Dialog Action-Aware Transformer for Dialog Policy Learning

Huimin Wang[1*], **Wai-Chung Kwan**[2,3*], **Kam-Fai Wong**[2,3]
[1]Jarvis Lab, Tencent, Shenzhen, China
[2]The Chinese University of Hong Kong, Hong Kong, China
[3]MoE Key Laboratory of High Confidence Software Technologies, China
{hmmmwang}@tencent.com
{wckwan,kfwong}@se.cuhk.edu.hk

Abstract

Recent works usually address Dialog policy learning DPL by training a reinforcement learning (RL) agent to determine the best dialog action. However, existing works on deep RL require a large volume of agent-user interactions to achieve acceptable performance. In this paper, we propose to make full use of the plain text knowledge from the pre-trained language model to accelerate the RL agent's learning speed. Specifically, we design a dialog action-aware transformer encoder (DaTrans), which integrates a new fine-tuning procedure named masked last action task to encourage DaTrans to be dialog-aware and distils action-specific features. Then, DaTrans is further optimized in an RL setting with ongoing interactions and evolves through exploration in the dialog action space toward maximizing long-term accumulated rewards. The effectiveness and efficiency of the proposed model are demonstrated with both simulator evaluation and human evaluation.

1 Introduction

A task-oriented dialog system that can serve users on certain tasks has increasingly attracted research efforts. Dialog policy learning (DPL) aiming to determine the next abstracted system output plays a key role in pipeline task-oriented dialog systems (Kwan et al., 2023). Recently, it has shown great potential for using reinforcement learning (RL) based methods to formulate DPL (Young et al., 2013; Su et al., 2016; Peng et al., 2017). A lot of progress is being made in demonstration-based efficient learning methods (Brys et al., 2015; Cederborg et al., 2015; Wang et al., 2020; Li et al., 2020; Jhunjhunwala et al., 2020; Geishauser et al., 2022). Among these methods, dialog state tracking (DST), comprising all information required to determine the response, is an indispensable module. However, DST inevitably accumulates errors from each module of the system.

Recent pre-trained language models (PLMs) gathering knowledge from the massive plain text show great potential for formulating DPL without DST. Recently, the studies on PLMs for dialog, including BERT-based dialog state tracking (Gulyaev et al., 2020) and GPT-2 based dialog generation (Peng et al., 2020; Yang et al., 2021) are not centred on DPL. To this end, we proposed the **D**ialog **A**ction-oriented transformer encoder termed as **DaTrans**, for efficient dialog policy training. DaTrans is achieved by a dialog act-aware fine-tuning task, which encourages the model to distil the dialog policy logic. Specifically, rather than commonly used tasks, like predicting randomly masked words in the input (MLM task) and classifying whether the sentences are continuous or not (NSP task) (Devlin et al., 2019), DaTrans is fine-tuned by predicting the masked last acts in the input action sequences (termed as MLA task). After that, DaTrans works as an RL agent which evolves toward maximizing long-term accumulated rewards through interacting with a user simulator. Following the traditional RL-based dialog policy learning framework, the main novelty of DaTrans is that it integrates a proposed dialog action-aware fine-tuning task (MLA), which helps to extract action-specific features from historical dialog action sequences to improve dialog policy learning. The empirical results prove the excellent performance of DaTrans. Our main contributions include 1) We propose the DaTrans that integrates the dialog act-aware fine-tuning task to extract the dialog policy logic from the plain text; 2) We validate the efficiency and effectiveness of the proposed model on a multi-domain benchmark with both simulator and human evaluation.

[*] Equal Contribution

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 142–148
September 11–15, 2023. ©2023 Association for Computational Linguistics

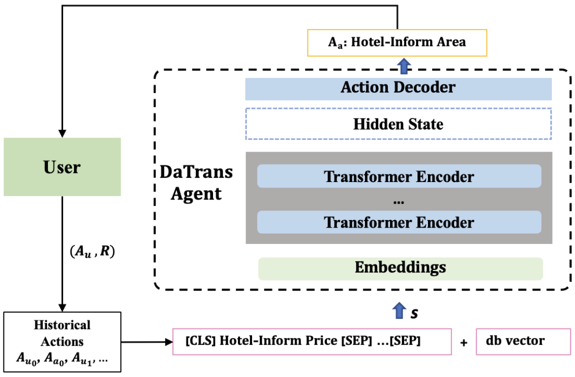


Figure 1: The Illustration of **D**ialog **A**ction-oriented Transformer Encoder (**DaTrans**). In this example, **DaTrans** generates the dialog action A_a based on historical actions.

