# Efficient Dialogue Complementary Policy Learning via Deep Q-network Policy and Episodic Memory Policy

**Yangyang Zhao**[†‡], **Zhenyu Wang**[†*], **Changxi Zhu**[‡] and **Shihan Wang**[‡]
[†]South China University of Technology
[‡]Utrecht University
msyyz@mail.scut.edu.cn; wangzy@scut.edu.cn;
cxzhu.cn@gmail.com; s.wang2@uu.nl

## Abstract

Deep reinforcement learning has shown great potential in training dialogue policies. However, its favorable performance comes at the cost of many rounds of interaction. Most of the existing dialogue policy methods rely on a single learning system, while the human brain has two specialized learning and memory systems, supporting to find good solutions without requiring copious examples. Inspired by the human brain, this paper proposes a novel complementary policy learning (CPL) framework, which exploits the complementary advantages of the episodic memory (EM) policy and the deep Q-network (DQN) policy to achieve fast and effective dialogue policy learning. In order to coordinate between the two policies, we proposed a confidence controller to control the complementary time according to their relative efficacy at different stages. Furthermore, memory connectivity and time pruning are proposed to guarantee the flexible and adaptive generalization of the EM policy in dialog tasks. Experimental results on three dialogue datasets show that our method significantly outperforms existing methods relying on a single learning system.

## 1 Introduction

Dialogue policy, one of the most critical modules of task-oriented dialogue systems, aims to determine system responses based on current states (Zhang et al., 2019a). One of the earliest methods is the rule-based policy (Litman and Allen, 1987; Bos et al., 2003). Although this method often has acceptable performance, handcrafting rules are expensive and non-extensible. Recently, deep reinforcement learning (RL) has become a mainstream method for training dialogue policies (Cuayáhuitl, 2016; Li et al., 2017; Peng et al., 2017, 2018). Since Deep RL-based methods are learning in an online fashion, a large amount of interaction with real users is required, which is generally infeasible in practical application (Fatemi et al., 2016; Dhingra et al., 2017; Su et al., 2018; Wu et al., 2019).

Intuitively, the reward-based learning mechanism in Deep RL (DRL) coincides with the dopamine-centered regulation in the human brain (O'Reilly et al., 2014). Human brains have two differentially specialized learning and memory systems for collaboration, allowing them to find good solutions without requiring copious examples (McClelland et al., 1995; Norman and O"Reilly, 2003; O'Reilly et al., 2014). However, most of the DRL-based dialogue policies (Chen et al., 2017b; Peng et al., 2018; Lipton et al., 2018; Takanobu et al., 2019; Wu et al., 2019; Wang et al., 2020) rely on a single learning system, which neglects the human brain's memory structure. Consequently, we imitate the human brain to construct an efficient complementary policy learning (CPL) model with learning and memory systems for cooperation.

Inspired by cognitive neuroscience studies (Sutherland and Rudy, 1989; Daw et al., 2005; Poldrack et al., 2001), some researches evidence that episodic memory (EM) plays a vital role in decision tasks. Thus, they incorporate the EM into RL to accelerate learning (Blundell et al., 2016; Young et al., 2018; Pritzel et al., 2017; Lin et al., 2018). Despite the effectiveness of these methods on video game tasks, there is little research validating the practical usage of EM in dialogue tasks.

In this paper, we investigate the roles of the EM policy and the DQN policy (a classic representative of the DRL-based dialogue policies) in dialogue policy tasks. We observed that the EM policy is similar to the human brain's memory system, which efficiently leans with little data and bridges interdependency between actions and results from past experience. It is of limited usefulness in novel situations, since it generalizes poorly. The DQN policy is analogous to the human brain's learning system. It effectively extracts and generalizes potential in-

*Corresponding author



*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4311–4323
November 7–11, 2021. 

formation from a large amount of experience to drive decisions and calibrate strategies stored in the EM. Its good generalization comes at a cost of learning inefficiency and the demand for massive data. These two policies complement each other. Nevertheless, directly combining the DQN policy with the vanilla EM policy is difficult to maintain effectiveness consistently in the dialog policy task. Thus we have the following considerations:

(1) A meta-controller should be proposed to coordinate between the two policies. Over-reliance on the DQN policy may not achieve the available performance quickly, while over-reliance on the EM policy is difficult to generalize to new situations. (2) In order to ensure that the EM policy remains consistently effective in the dialogue tasks, a mechanism for generalization to new situations is needed. The same situation may never be encountered twice in dialogue tasks, and it is impossible to record all the situations.

For question (1), we propose a confidence controller that allows the two policies to form a seamless hybridization by controlling the complementary timing according to their relative efficacy at different stages. Once the CPL is enabled, the EM policy provides diversified guidances for the DQN policy: an extra memory objective (EMO), an example memory action (EMA), and an extra intrinsic reward (EIR). For question (2), we define memory connectivity that allows the flexibility and generalization of the EM policy by associating past familiar memories. Then, time pruning prunes the outdated memories.

In summary, our main contributions are two-fold: (1) We present a novel CPL framework, which gets rid of collecting any demonstrations and does not rely on any experts. Preferably, it exploits the complementary superiorities of the EM policy and the DQN policy through the confidence controller. To the best of our knowledge, this is the first work to learn a dialogue policy, which integrates the learning and memory systems seamlessly and avoids being stuck on a single system. (2) We experimentally demonstrate that the effectiveness of our framework, and EM can be a crucial building block of effective dialogue policy learning. Our model is the first step in that direction, as far as we know.

## 2 Related Work

The research on the learning efficiency of dialogue policies is not new. Lipton et al. (2016, 2018) showed that pre-filling the replay buffer with few successful dialogue experiences at the beginning can accelerate learning. Prioritized experience replay improves the sample efficiency by increasing the replay probability of experiences with higher temporal difference errors (Schaul et al., 2016). Peng et al. (2018) proposed a world model to simulate users and integrated planning into policy learning. A lot of progress has been made in improving the effectiveness of dialogue policies by combining supervised learning (SL) Henderson et al. (2008). Su et al. (2016, 2017) and Williams et al. (2017) proposed to use SL to initialize the policy network and then fine-tune it within the RL process. Chen et al. (2017a,b), Liu et al. (2018), and Zhao et al. (2021) incorporated a teacher to guide policy learning. Nevertheless, these methods require extra effort to hire or design teacher models. Wang et al. (2020) proposed an efficient policy learning from demonstrations. However, these methods require the collection of human demonstrations, and their performance depends on the quality of the demonstrations. Parallelly, another solution is to increase the density of meaningful rewards (Takanobu et al., 2019; Lu et al., 2019; Zhao et al., 2020).

