





# **Deep Reinforcement Learning-based Dialogue Policy with Graph Convolutional Q-network**

Presenter: KaiXu

Email: [sekxu@mail.scut.edu.cn](mailto:sekxu@mail.scut.edu.cn)

# CONTENTS

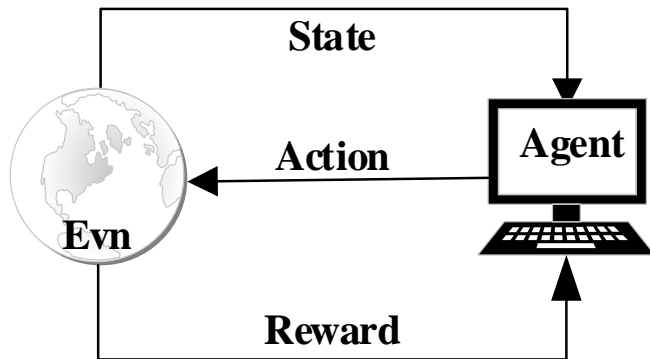
-  1. Reinforcement learning and dialogue policy
-  2. Limitations of existing DRL-based dialogue policy
-  3. Our Proposed method
-  4. Experimental results

CONTENT

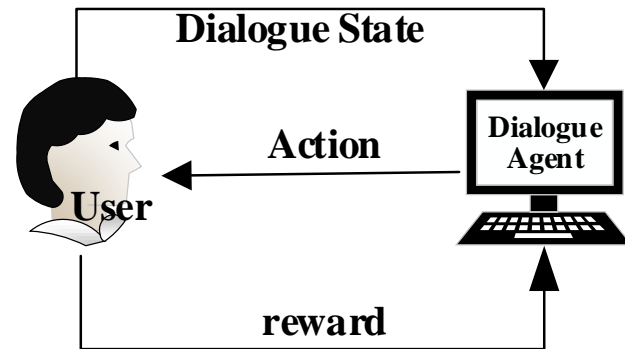
# 1. Reinforcement learning and dialogue policy

**Dialogue policy:** dialogue policy is a key component of a task-oriented dialogue systems, which select dialogue based on a given dialogue state. dialogue policy is mostly modelled through **Deep Reinforcement Learning (DRL)**, which is able to dynamically adjust the policy from real-time environmental (user) feedback.

## Reinforcement learning



## Dialogue policy modeled via RL



- Reinforcement learning dialogue policies view the user as an environment, model the policy between the dialogue agent and the user, receive dialogue state, and output dialogue action.
- DRL-based dialogue policies are better for maintaining large-scale dialogue state spaces and being robust to noise.

## 2. Limitations of existing DRL-based dialogue policy

The task-oriented dialogue policy based on DRL remains some limitations.

### 1. **discretely represent the states/actions**

the underlying relationships (e.g., behavioral similarities) between different states (or sub-states) are not effectively explored and exploited.

### 2. **require a high number of iterations to explore valuable state information**

DRL-based dialog policies conducted by trial-and-error, which require a large number interaction. Accelerating the learning of dialog policies through expert experiences is both a financially costly and laborious task. Some methods combine with supervised learning are often not available for a very large number of state spaces.

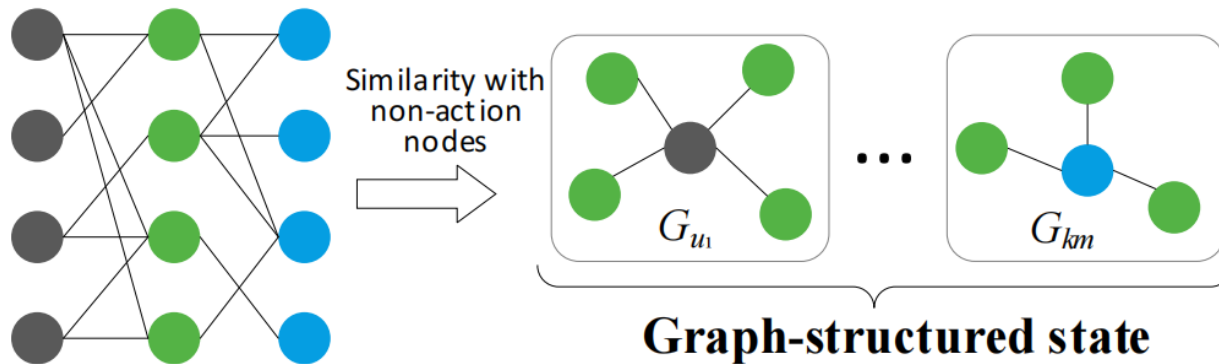
### 3. **Traditional DRL-based dialogue policy method are not sample-efficient.**

