Semi-Supervised Dialogue Policy Learning via Stochastic Reward Estimation

Xinting Huang,[1] Jianzhong Qi,[1] Yu Sun,[2] Rui Zhang[1]*
[1]The University of Melbourne, [2]Twitter Inc.
{xintingh@student., jianzhong.qi@, rui.zhang@}unimelb.edu.au, ysun@twitter.com

Abstract

Dialogue policy optimization often obtains feedback until task completion in task-oriented dialogue systems. This is insufficient for training intermediate dialogue turns since supervision signals (or *rewards*) are only provided at the end of dialogues. To address this issue, reward learning has been introduced to learn from state-action pairs of an optimal policy to provide turn-by-turn rewards. This approach requires complete state-action annotations of human-to-human dialogues (i.e., expert demonstrations), which is labor intensive. To overcome this limitation, we propose a novel reward learning approach for semi-supervised policy learning. The proposed approach learns a dynamics model as the reward function which models dialogue progress (i.e., state-action sequences) based on expert demonstrations, either with or without annotations. The dynamics model computes rewards by predicting whether the dialogue progress is consistent with expert demonstrations. We further propose to learn action embeddings for a better generalization of the reward function. The proposed approach outperforms competitive policy learning baselines on MultiWOZ, a benchmark multi-domain dataset.

1 Introduction

Task-oriented dialogue systems complete tasks for users, such as making a restaurant reservation or finding attractions to visit, in multi-turn dialogues (Gao et al., 2018; Sun et al., 2016, 2017). Dialogue policy is a critical component in both the conventional pipeline approach (Young et al., 2013) and recent end-to-end approaches (Zhao et al., 2019). It decides the next action that a dialogue system should take at each turn. Considering its nature of sequential decision making, dialogue policy is usually learned via reinforcement learning (Su et al., 2015; Peng et al., 2018; Zhang et al., 2019). Specifically, dialogue policy is learned by maximizing accumulated rewards over interactions with an environment (i.e., actual users or a user simulator). Handcrafted rewards are commonly used for policy learning in earlier work (Peng et al., 2018), which assigns a small negative penalty at each turn and a large positive/negative reward when the task is successful/failed. However, such reward setting does not provide sufficient supervision signals in each turn other than the last turn, which causes the sparse reward issues and may result in poorly learned policies (Takanobu et al., 2019).

Table 1: State Action Annotation and Utterance Example

User Side	**Utterance**
	I would like moderate price range please.
	Dialogue State annotation
	`Restaurant: {food=modern european, price range=moderate}`
System Side	**Utterance**
	I found de luca cucina and riverside brasserie. does either of them sound good for you?
	System action annotation
	`restaurant-inform:{name=de luca cucina, name=riverside brasserie}`

To address this problem, reward function learning that relies on *expert demonstrations* has been introduced (Takanobu et al., 2019; Li et al., 2019b). Specifically, state-action sequences generated by an optimal policy (i.e., expert demonstrations) are collected, and a reward function is learned to give high rewards to state-action pairs that better resemble the behaviors of the optimal policy. In this way, turn-by-turn rewards estimated by the reward function can be provided to learn dialogue policy. Obtaining expert demonstrations is critical to reward function learning. Since it is impractical to assume that an optimal policy is always available,

*Rui Zhang is the corresponding author.

a common and reasonable approach is to treat the decision makings in human-human dialogues as optimal behaviors. To accommodate the learning of reward function, human-human dialogues need to be annotated in the form of *state-action pairs* from textual utterances. Table 1 illustrates an example of human-human dialogue and its state-action annotation. However, obtaining such annotations require extensive efforts and costs. Besides, a reward function based on state-action pair might cause an unstable policy learning, especially with a limited amount of annotated dialogues (Yang et al., 2018).

To address the above issues, we propose to learn dialogue policies in a semi-supervised setting where the system action of expert demonstrations only need to be partially annotated. We propose to use an implicitly trained *stochastic dynamics model* as the reward function to replace the conventional reward function that is restricted to state-action pairs. Dynamics models describe sequential progress using a combination of stochastic and deterministic states in a latent space, which promotes an effective tracking and forecasting (Minderer et al., 2019; Sun et al., 2019; Wang et al., 2019a). In our scenario, we train the dynamics model to describe dialogue progress of expert demonstrations. The main rationale is that the reward function should give high rewards to actions that lead to dialogue progress similar to those in expert demonstrations. This is because dialogue progress at the early stage highly influences subsequent progress, and the latter directly determines whether the task can be completed. Since the learning of dynamics model maps observations to latent states and further reason over the latent states, we are no longer restricted to fully annotated dialogues. Using dynamics model as reward function also promotes a more stable policy learning.

Learning the dynamics model in the text space is, however, prone to compounding errors due to complexities and diversities of languages. We tackle this challenge by learning the dynamics model in an *action embedding* space that encodes the effect of system utterances on dialogue progress. We achieve action embedding learning by incorporating an embedding function into a generative models framework for semi-supervised learning (Kingma et al., 2014). We observe that system utterances with comparable effects on dialogue progress will lead to similar state transitions (Huang et al., 2019a). Therefore, we formulate the generative model to describe the state transition process. Using the generative model, we enrich the expert dialogues (either fully or partially annotated) with action embedding to learn the dynamics model. Moreover, we also consider the scenarios where both state and action annotations are absent in most expert dialogues, referred to as unlabeled dialogues. To expand the proposed approach to such scenarios, we further propose to model dialogue progress using action sequences and reformulate the generative model accordingly.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to approach semi-supervised dialogue policy learning.
- We propose a novel reward estimation approach to dialogue policy learning which relives the requirements of extensive annotations and promotes a stable learning of dialogue policy.
- We propose an action embedding learning technique to effectively train the reward estimator from either partially labeled or unlabeled dialogues.
- We conduct extensive experiments on the benchmark multi-domain dataset. Results show that our approach consistently outperforms strong baselines coupled with semi-supervised learning techniques.

