INTEGRATING PRETRAINED LANGUAGE MODEL FOR DIALOGUE POLICY EVALUATION

*Hongru Wang, Huimin Wang, Zezhong Wang, Kam-Fai Wong**

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong

ABSTRACT

Reinforcement Learning (RL) has been witnessed its potential for training a dialogue policy agent towards maximizing the accumulated rewards given from users. However, the reward can be very sparse for it is usually only provided at the end of a dialog session, which causes unaffordable interaction requirements for an acceptable dialog agent. Distinguished from many efforts dedicated to optimizing the policy and recovering the reward alternatively which suffers from easily getting stuck in local optima and model collapse, we decompose the adversarial training into two steps: 1) we integrate a pre-trained language model as a discriminator to judge whether the current system action is good enough for the last user action (i.e., *next action prediction*); 2) the discriminator gives and extra local dense reward to guide the agent's exploration. The experimental result demonstrates that our method significantly improves the complete rate (4.4%) and success rate (8.0%) of the dialogue system.

Index Terms— Reward Shaping, Dialogue Policy Learning, Pre-trained Language Model

1. INTRODUCTION

Reinforcement learning has revolutionized the way to model the dialogue policy which decides the next action of the dialogue system suited to the current state [1, 2, 3, 4, 5, 6, 7]. On the flip side, one notorious limitation of reinforcement learning is *reward sparsity* issue where here the system usually receive a positive (or negative) reward signal when the dialogue ends successfully (or unsuccessfully). Thus, the reward signal is delayed and sparse, making it extremely difficult to connect a long series of actions to a distant future reward especially for complex goals across multiple domains [6].

A typical reward function, for example, often apply a minor negative penalty (i.e., -1) in the middle of the session to encourage the system to accomplish the task in fewer turns, with a huge positive (i.e., +40) or negative (i.e., -40) reward at the end [1]. This kind of *global reward* based on $(state, action)$ pairs is not informative and lead to exploration in large action space inefficient [8]. To get more dense and enlightening reward signals, most previous works recovers the intrinsic *local reward* from expert demonstrations through *reward shaping* [5, 9]. More specifically, some works train a discriminator to differentiate $(state, action)$ generated by dialogue agents from $(state, action)$ by expert and then regards the discriminator as a reward estimator to provide intrinsic reward signals, where the dialogue policy model and discriminator updates alternately on the fly [10, 6]. A further line of work decomposes the whole training into two steps by firstly training the discriminator with an auxiliary dialogue generator and secondly incorporating it into common RL method, since the alternative update mechanism limits the policy model to policy-gradient-based algorithms [7]. However, the vast bulk of annotated $(state, action)$ pairs from expert demonstration is hard to acquire. Moreover, the reward model based on state-action pair might cause unstable policy learning and affect optimization speed with the limited amount of annotated dialogues [11].

Our work keeps in line with the methods to decompose the whole training into two sequential steps. We incorporate the pretrained language model as the reward model into common RL method to provide dense *local reward* signals, guiding the action decision of dialogue policy learning. Specifically, we re-formulate one of the pre-training sub-tasks of BERT *Next Sentence Prediction* as *Next Action Prediction* at the first step. Given current user action a_u, the classifier will distinguish whether or not the response system action a_s is suitable or acceptable. Intuitively, if the system chooses the right action to answer user's query at each turn, then the dialogue naturally succeeds at the end. Secondly, the trained classifier as the dialogue reward model will be incorporated into the RL process to guide the dialogue policy learning without updates. There are several advantages of our proposed method: 1) Our method is model-agnostic which can be incorporated in any RL algorithm to guide the policy learning, 2) Only action-pairs demonstration reduce the cost of annotation compared with state-action pairs demonstration, and 3) Pre-trained language model has been proved powerful and transferable in many NLP tasks which can capture the subtle difference of action-pairs, providing more dense reward signal at each turn.

The main contribution of this paper is two-fold: 1) we propose a simple yet effective reward estimator at the action-

*Corresponding Author

level to guide the action decision, and 2) We investigate the effects of *global reward* and *local reward*, and the experimental results on MultiWOZ [12] show that our methods outperform single *global reward* about 4.4%, and furthermore, almost 8% when combined with *global reward*, which indicates the complementary effects of these two types of reward.

