applied sciences

MDPI

Article

Variational Reward Estimator Bottleneck: Towards Robust Reward Estimator for Multidomain Task-Oriented Dialogue

Jeiyoon Park †, Chanhee Lee, Chanjun Park, Kuekyeng Kim ‡ and Heuiseok Lim *

Department of Computer Science and Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea; k4ke@korea.ac.kr (J.P.); chanhee0222@gmail.com (C.L.); bcj1210@naver.com (C.P.); overmind22@korea.ac.kr (K.K.)

* Correspondence: limhseok@korea.ac.kr

† Work partially done while the first author was an intern at NAVER WEBTOON Corp (Webtoon AI).

‡ Work performed while at Massachusetts Institute of Technology (MIT).

Abstract: Despite its significant effectiveness in adversarial training approaches to multidomain task-oriented dialogue systems, adversarial inverse reinforcement learning of the dialogue policy frequently fails to balance the performance of the reward estimator and policy generator. During the optimization process, the reward estimator frequently overwhelms the policy generator, resulting in excessively uninformative gradients. We propose the variational reward estimator bottleneck (VRB), which is a novel and effective regularization strategy 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. Quantitative analysis on a multidomain task-oriented dialogue dataset demonstrates that the VRB significantly outperforms previous studies.

Keywords: task-oriented dialogue; dialogue policy; reinforcement learning; inverse reinforcement learning

Citation: Park, J.; Lee, C.; Park, C; Kim, K.; Lim, H. Variational Reward Estimator Bottleneck: Towards Robust Reward Estimator for Multidomain Task-Oriented Dialogue. *Appl. Sci.* **2021**, *11*, 6624. https://doi.org/10.3390/app11146624

Academic Editor: Mauro Castelli

Received: 29 May 2021
Accepted: 16 July 2021
Published: 19 July 2021

1. Introduction

While deep reinforcement learning (RL) has emerged as a viable solution for complicated and high-dimensional decision-making problems [1], including games such as Go [2], chess [3], checkers [4], and poker [5,6], robotic locomotion [7,8], autonomous driving [9,10], and recommender system [11,12], the determination of an effective reward function remains a challenge, especially in multidomain task-oriented dialogue systems. Many recent studies have struggled in sparse-reward environments and employed a handcrafted reward function as a breakthrough [13–16]. However, such approaches typically are not capable of guiding the dialogue policy through user goals. For instance, as shown in Figure 1, the user cannot attain the goal because the system (S1) that exploits the handcrafted rewards completes the dialogue session too early. Moreover, as the dialog progresses, the user goal will frequently vary.

Due to these problems, systems that exploit the handcrafted rewards fail to assimilate user goals and guide users through user goals, achieving low performance, while humans self-judge from dialog context using well-defined reward function in their minds and generate appropriate responses despite multidomain circumstances.

MaxEnt-IRL [17] and Inverse reinforcement learning (IRL) [18,19] tackle the problem of recovering the reward function automatically and using this reward function to generate optimal behavior. Although generative adversarial imitation learning (GAIL) [20], which applies the GANs framework [21], 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) [22] enables GAIL to take advantage of disentangled rewards. Guided dialogue policy learning (GDPL) [23] uses the AIRL framework

to construct the reward estimator for multidomain task-oriented dialogues. However, these approaches often encounter difficulties in balancing the performance of the reward estimator and policy generator and produce excessively uninformative gradients.

In this paper, we propose the variational reward estimator bottleneck (VRB), a novel and effective regularization algorithm. The VRB uses information bottleneck [24–26] to constrain unproductive information flows between dialogue internal representations and state–action pairs of the reward estimator, thereby ensuring highly informative gradients and robustness. The experiments show that the VRB achieves state-of-the-art (SOTA) performances on a multidomain task-oriented dataset.

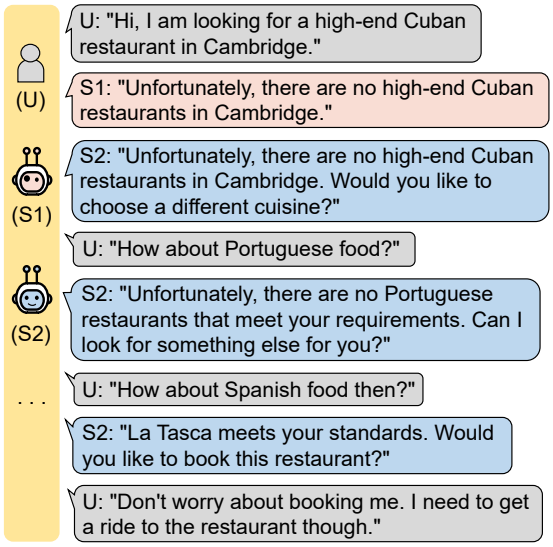

Figure 1. The system (S2) that uses well-specified rewards can guide the user through the goal, while S1 cannot.

The remainder of this paper is organized as follows: Section 2 presents the brief background to set the stage for our model. Section 3 describes the proposed method in detail along with mathematical calculations. Section 4 outlines the experimental setup, whereas Section 5 presents the experiments and the results thereof. Section 6 provide discussions and the conclusions of this study.

2. Background

2.1. Dialogue State Tracker

The dialogue state tracker (DST) [27–29], which takes dialogue action a and dialogue history as input, updates the dialogue state x and belief state b for each slot. For example, as shown in Figure 2, DST observes the user goal where the user aims to go. At dialogue turn t, the dialogue action is represented as a slot and value pair (e.g., *Attraction: (area, centre), (type, concert hall)*). Given the dialogue action, DST encodes the dialogue state as $x_t = [a_t^u; a_{t-1}; b_t; q_t]$.

2.2. User Simulator

Mimicking various and human-like actions is essential with respect to training task-oriented dialogue systems and evaluating these models automatically. The user simulator $\mu(a^u, t^u | x^u)$ [30,31] in Figure 2 extracts the dialogue action a^u corresponding to the dialogue state x^u. t^u stands for whether the user goal is achieved during a conversation. Note that the DST and the user simulator cannot meet the user in the absence of well-defined reward estimation.