The dialogue agent training with limited samples does not achieve a satisfactory performance, and it is also expensive and time-consuming to obtain data by interacting with real users.

### 3. Our Proposed method

propose a universal DPL framework with a graph-structured dialogue state. It based on a Graph convolutional Q-network to learn graph-Structured information for modeling Dialogue Policy, GSDP.

**Step1: Construct a graph-structured state**  $\longrightarrow$  **Step2: Dialogue policy learning**



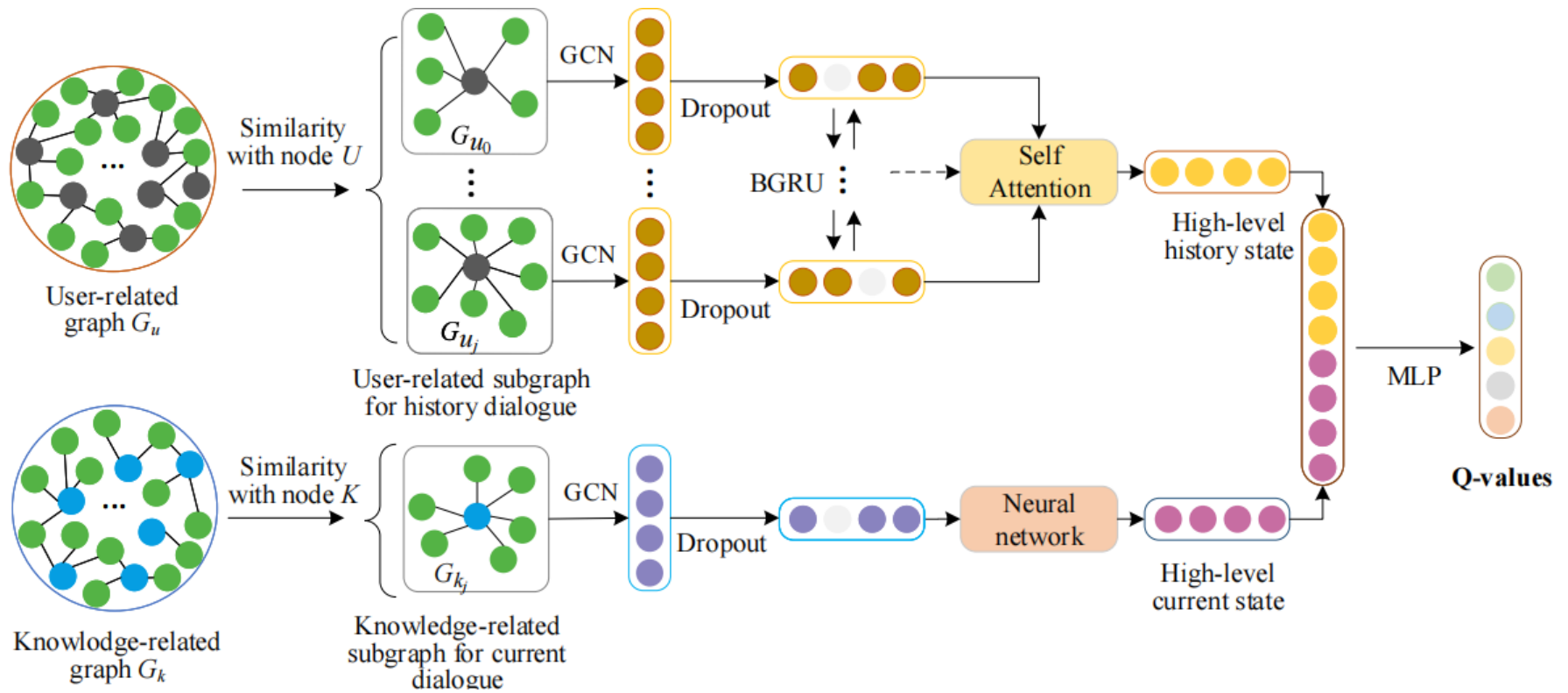
The gray circle  $\bullet$  is the user-related node. The green circle  $\bullet$  is the action node.