2. RELATED WORK

The first focus of reward shaping is recovering the intrinsic local reward from expert demonstration by inverse reinforcement learning or adversarial training. Peng et al. [10] proposed an adversarial advantage actor-critic (Adversarial A2C) method based on $(state, action)$ pairs from simulation and expert demonstration. Similarly, Takanobu et al. [6] utilizes Adversarial Inverse Reinforcement Learning to jointly estimate reward and optimize dialogue policy in multi-domain task-oriented dialog. Wang et al. [8] directly estimate potential-based reward function from demonstrations. Nevertheless, methods alternately learning the reward model and policy cannot be extended to off-policy methods which benefit from self-learned reward functions [7].

The second focus of reward shaping is to train reward model and dialogue policy consecutively, rather than alternatively as mentioned above [7]. Li et al. [7] appropriate the reward model with multilayer perceptron (MLP) to train a discriminator based on $(state, action)$ pairs. Besides that, Gabriel et al. [13] propose a No Label Expert (NLE) that uses an unannotated dialog dataset consisting of pairs of sentences (s_u, s_s), representing user utterances and the corresponding agent responses, guiding the dialogue policy. Distinguishing from these works, we tackle *reward sparsity* by incorporating pretrained language model as a reward estimator and train a discriminator at action-level rather than sentence-level or $(state, action)$ pairs.

3. METHODOLOGY

In this section, we firstly introduce the detail of next action prediction (namely, how to train the discriminator), and then presents an algorithm that incorporates the reward signal given in the first step (namely, how to update the dialogue policy). Figure 1 shows the overall framework of our proposed method.

3.1. Next Action Prediction

Next Sentence Prediction is one of pre-trained task in BERT, aiming to predict whether, for a given pair of sentences (s_1, s_2), s_2 is an next sentence to s_1. Similarly, *Next Action Prediction* task is to determine whether or not the current system action a_s is an appropriate response to the last user action a_u in dialogue. An action (i.e., a_u and a_s) consists of

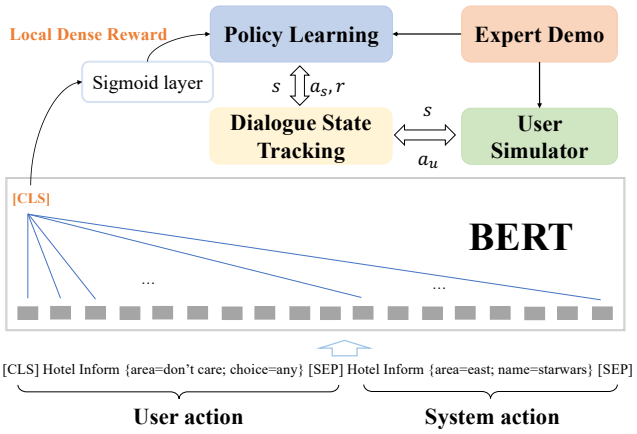

Fig. 1. Our proposed method: Integrating pre-trained language model into reinforcement learning as reward model to provide local dense reward signal based on action pairs (a_u, a_s), guiding the optimization of dialogue policy learning.

domain, intent, slots and its corresponding value as follows [14]:

$$\mathcal{A} = [\underbrace{\mathrm{D}}_{\text{Domain}}\ \underbrace{\mathrm{I}}_{\text{Intent}}\ \underbrace{\{s_1 = v_1; \cdots; s_P = v_P\}}_{\text{Slot-value pairs}}] \quad (1)$$

At each turn of multi-domain dialogue, the user and system may inquire and provide information across different domain. In this case, we represent the action as:

$$a_u = [\mathcal{A}_u^1, \mathcal{A}_u^2, ..., \mathcal{A}_u^n] \quad a_s = [\mathcal{A}_s^1, \mathcal{A}_s^2, ..., \mathcal{A}_s^n] \quad (2)$$

Then, we concatenate the user action a_u and system action a_s and feed it into BERT. Two special tokens $[CLS]$ and $[SEP]$ are added to indicate the start and separation of actions respectively. The input representations consist of token embedding, segment embedding and positional embedding.

$$H = \mathrm{BERT}(\mathrm{emb}(a_u, a_s)) \quad (3)$$

The embedding of first token (i.e. [CLS]) then is then feed into a sigmoid layer to do classification as defined:

$$f(x) = \mathrm{sigmoid}(Wh_1 + b) \quad (4)$$

The final output of the binary classifier $f(x)$ is a probability that indicates the confident score that the system action is suitable and appreciate given last user action. The objective of the classifier D_ϕ can be represented as follows:

$$L_D = \mathrm{E}(log(1 - D_\phi(a_u, a_s)_{sim})) - \mathrm{E}(D_\phi(a_u, a_s)_{real}) \quad (5)$$

After the discriminator is trained, we will keep it as the reward function for future dialogue policy learning without updates, which is illustrated at later section.

