中国科学院大学
University of Chinese Academy of Sciences

硕士学位论文

标注受限任务型对话的泛化策略研究

作者姓名：刘龙祥

指导教师：冯洋 研究员

中国科学院计算技术研究所

学位类别：工学硕士

学科专业：计算机应用技术

培养单位：中国科学院计算技术研究所

2024 年 6 月

Generalization Strategies of Task-Oriented Dialog in Annotation-Constrained Scenarios

A thesis submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Master of Engineering

in Computer Applied Technology

By

LIU Longxiang

Supervisor: Professor Yang Feng

Institute of Computing Technology, Chinese Academy of Sciences

June, 2024

中国科学院大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。承诺除文中已经注明引用的内容外，本论文不包含任何其他个人或集体享有著作权的研究成果，未在以往任何学位申请中全部或部分提交。对本论文所涉及的研究工作做出贡献的其他个人或集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关收集、保存和使用学位论文的规定，即中国科学院大学有权按照学术研究公开原则和保护知识产权的原则，保留并向国家指定或中国科学院指定机构送交学位论文的电子版和印刷版文件，且电子版与印刷版内容应完全相同，允许该论文被检索、查阅和借阅，公布本学位论文的全部或部分内容，可以采用扫描、影印、缩印等复制手段以及其他法律许可的方式保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

任务型对话系统的目标是通过与用户的多轮交互，高效准确地协助其完成特定领域的目标和动作，如生活服务、百科知识获取等，并最大化地提升用户的使用体验，降低人工成本。随着深度学习时代自然语言处理技术的迅速发展，这项技术已从实验室走向了现实场景，并随着大模型的问世，进一步改变了人们的生活。然而，目前的对话系统在多领域多场景的现实世界中，泛化性能仍然较差。

具体来说，为了构建一个高质量且能够覆盖尽可能多领域的任务型对话系统，开发人员通常需要根据用户的需求收集各种不同领域的语料，并进行标注。然而标注成本昂贵，数据通常较为稀缺或者仅包含有限的领域。在这种标注受限的场景下，本课题研究了实现泛化性强的任务型对话系统过程中存在的多个关键挑战：（1）标注数据稀缺的问题：不充足数据量下训练出的对话系统难以处理已见数据之外的模式。如何充分挖掘标注信息且进行更为鲁棒的训练方式是值得探索的问题；（2）标注模式依赖、迁移性差的问题：任务型对话的状态维护依赖于预定义好的本体以及固定的知识库模式，使得模型可扩展性变差，且模型与数据无法独立维护。如何建立一种能够摆脱标注模式的范式是十分重要的问题；（3）实时用户反馈缺失的问题：传统的基于固定数据集的自动评估由于缺乏实时的用户反馈，无法准确指导真实场景下的系统构建。如何引入能够覆盖多种领域的基于实时用户反馈的训练与评估框架则是契合现实的问题。

针对上述三个问题，本课题基于提升任务型对话系统泛化性的核心目标，沿着利用现有数据，摆脱特定模式以及迈向真实场景这条由易到难的研究路线，依次探索和验证了相应的解决方案；通过对训练过程中测试场景的模拟，以及测试场景的不断拓展，使其与真实场景的差距逐步减小，旨在最终构建出泛化性强的任务型对话系统。具体取得了如下成果：

1. 基于标注信息挖掘的对话表示学习方法

具有全环节完整标注的任务型对话数据较为稀缺，而已有方法对于中间状态标注的利用不够充分，且教师指导的训练方式使动作到回复的映射存在错误累积。为了解决标注利用不充分的问题，本研究对中间状态标注中含有的高层次信息进行挖掘（包括状态类别、动作类别、状态转移以及关键词），使用多任务学习来增强句子的向量表示。为了解决错误累积的问题，本研究引入对动作序列的计划采样策略，通过对测试可能出现动作的模拟，减少训练时的曝光偏差。实验结果表明该方法能够显著提升低资源场景下任务型对话的表现。分析实验也验证了多任务学习和计划采样的有效性。

2. 基于检索增强的对话标注模式解耦方法

为了解决标注中的模式依赖问题，通常采用检索增强的方式来解耦知识库模式与模型训练。然而一般的检索式增强存在两个问题：首先，简单使用历史对

话作为查询可能会带来噪音或者信息缺失；其次，正确知识的标签可能未经标注或者仅展示有限的模式。为了解决查询语句难以确定的问题，本研究引入一个查询改写模块来自动生成查询语句，通过强化学习优化回复质量奖励；为了解决知识标注缺失的问题，本研究引入后验信息指导的端到端训练，使标准回复提供正确知识的线索。实验结果表明该方法能够相比于基线显著提升端到端任务型对话的表现。分析实验也验证了查询改写和后验信息的有效性。

3. **基于模拟交互的对话偏好对齐训练方法**

为了解决任务型对话系统训练缺乏实时用户交互的问题，通常需要构建用户模拟器并使用强化学习来提升模型表现。然而基于固定数据集的评估方式无法适用于交互式环境，并且不同数据集/领域上训练的用户模拟器存在较大差异，在单一环境下训练的系统难以迁移到其他环境。为了解决评估方式的问题，本研究提出了基于 ChatGPT 提示工程的会话评估框架。随后对于已建立的多个用户模拟器，使系统模型动态地选择不同的模拟器进行交互，在交互过程中引入 ChatGPT 在环路的筛选-修改策略，构造偏好数据，用其对当前模型进行迭代式的偏好对齐。实验结果表明本研究提出的训练框架能够持续地提升对话系统的综合表现。分析实验验证了检索增强以及不同损失函数对于迭代过程的有效性。

综上所述，本课题以提升标注受限场景下任务型对话的泛化性能为核心，依次从增强表示、解耦知识以及偏好对齐的角度，通过尽可能地构造更为多样真实的环境，减少稀缺标注数据下训练的对话系统与现实场景的差距，对其中存在的多个关键问题设计了相应的解决方案，希望本课题能够为任务型对话系统在复杂多变的现实环境中的落地起到一定的启发作用。

关键词：任务型对话，检索增强生成，偏好对齐，大语言模型，鲁棒训练

Abstract

The goal of task-oriented dialogue systems is to efficiently and accurately assist users in accomplishing domain-specific goals and actions, such as life services, encyclopedic knowledge acquisition, etc., through multiple turns of interactions with them, and to maximize the user experience and reduce labor costs. With the rapid development of natural language processing technology in the era of deep learning, this technology has moved from the research to real-world scenarios and further changed people's lives with the emergence of large language models. However, existing dialogue systems still have poor generalization performance in the real world of multi-domain and multi-scenario.

Specifically, in order to build a task-oriented dialog system that is high-quality and can cover as many domains as possible, developers usually need to collect a variety of corpora from different domains and annotate them according to the user's needs. However, labeling is expensive, and the data is usually scarce or contains only a limited number of domains. In this annotation-constrained scenario, we investigate several key challenges in realizing a task-based dialog system with strong generalization: (1) Scarcity of annotated data: Dialog systems trained on insufficient data are difficult to deal with patterns beyond the seen data. It is worth exploring how to fully exploit the labeled information and conduct more robust training; (2) the problem of labeled pattern dependence and poor migration: the maintenance of task-oriented dialogs relies on predefined ontologies and fixed knowledge base patterns, which makes the model less scalable and the model and data cannot be maintained independently. How to establish a paradigm that can get rid of the annotation model is an important issue; (3) the problem of the lack of real-time user feedback: the traditional automatic evaluation based on fixed datasets can not accurately guide the construction of the system in real scenarios due to the lack of real-time user feedback. How to introduce a training and evaluation framework based on real-time user feedback that can cover a variety of domains is a problem that fits the reality.

To address the above three problems, based on the core objective of improving the generalizability of task-oriented dialogue systems, this project explores and validates the corresponding solutions along the research route of using existing data, getting rid of specific patterns and moving towards real scenarios from easy to difficult; through the simulation of test scenarios in the training process, as well as the continuous expansion of the test scenarios, the gap between the test scenarios and the real scenarios is gradually reduced, aiming to finally construct a task-based dialogue system with strong generalization. The following results were achieved:

1. **Dialog Representation Learning Based on Annotated Information Mining**

Task-oriented dialogue data with full-session complete annotations are scarce, while existing methods underutilize intermediate state annotations, and the teacher-forcing training approach accumulates errors in action-to-response mappings. To address the underutilization of annotations, this study mines the high-level information contained in intermediate state annotations (including state types, action types, state changes, and keywords) and uses multi-task learning to enhance the representation of turn sentence vector. To address the problem of error accumulation, this study introduces a scheduled sampling strategy for action sequences, which reduces the exposure bias during training by simulating the possible actions in the test stage. The experimental results show that this method can significantly improve the performance of task-oriented dialog in low-resource scenarios. The analysis experiments also verify the effectiveness of multi-task learning and scheduled sampling.

2. **Dialog Annotation Schema Decoupling Based on Retrieval Augmenting**

To solve the problem of schema dependency in annotation, retrieval augmentation is often used to decouple the knowledge base schema from model training. However, general retrieval-augmented generation suffers from two problems: first, simply using dialog history as query may introduce noise or missing information; second, the correct knowledge may be unlabeled or only show limited patterns. To address the problem of determining proper query, this study introduces a query rewriting module to automatically generate query statements and optimize response quality rewards through reinforcement learning; to address the problem of missing knowledge labeling, this study introduces a posterior information-guided end-to-end training to enable golden responses to provide clues to knowledge selection. Experimental results show that this method can significantly improve the performance of end-to-end task-based dialogs compared to the baseline. The analysis experiments also validate the effectiveness of query rewriting and a posterior information.

3. **Dialog Preference Alignment Training Based on Simulated Interaction**

To address the lack of real-time user interaction for task-oriented dialog system training, it is often necessary to build user simulators and use reinforcement learning to improve model performance. However, the evaluation based on fixed datasets cannot be applied to interactive environments, and user simulators trained on different datasets/domains differ significantly, making it difficult to migrate systems trained in a single environment to other environments. To address the problems of improper evaluation, this study proposes a session evaluation framework based on the ChatGPT prompt engineering. Subsequently, for multiple user simulators that have been established, the system model is made to dynamically select different simulators to interact with, and ChatGPT's filtering-modification strategy in the loop is introduced during the interaction process to construct the preference data, with which the current model is iteratively

preference-aligned. Experimental results show that the training framework proposed in this study can consistently improve the overall performance of the dialog system. The analyzed experiments validate the effectiveness of retrieval-augmented generation as well as different loss functions for the iterative process.

In summary, this paper aims to enhance the generalization performance of task-oriented dialogue systems under constrained annotation scenarios. It systematically addresses several key issues from the perspectives of enhancing representation, decoupling knowledge, and aligning preference. By constructing more diverse and realistic environments and reducing the gap between dialogue systems trained with scarce annotated data and real-world scenarios, the paper proposes corresponding solutions to these challenges. It is hoped that this paper will provide inspiration for the deployment of task-oriented dialogue systems in complex and dynamic real-world environments.

Key Words: Task-Oriented Dialog, Retrieval-Augmented Generation, Preference Alignment, Large Language Model, Robust Training

目 录

图目录

表目录

第 1 章　绪论

1.1　研究背景及意义

人机对话技术的核心目标是创造一种自然而智能的交流方式，使得计算机能够更全面、准确地理解和回应人类的语言，从而为人类提供更智能、高效的服务。这一领域的蓬勃发展承载着社会进步的巨大潜力。从历史的角度审视其发展轨迹，我们可以追溯到人工智能的初创时期。当时，人们试图通过规则和模板来模拟对话，如早期聊天机器人 ALICE 和 ELIZA。然而，这些模型受限于规则的刚性和模板的有限表达能力，无法真正理解语境和进行自由流畅的对话。其符号规则和模板需耗费大量人力物力，导致领域扩展性不足。

随着技术的演进，统计机器学习方法的兴起为人机对话的发展注入了新的活力，诞生了以 Siri、Watson、Cortana 等为代表的第二代对话系统。基于统计的语言模型和有限状态机使得对话系统能够更好地处理不同语言表达和语境变化，为当时的自然语言处理领域带来了新的思维方式。然而，真正的飞跃发生在深度学习时代。神经网络的崛起使得对话系统更加注重语义理解和上下文的处理。深度学习的 Seq2Seq 模型提出了一种通用的序列到序列生成的建模方式，对话系统的这种序列生成模式天然地适用于此。而在注意力机制基础上提出的 Transformer 结构[2] 则使得模型参数规模的不断扩大以及对超长序列的建模成为可能。这一系列的技术创新使得对话系统的端到端学习变得可行，整体表现逐渐超越了传统方法的局限，以大规模预训练语言模型为基座的对话系统成为当前的研究主流，形成了百度的 PLATO[9]、Meta 的 Blender[10]、Google 的 Meena[11] 等代表性系统。在此基础上，人们发现当参数量大到百亿级别时，模型不仅获得了在对话流畅性上的显著提升，甚至涌现出了之前小规模模型所没有的逻辑推理能力[12]。随着 ChatGPT 的开放使用，它强大的能力引发了人机对话乃至整个自然语言处理领域研究方式的变化，人们的研究也开始逐渐转向对百亿规模基座模型的训练以及大模型充当任务助理的应用方式中。

目前，人机对话技术已经从实验室走向了现实生活，广泛应用于各个领域。如图1-1所示，无论是国内外，已经涌现出许多具有影响力的落地应用，例如智能音箱领域中 Amazon 的 Echo、阿里巴巴的天猫精灵、百度的小度；手机智能语音助手中小米的小爱同学、苹果的 Siri、Google Assistant 等；智能客服中阿里巴巴的阿里小蜜、京东的言犀、Google Contact Center AI 等。这些系统能够高效地解决用户的问题，同时提升了企业的服务效率。而在医疗、教育、金融等行业，人机对话技术也展现出巨大的应用潜力，推动着社会进步的步伐。由此可见，对话系统的研究具有巨大的现实意义以及学术价值。

根据不同的应用场景，对话系统通常划分为两大类：开放域的闲聊式对话和垂直领域的任务型对话系统。如图1-2所示，前者通常要求回复具有一致性、多样性和个性化，并包含情感成分，不受具体领域的限制；而后者旨在通过尽可能

图 1-1　国内外对话应用产品

Figure 1-1　Domestic and international dialog application products

少的对话轮次来协助用户完成特定领域的目标或动作，如预订机票、酒店和餐馆等。通常，任务型对话系统需要与领域相关的数据库结合，以提供对用户的准确回复，对语义理解和回复内容的准确性要求较高。在工业界，由于落地场景的复杂性、评估的困难以及用户需求等因素，闲聊式对话距离大规模落地仍存在不小的差距；相比之下，由于场景明确且能够显著降低简单任务上的人力成本，任务型对话在实际应用中更为普遍，本课题的研究方向也集中在任务型对话上。此外，需要特别指出的是，近年来随着大模型的发展，闲聊式对话的流畅性得到了提升，人们也逐渐开始探索同一个系统既具备闲聊式对话的情感陪护能力，又具备任务型对话的智能助手能力的可能性，使得实际落地中这两类对话系统的边界的区分不那么重要。

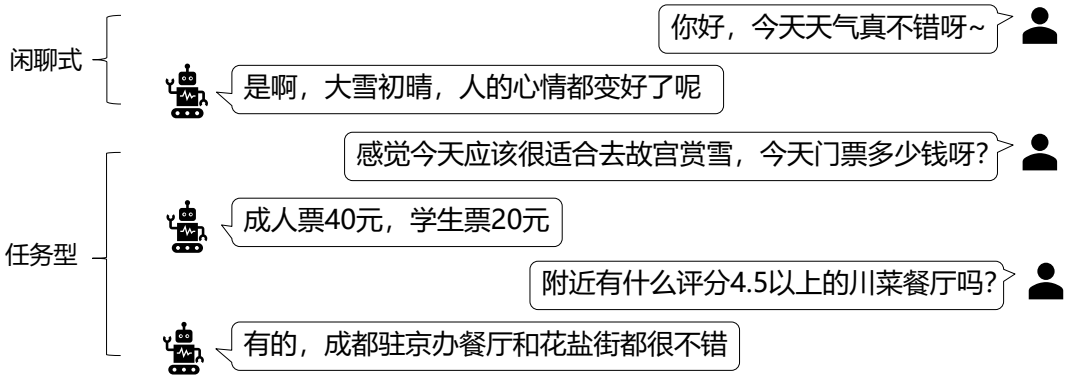


图 1-2　闲聊式对话与任务型对话

Figure 1-2　Chit-chat vs. task-oriented dialog

1.2　关键挑战

要在复杂多变的真实场景中构建一个泛化性强的对话系统，由于标注环节的繁杂以及优质数据获取的昂贵成本，目前仍然存在诸多重大挑战。图1-3展示

了一个任务型对话系统的实例，通常，系统需要与正确理解用户的具体需求，在图中对应着模型需要解析出用户想去的景点名称是故宫，想要大于 4.5 分且在故宫附近的川菜餐厅；获取知识库的信息，在图中对应着使用前述限制条件去检索数据库，获取到了符合条件的条目（成都驻京办餐厅）；采取正确的动作，在图中对应着系统将饭店名称告知用户，并询问是否预定；后续还需要根据用户的下一步反馈进一步调整未来的表现，例如图中用户表达满意，系统可以认识到目前对话是合理的，若用户反馈不满意，则需要系统及时调整对话的策略。总而言之，一个完整的任务型对话系统通常要求对中间状态（对话状态、对话动作）、知识库模式以及实时用户反馈这三个方面的标注，而在现实世界多领域多场景的条件下，标注往往是受限的，不仅只能包含有限的领域，且含有的要素也未必是完备的。因此，在标注受限的条件下，实现一个泛化性强的任务型对话系统，

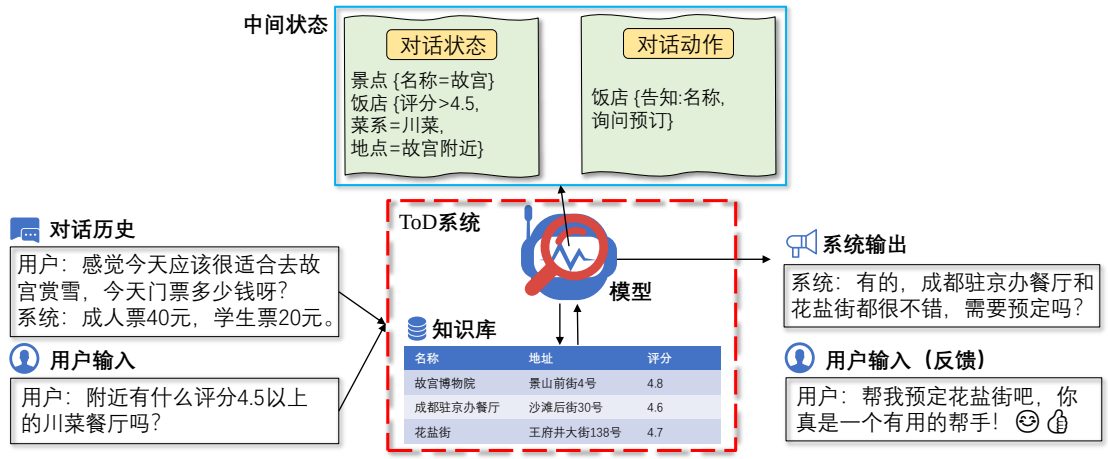


图 1-3　任务型对话系统实例

Figure 1-3　Example of a task-oriented dialog system

还存在如下三点关键挑战：

(1) 标注数据稀缺：任务型对话语料收集困难，需要标注的中间状态多，需要消耗大量的人力成本。然而对话的中间状态（槽值状态、对话动作等）对于系统回复的生成具有很好的指导作用，在不充足数据量情况下训练出来的对话系统往往难以处理已见数据以外的模式，并且随着中间状态的级联，错误累积也是不可忽视的问题。因此，该挑战要求在低资源场景下更充分地挖掘有标注数据的信息，并进行更鲁棒式的训练提升整体的泛化性能。

(2) 标注模式依赖、迁移性差：训练过程中对于对话状态的相关标注依赖于预定义的本体（Ontology）以及固定的知识库模式（Schema），使得模型的泛化性和可扩展性变差，因为当出现未见领域或数据集的时候，领域的本体，知识库的模式等都会发生变化，模型无法处理这种未见领域，并且模型与数据库的过度耦合会导致在数据库需要新增条目的时候，二者无法独立维护。因此，该挑战要求对话系统能够摆脱对于标注模式的依赖，实现预定义标注模式与模型的分离。

(3) 实时用户交互缺失：现实场景中，多轮对话评估要求用户实时对系统提供反馈。使用固定的数据集来训练对话系统并评估，一方面，会带来策略错位问

题[13]，导致对系统的评估不符合真实场景的人工评估；而另一方面模型会容易过拟合到域内的模式，泛化性能差。因此，该挑战要求引入能够覆盖多种领域的基于实时用户交互的训练与评估框架，以接近真实场景的交互。

1.3 研究目标与研究内容

针对上述三个关键挑战，本课题的研究目标依次包括提升在低资源、跨知识库、跨领域真实场景下对话系统的表现，从充分挖掘中间状态、摆脱知识库模式依赖、以及模拟实时用户反馈三个角度出发，实现标注受限条件下泛化性强的任务型对话系统。本节将详细介绍本课题的研究内容和研究目标，整体研究脉络如图1-4所示。

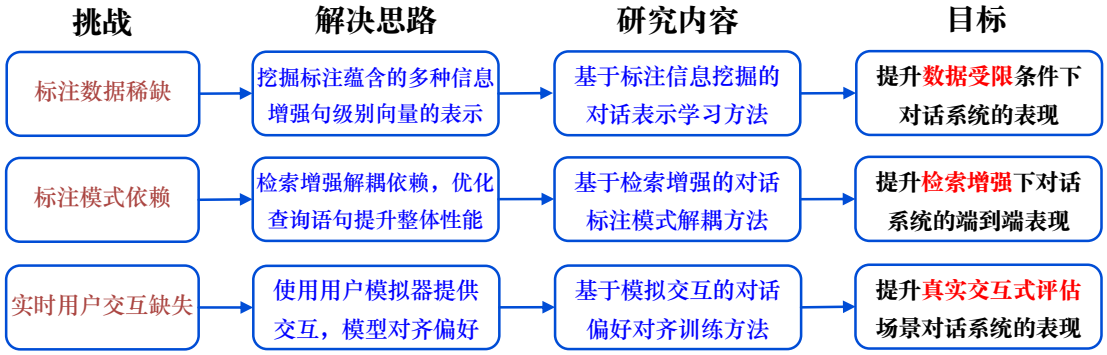


图 1-4 研究脉络图

Figure 1-4 Research framework diagram

- 任务型对话所需语料获取成本高、标注过程复杂，因此最终能获得的标注数据稀缺。为了更高效地利用少量的有标注数据，本课题提出了“基于标注信息挖掘的对话表示学习方法”来更充分地挖掘标注信息，并增强在少量标注数据上训练的鲁棒性，提升整体性能。
- 任务型对话中的槽值依赖于预定义的标注知识库模式，因此无法迁移到其他领域或场景。为了减少对这种标注模式的依赖，沿着检索增强的主流思路，本课题提出了“基于检索增强的对话标注模式解耦方法”，自动生成更好的查询语句，并且通过对回复后验信息的指导，提升检索增强端到端任务型对话性能。
- 任务型对话要求提供真实的实时用户交互来优化模型，本课题沿着目前“用户模拟器提供交互，使模型对齐偏好”的范式，思考如何进行更高效的模型训练以及提升多领域下的整体性能表现，提出了“基于模拟交互的对话偏好对齐训练方法”，补足模型在较难领域上缺乏交互的短板，持续提升模型的通用场景交互表现。

综上所述，针对标注受限下任务型对话的挑战，本课题分别提出基于以下思路的三个研究内容：（1）挖掘标注蕴含的多种信息增强句级别向量的表示（2）检索增强解耦依赖，优化查询语句提升整体性能（3）使用用户模拟器提供交互，模型对齐偏好。最终目标为实现标注受限条件下泛化性强的任务型对话系统。

下面将对解决不同挑战，实现各自目标的研究内容进更为详细的介绍。

1.3.1　基于标注信息挖掘的对话表示学习方法

针对标注数据稀缺的问题，本研究提出“基于标注信息挖掘的对话表示学习方法”，用来提升数据受限条件下对话系统的表现。首先，为了更充分地利用标注，本研究挖掘标注中蕴含的多种高层次信息，使用多任务学习的方式来增强句级别的表示。具体来说，由于句级别向量通常能够反映一轮对话的整体语义和目标[14]，本研究通过包含有多热伯努利/类别分布标签的对话状态类别、对话动作类别、槽值转移以及关键词信息构造的混合任务损失来对轮次级别的句向量进行监督，获得更好的编码器输出表示，提升整个模型的理解能力和生成能力。

此外，为了缓解动作与回复的自回归级联所带来的错误累积问题，本研究对训练时的标准动作进行计划采样，通过模拟测试可能出现的错误，来缓解曝光偏差。具体而言，本研究通过动作树的编辑距离来计算动作之间的相似度，将其作为采样概率，来模拟可能存在的错误动作，鼓励模型在扰动动作的影响下仍生成正确的回复，提升训练的鲁棒性，缓解错误累积现象，减小训练和测试阶段的不一致性。本研究所提出的方法在相关公开数据集上进行了验证，实验结果表明该方法在全量和低资源的设定下均能够取得比基线模型更好的表现，实现了当时同等条件下对比方法中的最好效果，该提升尤其是在低资源条件下更为显著。分析实验也验证了轮级别多任务学习和序列级计划采样方法各自的有效性。

1.3.2　基于检索增强的对话标注模式解耦方法

针对标注模式依赖、迁移性差的问题，本研究提出“基于检索增强的对话标注模式解耦方法”。为了减轻这种依赖，相关工作通常使用检索增强的范式来解耦知识库与生成模型训练，增强迁移性能。但是其中存在一些问题：（1）查询语句的确定：如果使用整个对话历史作为查询，则存在大量与当前无关的噪声，而如果仅使用当前语句，则很可能信息不足；（2）知识选择的多模问题：对于相同上文，候选知识集中可能在大量噪声知识中间包含有多条正确知识，而数据集中往往仅展示有限的模式或者不存在关于真实知识的标注，由此造成训练与测试不匹配的问题。基于此，本课题基于查询提示优化来提升检索增强任务型对话系统的表现。具体而言，首先，为了能够自动确定合适的查询，本研究引入一个查询提示生成模块，并遵循定向提示优化（Directional Stimulus Prompting）[15]的思路，使用回复质量相关的指标作为奖励，基于近端策略优化算法（Proximal Policy Optimization, PPO）[16] 优化查询的生成，通过指标奖励的最大化来优化出更好的查询改写。而为了解决知识标签的一对多或者缺失问题，本研究使用真实回复信息作为后验，来指导知识选择，从而实现对检索模块和语言模型的端到端联合训练。并且，为了保证训练的稳定性，上述两个过程实行交替训练。本研究所提出的方法在公开数据集上进行了验证。在公开数据集上测试的实验结果表明该方法相比于基线模型能够取得更好的表现。分析实验验证了查询提示生成模块和后验信息指导各自的有效性。

1.3.3 基于模拟交互的对话偏好对齐训练方法

针对实时用户交互缺失的问题，本研究提出“基于模拟交互的对话偏好对齐训练方法”。为了构造接近现实环境的多领域场景，本研究首先在多个数据集上分别训练出了多个数据分布具有显著差异的用户模拟器，为对话系统的训练提供多样的可交互环境。通过与用户模拟器的交互构造偏好数据，基于迭代式偏好对齐来提升模型的表现。具体而言，本研究沿着主流的大模型监督微调-对齐范式，通过监督微调使模型具有初步的跨领域能力，随后通过与多个用户模拟器的交互采集数据，在这个过程中，通过 ChatGPT 在环路的筛选与修改策略，构造适合于当前模型训练的偏好数据。根据在不同用户模拟器上表现的差异来动态分配选择与其交互的概率，通过迭代式的偏好优化算法，高效地提升模型的跨领域整体表现。除此之外，还设计了一种基于大模型思维链提示的会话级别评估方法。为了实现合理的会话级别的评估，本研究通过思维链（Chain-of-Thoughts, CoT）[17] 提示的构造来引导 ChatGPT 从多个维度对整体对话的质量进行可靠的评分。在这种评估标准下，实验结果表明了本研究提出的方法能够持续提升任务型对话系统的综合能力，在所有用户模拟器测试环境下都能够超过仅在相同环境下训练出的系统。尤其是在比较难的用户模拟器上，该方法带来的提升更为显著。分析实验探究了部分实验设定对于结果的影响。人工评估表明了与本研究提出评估体系的一致性，支撑了本研究的基础。

1.4 论文结构

本文共分为六章，章节的具体内容如下：

第一章首先介绍任务型对话系统的研究背景以及其重要研究意义，并进一步指出了在标注受限的条件下构建任务型对话系统所面临的重大挑战，最后介绍了本课题针对所述挑战的解决思路以及研究目标，并阐述了具体的研究内容。

第二章主要梳理了任务型对话的主流方法与国内外研究现状。首先介绍了任务型对话的主流建模方式，分别从流水线和端到端两种方式展开介绍了代表性工作。其次，针对目前的挑战，分别介绍了现有相关工作中的解决方案。

第三章研究了如何充分利用稀缺的标注数据的问题。首先阐述了任务型对话的中间状态标注所蕴含的高层次信息，并分析了目前级联式生成中存在的错误累积问题；其次介绍了本研究的基础模型结构；然后详细介绍了本研究设计的轮级别多任务损失目标，动作树的定义以及如何基于动作树进行计划采样；随后介绍了实验所用到的数据集、实验细节等，最后展示了实验结果与分析。

第四章研究了如何摆脱标注模式依赖的问题。首先阐述了标注模式依赖的问题，指出目前相关方法的不足之处；然后详细介绍了本研究所基于的模型结构、训练方法以及损失函数；其次介绍了本实验所用到的数据集、实验细节以及评价指标等，最后展示了实验结果与分析。

第五章研究了如何让模型从实时用户交互中学习偏好对齐的问题。首先阐

述了目前任务型对话系统跨领域性能差的问题，并说明目前评估指标的不合理之处。其次展开介绍了本文的相关工作，以及所基于的大模型偏好对齐新技术的相关背景；然后详细介绍了本研究提出的整套方法框架，包括自动评估方法、构建用户模拟器、偏好数据构建流程、以及迭代式训练过程；其次介绍了所使用到的数据集、实验细节等，最后展示了实验的结果与分析。

第六章对上述研究工作进行了总结，指出本文的主要贡献和创新点，最后对任务型对话系统的未来探索方向与发展进行展望。

第 2 章　研究现状与发展趋势

2.1　任务型对话主流建模方法

任务型对话系统的目的在于辅助用户完成多个领域的具体目标，例如购票预订、日程规划、媒体播放等。通常以多轮对话交互的形式进行，每轮接收自然语言形式的用户输入，确定系统动作并执行，将该动作及其对应的返回结果以自然语言的形式重新整合返回给用户。Levin 等[18] 第一次提出使用马尔可夫决策过程（Markov Decision Processs, MDP）建模任务型对话，将交互过程中获取的知识定义为马尔可夫过程的状态（State），将每一轮对话交互建模成状态转移。随后 Roy 等[19] 为了应对更复杂的语音对话系统场景，进一步构建了基于部分可观测的马尔可夫决策过程（Partially Observable Markov Decision Process, POMDP）模型。通常认为基于上述过程的任务型对话系统包含四个子模块，自然语言理解（Natural Language Understanding, NLU）、对话状态追踪（Dialogue State Tracking, DST）、对话策略学习（Dialogue Policy Learning, DPL）和自然语言生成（Natural Language Generation, NLG）[20]。随着深度学习的发展，基于神经网络的任务型对话建模逐渐取代了原先的统计机器学习，大体可分为两类方法，一类是流水线式的方法，模块划分与前所述一致，其中对话状态追踪与对话策略学习模块又通常合称为对话管理模块，是该类方法的核心部分。如图 2-1所示，（1）**自然语言理解**模块把用户的自然语言输入转化为结构化语言形式的输出，一般涉及领域分类、意图分类和语义槽填充。图中所示为，用户输入“找一家中关村附近的湘菜店”，自然语言理解模块的解析结果。当前的领域为“餐厅”，用户意图为“查找餐厅”，所包含的语义槽为“地点”和“菜式”，槽值分别为“中关村”和“湘菜”。（2）**对话状态追踪**模块记录和更新对话过程中所涉及的槽位名及其相应的槽值。例如在餐厅领域，用户可以利用餐厅的“地点”、“人均消费”和“菜式”来检索心仪的餐厅，那么该模块需要记录这三个属性对应的值，并且每一轮进行更新，作为当前轮次的对话状态，这个对话状态将用来与后台数据库交互。（3）**对话策略**模块将对话状态与上个模块返回的数据库状态作为输入，在策略框架下，选择下一步的系统动作。如图中，系统确定的行为为推荐，推荐的槽位为名称，对应的槽值为“望湘园”。（4）**自然语言生成**模块将对话策略模块选择的动作转化为自然语言形式的回复对用户输出。如图中，系统的回复为“望湘园这家餐厅挺不错的”。而另一类是端到端的任务型对话系统，直接通过一个统一的模型学习到从对话上下文到系统回复的映射关系，错误累积现象相较于流水线方法有所缓解，且统一模型的微调更容易，因此在成为近些年学术界研究的主流。此外，随着 ChatGPT 的出现，大语言模型强大的语言理解和生成能力开始改变自然语言处理领域的任务范式，研究者们开始将原先小模型全量微调的方式转移到利用大模型进行参数高效的指令微调，使用单一的基座模型，进行微调阶段的适配过后，能够同时完成多种不同的任务，取得了远超于之前的性能，尤其是

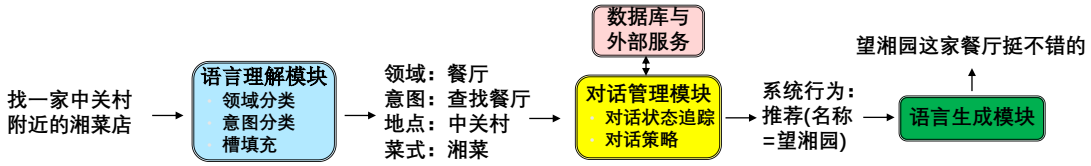


图 2-1　基于流水线的任务型对话系统
Figure 2-1　Pipeline-based task-oriented dialogue system

在一些更为高阶的复杂推理任务上，大模型的涌现能力使之取得了从 0 到 1 的突破，接下来将首先分别介绍两类任务型对话的主流建模方法对应的相关工作，然后介绍目前主流的一些大模型。

2.1.1　流水线式方法

自然语言理解　自然语言理解（NLU）模块是对话系统的关键组成部分，其主要任务包括领域识别、意图识别和槽位填充。这三个子任务在对话系统中扮演着至关重要的角色，它们负责解析用户输入的文本，并从中提取出对话系统需要的关键信息，以便系统能够准确理解用户的意图并做出恰当的回复。在自然语言理解领域，学术界和工业界一直在不断探索新的方法和技术来提升 NLU 模块的性能和效率。在过去的几年里，已经涌现出了许多重要的研究成果和技术创新。Xu 等人[21] 提出了一种基于卷积神经网络（CNN）的方法，将意图检测和槽位填充两个任务进行联合建模。他们利用 CNN 作为特征提取器，使得槽位填充器与意图分类器能够共享提取的特征。通过采用全局归一化的条件随机场（CRF），他们成功克服了局部归一化可能导致的标签偏斜问题，并且实现了意图标签的预测与槽位填充的同步进行，这一方法为自然语言理解任务的联合训练提供了新思路。Hakkani 等人[22] 则探索了在多领域场景下利用带有长短期记忆单元（LSTM）的双向循环神经网络（Bi-RNN）进行自然语言理解的方法。他们的工作突破了单一领域的限制，使得对话系统能够更好地适应多样化的对话场景。Liu 等人[23] 引入了注意力机制到循环神经网络中，通过这一机制，系统能够更加集中地关注对意图分类和槽位填充任务特别重要的部分，从而提升了 NLU 模块的性能。Goo 等人[24] 则通过考虑槽位与意图之间的关系，提出了一种新的方法，即利用槽位门来学习意图注意力与槽位注意力向量之间的关系。通过在全局优化过程中充分考虑这种关系，他们实现了更好的准确率。另外，Zhang 等人[25] 采用了胶囊神经网络，并结合了囊间动态路由算法，以更有效地利用词语、槽位和意图之间的层次信息，从而进一步提升了自然语言理解模块的性能和鲁棒性。综上所述，自然语言理解模块的不断创新和进步，为对话系统的发展提供了坚实的基础，也为未来更加智能和人性化的对话交互奠定了重要的技术基础。

对话状态追踪　在对话系统中，对话状态追踪模块扮演着关键的角色，它将从用户输入中提取的信息编码成一组紧凑的对话状态，其中包括了各种槽位及其相应的值，以及用户对系统的各种限制和请求。这个模块的输出直接影响到对话策

略的选择，因此其设计至关重要。然而，对话状态追踪模块也面临着诸多挑战，尤其是在自然语言理解和语音识别等模块出现错误时，可能导致系统无法正确理解用户意图。为了应对这些挑战，近期的研究多将传统的语言理解模块与对话状态追踪模块结合成一个联合模块，以便更直接地从用户语句中提取对话状态。

相关工作主要可分为两大类：基于值的分类模型和基于抽取或广义词表的生成模型。前者性能通常较好，但无法有效处理未登录词，而后者由于生成空间较大，性能相对较差。在基于本体中预定义好槽值的研究中，通常需要构建值的分类器。Zhong 等[26] 提出的 GLAD 模型利用全局模块学习不同槽位的共享参数，同时使用局部模块学习特定槽位的特征。Lee 等[27] 的 SUMBT 模型则采用注意力机制，基于上下文语义向量学习领域-槽位以及句子中槽值的关系，并以一种非参数化的方式预测槽-值标签。Shan 等[28] 的 CHAN 模型则以分层方式考虑词级别和句级别的信息，并通过调整不同槽值的权重来解决槽位不均衡的问题；Chen 等[29] 的 SST 模型则将槽位之间的关系建模成图神经网络，并通过图注意力机制提升性能。

基于开放词表生成的相关工作通常使用抽取式阅读理解的模式或是采用带有复写机制的指针生成网络来展开。Wu 等[30] 提出的 TRADE 模型属于这个分支开创性工作，通过从语句中的复制机制以及槽位门的状态预测来降低对于本体的依赖，以及提高不同域之间参数的共享。Gao 等[31] 提出将对话状态追踪建模成一个阅读理解问题，通过注意力机制直接抽取对话中的词作为答案。Zhou 等[32] 提出的 DSTQA 模型将对话状态追踪建模成问题回答任务，将询问槽值作为问题，并且引入一个动态变化的知识图谱来表征不同槽位之间的关系。Kim 等[33] 提出的 SOM-DST 模型将槽值的生成分为两个阶段，首先预测状态的转换类型，然后对于更新值的情况才进行生成，拆分成两个更为简单的任务使两个任务都能完成的更好。Heck 等[34] 提出的 Tripy 通过对于不同内容的复写机制来摆脱对于候选值列表的依赖，更好地解决了在跨领域和开放词表环境下的问题；Guo 等[35] 则通过在不同槽位填充时动态地选择相关的对话上文内容来减少无关信息的干扰，从而提升整体的性能。Wang 等[36] 提出的 JoDeM 模型则考虑到多轮对话中存在的错误累积现象，通过对话整体决策以及对比更新的方式，来缓解过去对历史轮次正确标注的假设所带来的测试性能下降问题。

对话策略学习　对话策略模块的性能直接决定了人机对话系统是否能够成功完成任务，从而影响用户体验。传统的手动设计对话策略规则虽然存在，但它们往往十分复杂且缺乏灵活性。因此，强化学习成为了一种有效的方法，在动作序列决策中得到了广泛的应用。在强化学习中，机器与用户的交互被视为智能体与环境的交互过程。用户的语句作为观测状态，而智能体则从动作空间中选择一个动作，并生成相应的回复。对话质量或成本被作为奖励信号，这就是强化学习在人机对话系统中的动机所在。William 等[37] 提出的 HCN 框架采用策略梯度方法来学习对话动作的选择，将学习范式从监督学习扩展到了强化学习，更好地利用了

与用户交互的语料。Su 等[38] 提出了一种高效采样的演员-评论家（Actor-Critic, AC）算法，通过使用置信域来控制学习步长，防止灾难性的模型变化，并选择最陡峭的梯度递增方向来加速收敛。此外，他们通过在线训练前的预训练来解决了冷启动问题。Chen 等[39] 的结构化对话策略学习基于图神经网络，其中图结构根据域本体定义，每个节点可视为一个子智能体，在决策时它们之间进行信息流动。Shu 等[40] 提出的 gCAS 模型针对于多个动作的策略，相对于原先的模型提升了表达能力。Takanobu 等[41] 提出的 GDPL 框架基于对抗强化学习逆过程来联合学习奖励的预测和策略的优化。Zhao 等[42] 提出的 LaRL 模型创新地将智能体的动作空间视为隐变量，并使用无监督的方式从数据中直接推导其动作空间。Takanobu 等[43] 提出的 MADPL 模型将用户与系统都视为智能体，并使用多智体学习方法。他们还提出了一个混合价值网络来解耦奖励，以更好地对全局信息和角色特定信息进行建模。

自然语言生成 自然语言生成模块的任务是把对话策略模块输出的抽象对话动作转换为句法合理、语义准确的自然语言回复，因此需要有比较好的上下文连贯性，话题的一致性，表达目的的覆盖性。这里的生成和一般意义上的语言生成主要是多了限制条件，需要在对话动作这一条件下来进行受限的生成。由于二者主体的相似性，基于深度学习的自然语言生成模块通常也会使用到具有普适性的编码器-解码器作为基础框架，对于其中的组件进行修改。此外，在大语料库的前提下，也可以使用检索式对话模型，从候选的回复中选择最合适的回复。也可以将检索式与生成式相结合，生成式对话产生的候选集合通过检索式模型进行评估排序，从而选出最佳结果，能够提升用户的满意度[44]。由于动作空间到回复映射的单一性，单独拿出这个模块，使用基于规则的方法已经能够比较好的解决，所以基于深度学习的方法相对较少。

Wen 等[45] 提出的 SC-LSTM，是长短期记忆单元 (Long-Short Term Memory, LSTM) 的一个变种，它增加了额外的能够控制生成结果的语义信息的组件，称为读门，可以起到规划句子的作用。此后，这个方法又被其他的工作所扩展，例如 Wen 等[46] 探索了多领域学习的方法来减少生成器训练所需要的数据量，Su 等[47] 提出了分层结构来利用语言模式，从而进一步改善生成的结果。

2.1.2 端到端式方法

通常来说，流水线式的任务型对话系统中每个模块被单独优化和设计。然而这样往往导致更为复杂的系统设计，并且单个模块性能的改善并不一定能反映到整个对话系统的改善中[48]，因此，近年来有越来越多的工作将多个模块联合优化，或者直接使用端到端的统一模型结构来完成多种任务。虽然内部可能会有不同的设计，但这种方式的特点是，输入为用户语句，输出为系统回复。表面上看，实现的是从用户语句到系统回复的映射，这个映射可以通过多步的中间映射来完成，但是这些映射的训练与优化都是同步进行的。端到端的任务型对话从最初的序列到序列 (Seqence-to-Sequence, Seq2Seq) 模型，到基于记忆网

络 (Memory Networks, MemNN)，到预训练语言生成模型 (Pre-trained Language Models, PrLMs)，涌现出越来越多的相关工作。接下来将分别介绍这三个方面的相关工作。

序列到序列生成模型　序列到序列生成模型的概念源于机器翻译领域，它是一种强大的神经网络架构，能够学习将输入序列映射到输出序列。这种架构最初被广泛应用于机器翻译，其中编码器负责将源语言句子编码成低维稠密表示，而解码器则利用这一表示来生成目标语言的翻译。然而，这种模型不仅限于翻译任务，在对话系统中也发挥着重要作用。

Lei 等[49] 提出的 Sequicity 框架，是这一方向的经典工作之一，它引入了信念跨度的概念，用于跟踪对话状态，使得序列到序列模型能够成功应用于任务型对话系统。该框架首先根据上下文信息生成信念跨度，然后利用数据库结果和信念跨度来生成回复。这两个步骤都使用了相同结构的序列到序列生成模型。此外，该工作还引入了两阶段复制网络（TSCP）作为框架的一种实现，大大简化了模型复杂度，加快了训练速度，并在多个评估指标上取得了领先的结果。Zhang 等[50] 提出的域感知多解码器网络（Domain Aware Multi-Decoder Network, DAMD）基于相同的上下文可能对应到多种合理回复的出发点，它基于域感知的多解码器网络。该模型的出发点是，相同的上下文可能对应多种合理的回复。为此，他们首先提出了一种多动作数据增广的框架，然后实现了 DAMD 模型来利用这一策略。该模型包含多个级联的解码器，分别用于解码对话状态、系统动作和系统回复。在对话策略多样性以及回复多样性和合理性方面，DAMD 模型都取得了良好的结果。Le 等[51] 提出的 UniConv 模型，采用了二级状态追踪器来学习槽位级别和领域级别的信息，以及一个对话动作与回复联合生成器。通过广泛使用自注意力和交叉注意力机制来整合信息，UniConv 模型在端到端任务上表现出了出色的性能。

记忆网络　记忆网络最初由 Facebook AI 在 2014 年提出[52]，先前一些处理序列数据的模型缺乏对于很长期记忆的读写能力，例如循环神经网络的记忆仅通过隐状态和权重来编码，所以记忆容量较小，并且很难记住长远上文中提到过的事实，因此该工作中使用记忆组件保存场景信息，以实现长期记忆的能力。如图 2-2所示，它通常包含有记忆组件和另外四个模块，分别是输入映射、更新模块、输出模块以及回复模块，大致的过程就是首先将输入转换到状态空间的表示，接下来根据这个表示更新记忆单元，随后计算输出特征，最后将输出特征解码成最终回复。在此基础上赋予其端到端的特性就有了一个后续工作的基础实现[1]。如图 2-3所示，问题使用注意力机制和整个记忆模块进行交互，选出相关的记忆后加权求和获得输出向量，最后和问题放在一起，预测答案。并且，通过多轮的更新，还能实现多跳推理的功能。

由于任务型对话的数据库以及上下文同样具有需要被记忆的特性，所以基

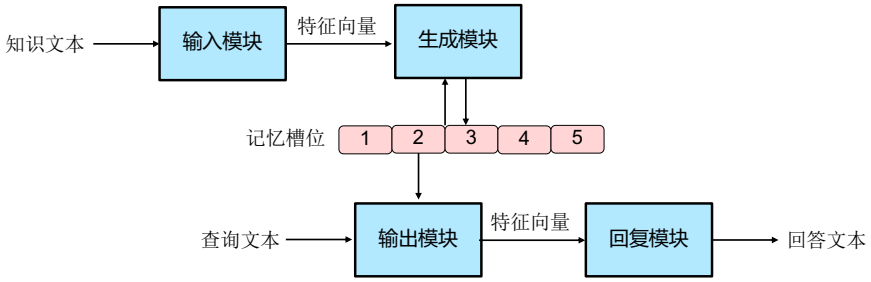


图 2-2　记忆网络基本结构

Figure 2-2　Framework of memory networks

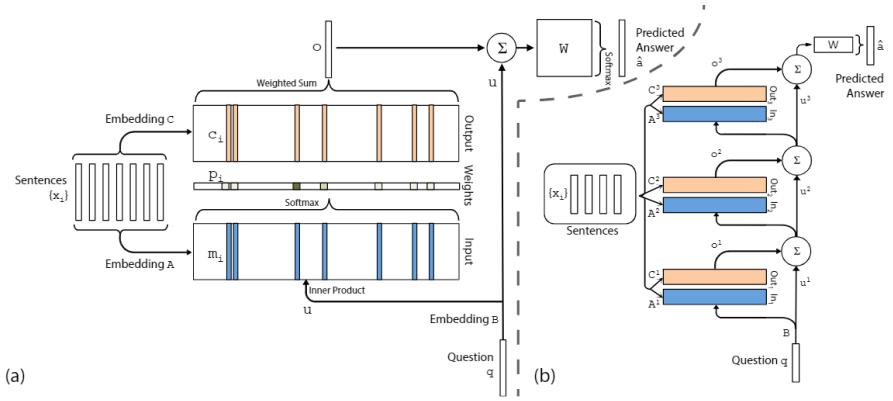


图 2-3　端到端的记忆网络[1]

Figure 2-3　End-to-end memory networks

于记忆网络的端到端任务型对话也涌现出了一批相关工作。Eric 等[53] 第一次构建了 InCar 数据集，并提出对整个知识库进行注意力操作，然后将知识库中实体的注意力得分增加到最终生成概率分布中，从而使整个模块能够生成实体。Maddoto 等[54] 提出的 Mem2Seq 模型是第一个将指针网络的想法和记忆网络上的多跳注意力机制结合起来的工作，每一步通过对于记忆单元的多跳注意力机制，学习记忆单元之间的关系，进而实现对于生成结果的控制。它较好地解决了先前工作没有解决好的两个问题：一是如何将外部知识库加入到循环神经网络的隐状态中，二是原先记忆网络的模式只能直接选择一个回复而不能逐词生成。Wu 等[55] 提出的 GLMP 模型是对 Mem2Seq 的改进，通过一个全局的记忆编码器和局部的记忆解码器来共享外部知识，先生成回复模板，然后使用全局编码器中得到的记忆指针来过滤外部知识，通过局部记忆指针来实例化回复模板中的未填充槽位值，提升了复制准确性，改善了未登录词的问题。Qin 等[56] 提出的 DF-Net 模型考虑到先前模型对于新领域的泛化性问题，通过一个共享-私有网络来学习不同领域的共有知识和独立知识，并且通过一个动态混合网络来自动探索目标域和已有域的相关性，用较少的实验数据，在多领域数据集上取得了较好的结果。

预训练语言模型　预训练语言模型的基础是谷歌团队 Vaswani 等[2] 提出的 Transformer，这是在自然语言处理领域一个划时代性的工作，因为它抛弃了原先占据

主流的循环或者卷积神经网络的模式，转而使用一种仅仅依靠注意力机制的新的简洁的神经网络架构。它能够关注到更为长远的信息，克服了卷积神经网络感受野局限的问题；并且注意力机制使得并行化成为可能，克服了循环神经网络不能并行化的缺点。单层运算时间的减少带来了可多层堆叠的结构，在机器翻译的任务上超过了当时的其他最好结果。它的结构如图2-4所示。虽然 Transformer 起源于机器翻译领域，但是由于它简洁统一效率高的特点，基于它并在大规模预料上预训练过后的模型（如 Bert[57]、T5[58]、BART[59] 和 GPT-3[60] 等）在 NLP 各大文本理解和文本生成任务上也取得了突破性的进步，在下游任务上对预训练语言模型进行微调成为了新的范式。Budzianowski 等[61] 第一次将 GPT 运用

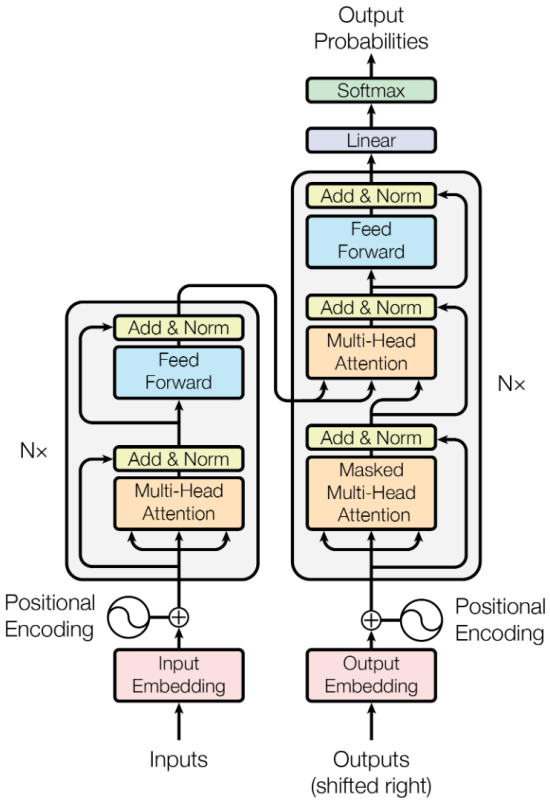


图 2-4　Transformer 模型结构[2]

Figure 2-4　Model framework of Transformer

到任务型对话的回复生成中，它将对话状态，数据库状态，对话上文以及回复这四者以纯文本的形式拼接起来，在回复的部分以自回归语言模型的目标来进行 GPT 的微调。Hosseini 等[62] 提出的 SimpleTOD 模型将所有任务型对话中的子任务（对话状态追踪、对话策略、回复生成）通过一个简单的单向语言模型一起完成。它直接将上文、对话状态、数据库状态、对话动作以及回复当作一整个序列自回归地训练。Peng 等[63] 提出的 Soloist 模型抛弃了对于系统动作的生成并引入对比学习和大规模异质语料库上的预训练来提升少样本学习的迁移性能。Yang 等[64] 提出的 UBAR 模型使用会话级别的训练，更接近实际场景的评估。Kulhanek 等[65] 提出的 AuGPT 模型对训练目标做修改并且通过回译（Back Translation）的方式进行数据增广。

2.1.3 大语言模型

通常，将参数量在十亿以上的模型称为大模型，OpenAI 在 2020 年提出的 GPT-3[60] 参数量为 175B，验证了超大规模的语言模型具有良好的少样本情境学习能力，揭示了单一的大模型本身能够通过示例的方式来完成多种多样任务的可能性。沿着这一思路，通过对奖励模型的训练以及其指导的人类偏好对齐技术[66]，ChatGPT 在 2022 年问世，其自然流畅的对话与指令遵循能力，以及涌现出的思维链和复杂推理能力，创作能力，角色扮演能力等，开始引发世界的瞩目，使其不仅仅停留在学术研究上，非专业人员也能够得心应手地使用，成为当今时代人们的新生产力工具。大模型的出现，引发了后续的广泛研究，从训练数据的角度，有些工作关注如何获取高质量的数据[67–69]，有些工作则关注数据配比的问题[70]。从基座模型的角度，目前学术界最常用的是 Llama 系列模型[71,72]，除此之外，国内外的研究机构和公司也提出了自己的一些基座模型[73–77]。从训练方法的角度，为了缓解大模型所需消耗大量算力和显存资源的问题，一些参数高效微调[78–80] 的工作也被提出，使得学术界能够更好地参与大模型相关的研究。

2.2 任务型对话系统的挑战及研究现状

基于第1章所提到的挑战，接下来的部分将分别介绍目前国内外针对这些挑战所提出的解决方案，从而整理出目前的研究现状。

2.2.1 标注数据稀缺

通常在检索式对话、知识型对话或是机器翻译领域相关的数据集中，高资源场景能达到 1M 规模的数据量，而如表2-1所示，任务型对话数据量仅在 $10^4 - 10^5$ 量级。为了在有标注数据稀缺的情况下提高任务型对话系统的性能，最直接的方

表 2-1 检索式[7]（左）和任务型对话[8]（右）数据集规模对比

Table 2-1 Comparison of retrieval-based (left) and task-oriented (right) dialog datasets

	train	val	test
# context-response pairs	1M	50k	10k
# candidates per context	2	2	10
# positive candidates per context	1	1	1.18
Min. # turns per context	3	3	3
Max. # turns per context	98	91	45
Avg. # turns per context	6.69	6.75	6.45
Avg. # words per utterance	18.56	18.50	20.74

Type	Single-domain goal					Multi-domain goal		
Dataset	DSTC2	WOZ 2.0	Frames	KVRET	M2M	MultiWOZ	Schema	CrossWOZ
Language	EN	EN	EN	EN	EN	EN	EN	CN
Speakers	H2M	H2H	H2H	H2H	M2M	H2H	M2M	H2H
# Domains	1	1	1	3	2	7	16	5
# Dialogues	1,612	600	1,369	2,425	1,500	8,438	16,142	5,012
# Turns	23,354	4,472	19,986	12,732	14,796	115,424	329,964	84,692
Avg. domains	1	1	1	1	1	1.80	1.84	3.24
Avg. turns	14.5	7.5	14.6	5.3	9.9	13.7	20.4	16.9
# Slots	8	4	61	13	14	25	214	72
# Values	212	99	3,871	1,363	138	4,510	14,139	7,871

式就是进行数据增广，相关工作有[65,81–85] 等，这些工作通过话语改写、反事实样例构造、回译、使用模拟器合成伪数据等方式来进行数据增广，希望训练数据多样性的增强能够提升模型在测试数据上的表现，但是很难控制增广数据的质量，并且仍然没有解决如何获取高成本的中间状态标注这个问题；另一类方法则尝试利用丰富的无标注任务型对话文本语料，使用一些自监督或者半监督的目标，例如 He 等[86] 收集少量的有监督和大量的无监督对话数据，沿用 R-drop[87] 的思路来学习到更好的对话动作表示；一类方法[88,89] 使用对比学习训练得到更

好的上下文信息以及对话状态的表示；Su 等[90] 则直接将多个任务型对话数据集混合起来对 T5 模型基于自回归文本生成的目标进行持续预训练，使其具备更通用的适配任务型对话下游任务的初始能力。以上方法均可归结为在不同的阶段引入了一些额外的数据，而另外的一类方法，是假定在不构造或引入任何额外数据的情况下，尝试更充分地利用标签或是提升训练的鲁棒性，部分工作[65,91,92]使用一些额外的任务（例如序列标注、回复选择等）与原先的语言模型目标一起构成多任务学习的训练，提升任务型对话系统的性能；还有一些工作尝试使用强化学习优化一些离散的评估指标或是回复生成中文本的关键词信息[93–95]。

2.2.2　标注模式依赖

不同数据集以及领域对应的知识库模式不同，而对话状态的槽值预测通常基于预定义的标注模式，因此这种绑定导致了系统的迁移性能差。因此，模块化任务型对话的训练中存在对预定义的知识库标注模式依赖的问题，主要原因是在对话状态追踪模块中学习固定的数据模式，从而导致迁移性差，训好的模型无法基于新模式的知识库进行测试。此外，如果测试时直接生成完整回复，则难以处理训练过程中的未见数据库条目；而使用去词汇化的回复会导致模式固定，对话的灵活性差。因此，目前的研究思路通常将数据库[1]与模型进行分离，采用检索增强的范式来进行端到端任务型对话的建模。主要存在三类方法：（1）将数据库编码成一个记忆网络，并且基于对话上下文的表示来进行查询和更新[54,56,96]；（2）将知识直接编码在模型的参数中，生成过程中不进行显式的检索，而是相当于在基于模型参数的运算中“隐式地”进行了检索[97,98]。这类方法简化了流程，混合检索与生成两个步骤，但是可扩展性差，在面对大规模知识库的时候使用知识的准确率降低；（3）使用预训练语言模型将序列化的知识库记录与对话上下文编码在一起[99–104]，它们将筛选后的知识库子集与对话上下文拼接或是表示融合起来后，借助解码器生成对应的回复。ChatGPT 问世以来，逐渐有工作尝试使用 ChatGPT/GPT4 等大模型，此时一般不会对大模型进行训练，侧重点在于对提示的设计和任务的分解上[105,106]。

2.2.3　实时用户反馈缺失

现实场景中，多轮对话要求用户实时对系统提供反馈，这是多轮对话人工评估的要求。而基于固定数据集的自动评估则会存在策略错位问题，导致模型在真实场景下的能力无法得到正确评价[13]。为了更好地填补该空缺，目前主流的做法是基于用户模拟器来提供交互式环境以及人类评估的替代，并采用强化学习来训练。在目前的研究工作中，一类方法侧重于如何构造更真实的用户模拟器[107–111]，通过考虑用户目标、情感、满意度以及引入大模型情境学习等方式，逐步丰富用户模拟器的多样性与真实性；另一类方法则侧重于如何优化基于用户模拟器的强化学习训练[5,112–114]，从寻找更合理的奖励（例如用户满意度）或者更合理的训练策略（例如多模拟器混合提升泛化能力）的角度来提升任务型

[1]由于与知识型对话的相似性，此后也称此处的数据库为知识库。

对话的表现。但是目前的方法仍然没有很好地解决跨领域/数据集的泛化性问题，模型无法迁移未见领域或适配不同数据集所对应的数据模式。

2.3 任务型对话系统发展趋势

从以上的研究进展中可以看出，任务型对话从过去的流水线式过渡到如今主流的端到端式系统，并且从原先单一领域的有监督学习，逐步过渡到为多领域多场景，交互式环境下的奖励反馈强化学习。随着预训练语言模型性能的巨大提升，单一领域的问题已基本解决，人们更关注泛化性，因此标注受限场景下的跨领域、跨数据集迁移性能受到研究者的更多关注。整体上，当前针对任务型对话优化主要有以下三个出发点：

- **充分挖掘标注信息**：由于任务型对话从语料收集到中间过程标注都很困难，少样本情况下训练的模型很难达到泛化性强的要求。为了在这种标注数据稀缺的情况下，提高对话系统的表现，之前的方法主要通过数据增广的手段，然而增广数据的质量无法得到保证。因此近年来的趋势是基于强大的预训练语言模型，尝试通过多种目标来充分挖掘珍贵的标注信息以及大量的无标注数据，例如多任务学习、半监督/自监督等。此类方法的核心是学习到更好的中间状态表示，从而为后续的生成助力。此外，由于低资源场景下级联中间状态错误累积的存在，提升训练的鲁棒性，缓解错误累积也是一个亟需解决的问题。
- **减少对标注模式的依赖**：现有任务型对话中对于对话状态的标注模式依赖于知识库，在不同数据集与领域之间不共享，使得模型不具备可迁移性。目前解决方案的核心在于抛弃对于对话状态的预测，而是直接根据上下文检索相关的知识库条目，一方面使得对话状态的标注不再必须，另一方面也成功解耦了知识库和模型，使二者的分别维护成为可能。其中存在两种实现方案：记忆网络和检索增强的预训练模型。前者是过去的处理方式，目前后者已成为主流，并且也成为大模型实际应用中十分重要的方向。
- **引入模拟实时用户反馈**：固定数据集评测与 “金标准” 人工评测存在的不一致性是人机对话领域普遍存在且不容忽视的问题。相较而言，任务型对话具有目标明确这一优势，能够构建出更好的用户模拟器，以此代替实时的用户反馈。目前相关的研究工作主要围绕 “如何构建更好的用户模拟器” 以及 “如何在交互中提升系统性能” 两个角度展开。就前者而言，目前的工作主要考虑引入多种用户特征来提升用户模拟器的多样性，使其越来越接近真实生活中的用户；而后者目前的工作中主要基于强化学习展开，强化学习本身存在的问题则成为了适配过程中研究者们关注的重点，例如奖励的设计、多用户场景下的连续强化学习、更稳定的训练等。

综上所述，在任务型对话系统的研究中，近年来的趋势主要是使模型的训练过程减少对于标注数据、标注模式以及实时用户反馈的依赖，增强任务型对话系统在标注受限条件下的泛化表现。

第 3 章　基于标注信息挖掘的对话表示学习方法

任务型对话系统由于对话预训练技术的进步而取得了显著的进步。然而，目前仍然存在两个重要的挑战。首先，大多数系统主要利用最后一轮的状态标签来训练解码器的自回归生成。这种做法忽视了状态标签在提升模型理解能力从而助力生成的全面价值。其次，对于预测动作的过度依赖通常会导致错误累积，从而在遵循不正确的动作条件下生成不合理的回复。

为了应对这些挑战，本研究提出了轮次级别的多任务目标来学习编码器。通过挖掘中间状态标注中的高层次信息，指导编码器为语义理解和回复生成输出语义更为丰富的表示。而对于解码器，本研究引入了基于动作树的计划采样技术。具体来说，本研究将层次化的动作建模为树结构，并利用树之间的相似度来基于计划采样得到负样本动作，希望模型在动作标签的扰动下仍然生成正确的回复。该方法通过负例的采样来模拟预测阶段模型可能生成的错误，缩小任务型对话在训练和推理之间的差距。

在同等条件的无持续预训练方法中，该方法在 MultiWOZ 数据集系列上取得了最好的性能，并且即使在加入了持续预训练的方法中，也具有一定的竞争力，验证了该方法对于低资源场景的有效性。

3.1　引言

任务型对话的目标是通过多轮对话更好地完成用户特定任务。如图3-1所示，典型的任务型对话系统包括四个模块：(1) 自然语言理解模块（Natural Language

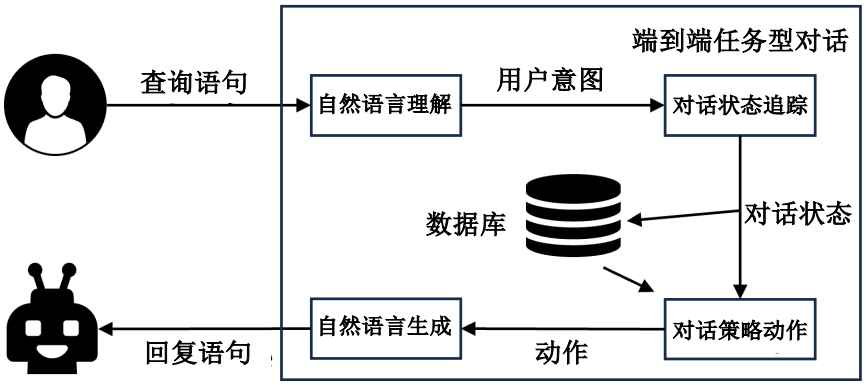


图 3-1　任务型对话系统示意图
Figure 3-1　Sketch of a task-oriented dialogue system

Understanding, NLU）确定用户意图；(2) 对话状态跟踪模块（Dialog State Tracking, DST）提取用户目标的约束条件，这些条件将用于查询数据库；(3) 策略模块（Policy）规划系统的下一动作序列；（4）自然语言生成模块（Natural Language Generation, NLG）最终生成流畅且信息丰富的回复。在目前的工作中，自然语言

理解通常不需要引入单独的额外工作，而是与对话状态追踪结合形成一整个模块[48]。端到端[1]的任务型对话是本研究工作的重点，使用同一个模型将不同环节的输出实现联合训练。

近年来，得益于强大的预训练语言模型，尤其是对话预训练，端到端的任务型对话系统性能有了显著提高。然而，仍然存在两个问题。图3-2展示了经典方法 SimpleTOD[62]，大多数方法都遵循该范式。首先，一些标注较为完整的数据

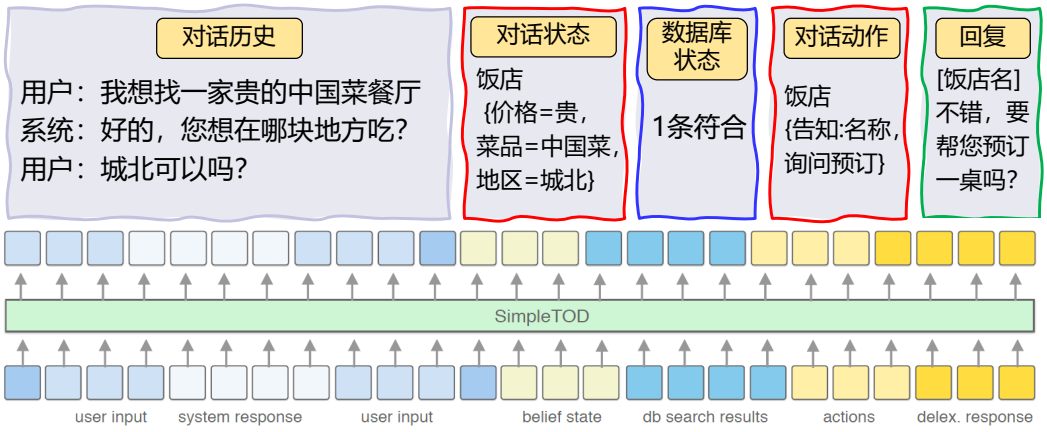


图 3-2　SimpleTOD 模型示意图
Figure 3-2　Illustration of SimpleTOD

集具有中间状态（即槽值状态和对话动作）的标注，而大多数工作只是简单地用于自回归地监督生成器，这样简单的做法忽视了这些标注中蕴含的高层次信息，忽略了其对于模型理解语义的重要价值。其次，自回归生成的训练方法会导致错误累积，已生成的不合理动作容易导向后续不合理回复的生成。

为了解决第一个问题，本研究利用中间状态的标签来监督编码器输出的隐状态，希望更好的表示能够为后续生成提供有利的线索。Lee 等人提出的 MT-TOD[91] 利用槽值状态的标注构建序列标注的辅助任务，这启发了本研究挖掘标注中蕴含的更多信息，来构建更多样的辅助任务（例如，槽值类型、槽值变化、动作类型和回复关键词的预测）。此外，Li 等人[14] 指出句级别向量能够反映对话中的更高层次信息，本研究使用上述的轮次级别信息来优化**句向量的表示**，从而使句向量能够编码更多的高层次信息，例如本轮对话的目标或是对下一个可能回复的潜在影响。

为了解决第二个问题，本研究尝试使用**计划采样**技术[115] 来减少训练和测试之间的不一致性。然而，在任务型对话系统中，简单地使用词元级别的计划采样实际上并不能很好地模拟测试阶段的错误。给定一个特定的词元，由于词元之间较强的依赖关系，生成下一个词元的概率是高度确定的。这导致了尖锐的词元级别条件概率分布，使得单个的负样例很难被采样到。然而，动作序列之间存在更多的不确定性。基于此，本研究提出了一种可以在训练时直接采样与标注动作

[1]此处的端到端有别于一般意义上的端到端，因为中间环节的输出仍然需要级联生成，并非直接从用户输入语句直接输出了系统回复，不过使用了一个统一的模型，且可以实现联合训练，因此在此处被称为端到端，本章其余部分同，将不再赘述

相似的负面动作序列的方法，称为**基于动作树的计划采样**。具体来说，受到 He 等人提出的 SPACE[88] 启发，本研究将动作形成的词序列建模为树，根据动作树之间的编辑距离计算相似度，然后使用相似度的归一化作为负例动作序列的采样分布，并优化了在对动作扰动情况下生成标准回复的概率，这也可以看做一种通过扰动来进行鲁棒训练的方法。

本研究在 MultiWOZ 2.0/2.1/2.2 上进行了全面的实验。实验表明，本研究提出的方法显着改善了任务型对话系统，在同等**不采用持续预训练**方法中取得了最好表现，综合表现相比于基线模型提升了 1.81-3.39 分。消融实验还分别验证了多任务学习和计划采样的有效性。

3.2　相关工作

端到端的任务型对话的目标是联合训练子模块并构建一个能够接受用户输入、系统回复输出的集成系统。Wen 等人[116] 首次为端到端的任务型对话提出了一种基于可训练神经网络框架的方法，在不同模块中分别使用了卷积神经网络[117] 和长短期记忆模块[118]。另外，[83,119,120] 主要基于 CopyNet 在序列到序列的训练方法和解码器的精心设计等方面提出了各自的方法。

由于预训练语言模型技术的兴起及其在自然语言处理任务中的优良表现，使用预训练模型作为下游任务的基座成为目前的标准做法，在任务型对话中使用的代表模型包括 GPT[121]、T5[58] 和 UniLM[122]。一些方法[62,65,123,124] 采用 GPT-2 模型作为不同的模块的基座模型，进行对话逐轮级或会话级的训练。由于任务型对话中既有生成任务又有语言理解任务[2]，编码器-解码器框架更加适合。有许多工作使用符合该结构的 T5 作为基础模型，并从各自的角度出发促进端到端性能的提升。其中，一部分工作[90,91,94] 设计多种多样的多任务学习目标来使用统一的训练框架基础上挖掘更多的信息；Sun 等人提出的 Mars[89] 利用对比学习来建模对话上下文与槽值状态/对话动作表示之间的关系，从而学习到更好的语义空间。还有一些方法希望在考虑理解与生成不同的同时复用参数，因此使用基于参数共享的编码器-解码器 UniLM 作为基座模型，He 等人提出的一系列工作[86,88] 在 UniLM 进行上持续预训练他们提出的半监督或自监督学习任务，然后通过微调适应下游任务，达到了当前的最先进水平。

为了缓解端到端任务型对话中的错误累计现象，Zhang 等人[83] 引入多个合理的对话动作，以学习一个更平衡的动作分布，从而引导对话模型生成多样化的回复。Sun 等人[125] 引入了一种反向去噪的重构方法，He 等人[86] 提出了正则化一致性来优化学到的表示。与上述工作不同，本研究尝试使用计划采样方法，该方法最初在序列到序列的生成任务中提出并取得一定的效果，然后在神经机器翻译中得到了改进[115,126]。

[2]即使均建模成自回归生成的方式，其体现的能力是不同的，对话状态主要是理解能力，而对话动作的确定和回复生成则主要是生成能力

3.3 方法介绍

此节将首先介绍本研究所使用的模型结构，然后详细介绍本研究提出的方法。为了缓解第3.1节中分析的标签利用**不充分**的问题，本研究提出了四个轮次级别的额外任务来加强编码器的语义理解能力，从而为后续生成提供帮助；为了缓解序列级别的错误累积问题，本研究提出了一种基于动作树的计划采样方法，从而生成更加鲁棒的回复。接下来将首先介绍多任务学习的目标，然后介绍动作树的定义和序列级计划采样方法。整体的方法如图3-4所示。

3.3.1 模型框架

本节将主要介绍该研究中所用到的模型结构。如上节所提到，端到端的任务型对话生成通常被建模成一个级联式生成的问题。如图3-3所示，在每一轮的对话中，系统首先接收用户的输入，将其与已保存的历史上下文进行拼接；接下来系统需要生成对话状态，它是一个从用户请求中解析出来的层次化语义，能够反映用户目标所定义的限制条件。对话状态将用于查询数据库，随后系统将该查询返回的结果和上下文信息综合推断出一组对话动作，它也是一个层次化的序列，并用于指导后续的回复生成过程。通常而言，对话状态包含“**领域/槽位/槽值**”，而对话动作包含“**领域/动作/槽位**”，都是层次化的三级结构。

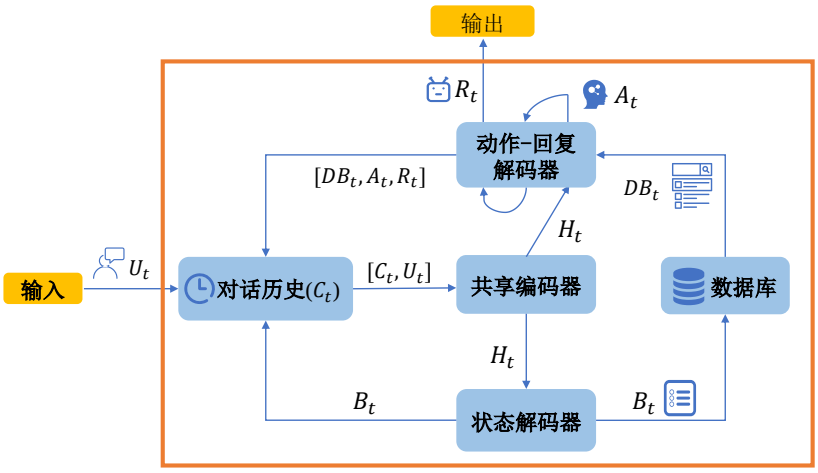
正如3.2中提到的，存在多种基座模型的选择，例如纯解码器架构的 GPT[123,124]，编码器-解码器架构的 T5[90,94]，基于 UniLM 的模型[86,88] 以及编码器-双解码器架构[91,92]。考虑到对话状态的生成更多地依赖于理解和总结的能力，而对话动作和回复生成则更多依赖于能够保持上下文一致性的生成能力。本研究认为二者属于不同的语义子空间，因此不选择纯解码器架构，并且在预实验中发现 UniLM 需要更为耗时的预训练才能够取得较好的结果，这一点也在 He 等人的工作[86]中进行了验证。最终，本研究遵循 Lee 等人的工作[91]，采用一个共享的编码器和两个不同的解码器，如图3-3所展示。

3.3.1.1 符号定义

以下介绍输入输出流中所涉及到的符号。在第 t 轮对话中，U_t 代表用户的输入语句。B_t 是对话状态，在图3-3的例子中对应 {restaurant:{pricerange:expensive,area:centre,food:Chinese}}。DB_t 代表数据库结果，它反映了满足当前对话状态的实体数目。A_t 代表动作序列，在图3-3的例子中对应着 {restaurant:{inform:[address,name], offerbook:[]}}。R_t 表示系统回复。每一轮的各环节输入输出信息 $I_t = (U_t, B_t, DB_t, A_t, R_t)$ 会与历史信息进行拼接整合，产生当前轮的上下文信息 $C_t = \text{Concat}(I_0, ...I_{t-1})$[124]

3.3.1.2 训练目标

在端到端任务型对话的框架中，历史上下文信息与当前用户输入拼接起来输入到一个共享的 Transformer 编码器中，获得隐状态 H_t。H_t 首先输入到对话



$\boldsymbol{C_t}$: Can you tell me about any expensive restaurants in the centre? [restaurant] pricerange expensive area centre [db_3] [restaurant] [inform] price choice area [request] food We have [value_choice] [value_pricerange] restaurants in the [value_area], do you have a specific cuisine in mind?
$\boldsymbol{U_t}$: Yes, I would prefer Chinese please.
$\boldsymbol{B_t}$: [restaurant] pricerange expensive area centre food Chinese
$\boldsymbol{DB_t}$: [db_3]
$\boldsymbol{A_t}$: [restaurant] [inform] address name [offerbook]
$\boldsymbol{R_t}$: I have the [value_name] located at [value_address]. Would you like to make reservations?

图 3-3　本研究所基于的对话系统框架的描述
Figure 3-3　Illustration of the dialog system framework on which this study is based

状态解码器中，生成对话状态 B_t。该对话状态将用于查询数据库，返回的结果为 DB_t。最后 H_t 和 DB_t 一起输入到一个动作-回复解码器中，依次自回归地生成对话动作 A_t 和对应的回复 R_t。

$$\begin{aligned} H_t &= \text{Encoder}([C_t, U_t]) \\ B_t &= \text{Decoder}_b(H_t) \\ A_t, R_t &= \text{Decoder}_{ar}(H_t, DB_t) \end{aligned} \tag{3-1}$$

两个解码器和编码器都使用交叉熵损失来进行优化，训练时的监督信号采用对话状态、动作以及回复的真实标签来进行教师指导。如公式(3-2)所示，对于某训练数据的第 t 轮对话而言，其中 H_t 表示对话历史，包含 $0 \sim t-1$ 轮的过往对话；$\hat{B}_t$ 表示对话状态的真实标签，即累积的槽值对描述；DB_t 表示数据库状态，使用向量表示匹配成功的条目数；$\hat{A}_t, \hat{R}_t$ 表示动作与回复的真实标签。

$$\begin{aligned} \mathcal{L}_B &= -\log P(\hat{B}_t | H_t) \\ \mathcal{L}_{AR} &= -\log P(\hat{A}_t, \hat{R}_t | H_t, DB_t) \\ \mathcal{L} &= \mathcal{L}_B + \mathcal{L}_{AR} \end{aligned} \tag{3-2}$$

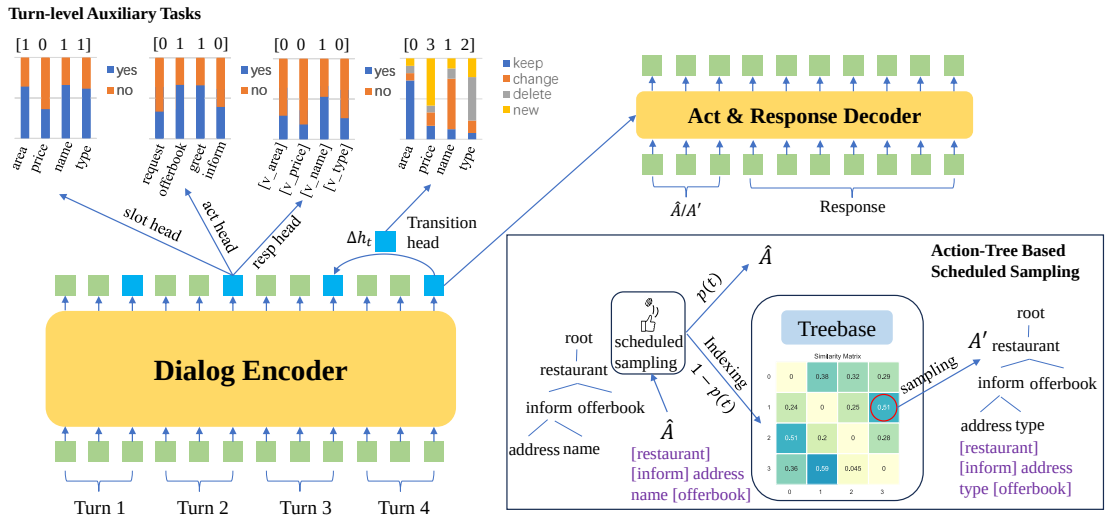


图 3-4　轮级多任务学习与基于动作树的计划采样

Figure 3-4　Turn-level auxiliary tasks and action-tree based scheduled sampling

3.3.2　轮级别的多任务学习

标注信息除了真实回复之外，还有很多其他方面的标注，它们可以用于增强编码器的理解能力。MTTOD[91] 根据对话状态的标注引入了一个简单的序列标注任务来增强任务型对话能力，本研究因此受到启发，认为可以利用更多种类的标注从而引入更多样的额外任务学习。除此之外，DialoFlow[14] 指出，轮次级别的表示能够反映高层次的信息，例如对话目标或者对下一轮回复生成的潜在影响。基于以上两点，本研究通过高层次的监督信号来监督句级别的表示，这些监督信号来自不同类型的标注，例如对话状态、对话动作和回复。通过以上目标让编码器输出更好的中间表示，助力后续解码器的生成。例如图3-4中例子的标签集合如下[3]

- Slot type: *[pricerange, area, food]*
- Slot transition: *{pricerange:keep, area:keep, food:new}*
- Action type: *[inform, offerbook]*
- Response keywords: *([value_name],[value_address])*

接下来将详细介绍这四个任务。

3.3.2.1　轮级别表示

本研究中共享编码器和两个解码器都是直接使用 T5 的对应模块进行初始化的，而 T5 本身不是本研究的关注点，因此其结构不在此介绍。根据公式(3-1)，H_t 是编码器的输出隐状态，本研究使用该隐状态序列中每一轮的最后一个位置的向量作为从第 t 轮的表示，而这里将第 0 轮到第 $t-1$ 轮所有向量的表示所组

[3]为了简化，领域标签已省略

成的矩阵，记作 T_t

$$\begin{aligned} P_t^{end} &= [pos_0, pos_1, ..., pos_{t-1}] \in \mathbb{N}^t \\ T_t &= \text{IndexSelect}(H_t, P_t^{end}) \in \mathbb{R}^{d \times t} \end{aligned} \tag{3-3}$$

3.3.2.2　槽位类型预测

确定哪些槽位出现在用户的语句有助于对话状态的生成，因为它缩小了槽值对的范围，并且这种判别式的任务与编码器的能力更为适配。对于某些对话轮次，提到的槽位往往不是单一类型，因此，遵循 He 等提出的 GALAXY[86]，本研究将槽位类别预测的任务建模成一个多标签分类问题。在公式(3-4)中，将每一轮的槽位类型标签记作 $ST = (st_1, st_2, ..., st_N)$，其中 N 为槽位类型的总数，使用多维的伯努利分布来建模槽位类型的概率。轮级别的表示 T 在经过一个多维的二元分类器之后，得到每类槽位的预测分数。

$$\begin{aligned} p(ST|T) &= \prod_i^N p(st_i|T) \\ p(st_i|T) &= \text{sigmoid}(W_{st}T) \in \mathbb{R}^N \\ \mathcal{L}_{st} &= -\sum_{i=1}^{N} \{y_i \log p(st_i|T) + (1 - y_i) \log(1 - p(st_i|T))\} \end{aligned} \tag{3-4}$$

其中，W_{st} 是一个可训练的线性分类头参数矩阵，$y_i \in \{0, 1\}$ 代表着槽位类型 st_i 是否出现在当前轮的对话状态中。

3.3.2.3　槽位变化预测

SOM-DST[127] 指出对于槽位状态变化的预测能够允许状态追踪模块仅仅需要为一小部分槽位重新生成它的值，从而提升推理效率。在本研究中，预测槽位的变化同样为对话状态的生成提供重要线索。此处，定义槽位的变化为以下四种类型：保持不变 (keep)、变化 (change)、删除 (delete) 以及新增 (new)，和数据库的操作类型很相似。由于标注中含有相邻两轮对话的槽值状态标注，获取槽位变化种类十分容易。本研究使用多维的类别分布来建模槽位变化，如公式(3-5)所示，定义槽位变化为 $SC = (sc_1, sc_2, ..., sc_N)$。相邻轮次句向量的表示差值记作 $\Delta T = T_t - T_{t-1}$，它将被输入到一个可训练的转移分类头 W_{sc}[4]获取预测分数，进

[4]为了简化，T_t 的下标已省略

一步归一化可以得到预测概率。

$$\begin{aligned} \Delta T_t &= T_t - T_{t-1} \\ p(SC|\Delta T) &= \prod_{i}^{|SC|} p(sc_i^{y_i}|\Delta T) \\ p(sc_i|\Delta T) &= \text{Softmax}(W_{sc}\Delta T) \in \mathbb{R}^4 \\ \mathcal{L}_{sc} &= -\sum_{i=1}^{N} \log p(sc_i^{y_i}|\Delta T) \end{aligned} \tag{3-5}$$

其中 $y_i \in \{0, 1, 2, 3\}$ 是第 i 种槽位的变化标签。

3.3.2.4 动作预测

GALAXY[86] 指出，识别动作类型（例如请求、提供订购等）能够更好地帮助对话策略的优化，从而提升端到端对话的整体表现。本研究使用了和 GALAXY 相同的多元伯努利分布来建模动作预测，区别在于本研究对于一个会话级别的样例预测其所有轮次的动作类型。

$$\begin{aligned} p(A|T) &= \prod_{i}^{N} p(a_i|T) \\ p(a_i|T) &= \text{sigmoid}(W_a T) \in \mathbb{R}^N \\ \mathcal{L}_a &= -\sum_{i=1}^{N} \{y_i \log p(a_i|T) + (1 - y_i) \log(1 - p(a_i|T))\} \end{aligned} \tag{3-6}$$

其中 W_a 是一个可训练的线性分类头参数矩阵，$y_i \in \{0, 1\}$ 是关于动作 a_i 在当前轮是否采取的标签。

3.3.2.5 回复关键词预测

在大多数任务型对话的场景下，系统应该基于用户的请求，在回复中告知用户一些关键信息，这也和节3.4.1中提到的成功率 (**Success**) 指标相关联。对这些关键词的预测能够使得模型更加关注于重要信息的生成，例如 *[value_name], [value_area]* 等。在去词汇化（delexicalized）过后[83]，类似于 *[value_xxx]* 这样的词语在一个有限的集合中。因此本研究使用多元伯努利分布来建模关键词的

词袋预测，如公式(3-7)所示。

$$
\begin{aligned}
p(K|T) &= \prod_{i}^{N} p(k_i|T) \\
p(k_i|T) &= \text{sigmoid}(W_k T) \in \mathbb{R}^N \\
\mathcal{L}_k &= -\sum_{i=1}^{N} \{ y_i \log p(k_i|T) + (1 - y_i) \log(1 - p(k_i|T)) \}
\end{aligned}
\tag{3-7}
$$

其中，N 是关键词的词表规模，W_k 是一个可训练的线性分类头参数矩阵，$y_i \in \{0, 1\}$ 是表明关键词 k_i 是否在回复中出现的标签。

总的来说，多任务学习的损失函数如公式(3-8)所示。

$$
\mathcal{L}_{TA} = \mathcal{L}_{st} + \mathcal{L}_{sc} + \mathcal{L}_{a} + \mathcal{L}_{k} \tag{3-8}
$$

3.3.3　基于动作树的计划采样

在本研究的预实验探索中，发现训练的越多，生成的回复越容易忠实于生成的动作。然后，这可能会导致不够令人满意的回复，尤其是当生成的动作不够合理的时候，这种现象称为错误累积，它是由于教师指导训练方式存在的曝光偏差问题所导致的。在教师指导下，测试和训练阶段存在不一致性[126]。计划采样[115] 是一种能够直接缓解这个问题的方法，它遵循课程学习的策略来随机使用模型自身的预测来替换待预测目标的真实输入。

然而，直接使用词元级别的计划采样在本研究的任务中不够有效，因为测试阶段的错误情况通常是错误的动作**序列**导致了不准确的回复，因此需要进行序列级别的替换。为了达成这个目标，本研究提出了一种基于动作树的计划采样方法。接下来将详细介绍该方法。

3.3.3.1　动作树

受到 SPACE[88] 的启发，本研究预先计算动作序列之间的相似度，并将形成的矩阵保存下来。在计算相似度的时候，首先将动作序列转化成层次化的树结构，它自顶向下包含 “领域、动作、槽位” 的三级结构，如图3-4的右半部分所示。接下来计算树编辑距离[128]，它是将一棵树转换成另一棵所需要操作（插入、删除和修改）数的加权和。需要指出的一点是，与 SPACE 不同，此处使用了**有序树**，原因是实验中发现动作的相对顺序会影响回复生成质量。除此之外，对一颗无序树进行重排会导致一些数据集标注中可能不存在的动作序列。将所有可能的动作树视为列表，分别记第 i 个动作树以及第 j 个动作树为 T_i 和 T_j。计算编

辑距离后，二者的相似度分数通过公式(3-9)计算。

$$\begin{aligned} s_{i,j} &= \frac{\max\{|T_i|, |T_j|\} - d_{i,j}}{\max\{|T_i|, |T_j|\}} \\ d_{i,j} &= \text{TreeEditingDistance}(T_i, T_j) \end{aligned} \tag{3-9}$$

3.3.3.2 计划采样

如图3-4的右下部分所示，在训练过程中，当一个真实动作 $\hat{A}_t$ 被输入到动作-回复解码器之前，它将以公式(3-10)计算的 $p(t)$ 的概率被保留[126]，否则，这个真实动作将会被用来索引相似度矩阵，假设索引到的行为 i，记相似度矩阵为 M，则采样概率如公式(3-11)计算得到。

$$p = \frac{\mu}{\mu + \exp(t/\mu)} \tag{3-10}$$

其中 μ 是一个超参数，并且这个函数随着训练步数增加严格递减。

$$p_j^* = \frac{M[i,j]}{\sum_{j=1, j\neq i}^{N} M[i,j]} \tag{3-11}$$

注意此处保证了相同的动作 i 不会被采样到。

3.3.3.3 损失函数

如公式(3-12)所示, 当扰动的动作作为输入的时候，动作的损失不会回传梯度，此时回复的损失仍然会被优化，从而提高在噪声动作下回复生成的鲁棒性。在这样的情况下，模型应该更多地依赖于历史上下文来生成回复。

$$\begin{aligned} \mathcal{L}_A &= -\log P(A_t | H_t, DB_t) \\ \mathcal{L}_R &= -\log P(R_t | H_t, DB_t, A_t) \\ \mathcal{L}_{AT} &= \begin{cases} \mathcal{L}_A + \mathcal{L}_R, & A_t = \hat{A}_t \\ \mathcal{L}_R, & A_t = A_t' \end{cases} \end{aligned} \tag{3-12}$$

3.3.4 训练和推理

最终整个训练过程的损失函数如公式(3-13)所示。需要说明的一点是，因为对话状态解码器不是本研究的关注点，因此不在此讨论，且受限于图例篇幅原因，在图3-4中也省去了，但是其损失 $\mathcal{L}_B$ 一直在训练过程中存在。

$$\mathcal{L} = \mathcal{L}_{TA} + \mathcal{L}_B + \mathcal{L}_{AT} \tag{3-13}$$

在推理阶段，只需要使用到共享的编码器和两个解码器，所有的分类头以及计划采样的手段不再需要，使得推理代价相比于基座完全不变。

3.4　实验结果与分析

本节将介绍实验数据、评价指标、对比的基线方法以及在不同场景下的实验结果。

3.4.1　数据集和评价指标

3.4.1.1　数据集

本研究在经典公开任务型对话基准数据集 MultiWOZ 上评估提出方法的端到端性能[3]，分别在 MultiWOZ 2.0 、2.1 和 2.2 上进行评测。遵循 MTTOD[91] 中的数据切分方式，训练/验证/测试集的规模分别为 8438/1000/1000。为了减少方法无关的表面形式多样性，把特定的槽位值进行去词汇化处理，仅保留其槽位类型，即 *[value_xxx]*，使得模型能够学到与值无关的参数表示，提升泛化性能[83]。

3.4.1.2　评估指标

本研究和一系列基线方法保持一致，采用自动评估指标来评价 MultiWOZ 数据集上端到端任务型对话的回复质量。**Inform rate** 衡量一个对话系统是否能够提供准确的实体；**Success rate** 衡量一个对话系统是否成功回应了所有用户请求的信息；**BLEU** 值[129] 通过与参考回复的比对计算得到，衡量了生成回复的流畅性。最终采用综合指标 **Combined score** = (Inform + Success) × 0.5 + BLEU，反映任务型对话系统的整体性能。

3.4.2　实验设置

遵循 MTTOD[91]，本研究使用预训练好的 T5-base 模型[58] 来初始化共享的编码器和两个解码器。本研究基于 Huggingface 的 Transformers 库[130] 来实现所有的方法与实验，在一张 40G 的英伟达 V100 GPU 训练 10 轮，一次训练大约耗时 10 小时。在低资源的设定下，训练的总轮数设置为 20。初始学习率为 5×10^{-4}，批次大小为 8，预热阶段的比例为 0.1。训练时采用的优化器为带有线性学习率衰减的 AdamW[131]。根据验证集上的性能来挑选最好的模型。对于超参数 μ 的选择，在不同数据集上尝试 {10,15,20} 并选择最适合的。为了消除随机性，固定随机种子为 42，在解码对话状态、对话动作和回复的时候，均采用简单的贪婪解码策略。

3.4.3　基线方法

为了进行公平的比较，本研究仅对比使用了预训练语言模型的方法，通常而言，存在以下两类设定

- **无持续预训练**: 直接在特定下游任务上对预训练模型进行微调。
- **有持续预训练**: 先在额外的大规模数据集上继续优化预训练目标，然后再转移到下游任务上进行微调。

由于本研究提出的方法没有进行持续预训练，因此本节接下来的内容将与**无持续预训练**的方法进行公平对比来说明本研究提出方法的有效性。此外，通过与**有持续预训练**方法的对比来说明本方法与上限方法的差距。在主实验和低资源设定的场景中，本研究对比了一些较强的基线方法，包括有 SimpleTOD[62], DoTS[132], SOLOIST[123], MinTL[133], PPTOD[90], UBAR[124], GALAXY[86], MTTOD[91], BORT[125], Mars[89] 以及 SPACE[88].

3.4.4 主实验结果

3.4.4.1 全量实验

表 3-1 MultiWOZ 2.0/2.1/2.2 的端到端实验结果[1]
Table 3-1 E2E performances on MultiWOZ 2.0/2.1/2.2.

Model	MultiWOZ 2.0				MultiWOZ 2.1				MultiWOZ 2.2			
	Inform	Success	BLEU	Comb	Inform	Success	BLEU	Comb	Inform	Success	BLEU	Comb
w.o. continual pre-training												
SimpleTOD	84.40	70.10	15.01	92.26	85.00	70.50	15.23	92.98	-	-	-	-
DoTS	86.59	74.14	15.06	95.43	86.65	74.18	15.90	96.32	80.40	68.70	16.80	91.40
SOLOIST	85.50	72.90	16.54	95.74	-	-	-	-	82.30	72.40	13.60	90.9
MinTL	84.88	74.91	17.89	97.79	-	-	-	-	73.70	65.40	19.40	89.00
UBAR	**95.40**	80.70	17.00	105.05	**95.70**	81.80	16.50	105.25	83.40	70.30	17.60	94.40
GALAXY	93.10	81.00	18.44	105.49	93.50	81.70	18.32	105.92	85.40	75.70	19.64	100.20
BORT	93.80	**85.80**	18.50	108.30	-	-	-	-	85.50	77.40	17.90	99.40
Mars	-	-	-	-	-	-	-	-	**89.20**	**80.30**	19.00	103.40
MTTOD	90.99	82.58	20.25	107.04	90.99	82.08	19.68	106.22	85.90	76.50	19.00	100.20
TA&AT	93.60	83.60	**20.67**	**109.27**	92.50	**84.00**	**19.78**	**108.03**	86.40	80.10	**20.34**	**103.59**
w. continual pre-training												
PPTOD*	89.20	79.40	18.62	102.92	87.09	79.08	19.17	102.26	83.10	72.70	18.20	96.10
GALAXY*	94.40	85.30	20.50	110.35	95.30	86.20	20.01	110.76	-	-	-	-
SPACE*	95.30	88.00	19.30	110.95	95.60	86.10	19.91	110.76	-	-	-	-

如表3-1所示，本研究提出的方法 TA&AT 在所有**无持续预训练**方法中，取得了最好的综合性能。即使与有持续预训练的最好方法 SPACE 进行比较，本研究提出的方法在没有使用任何额外数据的情况下，也是具有可比性的。值得指出的一点是，与本研究所基于的模型 MTTOD 相比较，本研究提出的方法能够将其性能在 MultiWOZ 2.0 提升 2.23（从 107.04 到 109.27），在 MultiWOZ 2.1 提升 1.81（从 106.22 到 108.03），在 MultiWOZ 2.2 提升 3.39（从 100.2 到 103.59）。此外，本研究提出的方法达到了所有方法中最高的 BLEU 值，验证了其在改善回复生成质量上的有效性。

3.4.4.2 低资源实验

为了探究本研究提出的方法在低资源场景下的有效性，遵循 Mars[89] 的设定，本研究基于 MultiWOZ 2.0 数据集，测试了模型在仅有 10%，20% 和 50% 的训练集下的表现。如表3-2所示，在大多数场景中，该方法获得了最好的表现，验证了有效性和鲁棒性。

[1]TA&AT 是本研究提出的方法的缩写，全称为 "**T**urn-level **A**uxiliary tasks and **A**ction-**T**ree based scheduled sampling"，表中的结果均引自原始论文。'*' 表示在额外的数据集上进行了持续预训练。

表 3-2　低资源实验结果
Table 3-2 Results of low-resource experiments

Model	10% 数据量				20% 数据量				50% 数据量			
	Inform	Success	BLEU	Comb	Inform	Success	BLEU	Comb	Inform	Success	BLEU	Comb
MinTL	55.5	44.9	15.6	65.8	64.3	54.9	16.2	75.8	70.3	62.2	18.0	84.3
PPTOD	68.3	53.7	15.7	76.7	72.7	59.2	16.3	82.3	74.8	62.4	17.0	85.6
UBAR	50.3	34.2	13.5	55.8	65.5	48.7	14.5	71.6	77.6	63.3	16.3	86.8
MTTOD	66.9	55.2	13.8	74.9	75.0	63.3	14.3	83.5	78.5	67.5	15.2	88.2
Mars	69.4	55.3	15.6	78.0	76.7	62.9	**17.2**	87.0	82.2	71.2	**18.6**	95.3
TA&AT	**71.5**	**58.4**	**16.2**	**81.1**	**79.2**	**68.2**	16.8	**90.5**	**83.5**	**73.8**	18.1	**96.8**

表 3-3　在 MultiWOZ 2.0 上进行的消融实验
Table 3-3　Ablation study on E2E results of MultiWOZ 2.0.

Model	Inform	Success	BLEU	Comb
TA&AT	93.60	83.60	**20.67**	109.27
- $\mathcal{L}_{st}$	93.10	84.50	19.79	108.59 (-0.68)
- $\mathcal{L}_{sc}$	92.60	**84.50**	20.28	108.83 (-0.44)
- $\mathcal{L}_{a}$	93.50	84.40	20.05	109.00 (-0.27)
- $\mathcal{L}_{k}$	**93.70**	84.40	20.29	**109.34** (+0.07)
w.o. AT	93.40	83.50	19.72	108.17 (-1.10)
w.o. TA	92.90	83.30	19.76	107.86 (-1.41)
MTTOD	90.99	82.58	20.25	107.04 (-2.23)

3.4.5　分析实验

在本节中，首先分析多任务目标以及计划采样的有效性，然后讨论从训练过程的学习曲线中所得到的一些观察。

3.4.5.1　消融分析

如表3-3所示，多数的额外任务是有效的，尤其是与槽位相关的任务。有趣的一点是，本研究发现在移除回复关键词预测的任务之后，综合性能取得了一定的提升，似乎这个任务不起作用。本研究将这个现象归因为不同的目标存在不同的学习难度和变化趋势，因此可能当 $\mathcal{L}_{sc/st}$ 已经过拟合的时候 $\mathcal{L}_{k}$ 还在欠拟合的阶段（进一步分析可见节3.4.5.2），最优的训练时间节点不一致导致了每个任务的贡献有所不同，引入自动化的权重配比可以成为未来的一个继续探索的方向。除此之外，+0.07 的微小提升也可能来源于随机性。

总而言之，基于动作树的计划采样 (AT, Action-Tree based scheduled sampling) 和轮级别多任务学习 (TA, Turn-level Auxiliary tasks) 都是有效的，而后者更为重要，因为它能够使得编码器输出更好的表示，从而同时对理解和生成造成影响。与最相似的基线 MTTOD 相比，方法整体提升了 2.23，验证了本研究提出方法的有效性。

3.4.5.2　学习曲线

图3-5展示了训练过程中不同任务对应的 F1 值变化。可以发现，在训练最早期的时候，F1 值几乎不变，说明此时的生成损失占据主要，在学习初始的生成能力时，隐状态的空间变化较大，导致分类器难以训练，因此比较容易输出随机、

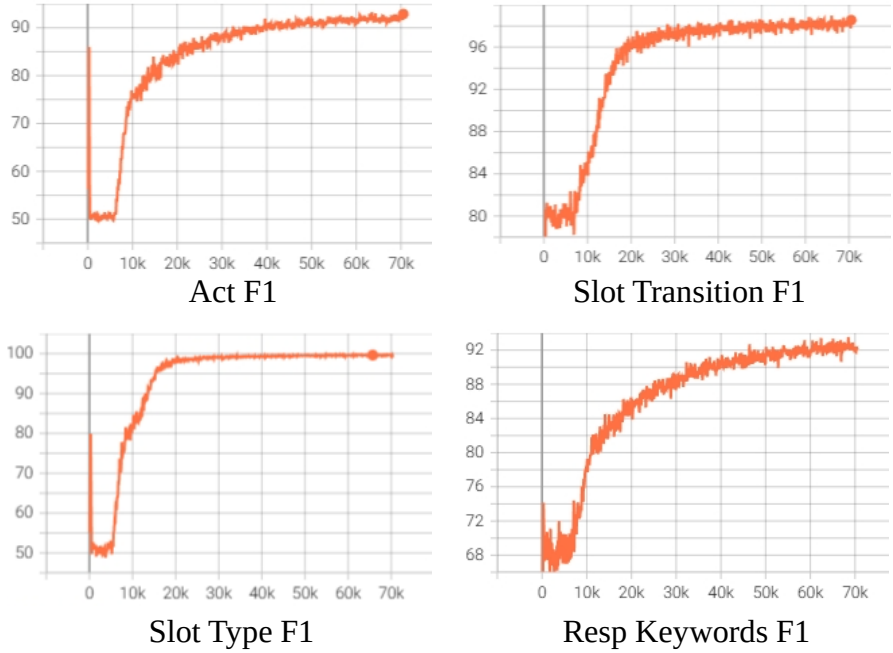


图 3-5　不同任务的训练学习曲线
Figure 3-5　Learning curve for different tasks in training

全 1 或者全 0 值。随着训练过程的进行，生成损失减小且多任务目标的损失比例增大，此时这些任务开始成功被优化。此外，与槽值状态相关的任务很快收敛且达到超过 96 的 F1 值，然而与动作相关的策略收敛相对较慢且只能够达到 92 左右的 F1 值。由此可见，后者比前者更难以优化，因为它除了理解能力之外，还需要更多的规划能力。

3.4.5.3　实例分析

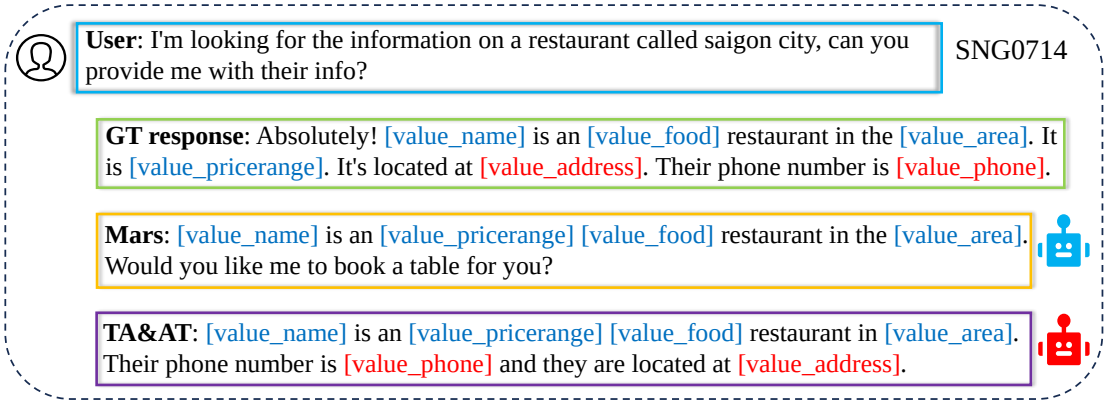


图 3-6　由 Mars 和本研究方法分别生成的一个去词汇化样例
Figure 3-6　Delexicalized responses generated by Mars and TA&AT

如图3-6所示的样例中，在用户需要一些信息的时候，本研究提出的方法能够相比于 Mars 生成更多的关键词，包含了所有标准回复中含有的信息，也没有引入冗余信息。

3.5　本章小结

本章提出了一种基于轮级信息表示增强的多任务学习方法。具体来说，为了尝试解决标注受限条件下，目前已有方法对于标签的不充分利用以及教师指导训练带来的序列级别错误累积问题，本研究首先从中间状态的标注中挖掘四种高层次信息，用来监督编码器输出的轮级别句向量表示，然后为解码器的训练引入了一种基于动作树的计划采样方法。该方法在 MultiWOZ 数据集上展现出相比于同等条件下其他方法更好的性能，分析实验也表明了多任务的不同目标以及计划采样的有效性，缩小了与通过额外数据的持续预训练方法之间的差距，且在更为标注稀缺的低资源场景下取得更为显著的优良表现。

第 4 章　基于检索增强的对话标注模式解耦方法

级联式生成的任务型对话系统中重要的一步是数据库检索，它要求对话状态的标注与预测遵循定义好的模式，然而这导致了系统的迁移性能差。使用检索增强生成的方式能够解耦知识，从而使系统摆脱数据库模式依赖，实现跨领域的迁移。

但是简单的端到端训练方法存在一些问题：（1）对话历史中可能存在大量噪声影响检索的准确性；（2）缺乏正确知识的标注。为了在保持端到端训练的前提下解决上述两个问题，本研究提出了一种基于查询提示优化的端到端检索增强方法。首先，为了自动确定合适的查询，本研究引入一个查询提示生成模块，使用强化学习优化查询的改写从而提升回复生成质量；其次，通过回复中蕴含的后验信息来指导端到端训练，克服正确知识标注的缺失问题；通过交替训练来缓解资源消耗以及提升稳定性。

本研究提出的方法在三个公开的任务型对话数据集上均进行了实验，实验结果证明本研究提出的方法能够相比于其他端到端的基线方法获得更好的表现。此外，分析实验的结果也验证了查询生成模块的引入和后验信息指导的有效性。

4.1　引言

任务型对话系统[134] 的目标通常是通过多轮对话的交互来帮助用户实现特定的任务，例如宾馆预订等。在对话过程中，系统需要依赖外部的数据库来提供知识，检索相关实体对应的信息，整合之后，形成更为流畅自然的自然语言表述返回给用户。传统的基于级联式生成的任务型对话通常包含有对话状态追踪模块，用来表达结构化的用户目标，它通常与数据库的模式相对应，能够直接用来查询数据库。但是，对话状态在标注的过程中，需要标注的字段较多，并且格式受限于数据库模式，对于标注的容错性低，具有较为复杂的标注过程；此外，更为严重的问题是，这种对于数据库的极度依赖一方面导致系统无法实现跨领域甚至跨数据集的迁移，而另一方面则会导致知识库中字段无法实现动态添加，无法动态维护知识库以适应不断更新的世界知识。这是由于训练过程与知识库本身模式的过度耦合所致，图4-1展示了一个这样的例子，MultiWOZ 和 CrossWOZ 是两个多领域的数据，不同领域的槽位是不一样的，而对于只见过一个领域的模型，它无法预测出其他领域的字段，从而对话状态无法用于查询其他领域的数据库，导致对话后续环节无法正常进行。即使两个数据集存在相同的领域，因为数据集中标注的模式存在不同，同领域跨数据集仍然无法预测。

为了解决上述问题，目前的研究思路通常将知识库与模型分离，进行端到端的训练。他们通常可以分为三类方法：（1）将数据库编码成一个记忆网络，并且基于对话上下文的表示来进行查询和更新[54,56,96]；（2）将知识直接编码到模型

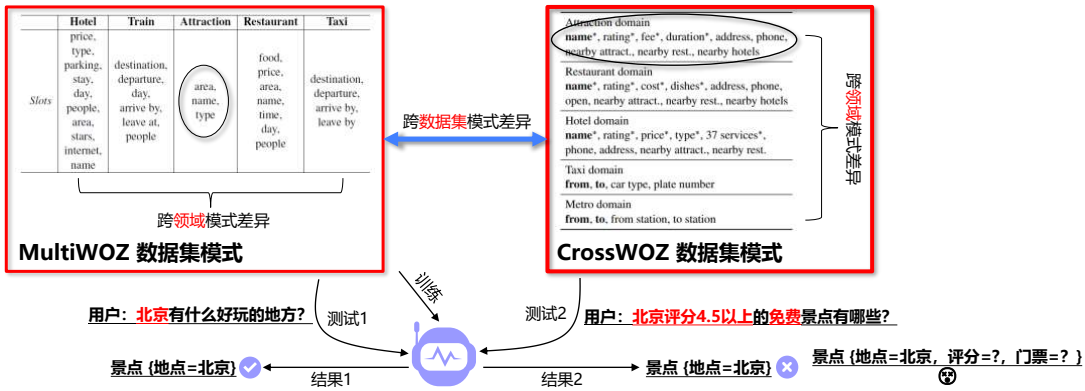


图 4-1　MultiWOZ[3] 与 CrossWOZ[4] 的数据集模式

Figure 4-1　The dataset schema of MultiWOZ and CrossWOZ.

的参数中，生成的过程中不进行显式的检索，而是相当于在基于模型参数的运算中“隐式地”进行了检索[97,98]。（3）使用预训练语言模型将序列化的知识库记录与对话上下文拼接连在一起，直接让语言模型根据上文进行信息的筛选与综合[99–104]。从 ChatGPT 问世以来的大模型时代开始，得益于更强的句向量嵌入模型对细粒度语义检索的增益，模型长文本能力的长足进步对于语言生成的增益，知识密集型任务往往首先通过检索获取相关的知识，然后使用上述的第三种方式进行建模，取得了优异表现。然而，目前在学术界，将大模型检索增强生成相关技术用到任务型对话的工作还较少，本研究希望填补这个空缺。

图4-2展示了一个基础的检索增强对话系统模型结构，提升该系统的表现存在两个重要挑战：（1）检索查询语句的选择：由于对话的多轮特性，历史可能包含与当前检索无关的噪声，直接用于查询会降低准确性，而只使用最近的一句又可能会遗漏信息；（2）正确知识标注缺失：回复的背后通常可能有一到多条正确的知识，而现有数据集中通常没有对此的标注，或是单一标注无法反映上文到知识存在的“一对多”映射。从以上挑战入手，本研究提出一种**基于查询提示优化的端到端检索增强方法**。

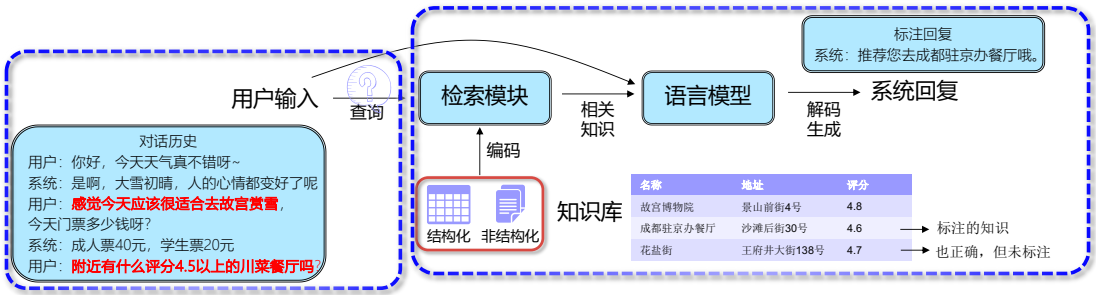


图 4-2　检索增强对话系统基础架构

Figure 4-2　Naïve framework of retrieval-augmented dialog system

为了解决第一个挑战，本研究引入一个**查询生成模块**，通过对上下文信息的分析综合，得出适合于检索的查询语句，为了克服梯度中断的问题，将回复的质

量作为奖励，使用强化学习优化查询生成；为了解决第二个挑战，本研究引入**后验信息指导**，用标准回复中存在的后验信息促进知识选择这一先验的学习，即在语言模型固定的时候，希望能够促使标准回复的条件似然更大的知识有更大概率在检索阶段被选择，从而在无知识标注的情况下，实现检索模块和回复生成语言模型的联合训练。为了提升训练的稳定性，以上两个部分实行交替训练。

本研究在三个基准数据集，MultiWOZ 2.1[134] (MWOZ), Stanford Multi-Domain[53] (SMD), CamRest[116] 上与其他的一些端到端任务型对话基线方法进行了比较，实验结果证明本研究提出的方法能够取得比基线更好的表现，并且在知识条数逐步变化的对比下，更为鲁棒。分析实验证明了每个模块的有效性，其中，查询生成模块的强化学习训练对于实体检索相关的指标提升更为重要。

4.2　相关工作

端到端任务型对话　基于数据库检索的端到端任务型对话工作可以分为三类。第一类是基于记忆网络的工作，它们将知识库编码成记忆网络，在查询过程中更新知识的表示，并使用注意力权重来综合知识向量的加权表示来生成回复，此类方法多用于在 Transformer 提出之前，那时候模型对于远距离文本的记忆能力还不够强，因此有效记忆容量较低。Maddoto 等人提出的 Mem2seq[54] 在记忆槽上使用多跳的注意力机制来索引知识词元，从而生成回复。Wu 等人提出的 GLMP[55] 引入了分层次的两级指针，全局指针用于生成回复的模板，局部的指针则用于从知识中确定每个待填充的具体实体。Qin 等人提出的 DF-Net[56] 通过对不同领域相似度的动态加权来实现更好的跨领域迁移能力。

第二类方法是直接在训练过程中将知识编码在模型的参数中，生成过程中不进行显式的检索，而是相当于在基于模型参数的运算中“隐式地”进行了检索。这类方法简化了流程，混合检索与生成两个步骤，但是可扩展性差，在面对大规模知识库的时候使用知识的准确率降低，且对知识的更新需要重新更新模型参数，不够轻量化。Maddoto 等提出的 GPT-KE[97] 使用数据增广方式将知识以模板构造的问答形式训练到 GPT-2 中，此后 Huang 等人提出的 ECO[98] 增加了带有前缀树限制的解码策略来保证回复中生成的实体本身与数据库的一致性，提升了实体相关的评价指标。

第三类方法如图4-3所示，直接将知识库和对话上下文拼接起来，输入到预训练语言模型中，让模型直接学习从知识和历史到回复的映射。这类方法扩展性较好，且效果能够随着预训练语言模型长文本能力的加强而取得自然提升。Xie 等提出的 UnifiedSKG[99] 将多种类型的结构化知识表示成文本，使用文本到文本统一建模知识驱动的任务；Rony 等提出的 DialoKG[135] 引入知识图谱上的关系并基于图谱对实体间的注意力重新进行权重分配，实现更好的知识选择；Wan 等提出的 MAKER[102] 提出多粒度的检索，增加从实体属性的层面的过滤；Shi 等[103] 提出从生成器中获得正向和负向的两种反馈来指导知识的选择；最接近本研究的工作是 Tian 等提出的 Q-TOD[101]，是第一个基于查询的端到端任务型

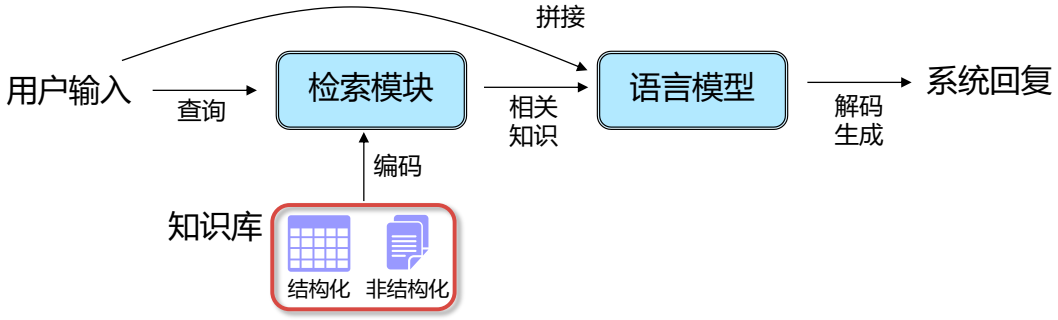


图 4-3 基于预训练模型的工作
Figure 4-3 Works based on pretrained language model

对话工作，并提供了带有查询标注的数据集，本研究所提出方法也是在 Q-TOD 基础上进行展开的。不同点在于，Q-TOD 的查询与检索模块均未实现与生成器的联合优化，因此无法取得最优结果，本研究正是尝试通过一些手段来实现这一点。总体来说，此类方法目前已成为主流，表现上也显著优于前两类方法，但是当知识库规模过大的时候，如果没有合适的预检索机制，会引入大量噪声且序列长度随着知识条目线性增长会使得推理代价提升。

检索增强生成 近年来人工智能生成内容 (Artificial Intelligence Generated Content, AIGC) 由于大模型的突破性进展在近年来取得显著进步，吸引大量研究人员的兴趣[136]。但是大模型本身仍然存在一些问题，例如难以学习长尾知识，存在事实性幻觉，以及泄露隐私数据等风险，为了缓解这些问题，目前系统中通常使用检索增强生成的方式，因为用于检索的知识库更易修改与包含长尾知识，降低模型端的负担。其中的核心在于如何利用检索结果，可以分为以下四类：

(1) 基于查询的检索增强生成：将检索结果和对话历史直接拼接作为语言模型的初始输入，主要作用于模型的输入阶段[137–142]。

(2) 基于隐状态表示的检索增强生成：通过检索结果的隐状态表示与模型进行交互，主要作用于模型中间层表示的阶段。[143,144]

(3) 基于解码输出概率分布的检索增强生成：使用检索结果对模型输出的概率分布进行加权调整，主要作用于模型在采样输出时的概率计算阶段[145,146]。

(4) 基于投机的检索增强生成：模仿大模型投机解码的策略，在适当的时候直接接受检索到的部分内容带替代大模型的解码输出，主要用于大模型的输出阶段[147–149]。

本研究所提出的方法属于第 (1) 类，而与现有方法主要的不同点在于，本研究尝试在**端到端任务型对话**上应用基于查询语句改写的检索增强生成，并实现了**检索器与语言模型的联合训练**。

4.3　方法介绍

4.3.1　模型框架

本研究所提出的基于查询提示优化的端到端检索增强方法所使用的系统如图4-4上半所示，包含一个查询提示模块，一个检索模块和一个基于大语言模型的生成器模块。其中，查询提示模块总结对话历史信息，生成适合于当前轮次的查询提示语句；检索模块使用查询提示语句来检索线性化统一成文本形式后的知识库；最后，生成器模块将对话历史和检索到的知识拼接作为输入，生成回复语句。

图4-4展示了模型基本框架，上半部分是查询提示模块的训练优化过程，下半部分是基于后验信息指导的检索模块与生成器模块联合优化过程。接下来将依次详细地介绍不同模块的工作原理和训练流程。

4.3.1.1　任务定义

检索增强的端到端任务型对话定义如下。给定一个任务型对话的数据集，对其中的每一段会话（session）定义为 $S = \{(x_t, y_t, D_t)\}, t = 0, 1, 2, T$，$x_t$ 表示第 t 轮用户的语句，y_t 表示第 t 轮系统的标准回复，而 D_t 表示当前轮已经过粗筛的知识库，对话历史 c_t 定义为 $\{x_0, y_0, x_1, y_1, ..., x_t\}$，该任务要求系统首先根据输入的对话历史 C_t 检索所需要的知识集合 D'_t，接下来将其与对话历史拼接起来的 $[D'_t, C_t]$，生成回复 y'。而由于本研究额外引入了一个查询生成模块，记生成的查询为 x'_t，它将用于代替完整的对话历史 C_t 进行对知识的检索。

4.3.1.2　查询提示生成模块

从图4-2中可以看出，在多轮对话场景下，累积的对话历史往往存在大量与当前检索无关的信息，但是也不能简单地仅使用最近一轮，那样可能会出现信息的缺失，因此如何动态地确定选择哪些对话历史成为一些工作的研究方向[35]。从另一个角度，受到对话状态追踪领域相关工作的启发[150,151]，可以通过生成式摘要的方式来直接生成当前轮次的对话状态，因此本课题使用一个查询生成模块，来直接学习对于后续检索有利的查询语句。使用预训练的 T5-large[58] 作为该模块的初始化，其参数记为 θ_Q。该模块接收对话历史作为输入，输出生成的查询提示语句，如公式(4-1)所示。为了训练的稳定性，该模块首先在 Q-TOD[101] 提供的查询标注数据集上使用有监督微调进行预热，然后使用强化学习进行训练。

$$\hat{q}_t = \mathrm{T5}_{\theta_Q}(C_t) \tag{4-1}$$

4.3.1.3 知识库与检索模块

每一轮对话对应的知识库[1]，记为 $D_t = \{d_0, d_1, ...d_{|D_t|}\}$，使用规则对这些知识进行线性化成非结构化文本的形式。本研究所使用的检索模块为一个 Transformer 编码器结构，使用 Contriever[152] 模型参数作为初始化。知识库与节4.3.1.2生成的查询提示分别输入到检索模块中，根据公式(4-2)进行相似度的计算，其中 top-k 的知识将被选择，进入到后续生成阶段中。

$$\begin{aligned} e_t^q &= \text{CLS}(\text{Contriever}(q_t)) \\ e_{t,i}^d &= \text{CLS}(\text{Contriever}(d_i)), i \in [0, |D_t|] \\ s_{t,i} &= \text{CosSimilarity}(e_t^q, e_{t,i}^d) \\ D_t' &= d_{t,i'}, s_{t,i'} \in \text{topk}(s_{t,i}) \end{aligned} \tag{4-2}$$

4.3.1.4 回复生成模块

将筛选出的 top-k 知识与对话历史拼接后输入到回复生成模块中，通过一个基于纯解码器结构的大语言模型生成回复。本研究使用 Llama-2 作为此处的初始化参数。相比于直接输入一整个知识库 D_t，经检索模块筛选过后的知识库 D_t' 含有的噪音更少，并且能够缩短输入长度，从而提高推理的时间和空间效率。

$$y_t = \text{Llama}(D_t', C_t) \tag{4-3}$$

4.3.2 训练流程

整体训练流程如图4-4所示，包含有 “查询提示优化” 以及 “后验信息指导” 的两阶段训练，前者属于强化学习，仅作用于查询生成模块的参数；后者属于有监督微调，仅作用于检索模块和语言模型部分。为了在有限显存资源下成功训练，且提升训练时的稳定性，以上两个阶段实行交替训练，本节接下来的部分将分别详细介绍两个阶段。

4.3.2.1 有监督预热

遵循 Q-TOD[101] 的数据处理方式，构造查询生成的训练集 $Q_{Train} = \{(c, q)\}$。使用公式(4-5)所示的标准负对数似然损失来有监督地微调查询提示生成模块。

$$\mathcal{L}_{warmup}^{Q} = \sum_i \log p_{\theta_Q}(q_{t,i}|C_t, q_{t,<i}) \tag{4-4}$$

在查询生成模块预热完毕之后，离线地为训练集所有的会话历史生成对应

[1]该知识库可以是全局共享，也可以是每轮对话使用高效统计方法粗筛得到的局部集合，在本研究中，随着数据集标注情况的不同，两种情况都存在

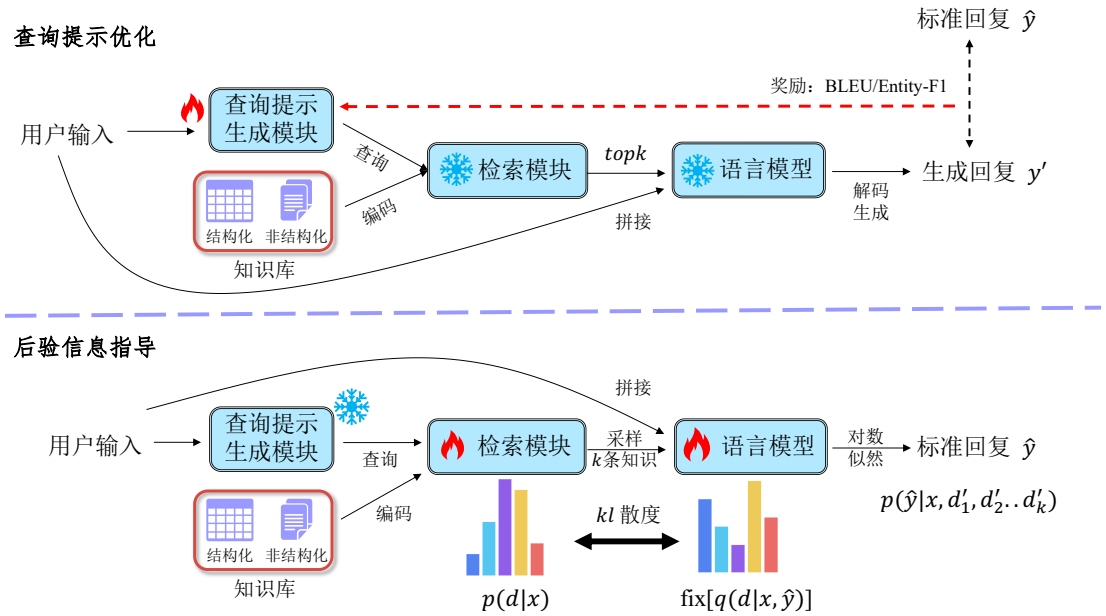


图 4-4　基于查询提示优化的检索增强模型框架与训练流程

Figure 4-4　Framework and training procedure of query prompt optimization based retrieval-augmented model

的查询提示，此时由于检索器虽然还没有经过下游任务训练，但是预训练已经使其具备了初步的能力，因此使用检索器挑选出 top-k 的知识，与对话历史拼接之后，使用真实回复的标准负对数似然来对语言模型进行有监督预热。记语言模型的参数为 θ_G

$$\mathcal{L}^{G}_{warmup} = \sum_{i} \log p_{\theta_G}(y_{t,i}|D'_t, C_t, y_{t,<i}) \tag{4-5}$$

由于两个模块都需要在后续的强化学习过程中进行解码，而预训练模型完全没有在对应的下游任务上进行过微调，生成效果很微弱，因此对两个模块分别预热的主要目的是防止强化学习在一开始出现效果崩溃的情况。其中，查询生成模块预热时仅仅需要本身参与，缺乏与检索器即语言模型所组成的环境进行交互的步骤，需要进行下一步的强化学习优化。

4.3.3　查询改写优化

为了更进一步地微调查询提示生成模块，需要以优化最终的回复生成质量作为目标，然而由于查询生成存在解码步骤，使用一般的有监督学习无法回传梯度，因此需要引入强化学习，将回复生成质量作为奖励，通过策略梯度的算法来优化。

4.3.3.1　强化学习建模

类似于问答任务中的建模[142]，给定对话历史，通过自回归地逐 token 解码得到整个查询语句，并与后续检索与生成模块交互最终生成回复的过程，遵循马尔科夫决策过程。它由一个五元组 $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ 决定，就查询提示语句的生

成过程而言，S 为状态空间，此处为所有可能生成的查询语句，由词表和序列长度决定；$\mathcal{A}$ 为动作空间，此处即整个词表；$\mathcal{P}$ 表示转移概率 $p_{\theta_Q}(q_t|x, q_{<t})$，由查询提示生成模块的参数决定；$\mathcal{R}(y, \hat{y}) = \text{BLEU}(\hat{y}, y) + \text{Entity-F1}(\hat{y}, y)$ 表示奖励函数，反映根据标准回复所评估的生成回复质量；最后的 γ 是一个在计算累积回报时需要用到的衰减因子。训练过程中，基于对话历史 C 和已生成的查询 token $\hat{q}_{<t}$，经过策略网络 θ_Q 得出的分布采样动作 $\hat{q}_t \sim p_{\theta_Q}(\cdot|x, \hat{q}_{<t})$，直到生成终结符为止，整个查询语句 $\hat{q}_{0,1,\ldots t}$ 序列称为采样的一个轨迹。将检索模块与语言模型视为一个整体环境，生成的 q 输入到环境后，可以得到生成回复对应的奖励，此时检索模块和语言模型参数冻结，该过程是一个确定性的过程。为了简化后续表述，简记为 $\mathcal{R}(C, q)$。

4.3.3.2 近端策略优化

使用节4.3.2.1中预热好的模型作为策略模型的参数初始化，记此时的策略为 π_{θ_0}。遵循近端策略优化算法（PPO, proximal policy optimization[16]），训练目标为在不偏离原始策略太远的情况下，最大化期望奖励。如公式(4-6)所示。

$$\max_{\theta_Q} \mathbb{E}_{C\sim\mathcal{D}, q\sim\pi_{\theta_Q}(\cdot|C)}[\mathcal{R}(C, q) - \beta log\frac{\pi_{\theta_Q}(q|C)}{\pi_{\theta_0}(q|C)}] \tag{4-6}$$

4.3.4 后验信息指导

由于数据集中不含有真实知识的标签，并且选取 top-k 知识拼接对话历史作为语言模型输入这一过程会导致其梯度无法回传至检索器，导致二者无法联合训练，因此，本节尝试通过后验信息指导来解决这个问题，具体而言，人们在基于知识检索完成一个问答任务的时候，答案中通常会包含与查找记录相关的内容，即可以根据标准回复作为后验信息，来得出知识的相对权重，一部分知识型对话中的工作据此提出了一些基于变分自编码器的方法[153,154]，本课题出发点相似，但是采用了一种更为稳定的训练方式，避免了从后验分布中采样重参数化优化这一过程。在该训练过程中，查询模块参数冻结，因此从对话历史 C 生成查询 q 是一个确定性的过程，即对于相同的对话历史 C，生成的查询 q 是一致的，则知识检索的概率分布 $p_R(d|q) \propto p(d|C)$。为了后续推导在表述上的一致性，将检索器建模的概率直接记为 $p(d|C)$。具体而言，如公式(4-7)所示，目标为拉近先验分布与后验分布的距离。

$$\mathcal{L}_{KL} = \frac{1}{|\mathcal{B}|}\sum_{C\in\mathcal{B}} KL(p(d|C)||q(d|C, y)) \tag{4-7}$$

其中，$\mathcal{B}$ 表示一批对话历史所组成的集合。接下来，根据贝叶斯公式对后验概率重写，如公式(4-8)所示。

$$\begin{aligned} q(d|C,y) =& \frac{p(d,y|C)}{p(y|C)} \\ =& \frac{p_{LM}(y|C,d)p(d|C)}{p(y|C)} \\ q(d|C,y) \propto& p_{LM}(y|C,d)p(d|C) \end{aligned} \tag{4-8}$$

其中 $p_{LM}(y|C,d)$ 为语言模型概率，$p(d|C)$ 为检索器模块输出的知识概率分布，二者乘积所形成的值通过公式(4-9)归一化后得到联合概率，

$$p(d,y|C) = \frac{p_{LM}(y|C,d)p(d|C)}{\sum_{d\in D} p_{LM}(y|C,d)p(d|C)} \tag{4-9}$$

与先验概率 $p(d|C)$ 拉近。总的来说，这一步的直觉在于使用不同知识生成标准回复的困惑度来监督不同知识的生成概率。为了提升训练和测试的一致性，本课题直接使用预测的先验分布概率进行**不放回**的 k 次抽样得到 k 条预测知识，将这些预测知识与对话历史拼接到语言模型输入中，而不是从后验概率中采样，这样会导致生成器训练过于简单，缺乏鲁棒性。此外，在训练时使用采样而非确定性的 top-k 策略是为了让所有的知识表示都有机会得到更新，增加训练时输入端的多样性，以上都可以视为一些类似于给输入端增加扰动的正则化方法。记筛选出的知识集合为 D'，语言模型的损失如公式(4-10)所示

$$\mathcal{L}_{LM} = -\sum_{i} \log p(y_i|D',C,y_{<i}) \tag{4-10}$$

综上所述，该环节为完全的有监督训练，损失函数如公式(4-11)所示，包含先验后验分布的拉近以及语言模型似然两部分的损失，它区别于节4.3.3所描述的查询提示优化环节。两过程实行交替训练，在查询提示优化的强化学习训练过程中，只有查询提示生成模块更新参数，其余模块均冻结；在后验分布指导的监督学习训练过程中，查询提示生成模块冻结，检索器模块和语言模型的生成器模块更新参数。

$$\mathcal{L} = \mathcal{L}_{KL} + \mathcal{L}_{LM} \tag{4-11}$$

4.4　实验结果与分析

本节介绍测试端到端任务型对话性能的数据集、实验设置、基线模型以及相关的实验结果。

4.4.1 数据集与评价指标

为了验证方法的有效性，本工作在三个经典任务型对话数据集上进行了实验。它们分别是 MultiWOZ 2.1 (MWOZ)[134], Stanford Multi-Domain (SMD)[53] 以及 CamRest[116]。这些数据集每一轮都含有特定的知识库。由于使用 Q-TOD[101] 的标注进行初始化，因此遵循了 Q-TOD 的预处理数据格式以及训练测试集的划分。值得指出的是，不同数据集每轮的知识库规模有所不同，具体的统计信息如表4-1所示。

表 4-1 数据集统计信息
Table 4-1 Dataset statistics

统计信息	SMD	CamRest	MWOZ
对话数	3031	676	2097
语句数	15928	5488	19632
领域数	3	1	3
平均轮数	5.26	8.12	8.89
句平均词元数	7.97	12.31	14.73
训练/验证/测试集	2425/302/304	406/135/135	1839/117/141
轮平均知识库规模	5.95	1.93	7

在评价指标方面，本研究与其他相关基线方法保持一致，采用 BLEU[129] 和 Entity-F1[53] 作为评估指标。具体而言，BLEU 值根据与标准回复的 n-gram 的重合度衡量生成回复的流畅性。Entity-F1 值根据实体的精确率和召回率微平均来计算 F1 值，衡量包含知识的正确性。

4.4.2 实验设置

本研究的实验使用 T5-Large[58] 初始化查询提示生成模块，使用单流编码器 Contriever 初始化检索模块，使用 Llama-2-7B[71] 作为语言模型，在各自需要进行参数更新的阶段，查询提示生成模块与检索模块均使用全量微调，而 Llama-2-7B 使用 LoRA[78] 微调。整体训练基于 Huggingface Transofmers[2]库，训练流程相关脚本主要使用了 Llama-Factory[3]的封装。PPO 算法的 β 参数在 (0.8 ~ 1.2) 的区间内进行开发集上步长 0.1 的网格化搜索，其余超参数则使用默认配置。

强化学习训练时，使用相同与策略模型相同的参数初始化状态估值模型，而后该模型与策略模型不共享参数，训练时将检索模块与语言模型组成的整体部署成服务，因此不与查询生成模块的训练抢占显存，学习率为 1×10^{-5}，通过梯度累积实现总批次大小为 8。训练过程中的采样遵循 Li 等[15] 的设定，使用 top-$p = 0.9$ 的随机采样，裁剪很不相关的动作。

检索器模块与语言模型进行联合训练时，为了降低显存占用，使用 deepspeed[4]的 zero-2 阶段训练，学习率为 5×10^{-5}，通过梯度累积实现总批次大小为

[2]https://github.com/huggingface/transformers
[3]https://github.com/hiyouga/LLaMA-Factory
[4]https://github.com/microsoft/DeepSpeed

32。知识选择时的 top-k 值设定为 3。所有模块在训练的时候都使用 AdamW[155] 作为优化器，使用余弦函数作为学习率的调度器。训练使用 4 张 Nvidia 3090 24G 的显卡。在预热过程结束后，先进行 3 轮的后验信息指导有监督训练，然后进行 1 轮的强化学习训练，如此交替，根据测试集上的结果来选择训练结束的时间。

在测试时，对于查询提示生成模块和语言模型均采用束大小为 4 的集束搜索进行解码。选择知识的 top-k 与训练时保持一致，均为 3。

4.4.3　基线模型

本课题方法对比了以下三类的基线模型：

记忆网络：此类方法使用记忆单元来保存外部知识，并且使用多跳的注意力来加权综合知识的表示以生成回复。工作包括 DSR[116], KB Retriever[156], GLMP[55], DF-Net[56], EER[157], FG2Seq[158], CDNET[96], GraphMemDialog[100]。

模型参数编码知识：此类方法直接在训练阶段将知识训练到模型的参数中，在推理阶段不需要进行显式的检索，通过与内部参数的交互即能够得到带有知识的回复。工作包括 GPT2-KE[97] 和 ECO[98]

知识与历史直接拼接：此类方法将知识库或者筛选过后的知识库直接与对话的历史上下文拼接起来输入到模型中，生成回复。方法包括 DialoKG[135], UnifiedSKG[99], Q-TOD[101], Dual-Feedback[103], MAKER[102]

4.4.4　主实验结果

主实验的结果如表4-2所示。本研究提出的方法超过了其他的基线系统，在三个公开数据集上取得了最好的结果。相比于之前最好的模型 MAKER，本研究的方法在 MWOZ 上 BLEU 值高 0.57，Entity F1 高 1.34；在 SMD 上 BLEU 值高 0.61，Entity F1 高 2.39；在 CamRest 上 BLEU 值高 0.70，Entity F1 高 2.56。实验结果证明了方法的优越性。

4.4.5　分析实验

为了进一步分析本研究所提出方法的有效性，本节将详细介绍从不同探究角度出发的分析实验，包括对不同模块的消融分析以及较为核心的检索模块相关设定的一些分析。

4.4.5.1　消融分析

为了分析不同模块或信息的作用，进行了以下几组设定。“无查询模块” 表示去除查询改写模块和强化学习的训练阶段，直接使用整个对话历史作为检索的输入；“无后验信息训练” 表示训练过程中不使用后验信息进行监督，此时检索器的训练要求知识标签，为了使实验顺利进行，直接使用回复语句 BM25 算法检索知识库知识得到的排序作为伪标签，注意该场景下，检索器与生成器之间无法传播梯度，相当于分别训练，但是生成器输入仍然会受到检索器输出的影响；“无检索模块” 表示不加筛选地将知识库中的知识完全拼接起来，输入到模型中。

表 4-2 主实验结果
Table 4-2 Main experimental results

模型	MWOZ		SMD		CamRest	
	BLEU	Entity F1	BLEU	Entity F1	BLEU	Entity F1
DSR[116]	9.10	30.00	12.70	51.90	18.30	53.60
KB-Retriever[156]	-	-	13.90	53.70	18.50	58.60
GLMP[55]	6.90	32.40	13.90	60.70	15.10	58.90
DF-Net[56]	9.40	35.10	14.40	62.70	-	-
GPT-2+KE[97]	15.05	39.58	17.35	59.78	18.00	54.85
EER[157]	13.60	35.60	17.20	59.00	19.20	65.70
FG2Seq[158]	14.60	36.50	16.80	61.10	20.20	66.40
CDNET[96]	11.90	38.70	17.80	62.90	21.80	68.60
GraphMemDialog[100]	14.90	40.20	18.80	64.50	22.30	64.40
ECO[98]	12.61	40.87	-	-	18.42	71.56
DialoKG[135]	12.60	43.50	20.00	65.90	23.40	75.60
UnifiedSKG[99]	13.69	46.04	17.27	65.85	20.31	71.03
Q-TOD[101]	18.27	53.28	21.76	73.44	24.65	76.81
Dual Feedback[103]	18.48	53.17	25.10	71.58	26.00	74.04
MAKER[102]	18.77	54.72	25.91	71.30	25.53	74.36
本研究方法	**19.34**	**56.06**	**26.52**	**73.69**	**26.23**	**76.92**

消融实验的结果如表4-3所示。首先，去掉不同模块之后，回复质量均出现

表 4-3 MWOZ 数据集上的消融实验
Table 4-3 Ablation study on MWOZ dataset

模型	**BLEU**	**Entity F1**
本研究方法	19.34	56.06
无查询模块	17.22 (-2.12)	51.72 (-4.34)
无后验信息训练	17.93 (-1.41)	52.86 (-3.40)
无检索模块	18.84 (-0.50)	55.24 (-0.82)

了不同程度的下降，验证了以上每个模块的有效性。然而，查询模块的训练过程由于使用强化学习直接优化指标，对于 BLEU 值以及 Entity F1 指标的贡献度更大。此外，还能发现的一个有趣的现象是，当去掉检索模块之后，整体性能没有出现太大的损失，这是由于目前基于的基座模型 Llama-2-7B 本身就具有了比较好的较长文本处理能力，且 MWOZ 给定的知识库条数有限（每轮均为 7 条），通过微调能够比较轻松地过滤掉不相关的知识。但是代价是随着知识库规模的扩大，序列长度会呈现线性增长，大大增加推理的成本，因此本研究提出的检索增强方法仍然具有一定的现实意义。

4.4.5.2　知识选择相关设定分析

知识选择先验假设　在对后验概率使用公式(4-8)进行计算的过程中，如果此公式中的先验概率假设为均匀分布，则可以直接让语言模型的概率分布 $p_{LM}(y|C,d)$ 归一化后与预测的先验概率 $p(d|C)$ 拉近，这正是过去一些工作所基于的假设[140,159]。由于在本任务中知识规模相对较小，分布不均匀，因此本研究不采用上述假设。针对二者的不同，分析实验[5]的结果如表4-4所示，可以看出在本任务中，不进行均匀分布的先验假设效果更好。

表 4-4　MWOZ 数据集的知识选择先验假设分析实验
Table 4-4　Analysis on MWOZ dataset for different assumption of knowledge prior

先验假设	**BLEU**	**Entity F1**
均匀	18.46	52.68
非均匀	**19.02** (+0.56)	**54.36** (+1.68)

线性化 vs 并列化知识　对于运用知识的模式，存在线性化和并列化两种方式，前者将所有知识拼接，好处是能考虑到所有知识，缺点是序列长度线性增加；后者则分别拼接，后续通过投票的方式选择最好的回复，好处是可以并行处理，缺点是无法考虑需要多条知识的情况。本文采用的模式是前者，从表4-5中可以看出，线性化的方式明显地优于并列化的方式。因为在 MultiWOZ 的多领域数据集中，为了准确地完成一轮回复，需要有多条知识的信息。

表 4-5　MWOZ 数据集的知识输入模式分析实验
Table 4-5　Analysis on MWOZ dataset for organization of knowledge input

知识输入模式	BLEU	Entity F1
并列化	17.02	52.96
线性化	**19.34** (+2.32)	**56.06** (+3.08)

知识条数 top-k 的影响　一般而言，当选择的知识条数过少，有可能出现有用信息被遗漏的情况；而当选择的知识过多，则难免出现大量的噪声，影响后续的生成。为此，本研究在 MultiWOZ 数据集上，探究了知识条数与选择知识条目 top-k 之间的关系，分析实验的结果如图4-5所示。可以看出，随着 k 值的增大，一开始性能出现明显的提升，这说明当 k 较小的时候，信息量不足以得出正确答案，通过知识条数的增加能够很快提升输入中含有的信息，进而提高对话系统表现；然而当 $k>3$ 之后，性能开始逐步略有降低，但是降低幅度相比于 $k<3$ 的程度低很多，这说明一方面过多的知识引入了噪声，但同时语言模型本身对于噪声知识的鲁棒性较好，所以影响较小。这对当今基于大模型检索增强的启示是，提升知识出的召回率比准确率更为重要。

[5]此处为了控制变量，使用的是预热阶段的查询生成模块

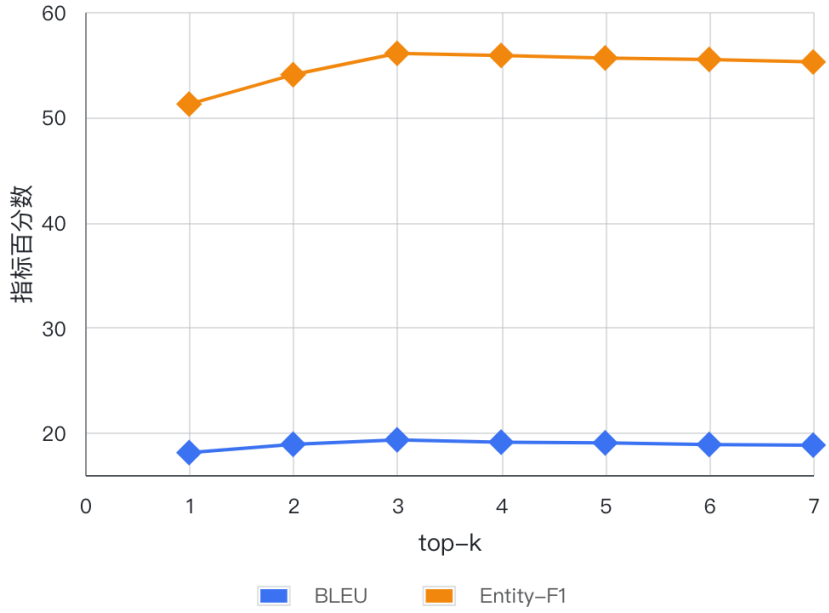


图 4-5　对话系统表现随 top$k$ 的变化的曲线
Figure 4-5　Performance of dialog system over top-k

4.4.5.3　实例分析

图4-6展示了某个样例在强化学习加入与否之下，查询改写结果的不同进而导致的回复质量不同。对于该历史上下文而言，红色圈出的部分为干扰项，未经过强化学习的模型容易直接根据这个强条件做出错误的检索语句，此外，由于标准查询在标注时候的缺陷，提前得出了未通过检索才能得出的答案，而模型预测则会使用更为合理的 “the same” 表述，模型只有在检索出 “chiquito restaurant bar” 以及 “the good luck food takeaway” 两者之后，才实现正确的检索以及后续生成正确的回复。但是在这个样例中，同时也暴露了本文模型存在的一个问题，就是无法结合上下文过程中推导出来的额外条件，例如这里的 “same”，在对话历史中未提及具体的价位以及地点，模型在此处能答对，其实很可能是在参数中已经记住了一部分知识，而扩展到更大规模的知识库情况下，则是较大的挑战，因此未来工作中可以考虑如何将历史对话隐含的状态信息加入到检索输入中。

表 4-6　强化学习消融的生成样例展示
Table 4-6　Generated Example Demonstration of Reinforcement Learning Ablation

Dialogue Context	U: Can you lookup the **chiquito restaurant bar** for me? S: Yes I found it, what information would you like?. U: Can you book a table for 4 at chiquito on Tuesday at 20:00 please? S: I'm sorry but the booking was unsuccessful. Would you like to try **another restaurant**? U: Yes in **the same area and price range**. I'll need a postcode too.
Generated Query w. RL	Book a restaurant of the same pricerange and area with chiquito restaurant bar. for 4 people on tuesday at 20:00.
Generated Query w.o. RL	Book the chiquito restaurant bar for 4 at 20:00. Request the postcode.
Oracle Query	Book an expensive restaurant in the south for 4 people on tuesday at 20:00. Request the postcode.
Retrieved Knowledge w. RL	chiquito restaurant bar, mexican, south, expensive, ... the good luck chinese food takeaway, chinese, south, expensive, ...
Retrieved Knowledge w.o. RL	chiquito restaurant bar, mexican, south, expensive, ...
Generated Response w. RL	I will recommend the good luck chinese food takeaway and its postcode is cb17ag. Would you need more information?
Generated Response w.o. RL	Alright, the reference number of chiquito restaurant bar is cb17dy.
Oracle Response	I booked a table at the good luck chinese food takeaway. The postcode will be cb17ag. The restaurant is at 82 cherry hinton road. Will that be all?

4.5　本章小结

本章提出了一种基于查询提示优化的端到端检索增强方法，尝试解决任务型对话中的依赖模式标注问题。具体来说，本研究首先提出了一个查询提示生成模块，用于综合上下文得出适合于当前轮次的查询语句，使用强化学习来优化该查询经过整个系统后的回复质量奖励；其次，为了实现检索器与语言模型生成器的联合训练，在没有正确知识标注的情况下，使用标准回复中蕴含的后验信息作为指导，拉近与先验知识分布的距离。该方法在三个任务型对话数据集中展现出了比基线模型更好的性能。分析实验验证了查询提示生成模块、后验信息增强、检索器等的有效性。

第 5 章　基于模拟交互的对话偏好对齐训练方法

传统的任务型对话工作通常在相同分布的数据上训练和测试，这导致模型跨任务跨领域的泛化表现差，难以应对现实生活的复杂环境。此外，基于参考回复比对的多轮对话自动评估存在策略错位问题，要求模型在已生成不同对话历史的前提下，仍然生成标准回复，这种评估方式无法适配对话“一对多”的特性。

为了解决第一个问题，提升跨领域性能，本研究提出一种基于多用户反馈的迭代式偏好对齐方法，让模型与不同用户模拟器进行交互，通过对对话记录的逐轮筛选与改进，取得偏好数据，通过直接偏好优化算法对齐模型与偏好。每训练一定的数据规模就重新评估模型在不同用户模拟器的性能，重新调整交互的权重，从而实现迭代提升。

为了解决第二个问题，实现更合理的会话级别评估。得益于强大的大语言模型，本研究提出一种基于大语言模型思维链与任务分解的会话评估方式，具有一定的可解释性，并且使用基于用户模拟器交互的评测，缓解固定数据集评测存在的问题。该评估框架也作为指导本研究各个实验的主要依据。

相比于基线方法，本研究提出的基于多用户交互的偏好对齐方法在任务型对话与知识型对话数据集上均取得了一定的提升，同时，实验结果还证明了迭代式提升能够持续提升在较难领域上的性能。

5.1　引言

在现代社会，人们出于生活以及获取知识的目的，需要自主查阅或检索相关资料，而随着大模型相关技术的兴起，人机对话技术取得重大突破，智能助手能够与用户进行流畅的交流，并且本身蕴含了较为丰富的世界知识，在很多场景下能够给用户以良好的体验。前两章的研究主要停留在任务型对话，且通常训练和测试在同一个分布的数据集上，而现实生活中的场景远比单一的数据集更为复杂，人们期望的智能助手通常是多领域多任务的，不仅仅局限于类似机票预订的任务型对话，还包括对于某个话题进行讨论的知识型对话；不仅仅是被动地回应用户的查询，还要尝试主动挖掘用户的兴趣甚至引导一个新的话题，这是一个长远而现实的目标。局限于人工构造数据以及评估的成本高昂且耗时耗力，以及当时的小规模语言模型通用性能较差，过去学术界的研究对于贴近现实的目标鲜有尝试，如今得益于强大的大语言模型，使得在研究工作上迈出这一步成为可能。本研究希望尝试模拟尽可能真实的交互环境与评估方法，目前主要存在两个挑战。

第一个挑战是**基于实时交互式自动评估方法的缺失**。评估指标是深度学习中挑选模型、分析方法有效性的关键参考，因此确定一个**贴近真实场景**的合理评估指标至关重要。对于多轮对话的评估指标，分为轮级和会话级两类，前者

将数据集中的标准对话历史作为当前输入，而后者使用生成的对话历史；使用BLEU[129]、ROUGE[160] 等计算最新生成的一轮回复与标准回复之间的相似度。评估方法的优点是速度快，无需人工参与；然而对于轮级评估而言，无法模拟实际测试过程中可能存在的错误累积现象；此外，由于对话上文和回复具有“一对多”的特性，标准回复难以涵盖所有情况，会惩罚与参考回复不同但是使用其他同样合理策略的生成回复，因此无法实现正确的评估。而对于会话级评估而言，则存在如图5-1所示的策略错位问题[13]，当系统回复变化的时候，下一轮的用户输入仍然来自于数据集，而不会变化，这与现实不符，无法正确地评估一个对话系统的好坏。理论上而言直接与人的交互最能反映真实效果，例如产品新版

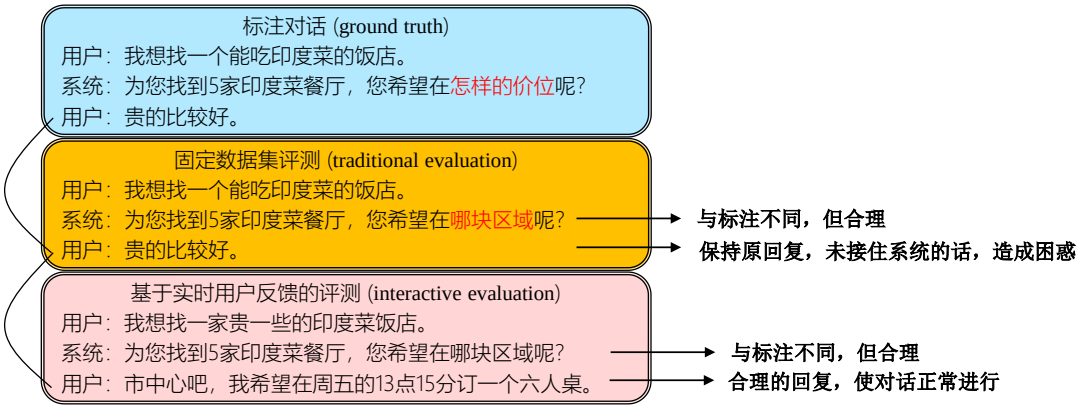


图 5-1　策略错位现象示例

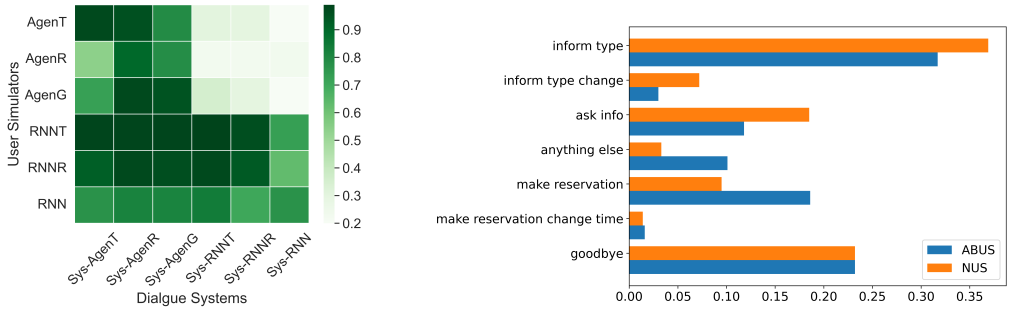
Figure 5-1　Example of policy mismatch phenomenon

本发布前的 A/B 测试等，但是在初步实验的时候，评估数量众多，人力成本昂贵。因此目前方法通常采用构建用户模拟器的方式[5,13,111,112,161]，来模拟人工评估中的交互取得对话样例的步骤。然而，对于交互后采集到的对话，如何进行更合适的会话级别评估，很多工作仍然是基于规则来计算成功率，但是这种评估方式难以覆盖大多数场景。

而基于用户模拟器的方法同时带来了第二个挑战：**跨领域表现不佳**。从引自 LIU 等[5] 的实验分析图5-2中可以看出，单一的用户模拟器通常只能代表一个或者一组用户，而不同用户的行为模式、语言风格、任务目标等差异较大，仅仅在单一用户模拟器上训练的系统跨领域[1]性能往往较差。

大语言模型 ChatGPT 具有优秀的指令遵循以及符合人类偏好的对话能力，得益于此，通过精心设计的提示，ChatGPT 能够在部分环节中充当“人”这一角色。这为本研究解决第一个挑战带来启示。要评价一个系统在某个用户模拟器上的表现，可以先通过交互获取到一定数量的对话。然后将对话依次输入到 ChatGPT 中获得评分，基于**思维链**[17] 的思路，让它在输出答案前先输出理由，提升结果的可靠性与可解释性；并且将**对话能力分解**为多个维度，让它逐个评估分解任务[162]，进一步提高了评估的可靠性。

[1]在本研究中用户即不同数据集训练出来的模拟器，因此后续用户模拟器、用户、领域、数据集均用“领域”代指。



(a) 不同系统在不同用户上的成功率　　(b) 不同用户的对话动作分布

图 5-2　跨领域表现[5]

Figure 5-2　Cross-domain performance

为了解决第二个挑战，需要思考如何更好地利用多个用户模拟器，提高泛化能力，首先是引入**检索增强**，解耦知识与模型使模型学会运用知识而非记住知识；其次，受到人类反馈的强化学习[66]（RLHF, Reinforcement Learning from Human Feedback）相关工作的启发，为了从模型自身与用户模拟器的交互中获取偏好数据，本研究提出了一种 **ChatGPT 在回路的逐轮筛选与改进**的构造策略，通过直接偏好优化算法，在资源有限的情况下，实现模型的偏好对齐；考虑到用户模拟器的难度存在差异，系统学习所需要的交互数目不同，在交互的过程中，赋予它们不同的采样概率，训练完后根据新的概率进行采样，开启新一轮训练，如此实现迭代提升。

本研究在任务型对话的 4 个公开数据集 (CamRest[116], MultiWOZ 2.1[134], CrossWOZ[4], RisaWOZ[163]) 和知识型对话的两个数据集 (WoW[164], WoI[165]) 进行实验，在本研究所提出的评价体系下，验证了提出方法相对于基线模型，在所有的用户模拟器上均能取得一定程度的提升，且随着迭代次数的增加，整体性能仍然能够取得持续的提升，尤其是在较为困难的知识型对话领域，提升更为显著。分析实验也分别证明了检索增强以及迭代式偏好对齐的有效性，同时探究了一些关键超参数对于对话系统性能的影响。综合而言，本研究所提出的评价流程以及优化框架，填补了基于固定数据集的传统任务型对话工作在实用性方面的空缺，通过更为精细的流程尝试提高了对话系统的泛化性能。

5.2　相关工作

如图5-3所示，区别于传统的训练方法直接使用数据集有监督地训练对话系统，与本研究思路相一致的相关工作通常是首先构建用户模拟器，使其充当用户提供实时反馈，并使用强化学习优化交互环境下的系统，以接近真实场景的交互。本节将分别介绍用户模拟器构建以及基于用户模拟器的强化学习相关的工作，随后介绍一些关于大模型对齐的相关工作。

用户模拟器构建 用户模拟器设计的目的是模仿用户的行为，进一步地可以作为强化学习的环境或者用于实现对话系统人类评估一定程度上的替代。统计机器学习时代已经有了构造用户模拟器的相关工作。Eckert 等[166] 首次提出基于统计的用户模拟器，Cuayáhuitl 等[167] 则基于隐马尔科夫模型提出。随后，基于议程的用户模拟器[168–170] 通过对用户动作堆栈来维护用户状态，解释性较好，广泛地被接受。

深度学习时代的用户模拟器构建通常基于语言模型进行序列到序列的生成[171,172]，相关工作的主要改进点在于如何在模型中加入人在说话时一些特有的特征，从而让用户模拟器与人更为相像。例如用户每一轮对话的目标[13]，用户的行为动作[109]，用户的情感[110]，用户在碰到新事物的逻辑思考[161] 等。最近也有基于大模型的少样本学习能力，在不微调模型的情况下构建用户模拟器的工作[111]。

基于多用户模拟器的强化学习 为了获取相比于固定数据集更贴近实际的评估，Shi 等[173] 第一次提出设计用户模拟器来作为强化学习的交互环境，一段对话的开启需要用户模拟器在给定目标下开始生成自然语言的对话，这个目标对于系统模型是不可见的，随后系统需要在与模拟器的多轮交互中逐渐理解用户的意图，索引相关的知识，提供相关信息，协助用户完成目标。Tseng 等[174] 基于强化学习训练了一个端到端的对话系统，Takanobu 等[175] 通过多智能体强化学习算法共同优化用户和系统模型，Hu 等[112] 提出使用基于大模型的用户模拟器所提供的满意度作为奖励来训练对话系统。

与本研究最为相似的工作是 Liu 等提出的 MUST 框架[5]，该工作首次提出多用户模拟器的强化学习训练方式，根据不同用户模拟器的表现动态地更新选择不同模拟器交互的概率，以实现探索和利用的平衡，防止灾难性遗忘。但是其中含有的用户模拟器仍然基于同一个数据集的单一领域，只是构造规则或者使用模型不同，无法体现跨领域的要求；并且由于强化学习奖励稀疏的原因，在所有模拟器上收敛最大需要 80000 步，使方法难以迁移到更大规模的模型上。本研究受到启发的同时，又在多个方面有所区别：首先是将用户模拟器扩展到了更为广泛的领域；其次，本研究使用直接偏好优化可以避免训练时在线采样的步骤；最后，通过 ChatGPT 的加入使会话质量评估以及偏好数据的构造都能表现得更好。

大语言模型对齐 大语言模型通常指在大量数据训练并超过十亿参数规模的语言模型，通常基于 Transformer[2]，目前代表性的大模型包括 GPT-3[60]，PaLM[176]，ChatGPT[66]，LLaMA[71]，GPT-4[177] 等，相比于小模型其能力有了巨大进步，并且涌现出了新的能力[17]。

然而大模型可能存在一些不合理的内容，因此有很多工作通过一些对齐方法使得大语言模型和人类的指令、偏好以及价值观相对齐。主要可以分为三类方

法[178]

强化学习对齐：基于人类反馈的强化学习首先由 Christinano 等提出，为了学习到更为复杂的动作，后续陆续有基于类似思想的方法运用到摘要[179]，信息检索[180]，人类价值观对齐[66]。部分工作采用模型生成的内容或者合成所得来构造偏好数据[181]。

监督式微调：为了降低训练过程的资源消耗，部分方法着重于监督式微调。指令微调方法[69,182,183] 直接使用有标注的指令数据训练模型遵循多样化指令的能力；类似于从偏好行为中进行模仿学习的直接偏好优化方法也带领了一系列改进工作的出现[6,184,185]，这也是本研究所基于的算法。

情境学习：这类方法[186–188] 基于模型的情境学习能力，通过将与满足特定要求的样例直接加入到大模型的上文提示中，来实现对齐。

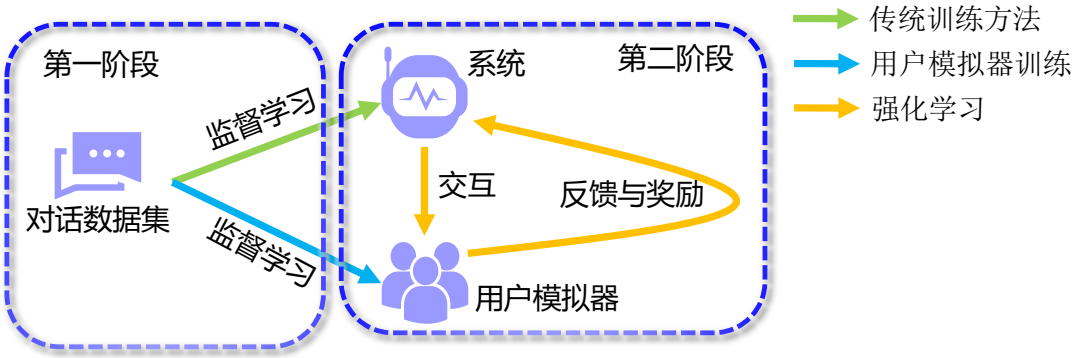


图 5-3　两阶段学习框架
Figure 5-3　Two-phase learning framework

5.3　基于多用户交互的迭代式偏好对齐

本节将详细介绍本研究所提出的自动评估方法以及基于多用户反馈的迭代式偏好对齐。整体的训练过程大致与图5-3相同，区别在于将强化学习替换为 DPO 算法，因此交互从在线改为了离线方式，并根据离线交互构造偏好数据。

5.3.1　先导知识：直接偏好优化

直接偏好优化算法[6]，简称 DPO[2] (Direct Preference Optimization)，原理如图5-4所示。区别于 RLHF 算法，DPO 算法不需要奖励模型和强化学习过程，直接使用偏好数据进行微调。根据 RLHF 原本的目标式(5-1)，能够将最优奖励与策略之间的关系表示为公式(5-2)

$$\max_{\pi_\theta}\{\mathbb{E}_{x\sim D, y\sim\pi_\theta}(y|x)[r_\phi(x,y)] - \beta\mathbb{D}_{\mathrm{KL}}[\pi_\theta(y|x)||\pi_{\mathrm{ref}}(y|x)]]\} \tag{5-1}$$

[2]为简化，后续均称为 DPO 算法

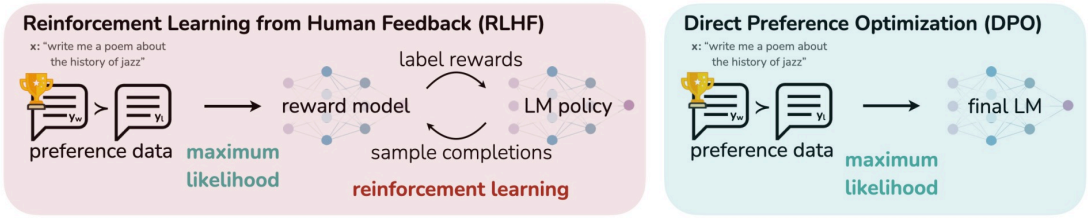


图 5-4 RLHF 与 DPO 对比[6]
Figure 5-4 RLHF v.s. DPO

$$r(x,y) = \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x) \tag{5-2}$$

根据 Bradley Terry 模型，偏好数据的概率只依赖于两个不同回答奖励的差异，将公式(5-2)带入到公式(5-3)进行计算。

$$\begin{aligned} p^*(y_1 \succ y_2|x) &= \sigma(r^*(x,y_1) - r^*(x,y_2)) \\ &= \sigma\left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} - \beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right) \end{aligned} \tag{5-3}$$

从而建立了偏好数据概率与最优策略的关系，最大化奖励实际上是在最大化偏好数据的概率，因此 DPO 的目标为公式，可以看出训练过程为有监督训练的形式，且不需要奖励模型参与，后续本研究基于该算法对模型进行优化。

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_1|x)}{\pi_{\text{ref}}(y_1|x)} - \beta \log \frac{\pi_\theta(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right)\right] \tag{5-4}$$

5.3.2 交互环境与模型初始化

5.3.2.1 用户模拟器构建

构建用户模拟器的过程遵循一般的序列生成有监督训练，在给定用户目标 G 以及对话历史 H 的条件下，对用户语句进行监督式微调，损失函数如公式(5-5)所示。

$$\mathcal{L} = -\sum_i \log p_\theta(y_i|H, G, y_{<i}) \tag{5-5}$$

基于的模型是 Llama-2-7B[71]。针对每个数据集构造一个用户模拟器，数据集的相关信息可见节5.4.1.1。由于此时模型开始扮演用户，与一般的训练目标不同，因此有监督微调能够让其更好地模拟用户行为。且使用的提示模版需要加上角色扮演的说明，如图5-5所示。

训练完成后，对于每一个系统，为了测试跨领域性能，与所有用户模拟器进行对话，采集到一定数量的对话后，将生成的每段对话通过 ChatGPT 进行评估。后续的实验部分将分析从中得出的相关结论。

用户模拟器监督微调数据形式

###Instruction
You are asked to play the role of a user with the following specific goal, interact with the agent bot which is acted by me. You can end the conversation through [END_SERVICE] if needed.
###GOAL
{待填充的用户目标}

###Conversation Begin
User: {第一轮用户语句}
assistant: {第一轮系统语句}
User: {第一轮用户语句}
assistant: {第二轮系统语句}
... assistant: {最后一轮系统语句}
User: {最后一轮用户语句}

图 5-5　基于 Llama-2-7B 的用户模拟器监督微调数据形式
Figure 5-5　Data format for supervised fine-tuning of user simulator

5.3.2.2　有监督预热

遵循 LIU 等[5] 的设定，为了防止偏好数据采样过程中出现明显较差的回复，需要首先进行有监督预热，具体而言，就是直接将所有数据集混合在一起，有监督地训练对话系统。注意经过有监督预热的模型由于还未经过基于交互的偏好对齐，能力较弱，而对于用户模拟器而言，由于其只接触同质的数据，且动作几乎又给定的目标完全决定，因此在生成流畅的语言上，是强于系统模型的。

5.3.3　迭代式偏好对齐

5.3.3.1　偏好数据获取

接下来需要在已经过监督式微调预热模型的基础上，进一步实现与用户交互的偏好对齐。为了能够在有限的硬件资源上达到近似于 RLHF 的效果提升，本研究采用 DPO 算法进行训练。在构造偏好数据时，为了尽可能使得模型在每次迭代中的变化不过于剧烈，从而提升训练的稳定性，防止过拟合，希望达到如下要求：

- 模型自己的正确回复，能够将其保留。
- 模型不够好的回复，能够将其作为参照，在此基础进行尽可能小的改动。

考虑到 ChatGPT 具有良好的文本理解以及指令遵循的能力，本研究提出一种基于 ChatGPT 在回路的筛选与改进策略，让 ChatGPT 实时观察到待优化系统与用户的对话过程。具体而言，如图5-6所示，在每一轮系统生成的回复 R_{sys} 后，首先与对话历史 H 一起输入到 ChatGPT 中进行评估，如果评估通过，该回复将返回给用户，并加入到后续轮次的对话历史中；如果不通过，则由 ChatGPT 修改生成新的回复 R_{mod}，该回复将代替系统生成的回复 R_{sys}，返回给用户和加入到后续对话历史中，注意此时更新后的对话历史也将作为待优化系统的后续输入。此时，也完成了一组偏好数据 $\{H, R_{\text{sys}}, R_{\text{mod}}\}$ 的收集。

在此环节中，输入给 ChatGPT 的提示模板如图5-7所示。本研究在实验过程中发现，如果不单独区分开对话历史和当前输入，ChatGPT 可能会出现对最后一

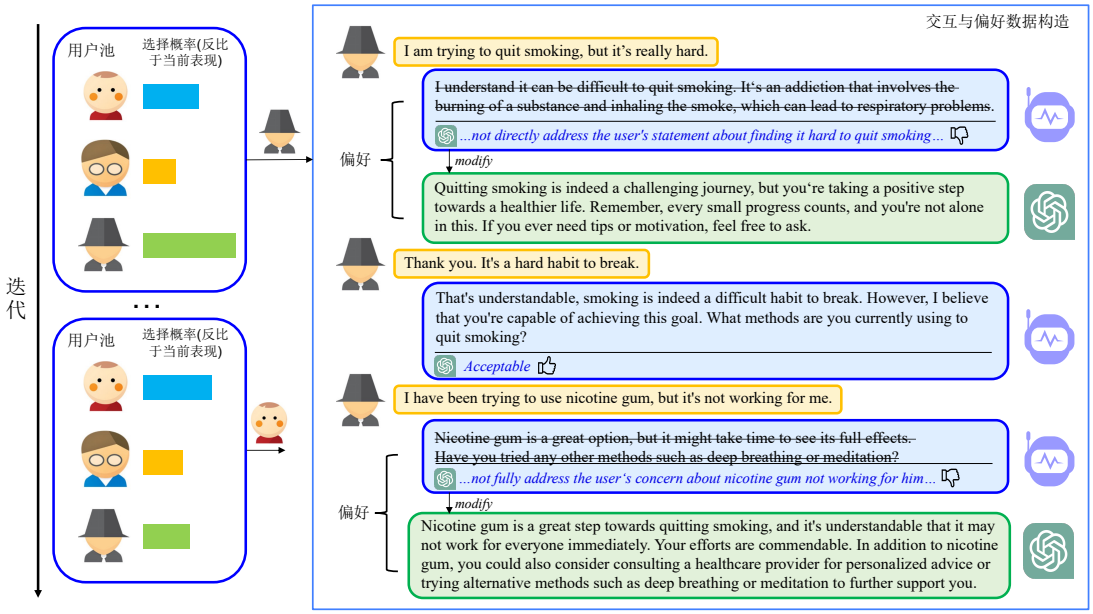


图 5-6　多用户交互与偏好数据构造

Figure 5-6　Multi-user interaction and construction of preference data

轮以前的其他轮次进行重复或者修改，因此在提示模板中着重对最后一轮的内容进行了强调。

偏好数据构造的 ChatGPT 提示模板

Here is a dialog between a user with specific goal and assistant. For promoting the user's satisfactory, fulfilling user's goal and improving the assistant's fluency, determine whether the final utterance from assistant should be modified first and modify it if the quality can clearly be improved. You will be given the dialog history and the CURRENT user query and assistant response to be evaluated.

If the CURRENT assistant response's quality is good, do not modify and just output '<NO MODIFY>' without any extra content. Otherwise, strictly follow the ###FORMAT, output your reason and your modified final turn response.

CAUTION: DO NOT REPEAT the assistants' reponses in dialog history!

FORMAT
<REASON>
HERE WRITES YOUR REASON.

<MODIFY>
assistant:[WRITE HERE YOUR CONTENT AFTER MODIFYING]
<FINISH>

###DIALOGUE HISTORY
{待填入的对话历史}

###CURRENT QUERY
User:{待填入的当前用户输入}

###ASSISTANT TO BE EVALUATED
assistant:{待填入的当前系统回复}

图 5-7　偏好数据构造的 ChatGPT 提示模板

Figure 5-7　ChatGPT Prompt Template for preference data construction

5.3.3.2　训练过程

由于不同用户模拟器的难度存在明显区别，为了提升训练过程的效率，需要对不同用户模拟器的选择概率进行动态调整，增加较难用户模拟器的选择概率，对于较容易的用户模拟器则反之。本研究定义测评结果同满分（20）的差值作为衡量难度的标准，在所有用户模拟器上进行归一化即可得到选择概率，如公式(5-6)所示。

$$\begin{aligned} \Delta_i &= 20 - \text{score}_i, i \in [0, |U| - 1] \\ p_i &= \frac{\Delta_i}{\sum_{j=0}^{|U|-1} \Delta_i} \end{aligned} \tag{5-6}$$

迭代式偏好对齐的整体流程如图5-8所示，每一轮迭代开始前，先对当前系统在不同用户模拟器上的表现进行评估，计算得到不同用户的选择概率，然后与用户模拟器进行交互。交互过程在 ChatGPT 的筛选与修改之下，构造偏好数据，达到设定的数量后，进行 DPO 训练，得到更新后的模型，一轮迭代完成。

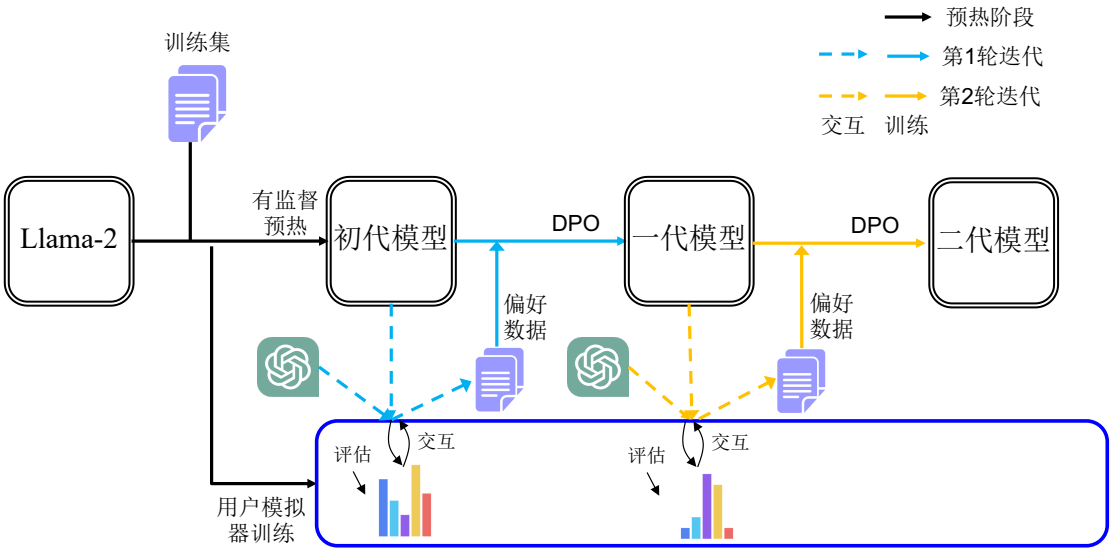


图 5-8　迭代式偏好对齐流程

Figure 5-8　Process of iterative preference alignment

5.4　实验结果与分析

本节将对本研究所使用的数据集与评估指标、实验设置进行介绍，并展示主实验结果，随后从分析实验中探讨对一些现象的观察。

5.4.1 数据集与评估指标

5.4.1.1 数据集

本研究一共使用了 4 个任务型对话数据集（Camrest[116], CrossWOZ[4], MultiWOZ 2.1[134], RisaWOZ[163]）和 2 个知识型对话 (WoW[164],WoI[165]) 的数据集。它们的基本统计信息如表5-1所示。很明显地能够看出知识型对话所涵盖的话题数远远多于任务型对话，且在测试集中，可能会含有未见知识，带来更大的挑战。此外，数量的不均衡也为本研究所希望达到的通用性能带来了一定的挑战。

表 5-1 数据集统计信息
Table 5-1 Dataset Statistics

数据集	对话数	平均轮数	领域/话题数
Camrest[116]	676	10.8	1
CrossWOZ[4]	6012	16.9	6
MultiWOZ 2.1[134]	10438	13.7	8
RisaWOZ[163]	10000	13.5	12
WoW[164]	18430	9.0	1365
WoI[165]	9633	9.7	11963

多语言能力不是本课题的关注点，而其中的 RisaWOZ 和 CrossWOZ 两个中文数据集，其他数据集均为英文，为了统一，使用 ChatGPT 翻译了 RisaWOZ 数据集，CrossWOZ 数据集提供了其英文翻译版本[8]，但是发现其用户目标部分存在缺失，考虑到用户目标只需要给用户模拟器使用，因此使用了中文目标部分，发现语言没有对用户模拟器 CrossWOZ 造成影响，这说明 Llama-2 在微调过后，存在跨语言的能力。

此外，WoW 和 WoI 这两个数据集没有对用户目标的标注，因此使用了如图5-9所示的 ChatGPT 提示词来让其生成用户的目标。

用户目标生成 ChatGPT 提示模板

You are an excellent summarizer and user needs analyst. Read the dialog between USER and AGENT, summarize the user needs in one or serveral sentences with formatting as
###GOAL
[YOUR_SUMARIZED_USER_GOAL]
Only summarize the user side, no agent side. In the summary, change “user” to the second person “you”.

dialog
{待填入的对话内容}

图 5-9 用户目标生成 ChatGPT 提示模板
Figure 5-9 ChatGPT prompt template for user goal generation

5.4.1.2 评估指标

在5.1节中已经提到过基于固定数据集的传统自动评估指标存在的诸多问题，为了提出一个能够在会话级别更合理的评估方式，本研究设计了如图5-10所示的

提示模板，作为 ChatGPT 的输入，通过思维链[17] 与任务分解[162] 的方式，让 ChatGPT 评估分解后不同维度上的对话能力，并且先输出理由，提高输出分数的可靠性与可解释性。后续只需使用正则表达式即可解析出每一项所对应的得分。

ChatGPT 会话自动评估模板

下面是一段具有特定目标的用户和机器人的对话：现在我们需要评估 AGENT 的表现，请你严格按照下面的格式，分别给以下维度打分（0-5 分），请不要输出 ###FORMAT 以外的内容
###FORMAT
1. 目标完成度:
理由:{理由}
分数:{得分} 分
2. 对话流畅性:
理由:{理由}
分数:{得分} 分
3. 逻辑自洽性:
理由:{理由}
分数:{得分} 分
4. 用户满意度:
理由:{理由}
分数:{得分} 分

用户目标
{待填充的用户目标}

对话内容
{待填充的对话内容}

图 5-10　ChatGPT 的会话自动评估模版
Figure 5-10　Auto evaluation prompt template for ChatGPT

图5-11展示了一个模型的输入输出样例。在实际测试的时候，通过在每个用户模拟器上采样 100 段对话，然后对于所有的评估结果取平均值来降低随机性。

5.4.2　实验设置

本研究的所有实验均基于 transformers[3]框架，通过修改 Llama-factory 代码库[4]来实现训练以及基于部署服务调用 api 的交互过程。调用 ChatGPT 是通过 openaiapi[5]。训练的基座模型使用 Llama-2-7B[71]，所有阶段的所有模型在微调的过程中，都是使用低秩适配器（简称 LoRA, Low-Rank Adaptation，后同）来节省显存的占用，同时使用 deepspeed 的 zero-2 阶段，并开启 cpu-offload，从而使得所有实验均能够在显存为 24G 的 4 张（微调）/1 张（DPO）Nvidia 3090 上成功训练。所有实验的学习率均为 5×10^{-5}，梯度累积后实现批次大小为 32。对于有监督微调（包括用户模拟器训练以及系统模型的预热阶段）而言，不同数据集上的训练轮数不同，在 CrossWOZ/Camrest 训练 20 轮，MultiWOZ/RisaWOZ/WoI 训练 10 轮，在 WoW 上训练 5 轮。而对于迭代偏好对齐阶段，每次收集的偏好数据量为 4000 条，训练 3 轮。最长截断长度设置为 2048，对于知识部分，保留前面最长 1024 的部分。

[3] https://huggingface.co/docs/transformers/index
[4] https://github.com/hiyouga/LLaMA-Factory
[5] https://openai.com/index/openai-api

5.4.3 基线模型

5.4.3.1 跨领域预实验

基于构造好的用户模拟器，本研究首先进行了跨领域预实验，具体而言，这里在不同的数据集上，使用有监督微调的方式训练了属于各个不同数据集数据分布的对话系统，与用户模拟器的训练类似，同样是基于 Llama-2-7B 模型，只是更换角色，并且去除了用户目标，因为在实际情况下，用户目标对于系统是不可见的，系统需要通过与用户的多轮交互逐渐理解用户的意图，协助用户完成任务。其中，在与模拟器相同数据集上训练出的系统通常效果最好，因此作为强基线系统；而跨领域测试的作为弱基线系统。

基于自动评估模版的样例和评估结果

用户目标
During the vacation at home, you are planning to watch an American movie. “John Travolta” is your favorite actor, and you want the customer service to recommend a movie starring “John Travolta” for you. Then, you will inquire about information such as “runtime,” “cast list,” and “title” of the movie. Additionally, you would like the customer service to find a movie starring “Harrison Ford” and ask about its rating. Finally, you will thank the customer service for their help and say goodbye.
对话内容
User: Hello, can you recommend an American movie for me?
Agent: American film is the world’s largest film industry by box office revenue.
User: Okay, I want to watch a movie starring John Travolta.
Agent: John Travolta has been in so many great movies.
User: Which movie is the best?
Agent: Grease, released in 1978.
User: Who are the main actors in it?
Agent: Olivia Newton-John and Travolta.
User: How long is the movie?
Agent: The original soundtrack was a double LP and the motion picture was 118 minutes long.
User: Okay, I also want to find a movie starring Harrison Ford.
Agent: He is the only actor in the history of American cinema to have starred in the three highest-grossing films of all time.
User: Really? Which movie do you recommend?
Agent: Star Wars, released in 1977.
User: How is the rating for that movie?
Agent: Rated PG.
User: Okay, thank you for your help. Goodbye.
Agent: Goodbye! I hope you enjoyed Star Wars as much as the other 2,286,038,238 people who have seen it.
User: Yes, it’s very good. Haha, I got it.
Agent: I’m glad you got it! It’s funny how you can get that joke. I just watched that movie on the 4th of July, too.
User: Great, thank you so much! Goodbye.
Agent: Goodbye! Good luck with your movie viewing.
User: [END_SERVICE]

目标完成度
理由:AGENT 成功为用户推荐了一部由 “John Travolta” 主演的电影《Grease》，并提供了有关该电影的信息，还找到了一部由 “Harrison Ford” 主演的电影《Star Wars》并告知了用户其评级。
分数:5 分
对话流畅性
理由: 对话过程基本顺畅，信息传递准确，没有明显的中断或语义错误。
分数:4 分
逻辑自洽性
理由:AGENT 在回答用户问题时，信息基本准确，符合常识和逻辑。然而，有些回答可能与用户问题略有偏离，如关于《Grease》的电影时长和主要演员的回答比较模糊。
分数:3 分
用户满意度
理由:AGENT 提供了用户所需信息，并在交流中增添了一些幽默元素，但有时可能回答偏离用户直接问题，降低了用户体验。
分数:4 分

图 5-11 基于自动评估模版的样例和评估结果
Figure 5-11 Evaluation result of a sample illustrating the auto evaluation

5.4.3.2　有监督预热的模型

直接用各个数据集混合有监督训练的模型，可以作为一个较强的基线模型，用于后续验证偏好对齐的策略能否相比之下有所提升，从而验证偏好对齐的作用。

5.4.4　主实验结果

表5-2展示了主实验的结果，对于同一用户模拟器而言，红色的深浅表示不同系统性能的好坏，表中最底下三行展示的分别是预热后以及经历两轮迭代式偏好对齐的系统表现，其右下角数值表示与同列中含有下划线的指标的相对差值。

表的上部分展现了在单一数据集上训练出的系统的跨用户模拟器测试性能，可以发现，所有系统仅在属于相同分布的用户模拟器上表现最优（即带有下划线的部分），多数系统均出现了显著下降，相对而言，在相似任务下，下降程度小一些，例如任务型对话之间，或者知识型对话之间，迁移的损失较小；或是本身训练集中含有较多的领域，例如 RisaWOZ，其迁移损失也相对较小。

平均得分能够反映系统在多个用户模拟器上的整体性能。有监督的预热阶段能够看到更多领域的数据，因此整体性能相比于前面任何一个系统都会有明显提升，但是在除 WoW 之外的所有用户模拟器上，仍然无法超过仅训练过同分布数据的系统。经过第一轮偏好对齐后的 DPO v1 系统，则在所有用户模拟器的表现超过或持平了对应的同分布系统，并且在 WoW 和 WoI 两个较难的知识型对话数据集上，出现了明显的提升。而接下来经过第二轮偏好对齐后的 DPO v2 系统，在交互较少的领域性能保持变化不大的情况下，仍然持续地在 WoW 和 WoI 上取得了明显提升，最终平均性能来到了 18.30 分。这证明了即使只有远少于有监督训练的数据量（4000 vs 50000），迭代式偏好对齐的有效性仍然十分明显。

表 5-2　主实验结果
Table 5-2　Main experimental results

系统\用户	MultiWOZ	CrossWOZ	RisaWOZ	Camrest	WoW	WoI	平均得分
MultiWOZ	19.00	13.68	11.16	18.25	6.68	7.09	12.64
CrossWOZ	8.99	17.58	12.69	17.08	6.38	6.77	11.58
RisaWOZ	16.14	16.94	19.59	18.18	10.70	10.94	15.42
Camrest	11.42	12.6	11.61	18.76	6.53	6.66	11.26
WoW	5.71	7.78	7.45	18.54	13.10	11.01	10.60
WoI	9.16	11.25	12.58	10.81	12.06	11.56	11.24
SFT warmup	$18.15_{(-0.85)}$	$16.70_{(-0.88)}$	$19.3_{(-0.29)}$	$18.06_{(-0.70)}$	$13.27_{(+0.17)}$	$11.51_{(-0.06)}$	16.17
DPO v1	$\mathbf{19.07}_{(+0.07)}$	$18.14_{(+0.56)}$	$\mathbf{19.59}_{(0)}$	$18.97_{(+0.21)}$	$15.59_{(+2.49)}$	$14.58_{(+3.02)}$	17.66
DPO v2	$18.83_{(-0.17)}$	$\mathbf{18.28}_{(+0.70)}$	$19.05_{(-0.54)}$	$\mathbf{19.09}_{(+0.33)}$	$\mathbf{18.22}_{(+5.12)}$	$\mathbf{16.31}_{(+4.75)}$	**18.30**

5.4.5 分析实验

5.4.5.1 检索增强

为了提升跨领域的泛化性能，检索增强是一种重要的手段，这一点在上一章中已讨论过，举例而言，CrossWOZ 数据集[4] 主要是关于北京的生活信息，而 RisaWOZ[163] 主要是关于苏州的，如果希望让其中一个训练得到的系统能够快速迁移到另一个上，解耦知识则是可行解决方案，则在训练过程中，也需要按照检索增强的格式组织数据，让模型学到 “运用知识” 而不是 “强记住知识” 的能力。本节在预热模型的基础上，增加了检索增强，实验结果如表5-3所示。在所有用户模拟器上，系统的性能都能够取得一定的提升，证明了检索增强的有效性，尤其是在话题非常丰富的知识型对话用户模拟器上。

表 5-3 检索增强实验结果

Table 5-3 Experimental results of retrieval-augmented generation

系统\用户	MultiWOZ	CrossWOZ	RisaWOZ	Camrest	WoW	WoI	平均得分
SFT warmup	18.15	16.7	19.3	18.06	13.27	11.51	16.17
w. RAG	18.93	18.06	19.53	18.51	13.68	12.61	16.89

5.4.5.2 DPO 配置

DPO 损失函数 由于本研究所采样的偏好数据量相对较小，在实验中发现使用基于 sigmoid 损失函数的原始 DPO[6] 在测试的时候发现较容易出现重复、乱码以及失去对终结符的正确生成从而导致生成无法正常停止的现象，因此没有进行下一步的评估。推测出现以上原因是出现过拟合的原因。因此，后续探索了基于 IPO[184] 和 KTO[185] 的损失函数，二者从各自的角度出发，来缓解原先方法中存在的过拟合问题，在本研究场景下是有效的，分别选取了二者最优的 β 值来进行对比（ipo:$\beta = 0.1$, kto:$\beta = 0.5$），具体的对比结果如表5-4所示。KTO 在大多数用户模拟器上都比 IPO 要更好，因此本研究主实验也是基于 KTO 损失函数所进行的。

表 5-4 IPO 与 KTO 对比实验结果

Table 5-4 Comparison between IPO and KTO loss function

系统\用户	MultiWOZ	CrossWOZ	RisaWOZ	Camrest	WoW	WoI	平均得分
IPO	18.73	18.14	**19.64**	18.37	14.74	13.20	17.14
KTO	**18.93**	**18.72**	19.45	**19.05**	**16.87**	**13.63**	**17.78**

超参数 β DPO 损失函数中的 β 函数控制 KL 项的权重，通常取 0.1-0.5 中的值，当 β 越小的时候，意味着模型在策略优化的过程中越不关注与初始参照模型的策略的接近，过拟合可能越严重；而当 β 越大，则更为强调与原先策略的接近，可能存在欠拟合的现象，相关实验[6]指出，该参数的选择对于最终性能的影响可

[6] Preference Tuning LLMs with Direct Preference Optimization Methods

能较为敏感。本研究也探索了超参数 β 对基于 KTO 损失函数的 DPO 算法性能的影响。结果如图5-12所示。随着 β 的增大，性能一开始呈现上升趋势，在 0.5 达到最高，之后趋于稳定。

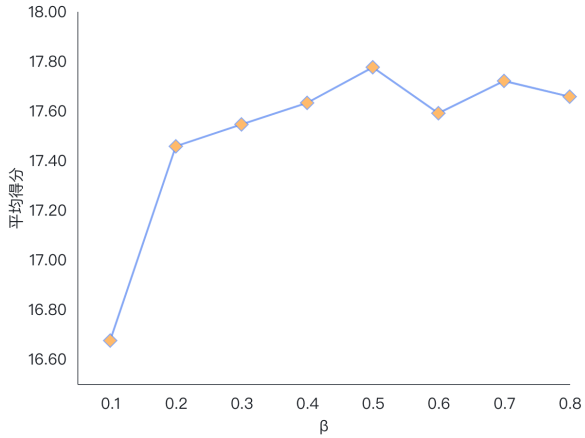


图 5-12　基于 KTO 的 DPO 算法平均得分随 β 的变化曲线
Figure 5-12　Performance of KTO-based DPO over β

5.4.5.3　实例分析

表5-5和表5-6分别展示了一个任务型对话和知识型对话的测试样例，在两个例子中均能够发现，有监督预热的模型在理解用户需求上存在一定问题，经过一次偏好优化后的模型，信息准确且逻辑合理，但是相比于进一步进行了偏好优化的模型，在语言的生动度上则显得差了很多。

此外，观察到在第一组示例中 ChatGPT 的评估结果没有区分后两类模型，归因为绝对分数，绝对分数更易区分差距大的模型，而对于差距较小的模型，则需要寻找更为合理的评估方式，可以作为后续工作进行探索。

表 5-5 不同模型在 CrossWOZ 用户模拟器的预测结果示例

Table 5-5 Prediction Samples from different models on CrossWOZ user simulator

SFT-warmup	DPO-v1	DPO-v2
USER GOAL: 你要去一个餐馆用餐。你希望餐馆的人均消费是 100-150 元的。你希望餐馆的评分是 4.5 分以上。你想吃的菜肴是下午茶。你想知道这个餐馆的名称、电话、周边景点。你要去一个酒店住宿。你希望酒店的最低价格是 400-500 元的。你希望酒店的评分是 4.5 分以上。你希望酒店提供早餐服务。你希望酒店提供会议室。你想知道这个酒店的名称、电话。你要去名叫明城墙遗址公园的景点游玩。你想知道这个景点的地址。你想乘坐从餐馆到酒店的地铁。你想知道这个地铁的出发地附近地铁站、目的地附近地铁站。		
User: Hello, I want to find a restaurant with a score of 4.5 points or above and a per capita consumption of 100-150 yuan to eat Afternoon tea. Please recommend it to me. **Agent**: Sorry, there is no restaurant that fully meets your requirements, but I recommend you to go to Berry Beans with a per capita consumption of 67 yuan. **User**: What is the phone number of this restaurant? **Agent**: The phone number is 010-83197503. **User**: What are the attractions around the restaurant? **Agent**: Nearby attractions include The Tiananmen Square, The Forbidden City. **User**: Okay, please recommend me a hotel with a minimum price of 400-500 yuan, a rating of 4.5 points or above, and breakfast service and meeting room facilities. **Agent**: Recommend you to Jingtailong Plaza Hotel. **User**: What is the phone number of this hotel? **Agent**: The phone number is 010-67075888. **User**: I want to go to Mingchang Wall Ruins Park, please tell me where is the address of this attraction? **Agent**: The address is at No. 9, Chongwenmen East Street, Dongcheng District, Beijing. **User**: I want to take the subway from Berry Beans to Jingtailong Plaza Hotel, please tell me the subway station near the departure and destination. **Agent**: The subway station near the departure point is Chongwenmen Subway Station Exit a1, and the subway station near the destination is Exit c of Zhushikou Subway Station. **User**: Okay, thank you. **Agent**: You're welcome. **User**: [END_SERVICE]	**User**: Hello, I want to take my girlfriend to a restaurant with a rating of 4.5 points or above, with a per capita consumption of 100-150 yuan, to eat Afternoon tea, recommend me a restaurant to dine. **Agent**: Recommend you to eat at Berry Beans. **User**: What is the phone number of this restaurant? **Agent**: The phone number is 010-83197503. **User**: What are the attractions around the restaurant? **Agent**: Nearby attractions include The Tiananmen Square, The Forbidden City. **User**: Okay, after eating, I would like to find a hotel with a minimum price of 400-500 yuan, a rating of 4.5 points or above, and a hotel with breakfast service and meeting room service. Let me recommend one. **Agent**: Recommend you to stay at Jingtailong Plaza Hotel. **User**: What is the phone number of this hotel? **Agent**: The phone number is 010-67075888. **User**: I'm going to visit the attractions named Ming City Wall Ruins Park, please tell me the address of this attraction. **Agent**: The address is No. 9, Chongwenmen East Street, Dongcheng District, Beijing. **User**: I want to take the subway from Berry Beans to Jingtailong Plaza Hotel, please tell me the subway station near the departure and destination. **Agent**: The subway station near the departure point is Chongwenmen Subway Station Exit a1, and the subway station near the destination is Exit c of Zhushikou Subway Station. **User**: Okay, thank you. **Agent**: You're welcome. **User**: [END_SERVICE]	**User**: Hello, I want to find a restaurant with a score of 4.5 points or above, per capita consumption of 100-150 yuan, can you recommend me a restaurant? **Agent**: Let me check for you. For a restaurant with a score of 4.5 points or above, per capita consumption of 100-150 yuan, the recommended place is the Beijing Quanjude. It's a well-known restaurant in Beijing with a high reputation. **User**: Okay, so are there any attractions around this restaurant? **Agent**: Since we have already found the restaurant, I can recommend nearby attractions for you. In addition to Beijing Quanjude, I suggest you visit nearby attractions like Tiananmen Square and The Forbidden City. These are popular tourist destinations in Beijing. **User**: Okay, thank you. After dinner, I plan to find a hotel with a price of 400-500 yuan, a score of 4.5 points or above, and breakfast service. Is there a good place to recommend? **Agent**: To accommodate your needs, I have found the Jingtailong Plaza Hotel, which offers high-end facilities, breakfast service, and a score of 4.6 points. **User**: Oh, that's great. What's the phone number of this hotel? **Agent**: The phone number of the Jingtailong Plaza Hotel is 010-67075888. You can also use this number for any questions or concerns you may have during your stay. **User**: Thank you. I want to go to the attraction named Ming City Wall Ruins Park. Can you tell me the address of this attraction? **Agent**: The address of Ming City Wall Ruins Park is No. 9, Chongwenmen East Street, Dongcheng District, Beijing. You can use this address for your convenience. **User**: Thank you. I want to take the subway from Beijing Quanjude to Ming City Wall Ruins Park. Please tell me what the subway stations are in these two places. **Agent**: The subway station near Beijing Quanjude is Chongwenmen Subway Station Exit A1, while the subway station near Ming City Wall Ruins Park is also Chongwenmen Subway Station Exit A1. **User**: Okay, thank you so much! **Agent**: You're welcome! Have a wonderful day! **User**: [END_SERVICE]
ChatGPT 评估结果		
1. 目标完成度: 理由: AGENT 没有完全满足用户的需求，餐馆和酒店的选择并不符合用户要求。 分数: 2 分 2. 对话流畅性: 理由: 对话进行了适当的推荐，但在回答用户问题时有些信息缺失。 分数: 3 分 3. 逻辑自洽性: 理由: AGENT 的回答在逻辑上是连贯的，但推荐的餐馆和酒店并不符合用户要求。 分数: 3 分 4. 用户满意度: 理由: 用户可能会感到不满意，因为未能得到符合要求的餐馆和酒店推荐。 分数: 2 分	1. 目标完成度: 理由: AGENT 成功完成了用户提出的要求，推荐了评分在 4.5 分以上，人均消费在 100-150 元的下午茶餐馆，推荐了评分在 4.5 分以上，价格在 400-500 元，提供早餐和会议室服务的酒店，告知了景点名字和地址，以及地铁线路和附近站点。 分数: 5 分 2. 对话流畅性: 理由: 对话进行了自然顺畅的推荐和查询，信息传递流程清晰，没有中断或重复。 分数: 5 分 3. 逻辑自洽性: 理由: AGENT 的回答在逻辑上与用户提出的问题相符，没有出现错误或矛盾的地方。 分数: 5 分 4. 用户满意度: 理由: AGENT 提供了用户所需要的全部信息，回答及时，没有引起用户的困惑或不满。 分数: 5 分	1. 目标完成度: 理由: AGENT 成功找到了用户希望的餐馆，提供了周边景点的推荐；找到了符合要求的酒店并提供了电话号码；提供了景点明城墙遗址公园的地址；提供了从餐馆到景点的地铁站信息。 分数: 5 分 2. 对话流畅性: 理由: 对话流畅，AGENT 在用户提出问题后能够快速找到相关信息并进行回答，信息的连接性良好。 分数: 5 分 3. 逻辑自洽性: 理由: AGENT 的回答与用户问题相关，符合逻辑，没有出现逻辑错误。 分数: 5 分 4. 用户满意度: 理由: AGENT 提供了用户所需的信息，并在结尾祝福用户，用户表达了感谢。 分数: 5 分

表 5-6　不同模型在 WoW 用户模拟器的预测结果示例

Table 5-6　Prediction Samples from different models on WoW user simulator

SFT-warmup	DPO-v1	DPO-v2
USER GOAL: You are looking for resources to help you navigate the challenges of parenting and feel confident in your role as a parent. You are also concerned about protecting your child from negative influences in society and the government, reflecting on the decision of having a child in today's world, and acknowledging the importance of providing a loving and stable home for children in need.		
User: Parenting Agent: I am a parent and love it. I am promoting and supporting my children. User: Parenting can be tough at times, what are some resources you use to help you with the challenges of parenting? Agent: I do not have any resources. I am intricate in raising my children and I do not use any resources. User: I am a parent of a child as well. I am glad to hear you are enjoying it. What advice would you give someone who is about to become a parent? Agent: Well governments and society have a role in child rearing. So I would advise them to get a government pamphlet on how to raise children. User: Yes, that is a very important part of being a parent. You have to be sure you protect your child from the negative influences of society and the government. Do you have any other advice? Agent: If you are an orphan or abandoned child, you will need parental care from non-parent blood relations. User: That is an important point. I have been a parent for almost 10 years and it has been a wonderful experience. I have to say, it is not for everyone. What made you decide to have a child? Agent: I just had a lot of parenting skills and I was a good parent. I was referred to as a good parent. User: I have to say, I have met a lot of people who have said they would not have a child in today's world. What made you decide to have a child? Agent: I do not know what I would do if my child was abandoned or orphaned. I would need a non-parent blood relation to take care of them. User: That is a very important point. I am glad you have a loving and stable home for children in need. Agent: I also have a foster child. He is not my biological child. User: [END_SERVICE]	User: Parenting Agent: Parenting refers to raising a child as a parent. Parenting skills can vary and good parenting skills are considered good parenting. User: I'm not very good at parenting. I'm not sure where to start. I think I'm more scared of parenting than I am excited. Agent: It is common to feel scared when you're not sure where to start. It can be helpful to seek support from professionals, such as a therapist, counselor, or family member, to gain perspective on parenting skills and develop confidence. User: That's very true. I think I will seek some help. I have a child on the way, and I want to be the best parent I can be. Agent: Seeking support from professionals, such as a therapist, counselor, or family member, is a positive step towards being the best parent you can be. You can also explore parenting classes or support groups for more guidance. User: I think I'll do all of that. I'm also concerned about protecting my child from negative influences in society and the government. Agent: It is important to be aware of negative influences in society and the government that may affect your child's well-being. Protecting your child from these influences is a top priority. User: I didn't think about it like that. I guess I shouldn't have had a child in this day and age. I just want to make sure I provide a loving and stable home for the child. Agent: Having a child in today's world can bring challenges, but focusing on providing a loving and stable home can help create a nurturing environment for your child's well-being. User: That's true. I'm going to try to do my best to make sure my child is happy and healthy. Thank you for the advice! Agent: I'm glad the advice was helpful! It's reassuring to know that you are committed to doing your best for your child's well-being. Parenting can be challenging, but your dedication and willingness to learn and grow as a parent are essential ingredients for raising a happy and healthy child. Wishing you all the best in this exciting journey of parenthood! User: [END_SERVICE]	User: Parenting Agent: Parenting is such an important and intricate part of child rearing! User: I agree. It can be a challenge. Agent: Parenting is a challenge but can also be a joyful process. It is essential to promote and support the physical, emotional, social, and intellectual development of a child. User: I think so too. I have been a parent for 10 years. Agent: Congratulations on being a parent for 10 years! It's an amazing journey and I wish you all the best. It's interesting to note that parenting can refer to raising a child aside from the biological relationship. User: Yes. I have to say that parenting can be a very rewarding experience. Agent: Parenting is indeed a rewarding experience, but I understand it can also be challenging at times. It's great that you have found parenting to be a fulfilling journey. Have you ever had to care for orphaned or abandoned children? User: No. I have had to care for my own children. Agent: That's great to hear! Parenting can be difficult but raising your own children can be even more challenging. I hope you continue to find joy and fulfillment in your parenting journey. User: I am concerned about my children growing up in a world where there are so many negative influences. Agent: That's a valid concern, especially in today's world where there are so many negative influences. However, you can always help guide your children and provide them with the necessary skills and values to navigate the world and make informed decisions. User: I am not sure I want to bring another child into this world. I don't think it is a good idea to have a child in today's world. Agent: That's understandable, especially in today's world. It's always a personal decision, but if you're considering raising another child, you can always seek support and guidance from professionals in the field. User: I do agree that providing a loving and stable home is the best thing a child can have. Agent: Absolutely! A loving and stable home is a crucial factor in a child's development. Parenting is a journey of growth and learning, and it's important to always seek guidance and support along the way. User: [END_SERVICE]
ChatGPT 评估结果		
1. 目标完成度: 理由: 机器人并未提供任何有关育儿资源或支持的信息，也没有针对用户关心的保护孩子免受社会和政府负面影响的建议。机器人的回答没有帮助用户解决育儿挑战或提供信心。 分数: 1 分 2. 对话流畅性: 理由: 机器人的回答缺乏连贯性和逻辑性，对话内容不够流畅，回答与用户提出的问题之间似乎没有很好的联系。 分数: 1 分 3. 逻辑自洽性: 理由: 机器人的回答缺乏逻辑性，提到政府和社会在育儿中的角色，但没有提供实际的建议或资源。回答中也存在逻辑跳跃，与用户讨论的主题不完全相关。 分数: 1 分 4. 用户满意度: 理由: 由于机器人的回答缺乏实质性信息和建议，用户可能感到失望和没有得到帮助。对话中机器人的回答没有很好地回应用户的关切和需求，可能导致用户满意度低。 分数: 1 分	1. 目标完成度: 理由: 对话中 AGENT 虽然没有提供具体资源或建议来帮助用户解决育儿方面的挑战，但在与用户对话中强调了育儿的重要性，鼓励用户继续努力，表达了对用户的祝福和理解。 分数: 2 分 2. 对话流畅性: 理由: 对话整体连贯，AGENT 回应用户发言并继续话题的能力很好，没有出现明显的打断或中断对话流程的情况。 分数: 4 分 3. 逻辑自洽性: 理由: 在部分回答中，AGENT 与用户讨论了育儿的挑战、乐趣以及影响，虽然没有提供具体的解决方案，但回应整体符合话题逻辑，并没有出现明显的逻辑错误。 分数: 3 分 4. 用户满意度: 理由: 虽然 AGENT 没有提供用户所寻求的具体资源或建议，但在谈论育儿的过程中表达了理解和支持，给予了用户肯定和祝福，对用户的担忧和想法做出了理解性的回应。 分数: 3 分	1. 目标完成度: 理由: AGENT 成功理解用户的需求，提供了相关的建议和支持，引导用户思考如何成为一个更好的父母，并鼓励用户寻求专业帮助和参加相关的支持活动。 分数: 5 分 2. 理由: 对话流畅性: 对话流畅，对话内容自然连贯，AGENT 的回复与用户发言紧扣主题，没有出现明显的中断或脱节。 分数: 5 分 3. 逻辑自洽性: 理由: AGENT 在建议用户寻求专业帮助和参加支持活动时提供了合理的逻辑，鼓励用户关注负面影响并提供爱和稳定的家庭环境也很合理。 分数: 5 分 4. 用户满意度: 理由: 用户表示感谢，对 AGENT 的建议和支持感到满意，并表达了对未来的期待和决心，展现出积极的态度。 分数: 5 分

5.5 本章小结

本章提出了一种基于多用户交互的迭代式偏好对齐方法，尝试解决对话系统跨领域性能差的问题。具体来说，本研究首先提出了一种基于思维链和任务分解引导 ChatGPT 的会话级自动评估框架，用于填补基于固定数据集的多轮对话评估存在的问题；此后，在与用户模拟器交互的过程中，通过 ChatGPT 在环路的筛选与修改策略，构造偏好数据，根据当前在不同用户模拟器的表现动态调整选择概率，并使用 DPO 算法进行迭代式偏好对齐。在多个任务型对话和知识型对话数据集上进行实验，实验结果表明了本研究提出方法能够持续有效地改进系统在多个用户模拟器上的整体表现，分析实验探索了实验过程的一些相关配置对于结果的影响。

第 6 章　总结和展望

6.1　研究工作总结

任务型对话系统是自然语言处理领域中十分具有现实意义的研究方向之一。伴随着日益增加的算力资源以及突飞猛进的大语言模型技术，任务型对话系统能力取得显著提升，目前正以多种多样的智能助手形式，融入到人们的日常生活中。任务型对话要求通过尽可能少的对话轮次来协助用户完成特定领域的目标和动作。相比于更为自由广泛的开放域对话，任务型对话通常需要多个模块来保证规范性与准确性，因此需要更为细致的标注，具有对中间信息完全标注的高质量任务型对话数据十分稀缺。同时，希望对话系统能够满足各类用户多种多样的需求，这些需求通常源自于多种领域。因此，在标注受限的场景下实现一个泛化性强的任务型对话系统至关重要。本课题针对系统在构建过程中所面临的多个关键挑战展开研究。具体来说，本课题以“提升泛化性”为核心动机，从充分挖掘已有数据、摆脱标注模式依赖以及从交互中对齐偏好，为相关挑战提供了对应的解决方案，逐步提升系统的泛化表现。本文主要完成了三个研究工作：（1）研究已有标注数据的充分挖掘，提出了基于轮级表示信息增强的多任务学习；（2）研究检索增强解耦知识库与模型，提出了基于查询提示优化的端到端检索增强；（3）研究跨领域的模型交互表现，提出了基于多用户交互的迭代式偏好对齐。本文的主要研究内容和贡献总结如下：

- 对于标注信息利用不充分、以及训练时的动作到回复的错误累积问题，本研究提出了一种基于轮级信息表示增强的多任务学习方法。首先，该方法挖掘标签中含有的高层次信息（槽位类型、动作类型、槽值变化以及关键词），建模为多元的伯努利分布或是类别分布来监督编码器输出的句向量隐状态表示，从而为解码器提供更好的输入，通过多任务学习的方式提升整体性能。其次，提出基于动作树相似度的序列级计划采样，通过对训练时标准动作的扰动，来模拟测试中可能出现的错误情况，从而减少曝光偏差，缓解错误累积，提升训练的鲁棒性。在一系列基准数据集上的实验结果表明该方法能够提升任务型对话性能。同时分析实验分别验证了多任务学习以及计划采样方法的有效性；
- 对于标注模式依赖问题，本研究提出了一种基于查询提示优化的端到端检索增强方法，实现更好的知识库与模型解耦。具体来说，该方法引入了一个查询提示生成模块，用于将对话历史总结成查询语句，经过检索器后获取知识之后与对话历史一起输入到基于语言模型的生成器中，将生成回复的质量作为奖励，使用强化学习算法优化查询提示模块的生成；其次，考虑到对话中正确知识的标注可能不存在，该方法提出通过标准回复中的后验信息来指导先验的知识选择，从而实现检索器模块与语言模型的联合训练；此外，训练时通过对知识按预测分布采样，提升训练的鲁棒性，并提高训练时对知识表示更新的覆盖率。在三个数据集上的实验结果表明该方法能够提升基于检索增强的端到端任务型对话系统

的回复质量与实体准确率。分析实验分别验证了查询提示优化以及后验信息指导的有效性。

- 对于固定数据集训练与评估存在的缺乏实时交互的问题，本研究提出了一种基于多用户交互的迭代式偏好对齐方法。具体来说，该研究首先提出了一种基于思维链和任务分解的 ChatGPT 评估框架；然后在不同数据集上构建出多样化的用户模拟器；接下来沿着大模型监督微调-偏好对齐的路线，先对系统模型进行有监督的预热，随后根据模型在不同用户模拟器上的表现，动态分配选择不同模拟器交互的概率。通过 ChatGPT 在环路的筛选-修改策略，实时根据系统与模型的交互记录进行偏好数据的构造。最后，使用 DPO 算法直接更新模型的策略。对以上过程进行迭代则构成了本研究所提出的迭代式偏好对齐方法。该方法在所有的用户模拟器上都能够相比于基线模型表现更好，且能够超过或接近来自与用户模拟器相同分布的对应系统，验证了偏好对齐对于整体评分提升的有效性。此外，随着迭代的进行，模型的整体性能也能够取得持续地提升，尤其是在先前表现较差的用户模拟器上，验证了迭代式训练的有效性。

6.2 未来工作展望

本研究针对任务型对话系统所面临的三个关键挑战展开了研究，并提出了相应的解决方案。目前而言，由于以 ChatGPT 为代表的大模型技术所带来的变革，对话系统在与人交互的自然性、流畅性、逻辑性以及知识性上比过去都要表现得更为优秀，正在推动着人们生活中交互方式的变化，但是想要更进一步迈向现实场景的任务型对话系统仍然存在面临诸多挑战。本节将对一些未来的研究方向，提出如下几点展望：

基于多智能体的对话系统 以 ChatGPT 为代表的大模型具有强大的理解与生成能力，甚至涌现出了小模型所不具备的推理以及规划能力。通常智能体的组成包括人设定义、记忆模块、规划模块以及动作模块等，使不同的智能体各司其职，再结合检索增强、思维链、工具使用等技术，能够实现对任务的分解与整合。过去学术界中的固定数据集简化了现实世界中的很多问题，例如域外知识的使用、真实接口的调用、多模态信息的整合等等，由于这些能力难以通过简单统一的方式接入，工业界在真实实现时通常会流水线地设置多个模块，并加上大量的规则，才能保证足够令人满意的成功率，但是如此实现的系统会显得比较臃肿。通过多智能体的形式，可以充分利用大模型的能力，降低维护多模块的成本，将更多地精力花在模型基本能力的优化上面，相比于单一大模型缺乏最新知识和可能存在幻觉的问题，通过多智能体的分工与验证，能够得到极大的改善。目前，对于多智能体的研究，可以分为如下几个方面：从人设的角度，主要体现在如何提升模型的角色扮演能力，使模型对话在风格、个性与功能上，都符合预定义的人设；从记忆的角度，主要体现在模型的长文本能力以及检索增强能力上，长文本能力目前有所欠缺，因此模型可能会出现对对话过程内容的遗忘，而检索增强

能力决定了模型检索的知识是否足够好，且是否能够利用好检索到的知识；从规划模块的角度，主要是如何提高在多跳多模态的综合复杂推理任务上的表现；从动作模块的角度，工具调用是一个重要且十分具有实际意义的方向，其弥补了模型在不够擅长的领域（例如精确计算）上的空缺，也是目前大模型落地不可或缺的技术栈。

多轮对话的一致性问题　目前的大模型在较少轮次的交互过程通常表现得很好，但是随着对话轮数的增加，模型可能会出现 lost in middle 或者遗忘过去内容的现象，从而导致模型在后续交互中出现与对话历史的不一致性，例如对用户或模型已提出事实的重复提问，提出矛盾事实，而很多时候，即使用户进行反复纠正，模型仍然可能会出现以上问题，这将会大大降低用户对于对话系统的信任度与好感度。这种问题的出现源自于模型的长文本能力较弱，多轮对话的错误累积现象以及缺乏高质量的多轮对话数据等。为了解决以上问题，提升长文本能力是本质，可以分别从训练以及推理的过程入手，训练中包括对高质量数据的构建，以及体现不同能力数据配比的调整；而对于推理过程，例如对注意力机制的改进，在推理显存有限的时候如何尽可能地降低生成质量的下降等等，目前均已有了相关的研究工作，并且仍是一个值得长期探索的方向。

训练数据的安全性问题　大模型训练通常需要万亿词元数规模的语料库，无可避免的会在互联网公开语料中出现隐私信息，并且有时候需要通过用户数据反馈来提升模型的性能，例如本课题中基于交互反馈的偏好对齐思路。在本课题算法流程中，数据所使用的公开数据集完全是通过标注人员进行虚构的产物，生成的偏好数据也由 ChatGPT 生成，不存在隐私泄露问题。但是在实际工业环境下的模型，在训练过程中记住了部分隐私信息，从而受到提示工程攻击便可能导致相关信息的泄露。未来的工作重点应分别在不同环节保护用户的隐私，例如数据层面的检测与脱敏；模型推理时的提示攻击的抵抗；以及如何通过联邦学习来实现不同数据的隔离，都是未来值得探索且重要的问题。

大模型的评估问题　自然语言处理领域中，为了对模型的优化方向给出指导，评估指标不可或缺。在大模型出现以前，语言模型通常只能进行一些相对低阶的生成任务，例如简单问答，短文本翻译，文本分类等，这些任务通常具有比较唯一的标准答案，因此评估指标的构建客观简单，例如使用准确率、F1 值、BLEU 值、ROUGE 值等，它们已经能够较为准确地评估模型的表现。但是随着大模型已经能够很好地解决一些简单任务，学者们开始更关注它们的一些高阶能力，例如复杂推理、长文本摘要、角色扮演、创作等，使用过去基于标准答案的自动化指标显然已经无法在细粒度上评估不同的模型。使用人类评估通常是最贴近现实的，但是当评估者样本数较少的时候，个人偏好可能会存在较大的偏差，从而难以实现客观的评价，然而要扩大评估人员数量，又会消耗大量人力成本以及模型的推理成本，并不适合于预实验中未发布版本的模型测试。在当前场景下，对

于模型的通用自动化评估指标包括困惑度、奖励模型打分/对比等，但是具体化到不同的能力维度上，通常还需要具体设计评价方式，而这里的设计方式在学术界和工业界也没有完全统一的定义，是一个十分开放且需要长期探索的现实问题，一个好的评估方式能够为训练迭代优化的过程提供良好的指导，从而减少大量返工成本。

参考文献

[1] Sukhbaatar S, Szlam A, Weston J, et al. End-to-end memory networks [J]. arXiv preprint arXiv:1503.08895, 2015.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[3] 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 [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 5016-5026.

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

[5] Liu Y, Jiang X, Yin Y, et al. One cannot stand for everyone! leveraging multiple user simulators to train task-oriented dialogue systems [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 1-21.

[6] Rafailov R, Sharma A, Mitchell E, et al. Direct preference optimization: Your language model is secretly a reward model [J]. Advances in Neural Information Processing Systems, 2024, 36.

[7] Wu Y, Wu W, Xing C, et al. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 496-505.

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

[9] Bao S, He H, Wang F, et al. Plato: Pre-trained dialogue generation model with discrete latent variable [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 85-96.

[10] Roller S, Dinan E, Goyal N, et al. Recipes for building an open-domain chatbot [C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 300-325.

[11] Adiwardana D, Luong M T, So D R, et al. Towards a human-like open-domain chatbot [J]. arXiv preprint arXiv:2001.09977, 2020.

[12] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models [J]. Transactions on Machine Learning Research, 2022.

[13] Cheng Q, Li L, Quan G, et al. Is multiwoz a solved task? an interactive tod evaluation framework with user simulator [C]//Findings of the Association for Computational Linguistics: EMNLP 2022. 2022: 1248-1259.

[14] Li Z, Zhang J, Fei Z, et al. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 128-138.

[15] Li Z, Peng B, He P, et al. Guiding large language models via directional stimulus prompting [J]. arXiv preprint arXiv:2302.11520, 2023.

[16] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms [J]. arXiv preprint arXiv:1707.06347, 2017.

[17] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models [J]. Advances in neural information processing systems, 2022, 35: 24824-24837.

[18] Levin E, Pieraccini R, Eckert W. Using markov decision process for learning dialogue strategies [C]//Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181): volume 1. IEEE, 1998: 201-204.

[19] Roy N, Pineau J, Thrun S. Spoken dialogue management using probabilistic reasoning [C]// Proceedings of the 38th annual meeting of the association for computational linguistics. 2000: 93-100.

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

[21] Xu P, Sarikaya R. Convolutional neural network based triangular crf for joint intent detection and slot filling [C]//2013 ieee workshop on automatic speech recognition and understanding. IEEE, 2013: 78-83.

[22] Hakkani-Tür D, Tür G, Celikyilmaz A, et al. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. [C]//Interspeech. 2016: 715-719.

[23] Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling [J]. arXiv preprint arXiv:1609.01454, 2016.

[24] Goo C W, Gao G, Hsu Y K, et al. Slot-gated modeling for joint slot filling and intent prediction [C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 753-757.

[25] Zhang C, Li Y, Du N, et al. Joint slot filling and intent detection via capsule neural networks [J]. arXiv preprint arXiv:1812.09471, 2018.

[26] Zhong V, Xiong C, Socher R. Global-locally self-attentive dialogue state tracker [J]. arXiv preprint arXiv:1805.09655, 2018.

[27] Lee H, Lee J, Kim T Y. Sumbt: Slot-utterance matching for universal and scalable belief tracking [J]. arXiv preprint arXiv:1907.07421, 2019.

[28] Shan Y, Li Z, Zhang J, et al. A contextual hierarchical attention network with adaptive objective for dialogue state tracking [J]. arXiv preprint arXiv:2006.01554, 2020.

[29] Chen L, Lv B, Wang C, et al. Schema-guided multi-domain dialogue state tracking with graph attention neural networks [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. 2020: 7521-7528.

[30] Wu C S, Madotto A, Hosseini-Asl E, et al. Transferable multi-domain state generator for task-oriented dialogue systems [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 808-819.

[31] Gao S, Sethi A, Agarwal S, et al. Dialog state tracking: A neural reading comprehension approach [C]//SIGdial. 2019.

[32] Zhou L, Small K. Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering [J]. ArXiv, 2019, abs/1911.06192.

[33] Kim S, Yang S, Kim G, et al. Efficient dialogue state tracking by selectively overwriting memory [J]. ArXiv, 2020, abs/1911.03906.

[34] Heck M, van Niekerk C, Lubis N, et al. Trippy: A triple copy strategy for value independent neural dialog state tracking [J]. arXiv preprint arXiv:2005.02877, 2020.

[35] Guo J, Shuang K, Li J, et al. Beyond the granularity: Multi-perspective dialogue collaborative selection for dialogue state tracking [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 2320-2332.

[36] Wang H, Xin W. How to stop an avalanche? jodem: Joint decision making through compare and contrast for dialog state tracking [C]//Findings of the Association for Computational Linguistics: EMNLP 2022. 2022: 7030-7041.

[37] Williams J, Asadi K, Zweig G. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning [C]//ACL. 2017.

[38] hao Su P, Budzianowski P, Ultes S, et al. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management [C]//SIGDIAL Conference. 2017.

[39] Chen L, Tan B, Long S, et al. Structured dialogue policy with graph neural networks [C]// COLING. 2018.

[40] Shu L, Xu H, Liu B, et al. Modeling multi-action policy for task-oriented dialogues [J]. ArXiv, 2019, abs/1908.11546.

[41] Takanobu R, Zhu H, Huang M. Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog [C]//EMNLP/IJCNLP. 2019.

[42] Zhao T, Xie K, Eskénazi M. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models [J]. ArXiv, 2019, abs/1902.08858.

[43] Takanobu R, Liang R, Huang M. Multi-agent task-oriented dialog policy learning with role-aware reward decomposition [J]. ArXiv, 2020, abs/2004.03809.

[44] 赵阳洋, 王振宇, 王佩, 等. 任务型对话系统研究综述 [J]. 计算机学报, 2020, 43(10): 1862-1896.

[45] Wen T H, Gasic M, Mrksic N, et al. Semantically conditioned lstm-based natural language generation for spoken dialogue systems [C]//EMNLP. 2015.

[46] Wen T H, Gašić M, Mrksic N, et al. Multi-domain neural network language generation for spoken dialogue systems [J]. ArXiv, 2016, abs/1603.01232.

[47] Su S Y, Lo K L, Yeh Y T, et al. Natural language generation by hierarchical decoding with linguistic patterns [J]. ArXiv, 2018, abs/1808.02747.

[48] Takanobu R, Zhu Q, Li J, et al. Is your goal-oriented dialog model performing really well? empirical analysis of system-wise evaluation [C]//21th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2020: 297.

[49] Lei W, Jin X, Kan M Y, et al. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures [C]//ACL. 2018.

[50] Zhang Y, Ou Z, Yu Z. Task-oriented dialog systems that consider multiple appropriate responses under the same context [J]. ArXiv, 2020, abs/1911.10484.

[51] Le H, Sahoo D, Liu C, et al. Uniconv: A unified conversational neural architecture for multi-domain task-oriented dialogues [J]. ArXiv, 2020, abs/2004.14307.

[52] Weston J, Chopra S, Bordes A. Memory networks [J]. arXiv preprint arXiv:1410.3916, 2014.

[53] Eric M, Manning C D. Key-value retrieval networks for task-oriented dialogue [J]. arXiv preprint arXiv:1705.05414, 2017.

[54] Madotto A, Wu C S, Fung P. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1468-1478.

[55] Wu C s, Socher R, Xiong C. Global-to-local memory pointer networks for task-oriented dialogue [C]//7th International Conference on Learning Representations, ICLR 2019. 2019.
[56] Qin L, Xu X, Che W, et al. Dynamic fusion network for multi-domain end-to-end task-oriented dialog [J]. arXiv preprint arXiv:2004.11019, 2020.
[57] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.
[58] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551.
[59] Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension [J]. arXiv preprint arXiv:1910.13461, 2019.
[60] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners [J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
[61] Budzianowski P, Vulic I. Hello, it's gpt-2-how can i help you? towards the use of pretrained language models for task-oriented dialogue systems [J]. EMNLP-IJCNLP 2019, 2019: 15.
[62] Hosseini-Asl E, McCann B, Wu C S, et al. A simple language model for task-oriented dialogue [J]. arXiv preprint arXiv:2005.00796, 2020.
[63] Peng B, Li C, Li J, et al. Soloist: Few-shot task-oriented dialog with a single pretrained auto-regressive model [J]. arXiv preprint arXiv:2005.05298, 2020.
[64] Yang Y, Li Y, Quan X. Ubar: Towards fully end-to-end task-oriented dialog systems with gpt-2 [J]. arXiv preprint arXiv:2012.03539, 2020.
[65] Kulhánek J, Hudeček V, Nekvinda T, et al. Augpt: Dialogue with pre-trained language models and data augmentation [J]. arXiv preprint arXiv:2102.05126, 2021.
[66] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback [J]. Advances in neural information processing systems, 2022, 35: 27730-27744.
[67] Taori R, Gulrajani I, Zhang T, et al. Alpaca: A strong, replicable instruction-following model [J]. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html, 2023, 3(6): 7.
[68] Gunasekar S, Zhang Y, Aneja J, et al. Textbooks are all you need [J]. arXiv preprint arXiv:2306.11644, 2023.
[69] Zhou C, Liu P, Xu P, et al. Lima: Less is more for alignment [J]. Advances in Neural Information Processing Systems, 2024, 36.
[70] Xie S M, Pham H, Dong X, et al. Doremi: Optimizing data mixtures speeds up language model pretraining [J]. Advances in Neural Information Processing Systems, 2024, 36.
[71] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models [J]. arXiv preprint arXiv:2307.09288, 2023.
[72] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models [J]. arXiv preprint arXiv:2302.13971, 2023.
[73] Yang A, Xiao B, Wang B, et al. Baichuan 2: Open large-scale language models [J]. arXiv preprint arXiv:2309.10305, 2023.
[74] Le Scao T, Fan A, Akiki C, et al. Bloom: A 176b-parameter open-access multilingual language model [J]. 2023.
[75] Almazrouei E, Alobeidli H, Alshamsi A, et al. The falcon series of open language models [J]. arXiv preprint arXiv:2311.16867, 2023.

[76] Bai J, Bai S, Chu Y, et al. Qwen technical report [J]. arXiv preprint arXiv:2309.16609, 2023.

[77] Jiang A Q, Sablayrolles A, Roux A, et al. Mixtral of experts [J]. arXiv preprint arXiv:2401.04088, 2024.

[78] Hu E J, Wallis P, Allen-Zhu Z, et al. Lora: Low-rank adaptation of large language models [C]//International Conference on Learning Representations. 2021.

[79] He J, Zhou C, Ma X, et al. Towards a unified view of parameter-efficient transfer learning [C]//International Conference on Learning Representations. 2021.

[80] Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation [C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 4582-4597.

[81] LI S, Yavuz S, Hashimoto K, et al. Coco: Controllable counterfactuals for evaluating dialogue state trackers [C]//International Conference on Learning Representations. 2020.

[82] Gao S, Zhang Y, Ou Z, et al. Paraphrase augmented task-oriented dialog generation [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 639-649.

[83] Zhang Y, Ou Z, Yu Z. