

Original Paper

Robust Multi-Domain Multi-Turn Dialogue Policy via Student-Teacher Offline Reinforcement Learning

Mahdin Rohmatillah and Jen-Tzung Chien*

Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Taiwan

ABSTRACT

Dialogue policy plays a crucial role in a dialogue system as it determines the system response given a user input. In a pipeline system, the dialogue policy is susceptible to the performance degradation when the preceding components fail to produce correct output. To address this issue, this paper proposes a new method to train a robust dialogue policy that can handle noisy representation due to the mispredicted user dialogue acts from natural language understanding component. This method is mainly designed with two strategies, which are student-teacher learning and offline reinforcement learning. Student-teacher learning aims to force the student model to map the extracted features of the noisy input to be close to the clean features extracted by teacher model. Meanwhile, the offline reinforcement learning with multi-label classification objective is used to train the dialogue policy to provide appropriate response given user input by only utilizing the trajectories stored in the dataset. The experimental results show that the proposed hybrid learning can substantially improve the multi-turn end-to-end performance in a pipeline dialogue using MultiWOZ 2.1 dataset under ConvLab-2 evaluation framework. Furthermore, competitive results are obtained when compared to the end-to-end performance

*Corresponding author: Jen-Tzung Chien, jtchien@nycu.edu.tw

by using the pre-trained GPT-2 model with lower computational cost.

Keywords: Dialogue system, dialogue policy optimization, student-teacher learning, offline reinforcement learning

1 Introduction

Designing a faultless dialogue system is challenging, especially in the case of multi-domain multi-turn dialogue tasks where each conversation with multiple turns may comprise multiple domains such as finding the best Italian restaurant and the tourist attraction in the nearby location. Recent progress has shown two different approaches to design a dialogue system. The first one is the end-to-end approach in which all dialogue system components are represented by a single learning machine. This approach shows state-of-the-art results in the multi-domain dialogue task with convincing performance [17, 21, 51]. However, this approach is computationally expensive due to the utilization of large pre-trained language models (PLMs) such as the generative pre-trained transformer 2 (GPT-2) [28]. Furthermore, completely relying on large PLMs for end-to-end dialogue system is problematic, especially in case of the trustworthy issues. It is because that PLMs are prone to generate out-of-context sentences even given by in context input. Addressing this problem is required. For example, the two best models in DSTC9 track 2 applied the complicated pre-processing and post-processing stages to make sure the model could handle multi-domain dialogue task [21, 51]. The pre-processing stage involved fine-tuning the PLMs with several dialogue datasets, like Schema-Guided [29], Taskmaster [3] and CamRest676 [47]. Meanwhile, the post-processing stage introduced some special modules like fault tolerance mechanism and customized user interface that allowed manual revision.

Due to the aforementioned problems, a practical solution is to construct a pipeline dialogue system. As we can easily integrate any trustable rule function in a pipeline setup, the potential trustworthy issues can be resolved. Importantly, pipeline dialogue systems offer flexibility in optimizing each dialogue system component. Each component of the dialogue system can be optimized individually or jointly. Furthermore, the computation cost of the pipeline system is relatively cheaper compared to the end-to-end approach. Unfortunately, designing a high-performance pipeline dialogue system is demanding. It is because the performance of earlier components has a significant impact on output quality in later components. Especially in the multi-domain multi-turn dialogue task, like in MultiWOZ 2.1 [12], the errors from the natural language understanding (NLU) component considerably degrades the performance of

the dialogue system. Although the dialogue policy is well trained given a clean dataset, such a dialogue policy could not handle the noisy input due to the mispredicted user dialogue acts by the NLU component which contain unseen information. In Yeung *et al.* [50], a learning to learn method was proposed to handle noisy input in a classification task.

To cope with this challenge, this paper presents a robust dialogue policy which is trained to handle the noisy representation due to the errors caused by NLU component. This is different from Henderson *et al.* [14] where the noise robustness was developed on dialogue state tracking. This work focused on policy learning. The resulting dialogue policy is merged in the pipeline dialogue system that considerably saves computation time. The robustness of dialogue policy is pursued by using offline reinforcement learning (RL) with two different objectives which are optimized to improve noisy feature representation and action decision. To achieve the first objective, this study conducts a student-teacher learning where the student model is trained to map the extracted noisy features to be near to the clean features produced by the teacher model. A new kind of knowledge distillation [15, 20] is implemented for reinforcement learning of dialogue policy. Meanwhile, the second objective is pursued by using focal loss to deal with the multi-label classification problem. Before optimizing the dialogue policy, this paper introduces a data augmentation process by following a dialogue self-play [40] using the ConvLab-2 framework [54]. Such data augmentation is performed to collect additional expert trajectories so as to implement offline RL to train the policy. This study proposes a new approach to enhance the robustness of dialogue policy learning due to the errors introduced by NLU component. The merit of this work is illustrated by the experiments on multi-domain multi-turn dialogue policy optimization under various evaluation metrics.

The remainder of this paper is organized as follows. In Section 2, the multi-domain task-oriented dialogue system is introduced and the previous dialogue policy optimization approaches are surveyed. Section 3 presents the proposed robust dialogue policy learning. The overall learning process including data pre-processing, data augmentation and student-teacher model training are explained in details. Section 4 addresses the experimental settings for evaluation of multi-domain multi-turn dialogue policy followed by the experimental results to illustrate the benefits of the proposed work. The summary of findings from this study is given in Section 5.

2 Multi-Domain Task-Oriented Dialogue

The current methods to build multi-domain dialogue system along with dialogue policy learning [35] are surveyed.

2.1 Multi-Domain Dialogue System

Multi-domain task-oriented dialogue is a dialogue task in which there may exist more than one domain in each dialogue session. For example, in the travel assistant task, a user may request information about a train ticket, restaurant and hotel in a dialogue session. It is practical and interesting to work out a multi-domain dialogue system compared to a single domain dialogue [40, 1, 18, 8]. However, achieving a desirable performance for a multi-domain dialogue system is challenging. One of the main difficulties arise by the complexity of dialogue structure due to the information from several domains. Another challenge is the problem definition itself where the dialogue system must satisfy the user goal in a limited number of turns or steps, commonly under 10 conversation turns. In order to promote researches on such a dialogue system scenario, DSTC9 track 2 has provided ConvLab-2 [54] as an evaluation framework that integrates with the MultiWOZ 2.1 dataset [12]. Compared to the other frameworks designed for multi-domain dialogue task [44, 12], Convlab-2 offers two main advantages for system evaluation. Firstly, ConvLab-2 provides a simulated user which enables an end-to-end system evaluation with multi-turn evaluation. Such an evaluation can represent real-world phenomena compared to the component-wise and single-turn evaluation similar to the traditional evaluation [43]. Secondly, this framework offers the flexibility to train the dialogue system components consisting of natural language understanding (NLU), dialogue state tracking (DST), dialogue policy (POL) and natural language generation (NLG) [25], either individually or jointly trained. Figure 1 displays an illustration of how the simulated user and the multi-domain dialogue system interact with each other using the MultiWOZ 2.1 dataset under the ConvLab-2.

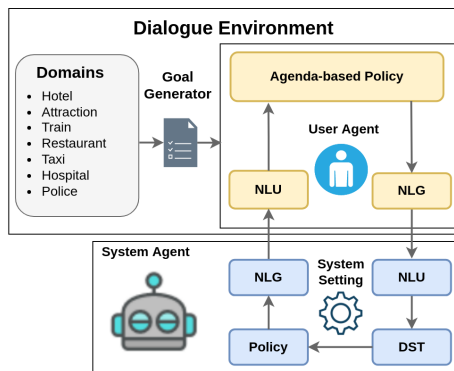


Figure 1: Interaction between system agent and simulated user in a multi-domain dialogue via an agenda-based policy using ConvLab-2 framework. The system agent can be set with different configurations.

2.2 Multi-Domain Dialogue Policy

Dialogue policy is an important component of the dialogue system which serves as a system brain that determines the response conditioned on dialogue context [32]. Dialogue policy optimization can be formulated under partially observable Markov decision process (POMDP) where the dialogue agent does not have any access to the complete state information $\mathbf{s} = \{g, \mathbf{o}\}$ consisting of user goal g and observation \mathbf{o} . Instead, the dialogue agent can only obtain \mathbf{o} to optimize its policy. Basically, dialogue policy π is optimized as π^* by using RL where the main objective is to maximize the discounted reward accumulation r given by state $\mathbf{s}_t = \{g, \mathbf{o}_t\}$ and system action \mathbf{a}_t along a whole trajectory over different time steps t via

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \gamma^t r(g, \mathbf{o}_t, \mathbf{a}_t) \right] \quad (1)$$

with a discount factor γ and termination time step T . \mathbf{a}_t is a sparse vector where the positive labels or one-hot values represent specific system dialogue acts in the current time step. Previously, the dialogue policy is optimized by utilizing deep RL techniques [7] such as the proximal policy optimization (PPO) [39], adversarial inverse reinforcement learning (AIRL) [13], generative adversarial imitation learning (GAIL) [16] and deep RL with human-in-the-loop paradigm [34, 33]. Methods such as deep Q network [27, 6] or REINFORCE [48] are also feasible to train dialogue policy. However, the dialogue policies optimized using those two methods did not perform well in the multi-domain case due to their inability to handle large state and action spaces. In the Convlab-2 framework, the observation vector \mathbf{o} is described as a 340-dimensional vector composed of six distinct components which are user action, system last action, belief state, book info, database pointer and termination with the detailed dimensions shown in Figure 2. Meanwhile, the system action \mathbf{a} is specified as a 209-dimensional vector which reflects the system dialogue acts.

Commonly, before starting RL procedure, the dialogue policy must be initially trained by using the behavior cloning method through leveraging the observation-action pairs $\mathcal{D} = \{\mathbf{o}, \mathbf{a}^*\}$ in MultiWOZ 2.1 dataset. Here, \mathbf{a}^* is the ground truth action provided by the dataset. Behavior cloning aims to find the policy network π_{θ} with parameter θ by minimizing the cross-entropy loss. Due to the multi-domain case, the learning process needs to deal with the multi-label classification task with sparse one-hot values. Therefore, most of the previous works applied the cross-entropy (CE) loss that merges a weight β to the positive labels or one-hot values in a form of

$$\begin{aligned} \mathcal{L}_{\text{CE}}(\mathbf{a}^*, \pi_{\theta}(\mathbf{a}|\mathbf{o})) = & -\mathbb{E}_{(\mathbf{o}, \mathbf{a}^*) \sim \mathcal{D}} [\beta \mathbf{a}^{*\top} \log \pi_{\theta}(\mathbf{a}|\mathbf{o}) \\ & + (\mathbf{1} - \mathbf{a}^*)^{\top} \log (\mathbf{1} - \pi_{\theta}(\mathbf{a}|\mathbf{o}))]. \end{aligned} \quad (2)$$

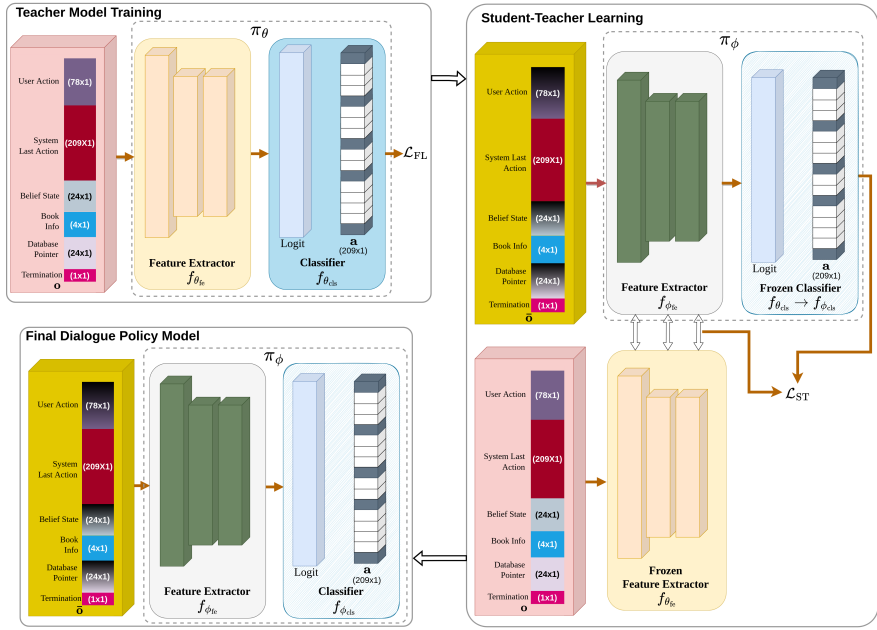


Figure 2: Student-teacher offline reinforcement learning process. It initially starts with teacher model trained with clean data, followed by student-teacher learning using both of clean and noisy data by the frozen teacher model. The well-trained student model is used as the final dialogue policy model where the student classifier parameter ϕ_{cls} is directly copied from the teacher classifier parameter θ_{cls} .

Once the behavior cloning or training is done, the dialogue policy is fine-tuned by using the RL algorithm via interacting with the simulated user in a pre-defined environment. In this setting, the reward is very sparse which may cause an unstable learning process. The dialogue agent will receive -1 in every conversation it makes, +5 if the current domain is satisfied, and +40 if all domains in the user goal are satisfied.

Other approaches were developed by jointly training the dialogue policy and the NLG component [53, 5] or even jointly training all of the dialogue components in an end-to-end fashion. In this setting, the action space is the number of all possible vocabularies. Therefore, the dialogue policy is trained with autoregressive learning. Notably, most of the previous works only showed good performance in the component-wise and single-turn evaluation [43]. Meanwhile, this work considers end-to-end system evaluation with the *multi-turn* evaluation as done in the DSTC9 track 2. Therefore, dealing with the error propagation from the preceding component is essential for a practical dialogue system. Based on the results in DSTC9 track 2, only limited end-to-

end approaches showed desirable performance [21, 51]. Those methods also required a huge amount of computational cost due to the implementation of large-sized PLM, which is GPT-2.

To relax the computational issue, this paper presents a new approach to train a robust dialogue policy for pipeline dialogue system that offers a low computational cost. The dialogue policy is trained by using student-teacher offline RL to improve its robustness when dealing with noisy input representation. The teacher is designed to guide the student to map the noisy input into the same representation as the clean input. Meanwhile, the offline RL is exploited to show a potential dialogue learning technique when the abundance of trajectories from data augmentation [9] is available.

3 Robust Multi-Domain Multi-Turn Dialogue Policy Learning

This section addresses the proposed policy learning strategy for multi-domain multi-turn dialogues. This strategy is generally divided into three stages, as shown by Figure 3. The first is the data pre-processing followed by data augmentation using the dialogue self-play [40]. After the dataset is augmented, the student-teacher dialogue policy model is trained by using the offline reinforcement learning.

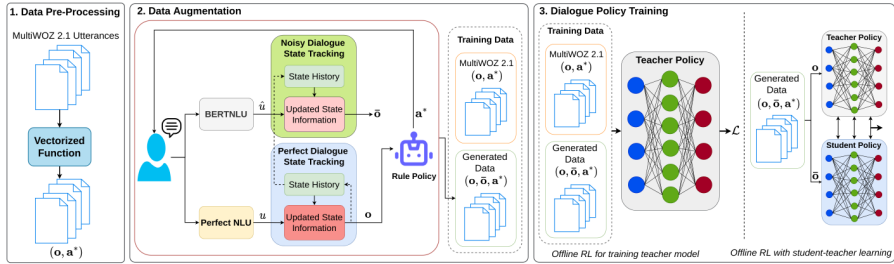


Figure 3: Overview of the proposed learning strategy for robust dialogue policy consisting of the stages of data pre-processing, data augmentation and dialogue policy training. Perfect NLU means that the true user dialogue acts are directly passed to the DST component.

3.1 Data Pre-Processing and Data Augmentation

Before starting the training procedure of dialogue policy, the conversations in each dialogue session in the MultiWOZ 2.1 dataset need to be processed by converting the raw text data into the observation-action pairs $\{o, a^*\}$ by using the vectorized functions provided by ConvLab-2. This pre-process results in the 340-dimensional observation and 209-dimensional action vector representations, $o \in \mathbb{R}^{340}$ and $a \in \mathbb{R}^{209}$. The total number of the generated

pairs reaches around 50K observation-action pairs. This number may not be large enough for training an offline RL algorithm that commonly requires huge numbers of trajectories as the training data. Therefore, this work introduces a data augmentation process by using the dialogue *self-play* method [40]. In this setting, the rule-based policy will interact with the simulated user driven by the agenda-based policy. Besides adding the augmented data for offline training, the data augmentation process aims to generate noisy data to improve the robustness of dialogue policy in the subsequent offline RL training stage. The dialogue *self-play* is implemented by utilizing the provided RL environment in the ConvLab-2 where all of the stored trajectories in the replay buffer are saved as the augmented data. There will be no overlap with test data which follows the DSTC-9 track 2 evaluation where the evaluation was done under an end-to-end system evaluation environment in the Convlab-2.

However, generating meaningful noisy data in the multi-domain dialogue is non-trivial due to the enriched complexity of observation representation. In the multi-domain dialogue tasks, the user dialogue act consisting of four entries is formed as $[intent, domain, slot, value]$. Randomly introducing noise to observation \mathbf{o} is not helpful since this may produce irrelevant domain information. For instance, considering a user input ‘‘I want to find an Italian restaurant.’’, the correct user dialogue act would be [‘Inform’, ‘Restaurant’, ‘Type’, ‘Italian’]. However, the dialogue act due to random noise might be produced as [‘Inform’, ‘Taxi’, ‘Type’, ‘Indian’], which bears no relevant information to user input. Particularly, the entries of *slot* and the *value* have no connection to that of *domain* within this noisy dialogue act.

As the main objective of this work is to compensate the noise from NLU component which is mostly represented by BERTNLU, an NLU that uses BERT [11, 36] as a base language model, then the noisy observation $\bar{\mathbf{o}}$ is generated by using the user dialogue acts $\{\hat{u}_1, \dots, \hat{u}_m\}$ predicted by BERT as shown by Figure 3 in the middle part. Each predicted dialogue act \hat{u} contains various information in four entries *intent*, *domain*, *slot* and *value* from the current user input. The BERTNLU predicted user dialogue acts will change the information of user action, belief state and database pointer stored in observation vector \mathbf{o} . The process of generating user dialogue acts can be expressed by

$$\{\hat{u}_1, \dots, \hat{u}_m\} = f_{\text{MLP}}(f_{\text{BERT}}(x_1, \dots, x_m))$$

where f_{MLP} is the multi-layer perceptron (MLP) network added on the top of BERT model f_{BERT} for intent classification and slot tagging. The input of BERT $\{x_1, \dots, x_m\}$ is the ground-truth conversation history between simulated user and dialogue system with perfect NLU. To ensure that the observation $\bar{\mathbf{o}}$ may reflect reasonable noise, the state history in the noisy DST is set to be identical with the state history in the perfect DST, as shown by the dashed line. The generated noise can be said as the noise that occurs in the

current time step only. The action \mathbf{a}^* is the output of rule-based policy given clean observation \mathbf{o} . After the data augmentation process is completed, both MultiWOZ 2.1 and the generated dataset are used to train the dialogue policy.

3.2 Student-Teacher Offline Reinforcement Learning

Before training the dialogue policy by using student-teacher offline RL, we initially train the teacher model by using the behavior cloning to minimize the action probability discrepancy between the learned policy π_θ and the expert policy π_E that can be measured by using the Kullback-Leibler (KL) divergence with approximation as expressed by

$$\begin{aligned} \min_{\pi_\theta} \mathbb{E}_{\mathbf{o} \sim \eta_{\pi_E}} [D_{\text{KL}}(\pi_E(\cdot|\mathbf{o}) \parallel \pi_\theta(\cdot|\mathbf{o}))] \\ \approx \mathbb{E}_{(\mathbf{o}, \mathbf{a}) \sim \rho_{\pi_E}} \left[\log \left(\frac{\pi_E(\mathbf{a}|\mathbf{o})}{\pi_\theta(\mathbf{a}|\mathbf{o})} \right) \right] \end{aligned} \quad (3)$$

where η_{π_E} and ρ_{π_E} denote the state and trajectory distributions under the expert dialogue policy π_E , respectively. Both η_{π_E} and ρ_{π_E} can be represented as the dataset used for offline RL training. The most common offline RL method used in this pre-training stage is the behavior cloning which is the simplest approach to learn the expert policy. Although the previous works mentioned that behavior cloning is prone to the compounding errors [26, 46], as the empirical evidences shown in the experimental results, we argue that the compounding errors which can degrade the performance of the dialogue policy do not really exist in the task-oriented dialogue, as long as we can make sure that the policy discrepancy between π_E and π_θ in Eq. (3) is very small with a bound ϵ , namely

$$\mathbb{E}_{\mathbf{o} \sim \eta_{\pi_E}} [D_{\text{KL}}(\pi_E(\cdot|\mathbf{o}) \parallel \pi_\theta(\cdot|\mathbf{o}))] \leq \epsilon. \quad (4)$$

It is because each conversation in the task-oriented dialogue is made under *very short trajectory* in a *finite horizon* which accordingly minimizes the probability of occurring distributional shift issue leading to severe compounding errors.

It can be verified by deriving the difference of value functions between expert policy π_E and imitated policy π_θ in the finite horizon of POMDP. First of all, consider the value function of a finite horizon as the expected total reward obtained by policy π without any discount factor γ

$$V_\pi = \mathbb{E}_{\tau \sim \rho_\pi} [r(g, \mathbf{o}, \mathbf{a})]. \quad (5)$$

τ is a trajectory consisting of (\mathbf{o}, \mathbf{a}) pairs from initial until terminal time steps. By using the definition of $\|\mu - \nu\|_1 = 2D_{\text{TV}}(\mu \parallel \nu)$ where $D_{\text{TV}}(\mu \parallel \nu)$ denotes the total variance distance over two probability distributions μ and

ν , the difference of values between policies π_E and π_θ can be expressed and manipulated to derive an upper bound as follows

$$\begin{aligned}
V_{\pi_E} - V_{\pi_\theta} &= \mathbb{E}_{\tau \sim \rho_{\pi_E}} [r(g, \mathbf{o}, \mathbf{a})] - \mathbb{E}_{\tau \sim \rho_{\pi_\theta}} [r(g, \mathbf{o}, \mathbf{a})] \\
&= \sum_{(\mathbf{o}, \mathbf{a}) \in \mathcal{O} \times \mathcal{A}} (\rho_{\pi_E}(\mathbf{o}, \mathbf{a}) - \rho_{\pi_\theta}(\mathbf{o}, \mathbf{a})) r(g, \mathbf{o}, \mathbf{a}) \\
&= 2 \cdot r(g, \mathbf{o}, \mathbf{a}) \cdot D_{\text{TV}}(\rho_{\pi_E}(\mathbf{o}, \mathbf{a}) \| \rho_{\pi_\theta}(\mathbf{o}, \mathbf{a})) \\
&\leq 2 \cdot r(g, \mathbf{o}, \mathbf{a}) \cdot \sum_{t=0}^{T-1} \mathbb{E}_{\mathbf{o} \sim \eta_{\pi_E}^t} [D_{\text{TV}}(\pi_E(\cdot | \mathbf{o}) \| \pi_\theta(\cdot | \mathbf{o}))] \\
&\leq (2T) \cdot r(g, \mathbf{o}, \mathbf{a}) \cdot \mathbb{E}_{\mathbf{o} \sim \eta_{\pi_E}} [D_{\text{TV}}(\pi_E(\cdot | \mathbf{o}) \| \pi_\theta(\cdot | \mathbf{o}))] \triangleq \mathcal{U}_V.
\end{aligned} \tag{6}$$

The last two inequalities were provided in Ross *et al.* [37] and further explained in Ke *et al.* [19]. Next, by using the Pinsker's inequality [10] that states $D_{\text{TV}}(\mu \| \nu) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(\mu \| \nu)}$, the upper bound \mathcal{U}_V in Eq. (6) for finite horizon case can be further tightened by considering the bound ϵ in Eq. (4) in a form of

$$\begin{aligned}
\mathcal{U}_V &\leq (2T) \cdot r(g, \mathbf{o}, \mathbf{a}) \cdot \mathbb{E}_{\mathbf{o} \sim \eta_{\pi_E}} \left[\sqrt{\frac{1}{2} D_{\text{KL}}(\pi_E(\cdot | \mathbf{o}) \| \pi_\theta(\cdot | \mathbf{o}))} \right] \\
&\leq (\sqrt{2}T) \cdot r(g, \mathbf{o}, \mathbf{a}) \cdot \sqrt{\mathbb{E}_{\mathbf{o} \sim \eta_{\pi_E}} [D_{\text{KL}}(\pi_E(\cdot | \mathbf{o}) \| \pi_\theta(\cdot | \mathbf{o}))]} \\
&\leq (\sqrt{2\epsilon}T) \cdot r(g, \mathbf{o}, \mathbf{a}).
\end{aligned} \tag{7}$$

By considering the fact that finite number of turns T in dialogue task is likely small, around 10 turns at most which only happens in the dialogues containing three domains and at the average only 6 turns in the dialogues containing one and two domains, it is reasonable that the compounding errors of behavior cloning in a short horizon will not be really harmful as long as we can guarantee that the discrepancy of policies between π_E and π_θ is small enough.

Unlike the previous works [34, 16, 42] which employed the cross-entropy or balanced cross-entropy loss to minimize Eq. (3), in this work, the dialogue policy is trained by minimizing the focal loss (FL) [22] which addresses the sparse multi-label classification via

$$\begin{cases} \mathcal{L}_+^{\text{FL}} = (\mathbf{1} - \pi_\theta(\mathbf{a} | \mathbf{o}))^\omega \odot \log(\pi_\theta(\mathbf{a} | \mathbf{o})) \\ \mathcal{L}_-^{\text{FL}} = (\pi_\theta(\mathbf{a} | \mathbf{o}))^\omega \odot \log(\mathbf{1} - \pi_\theta(\mathbf{a} | \mathbf{o})) \end{cases} \tag{8}$$

where ω is a focusing parameter, \odot is the element-wise product, \mathcal{L}_+ and \mathcal{L}_- are the losses for positive and negative labels. The total loss can be calculated through

$$\mathcal{L}_{\text{FL}}(\mathbf{o}, \mathbf{a}^*; \theta) = -\mathbf{a}^\top \mathcal{L}_+^{\text{FL}} - (\mathbf{1} - \mathbf{a})^\top \mathcal{L}_-^{\text{FL}}. \tag{9}$$

If we set $\omega > 0$, the focal loss will down weight the easy negatives which have low probability ($\pi_\theta(\mathbf{a}|\mathbf{o}) \ll 0.5$). This means that the learning process will focus more on the harder samples. Meanwhile, if we set $\omega = 0$, it will be reduced to the standard binary cross-entropy loss. Even though the balanced cross-entropy loss introduces a weighting factor β to balance the importance of positive and negative samples, the weighting factor in the focal loss is adaptive to the current learning progress to distinguish between easy and hard examples.

Since the main objective of this paper is to train a robust dialogue policy, the objective stated in Eq. (3) should be changed by considering the noise from BERTNLU model that can be simply formulated as $\bar{\mathbf{o}} = \text{BERTNLU}(\mathbf{o})$. The probability discrepancy between π_E and π_θ is then minimized and approximated as

$$\begin{aligned} & \min_{\pi_\theta} \mathbb{E}_{\mathbf{o} \sim \eta_{\pi_E}} [D_{\text{KL}}(\pi_E(\cdot|\mathbf{o}) \| \pi_\theta(\cdot|\text{BERTNLU}(\mathbf{o})))] \\ & \approx \mathbb{E}_{(\mathbf{o}, \mathbf{a}^*) \sim \rho_{\pi_E}} \left[\log \left(\frac{\pi_E(\mathbf{a}^*|\mathbf{o})}{\pi_\theta(\mathbf{a}^*|\text{BERTNLU}(\mathbf{o}))} \right) \right]. \end{aligned} \quad (10)$$

In order to mitigate such noise, this work introduces a student-teacher offline RL which mainly aims to improve the representation learning. Under this learning stage, we define the teacher parameters as $\theta = \{\theta_{\text{fe}}, \theta_{\text{cls}}\}$ and student parameters as $\phi = \{\phi_{\text{fe}}, \phi_{\text{cls}}\}$ which consist of the parameters of feature extractor and dialogue act classifier. Then, the learning objective can be defined to force the noisy features extracted by the student model π_ϕ given $\bar{\mathbf{o}}$ to be similar to the clean features extracted by the teacher model π_θ given \mathbf{o} . The learning process of this stage is depicted by Figure 2. As the teacher classifier can show convincing performance given clean input, then the student-teacher offline RL only focuses on how to improve the feature extractor. The objective of this learning can be formulated by minimizing the following student-teacher (ST) loss function

$$\mathcal{L}_{\text{ST}}(\mathbf{o}, \bar{\mathbf{o}}, \mathbf{a}^*; \phi, \theta) = \mathbb{E}_{(\mathbf{o}, \mathbf{a}^*, \bar{\mathbf{o}}) \sim \mathcal{D}} [\mathcal{L}_{\text{FL}}(\bar{\mathbf{o}}, \mathbf{a}^*; \phi) + \lambda \|f_{\theta_{\text{fe}}}(\mathbf{o}) - f_{\phi_{\text{fe}}}(\bar{\mathbf{o}})\|_2^2]. \quad (11)$$

The first term is the focal loss given noisy input $\bar{\mathbf{o}}$ that will be used to update ϕ_{fe} to find the best representation for $f_{\phi_{\text{cls}}}$ to correctly predict \mathbf{a} , which is the same label for \mathbf{o} . ϕ_{cls} directly copies θ_{cls} and the parameters are frozen during training. Meanwhile, the second term in Eq. (11) is a feature distance loss that will be directly used to train $f_{\phi_{\text{fe}}}$ to map the noisy input representation to be as close as possible to the clean data representation extracted by $f_{\theta_{\text{fe}}}$. The overall learning process of student-teacher offline RL is shown by Algorithm 1 which first calculates teacher model and then student model.

Algorithm 1: Student-teacher offline reinforcement learning for robust dialogue policy.

Require : dialogue dataset \mathcal{D} containing tuples of $(\mathbf{o}, \bar{\mathbf{o}}, \mathbf{a}^*)$,
hyperparameters $\{\alpha, \omega\}$
Initialize teacher parameters $\theta = \{\theta_{\text{fe}}, \theta_{\text{cls}}\}$
for each updating iteration **do**
 sample minibatch of $\{\mathbf{o}_i, \mathbf{a}_i^*\}_{i=1}^N$ from \mathcal{D}
 calculate gradient of \mathcal{L}_{FL} (Eq. (9)) by $g_\theta \leftarrow \nabla \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\text{FL}}(\mathbf{o}_i, \mathbf{a}_i^*; \theta)$
 $\theta \leftarrow \text{AdamW}(\theta, \alpha, g_\theta)$
Initialize student parameters $\phi \leftarrow \theta$
for each updating iteration **do**
 sample minibatch of $\{\mathbf{o}_i, \bar{\mathbf{o}}_i, \mathbf{a}_i^*\}_{i=1}^N$ from \mathcal{D}
 calculate gradient of \mathcal{L}_{ST} (Eq. (11)) by
 $g_{\phi_{\text{fe}}} \leftarrow \nabla \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\text{ST}}(\mathbf{o}_i, \bar{\mathbf{o}}_i, \mathbf{a}_i^*; \phi, \theta)$
 $\phi_{\text{fe}} \leftarrow \text{AdamW}(\phi_{\text{fe}}, \alpha, g_{\phi_{\text{fe}}})$

4 Experiments

The experiments were done by using ConvLab-2 framework. ConvLab-2 provides a framework to allow interactions between simulated user and dialogue agent in an environment by using MultiWOZ 2.1 [12] dataset. MultiWOZ 2.1 was an updated version of MultiWOZ 2.0 [2], known as a multi-domain, multi-intent task-oriented dialog corpus [41] that contained 7 domains which are hotel, attraction, restaurant, train, taxi, police and hospital, 13 user intents, 25 slot types, 10,483 dialog sessions, and 71,544 dialog turns. By using ConvLab-2, the end-to-end system evaluation was performed to sufficiently reflect real-world scenarios.

4.1 Experimental Settings

The MultiWOZ 2.1 dataset was initially split into the training, validation and test data with 8434, 999 and 1000 samples, respectively. The feature extractor of both student model $f_{\phi_{\text{fe}}}$ and teacher model $f_{\theta_{\text{fe}}}$ were formed by MLP networks with two hidden layers consisting of 100 neurons with activation function ReLU. Meanwhile, the teacher and student classifiers, $f_{\theta_{\text{cls}}}$ and $f_{\phi_{\text{cls}}}$, respectively, were based on MLP networks with sigmoid activation function. The focusing parameters ω and λ in the feature distance loss Eq. (11) were set to 2 and 1, respectively. The dialogue policy parameters $\{\theta, \phi\}$ were optimized by using Adam with the decoupled weight decay (AdamW) [24] with learning rate $\alpha = 5\text{e-}4$ and batch size $N = 32$. The data generated in the data augmentation process were around 36000 observation-action pairs.

The benefit of the proposed method for improving dialogue system robustness was evaluated with the end-to-end system evaluation over a set of 1000 dialogues, consisting of 337, 523 and 140 dialogues containing 1, 2 and 3 domains, respectively. The dialogue system performance was measured through the interactions with the simulated user where the NLU, DST and NLG in a pipeline system were identical to the ConvLab-2 default settings, which included BERT [11] based NLU (BERTNLU), rule-based DST and template NLG, respectively. The experimental results were carried out on a PC with a CPU i9-10900K, 128GB of RAM, and a GPU NVIDIA RTX 2080Ti. Comparative studies were performed by using five primary metrics listed below

- success rate: judges whether constraints (e.g. hotel location or hotel price) and requests (e.g. hotel phone number) in the user goals have been satisfied by system.
- F1 score: judges if all requested information like taxi type or taxi phone number has been informed. This score is computed from precision and recall.
- complete rate: ratio of the completed user constraints.
- booking rate: ratio of the successful dialogues for booking a request which is only available in the domains of hotel, restaurant and train.
- average conversation turns between dialogue system and user for successful and all dialogues.

For the first four metrics, the higher the better. For the last metric, the lower the better. The proposed method was compared with two types of baseline methods. The first type of baselines was the methods which only optimized the dialogue policy, mentioned as follows

- behavior cloning optimized with balanced cross-entropy loss (Eq. (2)) and asymmetric loss (ASL) [30]. ASL was shown good benefit in the multi-label image classifications, leading to state-of-the-art results. ASL is an extension of focal loss where the weights for positive and negative labels are different, ω^+ and ω^- , respectively, by

$$\begin{cases} \mathcal{L}_+^{\text{ASL}} = (\mathbf{1} - \pi_\theta(\mathbf{a}|\mathbf{o}))^{\omega^+} \odot \log(\pi_\theta(\mathbf{a}|\mathbf{o})) \\ \mathcal{L}_-^{\text{ASL}} = (\pi_\theta^m(\mathbf{a}|\mathbf{o}))^{\omega^-} \odot \log(\mathbf{1} - \pi_\theta^m(\mathbf{a}|\mathbf{o})) \end{cases} \quad (12)$$

The total loss is accumulated just like Eq. (9). $\pi_\theta^m(\mathbf{a}|\mathbf{o})$ is an asymmetric probability with a shift m calculated by $\pi_\theta^m(\mathbf{a}|\mathbf{o}) = \max(\pi_\theta(\mathbf{o}) - m, 0)$. There are three hyperparameters ω^+ , ω^- and m in this method.

- policy gradient (PG) [48]: a standard policy based method in RL where the gradient with respect to the cumulative reward is calculated to update $\pi_{\theta}(\cdot)$
- proximal policy optimization (PPO) [39]: an actor-critic method implemented by maximizing the clipped surrogate objective to train the actor and minimizing the regression error to train the critic.
- guided dialogue policy learning (GDPL) [42]: a solution based on adversarial inverse RL [13] method which learns a reward function by using the expert trajectories and uses this information to train the dialogue policy agent sequentially in the same loop.

Another type of baselines conducted the optimization in an end-to-end fashion. All components from NLU until NLG were optimized jointly.

- domain aware multi-decode (DAMD) [52]: a multi-action data augmentation scheme to produce diverse response by using additional state-action pairs.
- minimalist transfer learning (MinTL) [23]: a transfer learning framework offering plug-and-play approach for task-oriented dialogue system.
- UBAR [49]: a task-oriented dialogue model which used distilGPT-2 model [38] as the base model. The model was fed not only with the user and response sentences, but also with database search result and belief state from the previous steps.
- offline RL methods, which are GPT-critic [17], critic regularized regression (CRR) [45] and decision transformer [4]. All of them used GPT-2 as the base model.
- GPT-TDAPT (GPT-2 model with task and domain adaptive pretraining) [51]: proposed five stages of learning. The first was the domain adaptation using the pre-trained GPT-2 where the datasets including Schema [29], Camrest [47], Taskmaster 2019, Taskmaster 2020 [3] and MSR-e2e were used. Multi-task fine-tuning using MultiWOZ 2.1, data pre-processing and post-processing, fault tolerance mechanism, and rule-based post-processing for refining the agent utterances were the other four stages.
- AUGPT (GPT-2 finetuned with auxiliary tasks) [21]: conducted similar implementation as the GPT-TDAPT with two distinctions. First, there was no post-processing approach in this work. Second, the auxiliary tasks were employed to increase the consistency in sentence generation given the identical system action responses.

4.2 Experimental Results

Firstly, the evaluation was carried out to investigate the effect of data augmentation by utilizing the dialogue self-play method under different behavior cloning objectives. The configuration of the dialogue system was identical for the methods including BERTNLU, rule-based DST and template NLG. We set $\beta, m, \omega^+, \omega^-$ to be 5, 0.05, 0 and 2, respectively. As shown in Table 1, the proposed data augmentation process successfully produced the meaningful data, as indicated by the performance gaps between the models trained by using only MultiWOZ 2.1 dataset and the models trained with the generated data. One of the main reason of this phenomena is the simplicity offered by the generated data indicated by the number of system dialogue acts in the action vector \mathbf{a} , as shown by Figure 4. Compared to the generated data, MultiWOZ 2.1 contains more complicated observation-label pairs as more than one-fifth of data samples have at least 4 positive labels. Furthermore, it also contains the data with more than 8 positive labels which might be hard for the model to understand them. Therefore, dialogue policy could learn better from the generated dataset than the original dataset. Although the performance gap between models trained using MultiWOZ 2.1 and generated data is significant, the best results were obtained by combining those two datasets. Training the models solely by using the generated data only resulted in suboptimal performances. This evidence implies that each dataset has unique important information that must be learned by the dialogue policy. Another interesting finding is the suboptimal performance of the behavior cloning with asymmetric loss in comparison to the behavior cloning with focal loss, even after adjusting the hyperparameters in the Eq. (12). The results are shown by Table 2. Since the behavior cloning with asymmetric loss is quite sensitive to the hyperparameter tuning, the resulting performance is worse than that with focal loss. Using asymmetric loss, the most notable evidence is seen when the precision score drastically declined after the value of ω^- was raised. Despite the fact that the success rate difference between behavior cloning trained with asymmetric loss and focal loss is relatively not significant, the focal loss effectively helped the model in achieving a high F1 score of 88.6 as compared to 85.1 for the model trained with asymmetric loss.

The next experiment was performed in order to demonstrate the benefit of the proposed method, called the student-teacher offline RL (denoted by STORL). In this evaluation, the proposed method was compared with both pipeline and end-to-end dialogue system. The evaluation results are shown by Table 3. From the result, it can be seen that the proposed method properly learned how to map the noisy input $\bar{\mathbf{o}}$ into the better representation that was closer to the representation of teacher model given clean input \mathbf{o} . As a result, the dialogue policy could reduce the impact of the noisy input while still generating appropriate actions, indicated by absolute improvements 1.4%

Dialogue	Success	Inform			Complete	Booking	Average
Policy Objective	Rate (%)	Precision	Recall	F1	Rate (%)	Rate (%)	Turn (Succ/All)
MultiWOZ 2.1							
Balanced cross-entropy	48.8	71.3	76.7	71.1	53.0	13.2	13.0/18.7
Asymmetric loss	52.1	81.2	81.0	78.7	58.1	18.3	12.6/15.6
Focal loss	41.1	86.6	73.5	77.0	44.5	0.03	10.9/22.3
Generated Data							
Balanced cross-entropy	78.9	79.8	89.2	82.3	81.3	89.5	11.8/12.7
Asymmetric loss	80.0	80.0	92.0	83.1	88.0	90.2	11.2/12.8
Focal loss	81.3	85.0	92.2	86.7	90.0	91.0	11.2/13.7
MultiWOZ 2.1 + Generated Data							
Balanced cross-entropy	79.1	83.2	91.4	84.7	86.8	90.2	11.2/13.4
Asymmetric loss	82.6	82.4	93.0	85.1	91.4	91.1	11.4/12.2
Focal loss	83.3	86.7	93.4	88.6	92.0	91.8	11.4/12.3

Table 1: The performance comparison of behavior cloning trained with three different loss functions under three different datasets. The bold numbers indicate the best score in the metrics.

Dialogue Policy Objective	Hyperparameter Setting	Success Rate (%)	Inform			Complete Rate (%)	Booking Rate (%)	Average Turn (Succ/All)
			Precision	Recall	F1			
Asymmetric loss	$\omega^+ = 0, \omega^- = 2, m = 0.05$	83.0	82.4	93.0	85.1	91.4	91.1	11.4/12.2
	$\omega^+ = 0, \omega^- = 4, m = 0.05$	82.0	75.3	90.8	79.4	87.7	89.1	11.1/12.0
	$\omega^+ = 0, \omega^- = 5, m = 0.05$	82.5	79.2	93.1	83.1	90.8	90.7	11.4/12.3
	$\omega^+ = 1, \omega^- = 2, m = 0.05$	82.1	80.4	92.8	83.8	90.4	90.1	11.3/12.6
	$\omega^+ = 1, \omega^- = 4, m = 0.05$	82.0	73.1	92.4	78.9	90.0	89.9	11.3/12.0
	$\omega^+ = 1, \omega^- = 5, m = 0.05$	80.2	64.6	90.2	72.2	86.7	90.0	11.5/12.0
Focal loss	$\omega^+ = 2, \omega^- = 2, m = 0$	83.3	86.7	93.4	88.6	92.0	91.8	11.4/12.3

Table 2: The performance comparison between behavior cloning trained with asymmetric loss and focal loss with different hyperparameter settings.

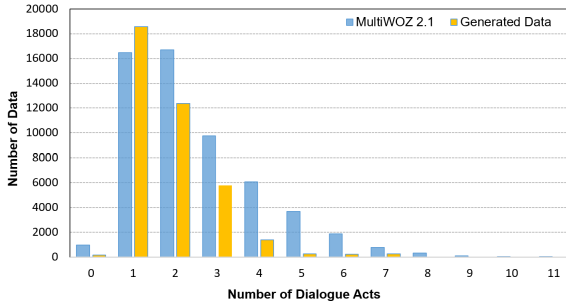


Figure 4: A comparison between MultiWOZ 2.1 data and the generated data in terms of dialogue act number in the action or label vector \mathbf{a} .

and 0.8% in the success rate and complete rate, respectively, when compared to the teacher model. Conditioned on the clean input, which means the true user dialogue acts were directly passed to the DST component, the learned student encoder $f_{\phi_{fe}}$ could maintain to map the clean input to the appropriate representation indicated by the nearly identical performance between teacher model and the proposed method given by the perfect NLU setting. Furthermore, compared to the rule-based policy which is always set as the upper bound of

Table 3: End-to-end system evaluation results of the proposed student-teacher offline RL (STORL) compared to the previous approaches. The bold numbers indicate the best score in the metrics without considering system with perfect NLU. AUGPT [21] is a model participating in DSTC9 track 2. * means the scores were obtained by running the provided models. ** means the scores were obtained from the results mentioned in GPT-critic paper [17].

Configuration				Success	F1	Complete	Booking	Average
NLU	DST	Policy	NLG	Rate (%)	Score	Rate (%)	Rate (%)	Turn (Succ/All)
BERT	Rule	PG*	Template	44.7	60.6	47.1	29.7	12.5/20.1
BERT	Rule	GDPL*	Template	47.2	64.6	50.0	26.8	11.9/19.3
BERT	Rule	PPO*	Template	61.2	68.2	64.7	62.4	13.0/18.1
BERT	Rule	PPO HITL*	Template	81.4	84.6	86.2	88.4	11.3/12.4
BERT	Rule	Rule	Template	83.8	86.2	92.7	91.5	11.4/11.9
		End-to-End (DAMD)*		34.2	56.9	39.6	52.0	15.6/30.2
		End-to-End (MINTL)**		68.1	69.0	71.4	65.4	15.7/20.7
		End-to-End (UBAR)**		74.3	76.0	79.8	80.8	14.2/18.1
		End-to-End (CRR)**		72.6	76.0	78.2	82.2	13.6/17.9
		End-to-End (Decision Transformer)**		75.3	77.0	81.3	83.5	14.8/18.0
		End-to-End (GPT-Critic)**		77.7	79.0	84.3	85.4	16.3/19.4
		End-to-End (AUGPT (DSTC9 track 2))*		60	70.2	89.3	86	12.7/13.9
BERT	Rule	STORL	Template	84.7	86.9	92.8	92.1	11.5/12.3
BERT	Rule	STORL (Teacher)	Template	83.3	88.6	92.0	91.8	11.4/12.3
Perfect	Rule	STORL	Template	93.0	89.5	96.1	97.7	11.6/12.0
Perfect	Rule	STORL (Teacher)	Template	92.6	91.2	96.0	98.0	11.6/12.0

Table 4: End-to-end system evaluation results on STORL compared to the previous approaches which were optimized by using the augmented data during pre-training stage.

