基于强化学习的任务型对话策略研究综述

徐 恺 王振宇 王 旭 秦 华 龙宇轩

(华南理工大学软件学院 广州 510006)

摘 要 对话系统在自然语言处理中发挥着重要作用,具有较好的实际应用前景和许多值得研究的方向.对话策略是基于管道方法的人机对话系统的核心组件,能够根据对话状态生成响应动作,进而指导对话生成.对话策略学习常建模为(半)马尔可夫决策过程,然后通过强化学习求解.近年来,基于强化学习算法解决任务型对话策略问题的研究层出不穷,而相关综述缺乏.因此,本文对基于强化学习的任务型对话策略进行分析、归类、总结.首先,介绍分类强化学习的一般模型,并基于强化学习的分类,分析并总结现有对话策略学习的一般思路和存在问题;其次,基于不同的研究热点,包括多领域、多模态、多代理和共情对话策略,深度剖析新近研究的理论模型、研究进展和存在的问题;接着,针对对话策略的相关研究,包括用户模拟器、对话策略评估、对话策略平台与数据集以及大语言模型与对话策略等进行介绍;针对现有研究的不足,本文从 5 种不同的角度分析了对话策略的未来研究方向;最后,对全文进行总结与展望.本文不仅从强化学习分类上概述任务型对话策略,而且从应用的角度分类任务型对话策略,全方面、多角度地综述了任务型对话策略,为未来的任务型对话策略的研究提供启示.

关键词 对话策略;强化学习;任务型对话系统;深度强化学习;多领域;多模态

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2024.01201

A Survey of Task-Oriented Dialogue Policies Based on Reinforcement Learning

XU Kai WANG Zhen-Yu WANG Xu QIN Hua LONG Yu-Xuan

(*School of Software Engineering, South China University of Technology, Guangzhou* 510006)

Abstract The dialogue system holds a crucial position within the realm of natural language processing (NLP), serving as a significant and valuable component in facilitating human-machine interaction. At present, the dialogue system has attracted more and more attention in both academic and industrial communities because it is conversational for real-world applications as well as valuable in academic prospects. The pipeline-based human-computer dialogue systems consist of four distinct modules, with dialogue policy learning serving as a central component. In the pipeline framework, dialogue policy learning is responsible for selecting suitable dialogue actions based on the dialogue states obtained from the modules of natural language understanding and dialogue state tracking. These selected actions subsequently drive the natural language generation process to produce a coherent and complete response. Dialogue policy learning is commonly formulated as either a Markov decision process (MDP) or a semi-Markov decision process (SMDP). These processes are subsequently addressed by the means of reinforcement learning methods as a sequential decision problem. In recent years, there has been a rapid expansion of research methods focused on studying task-oriented dialogue policy learning using reinforcement learning methods. However, to the best of our knowledge, the existing reviews on dialogue policy learning based on reinforcement

收稿日期:2023-04-01;在线发布日期:2024-01-17.本课题得到广东省重点领域研发计划项目(2021B0101190002)资助.**徐 恺**,博士研究生,主要研究方向为任务型对话策略、强化学习和深度强化学习.E-mail:sekxu@mail.scut.edu.cn.**王振宇**(通信作者),博士,教授,博士生导师,主要研究领域为自然语言处理、对话系统和深度强化学习.E-mail:wangzy@scut.edu.cn.**王 旭**,博士研究生,主要研究方向为对话系统、自然语言生成和深度强化学习.**秦 华**,硕士研究生,主要研究方向为对话策略、强化学习和深度强化学习.**龙宇轩**,博士研究生,主要研究方向为深度学习、云计算和深度强化学习.

learning fall notably short in terms of comprehensiveness and depth. Therefore, the primary focus of this paper revolves around task-oriented dialogue policy learning utilizing reinforcement learning methods. We undertake an all-sided analysis, categorization, and comprehensive synthesis of task-oriented dialogue policy learning based on reinforcement learning techniques. First, we classify the reinforcement learning algorithms that are commonly used in dialogue policy learning. Then, based on the classification of reinforcement learning, we analyze the concept of dialogue policy learning in general, and summarize the problems or limitations in the existing dialogue policy learning methods. Furthermore, we present a comprehensive examination of current research directions and obstacles in the field of dialogue policy learning, which encompass various prominent areas of investigation such as multi-domain, multi-modal, multi-agent, and empathetic dialogue policies. Next, we proceed to introduce additional pertinent studies pertaining to dialogue policy learning. These encompass investigations on user simulators, methodologies for evaluating dialogue policy learning, dialogue policy platforms, datasets tailored for dialogue systems, as well as the interplay between large language models and the learning of dialogue policies. In order to rectify the deficiencies found in current research on dialogue policy learning, this paper undertakes an analysis of the prospective research directions for dialogue policy learning from five distinct vantage points. These perspectives encompass the realms of reinforcement learning technology and various applications. In conclusion, we wrap up this article and turn our gaze toward the future of dialogue policy learning. This paper not only provides a classification and comprehensive overview of task-oriented dialogue policy learning based on reinforcement learning algorithms but also categorizes it from different application perspectives. It offers a multi-dimensional, comprehensive, and systematic synthesis of task-oriented dialogue policy learning. We believe that this paper can provide valuable insights and inspiration for future research in task-oriented dialogue policy learning, and promoting the development of human-machine dialogue systems.

Keywords dialogue policy; reinforcement learning; task-oriented dialogue systems; deep reinforcement learning; multidomain; multimodal

1 引 言

任务型对话系统旨在帮助用户完成一个或多个特定的任务,如餐厅预订、航班预订、智能客服等.它以完成特定任务为导向,并希望以较少的对话轮数完成对话任务.任务型对话系统的实现方式有两种,分别是管道(Pipeline)方法(也称管方法)和端到端(End to End,E2E)的方法.管道方法将任务型对话系统看成管道形式,构建多个子模块,通过各个子模块的级联实现对话[1].端到端的方法则直接根据对话文本建模并生成系统回复[2].端到端的方法和管道方法各有优劣.端到端的方法在新的对话场景上更加灵活,使用单一模块,具备更低的开发时间和更小的维护成本.但端到端的方法需要大量的数据进行训练,对设备的依赖程度较高.此外,由于端到端方法直接映射输入到输出,使得该类方法解释性不足且不易控制.管道方法具有更好的解释性且易于实施,是近年来商业对话系统的首选[3],但管道形式的模块化设计通常较为复杂,且在模块集成时较难保证全局最优.

管道方法将任务型对话系统分成四个模块,分别是自然语言理解、对话状态跟踪、对话策略(也称对话策略学习)和自然语言生成[4].其中,一些研究将对话状态跟踪和对话策略统称为对话管理[5-6].图1给出了基于管道方法的任务型对话系统的整体过程.自然语言理解将原始用户话语(Utterance)转换为语义信息,获取领域内用户意图和槽值信息.对话状态跟踪根据自然语言理解的信息评估得到用户目标和请求,构建并记录对话状态.对话策略根据对话状态来决策系统采取的动作.自然语言生成将对话策略生成的对话动作转换为最终的自然语言.由于对话策略根据自然语言理解的结果和对话状态跟踪的输出来制定对话动作,而自然语言生成可以根

据对话动作基于模板生成对话.因此,对话策略学习在自然语言理解和自然语言生成中发挥着重要作用.本文主要关注对话策略.

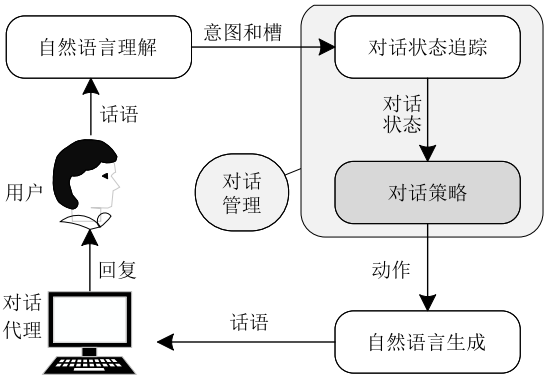


图 1　基于管道方法的对话系统一般过程

早期对话策略从对话状态中映射规则来得到对话动作,被称为基于规则的对话策略[7],该类方法首先确定对话的终止状态和各个状态的转移条件,然后为每个状态定义对应的回复动作.在对话过程中从初始对话状态出发,根据对话动作和状态转移的条件选择下一状态,直到对话到达终止状态.基于规则的对话策略可通过槽填充的方法[8]、有限状态机的方法[9]和监督学习的方法[10-11]进行实施.基于规则的对话策略具有易于分析和调试的优点.然而,由于高度依赖专家干涉,它们的灵活性和可扩展性较差.

在真实场景中,用户的行为是难以预测的,而语音识别和自然语言理解会不可避免地存在误差,这使得根据真实对话状态做出决策变得困难.且在数据规模巨大的情况下,基于规则的对话策略存在查询结果慢的问题.因此,一些研究将对话策略看成马尔可夫决策过程(Markov Decision Process,MDP),然后基于强化学习进行求解.Levin 等人[12]最早将对话策略学习视为 MDP,他们概述了对话策略建模为 MDP 问题的复杂性,这奠定了基于强化学习求解对话策略的基础.之后,对话策略的研究大多基于强化学习进行.基于线性强化学习建模对话策略是早期的研究方法,其将对话策略学习看成是一个从对话状态到对话动作的线性映射[13].基于线性强化学习的对话策略不能获取非线性结构,且计算代价会随着训练数据的增加而增加.尽管有研究将线性强化学习和监督学习进行结合[14],但忽略了当前策略对未来的影响,容易导致次优的结果.

随着深度学习的兴起,将深度学习和传统强化学习相结合构建深度强化学习成为了对话策略的主流研究,因为深度学习具备更强的非线性拟合能力.深度强化学习按强化学习的类别可分为基于模型(Model-Based)和免模型(Model-Free)的方法.基于模型的方法如 DDQ(Deep Dyna-Q)[15],是第一个将对话规划(Planning)应用于任务型对话策略的深度强化学习框架.免模型的方法包括基于值函数逼近的方法、基于策略梯度的方法、基于层次强化学习的方法等,后文将详细介绍.

目前,对话策略的研究得到了快速发展,研究者们从不同的角度撰写与之相关的研究综述.Zhao 等人[16]综述了对话管理,他们在对话策略上仅讨论了经典的深度 Q 网络(Deep Q-learning Network,DQN)和 A2C(Adversarial Advantage Actor-Critic)算法,讨论的模型有限且没有很好地进行模型局限性分析.Dai 等人[17]综述了基于层次强化学习、基于封建强化学习和基于 DDQ 的对话策略模型,并没有从应用的角度对对话策略进行分类介绍.Kwan 等人[18]介绍基于强化学习的任务型对话策略,但他们的分类方式是按照强化学习的模块进行,因此,他们的阐述围绕在基于强化学习的环境、策略、状态空间、动作空间和奖励学习.这种分类方法忽略了目前对话策略的主流研究领域,缺乏横向比较,使得特定领域的研究人员较难提取领域内的价值信息.此外,有一些对话系统的综述也讨论了对话策略学习,如:赵等人[1]综述了任务型对话系统中的管道方法和端到端的方法;Zhang 等人[3]对任务型对话系统的不同模块分别阐述,并从数据利用率和多轮对话的动态特性上描述了任务型对话系统中存在的挑战.但他们的综述涉及到较少的对话策略学习.随着人机对话的快速应用,对话策略作为其核心组件,在近些年涌现出较多的新技术和新应用,急需对该领域的知识进行梳理、归类和总结.本文对基于强化学习的对话策略进行细粒度讨论,分析新近研究热点,希望为基于强化学习的任务型对话策略的进一步研究提供指导意义.本文的主要贡献为:

(1)分类了对话策略所用到的相关强化学习算法,分析了不同强化学习算法的问题及解决思路;按照强化学习算法的分类分别介绍任务型对话策略的研究进展、可用性、存在问题与挑战.

(2)从应用场景角度将新近的基于强化学习的对话策略分成多领域、多模态、多代理和共情对话策略;针对每个应用场景进行深度剖析,分析新近具有代表性的研究方法和它们存在的局限性.

(3)调研对话策略的辅助模块,包括对话策略

的用户模拟器、对话策略平台、对话策略的评估方法和相关的数据集.从技术、多领域、多模态、多代理和共情的角度分别讨论了任务型对话策略未来的研究方向.有助于读者从整体上深入了解对话策略研究并选择新的研究方向.

为了让读者快速定位到相关内容,本文绘制图 2 的组织结构.第 2 节介绍分类强化学习的一般方法,并分析它们的优势和劣势;第 3 节基于强化学习的分类详细综述对话策略的研究方法;第 4 节从应用场景角度对基于强化学习的多领域、多模态、多代理、共情对话策略进行综述;第 5 节介绍对话策略的用户模拟器、评估方法、对话策略平台和数据集;第 6 节对基于强化学习的任务型对话策略的未来研究方向进行分析;最后,第 7 节对本文的工作进行总结与展望.

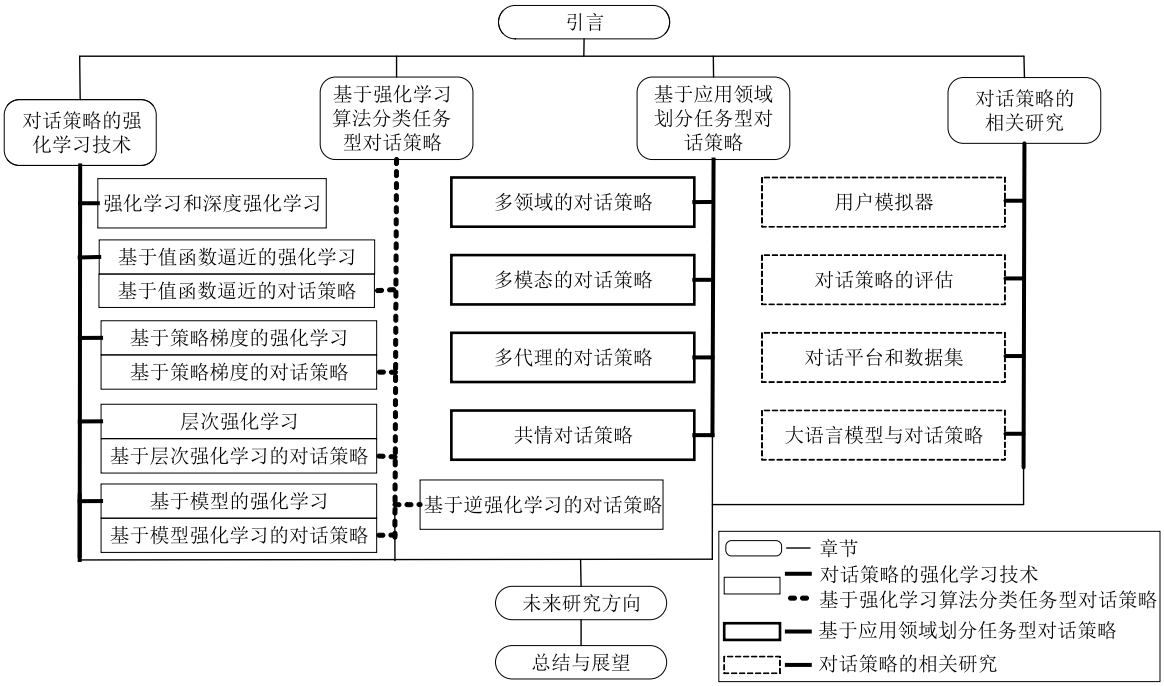


图 2 本文的总体结构图

2 对话策略的强化学习技术

本节分类介绍强化学习模型,它们与新近的对话策略研究息息相关.对这些模型归类介绍有利于探索不同强化学习算法在任务型对话策略上的应用.本节首先对强化学习、深度强化学习,以及基于强化学习建模的对话策略进行介绍.接着对强化学习算法进行分类,介绍基于值函数逼近的强化学习、基于策略梯度的强化学习、层次强化学习和基于模型的强化学习,评估各类强化学习模型的优缺点,以及分析现有的改进方案.

2.1 强化学习和深度强化学习

强化学习(Reinforcement Learning,RL)被定义为在环境的不同状态下采取行动,使累积的奖励最大化,用于解决序贯决策问题[19],常常被建模成马尔可夫决策过程 MDP[20],图 3(a)展示了强化学习的一般流程.MDP 用于描述随机控制过程,被定义成一个五元组$\langle S,A,P,R,\gamma\rangle$,其中 S 表示状态空间,$s_t\in S$ 表示 t 时刻的一个状态.A 是动作空间,$a_t\in A$ 表示代理在 t 时刻采取的动作.P 为状态转移概率,代理在当前状态 s_t 下采取动作 a_t 转移到下一状态 s_{t+1} 的概率为 $P(s_{t+1}\mid s_t,a_t)$.如果 MDP 状态转移是随机的,则 $P(s_{t+1}\mid s_t,a_t)\neq 1$.$R$ 是奖励函数,$R(s_t,a_t,s_{t+1})$表示代理在状态 s_t 下执行动作 a_t 转移到状态 s_{t+1} 所获得的奖励,简写为 r_t.γ 是折扣因子,用于惩罚远距离的奖励.对于一个给定的 MDP,代理和环境进行交互得到观测样本(s_t,a_t,r_t,s_{t+1}).强化学习的求解通常需要计算状态-动作值函数(也称为 Q 值函数).它是从任意一个状态 s_t 开始,通过策略 π 采取动作 a_t 所获得累计奖励的期望,表达式为

$$Q^{\pi}(s_t,a_t)=\mathbb{E}_{\pi}\Big[\sum_{k=0}^{T-1}\gamma^k r_{t+k}\mid s_t=s,a_t=a\Big],\quad \forall s\in S,\forall a\in A,\forall t\geq 0 \tag{1}$$

其中,π 为策略,折扣因子 $\gamma\in[0,1)$.T 是完成序贯决策任务所需的步骤.