Episodic memory has been used outside of dialogue research to improve data efficiency (Lengyel and Dayan, 2007). Blundell et al. (2016) proposed table-based model-free episodic control to learn past good experiences in a one-shot learning fashion. Pritzel et al. (2017) proposed neural episodic control, which uses differentiable neural dictionaries to store and lookup beneficial memories for decision. However, these table-based methods lack good generalization capabilities. Young et al. (2018) proposes a EM integrated into RL agent. But its computing time increases with the history length. Based on this, Lin et al. (2018) proposed episodic memory deep Q-network in the video game domains with high-dimensional. However, these researches focused on the video game fields, how effectively use the EM in the dialogue domain, and whether it is feasible are less explored.

## 3 Proposed Framework

The CPL framework is described in Figure 1, which mainly includes three modules: (1) The *episodic memory policy* quickly latches familiar experiences from the past to provide auxiliaries for the DQN policy. It includes two operations. *Writing* effectively retains memories while minimizing the reten-

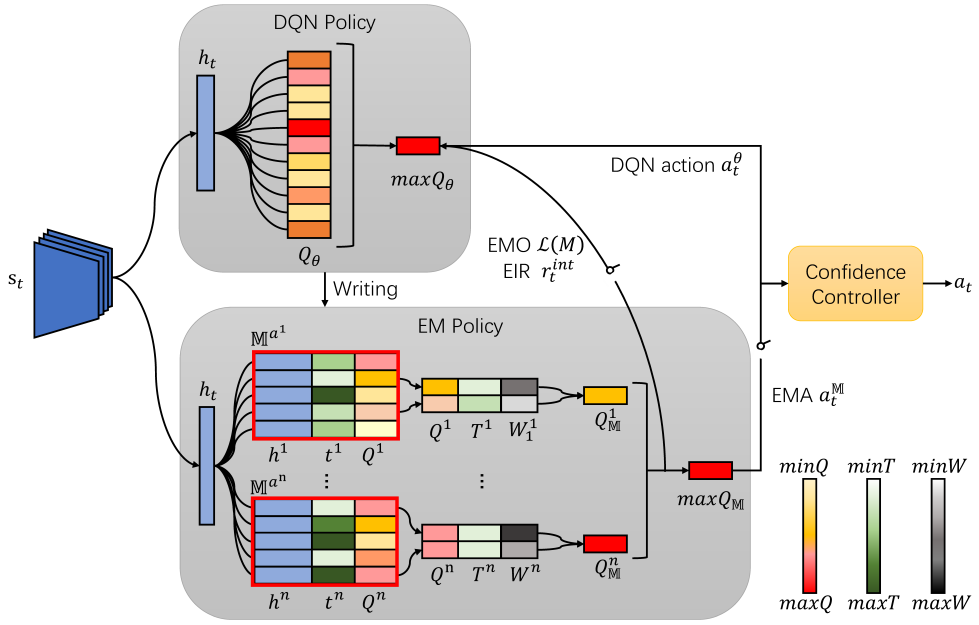



Figure 1: Complementary policy learning framework. Taking $M = 2$ as an example, it contains three colors, in which red represents the Q-value, green represents the update time, and black represents the importance.

tion of obsoleting memories. *Lookup* with *memory connectivity* and *time pruning* selectively associates relevant memories while casting aside irrelevant or obsolete memories; (2) The *DQN policy* effectively extracts and generalizes potential information from a large amount of experience to drive decisions and calibrate strategies stored in the EM; (3) The *confidence controller* choose an appropriate time to perform complementary policy learning according to their relative efficacy at different stages.

### 3.1 Episodic Memory Policy

Episodic memory policy is a memory system based on past experience. It can quickly record and replay the empirical decisions of the dialogue agent, containing two operations, as depicted in Figure 2:

**Writing**: we adopt the similar architecture as previous EM (Pritzel et al., 2017) to record past experience. For each action $a \in A$, the EM policy has a separate memory, which is indexed by states and actions, $\mathbb{M}^a = (H^a, T^a, Q^a)$. After the episode ends, the EM policy will write each $(h, t, Q(s, a))$ into the corresponding $\mathbb{M}^a$ through a backward replay process according to the following equation:

$$\mathbb{M}^a \leftarrow \begin{cases} add(h, t, Q_\theta(s, a)) if (s, a) \notin \mathbb{M}^a \\ update(h, t, Q_\mathbb{M}(s, a) \\ \quad + \alpha(Q_\theta(s, a) - \\ Q_\mathbb{M}(s, a))) \qquad otherwise \end{cases} \tag{1}$$

where $h$ is the representation of state $s$, $t$ are the update time, $Q_\theta(s, a)$ is the current Q-value estimated by the DQN policy, $Q_\mathbb{M}(s, a)$ is the past Q-value recorded by $\mathbb{M}^a$, $\alpha$ is the learning rate.

In theory, each $\mathbb{M}^a$ in vanilla EM is constantly growing, so they need to consume a large amount of memory to record. Therefore, we add the update time $T^a$ and overwrite the entry that has the least recently updated to minimize the retention of obsoleting memories and limit the size of the memory for each action. This is in line with the law that the human brain is more likely to forget older memories (Hardt et al., 2013).

**Lookup**: the query key $h$ is used to lookup similar experiences from the $\mathbb{M}^a$. For large-scale dialog tasks, novel states are common. However, the lookup methods used in the video game domain are not applicable. Generalizing familiar experiences to novel situations in our tasks is essential. Therefore, we define ***memory connectivity*** to lookup $M$ memories as similar memories [1] :

$$C(h||h_i) = \sum_{j=0}^{n} h(j) \cdot \log \frac{h(j)}{h_i(j)} \tag{2}$$

where $h$ and $h_i$ are two probability distributions of the state. The smaller $C(h||h_j)$, the stronger the connectivity of memories. Consequently, we use it to indicate the ***importance weighting*** of the

[1] We tested the influence of the $M$ value in the subsequent simulation experiments.

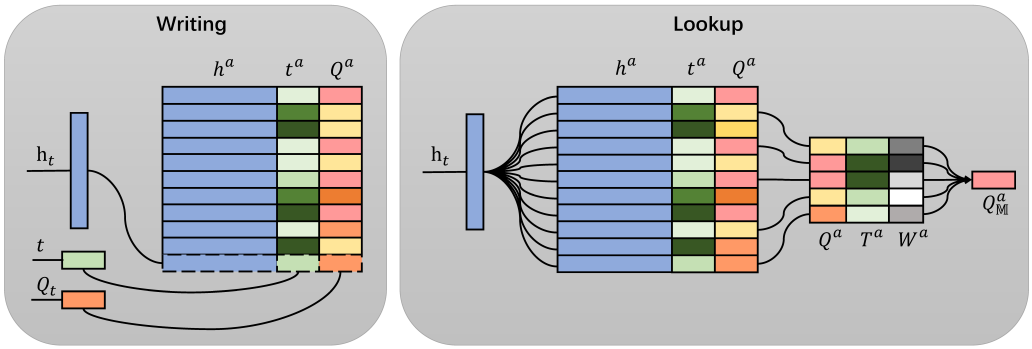



Figure 2: Illustration of writing and lookup operations on episodic memory.

selected memories, $W^a = 1 - C(h||h_i)$. This is in line with the law that the past familiar experiences have profound implications for humans (Carbonell, 1983). In order to cast aside the outdated non-optimal policies, we further propose ***time pruning*** for corresponding entry, which a monotonically decreasing function:

$$TP^a(t^a) = \begin{cases} 0 & if\ t' - t \geq T \\ 1 - \dfrac{(t' - t)}{T} & otherwise \end{cases} \quad (3)$$

where $T$ is the maximum valid time (set to 15 in this paper), $t'$ is the current time, and $t$ is the update time of the memory.

Therefore, for each the $\mathbb{M}^a$, the corresponding $Q^a_{\mathbb{M}}$ obtained by lookup operation can be rewritten as follow:

$$Q^a_{\mathbb{M}} = \begin{cases} Q^a & if (s,a) \in M^a \\ \sum_{j=1}^{M} [Q^a_j \cdot W^a_j \cdot & \\ TP^a_j(t^a_j)] & otherwise \end{cases} \quad (4)$$

where $W^a_j \cdot TP^a_j(t^a_j)$ is a normalized value belonging to $[0, 1]$, and $\sum_{j=1}^{M}[W^a_j \cdot TP^a_j(t^a_j)] = 1$. The EM policy select the corresponding action with the maximum $Q^a_{\mathbb{M}}$ as the memory action $a^{\mathbb{M}}_t$ for subsequent auxiliaries.

Overall, the EM policy is different from the DQN policy, which does not correspond to estimate the expected return, rather than looking up the highest potential return for a given state based on the previous memories.

### 3.2 DQN Policy

The task-oriented dialogue policy learning is typically formulated as an MDP problem. We employ the vanilla DQN (Mnih et al., 2015) [2]to train the dialogue policy based on experience from the interaction between agents and users.

At each step, the agent uses $\epsilon$-greedy to select a DQN action based on the dialogue state $s$. Afterward, the agent obtains a reward $r$, observes a corresponding user response, and updates the dialogue state to the next $s'$ until the end of the conversation. Finally, we store the experience $(s, a, r, s')$ into the experience replay buffer $D$. We optimize the parameter $\theta$ by minimizing the mean-squared loss function. It is worth noting that here we only consider the vanilla inference objective function:

$$\begin{aligned} \mathcal{L}(\theta) &= \mathbb{E}_{(s,a,r,s')\sim D}[(y_i - Q_\theta(s,a))^2] \\ y_i &= r + \gamma \max_{a'} Q_{\theta'}(s', a') \end{aligned} \quad (5)$$

where $\gamma \in [0, 1]$ is a discount factor, and $Q_{\theta'}$ is the target value function that is updated periodically. $Q_\theta$ is optimized through back-propagation and mini-batch deep Q-learning.

### 3.3 Confidence Controller

We use the confidence controller to control the complementary timing by judging the confidence of the DQN policy.

We use the DropoutQNetwork [3] (Hinton et al., 2012; Srivastava et al., 2014; Chen et al., 2017b) to estimate the confidence of the DQN policy at $t$-th turn $c^D_t$ (lines 7-12 in Algorithm 1). The DQN policy has more confidence when the $c^D_t$ is greater than the confidence threshold $\xi$, and its $Q_\theta$ is greater than $Q_{\mathbb{M}}$. Otherwise, the EM policy is more confident. When the DQN has less confidence, the CPL is enabled, where the EM policy provides three guidances for the DQN policy:

[2]Obviously, our approach works for any policy optimizer.

[3]The output element of each hidden layer $h$ is randomly set to 0 with probability $p$ and then fed to the next layer.

**Algorithm 1** The Procedure of CPL.

**for** *each episode* **do**
  **for** $t = 1 \rightarrow T$ **do**
    Initialize the probability vector $p = [p_1, p_2, ..., p_n]$ with zeros, where $n$ is the number of actions.
    Receive observation $s_t$ from the environment
    $a_t^D \leftarrow \epsilon$-*greedy* policy based on the DQN policy
    **for** $i = 1 \rightarrow N$ **do**
      $Q_i(s_t, a) \leftarrow$ *DropoutQNetwork*$(s_t)$
      $a_{ti} = \arg\max_a Q_i(s_t, a)$
      $p[a_{ti}] \leftarrow p[a_{ti}] + \frac{1}{N}$
    **end for**
    Get confidence of $a_t^D$: $c_t^D = p[a_t^D]$
    **if** $c_t^D > \xi$ & $Q_\theta > Q_\mathbb{M}$ **then**
      Take action $a_t^D$, receive environment rewards $r_t^{ext}$ and next state $s_{t+1}$.
      Append $(s_t, a_t^D, r_t^{ext}, s_{t+1})$ to $D$
      Train DQN policy via single inference objective (Eq.(5))
    **else**
      Lookup $Q_\mathbb{M}^a$ for each action via Eq.(4), $a_t^\mathbb{M} = \arg\max_a Q_\mathbb{M}$
      Take action $a_t^\mathbb{M}$, receive environment rewards and EIR $r_t \leftarrow r_t^{ext} + r_t^{int}$ and next state $s_{t+1}$.
      Append $(s_t, a_t^\mathbb{M}, r_t, s_{t+1})$ to $D$
      Train DQN policy via two inference objectives (Eq.(6))
    **end if**
  **end for**
  **for** $t = 1 \leftarrow T$ **do**
    Update $\mathbb{M}_a$ using $(h_t, t, Q(s_t, a_t))$ via Eq.(1)
  **end for**
**end for**

a) **Extra Memory Objective (EMO)**: the EM policy provides an extra memory objective $\mathcal{L}(\mathbb{M})$ to reconciles the loss function of DQN policy. We propose a new objective function combining the two objectives:

$$\mathcal{L}(\mathbb{M}) = \mathbb{E}_{(s,a,r,s')\sim D}[(Q_\theta(s_t, a_t) - Q_\mathbb{M}(s_t, a_t))^2] \quad (6)$$
$$\mathcal{L} = \mathcal{L}(\theta) + \lambda\mathcal{L}(\mathbb{M})$$

where $Q_\theta(s_t, a_t)$ is the same as $Q_\theta(s, a)$ in Eq. (5). $Q_\mathbb{M}(s_t, a_t)$ is the Q-value looked up by the EM policy in the same action. And we weigh the two policies by adjusting the value of $\lambda$. In this way, we make flexible use of two policies in the learning process.

b) **Example Memory Action (EMA)**: the memory action $a_t^\mathbb{M}$ with the highest potential reward replaces the DQN action $a_t^\theta$ for responding.

c) **Extra Intrinsic Reward (EIR)**: exploitation rewards and exploration rewards are composed of the extra intrinsic reward $r_t^{int}$ to encourage the DQN policy to explore and exploit effectively. If the DQN action $a_t^\theta$ is the same as the memory action $a_t^\mathbb{M}$ with the highest potential rewards, exploitation rewards are provided. If the DQN action $a_t^\theta$ does not appear in the corresponding $\mathbb{M}^a$, exploration rewards are provided.

At each turn, the dialogue state $s_t$ is transmitted to both the DQN policy and the EM policy. The DQN policy first generates a DQN action $a_t^\theta$. Then the confidence controller judges whether the DQN policy has sufficient confidence. When it has less confidence, the EM policy provided auxiliaries for it: EMO, EMA, and EIR. After the episode ends, the memories of the EM policy will be updated through a backward replay process. The full procedure of the CPL is described in Algorithm 1.

| Task | Intents | Slots | User goals |
|---|---|---|---|
| Movie-Ticket Booking | 11 | 16 | 128 |
| Restaurant Reservation | 11 | 30 | 3525 |
| Taxi Ordering | 11 | 29 | 2830 |

Table 1: The number of intents, slots and user goals in three datasets.

## 4 Performance Evaluation

We conduct sufficient experiments on three public task-oriented datasets in both simulation and human evaluation: movie-ticket booking, restaurant reservation, and taxi ordering [4].

### 4.1 Dataset

The movie-ticket booking task is collected from Amazon Mechanical Turk and annotated by Li et al. (2017), and the other two tasks are provided by Microsoft Dialogue Challenge (Li et al., 2016, 2018). Each domain has its domain-specific intents, slots,

[4]We consider that these tasks have been widely used in the research of dialogue policy (Li et al., 2017; Wang and Chen, 2019; Zhang et al., 2019b; Wang et al., 2020). Hence, we use these three datasets as all benchmark task-oriented dialogue environments to evaluate our model.

and labeled dialogues, and the statistics are shown in Table 1. Readers can refer to the details of the three domains from Appendix A.

### 4.2 Baselines

To benchmark the performance of our method, we have developed different versions of task-oriented dialogue agents as baselines for comparison:

- **DQN** agents are learned with standard DQN with only direct reinforcement learning [5].
- **DQN($K$)** agents are learned by DQN, but with $(K-1)$ times more real experiences than the DQN agent (Peng et al., 2018; Su et al., 2018; Wu et al., 2019). [6].
- **EPAC** agents introduce a human teacher in the training process to teach dialogue policy learning via providing example actions and extra rewards (Chen et al., 2017a).
- **$S^2$Agent** learns the dialogue policy from demonstrations through policy shaping and reward shaping (Wang et al., 2020).

In order to further analyze the effectiveness of each component in our method, we construct ablation tests:

**Proposed CPL**

- **CPL** is our proposed approach which learns policy by complementary policy learning.
- **CPL w/o EMP** is a variant of CPL which learns policy by DQN policy with two guidances (without EMA).
- **CPL w/o DQN** is a variant of CPL, but only uses EMP to make quick decisions (without the DQN action).
- **CPL w/o W** is our proposed method which learns policy by complementary policy learning without importance weights.
- **CPL w/o T** is our proposed method which learns policy by complementary policy learning without time pruning.

[5]For a fair comparison, all baselines are based on DQN rather than DQN($K$).

[6]Since the performance of DQN($K$) can be viewed as the upper bound of DDQ($K$) (Peng et al., 2018), D3Q($K$) (Su et al., 2018), and Switch-DDQ($K$) (Wu et al., 2019) with the same planning steps, we directly use DQN($K$) instead of above methods as the baseline model.

### 4.3 Implementation Details

For all RL-based agents, value network $Q(\cdot)$ has one hidden layer MLPs with 80 hidden nodes, $ReLU$ is used as the activation function in three domains. All the NN models are warm start 100 epochs and trained with the same hyper-parameters settings. $\epsilon$-greedy is applied for policy exploration which starts from $0.2$ and decays every episode with a decay rate of $0.95$. We set the discount factor $\gamma = 0.9$. The size of the experience relay in the movie domain and other domains is set to 5000 and 10000, respectively. The batch size is 16, and the learning rate is 0.001. We set $K$ as 10 in the movie domain and 50 in other domains. For a fair comparison, all baselines (except DQN($K$)) are based on DQN rather than DQN($K$).

In terms of hyperparameters for EM policy, the memories are stored up to 5000 per action. We do a backward replay update for each action after the end of each episode. The $M = 5$ unless indicated. The learning rate $\alpha$ in Eq.1 is set to 0.1. We fix $\lambda$ in Eq.6 at $0.1$. The confidence threshold $\xi$ is set $0.7$. The $N$ is set to $50$. The dropout rate is set to 0.25. Exploration in the EM Policy is applied by using $\epsilon$-greedy with $\epsilon = 0.005$. The maximal extra intrinsic reward $r^{int}$ is 5. Appendix B shows detailed information about the user simulator.

### 4.4 Simulation Evaluation

#### 4.4.1 Main Results

The main simulation results are shown in Table 2 and Figure 3. From the results, it is clear that through complementary policy learning, the CPL agents are much faster and consistently better than other strong methods in all domains.

Figure 3 shows the learning curves of different agents in three domains. It can be seen that the DQN($K$) performs better than the DQN in all domains since its experiences have $K-1$ times more than the DQN. With the same number of experiences, EAPC and $S^2Agent$ consistently perform better than the DQN in all domains. And even in the case of less experience, they are still superior to the DQN($K$) in restaurant and taxi domains. But their performance hardly exceeds the DQN($K$) in the movie domain. The reason might be that, in the simpler movie domain where dialogues are easier to succeed, simply increasing experiences makes efficiency improvement more obvious. By contrast, in the relatively complex domains where successful dialogues are relatively rare, it is difficult to provide

| Agent | domain | Epoch = 100 | | | Epoch = 200 | | | Epoch = 300 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Success | Reward | Turns | Success | Reward | Turns | Success | Reward | Turns |
| DQN | Movie | 0.4012 | -6.477 | 31.24 | 0.5242 | 10.36 | 27.08 | 0.6448 | 26.17 | 24.40 |
| DQN(10) | | 0.7796 | 46.80 | **15.52** | 0.8136 | 51.75 | **13.76** | 0.8002 | 50.19 | **13.68** |
| EAPC | | 0.4930 | 26.07 | 28.18 | 0.6685 | 30.18 | 22.08 | 0.7180 | 58.81 | 22.60 |
| $S^2Agent$ | | 0.5867 | 35.84 | 29.61 | 0.6978 | 49.29 | 28.46 | 0.6982 | 51.80 | 29.75 |
| CPL* | | **0.8386** | **62.93** | 28.05 | **0.8448** | **66.80** | 22.28 | **0.8446** | **67.54** | 20.18 |
| *CPL w/o EMP* | | 0.6214 | 33.11 | 24.91 | 0.7728 | 40.95 | 19.86 | 0.8198 | 45.52 | 18.37 |
| *CPL w/o DQN* | | 0.3881 | 34.22 | 33.26 | 0.5894 | 55.27 | 32.45 | 0.5969 | 58.69 | 33.16 |
| *CPL w/o W* | | 0.3935 | 33.86 | 30.19 | 0.3982 | 33.94 | 30.58 | 0.3997 | 39.18 | 30.44 |
| *CPL w/o T* | | 0.4774 | 40.99 | 27.86 | 0.5406 | 52.58 | 28.12 | 0.5283 | 51.82 | 28.19 |
| DQN | Rest. | 0.0358 | -55.13 | 40.85 | 0.0385 | -54.85 | 40.94 | 0.0439 | -54.14 | 40.82 |
| DQN(50) | | 0.0996 | -43.99 | 35.91 | 0.1201 | -41.98 | 35.59 | 0.1260 | -41.42 | 35.52 |
| EAPC | | 0.1882 | -32.58 | 33.33 | 0.2079 | -30.63 | 33.29 | 0.2294 | -28.13 | 32.83 |
| $S^2Agent$ | | 0.2327 | -21.08 | 29.54 | 0.2352 | -20.46 | 29.62 | 0.2378 | -20.39 | 29.58 |
| CPL* | | **0.4832** | **17.76** | 29.31 | **0.4846** | **24.03** | 28.32 | **0.4673** | 17.83 | **25.91** |
| *CPL w/o EMP* | | 0.3548 | 6.360 | 27.60 | 0.4141 | 15.03 | **27.59** | 0.4559 | **18.32** | 27.64 |
| *CPL w/o DQN* | | 0.3227 | 8.896 | 30.99 | 0.3399 | 13.02 | 32.18 | 0.3456 | 16.44 | 30.60 |
| *CPL w/o W* | | 0.2694 | 4.041 | 34.89 | 0.2775 | 11.97 | 36.12 | 0.2279 | 15.05 | 35.92 |
| *CPL w/o T* | | 0.2333 | 5.393 | **26.90** | 0.2973 | 11.32 | 29.28 | 0.3137 | 12.02 | 30.68 |
| DQN | Taxi | 0.0000 | -58.31 | 38.62 | 0.0015 | -59.68 | 41.73 | 0.0095 | -57.97 | 40.21 |
| DQN(50) | | 0.2534 | -25.82 | 34.45 | 0.2638 | -24.44 | 34.19 | 0.2748 | -22.97 | 33.89 |
| EAPC | | 0.3178 | -13.23 | 27.67 | 0.3172 | -13.30 | 29.70 | 0.3209 | -12.97 | 25.62 |
| $S^2Agent$ | | 0.3409 | -13.98 | 31.76 | 0.3743 | -9.394 | 30.63 | 0.4181 | -3.669 | 29.69 |
| CPL* | | **0.6086** | **30.55** | 28.88 | **0.6413** | **37.04** | 30.12 | **0.6822** | **44.04** | 31.20 |
| *CPL w/o EMP* | | 0.4967 | 5.472 | **26.53** | 0.5712 | 12.02 | **25.15** | 0.6441 | 20.89 | **23.61** |
| *CPL w/o DQN* | | 0.3742 | 11.21 | 30.70 | 0.4484 | 12.14 | 32.84 | 0.4631 | 26.09 | 33.52 |
| *CPL w/o W* | | 0.2839 | 17.93 | 33.36 | 0.2688 | 12.60 | 31.30 | 0.2832 | 12.07 | 29.87 |
| *CPL w/o T* | | 0.2857 | 13.17 | 36.64 | 0.2979 | 11.76 | 29.20 | 0.3279 | 14.51 | 31.94 |

Table 2: The results of different agents at training $epoch = \{100, 200, 300\}$. Each number is averaged over 10 runs, and each run is tested on 1000 dialogues. Best scores are labeled in blue. * denotes significant level $p < 0.05$ with other agents. Success: average success rate, Reward: average reward, Turn: average turn.

clear guidance for agents. The above observations are also confirmed in Table 2. With complementary learning, the CPL agent also alleviates reward sparsity issues, which is especially obvious in relatively complex domains. In the restaurant and taxi domains, the average rewards of all baselines are negative, while the CPL agent always learns meaningful positive actions. These actions are basically given in the form of EIR.

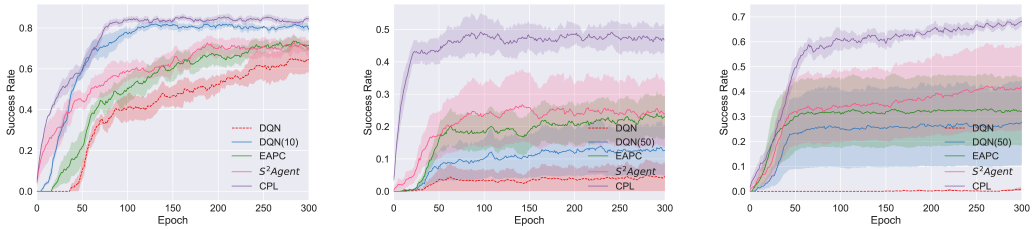
Moreover, an additional result is observed. Although the CPL agents have the highest average success rate and rewards, their average turns are longer than the *CPLw/oEMP* agents. We argue that the EMA from EM policy may be non-optimal, causing the CPL agents to complete user goals in a detour instead of the most effective way. The *CPLw/oEMP* agents explore a more efficient path through EMO and EIR.

#### 4.4.2 Training with varing number of M

Intuitively, the number of $M$ has a large impact on dialogue policy learning. $M$ represents the number of empirical decisions that the EM policy provides to the DQN policy for reference. Experiments with varying numbers of $M$ values were conducted in three domains. The moving averaged success rate is calculated at 300 epochs. Figure 4 shows that the moving average success rate of each agent during the learning. The agent with a small $M$ value still has better learning efficiency in the movie domain, while the agent performs worst in other domains. In all domains, the agent with a large $M$ value has an inferior learning efficiency. This is owing to the fact that the dialogue agent benefits from related memories in many aspects to consider the current state more comprehensively with the increase of $M$. After more than 9, irrelevant episodic memories are chosen to simply fill the post, which affects the efficiency and quality of dialogue policy learning. This experimental result also verified our assumption.

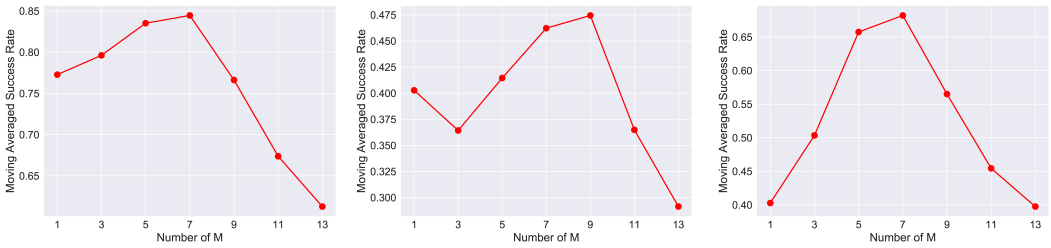
#### 4.4.3 Training with varing values of $\lambda$

Similarly, the $\lambda$ affects the performance of dialogue policies by controlling the use of two policies (EM policy and DQN policy) in the dialogue policy learning process. Therefore, experiments with




(a) Movie (b) Rest. (c) Taxi

Figure 3: The learning curves of different agents in Movie, Restaurant, and Taxi domains.




(a) Movie (b) Rest. (c) Taxi

Figure 4: The effect of the number of $M$ on performance in Movie, Restaurant, and Taxi domains.

varying $\lambda$ values were conducted in three domains to serve as a reference for CPL practitioners. The moving average success rate of each agent at 300 epochs is shown in Figure 6. It can be viewed that no matter whether the EM policy is completely non-participation or completely dominated, it seriously hurt the performance of dialogue policies. It performs better when the DQN policy is dominating with the EM policy auxiliary.

#### 4.4.4 Ablation Test

We conduct ablation experiments to analyze the effectiveness of each component in the CPL framework. As illustrated in Figure 5, although the average success rate of the *CPLw/oEMP* in the early stage is lower than the CPL, it can achieve approximate performance finally in three domains. The *CPLw/oDQN* achieves rapid learning in the early stage, but its later learning is limited when making decisions in novel situations. It can be seen that the involvement of the EM policy in the CPL framework tends to predominate early, while the involvement of the DQN policy predominates later. Although both the *CPLw/oW* and the *CPLw/oT* learn faster in the early stage, the performance in the later stage hardly improves. It is helpful to reference memories aggressively at the early stages regardless of their relevance and timeliness. With the increase of training time, the dialogue agent has been significantly improved, irrelevant and outdated memories often hurt the performance badly. The experiment verifies that the four components benefit the CPL to a large extent.

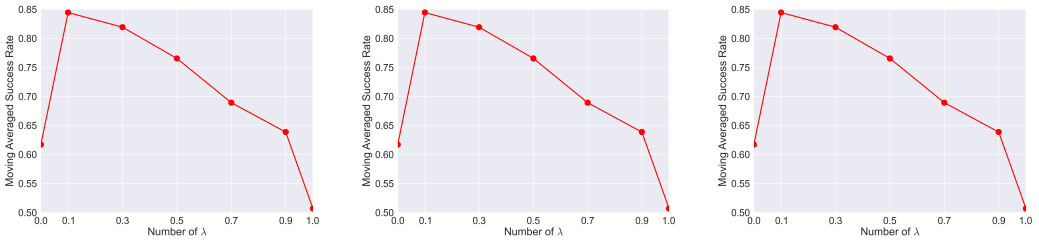
### 4.5 Human Evaluation

In order to further verify the feasibility of our method in real dialogue scenarios, we recruited 33 real users to interact with different agents in three tasks without know which one is behind. We collect 50 valid conversations for each agent in each domain. All evaluated agents have been trained for 300 epochs. In each conversation, users randomly select an agent to communicate with a user goal sampled from the corpus. Users have the right to abandon the task and terminate the conversation if they believe that the dialogue is unlikely to succeed. At the end of the conversation, in addition to requiring users to provide feedback on whether the conversation is successful, the datasets (Li et al., 2018, 2017) also needs users to evaluate the naturalness, coherence, and task completion ability of the agent with a score of 1 to 5 [7]. As illustrated in Table 3, the CPL and *CPLw/oEMP* are significantly outperforms other agents and the CPL is considered to be more slightly dilatory than *CPLw/oEMP*, which is consistent with what we have observed in simulation evaluation.

[7] 5 is the best, 1 is the worst


(a) Movie (b) Rest. (c) Taxi

Figure 5: The ablation experiment of four components of our method in Movie, Restaurant, and Taxi domains.




(a) Movie (b) Rest. (c) Taxi

Figure 6: The effect of the number of $\lambda$ on performance in Movie, Restaurant, and Taxi domains.

| Agent | Movie | | Rest. | | Taxi | |
|---|---|---|---|---|---|---|
| | Success | Rating | Success | Rating | Success | Rating |
| DQN | 0.48 | 3.14 | 0.03 | 1.71 | 0.01 | 1.09 |
| DQN($K$) | 0.59 | 3.40 | 0.10 | 2.48 | 0.24 | 2.38 |
| EAPC | 0.57 | 3.08 | 0.27 | 2.56 | 0.41 | 2.50 |
| $S^2Agent$ | 0.64 | 2.84 | 0.23 | 1.91 | 0.43 | 2.48 |
| CPL* | **0.75** | 3.61 | **0.42** | 2.85 | **0.64** | 2.69 |
| *CPL w/o EMP* | 0.72 | **3.88** | 0.40 | **3.65** | 0.62 | **3.36** |

Table 3: Human evaluation of different agents in Movie, Rest. and Taxi domains.

## 5 Conclusion

In this paper, we propose a novel complementary policy learning (CPL) framework that realized dialogue policy learning in a more effective and faster manner through direct use of its own experience without any extra cost. This framework exploits the complementary advantages of the EM policy and the DQN policy. Additionally, we proposed a confidence controller to coordinate between the two policies according to their relative efficacy at different stages. Further proposed memory connectivity and time pruning ensure the flexible and adaptive generalization of the EM policy in dialogue tasks. The results show that the CPL significantly outperforms baselines in three domains, and an episodic memory component is a crucial building block of effective dialogue policy learning. To the best of our knowledge, this is the first work to learn a dialogue policy, which integrates the learning and memory systems seamlessly and avoids being stuck on a single system. In the future, we plan to expand our method to multi-domain tasks, e.g., MultiWoz (Budzianowski et al., 2018) and evaluating it using other dialogue platforms, e.g., PyDial (Ultes et al., 2017), Convlab (Lee et al., 2019).

## 6 Acknowledgments

We would like to thank the reviewers for their comments and efforts towards improving our paper. And we would like to acknowledge volunteers of the South China University of Technology who help us with the human experiments. This work was supported by the Key-Area Research and Development Program of Guangdong Province, China (Grant No.2019B0101540042) and the Natural Science Foundation of Guangdong Province, China (Grant No.2019A1515011792).

## References

Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z. Leibo, Jack W. Rae, Daan Wierstra, and Demis Hassabis. 2016. Model-free episodic control. *CoRR*.

Johan Bos, Ewan Klein, Oliver Lemon, and Tetsushi Oka. 2003. DIPPER: description and formalisation of an information-state update dialogue system architecture. In *SIGDIAL*, pages 115–124.

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

Jaime G Carbonell. 1983. Learning by analogy: Formulating and generalizing plans from past experience. In *Machine learning*, pages 137–161.

Lu Chen, Runzhe Yang, Cheng Chang, Zihao Ye, Xiang Zhou, and Kai Yu. 2017a. On-line dialogue policy learning with companion teaching. In *EACL*, pages 198–204.

Lu Chen, Xiang Zhou, Cheng Chang, Runzhe Yang, and Kai Yu. 2017b. Agent-aware dropout DQN for safe and efficient on-line dialogue policy learning. In *EMNLP*, pages 2454–2464.

Heriberto Cuayáhuitl. 2016. Simpleds: A simple deep reinforcement learning dialogue system. In *IWSDS*, volume 427 of *Lecture Notes in Electrical Engineering*, pages 109–118.

N. D. Daw, Y. Niv, and P. Dayan. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *ACL*, pages 484–495.

Mehdi Fatemi, Layla El Asri, Hannes Schulz, Jing He, and Kaheer Suleman. 2016. Policy networks with two-stage training for dialogue systems. In *SIGDIAL*, pages 101–110.

Oliver Hardt, Karim Nader, and Lynn Nadel. 2013. Decay happens: the role of active forgetting in memory. *Trends in cognitive sciences*, 17(3):111–120.

James Henderson, Oliver Lemon, and Kallirroi Georgila. 2008. Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Comput. Linguistics*, 34(4):487–511.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.

Sungjin Lee, Qi Zhu, Ryuichi Takanobu, Xiang Li, Yaoqin Zhang, Zheng Zhang, Jinchao Li, Baolin Peng, Xiujun Li, Minlie Huang, et al. 2019. Convlab: Multi-domain end-to-end dialog system platform. *arXiv preprint arXiv:1904.08637*.

Máté Lengyel and Peter Dayan. 2007. Hippocampal contributions to control: The third way. In *NIPS*, pages 889–896.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Çelikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *IJCNLP*, pages 733–743.

Xiujun Li, Zachary C Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, and Yun-Nung Chen. 2016. A user simulator for task-completion dialogues. *arXiv preprint arXiv:1612.05688*.

Xiujun Li, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125*.

Zichuan Lin, Tianqi Zhao, Guangwen Yang, and Lintao Zhang. 2018. Episodic memory deep q-networks. In *IJCAI*, pages 2433–2439.

Zachary C. Lipton, Jianfeng Gao, Lihong Li, Xiujun Li, Faisal Ahmed, and Li Deng. 2016. Efficient exploration for dialog policy learning with deep BBQ networks \& replay buffer spiking. *CoRR*.

Zachary C. Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. 2018. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In *AAAI*, pages 5237–5244.

Diane J. Litman and James F. Allen. 1987. A plan recognition model for subdialogues in conversations. *Cogn. Sci.*, 11(2):163–200.

Bing Liu, Gökhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry P. Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *NAACL-HLT*, pages 2060–2069.

Keting Lu, Shiqi Zhang, and Xiaoping Chen. 2019. Goal-oriented dialogue policy learning from failures. In *AAAI*, pages 2596–2603.

McClelland, James, L., O'Reilly, Randall, and C. 1995. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nat.*, 518(7540):529–533.

Kenneth A. Norman and Randall C. O"Reilly. 2003. Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, 110(4):611–46.

Randall C. O'Reilly, Rajan Bhattacharyya, Michael D. Howard, and Nicholas Ketz. 2014. Complementary learning systems. *Cogn. Sci.*, 38(6):1229–1248.

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

Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Çelikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *EMNLP*, pages 2231–2240.

R. A. Poldrack, J. Clark, EJ Paré-Blagoev, D. Shohamy, J. C. Moyano, C. Myers, and M. A. Gluck. 2001. Interactive memory systems in the human brain. *Nature*.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. 2017. Neural episodic control. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 2827–2836.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized experience replay. In *ICLR*.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Pei-Hao Su, Pawel Budzianowski, Stefan Ultes, Milica Gasic, and Steve J. Young. 2017. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. In *SIGdial*, pages 147–157.

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

Shang-Yu Su, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Yun-Nung Chen. 2018. Discriminative deep dyna-q: Robust planning for dialogue policy learning. In *EMNLP*, pages 3813–3823.

R. J. Sutherland and J. W. Rudy. 1989. Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology*, 17(2):129–144.

Ryuichi Takanobu, Hanlin Zhu, and Minlie Huang. 2019. Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog. In *EMNLP-IJCNLP*, pages 100–110.

Stefan Ultes, Lina M Rojas Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Inigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gasic, et al. 2017. Pydial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78.

Huimin Wang, Baolin Peng, and Kam-Fai Wong. 2020. Learning efficient dialogue policy from demonstrations through shaping. In *ACL*, pages 6355–6365.

Yu-An Wang and Yun-Nung Chen. 2019. Dialogue environments are different from games: Investigating variants of deep q-networks for dialogue policy. In *ASRU*, pages 1070–1076.

Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *ACL*, pages 665–677.

Yuexin Wu, Xiujun Li, Jingjing Liu, Jianfeng Gao, and Yiming Yang. 2019. Switch-based active deep dyna-q: Efficient adaptive planning for task-completion dialogue policy learning. In *AAAI*, pages 7289–7296. AAAI Press.

Kenny J. Young, Richard S. Sutton, and Shuo Yang. 2018. Integrating episodic memory into a reinforcement learning agent using reservoir sampling. *CoRR*.

Rui Zhang, Kai Yin, and Li Li. 2020. Towards emotion-aware user simulator for task-oriented dialogue. *arXiv preprint arXiv:2011.09696*.

Zheng Zhang, Minlie Huang, Zhongzhou Zhao, Feng Ji, Haiqing Chen, and Xiaoyan Zhu. 2019a. Memory-augmented dialogue management for task-oriented dialogue systems. *ACM Trans. Inf. Syst.*, 37(3):34:1–34:30.

Zhirui Zhang, Xiujun Li, Jianfeng Gao, and Enhong Chen. 2019b. Budgeted policy learning for task-oriented dialogue systems. In *ACL*, pages 3742–3751.

Yangyang Zhao, Zhenyu Wang, and Zhenhua Huang. 2021. Automatic curriculum learning with over-repetition penalty for dialogue policy learning. *CoRR*.

Yangyang Zhao, Zhenyu Wang, Kai Yin, Rui Zhang, Zhenhua Huang, and Pei Wang. 2020. Dynamic reward-based dueling deep dyna-q: Robust policy learning in noisy environments. In *AAAI*, pages 9676–9684.

## A Appendices

Table 4 lists all annotated dialogue acts and slots in details.

These three datasets are not used to directly train the dialogue policy model, but to extract the user goals. Therefore, the movie-ticket booking task is simpler than the other two tasks. For each conversation, the user simulator (Li et al., 2016; Zhang et al., 2020) randomly samples a user goal from the user goal set to interact with the agent. The goal of each agent is to help them achieve specific user goals.

In order to verify the effectiveness of the proposed method, the datasets provide both automatic and human evaluations on three criteria (Li et al., 2018, 2017): success rate, average turns, and average reward [8]. Also, the datasets conducted a human evaluation: in addition to the above criteria, human users need to give a rating (1-5) at the end of each conversation according to the naturalness, coherence, and task completion ability of the agent. Specifically, the fulfillment degree of task (2 points), natural responses (1 point), timely and correct responses (1 point), and smoothly steer conversations (1 point). In this paper, we choose the success rate as our main evaluation criteria. If and only if the agent identifies all constraints provided by users and provides all information that users want, and finally successfully booking, the user goal is considered successful.

## B Appendices

The task-oriented dialogue system is designed to assist users to accomplish a specific goal $G$. The entire conversation revolves around this user goal $G$ implicitly, while the agent knows nothing about the user goal explicitly.

In order to make the user goal $G$ more clear, taking the movie-ticket booking domain as an example. A user may ask about the *theater* and *starttime* of a *today*'s movie-ticket about the *Enter the Dragon* of *Bruce Lee*,where the goal is in the form of:

$$\textbf{Goal} = \left( C = \begin{bmatrix} moviename = Enter \\ the\ Dragon \\ actor = BruceLee \\ date = today \end{bmatrix}, \quad R = \begin{bmatrix} theater = \\ starttime = \end{bmatrix} \right) \tag{7}$$

The user goals are generated from the annotated dataset mentioned in Section 4.1. The user goals extracted from the dataset are then aggregated into a user goal set. Whenever running dialogues, the user simulator randomly samples one user goal from this user goal set.

For the intrinsic rewards $r_int$, it includes exploitation rewards and exploration rewards to encourage the DQN policy to explore and exploit effectively. Exploitation rewards of 5 are provided, when the DQN action is the same as the memory action with the highest potential rewards. Exploration rewards of 5 are provided, if no memories are corresponding to the DQN action in EM policy. These two rewards do not appear at the same time. For the external reward function, in all domains, the agent receives $2L$ reward if the dialogue finishes successfully and $-L$ if it fails, where $L$ is the maximum of turns in each dialogue. A fixed $(-1)$ penalty is given to the agent at each turn to encourage the policy to reach the goal more efficiently. We set $L$ to 40 in three domains.

[8] In this paper, we choose the success rate as our main evaluation criteria.

Table 4: Number of intents, slots and dialogues in three dataset.

| Task | Intents | Slots | Dialogues |
|---|---|---|---|
| Movie | request, inform, confirm_question, confirm_answer, greeting, closing, deny,not_sure, multiple_choice, thanks, welcome | city, closing, data, greeting, distanceconstraints, moviename, price, numberofpeople, starttime, state, taskcomplete, theater, teater_chain, ticket, video_format, zip | 280 |
| Restaurant | request, inform, confirm_question, confirm_answer, greeting, closing, deny,not_sure, multiple_choice, thanks, welcome | address, atmosphere, choice, city, closing, cuisine, date, food, dress_code, greeting, distanceconstraints, numberofkids, mealtype, numberofpeople, other, personfullname, phonenumber, pricing, rating, restaurantname, restauranttype, seating, starttime, state, zip, result, occasion, taskcomplete, reservation | 4103 |
| Taxi | request, inform, confirm_question, confirm_answer, greeting, closing, deny,not_sure, multiple_choice, thanks, welcome | car_type, city, speed, closing, car_level, date, distanceconstraints, dropoff_location, zip, result, numberofkids, greeting, name, driver_id, numberofpeople, other, pickup_location, state, dropoff_location_city, pickup_location_city, pickup_time, cost, taxi_company, mc_list, taskcomplete, taxi, budget, emergency degree, drive_level | 3094 |