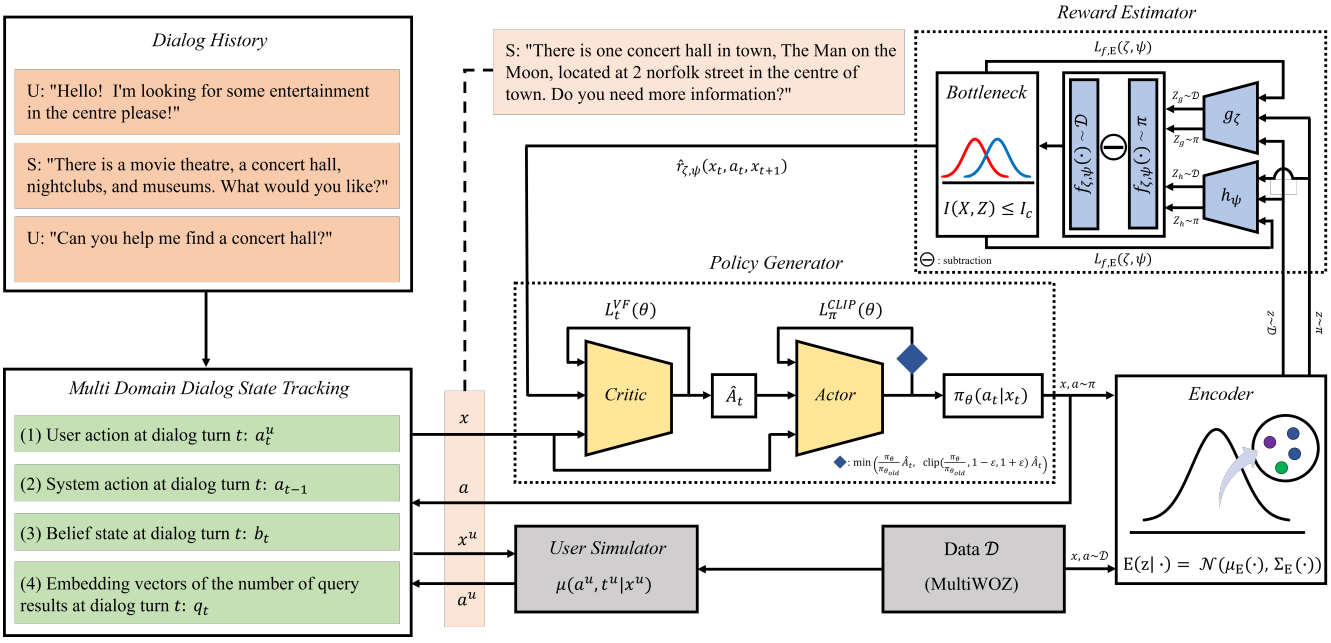

Figure 2. Schematic depiction of the variational reward estimator.

2.3. Policy Generator

The policy generator [32,33] encourages the dialogue 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 $\tilde{\xi}_t(\theta) = \mathrm{clip}(\xi_t(\theta), 1-\epsilon, 1+\epsilon)$, $\hat{A}_t = \delta_t + \gamma\lambda\hat{A}_{t+1}$, $\delta_t = \hat{r}_{\zeta,\psi} + \gamma V(x_{t+1}) - V(x_t)$, and δ is the TD residual [34]. $\xi_t(\theta) = \frac{\pi_\theta(a_t|x_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|x_t)}$ and V_θ is the state-value function. Epsilon and λ are hyperparameters. 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 [22]:

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

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 $\tau = \{x_0, a_0, x_1, a_1, \ldots, x_T, a_T\} \sim \pi^*$. However, building an effective reward function is challenging, especially in a multidomain task-oriented dialogue system.

3.2. Reward Estimator

The reward estimator [23], which is an essential component of multidomain task-oriented dialogue systems, evaluates dialogue state–action pairs at dialogue turn t and estimates the reward that is used for guiding the dialogue policy through the user goal. Based on MaxEnt-IRL [17], each dialogue session τ in a set of human dialogue sessions $\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_H\}$ can be modeled as a Boltzmann distribution that does not exhibit additional preferences for any dialogue sessions.

$$f_\zeta(\tau) = \log\left(\frac{\exp(\mathcal{R}_\asymp)}{Z}\right)$$

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 the reward function, and $\mathcal{R}_\zeta$ denotes a discounted cumulative reward. To imitate human behaviors, the reward estimator should learn the distributions of human dialogue sessions using the KL divergence loss:

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

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

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

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:

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

3.3. Variational Reward Estimator Bottleneck

The variational information bottleneck [24–26] is an theoretical information approach that restricts unproductive information flow between the discriminator and inputs. Inspired by this approach, we propose a regularized objective that constrains the mutual information between encoded original inputs and state–action pairs, thereby ensuring highly informative internal representations and a robust adversarial model. Our proposed method trains an encoder that is maximally informative regarding human dialogues.

To this end, we employ a stochastic encoder and an upper bound constraint on the mutual information between the dialogue states X and latent variables $\mathbf{Z}$:

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

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 [21], GAN-GCL [35], and AIRL [22]. 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 as follows:

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

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

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

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

Algorithm 1 Algorithm of Variational Reward Estimator Bottleneck

Initialize dialogue policy generator π_θ and reward estimator $f_{\zeta,\psi}$
for $i \leftarrow 0$ ***to*** N **do**
  Obtain Random Samples from Human Dialogue Corpus $\mathcal{D}$
  Gather Dialogue Sessions using User Simulator $\mu(a^u, t^u|x^u)$ and Policy Generator $\pi_\theta(a|x)$
  Encode Dialogue 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 $L_{f,\mathbf{E}}(\zeta,\psi)$
  Estimate Reward Function $\hat{r}_{\zeta,\psi}$ for each State–Action Pair
  Update State-Value Function $V(\mathcal{X})$ and Dialogue Policy π_θ given the Reward $\hat{r}_{\zeta,\psi}$

4. Experimental Setup

4.1. Dataset Details

We evaluated our method on multidomain wizard of oz [36] (MultiWOZ), which contained approximately 10,000 large-scale, multidomain, and multiturn conversational dialogue corpora. MultiWOZ consisted of 7 distinct task-oriented domains, 24 slots, and 4510 slot values. The dialogue sessions were randomly divided into training, validation, and test set. The validation and test sets contained 1000 sessions, respectively.