Configuration				Success	F1	Complete	Booking	Average
NLU	DST	Policy	NLG	Rate (%)	Score	Rate (%)	Rate (%)	Turn (Succ/All)
BERT	Rule	PG	Template	80.6	83.4	87.9	90.7	11.5/13/1
BERT	Rule	GDPL	Template	81.3	84.8	89.8	89.8	11.4/12/6
BERT	Rule	PPO	Template	82.0	84.8	90.1	90.7	11.3/12.4
BERT	Rule	PPO HITL	Template	82.8	84.8	91.2	90.9	11.3/12.4
BERT	Rule	STORL (Teacher)	Template	83.3	88.6	92.0	91.8	11.4/12.3
BERT	Rule	STORL	Template	84.7	86.9	92.8	92.1	11.5/12.3
BERT	Rule	Rule	Template	83.8	86.2	92.7	91.5	11.4/ 11.9

the dialogue policy optimization, the proposed learning scenario even performs better than the rule-based policy, especially in terms of success rate and booking rate by 0.9% and 0.6%, respectively. Unfortunately, there was a trade-off that should be paid in this learning strategy. Due to the enforcement to map noisy representation to be as close as possible to the clean representation, F1 score of the learned dialogue policy was degraded from the teacher model, by around 1.8.

In order to show further comparison with the dialogue policy baselines, additional experiments were conducted by training all of the dialogue policy baselines by utilizing a combination of MultiWoZ 2.1 with the augmented data. The augmented data were leveraged in the pre-training stage. The optimization process followed the default setting of each baseline method which mainly used the behavior cloning in accordance with the balanced cross-entropy loss. The results are shown in Table 4. As can be observed, all of the dialogue policy

baselines are significantly improved due to the addition of augmented data during the pre-training phase. The best baseline performance is obtained by PPO with human-in-the-loop (PPO HITL) which can reach the performance close to that using the STORL (Teacher) model. However, by using this setting, the learning strategy in PPO HITL introduced expert feedback twice. The first was in the pre-training stage which utilized the expert trajectories, and the second was in the PPO training in which an expert provided the action corrections in every interaction with the simulated user. As all of the dialogue policy baselines could not perform better than STORL (Teacher) model, this finding indicates that the offline RL is enough for solving the dialogue policy learning in the dialogue task. Also, the resulting training data are sufficient so that the learning strategy can guarantee to reach the derived bound in the Section 3.2 which is achieved by using the focal loss. Furthermore, even though all of the baselines were trained by using an additional dataset, all of them still could not outperform the rule-based policy, while the STORL showed better performance than the rule-based policy due to the well-trained encoder that can compensate for the errors from BERTNLU.

Compared to the other baselines that applied either RL or offline learning with large PLMs, the proposed method was only outperformed by the GPT-TDAPT [51] which achieved the best performance in the DSTC9 track 2. Table 5 shows the performance comparison between STORL and GPT-TDAPT which also includes the performance of each model given different test data categorized by the number of domain occurrences in each dialogue. Even though GPT-TDAPT outperformed STORL in overall testing, in terms of computational cost comparison, the proposed method only required 7.5 minutes to complete the testing consisting of 1000 test dialogues. GPT-TDAPT required nearly 74 minutes to complete the testing process. Moreover, the learning process introduced by GPT-TDAPT implemented some non-trivial tricks in both pre-processing and post-processing stages which required human manual design to make sure the model could achieve a convincing result. As a result, the other end-to-end approaches that employed GPT-2 model such as AUGPT [21] and GPT-critic [17] showed suboptimal performances due to the missing of the manual design tricks. On the other hand, this work proposes a much simpler and straightforward learning process to improve dialogue policy performance which makes the reproduction of the results becomes much easier. Conditioned on different categories of test data, the results show that the proposed method could achieve competitive performances in the simple and hard cases, where each dialogue consisted of 1 domain and 3 domains, respectively. In case of dialogues containing 2 domains, the STORL was significantly better than GPT-TDAPT by around 10% in success rate and 5% in both complete rate and booking rate. This empirical evidence may suggest a future research for representation learning methods that can handle all possible cases.

Table 5: End-to-end system evaluation results on STORL and GPT-TDAPT [51], a model participating in DSTC9 track 2, under different number of test dialogues in presence of different number of domains. GPT-TDAPT results were obtained by using the provided models.

Configuration				Success		Inform		Complete	Booking	Average	Computation
NLU	DST	Policy	NLG	Rate (%)	Precision	Recall	F1	Rate (%)	Rate (%)	Turn (Succ/All)	Time (Minutes)
1 Domain (337 test dialogues)											
End-to-End (GPT-TDAPT (DSTC9 track 2))*				95.2	84.2	97.4	88.3	98.2	96.7	8.6/8.9	11:23
BERT	Rule	STORL	Template	94.7	87.9	96.9	90.6	97.0	97.8	7.2/7.5	1:30
2 domains (523 test dialogues)											
End-to-End (GPT-TDAPT (DSTC9 track 2))*				91.2	78.2	98.6	85.5	97.7	96.3	18.0/18.1	46:13
BERT	Rule	STORL	Template	80.1	81.2	93.5	85.1	91.2	91.6	13.2/14.0	4:30
3 domains (140 test dialogues)											
End-to-End (GPT-TDAPT (DSTC9 track 2))*				80.0	80.2	95.5	86.0	90.7	92.9	21.5/22.7	16:21
BERT	Rule	STORL	Template	77.9	83.3	91.5	85.1	88.6	87.1	17.2/17.4	1:30
Overall (1000 test dialogues)											
End-to-End (GPT-TDAPT (DSTC9 track 2))*				91.0	80.4	97.8	86.4	96.9	95.8	15.1/15.7	73:57
BERT	Rule	STORL	Template	84.7	83.7	94.3	86.9	92.8	92.1	11.5/12.3	7:30

In comparison with the teacher model, Table 6 shows the benefit of using the trained encoder through STORL. In the goal IDs 304 and 357, the STORL model generated the responses identically to the teacher model given clean input although BERT did not predict the first user dialogue act “[‘Inform’, ‘Train’, ‘none’, ‘none’]”. On the other hand, the teacher model, fed with noisy input from BERT, produced the irrelevant response by offering the train schedule in the goal ID 304, and failed to generate any response in goal ID 357. In the goal ID 623, even though BERT outputs a wrong dialogue act “[‘Inform’, ‘Hotel’, ‘Type’, ‘hotel’]”, the STORL model could provide a nearly similar response to the teacher model given the clean input while the teacher model given noisy input could not find any hotel for the user. All of the findings are in line with the distance metrics (in both ℓ_1 and ℓ_2) and the distributions of latent features due to the introduction of the STORL and the teacher model given the noisy input (Teacher) relative to the teacher model given the clean or ground truth of user dialogue acts (Teacher*) as shown by Table 7. It can be seen that the STORL sufficiently maps the features conditioned on the noisy input to be closer to the teacher features given by the clean input.

More empirical evidence about the benefit of using STORL is shown by Table 8 which illustrates a qualitative comparison of the conversations made by STORL policy and rule-based policy given the noisy input from BERTNLU. Due to the mispredicted slot in the second conversation turn, ‘Destination’ to ‘Departure’, the rule-based policy repeated the question which asked for the user destination until it obtained slot ‘Destination’. However, the rule-based policy could not find any schedule for the train departing to Peterborough due to the inability for handling the conflicted DST information where the DST stored two different values for ‘Departure’ which were Cambridge and Peterborough. Meanwhile, the proposed STORL agent could handle the mispredicted user dialogue acts due to the well-trained encoder which could map the noisy input into a better representation. Therefore, although BERT mispredicted the slot ‘Destination’ to ‘Departure’, STORL could provide an appropriate response by offering a train schedule to Peterborough.

Table 6: Qualitative illustration using STORL. Teacher* represents the output of teacher model given clean input. Red colored texts indicate either unpredicted or mispredicted user dialogue acts.