2 Preliminaries

For task-oriented dialogues, a dialogue policy $\pi(a|s)$ decides an action $a \in \mathcal{A}$ based on the dialogue state $s \in \mathcal{S}$ at each turn, where $\mathcal{A}$ and $\mathcal{S}$ are the predefined sets of all actions and states, respectively. Reinforcement learning is commonly applied to dialogue policy learning, where the dialogue policy model is trained to maximize accumulative rewards through interactions with environments (i.e., users):

$$\mathcal{L} = -\mathbb{E}_{\tau_i \sim \pi}[r(\tau)] = -\mathbb{E}_{\tau_i \sim \pi}[\sum_t r(s_t, a_t)] \quad (1)$$

where $\tau_i = \{(s_t, a_t)|0 \leq t \leq n_\tau\}$ represents a sampled dialogue, and $r(\tau_i)$ is the numerical rewards obtained in this dialogue. Instead of determining $r(\tau_i)$ via heuristics, recent reward learning approaches train a reward function r_θ to assign numerical rewards for each state-action pair. The reward function is learned from expert demonstrations D_{demo} that are dialogues sampled from an optimal policy in the form of state-action pairs. Adversarial learning is usually adopted to enforces higher rewards to state-action pairs from expert demonstrations and lower rewards to those sam-

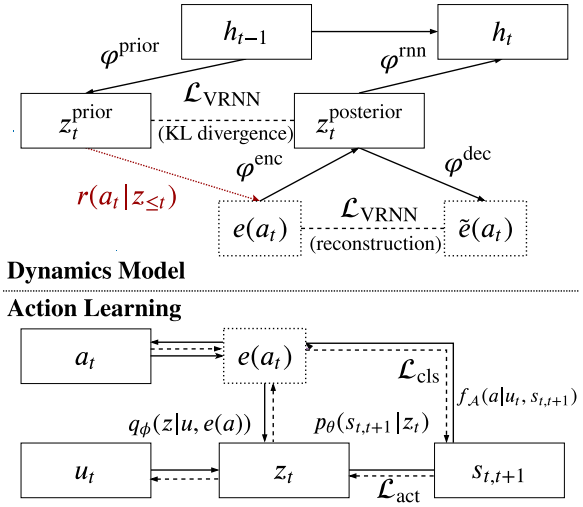


Figure 1: Overall framework of the proposed approach

pled from the learning policy (Fu et al., 2017):

$$\mathcal{L} = -\mathbb{E}_{\tau_j \sim D_{demo}}[r_\theta(\tau_j)] + \log \mathbb{E}_{\tau_i \sim \pi}\left(\frac{\exp r_\theta(\tau_i)}{q(\tau_i)}\right) \tag{2}$$

where π is the current dialogue policy, and q is the distribution of dialogues generated with π. In this way, the dialogue policy and reward function are iteratively optimized, which requires great training efforts and might lead to unstable learning results (Yang et al., 2018). Moreover, such a reward learning approach requires a complete dialogue state and system action annotation of expert demonstrations, which are expensive to obtain.

3 Proposed Model

3.1 Overview

We study the problem of *semi-supervised* dialogue policy learning. Specifically, we consider the setting that expert demonstrations D_{demo} consist of a small number of fully labeled dialogues $D_{\mathcal{F}}$ and partially labeled dialogues $D_{\mathcal{P}}$. For each fully annotated dialogue τ_i in $D_{\mathcal{F}}$, complete annotations are available: $\tau_i = \{(s_t, a_t, u_t)|1 \leq t \leq n_\tau\}$, where u_t is the system utterance at turn t. Meanwhile, each partially labeled dialogue τ_j in $D_{\mathcal{P}}$ only has state annotations and system utterances: $\tau_j = \{(s_t, u_t)|1 \leq t \leq n_\tau\}$.

Figure 1 illustrates the overall framework of the proposed approach. Rewards are estimated by a dynamics model that consumes action embeddings $e(a_t)$. Every action in the set $\mathcal{A}$ is mapped to a fix-length embedding via a learnable embedding function f_E. To obtain the action embeddings for $D_{\mathcal{P}}$ which has no action annotations, we first predict the action via a prediction model $f_{\mathcal{A}}$ and then transform the predicted actions to embeddings. To obtain effective action embeddings, we design a state-transition based objective to jointly optimize f_E and $f_{\mathcal{A}}$ via variational inference (Sec. 3.2). After obtaining the action embeddings, the dynamics model is learned by fitting the expert demonstrations enriched by action embeddings. Rewards are then estimated as the conditional probability of the action given the current dialogue progress encoded in latent states (Sec. 3.3). We also extend the above approach to unlabeled dialogues where both state and action annotations are absent (Sec. 3.4).

3.2 Action Learning via Generative Models

We aim to learn the prediction model $f_{\mathcal{A}}$ and action embeddings using both $D_{\mathcal{F}}$ and $D_{\mathcal{P}}$. We formulate the action prediction model as $f_{\mathcal{A}}(a|u_t, s_t, s_{t+1})$ which takes as input the system utterance u_t and its corresponding state transition (s_t, s_{t+1}). We then introduce an mapping function: $f_E : \mathcal{A} \rightarrow \mathcal{E}$, where $\mathcal{E} \subseteq \mathbb{R}^d$ is the action embedding space later used for learning the dynamics model.