4.2. Models Details

We used the agenda-based user simulator [30] and VHUS-based user simulator [31]. 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 used the ReLu activation function and Adam optimizer for the MLPs. We trained our model using a single NVIDIA GTX 1080ti GPU. Detailed hyperparameters are shown in Table 1.

Table 1. Detail description of VRB hyperparameters.

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

We compare the proposed method with the following previous studies: GP-MBCM [37], ACER [38], PPO [33], ALDM [39], and GDPL [23]. GP-MBCM [37] trains a number of policies on different datasets based on the Bayesian committee machine [40]. ACER [38] suggests the importance of weight truncation with bias correction for sampling efficiency. PPO [33] employs an effective algorithm that attains the data's robust and efficient performance using only a first-order optimizer. ALDM [39] shows an adversarial learning method to learn dialogue rewards directly from dialogue samples. GDPL [23] is the current SOTA model that consists of a dialogue reward estimator based on IRL.

4.3. Evaluation Details

To evaluate the performances of these models, we introduce four metrics: (i) *Turns*: we record the average number of dialogue turns between the user simulator and dialogue agent. (ii) *Match rate*: we conduct *match rate* experiments to analyze whether the booked entities are matched with the corresponding constraints in the multidomain environment.

For instance, in Figure 2, *entertainment* should be matched with *concert hall in the center*. The match rate ranges from 0 to 1 and scores 0 if an agent is unable to book the entity. (iii) *Inform F1*: we test the ability of the model to inform all of the requested slot values. For example, as shown 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 dialogue session scores 0 or 1. We obtain 1 if all required information is presented, and every entity is booked successfully.

5. Main Results

5.1. Experimental Results of Agenda-Based User Simulators

Table 2 presents the empirical results on both simulators and MultiWOZ. In the agenda-based setting, we observe that our proposed method achieves a new SOTA performance. Note that an outstanding model should obtain high scores in every metric, not just a single one, because to regard a dialogue as having ended successfully, every request should be informed precisely, thereby guiding a dialogue 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 close to the human evaluation score: 7.37, and provides more accurate slot values and matched entities than the other methods.

Table 2. Results on agenda-based user simulators.

Model	Agenda			
	Turns	**Match**	**Inform**	**Success**
GP-MBCM [37]	2.99	44.29	19.04	28.9
ACER [38]	10.49	62.83	77.98	50.8
PPO [33]	9.83	69.09	83.34	59.1
ALDM [39]	12.47	62.60	81.20	61.2
GDPL [23]	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

5.2. Experimental Results of VHUS-Based User Simulators

On the other hand, in the VHUS setting, though PPO behaves more human-like in *Turns*, PPO exhibits greater difficulty in providing accurate information, while our model does not because our approach constrains unproductive information flows. Results in Table 3 demonstrate that our proposed model outperforms existing models, providing more definitive information than the other methods. Similar to the agenda-based setting, the VHUS-based model also showed the best performance. It demonstrates that our methodology reflecting human-like characteristics is a very effective methodology.

Table 3. Results on VHUS-based user simulators.

Model	VHUS			
	Turns	**Match**	**Inform**	**Success**
GP-MBCM [37]	-	-	-	-
ACER [38]	22.35	33.08	55.13	18.6
PPO [33]	**19.23**	33.08	56.31	18.3
ALDM [39]	26.90	24.15	54.37	16.4
GDPL [23]	22.43	36.21	52.58	19.7
VRB (Ours)	20.96	**44.93**	**56.93**	**20.1**

5.3. Verification of Robustness

As shown in Figures 3 and 4, to evaluate the robustness of the models, we conduct experiments over 30 times for each model and visualize the results using a violin plot. Exper-

imental results show that our proposed method outperforms PPO in every metric, despite some negative outliers, and has a much lower standard deviation than PPO. An example of a dialogue session comparison between VRB and PPO is available in Table 4.