Goal ID: 304	
User:	I would just like to find a train first, and get the info. I need to leave after 14:15. I would like to leave on Friday.
User DA:	I would like to depart from Cambridge. [Inform , Train , none , none], [Inform, Train, Leave, '14:15'], [Inform, Train, Day, Friday], [Inform, Train, Departure, Cambridge]]
BERT Predicted DA:	[Inform, Train, Leave, '14:15'], [Inform, Train, Day, Friday], [Inform, Train, Departure, Cambridge]]
System (Teacher*):	What is your destination?
System (Teacher):	There is a train leaving at 15:00 would you like me to book this? What station would you like to arrive at?
System (STORL):	What is your destination?
Goal ID: 357	
User:	I would just like to find a train first, and get the info. I'm looking to travel on Thursday.
User DA:	Can you tell me when that train will be arriving in Norwich? I also need a train departing from Cambridge. [Inform , Train , none , none], [Inform, Train, Day, Thursday], [Inform, Train, Destination, Norwich], [Inform, Train, Departure, Cambridge]]
BERT Predicted DA:	[Inform, Train, Day, Thursday], [Inform, Train, Destination, Norwich], [Inform, Train, Departure, Cambridge]]
System (Teacher*):	What time do you need to arrive? Is there a certain time you are wanting to leave?
System (Teacher):	' ' (Note: Policy was unable to generate any response.)
System (STORL):	What time do you need to arrive? Is there a certain time you are wanting to leave?
Goal ID: 623	
User:	I need a hotel as well. I need one with Wifi. I'm looking for a place to stay in the centre.
User DA:	I would like it to have a 4 star rating [Inform , Hotel , none , none], [Inform, Hotel, Internet, yes], [Inform, Hotel, Area, centre], [Inform, Hotel, Stars, 4]]
BERT Predicted DA:	[Inform, Hotel, Internet, yes], [Inform , Hotel , Type , hotel], [Inform, Hotel, Area, centre], [Inform, Hotel, Stars, 4]]
System (Teacher*):	There are 2 of those. How about Alexander bed and breakfast?
System (Teacher):	I am sorry I have no listings for any with 4 stars. I'm sorry, there are no hotel that match your preferences.
System (STORL):	There are 1 of those. They have a 4 Star rating. University arms hotel looks like it would be a good choice.

Table 7: Comparison of the distance metrics (both ℓ_1 and ℓ_2) and the feature distributions between the latent features by using the teacher model given noisy input (Teacher) and the STORL relative to the teacher model given clean input (Teacher*).

Features	ℓ_1 Dist	ℓ_2 Dist	Distribution (mean \pm std)
Goal ID: 304			
Teacher-Teacher*	0.970	0.744	0.0084 \pm 0.081 - 0.0058 \pm 0.068
STORL-Teacher*	0.383	0.111	0.0066 \pm 0.064 - 0.0058 \pm 0.068
Goal ID: 357			
Teacher-Teacher*	1.200	0.332	0.0101 \pm 0.049 - 0.0120 \pm 0.055
STORL-Teacher*	0.602	0.115	0.0133 \pm 0.054 - 0.0120 \pm 0.055
Goal ID: 623			
Teacher-Teacher*	0.457	0.322	0.0058 \pm 0.064 - 0.0051 \pm 0.067
STORL-Teacher*	0.289	0.117	0.0054 \pm 0.061 - 0.0051 \pm 0.067

5 Conclusions

A novel method for improving the robustness of dialogue policy in addressing noisy input due to the error output from NLU component in the dialogue system, named as the student-teacher offline reinforcement learning (STORL) has been proposed. The proposed method was designed according to two strategies including student-teacher learning and offline reinforcement learning. Student-teacher learning aimed to force the student model to map the extracted features of the noisy input to be close to the clean features extracted by the teacher model. Meanwhile, the offline reinforcement learning which aimed to minimize the focal loss was used to train the dialogue policy to be able to provide appropriate response given the user input by only utilizing the observation-action pairs stored in the dataset. The experimental findings

Table 8: Example of the successful and failed dialogue generated by STORL and rule-based dialogue policy, respectively. **Red** colored texts indicate the mismatched user dialogue acts (DAs). **Blue** and **green** colored texts indicate the correct and wrong answers, respectively. The whole conversations are not shown to save space.

Goal:	{‘train’: {‘info’: {‘arrive by’: ‘11:45’, ‘day’: ‘Wednesday’, ‘departure’: ‘Cambridge’, ‘destination’: ‘Peterborough’}, ‘request’: {‘duration’: ‘?’, ‘price’: ‘?’}}, ‘restaurant’: {‘info’: {‘food’: ‘Indian’, ‘price range’: ‘expensive’}, ‘request’: {‘postcode’: ‘?’}}}
STORL Policy	
User:	Please book that train for me now. I want to get there by 11:45.
	The train should leave on Wednesday. I’m departing from Cambridge.
User DA:	[[‘Inform’, ‘Train’, ‘none’, ‘none’], [‘Inform’, ‘Train’, ‘Arrive’, ‘11:45’], [‘Inform’, ‘Train’, ‘Day’, ‘Wednesday’], [‘Inform’, ‘Train’, ‘Departure’, ‘Cambridge’]]
BERT Predicted DA:	[[‘Inform’, ‘Train’, ‘Arrive’, ‘11:45’], [‘Inform’, ‘Train’, ‘Day’, ‘Wednesday’], [‘Inform’, ‘Train’, ‘Departure’, ‘Cambridge’]]
System:	Where would you like to go to?
User:	I am leaving for Peterborough.
User DA:	[[‘Inform’, ‘Train’, ‘Destination’, ‘Peterborough’]]
BERT Predicted DA:	[[‘Inform’, ‘Train’, ‘Departure’, ‘Peterborough’]]
System:	Would you like me to book the 06:09 train?
User:	Can you tell me the price? Can you give me the travel time please?
User DA:	[[‘Request’, ‘Train’, ‘Ticket’, ‘?’], [‘Request’, ‘Train’, ‘Time’, ‘?’]]
BERT Predicted DA:	[[‘Request’, ‘Train’, ‘Time’, ‘?’], [‘Request’, ‘Train’, ‘Ticket’, ‘?’]]
System:	The price is 16.50 pounds per ticket. The trip is 50 minutes.
...	
Rule-Based Policy	
User:	Please book that train for me now. I want to get there by 11:45.
	The train should leave on Wednesday. I’m departing from Cambridge.
User DA:	[[‘Inform’, ‘Train’, ‘none’, ‘none’], [‘Inform’, ‘Train’, ‘Arrive’, ‘11:45’], [‘Inform’, ‘Train’, ‘Day’, ‘Wednesday’], [‘Inform’, ‘Train’, ‘Departure’, ‘Cambridge’]]
BERT Predicted DA:	[[‘Inform’, ‘Train’, ‘Arrive’, ‘11:45’], [‘Inform’, ‘Train’, ‘Day’, ‘Wednesday’], [‘Inform’, ‘Train’, ‘Departure’, ‘Cambridge’]]
System:	Where would you like to go to?
User:	I am leaving for Peterborough.
User DA:	[[‘Inform’, ‘Train’, ‘Destination’, ‘Peterborough’]]
BERT Predicted DA:	[[‘Inform’, ‘Train’, ‘Departure’, ‘Peterborough’]]
System:	Where is your destination?
User:	Can you also help me find a train going to Peterborough?
User DA:	[[‘Inform’, ‘Train’, ‘Destination’, ‘Peterborough’]]
BERT Predicted DA:	[[‘Inform’, ‘Train’, ‘Destination’, ‘Peterborough’]]
System:	I am sorry there are no trains to arrive at 11:45. There are no trains on Wednesday. There is no train leaving Peterborough.
...	

demonstrated that the proposed hybrid learning did enhance the performance of the pipeline dialogue system in the ConvLab-2 end-to-end system evaluation under MultiWOZ 2.1 dataset. Furthermore, the competitive results were also obtained when compared to the end-to-end approaches by using the pre-trained GPT-2 model with lower computational cost and simpler learning process. The source codes based on STORL can be accessed in <https://github.com/NYCU-MLLab/>. Future works will include the extension of this method to the other dialogue domain [31].

Acknowledgement

This work was supported in part by the National Science and Technology Council, Taiwan, under Contract NSTC 112-2634-F-A49-006.

Biographies

Mahdin Rohmatillah received his M.S. degree and Ph.D. degree from the National Sun Yat-sen University, Taiwan, and the National Yang Ming Chiao Tung University, in 2018 and 2024, respectively, both in electrical and computer engineering. His research interests include machine learning, large language models and dialogue systems.

Jen-Tzung Chien is currently the Lifetime Chair Professor with National Yang Ming Chiao Tung University. He has published extensively, including three books and more than 250 peer-reviewed articles, many on machine learning, deep learning and Bayesian learning with applications on natural language processing and computer vision. He was the recipient of the Best Paper Award in IEEE Workshop on Automatic Speech Recognition and Understanding in 2011 and IEEE International Workshop on Machine Learning for Signal Processing in 2023.

References

- [1] A. Bordes, Y.-L. Boureau, and J. Weston, “Learning End-to-End Goal-Oriented Dialog”, in *Proc. of International Conference on Learning Representations*, 2017.
- [2] 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 *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2018, 5016–26.
- [3] B. Byrne, K. Krishnamoorthi, C. Sankar, A. Neelakantan, B. Goodrich, D. Duckworth, S. Yavuz, A. Dubey, K.-Y. Kim, and A. Cedilnik, “Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset”, in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2019, 4516–25.
- [4] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, “Decision transformer: Reinforcement learning via sequence modeling”, *Advances in Neural Information Processing Systems*, 34, 2021, 15084–97.
- [5] W. Chen, J. Chen, P. Qin, X. Yan, and W. Y. Wang, “Semantically Conditioned Dialog Response Generation via Hierarchical Disentangled Self-Attention”, in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2019, 3696–709.
- [6] J.-T. Chien and P.-C. Hsu, “Stochastic Curiosity Exploration for Dialogue Systems”, in *Proc. of Annual Conference of International Speech Communication Association*, 2020, 3885–9.

