INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Causal Confusion Reduction for Robust Multi-Domain Dialogue Policy

Mahdin Rohmatillah, Jen-Tzung Chien

Dept of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Taiwan
{mahdin.ee08,jtchien}@nycu.edu.tw

Abstract

In the multi-domain dialogue system, dialog policy plays an important role since it determines the suitable actions based on the user's goals. However, in many recent works, most of the dialogue optimizations, especially that use reinforcement learning (RL) methods, do not perform well. The main problem is that the initial step of optimization that involves the behavior cloning (BC) methods suffer from the causal confusion problem, which means that the agent misidentifies true cause of an expert action in current state. This paper proposes a novel method to improve the performance of BC method in dialogue system. Instead of only predicting correct action given a state from dataset, we introduce the auxiliary tasks to predict both of current belief state and recent user utterance in order to reduce causal confusion of the expert action in the dataset since those features are important in every dialog turn. Experiments on ConvLab-2 shows that, by using this method, all of RL based optimizations are improved. Furthermore, the agent based on the proximal policy optimization shows very significant improvement with the help of the proposed BC agent weights both in policy evaluation as well as in end-to-end system evaluation.

Index Terms: causal confusion, behavior cloning, multi-domain dialogue system, reinforcement learning

1. Introduction

Task-oriented dialogue learning is getting high interest nowadays since this research topic practically reflects real-world problem [1, 2, 3]. In this task, system must have a robust dialogue policy in order to provide an appropriate system response. Among dialogue policy optimizations, reinforcement learning (RL) based optimization is the most popular method [4, 5, 6, 7, 8], because the system can be viewed as an agent who learns how to provide an appropriate action through learning by interactions with user in an environment. This agent basically transits or moves from one state s to another state s' after taking an action a which can be formulated as a Markov decision process. To stabilize the learning process for RL agent, the so-called behavior cloning (BC) has been popularly developed as the initial step in learning process. BC aims to find a mapping from a given state to a specific action based on a given dataset, which is trained in a supervised manner. Unfortunately, this method meet some problems which led to weak performance for generalization. One of the most common problems is the distributional shift between training and test data caused by the difference between expert and the learned agent trajectories. This issue is mainly induced by two prominent problems including the bias in dataset and the causal misidentification. Basically, the dataset bias emerges as the trajectories or the state-action pairs provided by dataset do not comprise all situations in real demonstrations. Therefore, in the heterogeneous environment, agent is prone to produce the failed trajectory once it makes a wrong action due to unseen situation. A straightforward solution is to conduct data augmentation, especially when uncommon state-action pairs are observed. However, for multi-domain dialogue system, providing augmented data is challenging and needs the complicated pre-processing steps to match the data format and style with the well-established environment setting. Another prominent problem is the causal confusion [9] which happens when the agent misidentifies true causes of an expert action in a given state during training. This problem is exacerbated by the stochastic gradient descent learning that typically assumes all data pairs are independent and identically distributed. In dialogue system, this problem considerably degrades the agent performance since the state of dialogue environment consist of various features or causes that may be irrelevant in some dialogue turns. For example, in the initial turn, the last system action and booking information are not important, but they are helpful in the upcoming turns. Meanwhile, the current belief state and recent user utterance are always important. The agent must consider them in every turn.

A lot of efforts have been made to deal with the aforementioned problems. Unfortunately, in the dialogue applications, BC problem is barely discussed. It is because the state and action spaces in the dialogue system are very large and the amount of real dialogues is very limited. High dimensionality and sparse data have created a huge obstacle for conducting researches. As an example, the most popular multi-domain dialogue dataset, MultiWOZ 2.1 [10], is a long-tailed dataset and the state and action dimensions have the size of 340 and 209, respectively. Therefore, based on the current benchmark table provided by ConvLab-2 [11, 12], all of the RL based solutions that include BC just perform moderately. Accordingly, this paper proposes a new approach to improve the performance of BC in the dialogue system. Instead of only predicting the correct action given a state from dataset, we introduce the auxiliary tasks [13, 14, 15, 16] to predict both current belief state and recent user utterance, which are seen as crucial features in every dialogue turn. The causal confusion of expert action is significantly reduced. Such auxiliary tasks act as the regularizer in the policy network updating because policy network considers the cumulative losses from both policy and auxiliary networks to obtain the optimal policy parameter. To illustrate the impact of proposed method, the BC agent weights are used as the pre-trained weights for other RL algorithms. The results show that by using BC agent weights all RL methods are improved in the multi-domain dialogue system. The proposed agent with proximal policy optimization (PPO) [5] can reach the performance close to rule policy in the policy evaluation and achieve state-of-the-art result in the dialogue pipeline system.