2 Approach

We cast the dialog policy learning problem as a Markov Decision Process and optimize the policy with deep reinforcement learning approaches. RL usually involves an interactive process (as shown in Figure 1), during which the dialog agent's behavior should choose actions that tend to increase the long-turn sum of rewards given by the user. It can learn to do this over time, by systematic trials and errors until reaches the optimal. In our setting, the dialog agent is encoded with the proposed DaTrans, which perceives the state s and determines the next action A_a. We consider a transformer decoder-based policy model, which takes text concatenating of tuples containing a domain name, an intent type, and slot names as input and determines the next action.

2.1 DaTrans

We apply Deep Q-learning to optimize dialog policy. $Q_\theta(s, a)$, approximating the state-action value function parameterized θ, is implemented based on DaTrans as illustrated in Figure 1. In each turn, perceiving the state s that consists of historical action sequences and a database vector denoting the matches of the current constraints, DaTrans determines the dialog action a with the generated value function $Q_\theta(\cdot|s)$. Historical action sequences are tokenized started from $[CLS]$, followed by the tokenized actions separated and ended with $[SEP]$. Then the transformer encoder gets the final hidden states denoted $[t_0..t_n] = encoder([e_0..e_n])$ (n is the current sequence length, e_i is the embedding of the input token). The contextualized sentence-level representation t_0, is passed to a linear layer named action decoder $\boldsymbol{T}$ to generate:

$$Q_\theta(s, a) = \boldsymbol{T}_a(encoder(Embed(s))) \qquad (1)$$

where $Embed$ is the embedding modules of transformer encoder, $\boldsymbol{T}_a$ denoted the a_{th} output unit of $\boldsymbol{T}$. Based on DaTrans, the dialog policy is trained with ϵ-greedy exploration that selects a random action with probability ϵ, or adopts a greedy policy $a = argmax_{a'} Q_\theta(s, a')$. In each iteration, $Q_\theta(s, a)$ is updated by minimizing the following square loss with stochastic gradient descent:

$$\begin{aligned} \mathcal{L}_\theta &= \mathbb{E}_{(s,a,r,s')\sim D}[(y_i - Q_\theta(s, a))^2] \\ y_i &= r + \gamma \max_{a'} Q'_\theta(s', a') \end{aligned} \qquad (2)$$

where $\gamma \in [0, 1]$ is a discount factor, D is the experience replay buffer with collected transition tuples (s, a, r, s'), s is the current state, r refers to the reward, and $Q'(\cdot)$ is the target value function, which is only periodically updated, and s' is the next state. By differentiating the loss function with regard to θ, we derive the following gradient:

$$\begin{aligned} \nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')\sim D}[(r+ \\ \gamma max_{a'} Q'_{\theta'}(s', a') - Q_\theta(s, a)) \nabla_\theta Q_\theta(s, a)] \end{aligned} \qquad (3)$$

In each iteration, we update $Q(.)$ using minibatch Deep Q-learning.