The blue circle  $\bullet$  is the knowledge related node.

We construct graph-related states by baseline reinforcement learning without introducing any expert experience. During dialogue, we build graph-structured states based on the similarity of non-action nodes.

### 3. Our Proposed method

#### Step2: Dialogue policy learning



We take into account the relationship of dialogue states and actions, and the user's historical feature information. This is our advantages.

## 4. Experimental results

### Performance of GSDP with different top-N similarity

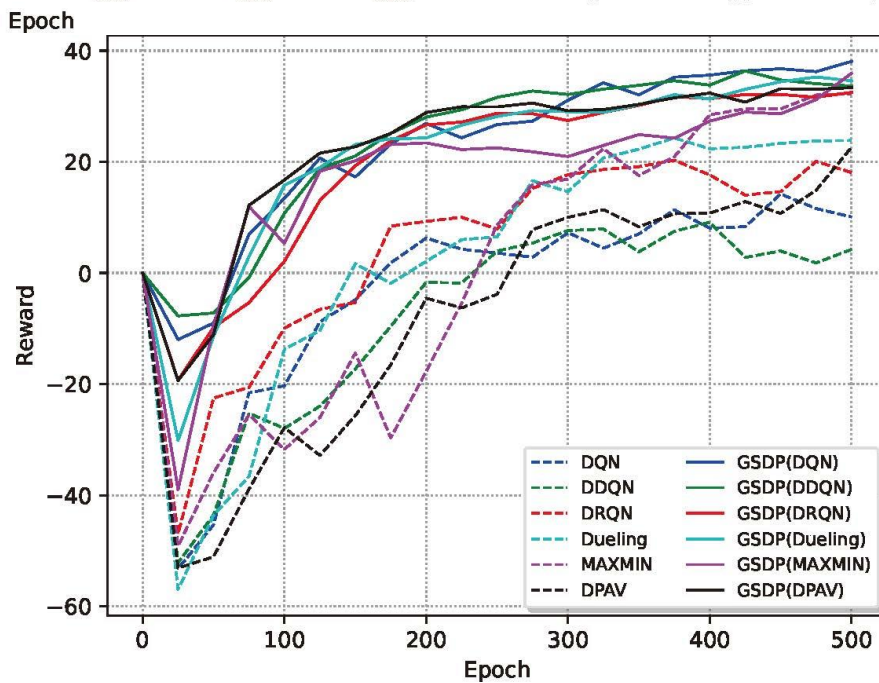
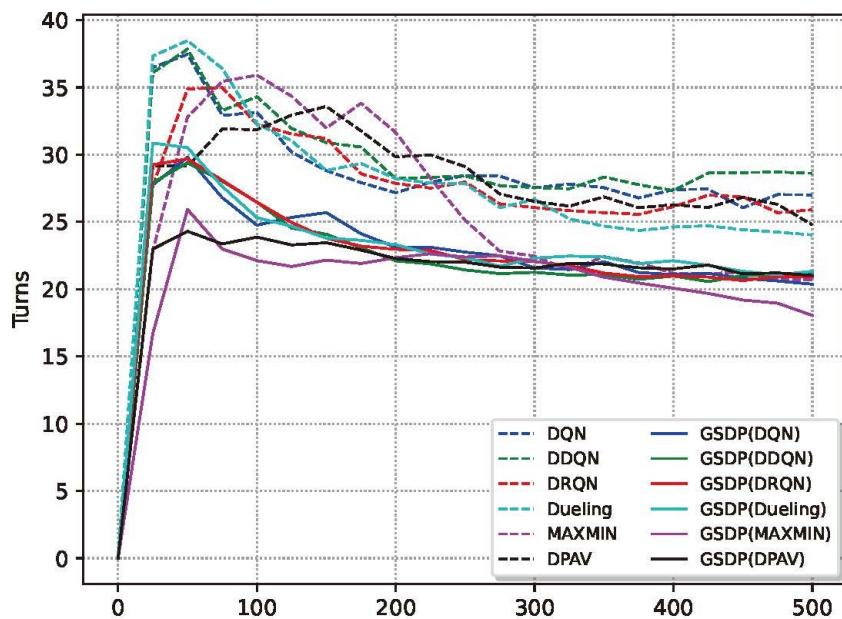
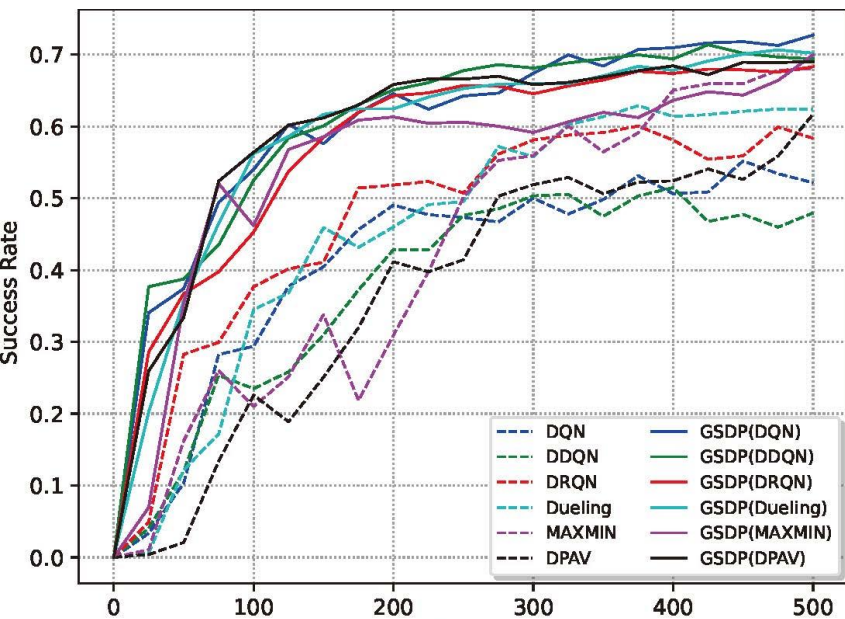
	top-1	top-2	top-3	top-4	top-5	top-6	top-7	top-8	top-9	top-10
GSDP(DQN)	0.661 $\pm 0.042$	<b>0.727</b> $\pm \mathbf{0.038}$	0.682 $\pm 0.041$	0.670 $\pm 0.053$	<u>0.683</u> $\pm \mathbf{0.040}$	0.651 $\pm 0.044$	0.661 $\pm 0.042$	0.678 $\pm 0.043$	0.662 $\pm 0.056$	0.664 $\pm 0.053$