We train the prediction model by proposing a variational inference approach based on a semi-supervised variational autoencoder (Semi-VAE) (Kingma et al., 2014). Semi-VAE describes the data generation process of feature-label pairs $\{(x_i, y_i)|1 \leq i \leq N\}$ via latent variables z as:

$$\log p(x) = \log \sum_y \int_z p_\theta(x, z, y) dz \tag{3}$$

where p_θ is a generative model parameterised by θ, and the class label y is treated as a latent variables for unlabeled data. Since this log-likelihood in Eqn. 3 is intractable, its variational lower bound for unlabeled data is instead optimized as:

$$\begin{aligned} \log p(x) &\geq \mathbb{E}_{q_{\phi,\psi}(y,z|x)}[\log \frac{p_\theta(x, z, y)}{q_{\phi,\psi}(y, z|x)}] \\ &= \mathbb{E}_{q_\psi(y|x)}[\mathcal{L}(x, y)] - \mathcal{H}(q_\psi(y|x)) = \mathcal{U}(x) \end{aligned} \tag{4}$$

where $q_\phi(z|x, y)$ and $q_\psi(y|x)$ are inference models for latent variable z and y respectively, which have a factorised form $q_{\phi,\psi}(y, z|x) = q_\phi(z|x, y)q_\psi(y|x)$; $\mathcal{H}(\cdot)$ denotes causal entropy; $\mathcal{L}(x, y)$ is the variational bound for labeled data, ans is formulated as:

$$\begin{aligned} \mathcal{L}(x, y) = \mathbb{E}_{q_\phi(z|x,y)}[p_\theta(x|z, y)] + \log p(y) \\ -\text{KL}(q_\phi(z|x, y)||p(z)) \end{aligned} \tag{5}$$

where KL is the Kullback-Leibler divergence, and $p(y)$, $p(z)$ are the prior distribution of y, z.

The generative model p_θ, inference model q_ϕ and q_ψ are optimized using both the labeled subset p_l and unlabeled subset p_u using the objective as:

$$\mathcal{L} = \sum_{(x,y)\sim p_l} \mathcal{L}(x,y) + \sum_{x\sim p_u} \mathcal{U}(x) \tag{6}$$

Semi-Supervised Action Prediction

We now describe the learning of action prediction model $f_\mathcal{A}$ using semi-supervised expert demonstrations. We extend the semi-supervised VAE by modeling the generation process of *state transitions*. State transition information is indicative for action prediction and is available in both fully and partially labeled demonstrations. Thus we choose to describe the generation process of state transitions, and the optimization objective is formulated as:

$$\begin{aligned} \log p_\theta(s_{t+1}, s_t) &= \log \sum_a \int p_\theta(s_{t+1}, z, s_t, a) dz \\ &= \log \sum_a \int p(s_{t+1}, s_t|, z, a) p(z) p(a) dz \end{aligned} \tag{7}$$

For partially labeled dialogues, we treat action labels as latent variables and use the action prediction model $f_\mathcal{A}(a|u_t, s_t, s_{t+1})$ to infer the value (which is denoted as $f_\mathcal{A}(a|\cdot)$ later for simplicity). The variational bound of Eqn. 7 is derived as:

$$\begin{aligned} \mathcal{U}(s_{t+1}, s_t) = \mathbb{E}_{f_\mathcal{A}(a|\cdot)}[\mathcal{L}(s_{t+1}, s_t, a)] \\ -\mathcal{H}(f_\mathcal{A}(a|\cdot)) \end{aligned} \tag{8}$$

where $\mathcal{L}(s_{t+1}, s_t, a_t)$ is the variational bound for demonstrations with action labels and is derived as:

$$\begin{aligned} \mathcal{L}(s_{t+1}, s_t, a) = \mathbb{E}_{q_\phi(z|u_t,a)}[p_\theta(s_{t+1}|s_t, z)] \\ -\mathrm{KL}(q_\phi(z|u_t, a)||p(z)) \end{aligned} \tag{9}$$

where $q_\phi(z|u_t, a)$ is the inference model for latent variable z. Lastly, we use fully annotated samples to form a classification loss:

$$\mathcal{L}_{\text{cls}} = \mathbb{E}_{\tau_i \in D_\mathcal{F}}[\log f_\mathcal{A}(a|u_t, s_t, s_{t+1})] \tag{10}$$

The overall objectives includes the loss of fully and partially labeled demonstrations:

$$\begin{aligned} \mathcal{L}_{\text{act}} = \sum_{\tau_i \in D_\mathcal{F}} \mathcal{L}(s_{t+1}, s_t, a) + \\ \sum_{\tau_i \in D_\mathcal{P}} \mathcal{U}_(s_{t+1}, s_t) + \mathcal{L}_{\text{cls}} \end{aligned} \tag{11}$$

Action Embeddings Learning

We then incorporate action embedding function f_E into the developed semi-supervised action prediction approach. The reason to introduce action embeddings is to make the learning of reward estimator more efficient and robust. Specifically, prediction error of the action prediction model might impinge the learning of reward estimator, especially for our semi-supervised scenarios where fully labeled dialogues are limited. By mapping actions to an embedding space, 'wrongly predicted' partially labeled demonstrations can still provide sufficient knowledge and thus we could achieve better generalization over actions for reward estimation.