- [7] J.-T. Chien and W. Lai, “Variational Skill Embeddings for Meta Reinforcement Learning”, in *Proc. of International Joint Conference on Neural Networks*, 2023, 1–8.
- [8] J.-T. Chien and W. X. Lieow, “Meta Learning for Hyperparameter Optimization in Dialogue System.”, in *Proc. of Annual Conference of International Speech Communication Association*, 2019, 839–43.
- [9] C.-T. Chu, M. Rohmatillah, C.-H. Lee, and J.-T. Chien, “Augmentation Strategy Optimization for Language Understanding”, in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, 7952–6.
- [10] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Cambridge University Press, 2011.
- [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in *Proc. of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, 4171–86.
- [12] M. Eric, R. Goel, S. Paul, A. Sethi, S. Agarwal, S. Gao, and D. Hakkani-Tur, “MultiWOZ 2.1: Multi-Domain Dialogue State Corrections and State Tracking Baselines”, *arXiv arXiv:1907.01669*, 2019.
- [13] J. Fu, K. Luo, and S. Levine, “Learning Robust Rewards with Adversarial Inverse Reinforcement Learning”, in *Proc. of International Conference on Learning Representations*, 2018.
- [14] M. Henderson, B. Thomson, and S. Young, “Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation”, in *Proc. of IEEE Spoken Language Technology Workshop*, 2014, 360–5.
- [15] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network”, *arXiv preprint arXiv:1503.02531*, 2015.
- [16] C.-E. Hsu, M. Rohmatillah, and J.-T. Chien, “Multitask generative adversarial imitation learning for multi-domain dialogue system”, in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, 954–61.
- [17] Y. Jang, J. Lee, and K.-E. Kim, “GPT-critic: Offline Reinforcement Learning for End-to-End Task-Oriented Dialogue System”, in *Proc. of International Conference on Learning Representations*, 2022.
- [18] X. Jin, W. Lei, Z. Ren, H. Chen, S. Liang, Y. Zhao, and D. Yin, “Explicit State Tracking with Semi-Supervision for Neural Dialogue Generation”, in *Proc. of International Conference on Information and Knowledge Management*, 2018, 1403–12.
- [19] L. Ke, S. Choudhury, M. Barnes, W. Sun, G. Lee, and S. Srinivasa, “Imitation learning as f-divergence minimization”, in *International Workshop on the Algorithmic Foundations of Robotics*, 2020, 313–29.

- [20] Y. Kim and A. M. Rush, “Sequence-Level Knowledge Distillation”, in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2016, 1317–27.
- [21] J. Kulhánek, V. Hudecek, T. Nekvinda, and O. Dušek, “AuGPT: Dialogue with Pre-trained Language Models and Data Augmentation”, *arXiv preprint arXiv:2102.05126*, 2021.
- [22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal Loss for Dense Object Detection”, in *Proc. of IEEE International Conference on Computer Vision*, 2017, 2980–8.
- [23] Z. Lin, A. Madotto, G. I. Winata, and P. Fung, “MinTL: Minimalist Transfer Learning for Task-Oriented Dialogue Systems”, in *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2020, 3391–405.
- [24] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization”, in *Proc. of International Conference on Learning Representations*, 2019.
- [25] T.-C. Luo and J.-T. Chien, “Variational Dialogue Generation with Normalizing Flows”, in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, 7778–82.
- [26] Y. Luo, H. Xu, and T. Ma, “Learning Self-Correctable Policies and Value Functions from Demonstrations with Negative Sampling”, in *Proc. of International Conference on Learning Representations*, 2020.
- [27] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, “Human-level control through deep reinforcement learning”, *Nature*, 2015.
- [28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners”, *OpenAI Blog*, 1(8), 2019, 9.
- [29] A. Rastogi, X. Zang, S. Sunkara, R. Gupta, and P. Khaitan, “Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset”, in *Proc. of AAAI Conference on Artificial Intelligence*, Vol. 34, No. 5, 2020, 8689–96.
- [30] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, “Asymmetric Loss For Multi-Label Classification”, in *Proc. of IEEE/CVF International Conference on Computer Vision*, 2021, 82–91.
- [31] M. Rohmatillah, B. Aditya, L.-J. Yang, B. Ngo, W. Sulaiman, and J.-T. Chien, “Promoting mental self-disclosure in a spoken dialogue system”, in *Proc. of Annual Conference of International Speech Communication Association*, 2023, 670–1.

- [32] M. Rohmatillah, J.-T. Chien, *et al.*, “Advances and Challenges in Multi-Domain Task-Oriented Dialogue Policy Optimization”, *APSIPA Transactions on Signal and Information Processing*, 12(e37), 2023, 1–52.
- [33] M. Rohmatillah and J.-T. Chien, “Causal Confusion Reduction for Robust Multi-Domain Dialogue Policy”, in *Proc. of Annual Conference of the International Speech Communication Association*, 2021, 3221–5.
- [34] M. Rohmatillah and J.-T. Chien, “Corrective Guidance and Learning for Dialogue Management”, in *Proc. of ACM International Conference on Information & Knowledge Management*, 2021, 1548–57.
- [35] M. Rohmatillah and J.-T. Chien, “Hierarchical Reinforcement Learning With Guidance for Multi-Domain Dialogue Policy”, *IEEE Transactions on Audio, Speech, and Language Processing*, 31, 2023, 748–61.
- [36] M. Rohmatillah and J.-T. Chien, “Revise the NLU: A Prompting Strategy for Robust Dialogue System”, in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, 10956–60.
- [37] S. Ross, G. Gordon, and D. Bagnell, “A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning”, in *Proc. of the International Conference on Artificial Intelligence and Statistics*, 2011, 627–35.
- [38] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”, *arXiv preprint arXiv:1910.01108*, 2019.
- [39] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms”, *arXiv preprint arXiv:1707.06347*, 2017.
- [40] 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 *Proc. of Conference of the North American Chapter of the Association for Computational Linguistics*, 2018, 41–51.
- [41] R. Takanobu, R. Liang, and M. Huang, “Multi-Agent Task-Oriented Dialog Policy Learning with Role-Aware Reward Decomposition”, in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2020, 625–38.
- [42] R. Takanobu, H. Zhu, and M. Huang, “Guided Dialog Policy Learning: Reward Estimation for Multi-Domain Task-Oriented Dialog”, in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2019, 100–10.
- [43] R. Takanobu, Q. Zhu, J. Li, B. Peng, J. Gao, and M. Huang, “Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation”, in *Proc. of Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2020, 297–310.

- [44] S. Ultes, L. M. Rojas-Barahona, P.-H. Su, D. Vandyke, D. Kim, I. Casanueva, P. Budzianowski, N. Mrkšić, T.-H. Wen, M. Gašić, and S. Young, “PyDial: A Multi-domain Statistical Dialogue System Toolkit”, in *Proc. of Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2017, 73–8.
- [45] Z. Wang, A. Novikov, K. Zolna, J. S. Merel, J. T. Springenberg, S. E. Reed, B. Shahriari, N. Siegel, C. Gulcehre, N. Heess, and N. de Freitas, “Critic Regularized Regression”, in *Advances in Neural Information Processing Systems*, 2020, 7768–78.
- [46] C. Wen, J. Lin, T. Darrell, D. Jayaraman, and Y. Gao, “Fighting Copycat Agents in Behavioral Cloning from Observation Histories”, in *Advances in Neural Information Processing Systems*, Vol. 33, 2020, 2564–75.
- [47] T.-H. Wen, M. Gašić, N. Mrkšić, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, D. Vandyke, and S. Young, “Conditional Generation and Snapshot Learning in Neural Dialogue Systems”, in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2016, 2153–62.
- [48] R. J. Williams, “Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning”, *Machine Learning*, 8(3-4), 1992, 229–56.
- [49] Y. Yang, Y. Li, and X. Quan, “UBAR: Towards Fully End-to-End Task-Oriented Dialog System with GPT-2”, in *Proc. of AAAI Conference on Artificial Intelligence*, Vol. 35, No. 16, 2021, 14230–8.
- [50] S. Yeung, V. Ramanathan, O. Russakovsky, L. Shen, G. Mori, and L. Fei-Fei, “Learning to learn from noisy web videos”, in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 5154–62.
- [51] B. Zhang, Y. Lyu, N. Ding, T. Shen, Z. Jia, K. Han, and K. Knight, “A Hybrid Task-Oriented Dialog System with Domain and Task Adaptive Pretraining”, *arXiv preprint arXiv:2102.04506*, 2021.
- [52] Y. Zhang, Z. Ou, and Z. Yu, “Task-Oriented Dialog Systems That Consider Multiple Appropriate Responses under the Same Context”, in *Proc. of AAAI Conference on Artificial Intelligence*, Vol. 34, No. 5, 2020, 9604–11.
- [53] T. Zhao, K. Xie, and M. Eskenazi, “Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models”, in *Proc. of Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, 1208–18.
- [54] Q. Zhu, Z. Zhang, Y. Fang, X. Li, R. Takanobu, J. Li, B. Peng, J. Gao, X. Zhu, and M. Huang, “ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems”, in *Proc. of Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, 142–9.