3.2. PPO-OFF

A trajectory $(s_0, a_0, s_1, a_1, ...)$ is generated by sampling actions according to the policy $a_t \sim \pi(a_t|s_t)$ consecutively,

until the terminal states is reached. Here, the action a_t can be further divided into the user action a_u and system action a_s in dialogue. A reward signal r_t is received at each time step.

$$\mathcal{T} = [(s_0, (a_u, a_s), r_0, s_1), ..., (s_n, (a_u, a_s), r_n, s_{n+1})] \quad (6)$$

where $s_t, a_u, a_s, r_t, s_{t+1}$ represents the dialogue state, user action, system action, reward and next state at time step t respectively. The main objective of agent is to maximize the cumulative reward $R = \sum_0^T \gamma^t r_t$, where γ is a discount factor. Given the objective function, the gradient can be computed as follows:

$$g = \mathbb{E}[\sum_0^T \Psi_t \nabla_\theta \log \pi_\theta(a_t|s_t)] \quad (7)$$

where Ψ_t can be estimated with different methods. We adapt generalized advantage function $A^\pi(s_t, a_t)$ here as follows [1]:

$$A_t^{GAE(\gamma,\lambda)} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^V \quad (8)$$

where $\delta_t^V = -V(s_t) + r_t + \gamma V(S_{t+1})$. Both γ and λ plays key roles in the bias-variance trade-off and serve different purposes when using an approximate value function. To stabilize the policy updates, we adapt the clipped surrogate objective as follows [15].

$$L^{CLIP} = \mathbb{E}[min(r_t(\theta)A_t^{GAE}, clip(r_t(\theta), 1-\sigma, 1+\sigma)A_t^{GAE})] \quad (9)$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$. In addition to policy, we also parametrize the value function and add entropy bonus to ensure sufficient exploration[15]. The whole training objective is defined as follows:

$$L^{CLIP+VF+S} = \mathbb{E}[L^{CLIP} - c_1 * L^{VF} + c_2 * S[\pi](s)] \quad (10)$$

where c_1, c_2 are coefficients, and S denotes an entropy bonus, and L^{VF} is a squared-error loss $(V_\theta(s_t) - V_t^{tar})^2$.

It is obvious that the reward model plays a key role in dialogue policy learning since it directly affect the training objective. To investigate the effects of different reward signals, we replace the r_t in the trajectory with three different types of reward: *global reward*, *local reward* and *combination*.

For *global reward*, we simply assign it a large positive (v.s. negative) reward +40 (v.s. -40) when the dialogue ends successfully (v.s. unsuccessfully), while -1 during the middle of session to encourage shorter session. For *local reward*, we remap the output confident score to a range of [-1, 1] as follows, aiming to encourage the policy to decide more high qualified actions.

$$r_{local} = -1 + 2 * D_\phi(a_u, a_s) \quad r_{comb} = r_{global} + r_{local} \quad (11)$$

[1] a_t refers the system action a_s while a_u is decided by a user simulator

4. EXPERIMENTS AND ANALYSIS

4.1. Dataset and Evaluation Metric

All models are evaluated on MultiWOZ [12], a multi-domain, multi-intent task-oriented dialog corpus that contains 7 domains, 13 intents, 25 different slots and 10483 dialog sessions. We report the average number of *dialog turns*, averaging over successful dialog sessions and all dialog sessions respectively, to measure the efficient of accomplishing a task. A dialog turn consists of a user utterance and a subsequent system (i.e. utterance pairs). *Precision*, *recall* and *F1* are calculated based on dialog acts (i.e. action pairs) [16]. *Match rate* assesses whether the offered entity meets all the constraints specified in a user goal. The dialog is marked as successful if and only if both inform recall and match rate are 1.