To this aim, we consider the inference steps in the semi-supervised learning process and utilize the ones that involve action labels, i.e., the inference models for latent variables z and a. We first specify how the action prediction model is modified to include action embeddings. Inspired by (Chandak et al., 2019), we model the action selection using Boltzmann distribution for stability during training:

$$\begin{aligned} f_\mathcal{A}(a|u_t, s_t, s_{t+1}) &= \frac{e^{z_a/\gamma}}{\sum_{a' \in \mathcal{A}} e^{z_{a'}/\gamma}} \\ z_a &= e(a)^\top g(u_t, s_t, s_{t+1}), e(a) = f_E(a) \end{aligned} \tag{12}$$

where γ is a temperature parameter, and $g(\cdot)$ is a function that maps the input into hidden states of the same dimension as action embeddings. We also modify the inference model for latent variable by incorporating action embeddings:

$$q_\phi(z|u_t, a) = q_\phi(z|u_t, e(a)) \tag{13}$$

After optimizing the action prediction model $f_\mathcal{A}$ and action embedding function f_E jointly using the objective function Eqn. 11, we use action embeddings to enrich the expert demonstrations. For fully labeled dialogues, we map the given system action labels to corresponding embeddings and obtain $\tau_i = \{(s_t, e(a_t))|1 \le t \le n_\tau\}$. For partially labeled dialogues, we first infer the action using prediction model: $\tilde{a}_t = f_\mathcal{A}(u_t, s_t, s_{t+1})$, and map the inferred action to its embedding to obtain: $\tau_j = \{(s_t, e(\tilde{a}_t))|1 \le t \le n_\tau\}$.

3.3 Reward Estimation by Dynamics Model

We aim to learn a reward estimator based on action representations obtained from the action learning module. To achieve a more stable reward estimation than adversarial reward learning, we propose

a reward estimator based on *dialogues progress*. Dialogue progress describes how user goals are achieved through multistep interactions and can be modeled as dialogue state transitions. We argue that an action should be given higher rewards when it leads to similar dialogue progress (i.e., state transitions) of expert demonstrations. To this aim, we learn a model to explicitly model dialogue progress without the negative sampling required by adversarial learning, and rewards can be estimated as the local-probabilities assigned to the taken actions.

To model dialogue progress, we use variational recurrent neural network (VRNN) (Chung et al., 2015). The reason to use a stochastic dynamics model is due to the 'one-to-many' nature of task-oriented dialogues. Specifically, both user and dialogue system have multiple feasible options to proceed the dialogues which requires the modeling of uncertainty. Thus, by adding latent random variables to an RNN architecture, VRNN can provide better modeling of dialogue progress than deterministic dialogue state tracking.

VRNN has three types of variables: the observations (and here we consider action embeddings), the stochastic state z, and the deterministic hidden state h, which summarizes previous stochastic states $z_{\leq t}$, and previous observations $a_{\leq t}$. We formulate the prior stochastic states to be conditioned on previous timesteps through hidden state h_{t-1}:

$$p(z_t|a_{<t}, z_{<t}) = \varphi^{\text{prior}}(h_{t-1}) \tag{14}$$

We obtain posterior stochastic states by incorporating the observation at the current step, i,e. action embeddings $e(a_t)$:

$$q(z_t|a_{\leq t}, z_{<t}) = \varphi^{\text{enc}}(h_{t-1}, e(a_t)) \tag{15}$$

Predictions are made by decoding latent states, including both the stochastic and deterministic:

$$p(e(a_t)|z_{\leq t}, a_{<t}) = \varphi^{\text{dec}}(z_t, h_{t-1}, s_t) \tag{16}$$

And lastly the deterministic states are updated as:

$$h_t = \varphi^{\text{rnn}}(e(a_t), z_t, h_{t-1}, s_t) \tag{17}$$

where φ are all implemented as neural networks. Note that we also make the prediction and recurrence step to condition on the dialogue state s_t to provide more information.

We train the VRNN by optimizing the evidence lower bound (ELBO) as:

$$\begin{aligned}\mathcal{L}_{\text{VRNN}} = \mathbb{E}_{q(z_t|a_{\leq t}, z_{<t})}\big[\sum_t \log p(e(a_t)|z_{\leq t}, a_{<t}) \\ - \text{KL}(q(z_t|a_{\leq t}, z_{<t})||p(z_t|a_{<t}, z_{<t}))\big]\end{aligned} \tag{18}$$

The rewards are estimated as the conditional probability given the hidden state of VRNN, which encodes the current dialogue progress:

$$r(s_{\leq t}, a_t) = \log p_{\varphi^{\text{dec}}}(a_t|a_{<t}, s_{\leq t}) \tag{19}$$

where $p_{\varphi^{\text{dec}}}$ is the probability given to the selected action based on the decoding step of VRNN (Eqn. 16). The larger this conditional probability is, the more similar the dialogue progress this action leads to imitates the expert demonstrations. The proposed reward estimation is agnostic to the choice of policy, and various approaches (e.g., Deep Q-learning, Actor-Critic) can be optimized by plugging into the policy learning objective (Eqn. 1).

3.4 Expanding to Unlabeled Corpus

We further describe how to expand the proposed model, including action learning and reward estimation modules, to utilize *unlabeled expert demonstrations*. Formally, we consider the setting that we have fully labeled dialogues $D_{\mathcal{F}}$ and unlabeled dialogues $D_{\mathcal{U}}$. For each dialogue in $D_{\mathcal{U}}$, only textual conversations are provided and neither of state and action labels are available: $\tau_j = \{(c_t, u_t)|1 \leq t \leq n_\tau\}$, where c_t is the context and consists of the dialogue history of both user and system utterances.

With the absence of dialogue state information, we formulate the action prediction model as $f_{\mathcal{A}}(a|u_t, u_{t-1}, u_{t+1})$. This formulation can be considered as an application of Skip-Thought (Kiros et al., 2015), which originally utilizes contextual sentences as supervision signals. In our scenarios, we instead utilize the previous and next system utterances to provide more indicative information for action prediction.

