.1). For example, in Figure 2, DST observes the user goal where the user wishes to go. At dialog turn t, the dialog action is represented as a slot and value pair (*e.g. Attraction: (area, centre), (type, concert hall)*). Given the dialog action, DST encodes the dialog state as $x_t = [a_t^u; a_{t-1}; b_t; q_t]$.

A.2 User Simulator

The user simulator $\mu(a^u, t^u|x^u)$ (Schatzmann et al., 2007; Gür et al., 2018) extracts the dialog action a^u corresponding to the dialog state x^u. t^u stands for whether user goal is achieved during conversation. Note that the DST and the user simulator can't achieve the user goal without well-defined reward estimation.

A.3 Policy Generator

The policy generator (Schulman et al., 2015; Schulman et al., 2017) encourages the dialog policy π_θ to determine the next action that maximizes the reward function $\hat{r}_{\zeta,\psi}(x_t, a_t, x_{t+1}) = f_{\zeta,\psi}(x_t, a_t, x_{t+1}) - \log \pi_\theta(a_t|x_t)$ (the full derivation is available in Appendix B.3):

$$\begin{aligned} L_\pi^{CLIP}(\theta) &= \mathbb{E}_{x,a\sim\pi}[\min(\xi_t(\theta)\hat{A}_t, \text{ clip}(\xi_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \\ L_t^{VF}(\theta) &= -\left(V_\theta - \sum_{k=t}^{T} \gamma^{k-t}\hat{r}_k\right)^2 \end{aligned} \tag{4}$$

where $\hat{A}_t = \delta_t + \gamma\lambda\hat{A}_{t+1}$, $\delta_t = \hat{r}_{\zeta,\psi} + \gamma V(x_{t+1}) - V(x_t)$, and δ is the TD residual (Schulman et al., 2016). $\xi_t(\theta) = \frac{\pi_\theta(a_t|x_t)}{\pi_{\theta_{\text{old}}}(a_t|x_t)}$ and V_θ is the state-value function. Epsilon and λ are hyper-parameters.

B Mathematical Details

B.1 Background and Notations on MDP

To represent Inverse reinforcement learning (IRL) as a Markov decision process (MDP), we consider a tuple $\mathcal{M} = (\mathcal{X}, \mathcal{A}, T, \mathcal{R}, \rho_0, \gamma)$, where $\mathcal{X}$ is state space and $\mathcal{A}$ is the action space. The transition probability $T(x_{t+1}|x_t, a_t)$ defines the distribution of the next state x_{t+1} given state x_t and a_t at time-step t. $\mathcal{R}(x_t, a_t)$ is the reward function of the state-action pair, ρ_0 is the distribution of the initial state x_0, and γ is the discount factor. The stochastic policy $\pi(a_t|x_t)$ maps a state to a distribution over actions. Supposing we are given an optimal policy π^*, the goal of IRL is to estimate the reward function $\mathcal{R}$ from the trajectory $\tau = \{x_0, a_0, x_1, a_1, ..., x_T, a_T\} \sim \pi^*$. However, constructing an effective reward function is challenging, especially in multi-domain task-oriented dialog system.

B.2 Gradient-Based Optimization

To imitate human behaviors, the reward estimator should learn the distributions of human dialog sessions using the KL-divergence loss:

$$\begin{aligned}
L_\pi(\theta) &\approx -\mathrm{KL}\left(\pi_\theta(\tau) \mid\mid \frac{\exp(\mathcal{R}_\zeta)}{Z}\right) \\
&= \sum \pi_\theta(\tau)\ \log\left(\frac{\frac{\exp(\mathcal{R}_\zeta)}{Z}}{\frac{\pi_\theta(\tau)}{1}}\right) \\
&= \mathbb{E}_{\tau\sim\pi}[\log\left(\frac{\exp(\mathcal{R}_\zeta)}{Z}\right) - \log\pi_\theta(\tau)] \\
&= \mathbb{E}_{\tau\sim\pi}[f_\zeta(\tau) - \log\pi_\theta(\tau)] \\
&= \mathbb{E}_{x,a\sim\pi}[f_{\zeta,\psi}(x_t, a_t, x_{t+1})] + \mathbb{E}_{s,a\sim\pi}[-\log\pi_\theta(x_t, a_t, x_{t+1})] \\
&= \mathbb{E}_{x,a\sim\pi}[f_{\zeta,\psi}(x_t, a_t, x_{t+1})] + H(\pi_\theta)
\end{aligned}$$