4.2. Baselines

MLE: One of representative work of supervised learning to learn dialogue policy, which employs a multi-class classification via imitation learning (i.e., behavioral cloning) with a set of compositional actions where a compositional action consists of a set of dialog act items.

GDPL: [6] A method which alternatively updates dialogue policy and reward estimation model by using adversarial inverse reinforcement learning. It is noted that reward estimator recovers the reward signal form the state-action pairs at each dialogue turn.

PPO [15] A policy-based reinforcement learning method which uses multiple epochs of stochastic gradient ascent and a constant clipping mechanism the soft constraint to perform each policy update. The dialogue policy model in GDPL is also PPO.

PPO-OFF-Comb The reward model offers the combination of *global reward* and *local reward* while the policy updated as illustrated in section 3.2. Similarly, **PPO-OFF-Local** only receives local dense reward from pre-trained language model (i.e., BERT) and **PPO-OFF-Global** receives only global reward respectively.

4.3. Implementation Details

For *next action prediction*, we firstly build the binary classification dataset from MultiWoZ [12] automatically and get 99370, 13157, 13073 labeled samples for training, testing and validation respectively [2]. And then we use BERT[17] as backbone and $Adam$ as optimization algorithm. The specific hyper-parameters are deployed as follows: batch size as 4, learning rate as 5e-5, epochs as 10, adam epsilon as 1e-8, the max sequence length as 512 [3].

[2] We sample user action a_u with system action a_s from same dialogue as positive sample, and randomly sample another system action from other dialogue to form negative sample.

[3] It is noted that the trained BERT model achieved 97.34% accuracy at the binary classification task, which proves the pre-trained language model is

Model	Agenda		
	Precision/Recall/F1	Match	Success
Human	81.9/93.4/85.3	92.1	82.7
MLE	63.1/72.6/64.5	50.1	47.0
GDPL	63.1/73.0/64.6	50.0	47.2
PPO	64.4/78.3/68.2	64.7	61.2
PPO-OFF-Global	63.0/81.2/67.8	70.8	63.2
PPO-OFF-Local	60.9/82.9/67.4	70.9	67.6
PPO-OFF-Comb	**67.2/85.7/72.7**	**79.6**	**71.4**

Table 1. Performance of different dialog agents on the multi-domain dialog corpus by interacting with the agenda-based user simulator. Success average turns / all average turns

For *dialogue policy learning*, we adapt ConvLab-2 [18] as our environment and evaluate the policy at sentence-level instead of action-level. In this case, other components in task-oriented dialogue system are *BERTNLU*, *RuleDST* and *Template-based NLG*. More specifically, we set the maximum turn of one conversation as 20, which means the dialogue will be terminated and regarded as failure when the turn excesses 20. We sample 1024 trajactories each epoch and the max epoch is set as 200. Other hyper-parameter settings follows Takanobu et al. [6].

4.4. Main Result

Table 1 demonstrates the performance of different methods on the MultiWoZ dataset [12]. Consistent with intuition, the combination (i.e., PPO-OFF-Comb) of global and local reward reaches highest performance in both match rate (9% improvement) and success rate (8% improvement). Besides that, PPO-OFF-Local achieves comparable match rate but much higher success rate compared with PPO-OFF-Global. We attribute this to more semantic frame are correct in the dialogue since success rate is much harder to improve than match rate. Compared with previous methods such as PPO, our proposed methods demonstrate superior performance by improving almost 10%.

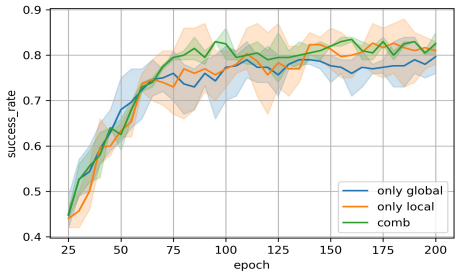

Fig. 2. Learning curves of our proposed method under different reward signals and action-level

capable to provide reliable reward signal.

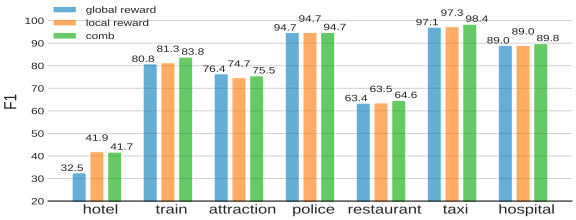

Fig. 3. F1 score for different domain by executing policy learned with different rewards

4.5. Analysis

We also conduct evaluation at the action-level (i.e., without NLU and NLG part) to learn the *converge speed* and *performance under different domains* respectively.

Converge Speed The learning curves under different reward types are presented in Fig 2. We can see that the comb reward converge faster than only local and only global reward, leading to higher performance. Besides, the error band of comb reward is dramatically smaller than the other two rewards which indicates more stable policy learning. Therefore, we conclude that *local reward* signals serve as a complement of *global reward* to better guide and stabilize the behavior of dialogue policy.

The Effects of Different Domains We also report the F1 score of different domains with policy learned from different rewards as shown in Fig 3. We noticed that there is still a large gap between different domains. The comb reward consistently outperforms single global or local reward signals, especially at the hotel, restaurant, taxi, and hospital domains. We conjecture this to the data distribution of the original dataset and the difficulty and complexity of different domains. These domains contain more slots and act types than other domains [12]. In this case, the combination of slots will be increasing exponentially which damages the effectiveness of the trained discriminator, since the combination of the slot that appears during training is much different from testing (i.e, the hotel domain has 15 slots which is the most).

5. CONCLUSION

In this paper, we integrates pre-trained language model (i.e. BERT) as a reward model to provide *local reward* that complements with *global reward*, guiding the dialogue policy learning. The experimental results demonstrate the combination of global reward and local reward reaches highest performance compares with only global or only local reward. We left more investigation such as pretraining in future work.

6. ACKNOWLEDGEMENTS

This work is partially supported by HK GRF#14204118 and HK RSFS#3133237.

7. REFERENCES

[1] Jianfeng Gao, Michel Galley, and Lihong Li, "Neural approaches to conversational ai," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 1371–1374.

[2] Marilyn A Walker, "An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email," *Journal of Artificial Intelligence Research*, vol. 12, pp. 387–416, 2000.

[3] Lihong Li, Jason D Williams, and Suhrid Balakrishnan, "Reinforcement learning for dialog management using least-squares policy iteration and fast feature selection," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[4] Pierre-Luc Bacon, Jean Harb, and Doina Precup, "The Option-Critic Architecture," *arXiv:1609.05140 [cs]*, Dec. 2016.

[5] Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong, "Composite Task-Completion Dialogue Policy Learning via Hierarchical Deep Reinforcement Learning," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Sept. 2017, pp. 2231–2240, Association for Computational Linguistics.

[6] Ryuichi Takanobu, Hanlin Zhu, and Minlie Huang, "Guided Dialog Policy Learning: Reward Estimation for Multi-Domain Task-Oriented Dialog," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 100–110, Association for Computational Linguistics.

[7] Ziming Li, Sungjin Lee, Baolin Peng, Jinchao Li, Julia Kiseleva, Maarten de Rijke, Shahin Shayandeh, and Jianfeng Gao, "Guided Dialog Policy Learning without Adversarial Learning in the Loop," *arXiv:2004.03267 [cs]*, Sept. 2020.

[8] Huimin Wang, Baolin Peng, and Kam-Fai Wong, "Learning efficient dialogue policy from demonstrations through shaping," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 6355–6365, Association for Computational Linguistics.

[9] Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong, "Deep Dyna-Q: Integrating planning for task-completion dialogue policy learning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018, pp. 2182–2192, Association for Computational Linguistics.

[10] Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Yun-Nung Chen, and Kam-Fai Wong, "Adversarial advantage actor-critic model for task-completion dialogue policy learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6149–6153.

[11] Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick, "Unsupervised text style transfer using language models as discriminators," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 7298–7309.

[12] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić, "Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," *arXiv preprint arXiv:1810.00278*, 2018.

[13] Gabriel Gordon-Hall, Philip John Gorinski, and Shay B Cohen, "Learning dialog policies from weak demonstrations," *arXiv preprint arXiv:2004.11054*, 2020.

[14] Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao, "Few-shot natural language generation for task-oriented dialog," *arXiv preprint arXiv:2002.12328*, 2020.

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

[16] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang, "Convlab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems," *arXiv preprint arXiv:2002.04793*, 2020.