We also build the joint learning of action prediction model the action embeddings on semi-supervised VAE framework. Instead of modeling state transitions, we choose the process of *response generation* to fully utilize unlabeled dialogues:

$$\begin{aligned}\log p_\theta(u_t) &= \log \sum_a \int p_\theta(u_t, z, a) dz \\ &= \log \sum_a \int p_\theta(u_t|z, a_t) p(z) p(a) dz\end{aligned} \tag{20}$$

System action labels are treated as latent variables for unlabeled dialogues, and the variational bond is derived as:

$$\mathcal{U}(u_t) = \mathbb{E}_{f_{\mathcal{A}}(a|\cdot)}[\mathcal{L}(u_t, a)] - \mathcal{H}(f_{\mathcal{A}}(a|\cdot)) \quad (21)$$

where $\mathcal{L}(u_t, a)$ is variational bound for fully labeled dialogues:

$$\mathcal{L}(u_t, a) = \mathbb{E}_{q_\phi(z|u_t,a)}[p_\theta(u_t|z, u_{t-1}, u_{t+1})] - \mathrm{KL}(q_\phi(z|a, u_t)||p(z)) \quad (22)$$

The objective to jointly train the prediction model and action embeddings is the same as Eqn. 11, where the terms for fully and partially labeled dialogues are replaced with the ones in Eqn. 22 and 21, respectively. Such expanding also enables a sufficient semi-supervised learning when expert demonstrations include all types of labeled dialogues: $D_{\mathcal{F}}$, $D_{\mathcal{P}}$ and $D_{\mathcal{U}}$. We notice that the posterior approximation $q_\phi(z|u_t, a)$ and action embedding function f_E can be sharing between the process of state transitions and response generation. Thus, by treating semi-supervised learning in $D_{\mathcal{F}}$ and $D_{\mathcal{P}}$ as auxiliary constraints, the learning over unlabeled corpus can also benefit from dialogues state information.

4 Experiments

To show the effectiveness of the proposed model (denoted as **Act-VRNN**), we experiment on a multi-domain dialogue environment under semi-supervised setting (Sec. 4.1). We compare against state-of-the-art approaches, and their variants enhanced by semi-supervised learning techniques (Sec. 4.2). We analyze the effectiveness of action learning and reward estimation of Act-VRNN under different supervision ratios (Sec. 4.3).

4.1 Settings

We use MultiWOZ (Budzianowski et al., 2018), a multi-domain human-human conversational dataset in our experiments. It contains in total 8438 dialogues spanning over seven domains, and each dialogue has 13.7 turns on average. MultiWOZ also contains a larger dialogue state and action space compared to former datasets such as movie-ticket booking dialogues (Li et al., 2017), and thus it is a much more challenging environment for policy learning. To use MultiWOZ for policy learning, a user simulator that initializes a user goal at the beginning and interacts with dialogue policy is required. For a fair comparison, we adopt the same procedure as Takanobu et al. (2019) to train the user simulator based on auxiliary user action annotations provided by ConvLab (Lee et al., 2019).

To simulate semi-supervised policy learning, we remove system action and dialogue states annotations to obtain partially labeled and unlabeled expert demonstrations, respectively. Fully labeled expert demonstrations are randomly sampled from all training dialogues with different ratios (5%, 10%, and 15% in our experiments). Note that the absence of action or state annotations only applies for expert demonstrations, while interactions between policy and user simulator are in dialogue-act level as (Takanobu et al., 2019) and not affected by semi-supervised setting.

We use a three-layer transformer (Vaswani et al., 2017) with a hidden size of 128 and 4 heads as our base model for action embedding learning, i.e., $g(\cdot)$ in Eqn. 12. We use grid search to find the best hyperparameters for the models. We choose the action embedding dimensionality among {50, 75, 100, 150, 200}, the stochastic latent state size in VRNN among {16, 32, 64, 128, 256}, and the deterministic latent state size among {25, 50, 75, 100, 150}.

We use **Entity-F1** and **Success Rate** to evaluate dialogue task completion. Entity-F1 computes the F1 score based on whether the requested information and indicated constraints from users are satisfied. Compared to inform rate and match rate used by Budzianowski et al. (2018), Entity-F1 considers both informed and requested entities at the same time and balances the recall and precision. Success rate indicates the ratio of successful dialogues, where a dialogue is regarded as successful only if all informed and requested entities are matched of the dialogue. We use **Turns** to evaluate the cost for task completion, where a lower number indicates the policy performs tasks more efficiently.

We compare Act-VRNN with three policy learning baselines: (1) **PPO** (Schulman et al., 2017) using hand-crafted rewards setting; (2) **ALDM**(Liu and Lane, 2018); (3) **GDPL** (Takanobu et al., 2019); We further consider using semi-supervised techniques to enhance the baselines under semi-supervised setting, and denote them as **SS-PPO**, **SS-ALDM**, and **SS-GDPL**. Specifically, we first train a prediction model based on semi-supervised VAE (Kingma et al., 2014), and use the predic-

Table 2: Semi-Supervised Policy Learning Results ($D_{\mathcal{F}}$ and $D_{\mathcal{P}}$)