2. Background Survey

2.1. Multi-domain Dialogue System

In the recent years, dialogue systems have been developed rapidly. Started from only handling single domain like flight

http://dx.doi.org/10.21437/Interspeech.2021-534

booking [17] and restaurant reservation [18], it is meaningful to implement the multi-domain tasks with complicated structure and performance evaluation [19, 10, 20, 21]. Figure 1 depicts the system structure in handling multi-domain dialogues based on ConvLab-2 framework [12]. This framework allows system diagnosis and performance evaluation when building a dialogue agent as a pipeline system which requires individual optimization for individual components including natural language understanding (NLU), dialogue state tracking (DST), dialogue policy (POL) and natural language generation (NLG). Furthermore, word-level optimizations like word-DST and word-POL which combine NLU-DST and POL-NLG components respectively or even end-to-end optimization that optimizes whole components jointly are able to be done in this framework [12].

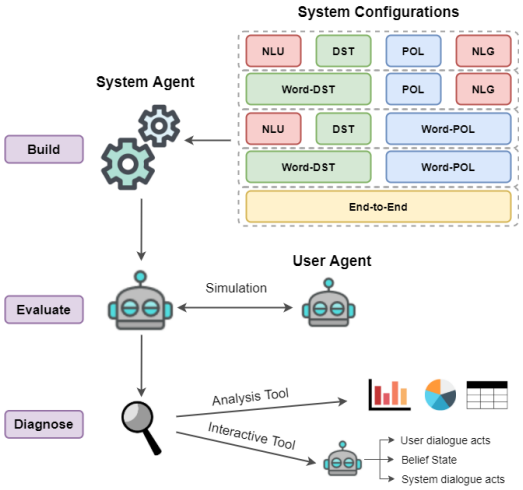

Figure 1: *ConvLab-2 framework for dialogue evaluation.*

A lot of efforts have been proposed to deal with dialogue system under different kinds of configuration [4, 6, 22, 23, 24]. However, most of them were assessed through the component-wise evaluation, which meant they were just judged individually within single turn evaluation rather than evaluated for the whole sequences in a dialogue [11]. Therefore, once those methods were applied to the real world case, they performed unsatisfactorily. This paper proposes a robust dialogue policy to provide state-of-the-art performance for pipeline system which is evaluated by the user simulated agent over the whole dialogue turns.

2.2. Behavior Cloning

Behavior cloning (BC) is one of the most popular realizations for imitation learning. BC learns an optimal policy from a given expert dataset, $\mathcal{D} = \{(s_1, y_1), \ldots, (s_N, y_N)\}$ where s and y are state and its corresponding expert expert action respectively. More specifically, the agent acquires the state-action pairs from expert dataset, then a regressor or classifier is estimated to mimic the expert behaviors. The optimal policy network $\pi(\cdot)$ with parameter θ is then trained by maximizing the data likelihood or minimizing the mean squared error or cross entropy error $\mathcal{L}(\cdot)$, expressed by

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}\left(\pi\left(s_i; \theta\right), y_i\right). \quad (1)$$

Due to its simplicity, BC has been used in many applications that mostly have abundant training samples to conduct supervised learning. For example, in the application of self-driving car, BC-based method obtained state-of-the-art performance [25]. Meanwhile, in the multi-domain dialogue system, BC is the first default step to provide the pre-trained weights for the other RL based optimization. This paper improves BC for dialogue system by means of resolving the causal confusion.