Table 4. A comparison between VRB and PPO with respect to the dialogue act.

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**

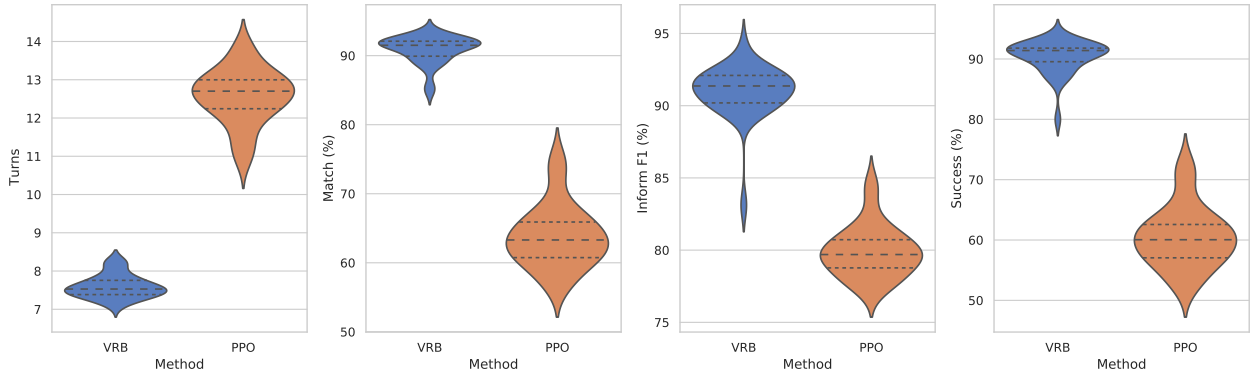

Figure 3. Performance on the MultiWOZ and the agenda-based user simulator. Higher is better except *Turns*. Quartiles marked with dashed lines.

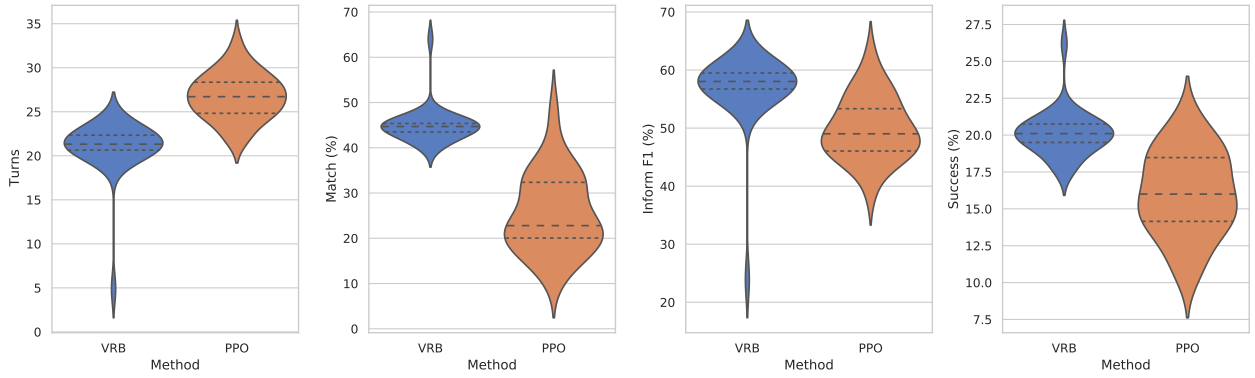

Figure 4. Performance on the MultiWOZ and the VHUS-based user simulator. Higher is better except *Turns*. Quartiles marked with dashed lines.

6. Conclusions

In this paper, we present a novel and effective regularization method known as the variational reward estimator bottleneck (VRB) for multidomain task-oriented dialogue systems. The VRB includes a stochastic encoder, which enables the reward estimator to be maximally informative, and provides information bottleneck regularization, which constrains unproductive information flows between the reward estimator and the inputs. The quantitative results show that VRB achieves new SOTA performances on two different user simulators and a multiturn and multidomain task-oriented dialogue dataset. Despite great improvements, training dialog policy via VHUS setting remains a hurdle to overcome. We leave this for future works.

Author Contributions: Conceptualization, J.P. and C.L.; data curation, J.P.; formal analysis, C.L.; funding acquisition/project administration/supervision, H.L.; investigation, J.P.; methodology/software, J.P.; visualization, J.P.; writing—review and editing, C.L., C.P. and K.K.; writing—original draft preparation, J.P., C.L. and C.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2018-0-01405) supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP), Institute for Information and communications Technology Planning and Evaluation (IITP), grant funded by the Korean government (MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques) and Ministry of Science and ICT (MSIT), Korea, under the ICT Creative Consilience program (IITP-2021-2020-0-01819) supervised by the Institute for Information and communications Technology Planning and Evaluation (IITP).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The data can be found here: MultiWOZ: https://arxiv.org/abs/1810.00278.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *CoRR* **2019**, doi:10.1038/s41586-020-03051-4.
2. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–503.
3. Campbell, M.; Hoane, A.; hsiung Hsu, F. Deep Blue. *Artif. Intell.* **2002**, *134*, 57–83, doi:10.1016/S0004-3702(01)00129-1.
4. Schaeffer, J.; Culberson, J.; Treloar, N.; Knight, B.; Lu, P.; Szafron, D. A world championship caliber checkers program. *Artif. Intell.* **1992**, *53*, 273–289, doi:10.1016/0004-3702(92)90074-8.
5. Brown, N.; Sandholm, T. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science* **2018**, *359*, 418–424, doi:10.1126/science.aao1733.
6. Moravcík, M.; Schmid, M.; Burch, N.; Lisý, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; Bowling, M.H. DeepStack: Expert-Level Artificial Intelligence in No-Limit Poker. *CoRR* **2017**, doi:10.1126/science.aam6960.
7. Peters, J.; Schaal, S. Reinforcement Learning of Motor Skills with Policy Gradients. *Neural Netw.* **2008**, *21*, 682–697, doi:10.1016/j.neunet.2008.02.003.
8. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Dy, J., Krause, A., Eds.; Volume 80, pp. 1861–1870.
9. Kiran, B.R.; Sobh, I.; Talpaert, V.; Mannion, P.; Sallab, A.A.A.; Yogamani, S.K.; Pérez, P. Deep Reinforcement Learning for Autonomous Driving: A Survey. *CoRR* **2020**, arxiv:2002.00444.
10. Wu, J.; Huang, Z.; Lv, C. Uncertainty-Aware Model-Based Reinforcement Learning with Application to Autonomous Driving. *arXiv* **2021**, arxiv:2106.12194.
11. Zhao, X.; Zhang, L.; Ding, Z.; Xia, L.; Tang, J.; Yin, D. Recommendations with Negative Feedback via Pairwise Deep Reinforcement Learning. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; doi:10.1145/3219819.3219886.
12. Zhang, S.; Yao, L.; Sun, A.; Tay, Y. Deep Learning Based Recommender System. *ACM Comput. Surv.* **2019**, *52*, 1–38, doi:10.1145/3285029.
13. Zhao, T.; Eskenazi, M. 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*; Association for Computational Linguistics: Los Angeles, CA, USA, 2016; pp. 1–10, doi:10.18653/v1/W16-3601.
14. Dhingra, B.; Li, L.; Li, X.; Gao, J.; Chen, Y.N.; Ahmed, F.; Deng, L. Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, AB, Canada, 30 July–4 August 2017; doi:10.18653/v1/p17-1045.
15. Shi, W.; Yu, Z. Sentiment Adaptive End-to-End Dialog Systems. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, QC, Australia, 15–20 July 2018; doi:10.18653/v1/p18-1140.
16. Shah, P.; Hakkani-Tür, D.; Liu, B.; Tür, G. 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)*; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 41–51, doi:10.18653/v1/N18-3006.
17. Ziebart, B.D.; Maas, A.L.; Bagnell, J.A.; Dey, A.K. *Maximum Entropy Inverse Reinforcement Learning*; AAAI: Chicago, IL, USA, 2008; Volume 8, pp. 1433–1438.
18. Russell, S. Learning agents for uncertain environments. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, Madison, WI, USA, 24–26 July 1998; pp. 101–103.
19. Ng, A.; Russell, S. Algorithms for Inverse Reinforcement Learning. In Proceedings of the ICML'00 Seventeenth International Conference on Machine Learning, Stanford, CA, USA, 29 June–2 July 2000.
20. Ho, J.; Ermon, S. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems 29, International Barcelona Convention Center*; Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016; pp. 4565–4573.
21. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 2672–2680.
22. Fu, J.; Luo, K.; Levine, S. Learning Robust Rewards with Adverserial Inverse Reinforcement Learning. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.

23. Takanobu, R.; Zhu, H.; Huang, M. 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)*; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 100–110, doi:10.18653/v1/D19-1010.
24. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. In Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing, Monticello, IL, USA, 22–24 September 1999; pp. 368–377.
25. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep Variational Information Bottleneck. *arXiv* **2016**, arxiv:1612.00410.
26. Peng, X.B.; Kanazawa, A.; Toyer, S.; Abbeel, P.; Levine, S. Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
27. Williams, J.; Raux, A.; Henderson, M. The Dialog State Tracking Challenge Series: A Review. *Dialogue Discourse* **2016**, *7*, 4–33.
28. Zhang, Z.; Huang, M.; Zhao, Z.; Ji, F.; Chen, H.; Zhu, X. Memory-Augmented Dialogue Management for Task-Oriented Dialogue Systems. *ACM Trans. Inf. Syst.* **2019**, *37*, doi:10.1145/3317612.
29. Wu, C.S.; Madotto, A.; Hosseini-Asl, E.; Xiong, C.; Socher, R.; Fung, P. Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; doi:10.18653/v1/p19-1078.
30. Schatzmann, J.; Thomson, B.; Weilhammer, K.; Ye, H.; Young, S. 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*; Association for Computational Linguistics: Rochester, NY, USA, 2007; pp. 149–152.
31. Gür, I.; Hakkani-Tür, D.; Tür, G.; Shah, P. User modeling for task oriented dialogues. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 900–906.
32. Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; Moritz, P. Trust Region Policy Optimization. In *Proceedings of the 32nd International Conference on Machine Learning*; Bach, F., Blei, D., Eds.; PMLR: Lille, France, 2015; Volume 37, pp. 1889–1897.
33. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arxiv:1707.06347.
34. Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; Abbeel, P. High-dimensional continuous control using generalized advantage estimation. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
35. Finn, C.; Levine, S.; Abbeel, P. 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. JMLR.org, ICML'16, New York, NY, USA, 19–24 June 2016; pp. 49–58.
36. Budzianowski, P.; Wen, T.H.; Tseng, B.H.; Casanueva, I.; Ultes, S.; Ramadan, O.; Gašić, M. 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*; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 5016–5026, doi:10.18653/v1/D18-1547.
37. Gašić, M.; Mrkšić, N.; Su, P.; Vandyke, D.; Wen, T.; Young, S. Policy committee for adaptation in multi-domain spoken dialogue systems. In Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, USA, 13–17 December 2015; pp. 806–812.
38. Wang, Z.; Bapst, V.; Heess, N.; Mnih, V.; Munos, R.; Kavukcuoglu, K.; de Freitas, N. Sample Efficient Actor-Critic with Experience Replay. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
39. Liu, B.; Lane, I. Adversarial Learning of Task-Oriented Neural Dialog Models. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 350–359, doi:10.18653/v1/W18-5041.
40. Tresp, V. A Bayesian Committee Machine. *Neural Comput.* **2000**, *12*, 2719–2741, doi:10.1162/089976600300014908.