A Review of the Research on Dialogue Management of Task-Oriented Systems

Yin Jiang Zhao, Yan Ling Li and Min Lin
College of computer science and technology, Inner Mongolia Normal University
liyanling7871397@163.com

Abstract. Scholars pay more and more attention to the spoken dialogue system after the emergence of deep learning technology. The task-based dialogue system has become one of the most important branches in the field of spoken dialogue systems. Dialogue management is the core of the task-based dialogue system, and its research theory and technology have been gained extensive attention. This paper summarizes the research progress and current situation of task-based dialogue system and dialogue management strategy. Firstly, it summarizes the status of the task-based dialogue system models and compares their advantages and disadvantages. Then it focuses on the analysis of various dialogue management research strategies from the perspective of model theory and research methods. Finally, it looks forward to the future research direction of task-oriented dialogue system combined with dialogue management.

1. Introduction

The spoken dialogue system mainly provides some kind of service for people, it is the bridge between human and intelligent machine. In recent years, the rapid development of big data and deep learning has provided great impetus for the scientific research fields such as dialogue system and computer vision. Moreover, it greatly encourages researchers to use deep learning and big data to study the spoken dialogue system from multiple perspectives. With the rise of the mobile Internet, spoken dialogue systems are tightly integrated with mobile products like Siri, Cortana, and Echo. Users can access information and services through the dialog system in a conversational manner.

Dialogue systems are generally divided into two types according to different application purposes: task-based spoken dialogue systems and non-task-based spoken dialogue systems (chattering systems). Non-task-based spoken dialogue systems are considered to be open domain chat systems, which provides gossip or entertainment to the user and does not accomplish a specific task. At present, researches on chattering systems usually include generation methods and retrieve-based methods. The purpose of a task-based dialog system is to help users complete certain tasks, such as booking train tickets, checking movie ticket, etc. It needs to collect data set for a specific domain. At present, the mainstream research methods of Task-based dialogue systems usually include pipeline method and end-to-end method. In addition, Task-oriented dialogue system has become the most prominent component in today's virtual personal assistant, which attracts more and more attention.

This paper has two objectives: (1) It summarizes the model of task-based dialog system and focuses on the latest development of dialog management; (2) It discusses possible research directions. The rest of this paper is arranged as follows: Section 2 introduces the research status of the task-based dialog system model, including pipeline method and end-to-end method; Section 3...
introduces several mainstream methods of dialogue management research technology, focusing on reinforcement learning and deep reinforcement learning. It elaborates and analyzes the model theory and research progress, and points out the advantages and disadvantages of each method. In the end, it summarizes and discusses several problems to be further studied and possible future research directions.

2. The research status of a task-based dialogue system model
The task-based spoken dialogue system is an important branch of the spoken dialogue system. This section describes the pipeline method and the end-to-end method.

2.1. Pipeline-based method
The pipeline model contains the following functional modules, as shown in figure 1: Natural Language Understanding (NLU), Dialogue Management (DM) and Natural Language Generation (NLG). First, it needs to understand the semantic information expressed by humans and translate it into an internal representation. Second, it takes a series of actions based on the strategy through internal representations. Third, it translates actions into natural language and feeds them back to the user.

![Figure 1. Traditional Pipeline for Task-oriented Systems.](image)

The pipeline model requires that the three modules be executed sequentially and independently of each other. In this way, the pipeline model has two advantages [1]: In each module, the problem is clear and the problems to be solved in each module are detailed. Although the pipeline model has obvious advantages, it also has some problems: (1) The three modules are strongly dependent, and errors in the previous modules may accumulate to the subsequent modules; (2) The change of the output structure of the previous module will affect the design and implementation of the subsequent modules; (3) The design of models is often determined according to specific tasks, so the portability of the domain is poor.

2.2. End-to-end method
The powerful distributed representation ability of neural networks in deep learning makes many natural language processing tasks successful [2, 3]. This makes it possible to build task-oriented end-to-end dialogue systems. The idea of the end-to-end model is to design a whole derivable model, and then transfer the gradient of the output end back to the whole neural network by back propagation. Finally, model parameters are jointly optimized. The overall framework adopts the encoder-decoder structure, as shown in figure 2.

![Figure 2. An Illustration of the Encoder-Decoder Model.](image)
Task-based dialogue systems often need to query the external knowledge base. Early models retrieved entries from the knowledge base by issuing a query symbol to the knowledge base retrieve entries from the knowledge base based on attributes [3, 4, 5]. But there were two problems in the end-to-end approach: (1) It does not model uncertainty in user input; (2) The end-to-end model cannot retrieve the knowledge base, because the end-to-end model neural network does not model the query symbol, so the parser and dialogue strategy are trained separately and it is difficult to End-to-end online learning from user feedback. Scholars have done a lot of research to solve the problem of interaction with the external knowledge base. Eric et al. [6] use the replication mechanism to retrieve the knowledge base and simultaneously record entities appearing in the dialogue history. It can accurately retrieve the items in the knowledge base and achieve good results. Later Eric et al. [7] expands the existing recursive network and establishes the key-value retrieval mechanism based on attention in the knowledge base items. The knowledge base is model in the form of triples (object, attribute name, attribute value). However, not all knowledge bases are presented in the form of key-value, so the applicability is not particularly wide. Wen et al. [8] uses internal state representation based on the attention mechanism to complete the retrieval of the knowledge base by combining the advantages of the pipeline model and the end-to-end model. Finally, the replication mechanism is used to extract entities from the knowledge base to generate responses, which significantly improves the retrieval efficiency.

In addition to the above research work, scholars have also studied a task-oriented dialogue system from another perspective. They see dialogue strategies as reinforcement learning strategies. This method can significantly improve the performance of the dialogue system. Li et al. [9] firstly provides a publicly available user simulator framework. Scholars can build different user simulators according to the characteristics of the dialogue task. This greatly promotes the application of reinforcement learning in a task-based dialogue system. Dhingra et al. [10] combines the soft attention retrieval process with reinforcement learning, which can achieve a higher task success rate and higher rewards. Although user simulator can reduce learning costs, there is no uniform standard to evaluate the advantages and disadvantages of user simulator. The effectiveness of training task-based dialog agents by simulating users has always been controversial.

3. Dialogue Management
Dialogue management controls the whole dialogue process, and its design directly affects the performance of the dialogue system. Dialogue management mainly includes two tasks: (1) Dialogue state tracking (DST); (2) Dialogue policy learning (DPL).

• Dialogue state tracking: It determines the current user target based on multiple rounds of conversations between the system and the user. Dialogue state tracking provides the basis for later decisions.

• Dialogue policy selection: It selects an executable action based on the results of the dialogue state tracing. The goal is to complete the task in a short conversation as possible.

3.1. Traditional research methods of dialogue management
Scholars have proposed many methods of dialogue management in early dialogue management study[11,12,13,14,15,16,17,18,19,20,21,22]. Among them, the main method includes Finite State Machine (FSM)[11,12], slot-fill Model (SFM)[13] and task-based structure method, etc. The FSM regards a dialogue process as a state transition path from the initial state to the termination state of the state machine. It can effectively control the dialogue flow and simple structure. But it has some problems such as incomplete state coverage, serious sequence dependence and poor portability[14]. The slot-filling method treats the dialog process as a process of filling a table with placeholders. This method extends the method based on finite automatic state machine, which has no strict order and the answer form is very flexible. However, SMF also has the problem of state explosion and cannot satisfy the task of multi-topic dialogue. Task-based structure method uses tree structure to express the element relations in the field, which can meet the multi-topic task [15]. But it is limited by the tree
structure and difficult to extend. In addition, there are other dialogue management research methods, such as the dialogue management method based on planning [16], the Information State Update method [17,18], the Agent method [19,20] and the self-organizing dialogue management planning model based on the theme forest structure [21,22]. These methods are widely used because they have the advantages of simple structure and easy implementation. However, these methods have common shortcomings and hinder the development of dialogue system: They cannot guarantee that all cases and dialogue rules will be covered and manual rules will need to be defined. Scholars propose a method based on reinforcement learning to solve the problems of traditional dialogue management.

3.2. Reinforcement learning

The method based on reinforcement learning has always been a hotspot in research of dialogue management. More classic models including Markov Decision Process (MDP) [23, 24] and Partially observables Markov Decision Process(POMDP)[25]. The idea of reinforcement learning is that agents interact with the environment by using their own experience and environmental feedback. It is the process of learning codes of conduct through trial and error. The goal of reinforcement learning is to select the “best” behavior by maximizing the discount reward of the expected value:

$$\text{Reward} = E\left[\sum_{t=0}^{\infty} \gamma^t (s_t, a_t)\right]$$

(1)

Where, $s_t$ is the state at time $t$, $a_t$ is the action at time $t$, and $\gamma$ is the discount factor. The optimal strategy can be obtained according to the maximum expected reward value:

$$\pi^* = \arg\max_{\pi} E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s\right]$$

(2)

Reinforcement learning method can effectively solve sequence decision-making problems so that it has the following advantages: (1) It gets rid of the dependence of manually defining rules in the decision-making process and improves the generalization ability; (2) It can fully explore the state space and solve the problem of insufficient coverage. In view of these advantages, reinforcement learning has been widely used in dialogue management, and it has been proved in experiments that POMDP model is superior to MDP model [26, 27, 28, 29, 30, 31, 32, 33]. Thomson et al. [32] used a POMDP model to construct a spoken dialogue system of home care robot, and it was proved that the POMDP model performed better than the MDP model in a certain noise environment. Young and Williams et al. also did a lot of research on dialogue management based on POMDP model [34, 35, 36]. Williams et al.[34] proposed a decomposed POMDP model based on Roy's model. The auto-generated policy of this model is directly compared with the MDP model and manually formulated state machine policy. The experimental results show that the POMDP model performs better than the MDP model in the case of high speech recognition error rate. Later, Williams et al. [35] constructs a tourism navigation system based on the mixed model, which combines the decomposition model in literature [34] with slot filling method. After that, Williams et al. [36] constructed an emotional dialogue system model based on POMDP. It added user emotion items into the decomposition model in literature [34], where user emotion is divided into stressed and unstressed.

Although the reinforcement learning is superior to the traditional method, it also has the following problems: (1) It has the problem of sparse data. The MDP model and POMDP model need a lot of dialogue data to train model parameters. Gasic et al. [27] provides samples to solve this problem by using the dialogue data of user simulator and systems; (2) How to design a reasonable reward function to guide the action selection. Young et al. [31] has adopted the Inverse Reinforcement Learning (IRL) method to learn Reinforcement function from dialogue corpus, but it also needs massive dialogue data. (3) As the state space increases, the computational complexity of the model solution will increase sharply[36]. Choi et al. [33] proposed Hidden information State Model to solve this problem; (4) How to automatically improve the dialogue policy in the interaction process based on the existing dialogue
policy. (5) The application of the POMDP dialog management model has limitations, and it cannot handle situations where slots or slot values are infinite.

3.3. Deep reinforcement learning techniques

3.3.1 DQN model. The traditional reinforcement learning algorithm is limited by the action space and sample space. When the state space is large or even infinite, the efficiency of learning will be greatly reduced. Later, with the development of deep learning, Mnih et al. [37] solves this problem by using deep reinforcement learning techniques. In particular, the proposal of DQN model is of ground breaking significance for the development of deep reinforcement learning techniques [38]. Before the appearance of DQN, when the neural network is used to approximate the action value function in RL, it will have the problem of unstable or non-convergence of Q value. DQN uses two techniques to solve this problem: the experience replay mechanism and the target network. The idea of the model is to approximate the value function of action with the function of a deep neural network:

$$Q(s,a) \approx f(s,a,\theta)$$  \hspace{1cm} (3)

Where, $\theta$ is the parameter of the neural network. State $s$ is used as input to output the Q value of each action.

Experience replay mechanism: It puts the collected samples $e = (s_i, a_i, r, s_{i+1})$ into the sample pool $D = \{e_1, e_2, ..., e_n\}$. It randomly selects small batches of samples from $D$ each time during training. This method can reduce the correlation between samples.

Target network: It establishes two neural networks with the same structure, denoted as Q network and target value network. The Q network is used to generate the current value, denoted as $Q(s,a,\theta)$. The target value network is used to generate the target Q value, denoted as $Q(s,a,\theta')$. The optimization objective of the target value network is:

$$y = r(s,a,s') + \gamma \max_a Q(s',a',\theta')$$  \hspace{1cm} (4)

Where $a'$ is the possible action, $s'$ is the state at the next moment, and $\theta'$ is the parameter of the target network. The loss function is expressed as follows:

$$loss(\theta) = [y - Q(s,a,\theta)]^2$$  \hspace{1cm} (5)

In this process, the parameters $\theta'$ in the target network are kept fixed, and then the parameters $\theta$ of Q network are updated with the stochastic gradient descent algorithm. The parameter of Q network is copied to the target value network after N rounds of iteration. The training process is shown in figure 3.

![Figure 3. DQN Training Process.](image)
The DQN model provides a new direction for the research techniques of dialogue management: (1) It provides two methods of experience replay mechanism and target network to stabilize the training process.; (2) It designed an end-to-end reinforcement learning method, which can input the state value and output the Q value of each action through the Q value network; (3) It uses the same model parameters, training algorithm and network structure in different tasks, which greatly improves the performance of the model.

Although DQN model improves the reinforcement learning algorithm and the performance of the dialogue system, it has the following disadvantages: (1) It has the problems of too long training time, low sample utilization and its reward function is hard to design; (2) It has the problem of Q network instability because Q network is sensitive to parameters; (3) It has the problem of over fitting the environment and it is difficult to adapt to the new environment.

Deep reinforcement learning technology greatly promotes the development of dialogue management research technology. For the first time, Li et al.[9] provides a publicly available framework of user simulator, which can interact with agents to learn dialogue strategies from samples. Among them, the success rate of DQN-based agents in the task of booking movie tickets is much higher than that of rule-based agents, which further illustrates the development potential of the deep reinforcement learning model in the task-based dialogue system. Volodymyr et al.[39] proposes a BBQN (Bayes-by-backdrop Q-network), which improves the DQN algorithm. It explores state space by using Thompson sampling, which greatly improves exploration efficiency.

3.3.2 Adversarial advantage actor-critic. Scholars are increasingly concerned about the use of deep reinforcement learning in task-based conversational systems [10, 27, 40, 41, 42, 43, 44, 45, 46] when the coming of deep reinforcement learning techniques. Zhao et al. [41] regards the retrieval operation of a knowledge base as a part of dialogue strategy learning through deep reinforcement learning. This method can be applied to any knowledge base and can be further optimized by supervised learning or reinforcement learning. Williams et al. [43] proposes an end-to-end task-oriented neural network model, which can improve training efficiency with only a small amount of effective data and has good mobility. However, this model requires a high-quality data set to achieve good results. Peng et al. [44] propose a hierarchical deep reinforcement learning model, which can learn dialogue management strategies in different time periods and complete multiple subtasks at the same time. Liu et al. [45] simulates the dialogue between two agents, and jointly optimizes the dialogue agent and the user simulator by using deep reinforcement learning. This approach achieves higher task success and rewards than single supervised learning and reinforcement learning.

Although these methods have achieved good results, they all have the problem of sparse reward. Scholars have come up with many different ways to solve this problem. These methods fall into two categories:(1) It learns transcendental knowledge from expert-generated dialogues and preliminary strategies from person-to-person dialogues. Peng et al. [46] and Lipton et al. [47] has proved that the pre-trained supervision strategy or the strategy based on weak rules can significantly improve the efficiency of exploration; (2) It introduces heuristic rules to guide the exploration through internal reward[48,49,50]. Jaderberg et al. [49] converts the training signals of three auxiliary tasks into intrinsic rewards, which can significantly improve the learning speed and robustness of agents. Houthoof et al. [50] added the intrinsic reward value into the reward function, which encouraged agents to explore those relatively unexplored regions. At present, the latest research method in the research field of dialogue management is the A2C model of confrontation learning proposed by Peng et al. [51]. The model combines the above two methods. The idea of the model is to use the dialogue generated by experts as the prior knowledge, but not to use the prior knowledge to construct the dialogue strategy. It trained a discriminator inspired by the antagonistic network (GAN) [52]. This discriminator can distinguish between action generated by agents and those generated by human experts. It uses the output of the discriminator as an intrinsic reward, encouraging agents to explore the state-action space like experts. This model easily leads to successful conversations and brings positive returns compared to other models. The adversarial A2C model consists of two parts, as shown in figure 4: (1) actor-critic
framework; (2) adversarial learning discriminator D. Please refer to the reference [51] for the specific implementation details of the model.

**Figure 4.** The adversarial advantage Actor-Critic.

### 4. Summary and Discussion
This paper mainly discusses the research progress of dialogue management of task-based dialogue system, and mainly introduces three kinds of dialogue management methods. At present, the end-to-end model has become a mainstream model of a task-oriented dialogue system, and the deep learning technology has become a basic technology to solve the problem of the dialogue system. In recent years, deep reinforcement learning technology has become one of the key technologies of dialogue management. Deep reinforcement learning technology can greatly improve the task success rate of end-to-end task-oriented dialogue system. At present, there are still the following problems in the task-based dialogue system:

- The end-to-end dialogue system model does not model the external knowledge base, which affects the performance of dialogue management using the external knowledge base. How to build a knowledge base suitable for a neural network to improve the performance of dialog management is a question worth exploring.

- The dialogue management based on deep reinforcement learning requires user simulators. But there is a big gap between user simulators and real users. How to improve the user simulator based on the neural network to have more human characteristics and how to find a unified standard to evaluate the user simulator needs further attention and exploration.

- There is a problem of poor domain portability in developing dialogue systems for different domains. How to use the transfer learning to transfer the original pre-trained dialogue management model to the new dialogue system management tasks, that is also an area worth exploring.

- Data set is an important factor affecting the performance of task-based dialogue system. However, at present, the data set of the task-oriented dialogue system is scarce, which affects the development of the dialogue system. How to build and collect conversational data sets with good universality and high dialogue quality is an urgent problem to be solved.

### 5. Acknowledgments
This work is partially supported by the National Natural Science Foundation of China(Nos. 61562068, 11704229, 61640204, 61806103); Natural Science Foundation of Inner Mongolia (No. 2017MS0607); Mongolian Informatization Special Support Sub-project of Inner Mongolia Nationalities Committee (MW-2014-MGYWXXH-01); Youth Innovation and entrepreneurship talents of Inner Mongolia “grassland talents” project; Inner Mongolia science and technology planning project (research and
development operation of online interactive learning platform of Mongolian language MOOC and analysis and research of national education big data); Graduate Innovation Fund of Inner Mongolia Normal University (CXJS18112).

References

[1] Xu, W., Hoffmann, R., Zhao, L., Grishman, R.: Filling knowledge base gaps for distant supervision of relation extraction. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 665–670 (2013).

[2] Bordes, Antoine, Y.L. Bourreau, and J. Weston. "Learning End-to-End Goal-Oriented dialogue" (2016)

[3] Wen TH, Vandyke D, Mrksic N, et al. A Network-based End-to-End Trainable Task-Oriented dialogue System[J]. 2016.

[4] Li X, Chen Y N, Li L, et al. End-to-End Task-Completion Neural dialogue Systems[J]. 2017.

[5] J. D. Williams and G. Zweig. End-to-end lstm-based dialogue control optimized with supervised and reinforcement learning. arXiv preprint arXiv:1606.01269.

[6] Eric M, Manning C D. A Copy-Augmented Sequence-to-Sequence Architecture Gives Good Performance on Task-Oriented dialogue[J]. 2017.

[7] Eric M, Manning C D. Key-Value Retrieval Networks for Task-Oriented dialogue[J]. 2017.

[8] Wen H, Liu Y, CheW, et al. Sequence-to-Sequence Learning for Task-oriented dialogue with dialogue State Representation[J]. 2018.

[9] Xiujuan Li, Zachary C. Lipton, Bhuvan Dhingra, Lihong Li, Jianfeng Gao, Yun-Nung Chen. 2016. A User Simulator for Task-Completion dialogues.

[10] Dhingra B, Li L, Li X, et al. Towards End-to-End Reinforcement Learning of dialogue Agents for Information Access[J]. 2016.

[11] Branton K, Lester J, Mott B. dialogue Management for Conversational Case-Based Reasoning [C]//Advances in Case-based Reasoning, European Conference, Eccbr, Madrid, Spain, August 30 September. DBLP, 2004.

[12] McTear M F. Modeling spoken dialogues with state transition diagrams: experiences with the CSLU toolkit[C]// Pro 5th International Conference on Spoken Language Processing. 1998: 5-7.

[13] Yuan Yan, Tian Huafeng, Du Bo, et al. Research and implementation of frame-based dialogue management model [J]. Computer engineering, 2005, 31(13):212-214.

[14] Abe K, Kurokawa K, Taketa K, et al. A new method for dialogue management in an intelligent system for information retrieval[C]// Proc of the 6th international conference on spoken language processing (ICSLP). Beijing, China, 2000: 118-121.

[15] Rosenblum DS, Wolf AL. A design framework for Internet-scale event observation and notification[J]. ACM SIGSOFT Software Engineering Notes, 1997, 22(6):344-360.

[16] Bohus D, Rudnicky A I. The RavenClaw dialogue management framework: Architecture and systems[J]. Computer Speech & Language, 2009, 23(3):332-361.

[17] Bos J, Klein E, Lemon O, et al. DIPPER: Description and formalisation of an information-state update dialogue system architecture [C]// Proc of 4th SIGdial Workshop on Discourse and dialogue, 2003: 115-124.

[18] Larsson S, Traum D R. Information state and dialogue management in the TRINDI dialogue move engine toolkit [J]. Natural language engineering, 2000, 6:323-340.

[19] Nguyen A. An Agent-Based Approach to dialogue Management in Personal Assistants[C]// International Conference on Intelligent User Interfaces. DBLP, 2005.

[20] Xu kaihua, Zhang deqin, yao Lin. Design and implementation of agent-based human-machine dialogue system [J]. Computer engineering, 2007, 33(16).

[21] Wu X, Zheng F, Xu M. Topic Forest: a plan-based dialogue management structure[C]// IEEE International Conference on Acoustics. 2001.

[22] Wu xiaojun, zheng fang, xu mingxing. A dialogue management model based on thematic forest structure [J]. Acta automatica sinica, 2003, 29(2):275-283.
[23] Levin E, Pieraccini R, Eckert W. Using Markov decision process for learning dialogue strategies. [C]// IEEE International Conference on Acoustics. IEEE, 2002.
[24] Howard, R. A. Dynamic programming and markov processes[J]. 1960.
[25] Cassandra A R. Exact and approximate algorithms for partially observable Markov decision processes[M]. Brown University, 1998.
[26] HENDERSON M S. Discriminative methods for statistical spoken dialogue systems[D]. [S.I]: University of Cambridge, 2015.
[27] Young S, Gasic M, Thomson B, et al. POMDP-Based Statistical Spoken dialogue Systems: A Review[J]. Proceedings of the IEEE, 2013, 101(5):1160-1179.
[28] WILLIAMS J D. A case study of applying decision theory in the real world: POMDPS and spoken dialogue systems[J]. Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions. 2010:315-342
[29] YASAVURU. Statistical dialogue management for health interventions[D]. [S. I.]: FLORIDA INTERNATIONAL UNIVERSITY, 2014.
[30] Roy N. Spoken dialogue Management Using Probabilistic Reasoning[J]. Acl, 2000.
[31] Choi J, Kim K E. Inverse Reinforcement Learning in Partially Observabll Environments[C] // International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc. 2009.
[32] Thomson B, Young S. Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems[J]. Computer Speech and Language, 2010, 24(4):562-588.
[33] Young S, Ga’I?l M, Keizer S, et al. The Hidden Information State model: A practical framework for POMDP-based spoken dialogue management[J]. Computer Speech and Language, 2010, 24(2):150-174.
[34] Williams J D, Young S. Partially observable Markov decision processes for spoken dialogue systems[J]. Computer Speech and Language, 2007, 21(2):393-422.
[35] Williams J D, Young S. Scaling POMDPs for Spoken dialogue Management[J]. IEEE Transactions on Audio, Speech and Language Processing, 2007, 15(7):2116-2129.
[36] Williams, Jason D. Challenges and Opportunities for State Tracking in Statistical Spoken dialogue Systems: Results From Two Public Deployments[J]. IEEE Journal of Selected Topics in Signal Processing, 2012, 6(8):959-970.
[37] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with Deep Reinforcement Learning [J]. Computer Science, 2013.
[38] Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540):529-533.
[39] Lipton Z C, Li X, Gao J, et al. BBQ-Networks: Efficient Exploration in Deep Reinforcement Learning for Task-Oriented dialogue Systems[J]. 2016.
[40] Fatemi M, Asri L E, Schulz H, et al. Policy Networks with Two-Stage Training for dialogue Systems[J]. 2016.
[41] Zhao T, Eskenazi M. Towards End-to-End Learning for dialogue State Tracking and Management using Deep Reinforcement Learning[J]. 2016.
[42] Su P H, Gasic M, Mrksic N, et al. Continuously Learning Neural dialogue Management[J]. 2016.
[43] Williams J D, Asadi K, Zweig G. Hybrid Code Networks: practical and efficient end-to-end dialogue control with supervised and reinforcement learning[J]. 2017.
[44] Peng B, Li X, Li L, et al. Composite Task-Completion dialogue Policy Learning via Hierarchical Deep Reinforcement Learning[J]. 2017.
[45] Liu B, Lane I. Iterative Policy Learning in End-to-End Trainable Task-Oriented Neural dialogue Models[J]. 2017.
[46] Baolinf Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and KamFai Wong, “Integrating planning for task-completion dialogue policy learning, “arXiv preprint arXiv:1801. 06176, 2018.
[47] Lipton Z C, Li X, Gao J, et al. BBQ-Networks: Efficient Exploration in Deep Reinforcement Learning for Task-Oriented Dialogue Systems[J]. 2016.
[48] Singh S, Lewis R L, Barto A G, et al. Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective[J]. IEEE Transactions on Autonomous Mental Development, 2010, 2(2):70-82.

[49] Jaderberg M, Mnih V, Czarnecki W M, et al. Reinforcement Learning with Unsupervised Auxiliary Tasks[J]. 2016.

[50] Houthooft R, Xi C, an D, et al. Variational Information Maximizing Exploration[J]. 2016.

[51] Peng B, Li X, Gao J, et al. Adversarial Advantage Actor-Critic Model for Task-Completion dialogue Policy Learning[J]. 2018.

[52] Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]// International Conference on Neural Information Processing Systems. 2014.