在一个情景(Episodic) MDP 下,完成一项任务从一个初始状态开始,到一个终止状态结束,所以状

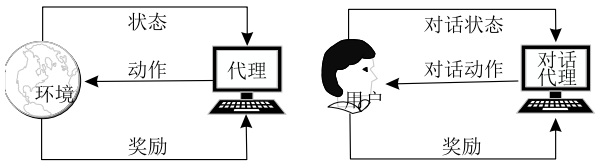
态通常在 T 步被重置；在一个非情景(Non-Episodic) MDP 下 $T=\infty$. 此时状态、动作和奖励的收集将在一个单独的情景下进行[21]. 强化学习的主要目的是找到一个最优的策略 π^*，满足：

$$\pi^*=\arg\max_{\pi} Q^{\pi}(s_t,a_t),\ \forall s_t\in S,\forall a_t\in A \quad (2)$$

强化学习是一个基于试错的算法，它通过增加正确预测的奖励和惩罚错误预测的奖励来学习. 在有限状态情况下，值函数(包括 Q 值和状态值)的计算通过一个表格进行，表格的索引是状态或者状态-动作对，此时，值函数的更新是对该表格的更新，这种基于表格更新的方法也被称为基于表格的强化学习[22]. 但是，当状态空间巨大或者状态空间连续的时候，基于表格的方法难以精确地评估 $Q^{\pi}(s_t,a_t)$.

深度神经网络(Deep Neural Network，DNN)也称深度学习，其通过梯度下降法自动地映射特征，相比于传统的手动提取特征的方式，DNN 能够大大减少对领域知识的依赖[23]. 深度强化学习是使用 DNN 近似强化学习的值函数、策略、状态转移函数或奖励函数等的一种方法[24]. 它能够解决强化学习状态空间巨大和状态连续的问题. 深度学习和强化学习分别在 2013 年和 2017 年被选为全球十强技术之一[25]，甚至有研究者构建了一个“人工智能＝深度学习＋强化学习”的公式[26]，由此可见深度强化学习的价值及重要性.

目前，对话策略的建模多通过深度强化学习进行. 图 3(b)是对话策略学习的一般流程，其中省略了自然语言理解和自然语言生成. 从图 3(b)可以看出，基于强化学习的对话策略将用户看成环境，并根据对话状态和对话奖励选择对话动作. 接下来对强化学习算法进行归类介绍.



(a) 强化学习流程　　(b) 对话策略学习流程

图 3　强化学习和对话策略学习的一般流程

2.2　基于值函数逼近的强化学习

由于任务型对话策略的状态空间巨大，直接基于表格型强化学习很难满足其需求. 基于值函数逼近的强化学习方法构建一个动作值函数 $Q(s,a)\approx f(s,a;\theta)$或状态值函数 $V(s)\approx f(s;\theta)$，其中 θ 为参数，来建模状态空间巨大的情况. 对值函数的逼近方法有两种形式，可分为线性逼近和非线性逼近. 线性逼近方法如 Sarsa 算法和 Q-Learning 算法[22]. 前者通过增量法近似值函数，通过迭代计算真实值函数和预测值函数的误差来学习参数，具有计算简单但利用率低的特点；后者通过批量法近似值函数，将一段时期内的数据集中处理，具有计算效率高但计算任务复杂的特点. 对于更为复杂的状态空间，通过线性方式逼近值函数往往并不能达到理想的效果. 而深度神经网络通过构建线性或非线性函数，理论上可以逼近任意形式的目标函数[22]. 对话策略中常使用深度学习结合强化学习方法逼近值函数.

深度 Q 网络(Deep Q-Network，DQN)最早由 Mnih 等人[27]提出，它是基于深度神经网络解决大规模状态空间中动作值函数近似的问题. 在每个时间步 t 时存储代理的经验 $e_t=(s_t,a_t,r_t,s_{t+1})$到经验池 $D=\{e_1,\cdots,e_t\}$，然后随机地从经验池中选取样本 $(s_t,a_t,r_t,s_{t+1})\sim D$，并使用损失函数

$$L_i(\theta_i)=\mathbb{E}_{(s_t,a_t,r_t,s_{t+1})\sim D}[(r+\gamma\max_{a'}(s_{t+1},a';\theta_i^-)-Q(s_t,a_t;\theta_i))^2] \quad (3)$$

通过 Q-learning 算法进行更新. 损失函数中 γ 为折扣因子，其作用是使算法更加关注短期奖励. θ_i 是第 i 次迭代的 Q 网络的参数，θ_i^- 是在第 i 次迭代目标网络的参数. 因为观测序列的相关性，传统的方式直接通过神经网络来逼近动作值函数会导致效果不稳定甚至是发散的，并且轻微地更新 Q 值可能会显著地改变策略和数据的分布. DQN 通过两种方法来解决上述问题，一是构建经验池，并随机从经验池中进行采样，这样做的目的是消除观测序列的相关性，并且平滑数据分布的变化. 二是定期地迭代 Q 值使其向目标值靠近，从而减少目标的相关性，在固定的迭代次数 K 后用 Q 网络更新目标网络，也就是每 K 步目标网络复制 Q 网络的参数值.

DQN 评估 Q 值时会选择最大的 Q 值进行迭代，存在过估计 Q 值的问题，为此，一些研究提出了改进方案. 如 DDQN(Double DQN)算法[28]通过 Q 网络值最大的动作来更新目标网络的 Q 值，在一定程度上缓解了 DQN 的过估计问题. Averaged DQN[29]通过平均先前学习到的 Q 值来使得训练更稳定. Maxmin DQN[30]利用 N 个动作值函数得到 N 个最大的 Q 值，然后从这 N 个 Q 值中选取最小值作为最终的 Q 值. 此外，TQC 算法(Truncated Quantile Critics)[31]认为 Maxmin DQN 的评估方式是粗糙的，因此 TQC 通过集成多个近似分布来促进 Q 函数的评估. Tian 等人[32]认为最优的 Q 值介于最大 Q

值和最小 Q 值之间，提出一个动态加权评估器，用于加权平均预测的最大 Q 值与最小 Q 值. 他们通过启发式搜索最大 Q 值与最小 Q 值的权重，通过加权获得最优的 Q 值，但启发式算法本身会加剧计算量.

DQN 算法存在探索低效性问题[33]，这是由于强化学习的监督信号较监督学习更为稀疏，而高维的稀疏状态空间加剧了该问题. 传统的深度强化学习通过 ϵ-greedy 算法[34]进行探索，通过探索率 ϵ 来进行探索，理论上可以在足够的时间内解决奖励稀疏问题，但这种方式并不是采样高效的. Agent57 算法[35]在训练过程中动态地调节折扣因子，使得训练过程中能够获取更高的奖励值. Lipton 等人[36]使用贝叶斯探索策略以促进代理探索不确定的动作，在对话动作的选择上优于 ϵ-greedy 算法. Li 等人[37]使用对话状态生成器训练一个判别器，将训练后的判别器作为对话奖励模型纳入强化学习的训练. 他们的方法能够间接地促进探索能力. Pong 等人[38]通过学习 Q 函数来预测到达目标(完成任务)的距离，并用它来指导即时奖励，有利于对目标的探索. 更多解决深度强化学习探索低效的问题可参见综述论文[39].

存在一些算法在 DQN 的网络结构上进行修改，如竞争 DQN 网络(Dueling DQN，简称 Dueling)[40]将 Q 网络的结构修改成状态值函数和状态下各动作的优势函数之和. 通过将 Q 值分解成两个部分，Dueling 可以学习到每个状态的价值，不需要了解每个状态对每个动作的影响. 分布 DQN(Distributional DQN)[41]将 DQN 的输出转化成 Q 值的分布，该算法认为在特定状态和动作下 DQN 的输出是随机变量，因此转化成 Q 值的分布将会更具稳定性，且能更好地进行策略学习. 彩虹 DQN(Rainbow DQN)[42]是一种融合多种 DQN 强化学习算法的模型，融合的模型包括竞争 DQN、分布 DQN 等，并且考虑了优先经验回放. 彩虹 DQN 在游戏领域进行了验证，取得了较单个基于 DQN 的算法更优的效果.

2.3 基于策略梯度的强化学习

基于值函数逼近的算法能解决状态空间巨大或连续的问题，然而，对于动作空间巨大或连续的情况，使用基于该类算法求解状态-动作所对应的值函数是不现实的. 因此，一些算法直接对策略进行优化，即对于策略 $\pi_\theta(a|s)=P(a|s;\theta)$，直接优化参数 θ，这种直接优化强化学习策略的方法称为基于策略的模型. 它们依据策略梯度(Policy Gradient，PG)理论[43]，当 $\pi_\theta(a|s)$ 选取的动作不确定时称为随机策略，反之则称为确定性策略. 策略梯度为

$$\nabla_\theta J(\pi_\theta)=\mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T}\nabla_\theta\log\pi_\theta(a_t|s_t)\boldsymbol{\Phi}_t\right] \quad (4)$$

其中，$\boldsymbol{\Phi}_t$ 为与环境交互获取的奖励，$\tau=\{s_0,a_0,\cdots,s_T,a_T\}$ 为对话轨迹，T 为轨迹长度.

REINFORCE[24]是一种经典的随机策略算法，其交互奖励 $\boldsymbol{\Phi}_t$ 考虑从时间 t 开始，到状态结束的整个完整轨迹的奖励. REINFORCE 采样的轨迹通常具有较高的方差，在奖励中添加基线可以降低方差，通常的操作是将状态值函数视为基线，因此调整交互奖励为

$$\boldsymbol{\Phi}_t=Q(s_t,a_t)-V(s_t) \quad (5)$$

策略 $\pi_\theta(a|s)$ 的参数更新方式为 $\pi_\theta=\pi_\theta+\alpha\nabla_\theta J(\pi_\theta)$，其中 α 是迭代的步长，其影响着对话策略更新的收敛速度. 置信域策略优化(Trust Region Policy Optimization，TRPO)[44]确定一个合适的步长，使得策略在更新的过程中单调不减.

然而，随机策略在计算策略梯度时需要对整个动作空间进行积分，且需要大量的采样轨迹，是计算昂贵的. 为了缓解该问题，Silver 等人[45]提出了确定性策略(Deterministic Policy Gradient，DPG)，并验证了在高纬度的动作空间中，DPG 较传统的随机策略方法更高效. 之后，DeepMind 团队结合 DQN 提出 DDPG(Deep DPG)[46]算法，使用离线策略数据和 Bellman 方程学习 Q 值函数，并使用 Q 值函数学习策略. 为了解决 DDPG 算法的目标策略平滑问题，TD3(Twin Delayed DDPG)[47]算法在策略中增加了噪声，使得策略网络能够更好地探索动作空间. SAC(Soft Actor-Critic)算法[48]也在 DDPG 的基础上做了改进，以进一步提高算法性能，在优化对象上增加了熵正则化项，鼓励策略保持随机性，从而增加探索能力. 目前，将基于 DPG 的方法应用于任务型对话策略的研究尚未被发现.

基于值的方法通常面临着高偏差的问题，而基于策略的方法通常面临着高方差的问题. 缓解上述问题通常可采用基于 AC(Actor-Critic)的算法[22]，它是一种混合了基于值和基于策略的方法. 基于 AC 算法定义了演员(Actor)和评论家(Critic)，可以用于单步更新，其中 Critic 用于评估在当前状态下采取特定动作的价值或优势；Actor 则按照 Critic 评估的值更新策略参数. Actor 的 Q 值函数可以通过线性或神经网络的方式逼近. A2C(Advantage Actor-Critic)算法[49]是在 AC 算法的基础上添加了一个优势函数，其 Critic 使用的是 Q 值函数和状

态值函数之差的优势函数. 之后，为了进行多线程处理，提出了 A3C(Asynchronous Advantage Actor Critic)[50]算法. 由于离线的策略梯度算法从固定大小的经验池中获取数据进行训练，当经验池中数据较少时会导致无法有效地估计 Q 值，BCQ(Batch Constrained deep Q-learning)[51]算法通过限定动作空间范围迫使智能体在给定的数据集子集上的动作和策略更加接近. 尽管基于 A2C 的算法能够学习到一个较稳定的策略，但是它们需要精心地设计 Actor 和 Critic 的网络结构.

2.4　层次强化学习

基于值函数逼近和基于策略梯度的方法可以看成是扁平化(Flat)的强化学习，其奖励信号在任务结束才能获得，这会使得该类方法的奖励是稀疏或延迟的. 层次强化学习(Hierarchical Reinforcement Learning，HRL)利用分而治之的思想将一个复杂的、困难的任务转化成多个容易求解的子任务，可以在更长的时间尺度上实现高效的信用分配[52]. 同时子任务的学习也会使得训练过程更具结构化[53]. 层次强化学习具有提供密集奖励信号和提高探索速率的优点[21].

层次强化学习基于半马尔科夫决策理论(Semi-Markov Decision Process，SMDP)[54]，类似于 MDP 的随机控制过程，但与 MDP 不同的是，它涉及动作被选择和执行的时间概念. 选项(Options)框架[55]是最早被提出的在时间维度上抽象的一种层次强化学习方法，传统的强化学习根据当前时刻采取某一动作获取下一时刻的状态奖励，而选项框架则是在不固定的时间步上根据策略执行动作，直到达到终止条件. 对于一个初始状态 s_t，代理选择一个选项(或对应子任务)$w_t\in\Omega$，其中 Ω 是选项的集合，从状态 s_t 开始，执行选项 w_t 获得的奖励为

$$R(s_t,w_t)=\mathbb{E}_{a\sim\pi_{w_t}(s)}\Big(\sum_{i=0}^{N_{w_t}-1}\gamma^i R(s_{t+i},a_{t+i})\,|\,s_t,a_t=\pi_{w_t}(s_t)\Big) \tag{6}$$

其中，N_{w_t} 为选项 w_t 达到终止条件时的总时间步数，i 为时间步数，π_{w_t} 是内部选项策略，它从时间 t 开始，在 N_{w_t} 个时间步(根据终止条件确定)之后终止. 那么其 Q 值函数可以被定义为

$$Q(s_t,w_t)=R(s_t,w_t)+\sum_{s_{t+N_{w_t}},N_{w_t}}\gamma^{N_{w_t}}P(s_{t+N_{w_t}},N_{w_t}|s_t,w_t)\max_{w_{t+N_{w_t}}}Q(s_{t+N_{w_t}},w_{t+N_{w_t}}) \tag{7}$$

其中，$P(s_{t+N_{w_t}},N_{w_t}|s_t,w_t)$ 是 SMDP 的转移函数，定义为

$$P(s_{t+N_{w_t}},N_{w_t}|s_t,w_t)=P(s_{t+N_{w_t}}|s_t,w_t,N_{w_t})P(N_{w_t}|s_t,w_t) \tag{8}$$

表示从状态 s_t 到达 $s_{t+N_{w_t}}$ 的转移概率，N_{w_t} 通常由终止条件决定.

层次强化学习中选项的发现并不容易，一些研究致力于自动提取选项. 如 Bacon 等人[56]结合策略梯度理论渐进地学习内部选项策略和终止函数. Machado 等人[57]基于原型值函数发现选项，该过程可以被看作是遍历状态特征表示中的每个维度. Zhang 等人[58]提出了自监督的内部选项发现框架，以选择的子轨迹为条件，通过熵最小化的方式学习选项，不需要依赖监督数据. Jin 等人[59]将符号知识嵌入到深度强化学习来发现选项，构建了符号选项. 符号知识的引入减轻了领域专家知识，并且具有更好的可解释性，尽管需要预先定义一些符号知识.

层次强化学习的工作还包括基于子目标. 通常构建两层的强化学习，顶层策略用于选择子目标，下层策略则选择动作用于完成子目标. 传统的方式是通过人工构建子目标，如 Kulkarni 等人[60]提出层次深度 Q 网络(Hierarchical-DQN，H-DQN)，使用两个 DQN，顶层策略需要使用预定义的子目标. 然而，人工定义子目标是耗时费力的，因此一些研究着重于发现子目标. Tang 等人[61]基于成功的对话数据将一个复杂的面向目标的任务划分为一组更简单的子目标，然后基于层次强化学习优化策略. Paul 等人[62]使用模仿学习基于专家轨迹分解复杂的任务得到子目标，然后使用这些子目标增强奖励函数，而不影响学习策略的最优性. Pateria 等人[63]提出了一种基于子目标图的规划方法，代理通过对子目标图进行剪枝，解决子目标之间连接错误的问题. Peng 等人[64]验证了子目标的引入能够减轻奖励稀疏，并且增加探索能力.

2.5　基于模型的强化学习

上述介绍的强化学习方法为免模型(Model Free)的强化学习，代理通过与环境交互来学习值函数或者策略参数. 免模型的强化学习直接与环境交互，通过试错的方式获取经验，而现实的世界中试错的代价是昂贵的. 基于模型(Model-Based)的强化学习建立了一个环境模型，使得试错可以在该环境模型中进行，从而减少了在真实环境中试错的代价. 基于模型的强化学习模拟人在想象的世界中做决策，具有提高采样性的优势[65]，可以利用学习到的环境模型生成模拟数据进行训练.

在基于模型的强化学习 MDP 的五元组$\langle S,A,P,R,\gamma\rangle$中，状态 S、动作 A 和折扣因子 γ 是预定义好的，状态的转移 P 和奖励函数 R 需要被学习. 其

训练的方式和免模型强化学习方式类似,可通过表格的方式直接统计得到状态转移概率和奖励函数,或通过计算损失的方式获取.基于模型的强化学习核心是构建模拟环境的模型.根据构建的模型被使用的情况,本文将其分两类.一类是直接交互,通过直接使用构建的模型去提升环境策略交互的过程,根据学习到的模型规划一系列的动作,然后将动作直接迁移到环境的相似场景中,如 MPC(Model Predictive Control)模型[66].另一类是间接的交互,通过构建的模型生成模拟的样本数据,然后利用模拟的数据进行策略或值函数的评估.该类方法如基于 Dyna[67]的方法,已经具有较好的性能和理论支撑.由于基于模型的强化学习构建的模型利用已有的训练数据得到,而策略通过与真实的环境交互得到训练数据,这会使得构建的模型并不总是在策略上取得高奖励[68].VAML(Value-Aware Model Learning)模型[69]对构建的环境模型进行优化,将值函数的信息并入到模型的学习,使环境和构建的模型之间的单步值函数估计差异最小化.Voelcker 等人[70]利用当前值函数的梯度信息重新调整均方误差损失函数,使得在大规模的状态动作对上学习更为准确.

基于模型的强化学习有一些优势,包括数据高效性、目标探索高效性、可迁移性、安全性和可解释性.但其需要逼近 MDP 的动力学模型,如何在环境随机性、有限数据的不确定性、部分可观测性等前提下构建一个较优的模型依然存在挑战.

3 基于强化学习算法分类任务型对话策略

上述总结了任务型对话策略的强化学习技术,本节介绍基于不同的强化学习算法研究对话策略的方法,探究它们在对话策略中的优势和不足.首先,介绍基于值函数逼近的任务型对话策略;然后,介绍基于策略梯度的任务型对话策略;接着,介绍基于层次强化学习的任务型对话策略和基于模型的强化学习构建任务型对话策略;最后,介绍基于逆强化学习的任务型对话策略.

3.1 基于值函数逼近的对话策略

对话策略是基于管道方法的任务型对话系统的核心组件之一,它根据对话状态跟踪得到的对话状态作为输入,通过强化学习输出对话动作.基于值函数逼近的对话策略通过拟合 Q 值函数来获取动作,获取的动作将用于后续对话生成.本节从两个方面考虑基于值函数逼近建模对话策略学习:(1)直接逼近,直接基于值函数逼近的强化学习方法建模对话策略;(2)联合逼近,将基于值函数逼近的强化学习方法和其它算法相融合建模对话策略.下面分别进行介绍.

直接逼近.直接基于值函数逼近的强化学习方法建模对话策略,该类方法主要是近似 Q 值函数.DQN 作为基于 Q 值函数近似的一种经典方法,将基于 DQN 的方法应用于对话策略是一种较为常见的方式.图 4 展示了基于 DQN 算法求解对话策略的一般过程,其中 s 为对话状态,a 为对话动作,r 为奖励,s'为下一对话状态.当前值网络和用户(或用户模拟器)进行交互,保存经验(s,a,r,s')到经验池,当前值网络从经验池采样状态-动作对(s,a),目标值网络采样 s 对应的下一状态 s',选择最大的状态值作为目标网络的 Q 值函数,并和当前值网络的 Q 值函数进行误差计算以更新参数 θ.在多轮迭代后,目标值网络复制当前值网络的参数 $\theta^- = \theta$.

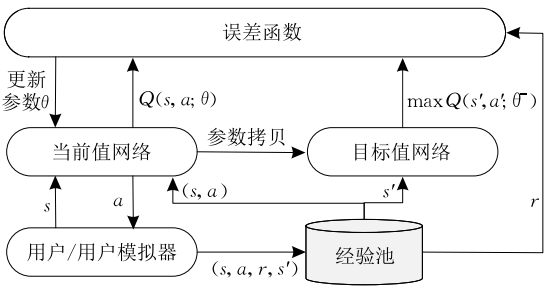


图 4 基于 DQN 算法求解对话策略一般过程

目前,任务型对话策略多基于 DQN 算法进行建模.Wang 等人[71]比较了不同基于 DQN 的算法在对话策略中的性能,涵盖了 DDQN、Dueling、分布 DQN 与优先经验回放的 DQN,在三种数据集,包括电影票预订、餐厅预订和出租车预订上进行测试,实验效果验证 Dueling 和分布 DQN 均能取得较好的效果,而其它算法较次.此外,在对话策略的探索上,该论文验证了基于好奇心探索[72]的 DQN 算法能够更好地提升对话策略学习的性能.然而上述研究并没有考虑对话状态的时序问题.Zhao 等人[73]考虑到对话状态可能是部分可观测的,他们将 DQN 的深度神经网络部分转化成循环神经网络,构建了 DRQN,并结合监督学习进行对话策略学习,其方法较传统的 DQN 算法更优.考虑到 DQN 中存在过估计 Q 值的问题,Tian 等人[32]提出了一个解决 DQN 过估计的方法,并将其应用到对话策略学习.他们在最大 Q 值和最小 Q 值之间找到一个平衡参数,平衡参数可通过启发式算法或者神经网络学到,然后通

过该平衡参数计算得到新的 Q 值.

联合逼近. 考虑到仅使用一种基于 DQN 的算法不能完全地解决对话策略中的不同问题，一些对话策略的研究基于 DQN 融合其它算法. 如 Cao 等人[74]让对话代理同时学习真实的、模拟的和事后经验，使用这些经验为对话策略学习提供更多的对话样本和更积极的反馈信息. 事后经验回放算法[75]是解决传统强化学习中奖励稀疏的一种常用方法. Zhao 等人[76]结合课程学习构建教师模型和学生模型，教师模型作为学生模型的控制器，通过监测对话代理的学习进度和重复惩罚来安排有序课程. 然而，得到的课程只适用于当前训练的模型，而重新训练可能会产生不同的课程顺序. 之后，他们又提出通用的全局课程，可以指导不同的对话策略学习[77]. Zhang 等人[78]构建了一个逆行为评估器用于评估代理的相反动作，然后将评估得到的动作融入到状态中用于策略学习. 因为评估动作的嵌入不仅增大了状态空间，而且最终的效果也受到错误的动作评估影响，该类方法需要严格地定义评估器. Wang 等人[79]基于 Dueling 网络构建蒙特卡洛的对话状态搜索树，通过差分学习选择最佳动作，他们的算法在低质量的模拟经验下具有较好的鲁棒性. Zhang 等人[80]从用户目标数据中学习用户偏好，并整合人类常识知识到深度强化学习模型中，实验结合多种强化学习算法均取得了较好的效果. 考虑到基于强化学习的对话策略可能在训练数据较少的情况下出现过拟合，Madusanka 等人[81]提出了一种合成议程的方法，采用基于奖励的抽样方法对失败的对话行为进行优先排序，避免了直接使用训练数据集进行训练导致模型过拟合的问题. 他们的算法与 DQN 等算法进行结合，验证了提出方法的有效性. 相比于仅使用基于 DQN 方法学习任务型对话策略，混合的方式在新近研究中更为常见. 表 1 概况了基于值函数逼近的任务型对话策略的优缺点.

表 1　基于值函数逼近的任务型对话策略

		文献	优势	不足
基于值函数逼近的对话策略	直接逼近	文献[32,71,73]包括了 DQN、Dueling、分布 DQN、DRQN、优先经验回放和好奇心算法等.	模型结构相对简单，易于实现，是对话策略模型的基础.	在性能上继承了传统强化学习模型的缺点，通常稳定性较差.
	联合逼近	文献[74,76-81]联合的方式包括结合不同经验、结合课程学习算法、人类常识知识等.	模型鲁棒性更强，能够缓解过拟合，性能较直接建模的方法更优，在对话策略中更为常见.	模型结构上更为复杂，进行联合时需考虑可解释性和巧妙地设计联合机制.

3.2　基于策略梯度的对话策略

基于值函数逼近的任务型对话策略通过提高值函数的估计来提升对话策略学习的性能. 考虑到策略梯度方法较基于值函数的强化学习方法具有收敛性快、搜索简单、可以学习随机梯度等优点[22]，基于策略梯度的强化学习在对话策略中也经常被使用. 我们将基于策略梯度的方法分成：(1) 直接建模，直接基于策略梯度建模对话策略；(2) 联合建模，基于策略梯度与其它模型相融合建模对话策略.

直接建模. 直接基于策略梯度建模对话策略，这类方法通常在 A2C 算法上进行改进. Fatemi 等人[82]提出了一种深度 A2C 网络，其优势函数为当前状态的值函数和上一状态的值函数之差. 他们在损失函数中添加 L2 范数正则化项来提高模型的稳定性. Zhao 等人[83]提出一种可调节的策略梯度方法，用于解决传统的基于策略的对话代理关注简单话语和形成次优策略的问题. 他们构建三种方案用于缓解上述问题，一是通过全局调节参数用于鼓励探索；二是并行地运行多个策略，以获取更加稳定的结果；三是基于动作频率动态地调节动作，以提高学习效率. 为了缓解基于策略的对话代理陷于局部最优解，同时提高模型的采样效率，LCPO(Loop-Clipping Policy Optimisation)模型[84]对对话轨迹中无效的动作进行剪除，通过降低无效动作的概率使得策略优化更加平滑和容易获得. 由于基于 A2C 的经验回放(A2C Experience Replay，A2CER)算法已经在动作集非常小的游戏环境中表现出优异的结果. Malviya 等人[85]基于 A2CER 进行对话策略学习，验证了它在对话策略上的有效性.

联合建模. 基于策略梯度与其它模型相融合建模对话策略. Li 等人[37]结合对抗学习网络，使用一个对话生成器训练判别器，用于标注对话是否成功，并通过反馈信号作为策略的奖励，然后将得到的奖励融入到基于策略梯度的 PPO(Proximal Policy Optimization)强化学习方法中，以指导对话策略的学习. Shah 等人[86]将用户具体的动作反馈信息整合到策略学习中，验证了使用动作反馈直接形成策略，相比于将动作反馈塑造奖励，可以使对话代理更快地学习新的交互. Su 等人[87]提出了两个步骤进行对话管理，先通过监督学习从对话数据中训练输出槽概率，然后使用策略梯度方法进行改进. 他们将策略梯度方法应用在监督学习之后是因为监督学习方法没有考虑选择的结果对未来对话的影响，且可能存在大量的对话状态不能生成合适的响应. Peng

等人[49]基于生成对抗网络的思想，提出了对抗 A2C 算法，从专家经验数据中训练一个判别器，然后将判别器整合进 A2C 算法中，用于鼓励对话代理产生类似于专家的动作. Cordier 等人[88]基于人类设计的次优专家演示数据来学习随机对话策略，专家演示数据用于代理模仿对话或者选择动作，实验验证演示数据可以提高训练开始时的性能. 表 2 给出了基于策略梯度的任务型对话策略方法的优势和不足.

表 2 基于策略梯度的任务型对话策略

		文献	优势	不足
基于策略梯度的对话策略	直接建模	文献[82-85]在模型的损失上或者对话轨迹数据上进行修改.	实施相对简单，易于实现.	难以保证鲁棒性，通常不能采样高效.
	联合建模	文献[37,49,86-88]联合的方式包括对抗学习、监督学习，或用户反馈等.	鲁棒性更强，性能较直接建模的模型更优.	其它模型的引入导致更加复杂的计算，模型融合时需要巧妙设计.

3.3 基于层次强化学习的对话策略

上述强化学习算法在值函数与采样效率上进行优化，对于更加复杂的任务，有价值的反馈信息更为稀少，这会增加对话代理探索的难度，从而导致策略性能的不稳定. 层次强化学习可以通过分解目标、动作空间或状态空间来缓解这种问题. 其中，在对话策略中最为常见的是进行目标分解，将对话要完成的任务看成多个子任务的组合，构建上层策略用于选择子任务，下层策略选择动作用于完成上层子任务. 上层策略接收外部奖励与对话状态，并完成子任务，下层策略则根据子任务（目标）学习动作，并接收内部评论家的奖励[89]. 图 5 展示了基于层次强化学习对话策略的一般框架.

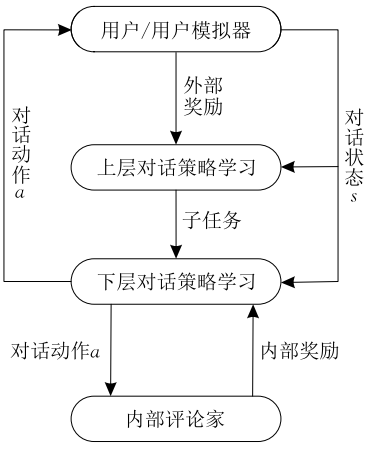


图 5 基于层次强化学习对话策略的一般框架

目标分解. Peng 等人[64]基于选项框架训练对话代理，他们构建了一个全局状态跟踪器用于确保交叉子任务之间的限制被满足. 但是，其用户目标需要根据人-人交互的对话数据，通过监督学习的方式提取. Wang 等人[90]使用选项框架同时建模对话策略和自然语言生成，上层策略用于建模对话动作，下层策略用于建模对话生成的单词，这种做法保证了动作和词的一致性. 他们还构建了一个辨别器在每一轮对话中评估生成的词，并根据评估结果给予内部奖励. Tiwari 等人[91]将层次强化学习应用于基于对话的疾病诊断领域，上层策略用于分配诊断的部门，下层策略用于采取动作，例如请求症状. 他们对疾病进行分类以缓解状态空间巨大导致的奖励稀疏问题. 考虑到不同用户的意图可能是一样的，Saha 等人[92]在用户意图上进行抽象，上层策略基于监督的方式决定用户的意图，下层策略用于提供动作，其用户意图相当于子目标. 基于子目标的层次对话策略学习通常需要一些专家数据来提取子目标，然后利用分层思想进行对话策略学习. 此外，Chen 等人[93]利用图卷积神经网络替换了传统层次深度强化学习中多层感知机部分，提出了基于图卷积的层次对话策略，上层策略接收状态之间构建的图和任务，下层策略用于输出动作. 实验验证他们提出的模型在对话策略上有更好的稳定性. Tang 等人[61]基于成功的对话轨迹挖掘用户子目标，构建了一个双层的循环神经网络用于发现用户的子目标，然后通过层次强化学习优化对话策略.

动作空间分解. Casanueva 等人[94]基于封建强化学习（Feudal RL）将策略分解成两部分，上层策略选择主要的动作子集，下层策略则从上层选择的动作子集中选取动作，使用领域内结构化的信息抽取状态空间，并在每一层使用不同的抽取状态进行决策. 但是他们使用外部奖励来学习下层策略，这会导致性能不稳定和较低的学习效率. Geishauser 等人[95]基于连续对话槽变化的概率计算信息增益，并将其作为内部奖励，缓解了上述问题. 动作空间分解通过压缩动作的搜索空间，对动作的预测更为高效. 但顶层动作子集的出错会导致错误的传播，从而使得无法学习到正确的动作.

基于层次强化学习对话策略使用子目标改进探索[53,96]，往往需要监督信息或者人为构建子任务或者动作子集. 此外，基于层次强化学习的任务型对话策略在状态空间抽取的研究尚不足. 关于非对话领域层次强化学习中的状态空间抽取方面的研究，读者可以参考文献[21,97]. 表 3 总结了基于层次强化学习的任务型对话策略的优势与不足.

表 3　基于层次强化学习的任务型对话策略

		文献	优势	不足
基于层次强化学习的对话策略	目标分解	文献[61,64,90-93]基于预定义的用户目标或根据专家经验挖掘用户目标，挖掘频繁访问的状态节点，将其视为子目标.	更好的稳定性，更高的采样效率，能够缓解奖励稀疏.	需要手动定义用户目标或需要专家数据.
	动作空间分解	文献[94-95]需要定义动作的子集.	压缩动作的搜索空间，对动作预测更为高效.	顶层动作子集的出错会导致错误传播而学不到正确的动作.

3.4　基于模型强化学习的对话策略

本节按照基于模型的强化学习对其构建的任务型对话策略进行分类，一类是直接规划动作用于策略提升，另一类则是生成模拟数据促进对话策略学习. 在任务型对话策略中，基于模型的强化学习创立一个世界模型，用世界模型模拟真实用户的经验数据，并根据模拟的经验数据促进策略学习，属于第二类方式，其流程图参见图 6. 其中，真实用户经验直接用于提升对话策略模型和训练世界模型，世界模型生成模拟经验用于提升对话策略模型.

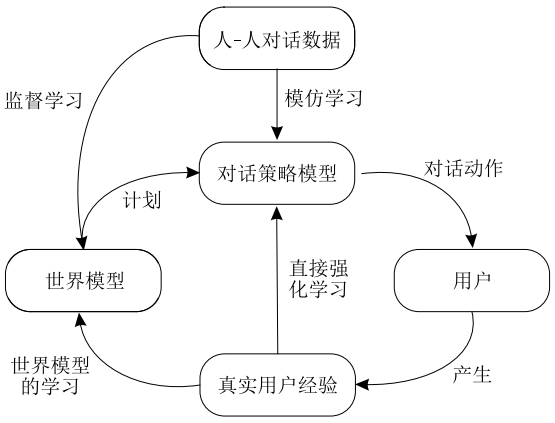


图 6　通过世界模型建模环境

基于世界模型模拟真实用户的经验数据建模对话策略. Peng 等人提出深度 Dyna-Q(Deep Dyna-Q，DDQ)[15]的方法，从真实对话数据和与用户交互的经验中训练一个基于深度神经网络的世界模型，并为代理提供模拟经验. 与仅使用真实经验的 DQN 相比，模拟经验提高代理可用样本总量. Huang 等人[98]基于获得的真实对话数据，使用监督学习的方式改进世界模型，并生成大量的模拟对话，然后让强化学习代理直接和世界模型交互，从而促进对话策略学习. 但是，DDQ 模型依赖高质量的模拟数据，低质量的模拟经验往往会严重影响其性能. Su 等人[99]提出了 D3Q(Discriminative Deep Dyna-Q，D3Q)算法，是一种有判别的 DDQ，他们基于循环神经网络构建了一个判别器，用来区分模拟经验和真实用户经验，只有被判别器检测过的高质量的模拟数据才会被用于训练对话策略. 但是，D3Q 没有兼顾不同训练阶段模型对模拟经验的不同需求，且 D3Q 对用户目标进行一致性采样，导致策略采样效率较低. Wu 等人[100]提出了 Switch-DDQ，加入了一个转换器来自动决定在对话训练的不同阶段是使用真实经验还是模拟经验，并在世界模型中增加了一个主动采样策略，在对话代理尚未充分探索的状态-动作空间中生成模拟经验. Zhang 等人[101]在 DDQ 中融入预算意识，提出了一种基于预算意识调度的 BCS-DDQ(Budget-Conscious Scheduling-Based Deep Dyna-Q)算法，构建了一个全局调度器，用于在不同的训练阶段分配预算，选择先前失败或未被探索的用户目标用来生成经验，并通过构建控制器来决定是收集人-人对话数据还是进行人-机交互获取数据，抑或与世界模型交互生成模拟经验. 在电影票预订的数据集上进行测试取得了较好的效果.

此外，考虑到基于 DDQ 的对话策略模型获取的模拟经验同真实用户交互中获取的经验是类似的，OPPA(OPPosite Agent Awareness)算法[78]将产生模拟经验的世界模型看成一个用户模拟器，用来预先估计用户的响应动作，从而指导策略模型进行下一个动作. OPPA 的算法属于规划动作迁移到场景中的范畴. Zhang 等人[102]构建了用户模型替换世界模型，并根据策略模型的输出动作和用户的真实经验生成模拟经验，然后将真实经验和模拟经验输入到策略模型中. 但是他们的奖励需要手动设计. 目前，基于模型的对话策略研究多集中于生成模拟样本或精炼模拟经验上，而在模型结构上进行优化的研究相对较少，将不同的强化学习算法融入到基于模型的算法中有望进一步提升对话策略学习性能. 表 4 总结基于模型的强化学习构建任务型对话策略的优势和不足.