Model		$D_{\mathcal{F}}$(5%) + $D_{\mathcal{P}}$(95%)			$D_{\mathcal{F}}$(10%) + $D_{\mathcal{P}}$(90%)			$D_{\mathcal{F}}$(20%) + $D_{\mathcal{P}}$(80%)		
		Entity-F1	Success	Turns	Entity-F1	Success	Turns	Entity-F1	Success	Turns
Handcrafted	PPO	41.8	34.1	13.3	45.3	36.7	12.5	50.6	41.2	11.2
Reward Learning	ALDM	38.7	35.6	15.2	42.1	38.6	14.9	44.9	42.1	13.7
	GDPL	49.5	47.5	12.8	54.9	53.2	12.1	60.4	59.1	10.8
Semi-VAE Enhanced	SS-PPO	45.2	36.2	13.6	47.4	37.2	12.4	53.1	43.6	11.5
	SS-ALDM	39.6	38.8	14.7	44.7	43.8	13.2	47.8	51.3	12.4
	SS-GDPL	53.7	51.2	11.1	61.3	58.4	10.5	66.5	68.7	9.2
Proposed	SS-VRNN	68.7	63.2	9.4	75.1	68.5	8.6	77.3	72.4	8.2
	Act-GDPL	70.6	65.6	9.5	78.8	71.1	8.4	80.9	78.0	8.2
	Act-VRNN	**76.2**	**72.7**	**9.1**	**83.0**	**81.8**	**8.0**	**85.5**	**86.7**	**7.9**

tion results as action annotations for expert demonstrations. [1] We also compare the full model **Act-VRNN** with its two variants: (1) **SS-VRNN** uses a VRNN that consumes predicted action labels instead of action embeddings; (2) **Act-GDPL** feeds expert demonstrations enriched by action embeddings to the same reward function as GDPL

4.2 Overall Results

Table 2 shows that our proposed model consistently outperforms other models in the setting that uses fully and partially annotated dialogues ($D_{\mathcal{F}}$ and $D_{\mathcal{P}}$). Act-VRNN improves task completion (measured by Entity-F1 and Success) while requiring less cost (measured by Turns). For example, Act-VRNN (81.8) outperforms SS-GDPL (60.4) by 35.4% under Success when having 10% fully annotated dialogues, and requires the fewest turns. Meanwhile, we find that both action learning and dynamics model are essential to the superiority of Act-VRNN. For example, Act-VRNN achieves 19.8% and 11.2% improvements over SS-VRNN and Act-GDPL, respectively, under Success when having 20% fully annotated dialogues. This validates that the learned action embeddings well capture similarities among actions, and VRNN is able to exploit such similarities for reward estimation.

We further find that the improvements brought by semi-VAE enhancement is limited for baselines, especially when the ratio of fully annotated dialogues is low. For example, SS-PPO and SS-GDPL achieve 6% and 7% improvements over their counterparts under Success when having 5% fully annotated dialogues. Similar results are also observed for pseudo-label approach. In general, the pseudo-label methods are outperformed by the counterparts of Semi-VAE and are even worse than the baselines without enhancement when the ratio of fully annotated dialogues is low. For example, in setting $D_{\mathcal{F}} + D_{\mathcal{P}}$, pseudo-label enhanced PPO performs worse than PPO under Entity-F1 when the ratio of fully annotated dialogues is 5% and 10% (37.2 vs 41.8, 39.2 vs 45.3), and only achieves slightly gain when the ratio is 20% (51.0 vs 50.6). This is largely because the prediction accuracy of Semi-VAE and pseudo-label approach might be low with a small amount of fully annotated dialogues, and the expert dialogues with mispredicted actions impinge reward function learning of baselines. Act-VRNN overcomes this challenge with the generalization ability brought by modeling dialogue progress in an action embedding space for reward estimation.

The results for policy learning using unlabeled dialogues ($D_{\mathcal{U}}$) are shown on Table 3. We consider two settings: (1) having fully labeled and unlabeled dialogues, i.e., $D_{\mathcal{F}}$ + $D_{\mathcal{U}}$; (2) having all three types of dialogues , i.e., $D_{\mathcal{F}}$ + $D_{\mathcal{P}}$ +$D_{\mathcal{U}}$. We can see that Act-VRNN significantly outperforms the baselines in both settings. For example, in setting $D_{\mathcal{F}}$ + $D_{\mathcal{U}}$, Act-VRNN outperforms SS-GDPL by 43% and 44% under Entity-F1 and Success, respectively. Similar results are also observed in setting $D_{\mathcal{F}}$ + $D_{\mathcal{P}}$ +$D_{\mathcal{U}}$. We further find that SS-VRNN outperforms Act-GDPL in these two settings while the results are opposite in setting $D_{\mathcal{F}}$ + $D_{\mathcal{P}}$, and we will conduct a detailed discussion in the following section. By comparing results of Act-VRNN and baselines in these two settings, we can see that Act-VRNN can better exploit the additional partially labeled dialogues. For example, SS-GDPL only achieves 2.3% under Success while Act-VRNN achieves more than 5%.

[1] We also experimented with the pseudo-label approach (Lee, 2013), and the empirical results were worse than Semi-VAE. Thus, we only report the Semi-VAE enhancement results in the table for simplicity.

Table 3: Semi-Supervised Policy Learning Results ($D_{\mathcal{F}}$, $D_{\mathcal{P}}$, and $D_{\mathcal{U}}$)

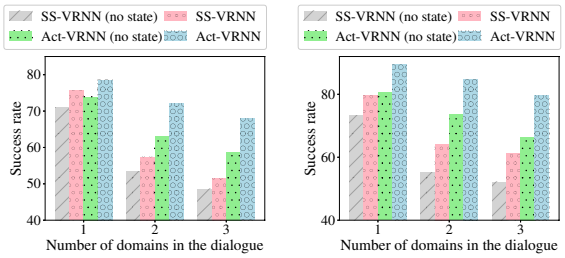
SUPERVISION	MODEL	Entity-F1	Success	Turns
$D_{\mathcal{F}}(10\%)$ + $D_{\mathcal{U}}(90\%)$	ALDM	40.0	34.9	15.9
	SS-PPO	44.7	33.8	12.9
	SS-ALDM	42.1	36.4	14.9
	SS-GDPL	56.3	50.2	11.8
	SS-VRNN	74.1	67.1	9.1
	Act-GDPL	72.9	66.7	8.5
	Act-VRNN	**80.6**	**72.4**	**8.4**
$D_{\mathcal{F}}(10\%)$ + $D_{\mathcal{P}}(10\%)$ + $D_{\mathcal{U}}(80\%)$	ALDM	41.7	35.2	15.7
	SS-PPO	44.9	34.6	12.8
	SS-ALDM	42.5	40.1	14.7
	SS-GDPL	57.1	51.4	10.7
	SS-VRNN	75.6	67.9	8.8
	Act-GDPL	73.3	67.1	8.5
	Act-VRNN	**81.1**	**76.3**	**8.2**