where $H(\pi_\theta)$ is the entropy of dialog policy π_θ. The reward estimator maximizes the entropy, which represents maximizing the likelihood of observed dialog sessions. Therefore, the reward estimator is trained to discern between human dialog sessions $\mathcal{D}$ and dialog sessions that are generated by the dialog policy:

$$\begin{aligned}
L_f(\zeta, \psi) &= -\mathrm{KL}\left(\mathcal{D}(\tau) \mid\mid \frac{\exp(\mathcal{R}_\zeta)}{Z}\right) - \left(-\mathrm{KL}\left(\pi_\theta(\tau) \mid\mid \frac{\exp(\mathcal{R}_\zeta)}{Z}\right)\right) \\
&= \mathbb{E}_{x,a\sim\mathcal{D}}[f_{\zeta,\psi}(x_t, a_t, x_{t+1})] + H(\mathcal{D}) - \mathbb{E}_{s,a\sim\pi}[f_{\zeta,\psi}(x_t, a_t, x_{t+1})]\ - H(\pi_\theta)
\end{aligned}$$

Note that $H(\mathcal{D})$ and $H(\pi_\theta)$ are not dependent on the parameters ζ and ψ. Thus, the reward estimator can be trained using gradient-based optimization as follows:

$$L_f(\zeta, \psi) = \mathbb{E}_{x,a\sim\mathcal{D}}[f_{\zeta,\psi}(x_t, a_t, x_{t+1})] - \mathbb{E}_{x,a\sim\pi}[f_{\zeta,\psi}(x_t, a_t, x_{t+1})]$$

B.3 Discriminative Reward Function

The reward function $\hat{r}_{\zeta,\psi}$ can be simplified in the following manner:

$$\begin{aligned}
\hat{r}_{\zeta,\psi}(x_t, a_t, x_{t+1}) &= \log\left[D_{\zeta,\psi}(x_t, a_t, x_{t+1})\right] - \log\left[1 - D_{\zeta,\psi}(x_t, a_t, x_{t+1})\right] \\
&= \log\left[-1 + \frac{1}{1 - D_{\zeta,\psi}(x_t, a_t, x_{t+1})}\right] \\
&= \log\left[\frac{\exp\left[f_{\zeta,\psi}(x_t, a_t, x_{t+1})\right]}{\pi_\theta(a_t|x_t)}\right] \\
&= f_{\zeta,\psi}(x_t, a_t, x_{t+1}) - \log\pi_\theta(a_t|x_t)
\end{aligned}$$

where $D_{\zeta,\psi}(x_t, a_t, x_{t+1})$ is the reward estimator which is defined as follows (Fu et al., 2018):

$$D_{\zeta,\psi}(x_t, a_t, x_{t+1}) = \frac{\exp[f_{\zeta,\psi}(x_t, a_t, x_{t+1})]}{\exp[f_{\zeta,\psi}(x_t, a_t, x_{t+1})] + \pi_\theta(a_t|x_t)}$$

C Experimental Setup

C.1 Training Details

The dialog sessions are randomly divided into training, validation, and test set. The validation and test sets contain 1,000 sessions each. We use the agenda-based user simulator (Schatzmann et al., 2007) and VHUS-based user simulator (Gür et al., 2018). The policy network π_θ and value network V are MLPs with two hidden layers. g_ζ and h_ψ are MPLs with one hidden layer each. We use the ReLu activation function and Adam optimizer for the MLPs. We train our model using a single NVIDIA GTX 1080ti GPU. The hyper-parameters are presented in Appendix C.2.

C.2 Hyperparameters

Hyperparameter	**Value**
Lagrange multiplier φ	0.001
Upper bound I_c	0.5
Learning rate of dialog policy	0.0001
Learning rate of reward estimator	0.0001
Learning rate of user simulator	0.001
Clipping component ϵ for dialog policy	0.02
GAE component λ for dialog policy	0.95

Table 2: VRB hyperparameters.