表 4　基于模型的强化学习构建任务型对话策略

		文献	优势	不足
基于模型强化学习的对话策略	模拟数据	文献[15,98-102]基于 DDQ 模型，根据真实数据生成模拟数据，通过结合判别器精炼数据.	增加了可用数据，通常能够进一步提升性能.	依赖少量的真实用户交互的数据，性能依赖生成数据的质量
	规划动作	文献[78]直接预测动作，用于规划策略的动作选择.	直接规划动作，模型更加高效.	依赖少量的真实用户交互的数据，性能难以保证.

3.5 基于逆强化学习的对话策略

逆强化学习(Inverse Reinforcement Learning, IRL)算法是根据给定策略或者专家经验反向推导出 MDP 的奖励函数,避免了手动设置奖励的局限[103-104].基于 IRL 的对话策略的一般过程可参见图 7,其根据专家演示数据来构建奖励.但是,逆强化学习算法的计算成本较高,Fu 等人[105]提出了对抗逆强化学习(Adversarial IRL, AIRL),证明了 AIRL 能够在环境变化的情况下恢复奖励函数,具有更强的鲁棒性.Liu 等人[106]直接基于专家样本,使用 AIRL 来学习对话奖励,构建了一个生成器与环境进行交互,一个辨别器用于标注样本是成功的还是失败的,并将对话成功可能性的值作为奖励值.Takanobu 等人[107]基于 AIRL 提出了一个在多领域中进行奖励评估和策略优化的方法,对状态-动作对进行奖励评估,为每轮对话提供奖励信号,能够更好地指导对话策略学习.Hou 等人[108]将对话状态分解为三个独立的子状态,分别表示领域、动作和槽,然后通过判别器学习子状态-动作对是真实的还是生成的,判别器还用于对状态-动作对进行奖励.基于 IRL 的奖励塑造方法缓解了奖励延迟问题,但是专家经验在其中不可或缺,尽管有算法通过半监督的方式进行奖励评估,但仍然依赖少量专家数据[109].

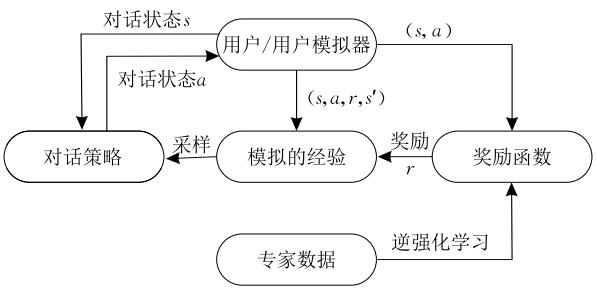


图 7 逆强化学习在对话策略中的学习奖励函数

此外,考虑到 IRL 在训练时需要在内部循环中进行强化学习,运行成本较高,这在一定程度上阻碍了基于 IRL 的对话奖励估计方法扩展到复杂的对话场景上.除 AIRL 外,相关的优化模型有限,因此,探索不同的、高效的 IRL 算法有助于其进一步应用于对话策略.

4 基于应用领域划分任务型对话策略

4.1 多领域的对话策略

对话策略从领域划分可分为单领域和多领域的研究.其中,单领域旨在解决一个领域内的对话任务,而多领域则主要解决多个不同领域的对话任务.多领域的对话策略较单领域对话策略具有更加稀疏的状态空间和规模更大的动作集合.任务型对话策略往往涉及多个领域.例如,用户的一个旅游任务会包括预订出租车(领域 1)和酒店(领域 2)等任务.然而,在多领域中,由于状态-动作维度较大,奖励信号变得更为稀疏.因此,目前多领域的对话策略研究主要集中在(1)领域融合;(2)奖励塑造;(3)基于 HRL 提升对话性能.图 8 展示了多领域中奖励塑造的基本思路,基于 HRL 提升多领域对话性能可参考图 5.

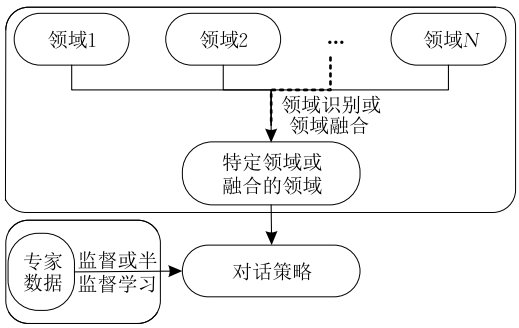


图 8 多领域的对话策略的基本思路

领域融合.多领域对话策略学习的一种思路是将其转化成多个单领域求解.例如,预订出租车和酒店的任务可以分别在领域识别后,按照单领域直接学习对话动作.Gašić 等人[110]提出的基于 BCM (Bayesian Committee Machines)的委员策略模型,每个委员在不同的数据集上进行训练,通过集合委员的决策得到多领域的决策.Cuayahuitl 等人[111]在两个领域的数据集上进行多个 DQNs 训练,对于不同的领域,选择不同的 DQNs 进行训练,多个 DQNs 之间可以相互转换以避免刚性设计结构;他们根据不同领域的特征选择领域,然后在选择的领域内选择具体的动作.由于他们考虑的领域比较少,这种根据特征选择领域的方式尚可,但当不同领域的交叉信息较多时,他们的方法难以适用.Mendez 等人[112]根据任务的索引来区分领域,在不同的领域中嵌入公共的动作和领域具体的动作,然后重新编码状态进行策略学习,他们验证动作的嵌入能帮助加速基于强化学习对话策略训练.此外,考虑到状态空间巨大且相互之间可能会纠缠,Peng 等人[113]构建了一个教师-学生模型,包含多个领域具体的教师和一个特定的学生.每个教师只关注一个特定的领域,并根据精确提取的单个领域对话状态表示,学习相应的领域知识和对话策略.然后,这些领域具体的教师将他们的领域知识和策略传授给特定的学生模型,使学生模型成为多领域对话专家.将多领域转化成单领域求解的思想缺乏考虑领域间的依赖关系,

例如，当用户需要订酒店和机票时，订酒店的时间往往会受到订机票时间的约束，而以上方法并不能考虑这种情况. Zhao 等人[114]构建了符号水平(Token-level)、对话轮数水平、领域水平和槽水平的关系图，然后通过图注意力网络挖掘它们之间的相关关系，但是他们提出的模型是一种端到端的对话策略. Cordier 等人[115]首先根据对话状态跟踪进行领域选择，然后在领域内确定依赖和独立的槽用于构建图，利用基于图卷积网络的消息传递预测动作，并基于专家轨迹通过行为克隆指导对话代理的探索. 之后，他们基于图卷积网络构建了少样本下提高多领域对话策略采样效率的算法[116]. 但是，基于专家轨迹的行为克隆可能会出现错误，因为在当前状态下代理可能会误解用户动作. Rohmatillah 等人[117]提出了一个混淆模型，根据当前对话状态预测动作的同时，还预测当前的置信度状态和用户最近的对话，而预测的结果可以被看作是当前轮对话的特征，从而促进行为克隆. 此外，Wu 等人[118]混淆多个领域的数据集进行训练，依然缺乏考虑领域之间的相关性，并且是基于 Seq2Seq 的模型. Qin 等人[119]构建了一个动态融合网络用于分析目标领域与其它领域之间的相关性，也是一种基于 Seq2Seq 的模型，并非基于强化学习算法.

奖励塑造. 不同领域的融合加剧了基于强化学习对话策略的奖励稀疏问题，因此，一些研究致力于多领域中的奖励塑造，基于专家数据进行监督或半监督的学习，也就是直接学习状态-动作对之间的关系，通过相似的状态选择对应的动作. 参见图 8 下半部分. Huang 等人[109]同时基于有标注和无标注的专家演示数据，通过预测对话进程是否与专家演示一致来计算奖励，这是一种半监督的学习方式，其无标注的数据通过半监督的变分自编码器预测状态转移之间的动作，得到状态-动作对后进行奖励塑造. Jeon 等人[120]提出了一种基于监督学习和强化学习混合的对话策略学习方法，监督学习部分通过预训练语言模型 Bert[121]学习用户当前对话轮的输入，获得领域状态、用户的信念和系统动作，并计算对话成功率. 得到领域状态与用户信念状态作为强化学习的输入，而系统动作和对话成功率用于奖励塑造. Takanobu 等人[107]基于逆强化学习构建了一个奖励评估器，和传统逆强化学习根据监督数据先训练奖励函数，再将奖励函数融入到策略中的不同方法，他们在方法中集成了对抗学习，使得策略学习和奖励评估同步. Wang 等人[122]为了防止奖励塑造陷入局部最优，使用 Bert 作为辨别器，来判断当前系统动作的优劣，辨别器能够给予一个额外的局部密集奖励来指导对话代理高效地探索. 上述方法在状态和对应的动作上进行奖励塑造，而多领域的对话策略中状态的组成往往包括槽、动作和领域. 因此，Hou 等人[108]分别在槽、动作和领域上构建奖励，只有对话属于正确的领域时，对话代理才会从给予的动作中获得奖励；只有当代理采取正确的动作时，才会获得槽水平上的奖励. 结果验证了他们的奖励机制能够更准确地进行奖励评估，并显著提高基于强化学习对话代理的性能，加快训练的收敛速度. 多领域中的奖励塑造通过监督或半监督的方式进行，其核心在于构建奖励评估器用于评估奖励，最后将得到的奖励应用于对话策略[107].

基于 HRL 提升对话性能. 考虑到多领域对话策略较单领域更为复杂，一些研究通过分层的思想解决多领域任务型对话策略的问题，并基于 HRL 的方法研究多领域对话策略[64,94]. 这些方法在每个领域中只考虑一个意图，而用户的目标可能并不局限于每个领域中的单个意图，有可能会包含领域中的多个意图. Saha 等人[123]构建了一个三层的对话策略学习网络，顶层策略基于用户的请求根据当前状态选择具体的领域，中间层策略基于顶层策略选择的领域选择意图，它在不同的时间步骤内完成领域的各个子任务，并根据用户的查询判断该域所有子任务是否已完成，而下层策略则基于中间层策略输出动作. Rohmatillah 等人[124]提出了一种计算高效的多领域对话策略，他们先基于模仿学习得到训练的权重，然后将得到的权重作为层次强化学习的下层策略权重，并且在层次强化学习的训练过程引入了基于规则的方式以指导对话代理. 多领域的层次对话管理的研究报告可参见文献[125]. 基于 HRL 的多领域对话代理简化了对话中的问题设定，能够有效地学习来自人类和环境的反馈. 表 5 展示了不同类别多领域对话策略的优势和不足.

表 5　多领域的对话策略

		文献	优势	不足
多领域的对话策略	领域融合	文献[110-119]对不同领域进行识别后再分别处理或者直接通过 Seq2Seq 方式进行领域融合.	多领域对话策略的主流方式，思想简单，实施方便.	需要预先定义好领域或进行领域识别，领域较多时直接融合效果可能会变差.
	奖励塑造	文献[107-109,120,122]通过监督或半监督的方式获取奖励信息.	能够解决奖励延迟问题，不仅仅局限于多领域，能够加快训练速度.	需要基于监督信息，通过奖励评估器评估，需要一定的领域知识.
	基于 HRL	文献[64,94,124-125]根据领域或意图进行分层.	简化了领域使得领域融合更为具体化.	需要预定义好领域或进行领域识别.

4.2 多模态的对话策略

传统的对话策略的对话状态是对话文本信息的编码，而在真实场景中，用户的交互并不限于文本内容，如用户的音频会隐藏着用户的音高、情感等信息；用户交互的图像（如表情包）会暗示用户的情绪等信息. 更多地融合用户信息有助于提高对话策略的性能. 多模态的对话策略是考虑多种模态信息，包括文本信息、音频信息或图像信息等进行策略学习的一种方式. 它在输入上融合了更多的特征信息，通常较单模态能够提供更为准确的动作响应. 图 9 展示了多模态对话策略的一般过程. 多模态对话策略需要考虑至少两种模态输入.

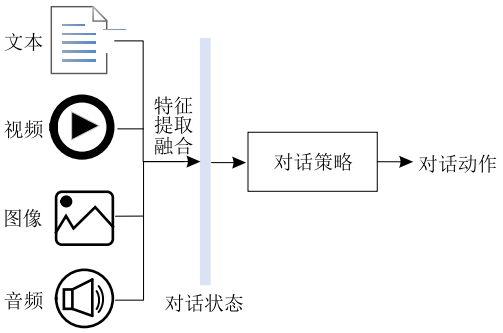


图 9 多模态的对话策略一般流程

在新近的研究中，Liao 等人[126]提出了一种知识感知的多模态框架，该框架使用层次循环编码器-解码器将可视化图转化为与之对应的文本. 代理不仅根据对话历史，而且提取与当前上下文相关的知识来调整响应动作. Saha 等人[127]在订餐领域考虑了文本和视觉信息，通过实验结果验证了多模态信息能够提高任务型对话的任务完成率和效率. Tiwari 等人[128]在纳入文本和视觉信息的同时，结合用户的动态目标进行策略学习. 在疾病诊断上，一个虚拟的诊断助理被提出[129]，它同时兼顾用户对话内容和患者的诊断图像进行人机对话，策略部分的状态信息包括患者自我的报告信息、患者提供的图像信息、文本或视觉的症状状态、对话轮数、代理之前的动作和奖励. 针对语音对话系统，传统的对话策略是将语音信号转化为文本信息，再通过自然语言理解、对话状态跟踪为对话策略提供输入. Zorrilla 等人[130]直接编码音频信息，验证了音频的嵌入可以促进对话策略学习. Zhang 等人[131]提出动态整合视觉和语言的多模态层次强化学习框架，该框架混合学习多模态对话状态表示，并进行层次对话策略学习. 尽管多模态的对话策略作为一种融合多种信息的新的对话策略学习，但是该类研究还相对较少. 对于高度拟人化的对话系统，考虑多模态信息是必不可少的. 目前，随着人机对话中的不同设备接入，使得多模态信息更为重要，因此研究多模态的对话策略有望在未来进一步加强. 表 6 总结了不同的输入模态信息进行对话策略学习的方法.

表 6 多模态的对话策略

		文献	优势	不足
多模态的对话策略	图像＋文本	文献[126-129,131]融入图片信息到对话状态.	较单模态状态信息嵌入，提高任务完成率.	状态联合编码时并不容易.
	音频＋文本	文献[130]融入音频信息到对话状态.	较单模态状态信息嵌入，提高任务完成率.	状态联合编码时并不容易，目前研究较少.

4.3 多代理的对话策略

基于强化学习的对话策略建立一个 MDP 来优化策略 $\pi(a|s)$，使得在任意对话状态 s 下，策略 π 都能给予一个恰当的动作 a 输出. 而多代理的对话策略则同时构建多个代理，这些代理共享一个环境，有各自的输入状态集合和输出动作集合，每一个代理都将最大化其累计折扣奖励[18]. 例如，在一个多领域的对话中，每一个领域可以对应一个代理，各代理接收该领域下的对话状态作为输入，并输出领域限定的动作，最终策略优化来自不同领域代理累计奖励之和最大化. 但由于两个或多个代理同时学习，每个代理都随着训练的进行而不断变化，因此多代理的对话策略环境不再是固定的. 多代理的强化学习算法可以按照代理之间的关系分成四类，包括(1)完全协作的、(2)完全竞争的、(3)既有协作又有竞争的和(4)既不协作也不竞争的[132]. 然而，目前对话策略的研究中仅仅考虑了完全协作型的多代理模型. 多代理的对话策略基本流程参见图 10，其中“-----〉”表示代理之间可直接进行交互，通过一个混淆网络来处理不同代理之间的协作和竞争关系.

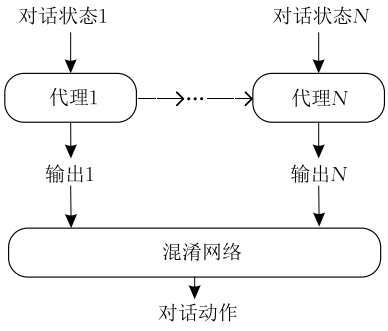


图 10 多代理对话策略的一般流程

完全协作. 多代理对话策略通过促进协作，能够提升不同代理的学习能力[133]. 目前，对话策略中

多代理的研究通过完全协作的方式进行. Liu 等人[134]首次将多代理的强化学习应用到任务型对话策略中，同时构建了系统代理和用户代理，用户代理被赋予一个要完成的目标，输出演示一致的用户动作. 系统代理评估用户的目标，并通过有意义的对话来完成用户的请求. 其系统代理和用户代理都通过策略梯度法进行迭代训练. 之后 Takanobu 等人[135]在多领域的数据集上做了和 Liu 等人[134]相似的工作. 这种建立用户代理的方案相比于传统对话策略使用模拟器的做法，能够解决模拟偏差的问题. Zhuang 等人[136]建立了一个视觉对话系统，在用户代理和系统代理交互的过程中融入了多模态信息，并构建了一个一致性方案用于处理代理交互之间的多模态信息，使得多模态信息的融合更为精确. 为了解决传统对话策略中采样效率不高和多领域中策略迁移不足的问题，Chen 等人[5]提出了一个代理图模型，他们构建了槽依赖和槽独立的有向图，每一个代理对应图中的一个节点，最后将各代理的输出连接在一起用于选择最终的动作. Wang 等人[137]在多领域上构建了协作的多代理模型，通过三个代理分别输出领域、槽类型和槽名称，同时多个代理之间进行横向交互，领域的输出作用于槽类型，槽类型的输出作用于槽名称，最终对话动作的输出来自整合当前状态、领域、槽类型和槽名称等信息. Papangelis 等人[138]在餐厅领域训练两个代理，一个用于提供信息，称为提供者，另一个用于寻找信息，称为寻找者. 每一个代理接收不同的输入状态，寻找者的状态包括用户目标和提供者提供的信息. 提供者的状态则包括寻找者提供的约束、请求信息和当前对话信息以及数据库查询结果. 实验验证了他们的模型优于基于监督学习的基线算法. Tiwari 等人[139]提出了一种用于说服领域的对话策略模型，其包含了两个代理，代理 1 负责槽的插值填充和动态目标设置，代理 2 在目标不可达的情况下通过提取其它目标来协助用户完成对话. 他们的模型提高了用户满意度和代理的效能.

尽管多代理的对话策略带来了很多优势，但对话策略中如何巧妙地设计代理之间的协作、竞争关系尚需要进一步挖掘. 据我们了解，尚未在基于强化学习的多代理对话策略中发现考虑代理之间竞争关系的研究. 此外，关于多代理强化学习的安全性问题，每个代理不仅要考虑自己的安全约束(比如代理之间可能存在奖励最大化的冲突)，还要考虑其它代理的安全约束，以保证它们的联合动作是安全的. 而每个代理在局部安全仍然不代表整体的安全[140]. 但很少有解决方案为多代理的安全性问题提供有效的算法模型.

4.4　共情对话策略

共情(Empathy)指的是能够理解、意识和间接地体验另一个人的过去或现在的感受、思想和经历. 它具有增强人们情感纽带的能力，在人类的交流中起着重要的作用. 研究表明在对话系统的设计中融入共情对提高人机交互中的用户体验至关重要[141]. 而且，具备共情能力的对话代理能够被用户感知到，从而改善用户参与感，使得沟通更为有效. 例如，用户抱怨说“我想要今天下午到北京的机票，你怎么就听不懂呢?”. 此时如果代理不能理解用户情绪，则会极大地降低用户体验. 本文将共情的对话策略分为两类，一类是理解用户情感并给予积极响应，以提升用户满意度. 另一类则为理解用户多意图以增强用户的“共鸣”感，旨在为用户提供“知其所想”的个性化响应.

理解用户情感并给予积极响应. 现有的研究将情感作为一种监督信号用于缓解奖励延迟和促进用户的满意度[142]. 在任务型对话策略中考虑用户的情感信息，有助于构建更加有效的策略模型，提高对话代理的响应能力[143]. Song 等人[144]为线上购物助手 AliMe 增加了情感，通过情感分类和回复选择为用户提供情感方面的问题解答，增强了用户的体验. Bui 等人[145]基于部分观测的马尔可夫决策过程构建了情感对话系统，将用户情感作为对话策略观测的一部分，并直接和用户的真实状态进行融合，尽管他们没有通过用户情感反馈来改变模型的动作选择，但是验证了融入情感的对话策略在性能上会超过普通的贪婪动作选择策略. Shi 等人[146]提出了情感自适应的端到端对话系统，通过语音文本等多模态信息识别用户情感，然后把用户情感特征融入到奖励函数中，尽管减少了对话长度并且提升了任务完成率，但由于情感的吸收态不易被定义，他们的做法会使得奖励函数的性能不稳定. Li 等人[147]提出了一种双向情感循环分类器进行对话情感分析，以上下文和当前的对话作为输入，能够建模任意轮的对话语义的情感特征，他们的模型可以整合进对话策略. Zhang 等人[148]考虑到与用户情感相关的线索，基于 DDQ 框架建模对话策略，引入了用户的情感，将用户情感信息作为监督信号，用于塑造奖励. Tu 等人[149]考虑用户细粒度的情感状态，探索用户的不同情感强度，同时根据不同的情感强度生成对

应的回复,从而降低用户困扰.但他们的研究只在开放域对话系统上进行.

理解用户多意图以增强用户的"共鸣"感.结合情感进行研究往往将情感作为监督信号,或直接并入到对话状态来提升任务完成率和用户满意度.然而,这些研究在整个对话过程中只使用单一的用户意图,也就是他们只预测用户的一个意图,并给予单一动作响应.而现实生活中多意图更为普遍.例如,在购物中,用户的单轮对话可能表现出需要一个价格低且体积小的商品(价格低和体积小被视为两种不同的意图),仅仅考虑用户的单意图会导致更多的对话轮数.Saha 等人[150]为对话动作的情感辅助分类开发了一个多任务框架.然而,他们的工作并没有讨论管道方法的对话策略学习框架.尽管 Saha 的团队在最新的研究中增加多意图,但是他们的多意图是根据不同领域来定义的,没有对用户潜在意图进行推理,只能挖掘已有意图[127].传统的对话策略识别用户单一意图,通过 One-hot 进行编码,如果一个新用户引入了新意图,则需要重新进行对话策略训练.这增加了模型的维护成本并且降低了可扩展性.Wang 等人[151]提出基于教师-学生模型解决该问题.在他们的方法中,旧模型和新用户意图的逻辑规则被视作教师,新模型被视为学生.对于过往的意图集合,旧模型直接指导新模型的训练.对于新意图,其逻辑规则被看成新的标记数据用来训练新模型.通过这种方式,新模型不再需要与环境进行交互从而不需要重新训练.此外,Chen 等人[152]在不需要标签数据和模型再训练的情况下辨别扩展的意图,他们直接学习意图的嵌入编码,将新意图嵌入到一个更高维的语义空间.因此,他们的方法可以直接根据新意图的描述生成相应的内容嵌入向量,然后识别意图.此外,一些研究在任务型对话系统中整合用户的配置信息实现端到端的对话系统[153-154],尽管该类研究具备共情能力,但是由于并没将对话策略单独研究,因此,本文不再赘述.

由于用户在对话过程中可能会切换意图,从而对后续的对话产生较大影响,因此识别和跟踪对话者的意图对多轮对话至关重要.此外,在现实中,用户通常会有一些特定的意图关联.如果代理能够对用户的潜在意图做出合理的"猜测",并在用户对话之前提供有用的信息,就可以减少对话重复,提高对话精度,也能够增强用户的"共鸣"感.Saha 等人[123]根据状态提取出不同的用户意图,训练多个网络,每个网络服务于一个特定的意图,用于完成一个特定的子对话.Bihani 等人[155]使用一个小规模的单一意图语料数据库来生成多意图语料,并解决多意图自然语言中的模糊性,但他们只适用于多意图分类.Shi 等人[156]使用层次循环编码-解码器进行意图的识别和预测,但他们并没有结合相似用户的意图转换来挖掘用户的潜在意图,也没有考虑情感因素对用户多意图的影响.表 7 总结了共情对话策略不同类别研究的优势和不足.

表 7 共情对话策略

		文献	优势	不足
共情对话策略	理解用户情感	文献[144-148]将用户情感信息融入到策略学习	能够解决奖励延迟问题,提升用户的满意度	需要包含情感信息的监督数据,或者根据对话状态识别得到情感信息.
	理解用户多意图	文献[127,150-152]考虑用户的意图信息进行策略学习	能够识别用户的潜在意图,挖掘用户的多意图	意图识别可能会引入其它误差,多意图研究不足.

5 对话策略的相关研究

5.1 用户模拟器

任务型对话策略通常可被视为监督学习或强化学习任务.在基于监督学习的方法中,训练策略模型来模拟专家的行为,这种方法通常需要领域专家标记大量数据进行训练.对于特定的任务,需要昂贵且耗时的数据收集和人工标注.此外,基于监督学习的方法缺乏对未知对话状态空间的探索能力,限制了对话策略寻找最优动作的能力.在基于强化学习的方法中,对话代理可以根据环境的奖励信号改进对话策略学习,但基于强化学习的方法需要大量的人机交互样本来进行模型优化,这会使得对话的成本过高且不易实施,特别是在对话策略模型样本不足的情况下.为了克服这个问题,研究人员使用用户模拟器来训练基于强化学习的对话代理.用户模拟的目的是生成自然合理的拟人化对话,使基于强化学习的对话代理能够高效探索对话轨迹并从中学习.图 11 展示了对话策略和用户模拟器交互的一般过

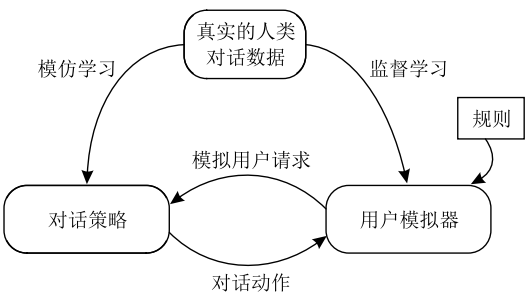


图 11 基于用户模拟器的对话策略学习

程.用户模拟器基于真实的人类对话数据进行监督学习,或根据人为制定的规则进行学习.其接收对话策略模型的对话动作,模仿真实用户对话,并将用户请求传递到对话策略,用于策略学习.

目前用户模拟器可以分为两类,包括基于议程的用户模拟器和基于模型的用户模拟器.基于议程的用户模拟器根据人为制定的规则构建议程.例如,根据堆栈来建模用户目标和更新状态,当用户目标完成时出栈,并随机选择一个新的用户目标,直到栈为空.基于模型的用户模拟器则根据特定模型构建一个模拟器,不需要人为地制定规则.基于议程的用户模拟器包括:Li 等人[157]在电影预订领域构建的用户模拟器,其根据是否预订了电影,以及电影是否满足用户的约束构建议程.Schatzmann 等人[158]提出可以应用于部分可观测马尔可夫决策过程中的用户模拟器,通过对议程的优先级排序来选择下一用户动作.Jain 等人[159]提出了一个社交用户模拟器,该用户模拟器利用社交线索,如利用用户和系统的对话策略和用户对话目标来估计交互过程中的融洽程度,并产生适当的任务和社交行为.

基于模型的用户模拟器包括:Liu 等人[160]基于预训练语言模型 GTP-2[161]建立了一个端到端的用户模拟器,优势是仅需要较少的对话历史.Kreyssig 等人[162]提出了神经用户模拟器(Neural User Simulator),从语料库中学习用户行为,然后直接生成自然语言,而不是对话动作等语义内容.

用户模拟器的评估的方式包括直接评估、间接评估和人类评估.直接评估中,通过自然语言生成直接进行评估,根据生成的词汇量或平均对话长度衡量语言多样性;间接评估中,基于强化学习对话策略的性能作为评估指标(对话策略的一般评估方式参见 5.2 节).人类评估的指标包括评估语言的流畅性、连贯性、生成的语句和用户目标的连接性、生成语言的多样性等.用户模拟器的最终目标是构建一个面向任务的强化学习系统,用来替换真实用户[163].因此,通过人类评估是一个理想的状态.

尽管现有的用户模拟器能够模拟真实用户,从而和对话代理进行交互,降低了对话策略的成本.但是,用户模拟器和真实用户之间是有差异的,这可能会导致次优对话策略结果.Dhingra 等人[164]证明了模拟器训练的对话代理在和真实用户进行评估时存在显著差异.特别是用户模拟器在大多数情况下都扮演着理想的用户角色,假定了用户模拟器会始终耐心地响应来自策略模型的所有请求.因此,应客观地看待用户模拟器的使用.

5.2　对话策略的评估

对话策略的评估不同于自然语言理解和自然语言生成的评估方式,对话策略的评估更多地关注对话的成功率和平均对话轮数、累计奖励等[165].主要包括两类评估方式,分别是自动的评估和人类的评估.由于用户模拟器的性能不一,在实验中两种类型的评估往往都需要涉及.其中,自动的评估指标有:

(1)对话的成功率(Dialog Success Rate).是在设定的有限对话轮数下,判断用户所有请求是否满足.

(2)对话的奖励值(Dialog Reward).是在设定的有限对话轮数下,对话轨迹累计奖励的平均值.

(3)平均对话轮数(Dialog Turns).是在设定的有限对话轮数下,对话成功的平均对话轮数.

(4)预订率(Book Rate).是用于评估对话代理是否预订了符合用户的所有约束条件.

(5)用户满意度(User Satisfaction).通常是在共情对话策略中的一个重要指标,往往以每轮对话中的用户情感变化的得分作为用户满意度.用户情感越积极则分值越大,代表用户越满意.

通常情况下,对话策略的对话成功率和奖励值越大、对话轮数越小说明当前对话策略模型越好.人类的评估指标有:

(1)对话成功/失败(Dialog Success/Failure).该指标通过真实用户判断任务目标是否完成,统计成功和失败的对话数量计算成功率/失败率.

(2)响应的恰当性(Response Appropriateness Score).用于衡量对话回复的恰当性,需要结合对话生成一起进行.该指标可人为设置 5 分制度量(整数),得分为 5 的响应在对话中表示非常恰当,而得分为 1 的则表示完全不恰当或离题.

(3)用户满意度(User Satisfaction).通过真实用户的主观体验进行满意度评估,设定不同的分值,当对话代理的回复非常符合用户需求时评分最高,反之则评分最低[165].

5.3　平台和数据集

近年来任务型对话系统受到越来越多的关注,促进了该领域的研究者开发一些对话平台.这些平台的出现大大简化了对话策略的搭建和测试环节.如 Pydial[166]是一个关注于强化学习对话策略的平台,该平台提供了一些常见的算法,如 DQN、A2C 和朴素 Actor-Critic 算法.目前,该平台已经更新到 Pydial3.0,但是,该平台对自然语言理解和自然语言生成关注较少.ParlAI[167]是 Facebook 旗下的一个"一站式对话研究"的平台.目前,该平台囊括了 25 种数据集,包括问答、开放域和任务型对话数据

集.支持的对话任务包括阅读理解、问答等.在该平台上可以做各种变换,但是需要用户自定义对话系统的各个模块.Rasa[168]和 Plato[169]是一种用于生产的对话平台,方便开发者用简单的代码快速搭建一个对话系统.但该平台并没有提供最新的数据集与最新的算法模型.CRSLab[170]是一个用于构建对话推荐系统的平台,包含了对话模型、策略模型和推荐模型等;基于 PyTorch 实现,提供的对话策略模型主要基于预训练语言模型而非基于管道方法的模型.LEGOEval[171]是一个对话的评估系统,允许研究人员轻松地为亚马逊土耳其机器人(Amazon Mechanical Turk,AMT)上的对话开发评估任务.ConvLab[172]是一个支持端到端的多领域对话平台,由清华大学团队提出,支持模块上而非集成后的评估,如对话状态跟踪评估,对话策略的评估等.此后,该团队提出了 ConvLab-2[173].ConvLab-2 优化了框架的易用性和扩展性,为对话系统的各个模块提供了最新的模型,并且支持更多的数据集,支持端到端和管道方法的对话系统.目前,ConvLab-3[174]已经发布,相比于之前的版本具有更多的数据集,且统一了数据格式.TaskMAD[175]是一个任务导向的多模态对话平台,以收集多模态信息为重点,支持消息推送、接收图像和视频数据,其计划在未来支持发布音频.表 8 给出了一些目前常用的对话平台的信息.

表 8 对话策略中一些常用的对话平台

对话平台	简介	编成语言	链接
Pydial[166]	集成了一些强化学习算法,包括 DQN、AC、A2C 等.	Python tensorflow	论文:https://aclanthology.org/P17-4013/ 平台地址:http://pydial.org
ParlAI[167]	包括问答、开放域、任务型对话的数据集,需要用户自定义对话系统的各模块.	Python	论文:https://aclanthology.org/D17-2014/ 平台地址:http://parl.ai
Rasa[168]	支持自然语言理解和对话管理,方便非专业人士快速搭建一个对话系统.	Python	论文:https://arxiv.org/abs/1712.05181 平台地址:https://github.com/RasaHQ/rasa
Plato[169]	支持在线和离线的交互,方便开发者快速搭建对话系统.	Python	论文:https://arxiv.org/abs/2001.06463 平台地址:https://github.com/uber-research/plato-research-dialogue-system
CRSLab[170]	用于对话推荐系统,包含了策略模型.	Python PyTorch	论文:https://aclanthology.org/2021.acl-demo.22/ 平台地址:https://github.com/RUCAIBox/CRSLab
LEGOEval[171]	对话评估系统,用于评估 AMT 交互的对话	Python Flask	论文:https://aclanthology.org/2021.acl-demo.38/ 平台地址:https://github.com/yooli23/LEGOEval
ConvLab-3[174]	包含多领域数据集的一个对话平台,支持策略评估,包括 PPO、DDPG 等算法.	Python PyTorch	论文:https://arxiv.org/pdf/2211.17148.pdf 平台地址:https://github.com/ConvLab/ConvLab-3/
TaskMAD[175]	多模态的对话平台,支持图像、视频的接收和推送.	JavaScript Node.js	论文:ttps://dl.acm.org/doi/10.1145/3477495.3531679 平台地址:https://github.com/grill-lab/TaskMAD

研究任务型对话策略离不开对应的数据集,任务型对话数据集的收集通常有三种形式,一是通过招募志愿者进行人机对话,获取数据集,称之为人-机对话数据.另外一种是通过现实世界中的真实任务收集数据,如预订电影和预订机票等,这类数据集称之为人-人对话数据.还有一种是直接基于机器-机器交互生成对话数据.目前任务型对话系统中一些常见的数据集包括 CrossWOZ[176]、MultiWOZ[177]等.CrossWOZ 是一个多领域的中文数据集,采集来自于真实的人-人对话.MultiWOZ 在对话策略领域常被用到,它包含了多个领域,目前已经更新到 MultiWOZ 2.4.表 9 给出了一些不同应用类别的任务型对话数据集.本文对其进行介绍,并给出了数据集的链接.

本文对一些基于强化学习的经典算法进行汇总,展示它们在不同数据集上测试的性能,以供读者进行整体性参考.表 10 给出模型和其在不同数据集上的性能,用黑体表示该模型在特定数据集上取得了先进效果.从中可以看出,在 Microsoft Dialogue Challenge movie 数据集上,MCTS-DDU[79]取得了最好的效果,这是由于其使用了蒙特卡罗树搜索去不断地模拟有希望的轨迹.Adversarial A2C[49]算法在该数据集上取得了仅次于 MCTS-DDU 算法的效果.在 Microsoft Dialogue Challenge Taxi 数据集上,DPAV[32]取得了最好的效果,其构建的强化学习代理能够更好地学习动作值函数.VACL[77]在 Microsoft Dialogue Challenge Restaurant 数据集上的表现较好,但在其它数据集上,VACL 并不能较 DPAV 更优,显示出 VACL 鲁棒性较弱.HRLagent[64]模型在 Frames 数据集上的对话成功率最高,但其对话轮数和平均对话奖励较 ES-DDQ[148]模型差,这两个模型在最大对话轮数上的设定并不一致.在 MultiWoz 2.0 的数据集上,DQN(GAN-VAE)[37]取得了最高的对话成功率 0.985,其平均对话轮数保持在

较高水平(11.04)；MADPL[135]的整体性能更好，对话成功率为 0.921，平均对话轮数为 7.62. 在 MultiWOZ 2.1 数据集上，HRLG[124]将人类参与到对话代理的循环训练中，效果最优. 在 DSTC2 数据集上，2016 年提出的 BatchDA2C[82]模型能够取得先进的效果，但其仅测试了平均对话轮数和平均对话奖励. 目前，对话策略研究的常用数据集是 MultiWOZ 和 Microsoft Dialogue Challenge. 不同的实验环境对各模型的性能会有一定的影响，对话策略的研究人员在构建对比实验时，应注意在相同环境下实施.

表 9　任务型对话数据集

名称	发布时间	介绍	领域	数据集大小	链接
CrossWOZ[176]	2020 年	大规模多领域的中文任务型对话数据集，涉及的领域包括酒店和出租车等；人-人对话数据集.	多领域，包括：attraction、restaurant、hotel、metro、taxi	对话规模：5012 对话轮数：84 692 平均对话轮数：16.9 语言：中文 槽个数：72 领域个数：5	论文： https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00314/96453/CrossWOZ-A-Large-Scale-Chinese-Cross-Domain-Task 数据集： https://github.com/thu-coai/CrossWOZ
MultiWOZ[177]	2018 年	包含了酒店、餐厅、出租车、火车等多个领域的对话数据集；最早版本为 MultiWOZ 2.0，为 2018 年发布. 目前，最新版本为 MultiWOZ 2.4(2022 年发布)，增加了对话状态评估；人-人对话数据集.	多领域，包括：Attraction、Hospital、Police、Hotel、Restaurant、Taxi、Train	对话规模：8438 对话轮数：115 424 平均对话轮数：13.68 语言：英文 槽个数：25 领域个数：7	论文： https://aclanthology.org/2022.sigdial-1.34/ 数据集： https://github.com/budzianowski/multiwoz (MultiWOZ 2.0) https://github.com/smartyfh/MultiWOZ-2.4 (MultiWOZ 2.4)
Microsoft Dialogue Challenge[178]	2018 年	包含三个领域的数据集，分别是电影票预订、餐厅预订和出租车预订.	三个单领域，包括：Movie、Restaurant、Taxi	Movie：对话规模：2890；意图：11；槽：29 Restaurant：对话规模：4103； 意图：11；槽：30 出租车：对话规模：3094； 意图：11；槽：29 语言：英文	论文： https://arxiv.org/pdf/1807.11125.pdf 数据集： https://github.com/xiul-msr/e2e_dialog_challenge
Frames[179]	2017 年	使用 Wizard-of-Oz 方法收集的一个多任务的数据集，包含航班和旅馆预订的任务型；人-人对话数据.	多领域，包含机票预订和旅馆预订	对话规模：1369 对话轮数：19 986 平均对话轮数：14.60 语言：英文 槽个数：61 领域个数：2	论文： https://aclanthology.org/W17-5526/ 数据集： http://datasets.maluuba.com/Frames
LEGO[180]	2012 年	该语料库为训练和测试提供数据，由 Let's Go 公交信息系统的；人-人对话数据.	单领域，公交信息	对话规模：200 对话轮数：4885	论文： https://aclanthology.org/L12-1157/ 数据集： https://www.ultes.eu/ressources/lego-spoken-dialogue-corpus/
DSTC2[181]	2014 年	对话系统技术挑战赛数据集，DSTC2 是任务型对话数据，通过使用亚马逊土耳其机器人进行收集. 餐厅领域的人-机对话数据.	单领域，领域为：restaurant	对话规模：1612 对话轮数：23 354 平均对话轮数：14.5 语言：英文 槽个数：8 领域个数：1	论文： https://aclanthology.org/W14-4337.pdf 数据集： http://camdial.org/~mh521/dstc/或 https://github.com/matthen/dstc
SGD[182]	2020 年	包含两个数据集，Reddit 和 MetaLWOz，其中 MetaLWOz 为任务型对话数据集，人-机对话数据.	多领域，包括：Banks、Buses、Calendar、Events、Flights、Homes、Hotels、Media、Movies、Music、RentalCars、Travel、Restaurants、Weather、RideSharing、Services	对话规模：16 142 对话轮数：329 964 平均对话轮数：20.4 语言：英文 槽个数：214 领域个数：16	论文： https://arxiv.org/pdf/2002.01359.pdf 数据集： https://github.com/google-research-datasets/dstc8-schema-guided-dialogue
Medical DS[183]	2018 年	包含上呼吸道感染、小儿功能性消化不良、小儿腹泻及小儿支气管炎的疾病诊断数据集；人-人对话数据集.	单领域，疾病诊断	疾病种类：4 类 症状：67 种 平均明确的症状：上呼吸道感染(2.15)，儿童功能性消化不良(1.7)，小儿腹泻(2.56)、小儿支气管炎(2.87)	论文： http://www.sdspeople.fudan.edu.cn/zywei/paper/liu-acl2018.pdf https://aclanthology.org/P18-2033.pdf 数据集： https://github.com/fantasySE/Dialogue-System-for-Automatic-Diagnosis

（续 表）

名称	发布时间	介绍	领域	数据集大小	链接
EmoWOZ[184]	2022 年	面向任务的对话的大规模手工情感标注的语料库.基于MultiWOZ多领域任务导向数据集;人-人对话数据集.	多领域,领域和MultiWOZ相同	对话规模:11434 对话轮数:167234 平均对话轮数:14.63 语言:英文 槽个数:214 领域个数:16 情感类型:7种,包括:中立、害怕、不满、内疚、辱骂、激动、满意	论文: https://paperswithcode.com/paper/emowoz-a-large-scale-corpus-and-labelling 数据集: https://zenodo.org/record/6506504#.ZBq8vnZByUk
JDDC[185]	2020 年	基于京东电子商务网站收集的售后服务中文对话语料;人-人对话数据.	单领域,电子商务	对话规模:1024196 平均对话轮数:20 最小对话轮数:2 最大对话轮数:83 意图:289 语言:中文	论文: https://arxiv.org/pdf/1911.09969.pdf 数据集: http://jddc.jd.com/auth_environment
PbAbI[186]	2017 年	整合了用户配置信息的任务型对话数据集,可以做用户个性化的对话策略研究;根据用户的配置信息进行个性化的餐厅预订(饮食偏好、最喜欢的食物等).	单领域,餐厅预订	对话任务个数:5 对话规模:24000(任务1、任务2、任务4),48000(任务3、任务5) 个性化特征:性别,年龄段 语言:英语 用户配置信息数量:6(任务1、任务2、任务4),180(任务3、任务5)	论文: https://arxiv.org/pdf/1706.07503.pdf 数据集: https://github.com/chaitjo/personalized-dialog
Vis-SentiVA[127]	2020 年	用于餐厅领域的多模态对话,包括视觉和情感辅助的信息,包括槽和情感标签,情感信息包括积极、消极和中性.提供的图像响应包括风格、人数、餐厅级别等.	单领域,餐厅预订;多模态,模态信息包括:文本(情感)和图像	对话规模:训练集(950)、测试集(450)、验证集(334) 情感:3种(中立、积极、消极) 图像类型:5种(菜肴、种类、人数、位置、价格) 图像个数:1500张 意图个数:1286 语言:英语	论文: https://link.springer.com/article/10.1007/s12559-020-09769-7 数据集:未公开
Code-mixed Medical Dataset[187]	2023 年	医疗诊断领域的多回合医患对话,任务是诊断患者的病情;根据人-人对话提取的医疗领域的多语言对话数据集,是一种英语和泰卢固语混淆的对话数据	单领域,医疗保健领域	诊断类型:10 对话规模:3005 对话轮数:29294 平均对话轮数:9.77 语言:英语,泰卢固语	论文: https://www.sciencedirect.com/science/article/pii/S0885230822000729 数据集: https://github.com/suman101112/Code-Mixed-TOD-Medical-Dataset

表 10 典型数据集的经典方法和性能

算法	发表年份	基础模型	数据集	对话成功率	平均对话轮数	平均对话奖励	人工评测(对话成功率)
VACL[77]	2022 年	DQN	Microsoft Dialogue Challenge	Movie:0.429 Restaurant:0.4256 Taxi:0.6513	Movie:29.02 Restaurant:24.14 Taxi:19.62	Movie:−2.03 Restaurant:−2.76 Taxi:19.80	Movie:0.390 Restaurant:0.280 Taxi:0.440
DPAV[32]	2022 年	DQN	Microsoft Dialogue Challenge	Movie:0.80 Restaurant:0.31 Taxi:0.68	—	Movie:50 Restaurant:21 Taxi:55	—
MCTS-DDU[79]	2020 年	Dueling	Microsoft Dialogue Challenge (movie)	**0.9314**	**12.13**	**55.87**	—
ACL-DQN[76]	2021 年	DQN	Microsoft Dialogue Challenge (movie)	0.8055	17.22	49.05	0.526
LHUA[74]	2020 年	DQN	Microsoft Dialogue Challenge (movie)	0.799	—	—	—
ES-DDQ[148]	2021 年	DDQ	Movie-ticket Booking (movie);Frames	Movie:0.7581 Frames:0.4763	Movie:18.55 Frames:21.06	Movie:41.10 Frames:6.36	Movie:**0.727** Frames:0.429

（续　表）

算法	发表年份	基础模型	数据集	对话成功率	平均对话轮数	平均对话奖励	人工评测(对话成功率)
HRLagent[64]	2017 年	HRL	Frames	0.632	43.0	33.20	0.750
DDQ[15]	2018 年	DQN	Microsoft Dialogue Challenge (movie)	0.7840	19.94	45.11	—
DPPO[98]	2023 年	PPO	Microsoft Dialogue Challenge (movie)	0.8693	17.32	56.65	—
D3Q[99]	2018 年	DDQ+LSTM	Microsoft Dialogue Challenge (movie)	0.7400	13.81	42.89	—
Switch-DDQ[100]	2019 年	DDQ+LSTM	Microsoft Dialogue Challenge (movie)	0.780	12.21	48.49	—
BCS-DDQ[101]	2020 年	DDQ	Microsoft Dialogue Challenge (movie)	0.7629	16.20	44.45	—
Adversarial A2C[49]	2018 年	A2C+GAN	Microsoft Dialogue Challenge (movie)	0.875	13.52	5.93	—
MADPL[135]	2020 年	A2C	MultiWOZ 2.0	0.921	**7.62**	—	0.833
JOIE[137]	2021 年	DQN	MultiWOZ 2.0	Domin2：0.980 Domin4：0.940 Domin7：0.910	Domin2：**5.82** Domin4：8.45 Domin7：9.45	Domin2：**66.71** Domin4：50.59 Domin7：40.82	**0.880**(整体)
GDPL[107]	2019 年	IRL	MultiWOZ 2.0	0.865	7.64	**1.4**	0.7500
Act-VRNN[109]	2020 年	PG	MultiWOZ 2.0	0.867	7.90	—	—
ACGOS[115]	2022 年	A2C (AC)	MultiWOZ 2.0	0.817	14.80	—	—
HDNO[90]	2021 年	HRL	MultiWOZ 2.0 MultiWOZ 2.1	MultiWoz 2.0：0.847 MultiWoz 2.1：0.830	—	—	—
DQN(GAN-VAE[37]	2020 年	DQN+GAN	MultiWOZ 2.0	**0.985**	11.04	—	
OPPA[78]	2020 年	DQN	MultiWOZ 2.0	0.816	8.47	—	0.750
HRLG[124]	2023 年	HRL	MultiWOZ 2.1	**0.928**	**13.1**	—	—
BatchDA2C[82]	2016 年	A2C	DSTC2	—	**4.05**	**0.73**	—
ALTD[106]	2018 年	PG+BiGRU	DSTC2	**0.588**	—	—	—

5.4　大语言模型与对话策略

自 2022 年 11 月 OpenAI 发布 ChatGPT① 以来，大语言模型(Large Language Models，LLMs)在对话领域蓬勃发展. 目前较为流行的大语言模型包括 ChatGPT、LLaMA[188]、PaLM-2②、Sparrow③ 以及百度的文心一言等. 这类大语言模型面向开放话题，称之为开放域语言模型，它们旨在开放话题中解决用户的各种需求，并不面向于特定的任务，对产生的回答有一定的容错性.

任务导向的大语言模型是面向特定任务的一类对话模型. 新近的研究如，TOD-Bert[189] 基于预训练语言模型构建任务导向的对话系统，其将对话策略的对话动作预测问题看成多标签分类问题. SOLOIST[190] 基于 GPT-2 在域外数据上进行训练，基于用户目标和现实世界知识生成响应，在完成特定任务时，该模型并不需要考虑对话动作. SPACE[191] 是一个集自然语言理解、对话策略、自然语言生成为一体的模型，其对话策略的输出并非动作，而是一个用于生成对话回复的向量，本质上是一种对特征的提取. Med-PaLM[192] 是一个医疗问答领域的大语言模型，其基于 PaLM-2 进行构建，是一种端到端的语言模型，并没有单独提取策略模块. 基于大语言模型的任务型对话系统不需要考虑模块之间的集成关系，构建统一的结构直接进行对话生成，避免了模块之间的错误累积. 但该类方法面临着模型参数巨大、训练时间较长、成本昂贵和数据饥渴等局限.

本文主要关注管道方法中基于强化学习的任务型对话策略. 在大语言模型兴起的时代，我们相信基于管道方法的任务型对话策略将会受到更多的重视，有助于中小企业构建数据安全、简单易实施、成

① https://openai.com/blog/chatgpt
② https://ai.google/discover/palm2
③ https://www.deepmind.com/blog/building-safer-dialogue-agents

本节约的对话系统，有助于学术界在不同领域内进一步研发新的模型.

6 未来研究方向

近年来任务型对话策略取得了长足发展，特别是一些新的强化学习技术的突破，使得基于强化学习的对话策略性能得以提升. 多领域的任务型对话策略更是在近些年中发展迅速，然而，依然有很多值得研究的方向在未来有望进一步发展.

(1) 基于强化学习技术的对话策略

强化学习算法常应用在游戏领域，在任务型对话领域中，其状态空间、动作空间较游戏领域更为庞大，而直接将游戏领域内效果较好的强化学习模型迁移到对话策略上时往往并不能达到理想效果. 此外，一些强化学习算法也存在采样低效、收敛较慢、数据利用率低等问题. 基于强化学习的任务型对话策略依然面临和传统强化学习类似的问题. 此处，我们更多地关注对话策略.

① 不同强化学习和其它模型的进一步混合优化：一些新近的强化学习算法和其它先进强化学习模型的混合，以在任务型对话策略中产生更具鲁棒性的效果，将会是近期和后期的一个研究方向. 比如多代理强化学习中结合策略梯度算法研究任务型对话策略相对较少，而该类研究在非对话策略中已经出现过[193].

② 基于模型的强化学习的进一步应用：基于模型的对话策略研究主要集中在基于 DDQ[15] 的算法上，它通过模拟用户的数据增强对话策略性能. 探索不同的、高效的基于模型方法，以促进数据利用率或直接指导动作存在诸多可能性. 此外，在环境随机性、有限数据的不确定性、状态的部分可观察性等前提下构建一个较优的基于模型的对话策略依然存在挑战.

③ 深层次的对话策略建模：在未来对话策略可以涉及到更多支持技术. 例如，为每一次对话建立一个模型，了解用户在整个对话过程中的对话状态变化，以更好地了解用户的需求[194]. 这可以通过应用更多的“语义感知”深度学习技术来实现，也可以通过将问题求解构建为所有相关子任务(目标)来实现. 因此，挖掘用户更深层次的对话特征信息以促进对话策略学习是非常有必要的.

(2) 多领域对话策略

目前，多领域的对话策略正处于时下热点，我们注意到多领域的研究层出不穷，包括结合迁移学习、逆强化学习等. 然而如何低成本、高效率地学习多领域的对话策略是一个核心问题. 而解决奖励稀疏问题时，依然需要基于专家经验进行. 因此，我们认为未来的研究趋势可以为：

① 多领域的并行化. 多领域的状态和动作空间更为巨大，强化学习的试错会使得其学习过程更为缓慢，构建能够并行化的、高效的强化学习是非常有价值的.

② 自动的奖励塑造. 在现阶段基于专家经验解决多领域对话策略奖励延迟问题依然是主流，而如何让强化学习代理能够自学习地挖掘有用价值的信息，或通过迁移学习的方式获取监督信号，从而减少专家经验或完全无需专家经验有待进一步研究.

③ 构建统一的多领域对话策略框架. 领域融合问题是目前多领域的主流，融合方式的不同往往能产生不同的对话策略结果，目前尚缺少一个统一融合框架能够自适应地调整多领域内的知识信息，从而在不同的角度上，包括互补、约束、增量式[195] 的学习对话策略.

(3) 多模态对话策略

多模态的研究为融合不同模态信息以使得状态信息更为具体，而如何在多轮对话中考虑不同模态信息的重要程度，以及如何考虑多模态信息之间的相关性以更好地促进策略决策是有待进一步研究的. 此外，着眼于人机交互的发展，多模态的交互是任务型人机对话的必然趋势，随之而来的将会是多模态接入后的安全问题，而该方向尚未得到充分的研究. 另外，在任务型对话中，多模态的数据集尚缺乏，现有的数据集包含的模态信息有限，这也在一定程度上阻碍了该领域的进一步研究，因此构建多模态的任务型对话数据集能够加快多模态对话策略的研究.

(4) 多代理对话策略

目前，多代理的任务型对话策略较多地关注代理之间的协作，通过分层或拆分处理过往状态等形式构建多代理，然而，多代理的研究并不仅仅局限于代理之间的协作，我们认为存在以下几个研究方向：

① 在多代理协作上. 现有的多代理对话策略在多代理的协作上尚不足，缺乏考虑不同代理的不同能力，探究自适应的多代理对话策略，实现不同优先级的多代理模型，有望增加代理的普适性和进一步提升对话策略的性能.

② 在多代理竞争上. 现有的方法缺乏考虑不同

代理之间的竞争关系，其中一个原因是他们为每一个代理输入了正向的状态信息. 目前，已经有一些研究探索对话策略中的逆行为代理[78]. 在多代理中考虑相反的状态嵌入，从而构建有竞争意识的多代理也是一个有趣的方向. 考虑多代理之间的矛盾信息是构建具备竞争能力的多代理的一个抓点.

③ 协作和竞争共存的多代理研究. 基于多代理之间的竞争和协作关系，构建混合优化的多代理是一种更为通用的方法. 然而，这需要对不同代理功能的深度分析，且由于对话状态的不断变化，如何较好地控制不同代理对状态的映射关系，从而更好融合竞争和协作是非常有意义的研究方向.

(5) 共情对话策略

共情是真正实现人工智能的保障，研究具备共情能力的对话策略，将为人机对话的发展起到积极的作用. 在共情对话策略的研究上，我们认为存在以下研究方向值得探究.

① 多目标的优化. 对话可能会因为多个目标而超载[196]. 将情感、个性和知识融合到对话系统中时，这种多目标的超载情况会更加明显. 对话代理应该尽可能多地考虑到所有不同的方面，具备对用户的内在状态感知能力，并能以最小的代价与用户进行交互. 因此，一个研究方向转变成如何有效地寻找一个最优的解决方案，能够同时感知用户内在状态以优化用户的多目标.

② 显性的情感策略. 存在的研究以情感促进动作的选择(以情感解决奖励延迟问题和提高用户满意度). 情感可以看作是动作空间中的显性行为，是用户直接表现的行为状态. 然而，现有的研究仅仅将情感作为监督信号来促进对话，而用户情感的变化或许会有迹可循. 目前还没有研究利用动作状态来促进用户的情感变化，以使代理具备更好的共情表现. 所以，探索动作和情感的显性映射关系(而非情感映射动作)有助于提高对话策略的性能.

③ 群体内共情对话响应. 具备共情能力的任务型对话策略带来的最大优势之一是动作响应的个性化. 也就是可以针对不同的用户配置文件(或已知的用户偏好特征)进行个性的动作响应，从而形成多样性的回复. 但是存在的研究往往在一组相似用户中生成相似的回复. 一个研究方向是如何挖掘细粒度的用户偏好，让对话代理在群体内产生个性化的对话响应.

此外，不同的类别中的交叉研究，如共情和多代理之间进行结合，用于探究不同代理之间的情感信息变化；多模态和多代理的混合，用于建立不同模态下不同代理协作关系，都是有趣且可以进一步挖掘的话题.

7　总结与展望

本文对基于强化学习的任务型对话策略进行综述，首先对任务型对话策略中常用的强化学习技术进行分类，介绍了对话策略中的常用强化学习模型，并分析了它们的优势和局限. 在此基础上，基于强化学习技术分类，介绍对话策略，包括基于值函数逼近的对话策略、基于策略梯度的对话策略、基于层次强化学习的对话策略、基于模型强化学习对话策略和基于逆强化学习的对话策略. 接着，本文介绍在应用领域的任务型对话策略，将其分成多领域、多模态、多代理和共情对话策略. 最后，介绍对话策略的用户模拟器、对话策略的评估以及对话策略平台和任务型对话系统的数据集. 从 5 种不同的角度总结了任务型对话策略的未来研究方向.

目前，人机对话正处于快速发展的阶段，各种对话平台和应用层出不穷. 在未来，随着各种技术产品的落地，将会吸引越来越多的研究者加入该领域. 强化学习是一种模仿人类学习过程的算法，研究基于强化学习的对话策略有利于人机对话领域的进一步发展.

参　考　文　献

[1] Zhao Yang-Yang, Wang Zhen-Yu, Wang Pei, et al. A survey on task-oriented dialogue systems. Chinese Journal of Computers, 2020, 43(10): 1862-1896(in Chinese)
(赵阳洋，王振宇，王佩等. 任务型对话系统研究综述. 计算机学报，2020，43(10)：1862-1896)

[2] Ni J, Young T, Pandelea V, et al. Recent advances in deep learning based dialogue systems: A systematic survey. Artificial Intelligence Review, 2023, 56: 3055-3155

[3] Zhang Z, Takanobu R, Zhu Q, et al. Recent advances and challenges in task-oriented dialog systems. Science China Technological Sciences, 2020, 63(10): 2011-2027

[4] Chen Y N, Celikyilmaz A, Hakkani-Tür D. Deep learning for dialogue systems//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts. Vancouver, Canada, 2017: 8-14

[5] Chen L, Chen Z, Tan B, et al. AgentGraph: Toward universal dialogue management with structured deep reinforcement learning. IEEE/ACM Transactions on Audio Speech and Language Processing, 2019, 27(9): 1378-1391

[6] Su P H, Budzianowski P, Ultes S, et al. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management//Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. Saarbrücken, Germany, 2017: 147-157

[7] Varges S, Quarteroni S, Riccardi G, et al. Leveraging POMDPs trained with user simulations and rule-based dialogue management in a spoken dialogue system//Proceedings of the SIGDIAL 2009 Conference. London, UK, 2009: 156-159

[8] Goddeau D, Brill E, Glass J R, et al. GALAXY: A human-language interface to on-line travel information//Proceedings of the International Conference on Spoken Language Processing. Yokohama, Japan, 1994: 707-710

[9] Huang Min-Lie, Zhu Xiao-Yan. A finite state automata approach based on slot-feature for dialogue management in spoken dialogue system. Chinese Journal of Computers, 2004, 27(8): 1092-1101(in Chinese)
(黄民烈, 朱小燕. 对话管理中基于槽特征有限状态自动机的方法研究. 计算机学报, 2004, 27(8): 1092-1101)

[10] Gasic M, Kim D, Tsiakoulis P, Breslin C, et al. Incremental on-line adaptation of POMDP-based dialogue managers to extended domains//Proceedings of the International Speech Communication Association. Singapore, 2014: 14-18

[11] Lemon O, Pietquin O. Machine learning for spoken dialogue systems//Proceedings of the 8th Annual Conference of the International Speech Communication Association, Interspeech. Anvers, Belgium, 2007: 1761-1764

[12] Levin E, Pieraccini R, Eckert W. Learning dialogue strategies within the Markov decision process framework//Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings. Santa Barbara, USA, 1997: 72-79

[13] Young B S, Gas M, Thomson B, et al. POMDP-based statistical spoken dialog systems: A review. Proceedings of the IEEE, 2013, 101(5): 1160-1179

[14] Shang L, Lu Z, Li H. Neural responding machine for short-text conversation//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015: 1577-1586

[15] Peng B, Li X, Gao J, et al. Deep Dyna-Q: Integrating planning for task-completion dialogue policy learning//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 2182-2192

[16] Zhao Y J, Li Y L, Lin M. A review of the research on dialogue management of task-oriented systems. Journal of Physics: Conference Series, 2019, 1267(1): 012025

[17] Dai Y, Yu H, Jiang Y, et al. A survey on dialog management: Recent advances and challenges. arXiv preprint arXiv: 2005.02233, 2021

[18] Kwan W-C, Wang H, Wang H, et al. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. Machine Intelligence Research, 2023, 20: 318-334

[19] Sutton R S, Barto A G. Reinforcement Learning: An Introduction. USA: MIT Press, 2018

[20] Bellman R. Some problems in the theory of dynamic programming. Econometrica, 1954, 22(1): 37-48

[21] Hutsebaut-Buysse M, Mets K, Latré S. Hierarchical reinforcement learning: A survey and open research challenges. Machine Learning and Knowledge Extraction, 2022, 4(1): 172-221

[22] Zou Wei, Ge Ling, Liu Yu-Shao. Reinforcement Learning. Beijing: Tsinghua University Press, 2020(in Chinese)
(邹伟, 鬲玲, 刘昱杓. 强化学习. 北京: 清华大学出版社, 2020)

[23] Liu W, Wang Z, Liu X, et al. A survey of deep neural network architectures and their applications. Neurocomputing, 2017, 234: 11-26

[24] Mousavi S S, Schukat M, Howley E. Deep reinforcement learning: An overview//Proceedings of the SAI Intelligent Systems Conference. London, UK, 2018: 426-440

[25] MIT Technology Review. 10 Breakthrough Technologies Archive. https://www.technologyreview.com/

[26] Li Y. Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274, 2018

[27] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529-533

[28] Xiao Y, Hoffman J, Xia T, et al. Deep reinforcement learning with double Q-learning//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, Arizona, USA, 2016: 2094-2100

[29] Ansehel O, Baram N, Shimkin N. Averaged-DQN: Variance reduction and stabilization for deep reinforcement learning//Proceedings of the 34th International Conference on Machine Learning. 2017, 70: 176-185

[30] Lan Q, Pan Y, Fyshe A, et al. Maxmin Q-learning: Controlling the estimation bias of Q-learning. arXiv preprint arXiv:2002.06487, 2021

[31] Kuznetsov A, Shvechikov P, Grishin A, et al. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics//Proceedings of the 37th International Conference on Machine Learning. Vienna, Australia, 2020: 5556-5566

[32] Tian C, Yin W, Moens M F. Anti-overestimation dialogue policy learning for task-completion dialogue system//Proceedings of the Association for Computational Linguistics: NAACL Findings. Seattle, USA, 2022: 565-577

[33] Yarats D, Zhang A, Kostrikov I, et al. Improving sample efficiency in model-free reinforcement learning from images//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2021: 10674-10681

[34] Tokic M. Adaptive ε-greedy exploration in reinforcement learning based on value differences//Proceedings of the Annual Conference on Artificial Intelligence. Berlin, Germany, 2010: 203-210

[35] Badia A P, Piot B, Kapturowski S, et al. Agent57: Outperforming the Atari human benchmark//Proceedings of the 37th International Conference on Machine Learning. Vienna, Australia, 2020: 484-494

[36] Lipton Z, Li X, Gao J, et al. BBQ-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Louisiana, USA, 2018, 32(1): 5237-5244

[37] Li Z, Lee S, Peng B, et al. Guided dialogue policy learning without adversarial learning in the loop//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. 2020: 2308-2317

[38] Pong V, Gu S, Dalal M, et al. Temporal difference models: Model-free deep RL for model-based control//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-14

[39] Ladosz P, Weng L, Kim M, et al. Exploration in deep reinforcement learning: A survey. Information Fusion, 2022, 85: 1-22

[40] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning//Proceedings of the 33rd International Conference on Machine Learning. New York, USA, 2016, 48: 1995-2003

[41] Bellemare M G, Dabney W, Munos R. A distributional perspective on reinforcement learning//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017: 693-711

[42] Hessel M, Modayil J, Van Hasselt H, et al. Rainbow: Combining improvements in DQN//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Louisiana, USA, 2018: 3215-3222

[43] Sutton R S, McAllester D, Singh S, et al. Policy gradient methods for reinforcement learning with function approximation. Advances in Neural Information Processing Systems, 2000: 1057-1063

[44] Schulman J, Levine S, Moritz P, et al. Trust region policy optimization//Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: 1889-1897

[45] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms//Proceedings of the 31st International Conference on Machine Learning. Beijing, China, 2014: 605-619

[46] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971, 2019

[47] Fujimoto S, Van Hoof H, Meger D. Addressing function approximation error in actor-critic methods//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018, 4: 2587-2601

[48] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018, 5: 2976-2989

[49] Peng B, Li X, Gao J, et al. Adversarial advantage actor-critic model for task-completion dialogue policy learning//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, Canada, 2018: 6149-6153

[50] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning//Proceedings of the 33rd International Conference on Machine Learning. New York, USA, 2016: 1928-1937

[51] Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration//Proceedings of the 36th International Conference on Machine Learning. California, USA, 2019: 3599-3609

[52] Vezhnevets A S, Osindero S, Schaul T, et al. FeUdal networks for hierarchical reinforcement learning//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017, 7: 5409-5418

[53] Nachum O, Tang H, Lu X, et al. Why does hierarchy (sometimes) work so well in reinforcement learning? arXiv preprint arXiv:1909.10618, 2019

[54] Baykal-Gursoy M. Semi-Markov decision processes. Wiley Encyclopedia of Operations Research and Management Science. John Wiley & Sons, Inc., 2010: 1-9

[55] Sutton R S, Precup D, Singh S. Between MDPs and Semi-MDPs: A framework for temporal abstraction in RL. Artificial Intelligence, 1999, 112(1): 181-211

[56] Bacon P, Harb J, Precup D. The option-critic architecture//Proceedings of the 31st AAAI Conference on Artificial Intelligence. California, USA, 2017: 1726-1734

[57] Machado M C, Bellemare M G, Bowling M. A Laplacian framework for option discovery in reinforcement learning//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017, 5: 3567-3582

[58] Zhang J, Yu H, Xu W. Hierarchical reinforcement learning by discovering intrinsic options. arXiv preprint arXiv:2101.06521, 2022

[59] Jin M, Ma Z, Jin K, et al. Creativity of AI: Automatic symbolic option discovery for facilitating deep reinforcement learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence. 2022, 36: 7042-7050

[60] Kulkarni T D, Narasimhan K R, Saeedi A. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation//Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 3682-3690

[61] Tang D, Li X, Gao J, et al. Subgoal discovery for hierarchical dialogue policy learning//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 2298-2309

[62] Paul S, van Baar J, Roy-Chowdhury A K. Learning from trajectories via subgoal discovery//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada, 2019, 32: 1-11

[63] Pateria S, Subagdja B, Tan A H, et al. Value-based subgoal discovery and path planning for reaching long-horizon goals. IEEE Transactions on Neural Networks and Learning Systems, 2023: 1-13

[64] Peng B, Li X, Li L, et al. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 2231-2240

[65] Luo F-M, Xu T, Lai H, et al. A survey on model-based reinforcement learning. arXiv preprint arXiv: 2206.09328, 2022

[66] Koller T, Berkenkamp F, Turchetta M, et al. Learning-based model predictive control for safe exploration. Annual Review of Control, Robotics, and Autonomous Systems, 2020, 3: 269-296

[67] Sutton R S. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming//Proceedings of the Seventh International Conference. Austin, USA, 1990: 216-224

[68] Lambert N, Amos B, Yadan O, et al. Objective mismatch in model-based reinforcement learning//Proceedings of the Machine Learning Research. Berkeley, USA, 2020, 120: 1-16

[69] Farahmand A M, Barreto A M S, Nikovski D N. Value-aware loss function for model-based reinforcement learning amir-massoud//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Lauderdale, USA, 2017, 54: 1486-1494

[70] Voelcker C, Liao V, Garg A, et al. Value gradient weighted model-based reinforcement learning//Proceedings of the 10th International Conference on Learning Representations. 2022: 1-19

[71] Wang Y A, Chen Y N. Dialogue environments are different from games: Investigating variants of deep Q-networks for dialogue policy//Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop. Singapore, 2019: 1070-1076

[72] Pathak D, Agrawal P, Efros A A, et al. Curiosity-driven exploration by self-supervised prediction//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 488-489

[73] Zhao T, Eskenazi M. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning //Proceedings of the Conference 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Los Angeles, USA, 2016: 1-10

[74] Cao Y, Lu K, Chen X, et al. Adaptive dialog policy learning with hindsight and user modeling//Proceedings of the Conference 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2020: 329-338

[75] Andrychowicz M, Wolski F, Ray A, et al. Hindsight experience replay//Proceedings of the Advances in Neural Information Processing Systems. CA, USA, 2017: 5048-5058

[76] Zhao Y, Wang Z, Huang Z. Automatic curriculum learning with over-repetition penalty for dialogue policy learning//Proceedings of the 35th AAAI Conference on Artificial Intelligence. 2021: 14540-14548

[77] Zhao Y, Qin H, Zhu C, et al. A versatile adaptive curriculum learning framework for task-oriented dialogue policy learning//Proceedings of the Association for Computational Linguistics: NAACL. Seattle, USA, 2022: 711-723

[78] Zhang Z, Liao L, Zhu X, et al. Learning goal-oriented dialogue policy with opposite agent awareness//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, 2020: 122-132

[79] Wang S, Zhou K, Lai K, et al. Task-completion dialogue policy learning via Monte Carlo tree search with dueling network//Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2020: 3461-3471

[80] Zhang H, Zeng Z, Lu K, et al. Efficient dialog policy learning by reasoning with contextual knowledge//Proceedings of the 36th AAAI Conference on Artificial Intelligence. 2022: 11667-11675

[81] Madusanka T, Langappuli D, Welmilla T, et al. Dialog policy optimization for low resource setting using self-play and reward based sampling//Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation. Hanoi, Vietnam, 2020: 178-187

[82] Fatemi M, Asri L El, Schulz H, et al. Policy networks with two-stage training for dialogue systems//Proceedings of the Conference 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Los Angeles, USA, 2016: 101-110

[83] Zhao R, Tresp V. Improving goal-oriented visual dialog agents via advanced recurrent nets with tempered policy gradient//Proceedings of the IJCAI 2018 Workshop Linguistic and Cognitive Approaches to Dialog Agents. Stockholm, Sweden, 2018: 1-7

[84] Wu Y-C, Rasmussen C E. Clipping loops for sample-efficient dialogue policy optimisation//Proceedings of the North American Chapter of the Association for Computational Linguistics. 2021: 3420-3428

[85] Malviya S, Kumar P, Namasudra S, et al. Experience replay-based deep reinforcement learning for dialogue management optimisation. ACM Transactions on Asian and Low-Resource Language Information Processing, to appear

[86] Shah P, Hakkani-Tür D, Heck L. Interactive reinforcement learning for task-oriented dialogue management//Proceedings of the Workshop on Deep Learning for Action and Interaction. Barcelona, Spain, 2016

[87] Su P-H, Gasic M, Mrksic N, et al. Continuously learning neural dialogue management. arXiv preprint arXiv: 1606.02689, 2016

[88] Cordier T, Urvoy T, Rojas-Barahona L M, et al. Diluted near-optimal expert demonstrations for guiding dialogue stochastic policy optimisation. arXiv preprint arXiv: 2012.04687, 2020

[89] Padmakumar A, Mooney R J. Dialog policy learning for joint clarification and active learning queries//Proceedings of the 35th AAAI Conference on Artificial Intelligence. 2021: 13604-13612

[90] Wang J, Zhang Y, Kim T-K, et al. Modelling hierarchical structure between dialogue policy and natural language generator with option framework for task-oriented dialogue system. arXiv preprint arXiv: 2006.06814, 2021

[91] Tiwari A, Saha S, Bhattacharyya P. A knowledge infused context driven dialogue agent for disease diagnosis using hierarchical reinforcement learning. Knowledge-Based Systems, 2022, 242: 108292

[92] Saha T, Gupta D, Saha S, et al. A hierarchical approach for efficient multi-intent dialogue policy learning. Multimedia Tools and Applications, 2021, 80(28-29): 35025-35050

[93] Chen Z, Liu X, Chen L, et al. Structured hierarchical dialogue policy with graph neural networks//Proceedings of the National Conference on Man-Machine Speech Communication. Santa Fe, USA, 2022: 264-277

[94] Casanueva I, Budzianowski P, Su P H, et al. Feudal reinforcement learning for dialogue management in large domains//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, USA, 2018: 714-719

[95] Geishauser C, Hu S, Lin H, et al. What does the user want? Information gain for hierarchical dialogue policy optimisation//Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop. Cartagena, Colombia, 2021: 969-976