2.2 Dialog Action-aware Fine-tuning

A vanilla transformer decoder without pretraining can encumber the learning of dialog policy since it is totally unaware of the text and dialog logic. Meanwhile, well-pre-trained models like BERT, due to the generality of pretraining tasks and corpus, are still difficult with competent in dialog modeling. The NSP task encourages BERT to model the relationship between sentences, which may benefit natural language inference, however, biased dialog policy learning due to the inconsistency between success and continuity of sentences, e.g. discontinuous sentences can form a successful dialog. Also, the MLM task allows the word representation to fuse the left and right context, while the dialog agent is only allowed to access the left one. Considering that the ability to reason the next dialog action plays a key role in dialog policy, we replace the MLM and NSP task with a novel fine-tuning task: predicting masked last dialog action (MLA). MLA is based on a dialog action-aware fine-tuning corpus, each piece of which is a dialog session composed of the annotated historical action sequences, for example, "*[CLS] Police-Inform Name [SEP] Police-Inform Phone Addr Post [SEP] general-thank none [SEP]*", (denoted as **sentence A**). Then we randomly cut between two consecutive actions of a session, and select the first half with the masked last act as input. For example, we cut **sentence A** between the 2_{nd} and the 3_{rd} action, and mask the last act to get the input: "*[CLS] Police-Inform Name [SEP] [MASK]..[MASK]*". The label for the masked tokens is "*Police - Inform Phone Addr Post*". Significantly, the proposed MLA task for BERT is actually different from auto-regression. The way auto-regression works is after each token is produced, that token is added to the sequence of inputs and this new sequence becomes the input to the model in its next step. However, in DaTrans, the MLA task works as predicting the last dialog action word by word without adding a new predicted word.

The goal of MLA is to minimize the cross-entropy loss with input tokens $w_0, w_1, .., w_n$:

$$\mathcal{L}^{mla} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=n-k+1}^{n} \log \boldsymbol{p}(w_j^i | w_{0:j-1,j+1:n}^i) \tag{4}$$

where $w_{0:j-1,j+1:n}^i = w_0^i \cdots w_{j-1}^i, w_{j+1}^i..w_n^i$, $\boldsymbol{p}$ is the action decoder head for predicting masked tokens. $w_j^i \in \{0 \cdots v-1\}$ is the label for the masked token, v is the required vocabulary size, and m is the number of dialog sessions. Besides, n and k are the length of the input and masked action sequences, respectively.

3 Experiments and Results

We first conduct the simulator evaluation to assess the DaTrans' performance of learning efficiency, the robustness of fine-tuning Corpus, and domain adaptation. Besides, the case study and human evaluation are conducted and the results are presented in Section D & E in Appendix. In our experiment, NLU and NLG modules are ignored since the interactions are made with dialog actions. Notably, DaTrans can be equipped with any NLU and NLG models. Two datasets, MultiWoz (Budzianowski et al., 2018) and Schema-Guided dialog (SGD) (Rastogi et al., 2020) are involved. We leverage a public available agenda-based user simulator (Zhu et al., 2020) setup on MultiWoz. The details of the dataset, implementation, and the user simulator are illustrated in the Appendix.

3.1 Baseline Agents

We compare the performance of the proposed DaTrans with the state-of-art model JOIE (Wang and Wong, 2021), vanilla BERT, and its variants of different optimization and fine-tuning settings. [1] **DQN** agent is trained with a deep Q-Network. **BERT** agent is equipped with BERT as the encoder that replaces the fully connected layer in DQN. **BERT$_{\texttt{MWoz}}$** agent is with BERT pre-trained with MLM and NSP tasks on MultiWoz. **JOIE** agent (Wang and Wong, 2021) is a collaborative multi-agent framework factoring the joint action space and learning each part by a different agent. **DaTrans$_{\texttt{MWoz}}$** is our proposed agent that is pre-trained with MLA task as described in Section 3.1 on MultiWoz dataset.

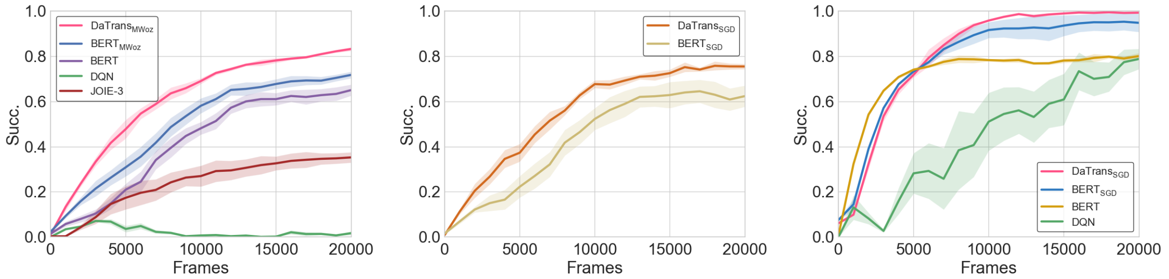
Table 1: The simulation performance of different agents. Succ. denotes the final success rate, Turn and Reward are the average turn and the average reward of the whole training process, respectively.