GSDP(DDQN)	0.644 $\pm 0.034$	0.659 $\pm 0.031$	<b>0.694</b> $\pm \mathbf{0.030}$	0.675 $\pm 0.034$	0.683 $\pm 0.053$	0.657 $\pm 0.029$	0.670 $\pm 0.038$	0.668 $\pm 0.043$	0.662 $\pm 0.048$	<u>0.684</u> $\pm \mathbf{0.051}$
GSDP(DRQN)	0.640 $\pm 0.034$	<u>0.688</u> $\pm \mathbf{0.022}$	0.683 $\pm 0.032$	0.679 $\pm 0.028$	0.687 $\pm 0.044$	0.651 $\pm 0.036$	0.686 $\pm 0.026$	0.67 $\pm 0.039$	0.643 $\pm 0.041$	<b>0.704</b> $\pm \mathbf{0.035}$
GSDP(Dueling)	0.635 $\pm 0.035$	<u>0.702</u> $\pm \mathbf{0.029}$	0.689 $\pm 0.031$	0.699 $\pm 0.027$	0.692 $\pm 0.032$	0.669 $\pm 0.033$	0.668 $\pm 0.038$	<b>0.703</b> $\pm \mathbf{0.036}$	0.692 $\pm 0.039$	0.683 $\pm 0.045$
GSDP(MAXMIN)	<u>0.648</u> $\pm \mathbf{0.027}$	0.602 $\pm 0.031$	0.630 $\pm 0.055$	<b>0.700</b> $\pm \mathbf{0.042}$	0.614 $\pm 0.059$	0.614 $\pm 0.059$	0.595 $\pm 0.036$	0.606 $\pm 0.037$	0.620 $\pm 0.042$	0.604 $\pm 0.048$
GSDP(DPAV)	0.665 $\pm 0.040$	0.632 $\pm 0.040$	<b>0.690</b> $\pm \mathbf{0.033}$	0.644 $\pm 0.034$	<u>0.668</u> $\pm \mathbf{0.033}$	0.653 $\pm 0.033$	0.653 $\pm 0.035$	0.667 $\pm 0.033$	0.661 $\pm 0.028$	0.623 $\pm 0.038$