[96] Jong N K, Hester T, Stone P. The utility of temporal abstraction in reinforcement learning//Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems. Estoril, Portugal, 2008: 299-306

[97] Pateria S, Subagdja B, Tan A H, et al. Hierarchical reinforcement learning: A comprehensive survey. ACM Computing Surveys, 2021, 54(5): 1-35

[98] Huang C, Cao B. Learning dialogue policy efficiently through Dyna proximal policy optimization//Proceedings of the International Conference on Collaborative Computing: Networking, Applications and Worksharing. Hangzhou, China, 2022: 396-414

[99] Su S, Li X, Gao J, et al. Discriminative deep Dyna-Q: Robust planning for dialogue policy learning//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 3813-3823

[100] Wu Y, Li X, Liu J, et al. Switch-based active deep Dyna-Q: Efficient adaptive planning for task-completion dialogue policy learning//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 7289-7296

[101] Zhang Z, Li X, Gao J, et al. Budgeted policy learning for task-oriented dialogue systems//Proceedings of the Conference 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2020: 3742-3751

[102] Zhang M, Shinozaki T. DNN-rule hybrid Dyna-Q for sample-efficient task-oriented dialog policy learning//Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Chiang Mai, Thailand, 2022: 1428-1434

[103] Arora S, Doshi P, Elsevier B V. A survey of inverse reinforcement learning: Challenges, methods and progress. Artificial Intelligence, 2021, 297: 103500

[104] Moshinsky M. Algorithms for inverse reinforcement learning//Proceedings of the 17th International Conference on Machine Learning. California, USA, 2000: 663-670

[105] Fu J, Luo K, Levine S. Learning robust rewards with adversarial inverse reinforcement learning//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-15

[106] Liu B, Lane I. Adversarial learning of task-oriented neural dialog models//Proceedings of the Conference 19th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Melbourne, Australia, 2018: 350-359

[107] Takanobu R, Zhu H, Huang M. Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog//Proceedings of the Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 100-110

[108] Hou Z, Liu B, Zhao R, et al. Imperfect also deserves reward: Multi-level and sequential reward modeling for better dialog management//Proceedings of the North American Chapter of the Association for Computational Linguistics. 2021: 2993-3001

[109] Huang X, Qi J, Sun Y, et al. Semi-supervised dialogue policy learning via stochastic reward estimation//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 660-670

[110] Gašić M, Mrkšić N, Su P-H, et al. Policy committee for adaptation in multi-domain spoken dialogue systems//Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding. Scottsdale, USA, 2015: 806-812

[111] Cuayahuitl H, Yu S, Williamson A, et al. Scaling up deep reinforcement learning for multi-domain dialogue systems//Proceedings of the International Joint Conference on Neural Networks. Anchorage, USA, 2017: 3339-3346

[112] Mendez J A, Liu B. Reinforcement learning of multi-domain dialog policies via action embeddings. arXiv:2207.00468v1, 2022

[113] Peng S, Ji F, Lin Z, et al. MTSS: Learn from multiple domain teachers and become a multi-domain dialogue expert//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 8608-8615

[114] Zhao M, Wang L, Jiang Z, et al. Multi-task learning with graph attention networks for multi-domain task-oriented dialogue systems. Knowledge-Based Systems, 2023, 259: 110069

[115] Cordier T, Urvoy T, Lefèvre F, et al. Graph neural network policies and imitation learning for multi-domain task-oriented dialogues//Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SigDial). Edinburgh, UK, 2022: 91-100

[116] Cordier T, Urvoy T, Lefevre F, et al. Few-shot structured policy learning for multi-domain and multi-task dialogues//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Croatia, 2023: 432-441

[117] Rohmatillah M, Chien J T. Causal confusion reduction for robust multi-domain dialogue policy//Proceedings of the Annual Conference of the International Speech Communication Association. 2021, 5: 3761-3765

[118] Wu C S, Socher R, Xiong C. Global-to-local memory pointer networks for task-oriented dialogue. arXiv preprint arXiv: 1901.04713, 2019

[119] Qin L, Xu X, Che W, et al. Dynamic fusion network for multi-domain end-to-end task-oriented dialog//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 6344-6354

[120] Jeon H, Lee G G. DORA: Towards policy optimization for task-oriented dialogue system with efficient context. Computer Speech and Language, 2022, 72: 101310

[121] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding //Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 4171-4186

[122] Wang H, Wang H, Wang Z, et al. Integrating pretrained language model for dialogue policy evaluation//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, 2022: 6692-6696

[123] Saha T, Gupta D, Saha S, et al. Towards integrated dialogue policy learning for multiple domains and intents using hierarchical deep reinforcement learning. Expert Systems with Applications, 2020, 162: 113650

[124] Rohmatillah M, Chien J-T. Hierarchical reinforcement learning with guidance for multi-domain dialogue policy. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 748-761

[125] Zhao T. ReinForest: Multi-domain dialogue management using hierarchical policies and knowledge ontology. https://api.semanticscholar.org/CorpusID:15445108, 2016

[126] Liao L, Ma Y, He X, et al. Knowledge-aware multimodal dialogue systems//Proceedings of the 26th ACM Multimedia Conference. New York, USA, 2018: 801-809

[127] Saha T, Saha S, Bhattacharyya P. Towards sentiment-aware multi-modal dialogue policy learning. Cognitive Computation, 2020, 14(1): 246-260

[128] Tiwari A, Saha T, Saha S, et al. Multi-modal dialogue policy learning for dynamic and co-operative goal setting//Proceedings of the International Joint Conference on Neural Networks. Shenzhen, China, 2021: 1-8

[129] Tiwari A, Manthena M, Saha S, et al. Dr. can see: Towards a multi-modal disease diagnosis virtual assistant//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. Atlanta, USA, 2022: 1935-1944

[130] Zorrilla A L, Torres M I, Cuayahuitl H. Audio embedding-aware dialogue policy learning. IEEE/ACM Transactions on Audio Speech and Language Processing, 2023, 31: 525-538

[131] Zhang J, Zhao T, Yu Z. Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog//Proceedings of the Conference 19th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Melbourne, Australia, 2018: 140-150

[132] Deng Y, Li Y, Ding B, et al. Leveraging long short-term user preference in conversational recommendation via multi-agent reinforcement learning. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(11): 11541-11555

[133] Naghizadeh P, Gorlatova M, Lan A S, et al. Hurts to be too early: Benefits and drawbacks of communication in multi-agent learning//Proceedings of the IEEE INFOCOM. Paris, France, 2019: 622-630

[134] Liu B, Lane I. Iterative policy learning in end-to-end trainable task-oriented neural dialog models//Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Okinawa, Japan, 2017: 482-489

[135] Takanobu R, Liang R, Huang M. Multi-agent task-oriented dialog policy learning with role-aware reward decomposition//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 625-638

[136] Zhuang Y, Yu T, Wu J, et al. Spatial-temporal aligned multi-agent learning for visual dialog systems//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal, 2022: 482-490

[137] Wang H, Wong K F. A collaborative multi-agent reinforcement learning framework for dialog action decomposition//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 7882-7889

[138] Papangelis A, Wang Y C, Molino P, et al. Collaborative multi-agent dialogue model training via reinforcement learning //Proceedings of the Conference the 20th Annual Meeting

of the Special Interest Group Discourse Dialogue. Stockholm, Sweden, 2019: 92-102

[139] Tiwari A, Saha T, Saha S, et al. A persona aware persuasive dialogue policy for dynamic and co-operative goal setting. Expert Systems with Applications, 2022, 195: 116303

[140] Gu S, Kuba J G, Wen M, et al. Multi-agent constrained policy optimisation. arXiv preprint arXiv:2110.02793, 2022

[141] Liu K, Picard R. Embedded empathy in continuous, interactive health assessment//Proceedings of the CHI Workshop on Challenges in Health Assessment. Portland, USA, 2005: 1-4

[142] Valentini S, Orsingher C, Polyakova A. Customers' emotions in service failure and recovery: A meta-analysis. Marketing Letters, 2020, 31: 199-216

[143] Broekens J. Emotion and reinforcement: Affective facial expressions facilitate robot learning//Proceedings of the Artifical Intelligence for Human Computing: ICMI 2006 and IJCAI 2007 International Workshops. Banff, Canada, 2007: 113-132

[144] Song S, Wang C, Chen H, et al. An emotional comfort framework for improving user satisfaction in E-commerce customer service Chatbots//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, 2021: 130-137

[145] Bui T H, Zwiers J, Poel M, et al. Affective dialogue management using factored POMDPs. Interactive Collaborative Information Systems, 2010, 281: 207-236

[146] Shi W, Yu Z. Sentiment adaptive end-to-end dialog systems //Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 1509-1519

[147] Li W, Shao W, Ji S, et al. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. Neurocomputing, 2022, 467: 73-82

[148] Zhang R, Wang Z, Zheng M, et al. Emotion-sensitive deep Dyna-Q learning for task-completion dialogue policy learning. Neurocomputing, 2021, 459: 122-130

[149] Tu Q, Li Y, Cui J, et al. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation //Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland, 2022: 308-319

[150] Saha T, Gupta D, Saha S, et al. Emotion aided dialogue act classification for task-independent conversations in a multimodal framework. Cognitive Computation, 2021, 13(2): 277-289

[151] Wang W, Zhang J, Zhang H, et al. A teacher-student framework for maintainable dialog manager//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 3803-3812

[152] Chen Y N, Hakkani-Tür D, He X. Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China, 2016: 6045-6049

[153] Siddique A B, Maqbool M H, Taywade K, et al. Personalizing task-oriented dialog systems via zero-shot generalizable reward function//Proceedings of the International Conference on Information and Knowledge Management. Atlanta, USA, 2022: 1787-1797

[154] Pei J, Ren P, De Rijke M. A cooperative memory network for personalized task-oriented dialogue systems with incomplete user profiles//The Web Conference 2021-Proceedings of the World Wide Web Conference. Singapore, 2021: 1552-1561

[155] Bihani G, Rayz J T. Fuzzy classification of multi-intent utterances//Proceedings of the North American Fuzzy Information Processing Society Annual Conference. 2022: 37-51

[156] Shi C, Chen Q, Sha L, et al. We know what you will ask: A dialogue system for multi-intent switch and prediction// Proceedings of the Natural Language Processing and Chinese Computing. Dunhuang, China, 2019: 93-104

[157] Li X, Lipton Z C, Dhingra B, et al. A user simulator for task-completion dialogues. arXiv preprint arXiv: 1612.05688, 2017

[158] Schatzmann J, Thomson B, Weilhammer K, et al. Agenda-based user simulation for bootstrapping a POMDP dialogue system//Proceedings of the North American Chapter of the Association of Computational Linguistics. New York, USA, 2007: 149-152

[159] Jain A, Pecune F, Matsuyama Y, et al. A user simulator architecture for socially-aware conversational agents//Proceedings of the 18th International Conference on Intelligent Virtual Agents. Sydney, Australia, 2018: 133-140

[160] Liu H, Ou Z, Huang Y, et al. Jointly reinforced user simulator and task-oriented dialog system with simplified generative architecture. arXiv preprint arXiv:2210.06706, 2022

[161] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. https://api.semanticscholar.org/CorpusID:160025533, 2019

[162] Kreyssig F L, Casanueva I, Budzianowski P, et al. Neural user simulation for corpus-based policy optimisation for spoken dialogue systems//Proceedings of the 19th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Melbourne, Australia, 2018: 60-69

[163] Shi W, Qian K, Wang X, et al. How to build user simulators to train RL-based dialog systems//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 1990-2000

[164] Dhingra B, Li L, Li X, et al. Towards end-to-end reinforcement learning of dialogue agents for information access//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 484-495

[165] Sun W, Zhang S, Balog K, et al. Simulating user satisfaction for the evaluation of task-oriented dialogue systems//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 2499-2506

[166] Ultes S, Rojas-Barahona L, Su P H, et al. Pydial: A multi-domain statistical dialogue system toolkit//Proceedings of the System Demonstrations 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 73-78

[167] Miller A H, Feng W, Fisch A, et al. ParlAI: A dialog research software platform//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Copenhagen, Denmark, 2017: 79-84

[168] Bocklisch T, Faulkner J, Pawlowski N, et al. Rasa: Open source language understanding and dialogue management. arXiv preprint arXiv:1712.05181, 2017

[169] Papangelis A, Namazifar M, Khatri C, et al. Plato dialogue system: A flexible conversational AI research platform. arXiv:2001.06463. 2020

[170] Zhou K, Wang X, Zhou Y, et al. CRSLab: An open-source toolkit for building conversational recommender system//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. 2021: 185-193

[171] Li Y, Arnold J, Yan F, et al. LEGOEval: An open-source toolkit for dialogue system evaluation via crowdsourcing//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. 2021: 317-324

[172] Lee S, Zhu Q, Takanobu R, et al. ConvLab: Multi-domain end-to-end dialog system platform//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Florence, Italy, 2019: 64-69

[173] Zhu Q, Zhang Z, Fang Y, et al. ConvLab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 142-149

[174] Zhu Q, Geishauser C, Lin H, et al. ConvLab-3: A flexible dialogue system toolkit based on a unified data format. arXiv preprint arXiv:2211.17148, 2022

[175] Speggiorin A, Dalton J, Leuski A. TaskMAD: A platform for multimodal task-centric knowledge-grounded conversational experimentation//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid, Spain, 2022: 3240-3244

[176] Zhu Q, Huang K, Zhang Z, et al. CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. Transactions of the Association for Computational Linguistics, 2020, 8: 281-295

[177] Budzianowski P, Wen T H, Tseng B H, et al. MultiWOZ—A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 5016-5026

[178] Budzianowski P, Wen T-H, Tseng B-H, et al. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. arXiv preprint arXiv:1807.11125, 2018

[179] Asri L El, Fine E. Frames: A corpus for adding memory to goal-oriented dialogue systems//Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. Saarbrücken, Germany, 2017: 207-219

[180] Schmitt A, Ultes S, Minker W. A parameterized and annotated spoken dialog corpus of the CMU let's go bus information system//Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey, 2012: 3369-3373

[181] Henderson M, Thomson B, Williams J. The second dialog state tracking challenge//Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Philadelphia, USA, 2014: 263-272

[182] Rastogi A, Zang X, Sunkara S, et al. Schema-guided dialogue state tracking task at DSTC8. arXiv preprint arXiv: 2002.01359, 2020

[183] Wei Z, Liu Q, Peng B, et al. Task-oriented dialogue system for automatic diagnosis//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 201-207

[184] Feng S, Lubis N, Geishauser C, et al. EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems//Proceedings of the Language Resources and Evaluation Conference. Marseille, France, 2022: 4096-4113

[185] Chen M, Liu R, Shen L, et al. The JDDC corpus: A large-scale multi-turn Chinese dialogue dataset for E-commerce customer service//Proceedings of the 12th International Conference on Language Resources and Evaluation. Marseille, France, 2020: 459-466

[186] Joshi C K, Mi F, Faltings B. Personalization in goal-oriented dialog. arXiv preprint arXiv:1706.07503, 2017

[187] Dowlagar S, Mamidi R. A code-mixed task-oriented dialog dataset for medical domain. Computer Speech and Language, 2023, 78: 101449

[188] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv: 2302.13971, 2023

[189] Wu C S, Hoi S, Socher R, et al. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue// Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2020: 917-929

[190] Peng B, Li C, Li J, et al. SOLOIST: Building task bots at scale with transfer learning and machine teaching. Transactions of the Association for Computational Linguistics, 2021, 9: 807-824

[191] He W, Dai Y, Yang M, et al. Unified dialog model pre-training for task-oriented dialog understanding and generation //Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid, Spain, 2022: 187-200

[192] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature, 2023, 620: 172-180

[193] Qiu Y, Member S, Jin Y, et al. Improving sample efficiency of multi-agent reinforcement learning with non-expert policy for flocking control. IEEE Internet of Things Journal, 2023, 10(16): 14014-14027

[194] Howard N, Cambria E. Intention awareness: Improving upon situation awareness in human-centric environments. Human-Centric Computing and Information Sciences, 2013, 3(1): 1-17

[195] Sodhani S, Faramarzi M, Mehta S V, et al. An introduction to lifelong supervised learning. arXiv preprint arXiv: 2207.04354, 2022

[196] Ma Y, Nguyen K L, Xing F Z, et al. A survey on empathetic dialogue systems. Information Fusion, 2020, 64: 50-70

XU Kai, Ph. D. candidate. His main research interests include task-based dialogue policy, reinforcement learning and deep reinforcement learning.

WANG Zhen-Yu, Ph. D., professor, Ph. D supervisor. His main research interests include natural language processing, dialogue systems and deep reinforcement learning.

WANG Xu, Ph. D. candidate. His main research interests include dialogue systems, natural language generation and deep reinforcement learning.

QIN Hua, M. S. candidate. Her main research interests include dialogue policy, reinforcement learning and deep reinforcement learning.

LONG Yu-Xuan, Ph. D. candidate. His main research interests include deep learning, cloud computing and deep reinforcement learning.

Background

Task-oriented dialogue systems are designed to help users solve domain-specific tasks, dialogue policy is a core component of task-oriented dialogue systems based on the pipeline approach. In the pipeline-based method, the dialogue policy takes the dialogue state and generates dialogue actions for natural language generation. The development of dialogue policy has seen rule-based dialogue policy, supervised learning-based dialogue policy, and reinforcement learning-based dialogue policy. Rule-based and supervised learning dialogue policies have the advantage of being easy to analyze and debug. However, they are relatively inflexible and poorly scalable due to their high dependence on expert interaction. Moreover, speech recognition and natural language understanding will inevitably be inaccurate, which makes dialogue policies usually based more on reinforcement learning modeling. In recent years, with the emergence of deep learning, combining deep learning and reinforcement learning has become a first choice for dialogue policy learning, and a large number of studies have implemented dialogue policy based on deep reinforcement learning techniques. The emergence of new techniques and methods in the field of dialogue policy is particularly prolific, and there is an imperative demand to review and summarize the frontier research methods of task-oriented dialogue policy.

In this paper, we introduce the reinforcement learning methods that are commonly used in dialogue policy and summarize the current frontier methods of task-oriented dialogue policy. We present the current state of research on reinforcement learning-based dialogue policies on technology-based and application-based perspectives, respectively. Finally, we discuss the limitations of dialogue policy and indicate some future trends, intending to provide a valuable reference for its future development.

This work is supported by the Guangdong Province Key Areas of Research and Development Program (No. 2021-B0101190002).