C.3 Performance on MultiWOZ and VHUS-based User Simulator

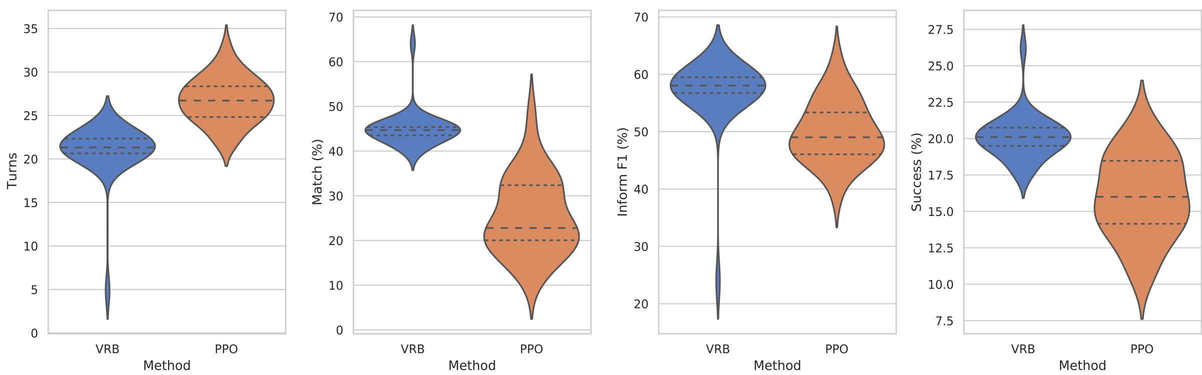

Figure 4: Performance on the MultiWOZ and the VHUS-based user simulator. Higher is better except *Turns*. Quartiles marked with dashed lines.

VRB	PPO
U: {'attraction-inform-area-1': 'south'}	U: {'attraction-inform-area-1': 'south'}
S: {'attraction-inform-choice-1': '8', 'attraction-request-type-?': '?'}	S: {'attraction-inform-choice-1': '8', 'attraction-request-type-?': '?'}
U: {'attraction-request-post-?': '?', 'attraction-request-phone-?': '?', 'attraction-request-addr-?': '?', 'attraction-request-fee-?': '?', 'attraction-request-type-?': '?'}	U: {'attraction-request-post-?': '?', 'attraction-request-phone-?': '?', 'attraction-request-addr-?': '?', 'attraction-request-fee-?': '?', 'attraction-request-type-?': '?'}
S: {'attraction-inform-name-1': 'the junction', 'attraction-recommend-name-1': 'the junction', 'attraction-recommend-addr-1': 'clifton way'}	S: {'attraction-inform-name-1': 'the junction', 'attraction-inform-fee-1': '?', 'attraction-recommend-name-1': 'the junction'}
U: {'attraction-request-post-?': '?', 'attraction-request-phone-?': '?', 'attraction-request-fee-?': '?', 'attraction-request-type-?': '?'}	U: {'attraction-request-post-?': '?', 'attraction-request-phone-?': '?', 'attraction-request-addr-?': '?', 'attraction-request-type-?': '?'}
S: {'attraction-inform-fee-1': '?', 'attraction-inform-phone-1': '01223511511', 'attraction-inform-post-1': 'cb17gx'}	S: {'attraction-inform-name-1': 'the junction', 'attraction-inform-addr-1': 'clifton way', 'attraction-inform-phone-1': '01223511511', 'attraction-inform-post-1': 'cb17gx'}
U: {'attraction-request-type-?': '?'}	U: {'attraction-request-type-?': '?'}
S: {'attraction-inform-type-1': 'museum'}	S: {}
U: {'hotel-inform-price-1': 'cheap', 'hotel-inform-area-1': 'centre'}	U: {'attraction-request-type-?': '?'}
S: {'hotel-inform-name-1': 'alexander bed and breakfast'}	S: {}
U: {'hotel-request-post-?': '?', 'hotel-request-phone-?': '?'}	U: {'attraction-request-type-?': '?'}
S: {'general-reqmore-none-none': 'none', 'hotel-inform-phone-1': '01223525725', 'hotel-inform-post-1': 'cb12de'}	S: {}
U: {'hotel-inform-stay-1': 'dont care', 'hotel-inform-day-1': 'dont care', 'hotel-inform-people-1': 'dont care'}	U: {'attraction-request-type-?': '?'}
S: {'booking-book-ref-1': 'none'}	S: {}
U: {'general-bye-none-none': 'none'}	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'attraction-request-type-?': '?'}
	S: {}
	U: {'general-bye-none-none': 'none'}
turn: 8 match: 1.0 inform (1.0, 1.0, 1.0)	turn: 22 match: 0.0 inform (0, 0, 0)
Success	**Failure**

Table 3: A comparison between VRB and PPO with respect to the dialog act.