Table 1: The dialogue success rate of GSDP with Top-N similarity under different models. The bolded font indicates the best results and the underline indicates the second best results.

The GSDP(DQN) obtained the highest dialog success rate with Top-2 similarity, and its performance is better compared with the other models, because GSDP (DQN) has more reliable actions with top- 2 similarity of bipartite graph non-action nodes

# 4. Experimental results

## Main Results



Models	Epoch 200			Epoch 500		
	<i>Suc</i>	<i>Rew</i>	<i>Tur</i>	<i>Suc</i>	<i>Rew</i>	<i>Tur</i>
DQN	0.491 ± 0.121	6.28 ± 16.02	27.17 ± 3.23	0.522 ± 0.089	10.09 ± 11.79	27.01 ± 2.40
DDQN	0.428 ± 0.164	-1.69 ± 21.23	28.22 ± 3.37	0.480 ± 0.094	4.24 ± 12.44	28.61 ± 2.51
DRQN	0.518 ± 0.099	9.28 ± 12.80	27.86 ± 2.29	0.584 ± 0.077	18.07 ± 10.20	25.90 ± 2.15
Dueling	0.460 ± 0.112	2.09 ± 14.82	28.22 ± 3.05	0.624 ± 0.067	23.85 ± 8.97	24.03 ± 2.01
MAXMIN	0.309 ± 0.138	-17.71 ± 17.93	31.65 ± 3.05	0.680 ± 0.054	32.28 ± 7.10	20.72 ± 1.56
DPAV	0.411 ± 0.115	-4.54 ± 14.57	29.84 ± 1.93	0.617 ± 0.077	22.63 ± 10.21	24.78 ± 2.19
GSDP(DQN)	<b>0.646 ± 0.031</b>	<b>26.91 ± 4.16</b>	<b>23.12 ± 1.07</b>	<b>0.727 ± 0.038</b>	<b>38.08 ± 6.17</b>	<b>20.37 ± 1.42</b>
GSDP(DDQN)	<b>0.651 ± 0.062</b>	<b>28.05 ± 8.04</b>	<b>22.11 ± 1.99</b>	<b>0.694 ± 0.030</b>	<b>33.65 ± 7.94</b>	<b>21.28 ± 1.41</b>
GSDP(DRQN)	<b>0.643 ± 0.056</b>	<b>26.65 ± 7.26</b>	<b>22.96 ± 1.83</b>	<b>0.688 ± 0.022</b>	<b>32.50 ± 7.47</b>	<b>20.94 ± 1.66</b>
GSDP(Dueling)	<b>0.591 ± 0.116</b>	<b>24.30 ± 7.98</b>	<b>23.31 ± 1.69</b>	<b>0.702 ± 0.029</b>	<b>34.59 ± 7.64</b>	<b>21.32 ± 1.88</b>
GSDP(MAXMIN)	<b>0.613 ± 0.023</b>	<b>23.42 ± 3.15</b>	<b>22.34 ± 1.21</b>	<b>0.700 ± 0.042</b>	<b>35.93 ± 5.32</b>	<b>18.05 ± 0.94</b>
GSDP(DPAV)	<b>0.658 ± 0.056</b>	<b>28.87 ± 7.59</b>	<b>22.28 ± 2.03</b>	<b>0.690 ± 0.033</b>	<b>33.34 ± 6.02</b>	<b>21.02 ± 1.43</b>

Table 2: Display of detailed experimental results. The bolded font indicates better results than its baseline model.

