Machine Intelligence Research
www.mi-research.net

DOI: 10.1007/s11633-022-1347-y

A Survey on Recent Advances and Challenges in Reinforcement Learning Methods for Task-oriented Dialogue Policy Learning

Wai-Chung Kwan* Hong-Ru Wang* Hui-Min Wang Kam-Fai Wong

The Systems Engineering and Engineering Management Department, The Chinese University of Hong Kong, Hong Kong 999077, China

Abstract: Dialogue policy learning (DPL) is a key component in a task-oriented dialogue (TOD) system. Its goal is to decide the next action of the dialogue system, given the dialogue state at each turn based on a learned dialogue policy. Reinforcement learning (RL) is widely used to optimize this dialogue policy. In the learning process, the user is regarded as the environment and the system as the agent. In this paper, we present an overview of the recent advances and challenges in dialogue policy from the perspective of RL. More specifically, we identify the problems and summarize corresponding solutions for RL-based dialogue policy learning. In addition, we provide a comprehensive survey of applying RL to DPL by categorizing recent methods into five basic elements in RL. We believe this survey can shed light on future research in DPL.

Keywords: Dialogue policy learning (DPL), task-oriented dialogue system (TOD), reinforcement learning (RL), dialogue system, Markov decision process.

Citation: W. C. Kwan, H. R. Wang, H. M. Wang, K. F. Wong. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intelligence Research*. http://doi.org/10.1007/s11633-022-1347-y

1 Introduction

Task-oriented dialogue (TOD) system aims to assist users in accomplishing tasks ranging from weather inquiries to schedule planning[1]. It can be classified into two approaches. The first is the end-to-end approach, which directly maps the current dialogue context to the system′s natural language response[2−5]. These works often adopt a sequence-to-sequence model and train in a supervised manner. The second is the pipeline approach, which separates the system into four interdependent components: Natural language understanding (NLU), dialogue state tracking (DST), dialogue policy learning (DPL) and natural language generation (NLG), as shown in Fig. 1[6].

Both of these methods have their own limitations and advantages. The end-to-end approach is more flexible and has fewer requirements for data annotations. However, it requires a large amount of data and its black box structure provides no interpretation and little control[7]. On the flip side, the pipeline approach is more interpretable and easier to implement. Although the whole system is harder to optimize globally, the pipeline approach is preferred by most commercial dialogue systems[6]. Our survey also falls under the pipeline category to investigate and summarize the current progress of dialogue policy learning. We will briefly introduce the different functions of these four modules and then look deeper into the dialogue policy learning module.

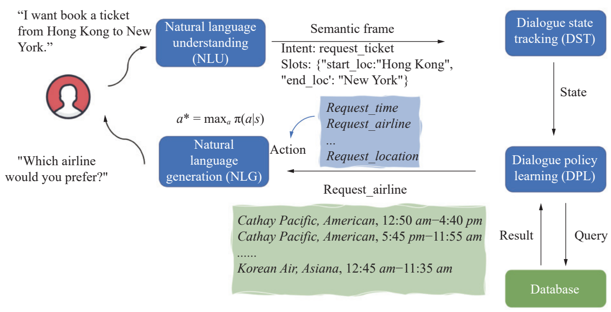


Fig. 1 An overview of a task-oriented dialogue system. All blue parts represent the four components in the pipeline dialogue system.

Among these four modules, NLU aims to identify the intentions and slots from the input sentence as the first module that interacts directly with the user. Then, the DST module represents all previous extracted intentions and slots as an internal dialogue state. Next, the DPL module performs an action to satisfy the user′s intent given the state as input. Finally, the NLG module transforms and outputs the action in natural language form. In

Review
Manuscript received May 2, 2022; accepted June 6, 2022
Recommended by Associate Editor Zhi-Yuan Liu
†These authors contributed equally to this work

this pipeline, DPL plays a key role in TOD as an intermediate connection between the DST and NLG modules, which directly affects the success of the dialogue system[6].

Recently, the progress in DPL has been significantly facilitated by the development of reinforcement learning (RL) algorithms[8–11]. Levin et al.[8] are the first to treat DPL as a Markov decision process (MDP) problem. They outline the complexities of modelling DPL as an MDP problem and justify the application of RL algorithms to optimize the dialogue policy[8]. Thereafter, the majority of studies attempt to investigate and resolve the technical issues that arise when applying RL algorithms to dialogue systems practically[9, 12, 13]. At the other end of the spectrum, several researchers explored the use of supervised learning (SL) techniques in DPL[10, 11, 14–16]. The main idea is to treat the dialogue policy learning as a multi-class classification problem, with actions and states acting as labels and inputs, respectively. However, SL techniques have a notorious and unaffordable flaw since they do not consider the future effects of the current decision, resulting in sub-optimal behaviour[14].

With the breakthroughs in deep learning, deep reinforcement learning (DRL) methods that combine neural networks with RL have recently led to successes in learning policies for a wide range of sequential decision-making problems. This includes simulated environments like the Atari games[17], the chess game Go[18], and various robotic tasks[19, 20]. Following that, DRL has received a lot of attention and achieved promising results, mainly in single-domain dialogue scenarios[21–24]. The neural models can extract high-level dialogue states and encode complicated and long language utterances. This was the biggest challenge that early works faced[8, 9]. As the focus of DPL research has slowly gravitated to more complicated multi-domain datasets, many RL algorithms face scalability problems[25].

Recently, there has been a flurry of works that focus on ways to adapt and improve RL agents in multi-domain scenarios. Few works attempt to review the vast literature on recent applications of reinforcement learning (RL) in DPL of TOD systems. Grassl surveyed the use of RL in the four types of dialogue systems, namely social chatbots, infobots, task-oriented, and personal assistant bots[26]. However, the progress and challenges of using RL in TOD systems were not well discussed. Similarly, Dai et al.[27] reviewed the recent progress and challenges of dialogue management, which only contained a limited discussion on RL methods in DPL due to its wide scope of interest. Furthermore, RL dialogue systems often have different settings in the five core RL elements, namely environment, policy, state, action, and reward. Previous surveys did not consider the inconsistent settings of different systems, which resulted in an unfair comparison among these systems.

In this survey, we describe the unique strengths of previous works and categorize them based on the five elements of RL. Then we focus our discussion on three main recent challenges of applying RL to DPL, namely exploration efficiency, cold start problem, and large state-action space. Most recent works using RL to optimize DRL attempt to address these challenges. The procedure which we use to shortlist these works for review is provided in Appendix. The remainder of this paper is organized as follows. In Section 2, we illustrate the problem definition of DPL and elaborate on the challenges of using RL to train a dialogue agent in TOD systems firstly. Then, we introduce our methodology to characterize recent DPL works. The methodology is motivated by the fact that the key differentiating aspect of recently proposed methods can be boiled down to the differences in these five fundamental elements of RL. In this case, it is easy and self-evident to find similarities and differences between different methods. Furthermore, this helps identify the key component of each work that contributed the most to its improvement. The state-of-the-art techniques of recent DPL works categorized by the five RL elements are discussed in detail separately in Sections 3–7. In Section 8, we discuss the current status of DPL research with RL. In Section 9, we present the challenges in applying RL dialogue agents in real-life scenarios and three promising future research directions. Finally, we conclude the survey in Section 10.

2 Overview

2.1 Problem definition and annotations

Given a dialogue state that encodes the previous interactions, the dialogue policy decides the next action to perform. Fig. 1 shows an example of a dialogue turn. After DST updated the belief state with the location information, the policy decided to request the airline preferred by the user. The goal of DPL is to learn a dialogue policy that generates the satisfactory next system action that answers the user′s query. DPL is often formulated as an MDP problem, and RL is often used to optimize the policy[6, 28–33]. Formally, an MDP is defined as a five-element tuple ($\mathcal{S}$, $\mathcal{A}$, P, R, γ). $\mathcal{S}$ refers to the dialogue state space that holds the necessary information for the policy to make a decision. $\mathcal{A}$ refers to the set of all system actions. $P(s'|a, s)$ refers to the transition model $S \times A \times S \to [0, 1]$ of the environment. $R(s, a)$ is the reward function $S \times A \to \mathbf{R}$ that provides an immediate reward for each turn. $\gamma \in (0, 1]$ is the discount factor that indicates the effect of future rewards. Sutton and Barto[34] provided a comprehensive introduction to RL methodologies.

A full turn of dialogue interactions can be viewed as a trajectory $(s_1, a_1, r_1, s_2, a_2, r_2, \cdots)$, which is generated by the following process at each step as depicted in Fig. 2. First, the dialogue agent observes the current dialogue states $s_t \in \mathcal{S}$ and responds with an action $a_t \in \mathcal{A}$. Second,

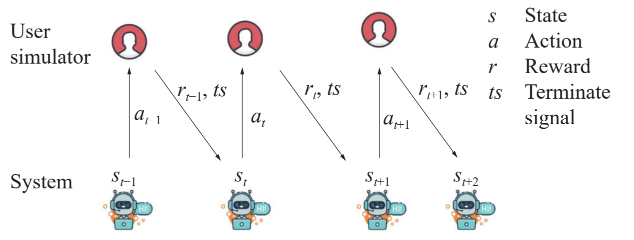


Fig. 2 Framework of Markov decision process in DPL. At time t, the system takes an action a_t, receiving a reward r_t and a terminate signal ts and then transferring to a new state s_{t+1}.

the environment[1] receives the action and transits to a new state $s_{t+1} \in \mathcal{S}$ according to the transition model P. Third, the environment provides a reward r_t and terminate signal ts after transiting to a new state. At each step t, this process gives us a tuple (s_t, a_t, r_t, s_{t+1}) which is called a transition. The goal of the RL agent is to learn an optimal deterministic policy $\pi : S \to A$ that maximizes the value function, which is the expected total discounted returns in a trajectory. It is formally defined as

$$V^{\pi}(s) = E\left[\sum_{t=0}^{T} \gamma^t r_t | s_0 = s\right]$$

where γ is the discounting factor and s_0 is the initial state. Equivalently, the policy can also maximize the Q-function, which is defined as

$$Q^{\pi}(s, a) = E\left[\sum_{t=0}^{T} \gamma^t r_t | s_0 = s, a_0 = a\right].$$

The value function can be derived from the Q-function by

$$V^{\pi}(s) = \max_{a \in \mathcal{A}} Q(s, a).$$

2.2 Recent challenges in applying RL

Recently, neural models have started to have a sufficient capacity to encode the long context in dialogues. It has played a big role in recent DRL methods in dialogue systems. However, moving towards more complicated dialogue scenarios have been difficult because the possible combinations of states grow exponentially with the number of actions[35]. More specifically, three major challenges appear and attract much attention: Exploration efficiency, cold start problem, and large state-action space.

Exploration efficiency. RL methods need to interact with an environment to collect enough interactions for training. In the dialogue system setting, the agent is required to interact with real users[36], which is expensive and time-consuming. In practice, the agent interacts with a rule-based user simulator[21, 37]. The exploration efficiency depends on how closely the simulator resembles human behaviour, which is not easy[28, 38]. It is laborious to build a high quality and specialized user simulator for each dataset.

Cold start problem. A poorly initialized policy may lead to low-quality interactions with users in online learning settings[39]. Having rare successful experiences causes the model to learn slowly in the beginning and discourages real users from interacting with the system[40, 41].

Large state-action space. DPL for some complex dialogue tasks, such as multi-domain involves a large state-action space[30, 42]. The dialogue agent is required to explore this large space and often takes many conversation turns to fulfil a task. The long trajectory results in a delayed and sparse reward, which is usually provided at the end of a conversation[29].

2.3 A method to characterize RL approaches

Recently, many researchers have been trying to tackle the three aforementioned challenges. The RL system comprises five elements: Environment, policy, state, action, and reward. Each work in DPL using RL can be summarized by how it configures these five elements. This motivates us to characterize the recent approaches in RL dialogue agents based on the five elements of RL. Since different RL dialogue agents usually have very different configurations of these five elements, it is difficult to compare them and identify the key components contributing to the improvement. Therefore, this survey breaks down the recent work into these five elements and describes the various configurations for each element separately. This method allows us to identify the focal points of recent advancements in RL methods in DPL research. Table 1 provides an overview of the different RL methodologies used in DPL.

3 Environment

In a typical scenario of a DPL, there are two speaker roles: user and system. Most of the current methods are single-agent that only model the system side, and treat the user side as the environment[21, 23, 33, 44, 47, 61, 63]. Some methods model two roles in n dialogues[28, 56, 60], and some works consider multi-person (more than two persons) dialogue. This section illustrates 1) different methods to build a user simulator (i.e., the environment) and 2) ways to model different agents simultaneously. The various work mentioned in this section directly tackled the exploration efficiency problem by improving the quality and efficiency of building a user simulator.

3.1 Single-agent/User simulator

Most previous works build a user simulator first and interact with the single system agent using the simulator

[1] Here, the environment is a user simulator.

Table 1 An overview of the configurations of recent works on DPL with RL approach

Model	Dataset	RL algorithm	Experience replay	Simulator		Annotations		Expert demo		Reward function
				Granularity	Methodology	Belief state	Dialogue act	IL	Supervised buffer	
TSL[43]	Calendar	Q-learning	√	Utterance level	Rule-based	√	√	√		Other
RNN reward shaping[44]	CamRes	GP-SARSA		Dialogue-act level	Agenda-based	√	√			Reward shaping
End-to-end RL[45]	20 Question game	DRQN	√	Utterance level	Agenda-based	√	√	√		Manually defined
Continuous learning[21]	CamRes	NAC	√	Dialogue-act level	Agenda-based	√	√	√		Manually defined
Two-stage training DQN[22]	DSTC2	GPSARSA, DA2C, TDA2C, DQN, DDQN	√	Dialogue-act level	Agenda-based	√	√	√		Manually defined
Option framework[46]	Pydial	HRL	√	Dialogue-act level	Agenda-based	√	√			Manually defined
BBQN[24]	Amazon movie-ticket	DQN	√	Dialogue-act level	Agenda-based	√	√	√		Others
IPLDM[28]	DSTC2	REINFORCE, multi-agent		Dialogue-act level	Multi-agent	√	√	√		Manually defined
CTCDS[47]	Frames	HRL	√	Dialogue-act level	Agenda-based	√	√		√	Manually defined
TRACER, eNACER[23]	CamRes	GPRL, TRPO	√	Dialogue-act level	Rule-based	√	√	√	√	Manually defined
CT[39]	DSTC2	DQN	√	Dialogue-act level	Agenda-based	√	√			Manually defined
ACER[48]	CamRes	Actor-critic/ TRPO/IS	√	Dialogue-act level	Agenda-based	√	√		√	Manually defined
ALDM[29]	DSTC2	Policy gradient		Dialogue-act level	Multi-agent	√	√	√		AL-IRL
Adversarial A2C[30]	Amazon movie-ticket	Actor-critic	√	Dialogue-act level	Agenda-based	√	√	√	√	AL-IRL
DDQ[31]	Amazon movie-ticket	Dyna-Q, actor-critic	√	Dialogue-act level	World model	√	√	√	√	Manually defined
HER[40]	Amazon movie-ticket	DQN	T-HER/ S-HER	Dialogue-act level	Agenda-based	√	√			Manually defined
FDQN[49]	PyDial	Feudal RL	√	Dialogue-act level	Agenda-based	√	√			Manually defined
Option framework[50]	Pydial	HRL	√	Dialogue-act level	Agenda-based	√	√			Manually defined
D3Q[51]	Amazon movie-ticket	Dyna-Q	√	Utterance level	World model	√	√	√	√	Manually defined
SDN[52]	Frames	HRL	√	Utterance level	Agenda-based		√			Manually defined
Switch-DDQ[53]	Amazon movie-ticket	Dyna-Q	√	Utterance level	World model	√	√	√	√	Manually defined
D3Q[51]	Amazon movie-ticket	Dyna-Q	√	Utterance level	Agenda-based	√	√	√	√	Manually defined
LaRL[54] ◇	DealOrNoDeal/ MultiWOZ	REINFORCE		Utterance level	Data-driven				√	Manually defined
Meta-DTQN[55]	MultiWOZ	DQN/Dual replay	√	Dialogue-act level	Agenda-based	√	√			Manually defined
WoLF-PHC[56]	DSTC2	WoLF-PHC		Dialogue-act/ Utterance level	Multi-agent	√			√	Manually defined
BCS-DDQ[57]	Amazon movie-ticket	Dyna-Q	√	Dialogue-act level	World model	√	√		√	Manually defined
GDPL[58]	MultiWOZ	PPO		Dialogue-act level	Agenda-based	√	√		√	AL-IRL

Table 1 (continued) An overview of the configurations of recent works on DPL with RL approach

Model	Dataset	RL algorithm	Experience replay	Simulator		Annotations		Expert demo		Reward function
				Granularity	Methodology	Belief state	Dialogue act	IL	Supervised buffer	
LHUA[32]	Amazon movie-ticket	DQN	T-HER/ S-HER	Dialogue-act level	Agenda-based	√	√			Manually defined
Act-VRNN[59]	MultiWOZ	ELBO	√	Dialogue-act level	Agenda-based	√	√		√	Others
OPPA[60]	MultiWOZ	DQN	√	Dialogue-act level	Agenda-based		√	√		Manually defined
GDPL w/o AL[61]	MultiWOZ	PPO, DQN	√	Dialogue-act level	Rule-based	√	√			AL-IRL
MADPL[62]	MultiWOZ	Actor-critic, multi-agent	√	Dialogue-act level	Multi-agent	√	√	√		Manually defined
DQfD[33]	MultiWOZ	DQN	√	Dialogue-act level	Agenda-based	√	√		√	Manually defined
RoFL[42]	MultiWOZ	DQN	√	Dialogue-act level	Agenda-based	√	√	√	√	Manually defined

to obtain a large number of simulated user experiences for RL algorithms. Building a reliable user simulator, however, is not trivial and often requires much expert knowledge or abundant annotated data[62]. There are two major methods to build a user simulator.

Agenda-based simulator. With the growing need for the dialogue system to handle more complex tasks, it is very challenging and laborious to build a fully rule-based user simulator, which requires extensive domain knowledge and expertise. An agenda-based simulator[37, 64–66] starts a conversation with a randomly generated user goal that is unknown to the dialogue manager. It keeps a stack data structure (i.e., user agenda) during the course of the conversation. Each entry in the stack maps to an intention the user aims to achieve, and the order follows the first-in-last-out operation of the agenda stack[67]. An agenda-based simulator stores all the information the user needs to inform and acquire. It acts according to pre-defined rules. An example of a dialogue and the corresponding agenda sequence are shown in Fig. 3. The C_0 refers to the user constraints on the venue, and R_0 specifies the information of the venue required by the user. Sys t and Usr t are the system response and user utterance at turn t, respectively. A_t is the user agenda at turn t. Usr t is generated based on the intention(s) popped from the top of the agenda stack A_t. For example, the user utterance at turn 1 (i.e., Usr 1) "I am looking for a nice bar serving beer." is based on the two intentions "inform (type = bar)" and "inform (drinks = bar)" popped from the user agenda at turn 1 (i.e., A_1).

Data-driven simulator. Another method to build a user simulator is to utilize a sequence-to-sequence framework. Its goal is to generate user responses (utterance or dialogue actions) based on the current dialogue context[68]. The dialogue context consists of historical dialogue content, dialogue goal, constraint status, and request status. This method can be learned and optimized directly from a large amount of human-human dialogue corpora[69–72].

Although the data-driven approach is able to construct a user simulator without much engineering, it is hard to evaluate the quality of a user simulator as it is unclear to define how closely the simulator resembles real user behaviours[73–75]. The gap between the user simulator and humans renders dialogue policy optimization difficult[67].

3.2 Multi-agents

The goal of RL is to discover the optimal strategy $\pi^*(a|s)$ of the MDP. It can be extended into the N-agents setting, where each agent has its own set of states S_i and actions A_i. In multi-agent reinforcement learning (MARL), the state transition $s = (s_1, \cdots, s_N) \longrightarrow s' = (s'_1, \cdots, s'_N)$ depends on the actions taken by all agents $(a_1, \cdots, a_N)$ according to each agent's policy $\pi_i(a_i|s_i)$ where $s_i \in S_i$, $a_i \in A_i$. Similar to single-agent RL, each agent aims to maximize its local total discounted return $R_i = \sum_t \gamma^t r_{i,t}$.

Instead of employing a user simulator, it was demonstrated that an user agent and dialogue agent learning concurrently by interacting with each other can achieve satisfactory performance in a negotiation scenario without a rule-based simulator[76]. Liu ang Lane[28] made the first attempt to apply MARL to the task-oriented dialogue policy to learn the system policy and user policy concurrently. It optimizes two agents from the corpus by iteratively training the system policy and the user policy with the policy gradient method. Thereafter, WoLF-PHC was applied within the MARL framework to the task-oriented dialogue policy[56], which is based on Q-learning for mixed policies to achieve faster learning. Following this line of research, Takanobu et al.[62] extended the MARL framework to handle multi-domain dialogue by using the

$$C_0 = \begin{bmatrix} type = bar \\ drinks = beer \\ area = central \end{bmatrix}$$

$$R_0 = \begin{bmatrix} name = \\ addr = \\ phone = \end{bmatrix}$$

Sys 0 Hello, how may I help you?

$$A_1 = \begin{bmatrix} inform(type = bar) \\ inform(drinks = beer) \\ inform(area = central) \\ request(name) \\ request(addr) \\ request(phone) \\ bye() \end{bmatrix}$$

Usr 1 I'm looking for a nice bar serving beer.

Sys 1 Ok, a wine bar. What pricerange?

$$A_2 = \begin{bmatrix} negate(drinks = beer) \\ inform(pricerange = cheap) \\ inform(area = central) \\ request(name) \\ request(addr) \\ request(phone) \\ bye() \end{bmatrix}$$

Usr 2 No, beer please!

Sys 2 A bar serving beer, correct?

$$A_3 = \begin{bmatrix} affirm() \\ inform(pricerange = cheap) \\ inform(area = central) \\ request(name) \\ request(addr) \\ request(phone) \\ bye() \end{bmatrix}$$

Usr 3 Yeah something cheap in the town centre.

Sys 3 Murphy's on Main Square serves cheap beer.

$$A_4 = \begin{bmatrix} request(phone) \\ bye() \end{bmatrix}$$

Usr 4 Ok, and what's the phone number?

Sys 4 The number is 796 69 94.

$$A_5 = \begin{bmatrix} bye() \end{bmatrix}$$

Usr 5 Thanks, goodbye!

Fig. 3 A dialogue sample and agenda. C_0 and R_0 specify the user's constraints and the information required by the user. Sys 0 refers to the first dialogue initiated by the system. A_t, Usr t and Sys t are the agenda stack, user utterance, and system response at turn t, respectively[37].

actor-critic framework instead of dealing with the large discrete action space in dialogue[62]. Recent work extended the traditional two-agent to three-agent, leading to a smaller action space and faster learning[77]. Another work explored the MARL framework from a different perspective[78]. They use MARL in the policy committee framework, where each policy decides an action on its own and is combined by a gating mechanism.

4 Policy

In this section, we firstly divide different DPL methods into two categories: Model-free reinforcement learning and model-based reinforcement learning. Furthermore, the former method is divided into hierarchical reinforcement learning (i.e., HRL)[79, 80] and feudal reinforcement learning (i.e., FRL)[81]. Noticeably, HRL and FRL alleviate the large state-action space problem by decomposing the state-action into smaller ones, while the Deep dyna-Q[31] models enhance the exploration efficiency by modelling the environment dynamics. In addition, most of these methods require warm-up before training, which alleviates the cold start problem. The details of the warm-up method are discussed below.

4.1 Model-free RL-HRL

Solving composite tasks, which consist of several inherent sub-tasks, remains a challenge in the research area of dialogue systems. For instance, a composite dialogue of making a hotel reservation involves several sub-tasks, such as looking for a hotel that meets the user's constraints, booking a room, and paying for the room. HRL decomposes complex tasks into several subtasks and learns different policies for these subtasks from top to low-level[46, 47, 50]. As shown in Fig. 4, the top-level policy decides which option (i.e., subtask) $w \in \Omega$ should be chosen, and the low-level dialogue policy selects the primitive actions $a \in \mathcal{A}$ to complete the subtask given by the top-level policy. It is noted that a primitive action is an action lasting for one time step, while an option is an action lasting for several time steps. HRL can be further divided into sub-domain or sub-goal hierarchical reinforcement learning.

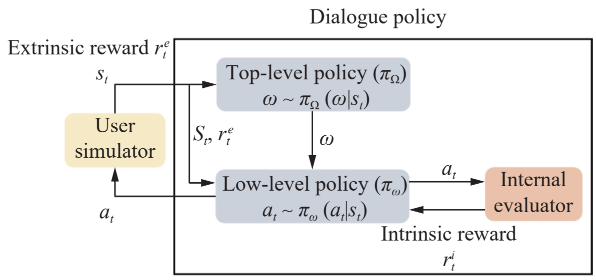


Fig. 4 The overview of two levels of policies in hierarchical reinforcement learning[47]

Sub-domain. Some works used the options framework[82] to solve the above problem with different approximators[46, 47]. However, each option (i.e., sub-task) and its property (e.g., starting and terminating conditions, and valid action set) had to be manually defined. Bacon et al.[83, 84] proposed a unified framework that integrated option discovery and achieved comparable performance with manually defined options framework[50].

Sub-goal. Instead of decomposing a task according to the corresponding domain, it is also an option to divide a complex goal-oriented task into a set of simpler subgoals. The subgoal discovery network (SDN)[52] was proposed to discover and exploit the hidden structure of the task to enable efficient policy learning inspired by the sequence segmentation model[85].

4.2 Model-free RL-FRL

Feudal reinforcement learning (FRL)[86] is another interesting attempt to solve the large state and action space problem. FRL decomposes a task spatially to restrict the action space of each sub-policy, whereas HRL decomposes a task temporarily to solve a different subtask in a different time step[27, 67]. Casanueva et al.[49] are the first to apply FRL to task-oriented dialogue systems and decomposes the decision into two steps based on its relevance with slots: A master policy is chosen to select a subset of primitive actions in the first step, and a primitive action is chosen from the selected subset at the second step. The decisions in different steps use different parts of the abstracted states. Furthermore, Casanueva et al.[87] showed that feature extraction could be learned jointly with the policy model. It obtained a similar performance with the handcrafted features in feudal dialogue management.

In contrast to HRL, which decomposes a task into temporally separated subtasks, FRL decomposes a complex decision spatially[67]. Although both HRL and FRL can be used to address large dimension issues, both have limitations. The decomposition in HRL often requires expert knowledge, while FRL does not consider the mutual constraints between sub-tasks[27].

4.3 Model-based RL

Different from model-free RL methods, model-based RL models the environment to decide the transition of states, enabling planning for dialogue policy learning[6]. Deep dyna-Q (DDQ)[31] is the first deep RL framework that integrates planning for task-completion DPL. It effectively leverages a small number of real conversations. Specifically, the environment is modelled as a world model to mimic the real user response and to generate a simulated experience. Recently, more DDQ variants have been proposed to improve the quality of simulated experience through adversarial training[51], active learning[53], and human teaching[57].

4.4 Warm-up by imitation learning

Imitation learning (IL) enables the policy to learn from the expert demonstrations without exploring the environment. Fig. 5 shows the architecture of imitation learning. The policy is first pretrained with the human demonstrations. Then, the pretrained policy interacts with the environment to collect experiences for RL fine-tuning. This leads to effective initialization in the warm-up stage[88]. With limited warm-up steps based on a small number of expert demonstrations, the learning speed of the dialogue RL agent can be accelerated[21–23, 28, 29, 31]. However, another line of work points out that IL requires expert demonstrations and the transition dynamics of the RL environment to have the same distribution, which is often not the case in DPL[22, 23]. Thus, it is critical to follow up on the IL with different RL methods[28, 31].

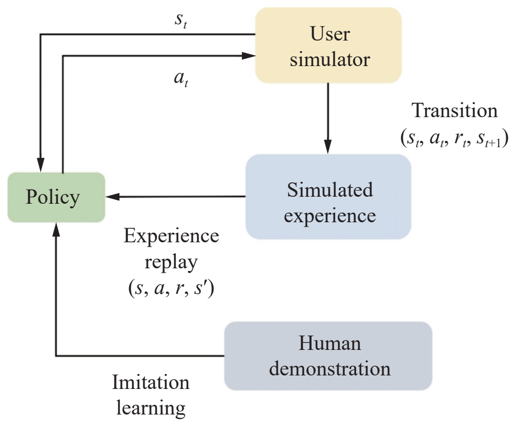


Fig. 5 RL architecture of using imitation learning

5 State space

The dialogue state encodes the essential information in the dialogue history for the dialogue policy to generate the next system action. There are mainly three types of state representations used in recent research, namely multi-hot, distributed, and multi-modal representations. This section explores the difference and effectiveness of these state representations and how they tackle the large state space problem.

5.1 Multi-hot representation

Most works using the multi-hot representation are based on a belief vector. This vector concatenates the one-hot vector based on the value for each slot[55, 58, 62, 89]. In addition to the belief vector, many of these works also incorporate the one-hot vector of the current user action, previous system action, database vector that indicates the number of query results, the repeated times of the last user action, etc.[58, 61]. These multi-hot representations are often simple to implement but require feature engineering. As the number of domains increases, the state space grows exponentially with the size of the one-hot representation of the actions. Therefore, the large state space problem exists in multi-hot representation.

5.2 Distributed representation

Some works avoid feature engineering by directly encoding the user's utterances as state representations[28, 31, 45, 53, 90]. Among these works, few adapt a feed-forward network that takes the *n*-gram features of the previous system response and the current user utterance as input[31, 53, 90]. Others use an long short term memory (LSTM) network[91] to encode the utterances of both parties[28, 45]. The latter approach is able to further capture the turn-level dynamics instead of just the current turn. By using a distributed representation, the state encoder and the policy network are jointly optimized together to achieve better performance. By utilizing the neural work to encode the dialogue history into a compact distributed representation, it can learn features that are invariant of the domain and enables the representation to scale with the increasing number of domains. Thus, the distributed representation can tackle the large state space problem.

5.3 Multi-modal representation

Conversation involves multiple modalities, especially in social media like Facebook[2] and WeChat[3]. For a dialogue system, understanding vision and language is one of the ultimate goals for creating intelligent conversational system[92]. Zhang et al.[93, 94] enrich the state representation with multi-modal information. Zhang et al.[93] proposed a framework to jointly learn the multi-modal dialogue state representation and the hierarchical dialogue policy. It improved the task success rate and enhanced the efficiency in an image guessing task. A later work incorporated sentiment representations in addition to image representation into the state and fed it to the policy as input[94].

6 Action space

Most works treat the action space as a set of dialogue acts. A dialogue act is specified by a dialogue act type that indicates the type of action the user/agent is performing, and a set of slot-value pairs that specify the imposed constraints[95]. There are two prominent problems in the action space of TOD systems. First, most methods can only predict a single dialogue act per turn. Second, the action space is large in complex dialogue scenarios. These two problems and the recent related research are discussed in detail below. In addition, a stream of works that use natural language as an action space and directly generate a system response to the user by integrating DPL and NLG is also covered.

6.1 Large action space

Chen et al.[35] pointed out that having a separate set of dialogue acts for each domain was not scalable as we worked toward multi-domain large-scale scenarios. Multi-domain dialogue scenarios involve a large action space since it includes multiple domains, and each dialogue act is represented as a (domain-action-slot) triple. The number of dialogue acts grows exponentially with the number of domains. To alleviate this problem, Chen et al.[35] proposed a multi-layer hierarchical graph that exploited the structure of dialogue acts. However, this work relies on having the dialogue act annotations. Zhao et al.[54] took another approach to treat the action space as a latent variable and used an unsupervised method to induce an appropriate action space from the data without having the dialogue act annotations. They optimize the model with RL by applying policy gradient methods in the latent action space. These works show promising results in multi-domain dialogue scenarios with large action space.

6.2 Multiple dialogue action

Most works did not address the one-to-many property of conversations where there might be multiple valid system actions that satisfy the same user query[96, 97]. An intelligent conversational agent should consider this multi-action characteristic. Shu et al.[98] formulated multiple action dialogue policy learning as a sequence to sequence problems and design a unique output format (e.g., continue, act, slots) to generate multiple actions per turn. Zhang et al.[96] proposed multi-action data augmentation (MADA) framework to enable dialogue models to learn a more balanced state-to-action mapping. Li et al.[97] modelled the one-to-many property by retrieving multiple candidate actions and selectively taking the candidates into consideration when generating system action. On the whole, these works enhance the expressiveness of the model with the ability to generate multiple dialogue acts in one turn.

6.3 Integrate DPL with NLG

At the other end of the spectrum, some researchers explored integrating DPL and NLG by using the policy to directly output utterance responses instead of dialogue actions[5, 99]. Wang et al.[5] treated dialogue act prediction as another sequence generation problem along with the response generation task. It used a share encoder to encode the previous utterances and fed it to the dialogue act generator and the response generator separately. While this work took a supervised learning approach, Wang et al.[99] modelled the hierarchical structure between DPL and NLG using the options framework[82] to improve the comprehensibility of generated system utterances. Both works demonstrated that by using natural language as the action space, the model was able to generate a natural and realistic response by exploring the semantic associations between dialogue acts and the output utterance.

[2] https://www.facebook.com/
[3] https://www.wechat.com/

7 Reward learning

Reward contains essential information which guides the RL agent towards the goal. Most of the works adopted manually designed reward functions that gave large positive and negative rewards for success and failed dialogue, respectively, and a small negative turn level reward to encourage shorter dialogue[22, 23, 31, 36, 40, 47, 48, 50–53, 63, 100]. However, the sparse reward signal of successes is one of the reasons that RL agents learn slowly, especially in the beginning stage[58, 101].

There are two streams of work that aim to learn a denser reward to encourage faster learning and tackle the cold start problem in RL, making use of the available expert demonstrations, namely inverse reinforcement learning (IRL) based methods and reward shaping. Figs. 6 and 7 show the pipeline of IRL methods and reward shaping, respectively. IRL learns a reward function given the human demonstrations, which is used to provide a reward for transitions in RL training, whereas reward shaping provides an additional reward given the human demonstrations to complement the environment reward.

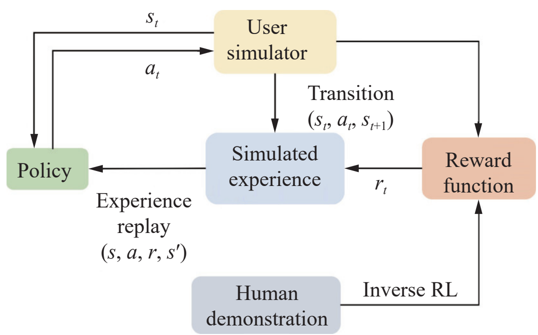


Fig. 6 An overview of inverse reinforcement learning

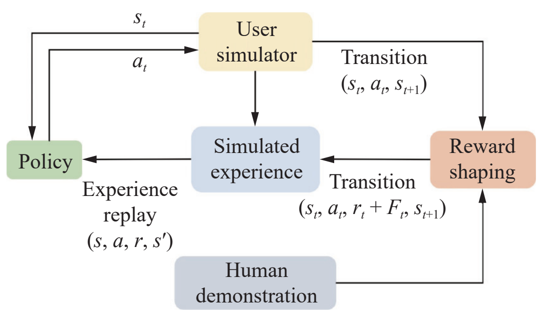


Fig. 7 Reward shaping

7.1 Inverse reinforcement learning method

IRL refers to the problem of learning a reward function given the expert demonstrations[102]. It is appealing since designing a good reward function is tedious and difficult in complicated domains. As a result, it has attracted a lot of research work[88, 88, 103, 104]. Boularias et al.[105] are the first to explore this idea in DPL to learn a reward function from a human expert in a Wizard-of-Oz setting. They proposed a reward function which is a linear combination of feature vectors with unknown weights. The weights can be first learned from the expert demonstrations, and then the learned reward function is used in RL. The learned reward function can provide meaningful feedback to the policy, which helps it learn effectively, especially in the early stage. Another work explored learning a classifier to estimate the expert policy. The reward function can be inferred by giving higher rewards $R(s, a)$ to those experts who agree more (i.e., higher $P(a|s)$ in the model)[43].

Despite the success of using IRL in the dialogue scenarios, IRL is often expensive to run, hindering it to scale to complex dialogue scenarios[106]. In the RL community, adversarial IRL (AL-IRL) is proposed to enhance the efficiency of learning the reward from expert demonstrations[106]. It avoided doing reinforcement learning in an inner loop of a training procedure. Liu and Lane[29] explored AL-IRL in DPL and used the discriminator to differentiate successful dialogues from unsuccessful ones. The discriminator′s output which is the probability of a given dialogue being successful, is used as the reward in policy optimization. Extending this line of research, Takanobu et al.[58] further combined AL with maximum entropy IRL to learn the policy and reward estimator alternatively.

7.2 Reward shaping

Reward shaping aims to incorporate domain knowledge into RL by introducing an extra reward in addition to the reward provided by the environment. Ng et al.[107] represented the reward shaping mathematically as the difference of any potential function $\phi(s)$ on two consecutive states s_t and s_{t+1}. The potential-based reward shaping does not affect the optimal solution of the MDP but speeds up learning.

In DPL, many works took advantage of additional information to formulate the potential function. The earliest work took advantage of the availability of the evaluation scores of the dialogues given by humans[108]. The potential function was inferred using distance minimization inverse reinforcement learning. Ferreira and Lefévre.[109] proposed learning an extra reward from the social cues of the user. In this work, they mainly consider the sentiment cues from the user-defined manually, including the type of dialogue acts, number of slots filled, agenda size, etc. While this method does not require extra annotated data, the manually defined features are not scalable to other domains. Wang et al.[101] took advantage of human demonstrations and used a multi-variate Gaussian to pick the most similar state-action pair to complement the main reward. They extended the potential function $\phi(s)$ to $\phi(s, a)$, which received the action as an additional input. They used a multi-variate Gaussian to compute and picked the highest similarity between the current state-action pair with the expert demonstrated state-action pair as the potential function.

Overall, these papers highlight the benefit of using a dense reward in DPL. An important difference between the inverse reinforcement learning method and reward shaping is that the former learns one single reward function, while the latter adds a reward function in addition to the main reward provided by the environment.

8 Discussions

TOD systems demonstrate satisfactory performance in many scenarios, including movie ticket booking[47], restaurant enquiry, and even multi-domain scenarios[110]. However, most of these techniques require tremendous expert demonstrations during training, as shown in Table 1. Moreover, most of the works rely on automatic evaluation through interacting with a simulator to validate the improvements. These limitations restrain us from applying these techniques to TOD systems in real-world scenarios.

On the one hand, the availability of high-quality annotated training data poses a major obstacle in applying TOD systems in real-world scenarios[111]. For some low-resource domains or complex tasks with significant costs to collect data, it is difficult to develop a robust TOD system with a limited budget. In this case, an effective and resource-saving method such as transfer learning is necessary[112]. On the other hand, the lack of solid evaluation criteria for automatic assessment of dialogue quality causes unstable optimization and performance[110]. For example, the discrepancies between the behaviours of real and simulated users inevitably lead to a sub-optimal dialogue policy. Most of the existing methods fail to generalize in an open and changing world (i.e., the real world) due to these problems.

9 Future direction

With the progress of RL methods, the three challenges are alleviated by a variety of techniques introduced in previous sections. The exploration efficiency is improved by multi-agents reinforcement learning techniques that learn the simulator without human interventions and model-based approaches which learn the environment dynamics. The cold start issue is alleviated by reward learning, which provides a more useful reward to guide the dialogue agent to learn effectively. The large state-action space problem is mitigated by HRL and FRL that decompose a task into subtasks and different methods that model the state-action space in a more compact manner.

Recent work demonstrated that TOD systems achieved satisfactory performance in many scenarios, including movie ticket booking[47], restaurant enquiry, and even multi-domain scenarios[110]. However, most of these techniques require tremendous expert demonstrations during training, as shown in Table 1. Moreover, most works rely on automatic evaluation through interacting with a simulator to validate the improvements. These limitations restrain us from applying these techniques to TOD systems in real-world scenarios. As the objective of a TOD system is to help users achieve their goals, future research should aim toward applying TOD in a real-world scenario. In this section, we elaborate on the two future directions (i.e., data scarcity and reliability of evaluation), as well as the latest work moving towards these directions.

Data scarcity. There are many different real-world dialogue scenarios such as restaurant booking, weather queries and flight booking. It is extremely costly to obtain a large amount of annotated data for different domains. However, the most recent methods presented in this survey often require a lot of expert demonstrations. As a result, for a TOD system to be practical, techniques and methods to learn a dialogue policy expeditiously and effectively in domains that have scarce data should be developed. Domain adaptation and meta policy learning are two effective and auspicious solutions to tackle this problem.

Reliability of evaluation. It is important to evaluate the performance of a dialogue policy in accomplishing human goals in different scenarios. Currently, the most widely used way to evaluate a dialogue policy is by interacting with a user simulator. However, there is often a behavioural discrepancy between the user simulator and the human user, as discussed in Section 3. Therefore, this evaluation method does not correctly reflect how well a dialogue policy can assist a human in completing his/her tasks. Two promising future directions for tackling the data scarcity problem and the key aspects of reliable evaluation methods are described below.

9.1 Data scarcity problem

Domain adaptation. Domain adaptation or policy transfer enables us to build a dialogue policy in a target domain with scarce data provided with a large amount of data in a source domain. In [113], they proposed a multi-agent dialogue policy (MADP) which consists of some slot-dependant agents that have shared parameters for every slot. The shared parameters can be transferred to a new domain for the common slots. Similarly, Ilievski et al.[111] matched the state space and the action space between the source domain and the target domain even if these actions/slots were never used in the source domain. The parameters of the common slots and actions are used in the target domain initially. However, different domains do not necessarily have common actions or consistent dialogue act naming. The PROMISE model is proposed to learn the similarity between slots and actions of different domains[114]. While these researches focus on domain adaptation between two domains, much work is required to adapt to multi-source domains.

Meta policy learning. To further extend the usage of DPL to a real-world scenario, we consider situations with even harsher data resource. In the previous section, we leveraged the abundant data in a source domain. In this section, the meta-learning paradigm tackles the situation in which all domains have scarce data. Recently, Mi et al.[115] adopted meta-learning in the NLG module in the spoken dialogue system (SDS) pipeline. Inspired by this work, some researchers proposed the deep transferable Q-network (DTQN), which leverages shareable features across domains[55]. They further combine DTQN with model-agnostic meta-learning[116] with a dual-replay mechanism to support effective off-policy learning, which helps models to adapt to an unseen domain quickly. In [57], they extended DDQ by incorporating budget-conscious scheduling to learn from a fixed, small amount of interactions. A decayed Poisson process is used to model the number of interactions allocated to each epoch, where the total number of epochs is pre-defined. More work is needed to explore efficient learning methods in TOD systems under the meta-learning paradigm.

9.2 Evaluation

In DPL research, Walker et al.[61] are the first to present a general framework to evaluate the performance of a dialogue agent[38]. They evaluate a dialogue from two aspects. One is the dialogue cost which measures the cost induced by the dialogue (e.g., number of turns), and the other one is task success which evaluates whether the dialogue agent can successfully accomplish the task from the user by comparing it with the user′s goal. In practice, the dialogue policy is often evaluated by having conversations with a simulated user with metrics, such as inform F1, success rate, and Bleu score[117]. The problem is that the simulator does not resemble human conversation behaviour well, as discussed in Section 3. Therefore, there is still a gap between human evaluation and simulated evaluation[117]. Much work is needed to provide a universal evaluation framework that can be used for any general TOD system. Instead of comparing the dialogue act with the simulated goal, a universal evaluation framework should emphasize the overall satisfaction of a human user. Such a framework should include, but not limited to, ways to measure how natural or helpful is the response of the dialogue agent to the user.

10 Conclusions

In this survey, we introduce the recent advances in RL approaches used for DPL in TOD systems. We focus on tackling three main challenges. Given the vast amount of work in such areas in recent years, a method to categorize these approaches is needed to identify the main focal research directions in applying RL in DPL. We propose to categorize recent methods based on the five RL elements and compare the different techniques in each element. As the DPL community is moving to apply TOD systems in real-world scenarios, the scarce data on various dialogue scenarios and the reliability of evaluating dialogue agents will be the most prominent obstacles. Three promising research directions that tackle these obstacles are discussed.

Appendix
Procedure for shortlisting papers

We used a two-step procedure to shortlist relevant papers for review. In the first step, we used two tools to search for relevant papers. The two tools were 1) AMiner[4] which provides literature dated back to 1922 for a given topic keyword, and 2) Connected papers[5], to provide us with a graph of strongly connected papers given a seed paper. We used Aminer with the keyword “dialogue policy” to search for papers within the last ten years. Among the returned list of papers, we used each one as a seed paper as input to Connected Papers and further selected related papers from the provided graph. Then, we went through the papers manually and selected those that applied RL methods in the DPL of TOD systems as the preliminary papers. In the second step, we reviewed the references of the preliminary papers and picked relevant ones.

Acknowledgements

This research was supported by Innovation and Technology Fund (ITF), Government of the Hong Kong Special Administrative Region (HKSAR), China (No. PRP-054-21FX).

[4] https://www.aminer.cn/

[5] https://www.connectedpapers.com/

[5] This paper proposed three models that work on data with belief state and dialogue act annotations, dialogue act annotations only and without any annotations respectively.

References

[1] H. S. Chen, X. R. Liu, D. W. Yin, J. J. Tang. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 2, pp. 25–35, 2017. DOI: 10.1145/3166054.3166058.

[2] M. Lewis, D. Yarats, Y. Dauphin, D. Parikh, D. Batra. Deal or no deal? End-to-end learning of negotiation dialogues. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2443–2453, 2017. DOI: 10.18653/v1/D17-1259.

[3] M. Eric, C. Manning. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, pp. 468–473, 2017.

[4] T. C. Chi, P. C. Chen, S. Y. Su, Y. N. Chen. Speaker role contextual modeling for language understanding and dialogue policy learning. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, Taipei, China, pp. 163–168, 2017.

[5] K. Wang, J. F. Tian, R. Wang, X. J. Quan, J. X. Yu. Multi-domain dialogue acts and response co-generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7125–7134, 2020. DOI: 10.18653/v1/2020.acl-main.638.

[6] Z. Zhang, R. Takanobu, Q. Zhu, M. L. Huang, X. Y. Zhu. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, vol. 63, no. 10, pp. 2011–2027, 2020. DOI: 10.1007/s11431-020-1692-3.

[7] S. Y. Gao, A. Sethi, S. Agarwal, T. Chung, D. Hakkani-Tur. Dialog state tracking: A neural reading comprehension approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Stockholm, Sweden, pp. 264–273, 2019. DOI: 10.18653/v1/W19-5932.

[8] E. Levin, R. Pieraccini, W. Eckert. Learning dialogue strategies within the Markov decision process framework. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, IEEE, Santa Barbara, USA, pp. 72–79, 1997. DOI: 10.1109/ASRU.1997.658989.

[9] S. Singh, M. Kearns, D. Litman, M. Walker. Reinforcement learning for spoken dialogue systems. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, Denver, USA, pp. 956–962, 1999. DOI: 10.5555/3009657.3009792.

[10] S. Gandhe, D. R. Traum. Creating spoken dialogue characters from corpora without annotations. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, pp. 2201–2204, 2007. DOI: 10.21437/Interspeech.2007-599.

[11] L. F. Shang, Z. D. Lu, H. Li. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, pp. 1577–1586, 2015. DOI: 10.3115/v1/P15-1152.

[12] M. A. Walker. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, vol. 12, pp. 387–416, 2000. DOI: 10.1613/jair.713.

[13] S. Singh, D. Litman, M. Kearns, M. Walker. Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 105–133, 2002. DOI: 10.5555/1622407.1622410.

[14] J. Henderson, O. Lemon, K. Georgila. Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Computational Linguistics*, vol. 34, no. 4, pp. 487–511, 2008. DOI: 10.1162/coli.2008.07-028-R2-05-82.

[15] D. DeVault, A. Leuski, K. Sagae. Toward learning and evaluation of dialogue policies with text examples. In *Proceedings of the SIGDIAL Conference*, Portland, USA, pp. 39–48, 2011.

[16] O. Vinyals, Q. Le. A neural conversational model. [Online], Available: https://arxiv.org/abs/1506.05869, 2015

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller. Playing Atari with deep reinforcement learning. [Online], Available: https://arxiv.org/abs/1312.5602, 2013.

[18] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, vol. 529, no. 7587, pp. 484–489, 2016. DOI: 10.1038/nature16961.

[19] A. Y. Ng, H. J. Kim, M. I. Jordan, S. Sastry. Autonomous helicopter flight via reinforcement learning. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 799–806, 2003. DOI: 10.5555/2981345.2981445.

[20] J. Peters, S. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, vol. 21, no. 4, pp. 682–697, 2008. DOI: 10.1016/j.neunet.2008.02.003.

[21] P. H. Su, M. Gasic, N. Mrksic, L. Rojas-Barahona, S. Ultes, D. Vandyke, T. H. Wen, S. Young. Continuously learning neural dialogue management. [Online], Available: https://arxiv.org/abs/1606.02689, 2016.

[22] M. Fatemi, L. El Asri, H. Schulz, J. He, K. Suleman. Policy networks with two-stage training for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Los Angeles, USA, pp. 101–110, 2016. DOI: 10.18653/v1/W16-3613.

[23] P. H. Su, P. Budzianowski, S. Ultes, M. Gašić, S. Young.

Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse And Dialogue*, Saarbrücken, Germany, pp. 147–157, 2017. DOI: 10.18653/v1/W17-5518.

[24] Z. C. Lipton, X. J. Li, J. F. Gao, L. H. Li, F. Ahmed, L. Deng. BBQ-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans, USA, pp. 5237–5244, 2018.

[25] H. Cuayáhuitl, S. Yu, A. Williamson, J. Carse. Deep reinforcement learning for multi-domain dialogue systems. [Online], Available: https://arxiv.org/abs/1611.08675, 2016.

[26] I. Gra\ssl. A survey on reinforcement learning for dialogue systems. [Online], Available: https://arxiv.org/abs/1903.0138, 2019.

[27] Y. P. Dai, H. H. Yu, Y. X. Jiang, C. G. Tang, Y. B. Li, J. Sun. A survey on dialog management: Recent advances and challenges. [Online], Available: https://arxiv.org/abs/2005.02233, 2020.

[28] B. Liu, I. Lane. Iterative policy learning in end-to-end trainable task-oriented neural dialog models. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, Okinawa, Japan, pp. 482–489, 2017. DOI: 10.1109/ASRU.2017.8268975.

[29] B. Liu, I. Lane. Adversarial learning of task-oriented neural dialog models. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse And Dialogue*, Association for Computational Linguistics, Melbourne, Australia, pp. 350–359, 2018. DOI: 10.18653/v1/W18-5041.

[30] B. L. Peng, X. J. Li, J. F. Gao, J. J. Liu, Y. N. Chen, K. F. Wong. Adversarial advantage actor-critic model for task-completion dialogue policy learning. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, pp. 6149–6153, 2018. DOI: 10.1109/ICASSP.2018.8461918.

[31] B. L. Peng, X. J. Li, J. F. Gao, J. J. Liu, K. F. Wong. Deep dyna-Q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, pp. 2182–2192, 2018. DOI: 10.18653/v1/P18-1203.

[32] Y. Cao, K. T. Lu, X. P. Chen, S. Q. Zhang. Adaptive dialog policy learning with hindsight and user modeling. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 329–338, 2020.

[33] G. Gordon-Hall, P. J. Gorinski, G. Lampouras, I. Iacobacci. Show us the way: Learning to manage dialog from demonstrations. [Online], Available: https://arxiv.org/abs/2004.08114, 2020.

[34] R. S. Sutton, A. G. Barto. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, vol. 9, no. 5, Article number 1054, 1998. DOI: 10.1109/TNN.1998.712192..

[35] W. H. Chen, J. S. Chen, P. D. Qin, X. F. Yan, W. Y. Wang. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 3696–3709, 2019. DOI: 10.18653/v1/P19-1360.

[36] P. H. Su, M. Gašić, N. Mrkšić, L. M. Rojas-Barahona, S. Ultes, D. Vandyke, T. H. Wen, S. Young. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, pp. 2431–2441, 2016. DOI: 10.18653/v1/P16-1230.

[37] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, S. Young. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, Rochester, USA, pp. 149–152, 2007.

[38] M. A. Walker, D. J. Litman, C. A. Kamm, A. Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, pp. 271–280, 1997. DOI: 10.3115/976909.979652.

[39] L. Chen, R. Z. Yang, C. Chang, Z. H. Ye, X. Zhou, K. Yu. On-line dialogue policy learning with companion teaching. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, pp. 198–204, 2017. DOI: 10.18653/v1/E17-2032.

[40] K. T. Lu, S. Q. Zhang, X. P. Chen. Goal-oriented dialogue policy learning from failures. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence and the 31st Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, Honolulu, USA, pp. 2596–2603, 2019. DOI: 10.1609/aaai.v33i01.33012596.

[41] K. T. Lu, S. Q. Zhang, X. P. Chen. AutoEG: Automated experience grafting for off-policy deep reinforcement learning. [Online], Available: https://arxiv.org/abs/2004.10698, 2020.

[42] G. Gordon-Hall, P. J. Gorinski, S. B. Cohen. Learning dialog policies from weak demonstrations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1394–1405, 2020. DOI: 10.18653/v1/2020.acl-main.129.

[43] L. H. Li, H. He, J. D. Williams. Temporal supervised learning for inferring a dialog policy from example conversations. In *Proceedings of IEEE Spoken Language Technology Workshop*, South Lake Tahoe, USA, pp. 312–317, 2014. DOI: 10.1109/SLT.2014.7078593.

[44] P. H. Su, D. Vandyke, M. Gašić, N. Mrkšić, T. H. Wen, S. Young. Reward shaping with recurrent neural networks for speeding up on-line policy learning in spoken dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Prague, Czech Republic, pp. 417–421, 2015. DOI: 10.18653/v1/W15-4655.

[45] T. C. Zhao, M. Eskenazi. Towards end-to-end learning for dialog state tracking and management using deep re-

inforcement learning. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Los Angeles, USA, pp. 1–10, 2016. DOI: 10.18653/v1/W16-3601.

[46] P. Budzianowski, S. Ultes, P. H. Su, N. Mrkšić, T. H. Wen, I. Casanueva, L. M. Rojas-Barahona, M. Gašić. Sub-domain modelling for dialogue management with hierarchical reinforcement learning. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, Saarbrücken, Germany, pp. 86–92, 2017. DOI: 10.18653/v1/W17-5512.

[47] B. L. Peng, X. J. Li, L. H. Li, J. F. Gao, A. Celikyilmaz, S. Lee, K. F. Wong. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2231–2240, 2017. DOI: 10.18653/v1/D17-1237.

[48] G. Weisz, P. Budzianowski, P. H. Su, M. Gašić. Sample efficient deep reinforcement learning for dialogue systems with large action spaces. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2083–2097, 2018. DOI: 10.1109/TASLP.2018.2851664.

[49] I. Casanueva, P. Budzianowski, P. H. Su, S. Ultes, L. M. Rojas-Barahona, B. H. Tseng, M. Gašić. Feudal reinforcement learning for dialogue management in large domains. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, USA, pp. 714–719, 2018. DOI: 10.18653/v1/N18-2112.

[50] G. Y. Kristianto, H. W. Zhang, B. Tong, M. Iwayama, Y. Kobayashi. Autonomous sub-domain modeling for dialogue policy with hierarchical deep reinforcement learning. In *Proceedings of the EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, Brussels, Belgium, pp. 9–16, 2018. DOI: 10.18653/v1/W18-5702.

[51] S. Y. Su, X. J. Li, J. F. Gao, J. J. Liu, Y. N. Chen. Discriminative deep dyna-Q: Robust planning for dialogue policy learning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 3813–3823, 2018. DOI: 10.18653/v1/D18-1416.

[52] D. Tang, X. J. Li, J. F. Gao, C. Wang, L. H. Li, T. Jebara. Subgoal discovery for hierarchical dialogue policy learning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 2298–2309, 2018. DOI: 10.18653/v1/D18-1253.

[53] Y. X. Wu, X. J. Li, J. J. Liu, J. F. Gao, Y. M. Yang. Switch-based active deep dyna-Q: Efficient adaptive planning for task-completion dialogue policy learning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Honolulu, USA, pp. 7289–7296, 2019. DOI: 10.1609/aaai.v33i01.33017289.

[54] T. C. Zhao, K. G. Xie, M. Eskenazi. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis USA, pp. 1208–1218, 2019. DOI: 10.18653/v1/N19-1123.

[55] Y. M. Xu, C. G. Zhu, B. L. Peng, M. Zeng. Meta dialogue policy learning. [Online], Available: https://arxiv.org/abs/2006.02588, 2020.

[56] A. Papangelis, Y. C. Wang, P. Molino, G. Tur. Collaborative multi-agent dialogue model training via reinforcement learning. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Association for Computational Linguistics, Stockholm, Sweden, pp. 92–102, 2019. DOI: 10.18653/v1/W19-5912.

[57] Z. R. Zhang, X. J. Li, J. F. Gao, E. H. Chen. Budgeted policy learning for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 3742–3751, 2019. DOI: 10.18653/v1/P19-1364.

[58] R. Takanobu, H. L. Zhu, M. L. Huang. Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 100–110, 2019. DOI: 10.18653/v1/D19-1010.

[59] X. T. Huang, J. Z. Qi, Y. Sun, R. Zhang. Semi-supervised dialogue policy learning via stochastic reward estimation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 660–670, 2020. DOI: 10.18653/v1/2020.acl-main.62.

[60] Z. Zhang, L. Z. Liao, X. Y. Zhu, T. S. Chua, Z. T. Liu, Y. Huang, M. L. Huang. Learning goal-oriented dialogue policy with opposite agent awareness. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China, pp. 122–132, 2020.

[61] Z. M. Li, S. Lee, B. L. Peng, J. C. Li, J. Kiseleva, M. de Rijke, S. Shayandeh, J. F. Gao. Guided dialogue policy learning without adversarial learning in the loop. In *Proceedings of Findings of the Association for Computational Linguistics*, pp. 2308–2317, 2020. DOI: 10.18653/v1/2020.findings-emnlp.209.

[62] R. Takanobu, R. Z. Liang, M. L. Huang. Multi-agent task-oriented dialog policy learning with role-aware reward decomposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 625–638, 2020. DOI: 10.18653/v1/2020.acl-main.59.

[63] P. H. Su, D. Vandyke, Gašić, D. Kim, N. Mrkšić, T. H. Wen, S. Young. Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, pp. 2007–2011, 2015. DOI: 10.21437/Interspeech.2015-456.

[64] J. Schatzmann, S. Young. The hidden agenda user simulation model. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 733–747, 2009. DOI: 10.1109/TASL.2008.2012071.

[65] X. J. Li, Z. C. Lipton, B. Dhingra, L. H. Li, J. F. Gao, Y. N. Chen. A user simulator for task-completion dialogues. [Online], Available: https://arxiv.org/abs/1612.05688, 2016.

[66] S. Ultes, L. M. Rojas-Barahona, P. H. Su, D. Vandyke, D. Kim, I. Casanueva, P. Budzianowski, N. Mrkšić, T. H. Wen, M. Gašić, S. Young. Pydial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL System Demonstrations*, Vancouver, Canada, pp. 73–78, 2017.

[67] J. F. Gao, M. Galley, L. H. Li. Neural approaches to conversational AI. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, Ann Arbor, USA, pp. 1371–1374, 2018. DOI: 10.1145/3209978.3210183.

[68] I. Sutskever, O. Vinyals, Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 3104–3112, 2014. DOI: 10.5555/2969033.2969173.

[69] W. Eckert, E. Levin, R. Pieraccini. User modeling for spoken dialogue system evaluation. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, Santa Barbara, USA, pp. 80–87, 1997. DOI: 10.1109/ASRU.1997.658991.

[70] E. Levin, R. Pieraccini, W. Eckert. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 11–23, 2000. DOI: 10.1109/89.817450.

[71] S. Chandramohan, M. Geist, F. Lefévre, O. Pietquin. User simulation in dialogue systems using inverse reinforcement learning. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, pp. 1025–1028, 2011.

[72] L. El Asri, J. He, K. Suleman. A sequence-to-sequence model for user simulation in spoken dialogue systems. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, San Francisco, USA, pp. 1151–1155, 2016. DOI: 10.21437/Interspeech.2016-1175.

[73] J. D. Williams. Evaluating user simulations with the cramér-von mises divergence. *Speech Communication*, vol. 50, no. 10, pp. 829–846, 2008. DOI: 10.1016/j.specom.2008.05.007.

[74] H. Ai, D. J. Litman. Assessing dialog system user simulation evaluation measures using human judges. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, Columbus, USA, pp. 622–629, 2008.

[75] O. Pietquin, H. Hastie. A survey on metrics for the evaluation of user simulations. *The Knowledge Engineering Review*, vol. 28, no. 1, pp. 59–73, 2013. DOI: 10.1017/S0269888912000343.

[76] K. Georgila, C. Nelson, D. Traum. Single-agent vs. multi-agent techniques for concurrent reinforcement learning of negotiation dialogue policies. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, USA, pp. 500–510, 2014. DOI: 10.3115/v1/P14-1047.

[77] H. M. Wang, K. F. Wong. A collaborative multi-agent reinforcement learning framework for dialog action decomposition. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, pp. 7882–7889, 2021. DOI: 10.18653/v1/2021.emnlp-main.621.

[78] M. Gašić, N. Mrkšić, L. Rojas-Barahona, P. H. Su, D. Vandyke, T. H. Wen. Multi-agent learning in multi-domain spoken dialogue systems. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems*, Montreal, Canada, 2015.

[79] R. Parr, S. Russell. Reinforcement learning with hierarchies of machines. In *Proceedings of Conference on Advances in Neural Information Processing Systems*, MIT Press, Denver, USA, pp. 1043–1049, 1998. DOI: 10.5555/302528.302894.

[80] T. G. Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, vol. 13, pp. 227–303, 2000. DOI: 10.1613/jair.639.

[81] S. Young, M. Gašić, B. Thomson, J. D. Williams. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013. DOI: 10.1109/JPROC.2012.2225812.

[82] R. S. Sutton, D. Precup, S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, vol. 112, no. 1–2, pp. 181–211, 1999. DOI: 10.1016/S0004-3702(99)00052-1.

[83] P. L. Bacon, J. Harb, D. Precup. The option-critic architecture. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco, USA, pp. 1726–1734, 2017.

[84] M. C. Machado, M. G. Bellemare, M. Bowling. A Laplacian framework for option discovery in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, pp. 2295–2304, 2017.

[85] C. Wang, Y. N. Wang, P. S. Huang, A. Mohamed, D. Y. Zhou, L. Deng. Sequence modeling via segmentations. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, pp. 3674–3683, 2017.

[86] P. Dayan, G. E. Hinton. Feudal reinforcement learning. In *Proceedings of the 5th International Conference on Neural Information Processing Systems*, Denver, USA, pp. 271–278, 1992. DOI: 10.5555/2987061.2987095.

[87] I. Casanueva, P. Budzianowski, S. Ultes, F. Kreyssig, B. H. Tseng, Y. C. Wu, M. Gašić. Feudal dialogue management with jointly learned feature extractors. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, Melbourne, Australia, pp. 332–337, 2018. DOI: 10.18653/v1/W18-5038.

[88] P. Abbeel, A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning*, ACM, Banff, Canada, 2004. DOI: 10.1145/1015330.1015430.

[89] M. Jhunjhunwala, C. Bryant, P. Shah. 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*, pp. 290–296, 2020.

[90] N. Mrkšić, D. Ó. Séaghdha, T. H. Wen, B. Thomson, S. Young. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of*

the Association for Computational Linguistics, Vancouver, Canada, pp. 1777–1788, 2017. DOI: 10.18653/v1/P17-1163.

[91] S. Hochreiter, J. Schmidhuber. Long short-term memory. *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[92] T. Winograd. Understanding natural language. *Cognitive Psychology*, vol. 3, no. 1, pp. 1–191, 1972. DOI: 10.1016/0010-0285(72)90002-3.

[93] J. P. Zhang, T. C. Zhao, Z. Yu. Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, Association for Computational Linguistics, Melbourne, Australia, pp. 140–150, 2018. DOI: 10.18653/v1/W18-5015.

[94] T. Saha, S. Saha, P. Bhattacharyya. Towards sentiment-aware multi-modal dialogue policy learning. *Cognitive Computation*, vol. 14, no. 1, pp. 246–260, 2022. DOI: 10.1007/s12559-020-09769-7.

[95] R. De Mori. Spoken language understanding: A survey. In *Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding*, Kyoto, Japan, pp. 365–376, 2007. DOI: 10.1109/ASRU.2007.4430139.

[96] Y. C. Zhang, Z. J. Ou, Z. Yu. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 9604–9611, 2020. DOI: 10.1609/aaai.v34i05.6507.

[97] Y. H. Li, Y. Y. Yang, X. J. Quan, J. X. Yu. Retrieve & memorize: Dialog policy learning with multi-action memory. In *Proceedings of Findings of the Association for Computational Linguistics*, pp. 447–459, 2021. DOI: 10.18653/v1/2021.findings-acl.39.

[98] L. Shu, H. Xu, B. Liu, P. Molino. Modeling multi-action policy for task-oriented dialogues. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 1304–1310, 2019. DOI: 10.18653/v1/D19-1130.

[99] J. H. Wang, Y. Zhang, T. K. Kim, Y. J. Gu. Modelling hierarchical structure between dialogue policy and natural language generator with option framework for task-oriented dialogue system. In *Proceedings of the 9th International Conference on Learning Representations*, 2020.

[100] L. El Asri, R. Laroche, O. Pietquin. Task completion transfer learning for reward inference. In *Proceedings of International Workshop on Machine Learning for Interactive Systems*, Québec, Canada, 2014.

[101] H. M. Wang, B. L. Peng, K. F. Wong. Learning efficient dialogue policy from demonstrations through shaping. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6355–6365, 2020. DOI: 10.18653/v1/2020.acl-main.566.

[102] S. Russell. Learning agents for uncertain environments (extended abstract). In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, Madison, USA, pp. 101–103, 1998. DOI: 10.1145/279943.279964.

[103] A. Y. Ng, S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, Stanford, USA, pp. 663–670, 2000. DOI: 10.5555/645529.657801.

[104] A. Boularias, J. Kober, J. Peters. Relative entropy inverse reinforcement learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, USA, pp. 182–189, 2011.

[105] A. Boularias, H. R. Chinaei, B. Chaib-Draa. Learning the reward model of dialogue POMDPs from data. In *Proceedings of NIPS Workshop on Machine Learning for Assistive Techniques*, 2010.

[106] J. Ho, S. Ermon. Generative adversarial imitation learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Barcelona, Spain, pp. 4572–4580, 2016. DOI: 10.5555/3157382.3157608.

[107] A. Y. Ng, D. Harada, S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning*, Bled, Slovenia, pp. 278–287, 1999.

[108] L. El Asri, R. Laroche, O. Pietquin. Reward shaping for statistical optimisation of dialogue management. *Statistical Language and Speech Processing*, A. H. Dediu, C. Martín-Vide, R. Mitkov, B. Truthe, Eds., Tarragona, Spain: Springer, pp. 93–101, 2013. DOI: 10.1007/978-3-642-39593-2_8.

[109] E. Ferreira, F. Lefévre. Social signal and user adaptation in reinforcement learning-based dialogue management. In *Proceedings of the 2nd Workshop on Machine Learning for Interactive Systems: Bridging the Gap Between Perception, Action and Communication*, Beijing, China, pp. 61–69. 2013, DOI: 10.1145/2493525.2493535.

[110] H. R. Wang, H. M. Wang, Z. H. Wang, K. F. Wong. Integrating pretrained language model for dialogue policy learning. [Online], Available: https://arxiv.org/abs/2111.01398, 2021.

[111] V. Ilievski, C. Musat, A. Hossman, M. Baeriswyl. Goal-oriented chatbot dialog management bootstrapping with transfer learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, pp. 4115–4121, 2018. DOI: 10.24963/ijcai.2018/572.

[112] S. J. Pan, Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010. DOI: 10.1109/TKDE.2009.191.

[113] L. Chen, C. Chang, Z. Chen, B. W. Tan, M. Gašić, K. Yu. Policy adaptation for deep reinforcement learning-based dialogue management. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, pp. 6074–6078, 2018. DOI: 10.1109/ICASSP.2018.8462272.

[114] K. X. Mo, Y. Zhang, Q. Yang, P. Fung. Cross-domain dialogue policy transfer via simultaneous speech-act and slot alignment. [Online], Available: https://arxiv.org/abs/1804.07691, 2018.

[115] F. Mi, M. L. Huang, J. Y. Zhang, B. Faltings. Meta-learn-

ing for low-resource natural language generation in task-oriented dialogue systems. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, Macao, China, pp. 3151–3157, 2019. DOI: 10.5555/3367471.3367479]

[116] C. Finn, P. Abbeel, S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, pp. 1126–1135, 2017. DOI: 10.5555/3305381.3305498.

[117] R. Takanobu, Q. Zhu, J. C. Li, B. L. Peng, J. F. Gao, M. L. Huang. Is your goal-oriented dialog model performing really well? Empirical analysis of system-wise evaluation. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 297–310, 2020.

Wai-Chung Kwan received the B. Sc. degree in computer science from Hong Kong Baptist University, China in 2019. He is currently a Ph. D. degree candidate in systems engineering and engineering management at Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong, China.

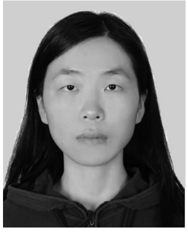
His research interests include natural language processing, reinforcement learning and dialogue systems.

E-mail: wckwan@se.cuhk.edu.hk (Corresponding author)
ORCID iD: 0000-0002-2942-4208

Hong-Ru Wang received the B. Sc. degree in computer science and technology from Communication University of China, China in 2019, received the M. Sc. degree in computer science from the Chinese University of Hong Kong, China 2020, respectively. He is a currently a Ph. D. degree candidate in systems engineering and engineering management at Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong, China.

His research interests include task-oriented dialogue system, controllable natural language generation, persona-knowledge enhanced dialogue system.

E-mail: hrwang@se.cuhk.edu.hk
ORCID iD: 0000-0001-5027-0138



Hui-Min Wang received B. Eng. and M. Eng. degrees in automation from Tsinghua University, China in 2014 and 2017, respectively, received the Ph. D. degree in systems engineering and engineering management from Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, China in 2021.

Her research interests include reinforcement learning, natural language processing, especially on dialogue system.

E-mail: hmwang@se.cuhk.edu.hk
ORCID iD: 0000-0002-6147-8310

Kam-Fai Wong received the Ph.D. degree in electrical engineering from Edinburgh University, UK in 1987. He was a post doctoral researcher in Heriot-Watt University, UK, UniSys, UK and ECRC, Germany. At present, he is professor in Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong (CUHK), China. He serves as the Associate Dean (External Affairs) of Engineering, the Director of the Centre for Innovation and Technology (CINTEC), and Associate Director of the Centre for Entrepreneurship (CfE), CUHK. He serves as the President of Asian Federation of Natural Language Processing (AFNLP, 2015–2016), President of the Governing Board of Chinese Language Computer Society CLCS (2015–2017). He has published over 250 technical papers in these areas in different international journals and conferences and books. He is Fellow of ACL (2020), Member of ACM, Senior Member of IEEE as well as fellow of the following professional bodies BCS (UK), IET (UK) and HKIE. He is the founding Editor-In-Chief of *ACM Transactions on Asian Language Processing* (TALIP), and serves as Associate Editor of *International Journal on Computational Linguistics and Chinese Language Processing*. He is the Publication Chair of ACL 2021, General Chair of AACL-IJCNLP 2020, Organization Chair of EMNLP 2019, Conference Co-Chair of NDBC 2016, BigComp 2016, NLPCC 2015 and IJCNLP 2011; the Finance Chair SIGMOD 2007; and the PC Co-chair of lJCNLP 2006. Also he is a Programme Committee Member of many international conferences.

His research interest focuses on Chinese computing, social media processing and information retrieval.

E-mail: kfwong@se.cuhk.edu.hk
ORCID iD: 0000-0002-9427-5659