Thesis for the Degree
of Master

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

by

Jeiyoon Park

Department of Computer Science
Engineering

Graduate School

Korea University

August 2021

林 希 錫 教授指導

碩 士 學 位 論 文

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

이 論文을 컴퓨터學 工學 碩士學位 論文으로
提出함

2021年 06月

高 麗 大 學 校 大 學 院

컴 퓨 터 學 科

朴 帝 玧

朴帝玧의 컴퓨터學 工學 碩士學位論文 審査를 完了함

2021年 06月

委員長 임 희 석 (印)

委 員 김 현 우 (印)

委 員 김 승 룡 (印)

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.

Contents

List of Figures

List of Tables

Chapter 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 [1, 2, 3, 4]. However, such approaches are often unable to guide the dialog policy through user goals. For instance, as illustrated in Figure 1.1, 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) [5, 6] and MaxEnt-IRL [7] tackles the problem of recovering reward function and using this reward function to generate optimal behavior. Although Generative adversarial imitation learning (GAIL) [8], which exploits the GANs framework [9], 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) [10] enables GAIL to take advantage of disentangled rewards. Guided dialog policy learning (GDPL) [11] 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

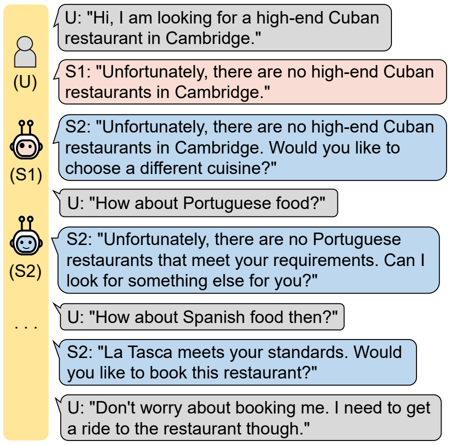

Figure 1.1: The system (S2) that uses well-specified rewards can guide the user through the goal while S1 can't.)

estimator, and produce excessively uninformative gradients.

In this paper, we propose the Variational Reward Estimator Bottleneck (VRB), an effective regularization algorithm. The VRB uses information bottleneck [12, 13, 14] 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.

Chapter 2

Background

2.1 Dialog State Tracker

The dialog state tracker (DST) [15], which takes dialog action a and dialog history as input, updates the dialog state x and belief state b for each slot. For example, in Figure 2.1, 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 $x_t = [a_t^u; a_{t-1}; b_t; q_t]$.

2.2 User Simulator

Mimicking diverse and human-like behaviors is essential, with respect to training task-oriented dialog systems and evaluating these models automatically. The user simulator $\mu(a^u, t^u|x^u)$ [16, 17] in Figure 2.1 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.

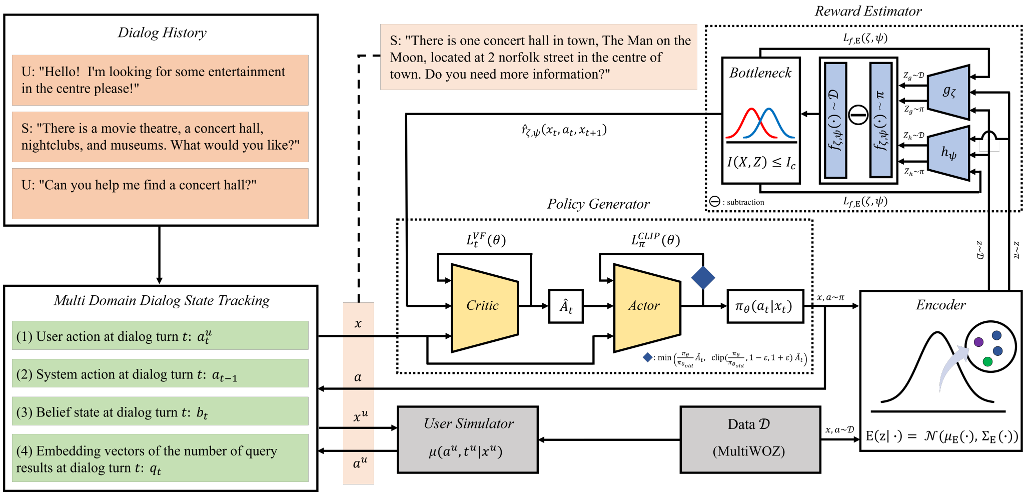

Figure 2.1: Schematic depiction of the Variational Reward Estimator.

2.3 Policy Generator

The policy generator [18, 19] encourages the dialog policy π_θ to determine the next action that maximizes the reward function $\hat{r}_{\zeta,\psi}(x_t, a_t, x_{t+1}) = f_{\zeta,\psi}(x_t, a_t, x_{t+1}) - \log \pi_\theta(a_t|x_t)$:

$$L_\pi^{CLIP}(\theta) = \mathbb{E}_{x,a\sim\pi}[\min(\xi_t(\theta)\hat{A}_t, \tilde{\xi}_t(\theta)\hat{A}_t)]$$

$$L_t^{VF}(\theta) = -\left(V_\theta - \sum_{k=t}^{T} \gamma^{k-t}\hat{r}_k\right)^2$$

where $\tilde{\xi}_t(\theta) = \text{clip}(\xi_t(\theta), 1-\epsilon, 1+\epsilon)$, $\hat{A}_t = \delta_t + \gamma\lambda\hat{A}_{t+1}$, $\delta_t = \hat{r}_{\zeta,\psi} + \gamma V(x_{t+1}) - V(x_t)$, and δ is the TD residual [20]. $\xi_t(\theta) = \frac{\pi_\theta(a_t|x_t)}{\pi_{\theta_{\text{old}}}(a_t|x_t)}$ and V_θ is the state-value function. Epsilon and λ are hyper-parameters. The reward function $\hat{r}_{\zeta,\psi}$ can be simplified in the following manner:

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

where $D_{\zeta,\psi}(x_t, a_t, x_{t+1})$ is the reward estimator which is defined as follows [10]:

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

Chapter 3

Proposed Method

3.1 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.

3.2 Reward Estimator

The reward estimator [11], 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. Based on MaxEnt-IRL [7], 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. To imitate human behaviors, the reward estimator should learn the distributions of human dialog sessions using the KL-divergence loss:

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

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:

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

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:

$$
\begin{aligned}
L_f(\zeta, \psi) = \mathbb{E}_{x,a\sim\mathcal{D}}&[f_{\zeta,\psi}(x_t, a_t, x_{t+1})] \\
-\mathbb{E}_{x,a\sim\pi}&[f_{\zeta,\psi}(x_t, a_t, x_{t+1})]
\end{aligned}
\tag{3.1}
$$

3.3 Variational Reward Estimator Bottleneck

The Variational information bottleneck [12, 13, 14] 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 $\mathbf{Z}$:

$$
\begin{aligned}
L_{f,\mathbf{E}}(\zeta, \psi) &= \mathbb{E}_{x,a\sim\mathcal{D}}[\,\mathbb{E}_{\mathbf{z}\sim\mathbf{E}(\mathbf{z}|x_t,x_{t+1})}[\,f_{\zeta,\psi}(\mathbf{z}_g, \mathbf{z'}_h, \mathbf{z}_h)]\,] \\
&\quad - \mathbb{E}_{x,a\sim\pi}[\,\mathbb{E}_{\mathbf{z}\sim\mathbf{E}(\mathbf{z}|x_t,x_{t+1})}[\,f_{\zeta,\psi}(\mathbf{z}_g, \mathbf{z'}_h, \mathbf{z}_h)]\,] \\
&\qquad\qquad\qquad \text{s.t.} \quad I(\mathbf{Z}, X) \leq I_c
\end{aligned}
\tag{3.2}
$$

Algorithm 1: Variational Reward Estimator Bottleneck

```
Initialize dialog policy generator π_θ and reward estimator f_{ζ,ψ}
for i ← 0 to N do
    Obtain random samples from human dialog corpus 𝒟
    Gather dialog sessions using user simulator μ(a^u, t^u|x^u) and policy generator
      π_θ(a|x)
    Encode dialog sessions using stochastic encoder E(z|·) = 𝒩(μ_E(·), Σ_E(·))
    Compute information bottleneck 𝔼_{x,a∼π}[KL[E(z|x)||r(z)]]
    Update reward estimator f_{ζ,ψ} by optimizing L_{f,E}(ζ, ψ)
    Estimate reward function r̂_{ζ,ψ} for each state-action pair
    Update state-value function V(𝒳) and dialog policy π_θ given the reward r̂_{ζ,ψ}
```

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 [9], GAN-GCL [21], and AIRL [10]. 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 $\mathbf{E}(\mathbf{z}|x_t, x_{t+1}) = \mathbf{E}_g(\mathbf{z}_g|x_t) \cdot \mathbf{E}_h(\mathbf{z}_h|x_t) \cdot \mathbf{E}_h(\mathbf{z}'_h|x_{t+1})$ 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:

$$
\begin{aligned}
L_{f,\mathbf{E}}(\zeta, \psi) = \mathbb{E}_{x,a\sim\mathcal{D}}[\,\mathbb{E}_{\mathbf{z}\sim\mathbf{E}(\mathbf{z}|x_t,x_{t+1})}[\,f_{\zeta,\psi}(\mathbf{z}_g, \mathbf{z}'_h, \mathbf{z}_h)]\,] \\
- \mathbb{E}_{x,a\sim\pi}[\,\mathbb{E}_{\mathbf{z}\sim\mathbf{E}(\mathbf{z}|x_t,x_{t+1})}[\,f_{\zeta,\psi}(\mathbf{z}_g, \mathbf{z}'_h, \mathbf{z}_h)]\,] \\
+ \varphi\,(\mathbb{E}_{x,a\sim\pi}\,[\,\mathrm{KL}[\mathbf{E}(\mathbf{z}|x_t, x_{t+1})]\,||\,r(\mathbf{z})\,] - I_c)
\end{aligned}
\tag{3.3}
$$

where the mutual information between dialog states X and latent variable $\mathbf{Z}$ is

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

In Equation 3.3, 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.

Chapter 4

Experimental Setup

4.1 Dataset

We evaluate our method on Multi-domain wizard-of-oz [22] (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.

4.2 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 [16] and VHUS-based user simulator [17]. 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 Table 4.1.

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 4.1: VRB hyperparameters.

4.3 Baselines and Evaluation Metrics

We compare the proposed method with the following existing methods: GP-MBCM [23], ACER [24], PPO [19], ALDM [25], and GDPL [11]. GP-MBCM [23] trains a number of policies on different datasets based on Bayesian committee machine [26]. ACER [24] introduces Importance weight truncation with bias correction for sampling efficiency. PPO [19] employs an effective algorithm that attains the data efficiency and robust performance using only first-order optimizer. ALDM [25] exploits an adversarial learning method to learn dialog rewards directly from dialog samples. GDPL [11] is current state-of-the-art model which consists of a dialog reward estimator based on IRL.

To evaluate the performances of theses models, we introduce 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 2.1, *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 1.1, 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.

Chapter 5

Results

5.1 Agenda-Based Setting

Table 5.1 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.

5.2 VHUS-Based Setting

Model	Agenda			
	Turns	Match	Inform	Success
GP-MBCM [23]	2.99	44.29	19.04	28.9
ACER [24]	10.49	62.83	77.98	50.8
PPO [19]	9.83	69.09	83.34	59.1
ALDM [25]	12.47	62.60	81.20	61.2
GDPL [11]	7.64	83.90	**94.97**	86.5
VRB (Ours)	**7.59**	**90.87**	90.97	**90.4**
Human	7.37	95.29	66.89	75.0

Table 5.1: Results on Agenda-based user simulators.

Model	VHUS			
	Turns	Match	Inform	Success
GP-MBCM [23]	-	-	-	-
ACER [24]	22.35	33.08	55.13	18.6
PPO [19]	**19.23**	33.08	56.31	18.3
ALDM [25]	26.90	24.15	54.37	16.4
GDPL [11]	22.43	36.21	52.58	19.7
VRB (Ours)	20.96	**44.93**	**56.93**	**20.1**

Table 5.2: Results on VHUS-based user simulators.

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. Results in Table 5.2 demonstrate that our proposed model outperforms existing models, providing more definitive information than the other methods.

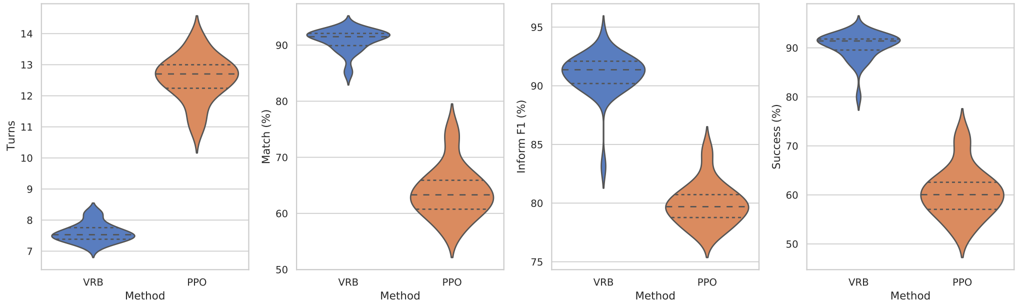

Figure 5.1: Performance on the MultiWOZ and the Agenda-based user simulator. Higher is better except *Turns*. Quartiles marked with dashed lines.

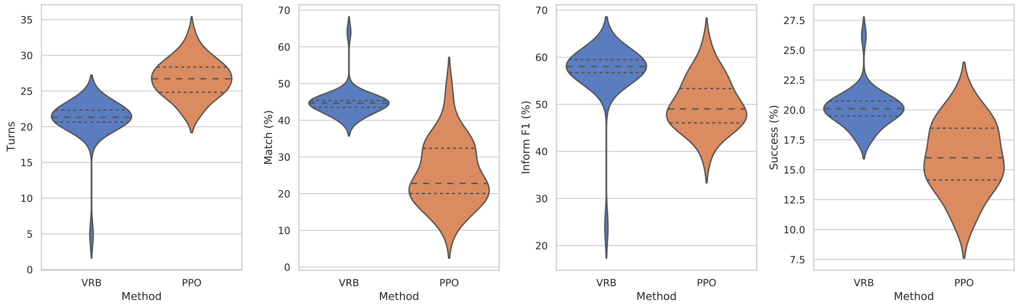

Figure 5.2: Performance on the MultiWOZ and the VHUS-based user simulator. Higher is better except *Turns*. Quartiles marked with dashed lines.

5.3 Repetitive Experiments

In Figure 5.1 and Figure 5.2, to evaluate the robustness of the models, we conduct experiments over 30 times for each model and visualize the results using a violin plot. In the experiments, our proposed method outperforms PPO in every metric despite some negative outliers and has much lower standard deviation than PPO[1].

[1] An example of dialog session comparison between VRB and PPO is available in Table 5.3

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	turn: 22
match: 1.0	match: 0.0
inform: (1.0, 1.0, 1.0)	inform: (0, 0, 0)
Success	**Failure**

Table 5.3: A comparison between VRB and PPO with respect to the dialog act.

Chapter 6

Conclusion

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.

Bibliography

[1] T. Zhao and M. Eskenazi, "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*, (Los Angeles), pp. 1–10, Association for Computational Linguistics, Sept. 2016.

[2] B. Dhingra, L. Li, X. Li, J. Gao, Y.-N. Chen, F. Ahmed, and L. Deng, "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)*, 2017.

[3] W. Shi and Z. Yu, "Sentiment adaptive end-to-end dialog systems," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[4] P. Shah, D. Hakkani-Tür, B. Liu, and G. Tür, "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)*, (New Orleans - Louisiana), pp. 41–51, Association for Computational Linguistics, June 2018.

[5] S. Russell, "Learning agents for uncertain environments," in *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 101–103, 1998.

[6] A. Ng and S. Russell, “Algorithms for inverse reinforcement learning,” *ICML ’00 Proceedings of the Seventeenth International Conference on Machine Learning*, 05 2000.

[7] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, “Maximum entropy inverse reinforcement learning,” in *AAAI*, vol. 8, pp. 1433–1438, Chicago, IL, USA, 2008.

[8] J. Ho and S. Ermon, “Generative adversarial imitation learning,” in *Advances in Neural Information Processing Systems 29* (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds.), pp. 4565–4573, Curran Associates, Inc., 2016.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 2672–2680, Curran Associates, Inc., 2014.

[10] J. Fu, K. Luo, and S. Levine, “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, 2018.

[11] R. Takanobu, H. Zhu, and M. Huang, “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)*, (Hong Kong, China), pp. 100–110, Association for Computational Linguistics, Nov. 2019.

[12] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” in *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

[13] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, “Deep variational information bottleneck,” 2016. cite arxiv:1612.00410Comment: 19 pages, 8 figures, Accepted to ICLR17.

[14] X. B. Peng, A. Kanazawa, S. Toyer, P. Abbeel, and S. Levine, "Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow," in *International Conference on Learning Representations*, 2019.

[15] C.-S. Wu, A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher, and P. Fung, "Transferable multi-domain state generator for task-oriented dialogue systems," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[16] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young, "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*, (Rochester, New York), pp. 149–152, Association for Computational Linguistics, Apr. 2007.

[17] I. Gür, D. Hakkani-Tür, G. Tür, and P. Shah, "User modeling for task oriented dialogues," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 900–906, IEEE, 2018.

[18] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 1889–1897, PMLR, 07–09 Jul 2015.

[19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[20] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *International Conference on Learning Representations*, 2016.

[21] C. Finn, S. Levine, and P. Abbeel, "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, p. 49–58, JMLR.org, 2016.

[22] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "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*, (Brussels, Belgium), pp. 5016–5026, Association for Computational Linguistics, Oct.-Nov. 2018.

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

[24] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas, "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, 2017.

[25] B. Liu and I. Lane, "Adversarial learning of task-oriented neural dialog models," in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, (Melbourne, Australia), pp. 350–359, Association for Computational Linguistics, July 2018.

[26] V. Tresp, "A bayesian committee machine," *Neural Comput.*, vol. 12, p. 2719–2741, Nov. 2000.

Acknowledgement

국내 최고수준의 연구기관인 고려대학교 대학원에 컴퓨터학과 석사과정으로 입학하여 정말 열심히 공부하였고 많은 것들을 배울 수 있었습니다. 하지만 지난 연구실 생활을 돌이켜보면 저 혼자 잘나서 이룬것들은 단 하나도 없었습니다. 오히려 많이 부족한 저에게 정말 과분할 정도의 도움과 격려를 주신분들이 계셨고, 그 분들이 계셨기에 제가 꿈을 갖고 훌륭한 학자로 성장할 수 있는 밑바탕을 그릴 수 있었습니다. 그 분들께 진심을 담아 감사의 말씀을 드리고자 합니다.

먼저, 항상 따뜻한 격려와 연구분야에 대한 아낌없는 조언을 주시고 부족한 저에게 많은 가르침을 주신 임희석 교수님께 진심을 담아 존경과 감사의 마음을 올립니다. 임희석 교수님께 연구지도를 받았다는 것은 저에게 정말 큰 영광이자 자랑입니다. 학생들에게 가르침을 주시고 연구하시느라 바쁘신 와중에도 석사 학위 심사를 맡아주시고 훌륭한 학자가 될 수 있도록 조언과 격려를 아낌없이 주신 김현우 교수님 그리고 김승룡 교수님께도 감사의 말씀을 드립니다. 석사과정동안 연구실에서 밤낮으로 동고동락(同苦同樂)했던 연구실 선배님들과 후배님들께도 진심으로 감사의 말씀을 전하고 싶습니다. 특히 제 연구에 대해 학자로서 정말 많은 가르침을 주셨던 찬희형, 제 옆자리에서 늘 조언과 도움을 아끼지 않으셨던 태선이형, 일이 잘 안되고 심적으로 힘들때마다 때로는 연구자로서, 때로는 친한 형동생으로서 격려해주고 다시 힘낼 수 있게 도와준 성진이형 그리고 기수에게도 감사의 말을 전합니다.

항상 뒤에서 제가 가는 길을 지지해주신 어머니, 아버지, 누나 사랑합니다. 부모님의 아들, 누나의 동생이라는게 늘 자랑스럽다고 생각하고 있습니다.

부족한 몸으로 정말 많은 분들께 과분할 정도의 은혜를 입었습니다. 덕분에 학자로서 올바른 방향으로 나아가고 있다고 생각합니다. 도움주신 모든 분들께 다시한번 감사의 말씀 올립니다. 저도 이 은혜를 잊지 않고 받은 만큼 베풀 수 있는, 더 나아가서 기술로 사람들에게 즐거움과 행복을 선사할 수 있는 사람이 되겠습니다.