* Note that PPO and GDPL achieve the same results as $D_{\mathcal{F}}(10\%)$+$D_{\mathcal{P}}(90\%)$ in Table 2 since they can only utilize dialogues in $D_{\mathcal{F}}$

4.3 Discussions

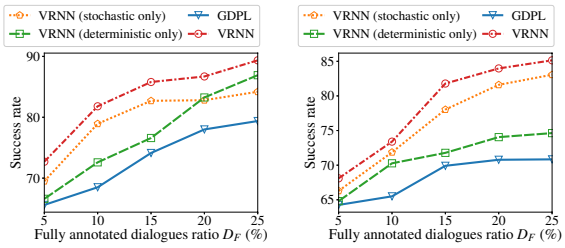
We first study the effects of action learning module in Act-VRNN. We compare Act-VRNN with SS-VRNN, and their counterparts that do not use state transition based objective in semi-supervised learning (i.e., optimizing Eqn. 3 instead of Eqn. 7). These two variants are denoted as Act-VRNN (no state) and SS-VRNN (no state). For a thorough investigation, under each setting, we further show the performances under dialogues spanning over different number of domains. Dialogues spanning over more domains are considered more difficult. The results under two supervision ratio setting are shown in Fig. 2(a) and Fig. 2(b). We can see that Act-VRNN outperforms other variants in each configuration, especially in the dialogues that include more than one domains. This is largely because the learned action embeddings effectively discover the similarities between actions across domains, and thus lead to better generalization of reward estimation. We further find that the state transition based objective we formulated fits well with the VRNN based reward estimator. Both Act-VRNN and SS-VRNN optimized considering state transitions achieve performance gains.

Last, we study the effects of dynamics model based reward function in Act-VRNN. We consider four different models as reward function: (1) our full dynamics model VRNN; (2) a dynamics model having only deterministic states (Eqn. 17); (3) a dynamics model having only stochastic states (Eqn. 15); (4) GDPL. All four models are learned based



(a) $D_{\mathcal{F}}(5\%) + D_{\mathcal{P}}(95\%)$ (b) $D_{\mathcal{F}}(20\%) + D_{\mathcal{P}}(80\%)$

Figure 2: Effects of action learning ($D_{\mathcal{F}}$ and $D_{\mathcal{P}}$)



(a) $D_{\mathcal{F}} + D_{\mathcal{P}}$ (b) $D_{\mathcal{F}} + D_{\mathcal{U}}$

Figure 3: Effects of dynamics model

on action embedding learned in the action learning module. The results under $D_{\mathcal{F}} + D_{\mathcal{P}}$ and $D_{\mathcal{F}} + D_{\mathcal{U}}$ are shown in Fig. 3(a) and Fig. 3(b), respectively. We can see that both stochastic and deterministic states in VRNN are important, since VRNN outperforms its two variants and GDPL in each configuration. We further find that the contribution of stochastic and deterministic states may vary in different setting. For example, VRNN (stochastic only) consistently outperforms VRNN (deterministic only) in $D_{\mathcal{F}} + D_{\mathcal{U}}$ while opposite results are observed in $D_{\mathcal{F}} + D_{\mathcal{P}}$ when ratio of $D_{\mathcal{F}}$ is over 20%. This is largely because modeling dialogue progress using stochastic states can provide more stable with less supervision signals, while the incorporation of deterministic can lead to more precise estimation can when more information of expert demonstrations are available.

5 Related Work

Reward learning aims to provide more effective and sufficient supervision signals for dialogue policy. Early studies focus on learning reward function utilizing external evaluations, e.g., user experience feedbacks (Gašić et al., 2013), objective ratings (Su et al., 2015; Ultes et al., 2017), or a combination of multiple evaluations (Su et al., 2016; Chen et al., 2019). These approaches often assume a human-in-the-loop setting where interactions with real users are available during training, which is

expensive and difficult to scale. As more large-scale high-quality dialogue corpus become available (e.g., MultiWOZ (Budzianowski et al., 2018)), recent years have seen a growing interest in learning reward function from expert demonstrations. Most recent approaches apply inverse reinforcement learning techniques for dialogue policy learning (Takanobu et al., 2019; Li et al., 2019b). These all require a complete state-action annotation for expert demonstrations. We aim to overcome this limitation in this study.

Semi-supervised learning aims to utilize unlabeled data to boost model performance, and is studied in computer vision (Iscen et al., 2019), item ranking (Park and Chang, 2019; Huang et al., 2019b), and multi-label classification (Miyato et al., 2015; Wang et al., 2018, 2019b). Many studies apply semi-supervised VAE (Kingma et al., 2014) for different classification tasks, e.g., sentiment analysis (Xu et al., 2017; Li et al., 2019a), text matching (Shen et al., 2018; Choi et al., 2019). While these work focus on prediction accuracies, we aim to enrich expert demonstrations via semi-supervised learning.

6 Conclusions

We study the problem of semi-supervised policy learning and propose Act-VRNN to provide more effective and stable rewards estimations. We formulate a generative model to jointly infer action labels and learn action embeddings. We design a novel reward function to first model dialogue progress, and estimate action rewards by determining whether the action leads to similar progress as expert dialogues. The experimental results confirm that Act-VRNN achieves better task completion compared with the state-of-the-art in two settings that consider partially labeled or unlabeled dialogues. For future work, we will explore the scenarios that annotations are absent for all expert dialogues.

Acknowledgement

We would like to thank Xiaojie Wang for his help. This work is supported by Australian Research Council (ARC) Discovery Project DP180102050, and China Scholarship Council (CSC).

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Yash Chandak, Georgios Theocharous, James Kostas, Scott Jordan, and Philip Thomas. 2019. Learning action representations for reinforcement learning. In *International Conference on Machine Learning*, pages 941–950.

Runyu Chen, Qili Wang, and Wen I Xu. 2019. Mining user requirements to facilitate mobile app quality upgrades with big data. *Electron. Commer. Res. Appl.*, 38.

Jihun Choi, Taeuk Kim, and Sang-goo Lee. 2019. A cross-sentence latent variable model for semi-supervised text sequence matching. In *Annual Meeting of the Association for Computational Linguistics*, pages 4747–4761, Florence, Italy. Association for Computational Linguistics.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, pages 2980–2988.

Justin Fu, Katie Luo, and Sergey Levine. 2017. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. *arXiv preprint arXiv:1809.08267*.

Milica Gašić, Catherine Breslin, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve Young. 2013. On-line policy optimisation of bayesian spoken dialogue systems via human interaction. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8367–8371.

Xinting Huang, Jianzhong Qi, Yu Sun, and Rui Zhang. 2019a. Mala: Cross-domain dialogue generation with action learning. *arXiv preprint arXiv:1912.08442*.

Xinting Huang, Jianzhong Qi, Yu Sun, Rui Zhang, and Hai-Tao Zheng. 2019b. Carl: Aggregated search with context-aware module embedding learning. In *IJCNN*, pages 101–108.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondej Chum. 2019. Label propagation for deep semi-supervised learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5065–5074.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302.

Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2.

Sungjin Lee, Qi Zhu, Ryuichi Takanobu, Xiang Li, Yaoqin Zhang, Zheng Zhang, Jinchao Li, Baolin Peng, Xiujun Li, Minlie Huang, and Jianfeng Gao. 2019. Convlab: Multi-domain end-to-end dialog system platform. In *Annual Meeting of the Association for Computational Linguistics*, page 6469.

Xuijun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *International Joint Conference on Natural Language Processing*, pages 187–196.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2019a. Semi-supervised stochastic multi-domain learning using variational inference. In *Annual Meeting of the Association for Computational Linguistics*, pages 1923–1934, Florence, Italy. Association for Computational Linguistics.

Ziming Li, Julia Kiseleva, and Maarten de Rijke. 2019b. Dialogue generation: From imitation learning to inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pages 6722–6729.

Bing Liu and Ian Lane. 2018. Adversarial learning of task-oriented neural dialog models. In *Annual SIGdial Meeting on Discourse and Dialogue*, pages 350–359.

Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin Murphy, and Honglak Lee. 2019. Unsupervised learning of object structure and dynamics from videos. *arXiv preprint arXiv:1906.07889*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. 2015. Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677*.

Dae Hoon Park and Yi Chang. 2019. Adversarial sampling and training for semi-supervised information retrieval. *ArXiv*, abs/1811.04155.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong. 2018. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. In *Annual Meeting of the Association for Computational Linguistics*, pages 2182–2192.

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

Dinghan Shen, Yizhe Zhang, Ricardo Henao, Qinliang Su, and Lawrence Carin. 2018. Deconvolutional latent-variable model for text sequence matching. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5438–5445.

Pei-Hao Su, Milica Gasic, Nikola Mrkšić, Lina M Rojas Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Annual Meeting of the Association for Computational Linguistics*, pages 2431–2441.

Pei-Hao Su, David Vandyke, Milica Gasic, Nikola Mrksic, Tsung-Hsien Wen, and Steve Young. 2015. Reward shaping with recurrent neural networks for speeding up on-line policy learning in spoken dialogue systems. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 417–421.

Chen Sun, Per Karlsson, Jiajun Wu, Joshua B Tenenbaum, and Kevin Murphy. 2019. Stochastic prediction of multi-agent interactions from partial observations. *arXiv preprint arXiv:1902.09641*.

Yu Sun, Nicholas Jing Yuan, Yingzi Wang, Xing Xie, Kieran McDonald, and Rui Zhang. 2016. Contextual intent tracking for personal assistants. In *SIGKDD*, pages 273–282. ACM.

Yu Sun, Nicholas Jing Yuan, Xing Xie, Kieran McDonald, and Rui Zhang. 2017. Collaborative intent prediction with real-time contextual data. *ACM Transactions on Information Systems (TOIS)*, 35(4):1–33.

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

Stefan Ultes, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, Lina M Rojas Barahona, Pei-Hao Su, Tsung-Hsien Wen, Milica Gasic, and Steve Young. 2017. Reward-balancing for statistical spoken dialogue systems using multi-objective reinforcement learning. In *18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 65–70.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Qili Wang, Wei Xu, Xinting Huang, and Kunlin Yang. 2019a. Enhancing intraday stock price manipulation detection by leveraging recurrent neural networks with ensemble learning. *Neurocomputing*, 347:46–58.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2018. Kdgan: knowledge distillation with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 775–786.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019b. Adversarial distillation for learning with privileged provisions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–12.

Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. 2017. Variational autoencoder for semi-supervised text classification. In *AAAI Conference on Artificial Intelligence*, pages 3358–3364.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, pages 7287–7298.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Zhirui Zhang, Xiujun Li, Jianfeng Gao, and Enhong Chen. 2019. Budgeted policy learning for task-oriented dialogue systems. *arXiv preprint arXiv:1906.00499*.

Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1208–1218.