Model	Succ.↑	Turn↓	Reward↑
DaTrans$_{\texttt{MWoz}}$	**0.84**	**10.21**	**27.35**
BERT$_{\texttt{MWoz}}$	0.72	12.14	14.21
BERT	0.64	14.75	-15.47
DQN	0.01	19.51	-53.66
JOIE-3	0.38	15.98	-21.42

3.2 Simulator Evaluation

All agents are evaluated with the success rate (Succ.) at the end of the training, average turn (Turn), average reward (Reward). The main simulation results are shown in Table. 1 and Figure 2(a). The results indicate that the proposed DaTrans$_{\texttt{MWoz}}$ learns faster and achieves

[1] "optimization" refers to the interactive training process with Reinforcement Learning. "pre-train" means the process of PLMs trained with massive plain text. Besides, we use both "pre-train" and "fine-tuning" to refer to the self-supervised training process of BERT with annotated historical action sequences.



(a) Main results.

(b) Domain adaptation.

(c) Altering fine-tuning corpus.

Figure 2: Comparison of the success rate evolving during the training process.

a better convergence in in-domain evaluation. DaTrans$_{\texttt{MWoz}}$, pre-trained with the mask last act task (MLA) on the MultiWoz corpus achieves the best Succ. (on average 0.84) with the highest learning efficiency in BERT-based models. The performance of DaTrans$_{\texttt{MWoz}}$ reveals that our MLA pre-training task can not only encode the characteristics of dialog policy for efficiency improvement but also show better transfer abilities because dropping it BERT$_{\texttt{MWoz}}$ degrades the performance of DaTrans$_{\texttt{MWoz}}$. Additionally, BERT is consistently the worst in BERT-based models, which is not surprising since it is only initialized with official BERT's pre-trained weights without in-domain fine-tuning. The generality of fine-tuning corpus and task, domain awareness, and knowledge transferability of BERT are poor. Furthermore, without any fine-tuning, JOIE and DQN are worse than BERT-based agents. Finally, the comparison results of Turn and Reward are illustrated in Table. 1. It depicts that DaTrans$_{\texttt{MWoz}}$ achieves the shortest average turn and highest average reward, which is consistent with the learning curves in Figure 2(a).

Effect of fine-tuning Corpus. We further test the effect of different fine-tuning corpus on the performance. The models are pre-trained on SGD and optimized on MultiWoz to investigate the influence of fine-tuning corpus. We denote DaTrans$_{\texttt{SGD}}$ as a variant of DaTrans which is pre-trained on SGD and optimized on MultiWoz. We only compared the results of fine-tuning on SGD, because the agents who have fine-tuned on MultiWoz have seen the dialogue logic of MultiWoz, so it is of little significance to optimize the comparison on MultiWoz. Besides, we don't optimize the models with RL on SGD because we didn't find an open-source simulator for SGD. Thus, we only take SGD to explore the effect of corpus and domain adaptation. The core conclusion indicated from Figure 2(b) is that DaTrans is robust to the different fine-tuning corpus. Firstly, the proposed MLA pre-training task does better in extracting the knowledge of dialog action sequence, especially the structure information that is invariant over domains. As a consequence, DaTrans$_{\texttt{SGD}}$ outperforming BERT$_{\texttt{SGD}}$.

Domain Adaptation. To assess the ability for new task adaptation, we compare the agents that continually learn a new domain Restaurant, starting from being well-trained on the other six domains (i.e. Train, Hotel, Hospital, Taxi, Police, Attraction). Figure 2(c) shows the performances of new task adaptation for dialog policy learning. The results confirm that DaTrans pre-trained with masked last action task is capable of quickly adapting to the new environment compared to DaTrans$_{\texttt{SGD}}$ and BERT$_{\texttt{SGD}}$. Besides, pre-training counts because removing it (BERT) damages the results.

4 Conclusion and Future Work

In this paper, we investigate the pre-trained language model enhancing the reinforcement learning agent for dialog policy learning. We propose DaTrans, which is equipped with a new fine-tuning task that masks the last dialog action to extract the dialog logic for efficient dialog policy learning. The evaluation results show the effectiveness of the proposed DaTrans in terms of learning efficiency and domain adaptation ability.

Limitations

Due to the high cost of interactions with human users, the dialog policy model was trained in a simulated environment rather than real-world scenarios. Our approach is able to construct a highly responsive dialog system because it shortens the required interaction turns, and reduces labour costs associated with interactive training with human users. However, it is worth noting that the model optimized in our experiments may not be suitable for dealing with real-world users, thus simulation evaluation results alone are not sufficient to prove DaTrans's superiority. Despite this limitation, as there are few studies dedicated to investigating PLMs advanced dialog policy learning, We hope that DaTrans will inspire further research in this field in the future.

Acknowledgements

We appreciate the constructive and insightful comments provided by the anonymous reviewers. This research work is partially supported by CUHK under Project No. 3230377.

References

Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E Taylor, and Ann Nowé. 2015. Reinforcement learning from demonstration through shaping. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Thomas Cederborg, Ishaan Grover, Charles L Isbell, and Andrea L Thomaz. 2015. Policy shaping with human teachers. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christian Geishauser, Carel van Niekerk, Hsien-Chin Lin, Nurul Lubis, Michael Heck, Shutong Feng, and Milica Gasic. 2022. Dynamic dialogue policy for continual reinforcement learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 266–284.

Pavel Gulyaev, Eugenia Elistratova, Vasily Konovalov, Yuri Kuratov, Leonid Pugachev, and Mikhail Burtsev. 2020. Goal-oriented multi-task bert-based dialogue state tracker. *arXiv preprint arXiv:2002.02450*.

Megha Jhunjhunwala, Caleb Bryant, and Pararth Shah. 2020. Multi-action dialog policy learning with interactive human teaching. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 290–296.

Wai-Chung Kwan, Hong-Ru Wang, Hui-Min Wang, and Kam-Fai Wong. 2023. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intelligence Research*, 20(3):318–334.

Sungjin Lee, Qi Zhu, Ryuichi Takanobu, Zheng Zhang, Yaoqin Zhang, Xiang Li, Jinchao Li, Baolin Peng, Xiujun Li, Minlie Huang, and Jianfeng Gao. 2019. ConvLab: Multi-domain end-to-end dialog system platform. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 64–69, Florence, Italy. Association for Computational Linguistics.

Ziming Li, Julia Kiseleva, and Maarten de Rijke. 2020. Rethinking supervised learning and reinforcement learning in task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3537–3546, Online. Association for Computational Linguistics.

Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2231–2240, Copenhagen, Denmark. Association for Computational Linguistics.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 172–182, Online. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696. Number: 05.

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Continuously learning neural dialogue management. *arXiv preprint arXiv:1606.02689*.

Huimin Wang, Baolin Peng, and Kam-Fai Wong. 2020. Learning efficient dialogue policy from demonstrations through shaping. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6355–6365.

Huimin Wang and Kam-Fai Wong. 2021. A collaborative multi-agent reinforcement learning framework for dialog action decomposition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7882–7889.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14230–14238.

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.

Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang. 2020. ConvLab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 142–149, Online. Association for Computational Linguistics.

A Dataset

Two datasets are involved: 1) MultiWoz (Budzianowski et al., 2018), a large-scale fully annotated corpus of human-human conversations; 2) Schema-Guided dialog (SGD) (Rastogi et al., 2020), multi-domain, task-oriented conversations between a human and a virtual assistant. MultiWOz contains 8,434 pieces of corpus covering 9 domains, while SGD consists of 16,142 pieces of dialog sessions involving 16 domains.

B Implementation Details.

We adopt BERT$_{base}$ (uncased) with default hyperparameters in Huggingface Transformers (Wolf et al., 2020) as the backbone transformer encoder model. We pre-train and optimize BERT-based models on one RTX 2080Ti GPU and GTX TITAN X. The pre-training batch size is 8. The learning rate for the BERT-based model is 0.00003. The action decoder of DaTrans is a linear layer with 400 output units corresponding to the 400 action candidates. Meanwhile, we set the discount factor γ as 0.9. Besides, we apply the rule-based agent from ConvLab (Lee et al., 2019) to warm start the policy with 1000 dialog epochs.

C User Simulator

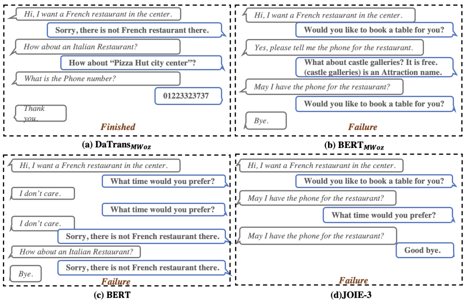
We leverage a public available agenda-based user simulator (Zhu et al., 2020) setup on MultiWoz. During training, the simulator initializes with a user goal and takes a system action as input and outputs the user action with a reward. The reward is set as -1 for each turn to encourage short turns and a positive reward $(2 \cdot T)$ for successful dialog or a negative reward of $-T$ for failed one, where T (set as 40) is the maximum number of turns in each dialog. A dialog is considered successful only if the agent helps the user simulator accomplish the goal and satisfies all the user's search constraints.

D Human Evaluation

We further conduct a human evaluation to validate the simulation results. We choose the agents trained with 10000 epochs. Before the test, all evaluators are instructed to interact with the agents to achieve their goals. In each session, a randomly selected goal and a random agent are assigned to a user. They can terminate the dialog if they think the session is doomed to fail. At the end of each session, the user is required to judge if the dialog is a success or a failure. We collect 50 conversations for each agent. The results are illustrated in Table. 2. We see that the human evaluation results further convince the simulator evaluation.

Table 2: The Human performance of different agents. The evaluation is conducted at 10000 epochs in Figure 2(a) for all agents. Succ. denotes success rate.

Model	Succ.↑
DaTrans$_{\texttt{MWoz}}$	0.68
BERT$_{\texttt{MWoz}}$	0.58
BERT	0.46
DQN	0.00
JOIE-3	0.24


Hi, I want a French restaurant in the center.
Sorry, there is not French restaurant there.
How about an Italian Restaurant?
How about "Pizza Hut city center"?
What is the Phone number?
01223323737
Thank you.
Finished
Hi, I want a French restaurant in the center.
Would you like to book a table for you?
Yes, please tell me the phone for the restaurant.
What about castle galleries? It is free. (castle galleries) is an Attraction name.
May I have the phone for the restaurant?
Would you like to book a table for you?
Bye.
Failure
Hi, I want a French restaurant in the center.
What time would you prefer?
I don't care.
What time would you prefer?
I don't care.
Sorry, there is not French restaurant there.
How about an Italian Restaurant?
Sorry, there is not French restaurant there.
Bye.
Failure
Hi, I want a French restaurant in the center.
Would you like to book a table for you?
May I have the phone for the restaurant?
What time would you prefer?
May I have the phone for the restaurant?
Good bye.
Failure

(a) DaTrans$_{MWoz}$ (b) BERT$_{MWoz}$ (c) BERT (d)JOIE-3

Figure 3: Sampled dialogue examples generated by DaTrans$_{\texttt{MWoz}}$, BERT$_{\texttt{MWoz}}$, BERT, DQN, JOIE3. The grey boxes convey the queries from the users while the blue boxes are the responses from the agents. At the bottom of the boxes, we marked whether the session is successful or not.

E Case Study

To further explore the performance of the agents after training, we randomly sampled some real examples generated for a shared restaurant goal. From the samples placed in Fig. 3, some explicable clues are found. In this example, BERT$_{\texttt{MWoz}}$ fails because it makes mistakes in the restaurant's dialogue logic though it recognizes the right domain. Besides, the response involving "castle galleries" indicates BERT$_{\texttt{MWoz}}$ suffers from disturbance from other task Attraction. As for BERT and JOIE3, it seems that the knowledge regarding restaurant has not been mastered. Only DaTrans$_{\texttt{MWoz}}$ systematically handles the issues by taking reasonable actions.