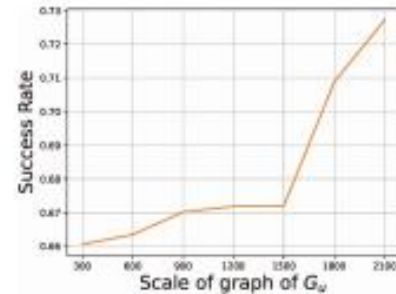
GSDP has lower variance in both the *Tur* and the *Rew*, which demonstrates the stability of GSDP.

## 4. Experimental results

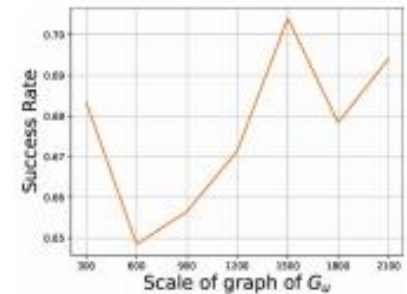
### Sub-graphs with different Sizes

We verify the influence of the size of the subgraph to the performance of proposed GSDP framework.

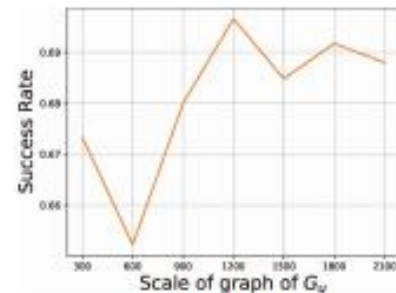
There is a positive correlation between  $G_u$  graph size and dialogue success rate. The larger of the size  $G_u$  have a higher dialogue success rate in general. The results of the dialogue success rate fluctuates with the sizes of  $G_u$  because the state encoding is based on the similarity to non-action nodes of  $G_u$ , whereas the experiments are randomly selected subgraphs of  $G_u$  to calculate the similarity.



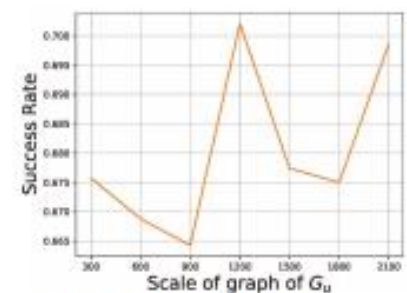
(a) GSDP(DQN)



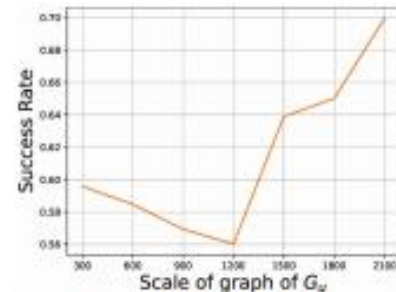
(b) GSDP(DDQN)



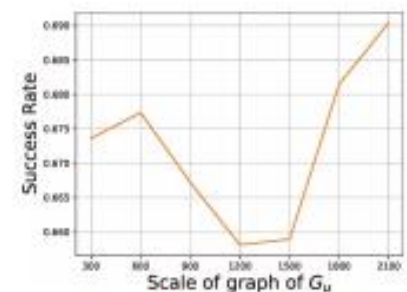
(c) GSDP(DRQN)



(d) GSDP(Dueling)



(e) GSDP(MAXMIN)



(f) GSDP(DPAV)

Figure 4: The dialogue success rate of GSDP with different Sizes of  $G_u$ .

## 4. Experimental results

### **graph-structured framework for learning dialogue policy. Contributions:**

1. This paper proposed a generic framework for DPL that predicts action values through bipartite graphs. The bipartite graphs are built via successful dialogs from baseline reinforcement learning algorithms and do not require expert experience.
2. We use similarity to generate dialogue subgraphs and construct a variant of graph neural network to extract information from different subgraphs and output graph-aware embeddings. The BGRU and self-attention are employed to process the graph-aware embeddings for more dialogue features.
3. Our experimental evaluations show that the GSDP framework is more robust, more efficient, and has high scalability compared to different DRL algorithms, performing excellently in the dataset of movie ticket booking.

**Thank you for your  
attention!**

presenter

**KaiXu**