2.3. Causal Confusion

Using the BC method, causal confusion is a severe problem that substantially degrades the system performance. In general, causal confusion has close connection with the causal misidentification. This issue is caused with the phenomenon that accessing more information results in worse performance. Therefore, this problem is found to have serious effects especially in the tasks with huge state space. The exploration in the environments, e.g. self-driving car and spoken dialogue, becomes very limited. For self-driving car, a number of works have been developed to overcome this problem. Broadly speaking, providing a certain additional condition [26] or imposing a regularization to policy network was helpful to enhance the knowledge [27] or enrich the planning [25, 28] for an agent. These works were clearly feasible since the augmented data were easy to be gathered. Furthermore, because of the input images, the well-trained image classification network like ResNet-50 [29] was utilized to improve the system performance. Different from self-driving car task, multi-domain dialogue task did not have the robust pre-trained models. It was challenging to carry out data augmentation since there were several pre-processing methods which were required to match the format between dialogue environment and dialogue dataset. Therefore, BC problem in multi-domain dialogue system was barely discussed. The RL-based methods [4, 5] only showed weak or moderate improvement as provided in the ConvLab-2 benchmark.

3. Proposed Method

This section addresses the details of the proposed method including the architecture, the loss functions and the training procedure. The procedure to achieve state-of-the-art result in policy optimization using the proposed behavior cloning (BC) is discussed thoroughly.

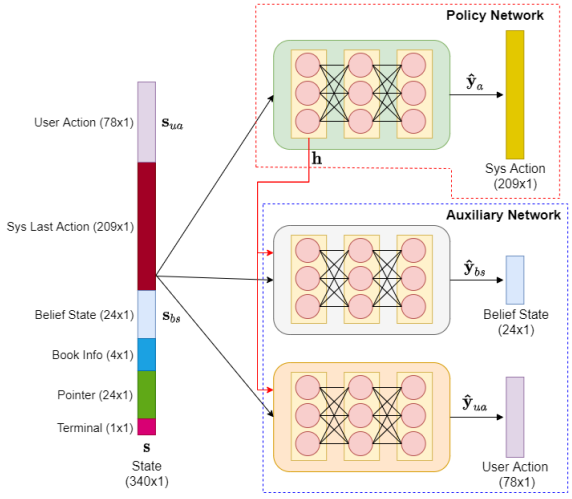

Figure 2: *Policy and auxiliary networks for prediction of system action $\hat{\mathbf{y}}_a$, belief state $\hat{\mathbf{y}}_{bs}$ and user action $\hat{\mathbf{y}}_{ua}$ in behavior cloning of dialogues. Various feature dimensions are shown.*

3.1. Behavior Cloning with Auxiliary Tasks

In the proposed BC method, the auxiliary tasks are introduced to predict both of current belief state and recent user utterance by utilizing the output features $\mathbf{h}$ from the first feedforward layer of policy network denoted as $p_a^{(1)}(\mathbf{s})$, which takes a state vector $\mathbf{s}$ from DST as an input. The introduction of auxiliary tasks is intended to provide prior knowledge of reasoning to our agent since current belief state and user action are the most important features in every dialogue turn. The remaining features may not be so important in every dialogue turn. Figure 2 shows the architecture of behavior cloning with policy network and auxiliary network where the corresponding loss functions $\mathcal{L}_{pol}$ and $\mathcal{L}_{aux}$ are introduced, respectively. Policy network is trained by minimizing the cross-entropy loss due to $\{\mathbf{s}_i, \mathbf{y}_{a,i}\}_{i=1}^{N}$.

$$\begin{aligned}\mathcal{L}_{pol}\left(p_a(\mathbf{s}), \mathbf{y}_a\right) = -\mathbb{E}_{(\mathbf{s},\mathbf{y}_a)\sim\mathcal{D}}\Big[\tilde{\mathbf{y}}_a^{\top}\log p_a(\mathbf{s}) \\ + (\mathbf{1}-\mathbf{y}_a)^{\top}\log(\mathbf{1}-p_a(\mathbf{s}))\Big].\end{aligned} \tag{2}$$

Meanwhile, $\mathcal{L}_{aux}$ is a summation of belief state loss, $\mathcal{L}_{bs}$, and current user action loss, $\mathcal{L}_{ua}$, which are formulated as

$$\begin{aligned}\mathcal{L}_{bs}\left(p_{bs}(\mathbf{s},\mathbf{h}), \mathbf{y}_{bs}\right) = -\mathbb{E}_{(\mathbf{s},\mathbf{y}_{bs})\sim\mathcal{D}}\Big[\tilde{\mathbf{y}}_{bs}^{\top}\log p_{bs}(\mathbf{s},\mathbf{h}) \\ + (\mathbf{1}-\mathbf{y}_{bs})^{\top}\log(\mathbf{1}-p_{bs}(\mathbf{s},\mathbf{h}))\Big]\end{aligned} \tag{3}$$

$$\begin{aligned}\mathcal{L}_{ua}\left(p_{ua}(\mathbf{s},\mathbf{h}), \mathbf{y}_{ua}\right) = -\mathbb{E}_{(\mathbf{s},\mathbf{y}_{ua})\sim\mathcal{D}}\Big[\tilde{\mathbf{y}}_{ua}^{\top}\log p_{ua}(\mathbf{s},\mathbf{h}) \\ + (\mathbf{1}-\mathbf{y}_{ua})^{\top}\log(\mathbf{1}-p_{ua}(\mathbf{s},\mathbf{h}))\Big].\end{aligned} \tag{4}$$

Finally, the total loss, $\mathcal{L}$, is a weighted sum of three loss functions defined as $\mathcal{L} = \alpha\mathcal{L}_{pol} + \beta\mathcal{L}_{bs} + \gamma\mathcal{L}_{ua}$. Since all of the predictions made by each network are sparse vectors in the form of multi-label classification, we apply a weighted positive target to each of the binary classification denoted by $\tilde{\mathbf{y}}_a$, $\tilde{\mathbf{y}}_{bs}$ and $\tilde{\mathbf{y}}_{ua}$. The weights are determined roughly based on the ratio between positive examples and negative examples. Furthermore, both of auxiliary network and policy network are trained jointly, and all losses are applied to Eq. (1) to update the policy network. This scenario is seen as making the auxiliary loss to act as a regularizer to the first layer of policy network.

3.2. Reinforcement Learning with Behavior Cloning

In order to achieve optimal result in task-oriented dialogue system, the best policy network weights from behavior cloning are re-used as the pre-trained weights to build a dialogue agent using reinforcement learning algorithms. The proposed BC with auxiliary tasks is employed in RL methods including standard policy gradient (PG), guided dialogue policy learning (GDPL) [4], and proximal policy optimization (PPO) [5]. All RL methods were implemented identically to the models provided by ConvLab-2 repository in order to evaluate the effect of the proposed BC implementation on different RL optimization results. For the reward function in the ConvLab-2 environment, the agent received -1 in every conversation it makes, +5 if current domain is satisfied, and +40 if the task succeeds. Meanwhile, the state and action representations were formed by using the vectorized function, provided by ConvLab-2, that took DST dictionary form as the input. From the experiments, it was found that, in determining the best model from BC learning, directly choosing the model with pre-trained weights that had the lowest validation loss did not work optimally. Instead, it was suggested that the model selection was performed according to the task success rate by using BC agent after applying it to real environment for at least 200 dialogue turns. This tricky condition might be caused by the task complexity in the multi-domain dialogue environment.

4. Experiments

4.1. Experimental Data and Setting

The experiments were conducted using ConvLab-2 framework [12] with MultiWOZ 2.1 dataset [10]. MultiWOZ [30] was a large-scale multi-domain Wizard-of-Oz dataset which was built by simulations of interaction between a computer system and a human user, involving seven domains which were attraction, taxi, restaurant, train, police, hospital, and hotel domain. The number of samples was 10,438 dialogues. This work only considered dialogue policy optimization. The final system was built in the pipeline system where NLU [31, 32], DST and NLG [33] were taken from the predefined modules provided by ConvLab-2. BERT [34] and template module for DST and NLG were used. For system performance evaluation, two different evaluation procedures including policy evaluation and end-to-end system evaluation were performed. For policy evaluation, the performance was evaluated only from task success rate. Basically, task success rate judged if all the domains were successfully completed. Agent counted for success rate if it met the goal from user and simultaneously either user make a book or all requested information are answered by system. Meanwhile, for end-to-end evaluation, the simulated user was involved to evaluate the proposed method in pipeline system using the ConvLab-2 default setting. Code of the implementation is available on https://github.com/NCTUMLlab/Mahdin-Rohmatillah.

4.2. Policy Evaluation

In the policy evaluation, the proposed method, denoted as "BC Aux", was applied by using four different RL algorithms including maximum likelihood estimator (MLE), PG, GDPL and PPO. MLE is the same as baseline BC. Rule policy implies the policy with human-defined rules. The baseline results, which are denoted as "Original", are obtained by using the original ConvLab-2 weights [12]. The upper bound of this evaluation is the performance of rule policy that is seen as a human-defined mapping function from a given state to an action. Table 1 compares task success rates by using different policy evaluations.

Table 1: *Task success rate over different dialogue policies*

Policy		Success Rate
MLE	Original	0.43
	BC Aux	0.54
PG	Original	0.48
	BC Aux	0.55
GDPL [4]	Original	0.48
	BC Aux	0.57
PPO [5]	Original	0.80
	BC Aux	0.94
Rule Policy	Original	0.96

The proposed BC Aux performs better than the original BC. This result indicates that introducing auxiliary tasks helps the agent to identify the components from a state that are important in each conversation turn. Those components eventually reduce causal misidentification of an expert action. As a result, all RL

Table 2: *Evaluation in end-to-end dialogue system. Up arrow symbol (↑) indicates absolute improvement of the policy that uses our BC Aux as pre-trained weights compared to the corresponding policy with original training implementation provided by ConvLab-2.*

Configuration				Average	F1	Complete	Success
NLU	**DST**	**Policy**	**NLG**	**Success Turn**	**Score**	**Rate**	**Rate**
Original ConvLab-2 (Original)							
BERT	Rule	MLE	Template	12.2	67.8	46.7	41.9
BERT	Rule	PG	Template	12.5	67.1	47.6	42
BERT	Rule	GDPL	Template	12.2	67.9	49.9	44
BERT	Rule	PPO	Template	15.3	71.1	64.9	63.8
BERT	Rule	Rule (Upper Bound)	Template	11.6	85.2	92.1	82.7
Behavior Cloning with Auxiliary Tasks (BC Aux)							
BERT	Rule	MLE	Template	11.9 (↑ 0.30)	74.6 (↑ 6.80)	52.9 (↑ 6.20)	47.4 (↑ 5.50)
BERT	Rule	PG	Template	12.0 (↑ 0.50)	73.7 (↑ 6.60)	56.6 (↑ 9.00)	48.3 (↑ 6.30)
BERT	Rule	GDPL	Template	11.3 (↑ 0.90)	72.9 (↑ 5.00)	55.0 (↑ 5.10)	49.0 (↑ 5.00)
BERT	Rule	PPO	Template	12.8 (↑ 2.50)	82.3 (↑ 11.2)	81.7 (↑ 16.8)	77.3 (↑ 13.5)
Perfect	Rule	PPO	Template	13.4 (↑ 1.90)	84.5 (↑ 13.4)	94.5 (↑ 29.6)	90.3 (↑ 26.5)

models that use the model weights of BC Aux as pre-trained weights, show the improved results when compared with the original models. Furthermore, the performance of PPO policy can reach the task success rate which is close to that of rule policy, known as an upper bound in this evaluation.

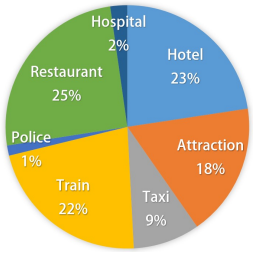

Figure 3: *Domain proportion in end-to-end system evaluation.*

4.3. End-to-End System Evaluation

The end-to-end system evaluation is an important evaluation process in dialogue system. This process sufficiently reflects the real conversations between user and system. Various policy optimizations in RL were applied into a pipeline system to show the improvements resulted by our proposed method. The simulated user pipeline was set to be identical to the ConvLab-2 default setting, which consisted of BERT-based NLU, rule policy and template-based NLG. Each system was evaluated in 1000 dialogues in which complicated domains like hotel, restaurant, attraction and train have more proportions compared to very simple domains like police and hospital as shown by Figure 3. Meanwhile, the overall results of various RL methods with their corresponding pipeline components are reported in Table 2. We can see that the proposed BC with auxiliary tasks improves all RL optimizations significantly in various metrics. These results explain that the introduction of auxiliary tasks can provide useful information to agent to understand the true causes of expert actions in the dataset such that eventually the robust pre-trained BC weights can be obtained to improve the real dialogue conversations using different RL methods. Furthermore, PPO agent that uses BC Aux as pre-trained weights can reach the performance close to rule policy in terms of F1 score and success rate. Detailed performance of PPO using BC Aux method in each domain is shown by Figure 4 as red and blue bars indicate with and without perfect NLU (BERT), respectively. Without perfect NLU, It is found that PPO with BC Aux performs very well only in four domains, attraction, train, hospital and police. The two domain that cannot be handled well are hotel and taxi where hotel is the most complicated domain since it has 12 different slots with various corresponding values and 5 corresponding actions.

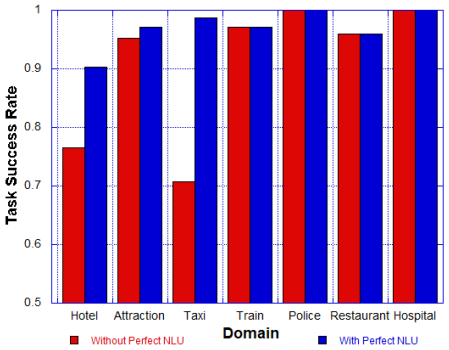

Figure 4: *Success rates over domains without/with perfect NLU.*

Based on our investigation, The moderate performance in hotel and taxi domain is highly affected by the imperfection from NLU component, generating wrong information in DST as a consequence. If the user dialog act output is directly fed to the system DST module, which means that a perfect NLU is given, then the performance is increased significantly as shown in the comparison in Figure 4 as well as Table 2 for perfect NLU in the last row. The highest improvement is shown in taxi domain which improves close to perfect score, as it is relatively not a complex domain, but will degrade our policy performance if it gets wrong information from DST due to defect in NLU. Thus, improvement of NLU part is required to get better result.

5. Conclusions

In this work, a new method that utilized auxiliary tasks to reduce causal confusion in behavior cloning has been proposed. Results both in policy evaluation and end-to-end system evaluation that involved the simulated user showed that this method could improve the standard BC performance and all RL-based optimizations with PPO agent achieved very well performance close to rule policy. However, more investigations are still required to handle multi-domains in complicated environments.

6. References

[1] Z. Zhang, R. Takanobu, Q. Zhu, M. Huang, and X. Zhu, "Recent advances and challenges in task-oriented dialog systems," *Science China Technological Sciences*, pp. 1–17, 2020.

[2] 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, pp. 839–843.

[3] J.-T. Chien and P.-C. Hsu, "Stochastic curiosity exploration for dialogue systems," in *Proc. of Annual Conference of International Speech Communication Association*, 2020, pp. 3885–3889.

[4] 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*, pp. 100–110, 2019.

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

[6] 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*, pp. 625–638, 2020.

[7] J.-T. Chien, W.-L. Liao, and I. El Naqa, "Exploring state transition uncertainty in variational reinforcement learning," in *Proc. of European Signal Processing Conference*, 2020, pp. 1527–1531.

[8] J.-T. Chien and P.-Y. Hung, "Multiple target prediction for deep reinforcement learning," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2020, pp. 1611–1616.

[9] P. de Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[10] 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 preprint arXiv:1907.01669*, 2019.

[11] 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*, pp. 297–310, 2020.

[12] 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*, pp. 142–149, 2020.

[13] M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu, "Reinforcement learning with unsupervised auxiliary tasks," *arXiv preprint arXiv:1611.05397*, 2016.

[14] C. Gelada, S. Kumar, J. Buckman, O. Nachum, and M. G. Bellemare, "DeepMDP: Learning continuous latent space models for representation learning," in *International Conference on Machine Learning*, 2019, pp. 2170–2179.

[15] A. Zhang, R. McAllister, R. Calandra, Y. Gal, and S. Levine, "Learning invariant representations for reinforcement learning without reconstruction," *in Proc. of International Conference on Learning Representations*, 2021.

[16] M. Tomar, A. Zhang, R. Calandra, M. E. Taylor, and J. Pineau, "Model-invariant state abstractions for model-based reinforcement learning," *in Proc. of International Conference on Learning Representations*, 2021.

[17] S. Seneff and J. Polifroni, "Dialogue management in the Mercury flight reservation system," in *ANLP-NAACL Workshop: Conversational Systems*, 2000.

[18] A. Bordes, Y.-L. Boureau, and J. Weston, "Learning end-to-end goal-oriented dialog," *arXiv preprint arXiv:1605.07683*, 2016.

[19] 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, pp. 73–78.

[20] S. Ultes, P. Budzianowski, I. Casanueva, L. M. Rojas-Barahona, B.-H. Tseng, Y.-C. Wu, S. Young, and M. Gašić, "Addressing objects and their relations: The conversational entity dialogue model," in *Proc. of Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2018, pp. 273–283.

[21] M. Gašić, N. Mrkšić, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, D. Vandyke, T.-H. Wen, and S. Young, "Dialogue manager domain adaptation using Gaussian process reinforcement learning," *Computer Speech & Language*, vol. 45, pp. 552–569, 2017.

[22] Z. Li, S. Lee, B. Peng, J. Li, J. Kiseleva, M. de Rijke, S. Shayandeh, and J. Gao, "Guided dialogue policy learning without adversarial learning in the loop," *Proc. of Conference on Empirical Methods in Natural Language Processing*, pp. 2308–2317, 2020.

[23] 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, pp. 9604–9611, 2020.

[24] 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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1208–1218, 2019.

[25] N. Rhinehart, R. McAllister, and S. Levine, "Deep imitative models for flexible inference, planning, and control," *in Proc. of International Conference on Learning Representations*, 2020.

[26] J. Hawke, R. Shen, C. Gurau, S. Sharma, D. Reda, N. Nikolov, P. Mazur, S. Micklethwaite, N. Griffiths, A. Shah, and A. Kndall, "Urban driving with conditional imitation learning," in *Proc. of IEEE International Conference on Robotics and Automation*, 2020, pp. 251–257.

[27] F. Codevilla, E. Santana, A. M. Lopez, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *Proc. of IEEE International Conference on Computer Vision*, 2019, pp. 9329–9338.

[28] J.-T. Chien and Y.-C. Chiu, "Stochastic temporal difference learning for sequence data," in *Proc. of International Joint Conference on Neural Networks*, 2021.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[30] 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, pp. 5016–5026.

[31] S. Watanabe and J.-T. Chien, *Bayesian speech and language processing*. Cambridge University Press, 2015.

[32] J.-T. Chien, "Deep Bayesian natural language processing," in *Proc. of Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2019, pp. 25–30.

[33] 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, pp. 7778–7782.

[34] 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*, pp. 4171–4186, 2019.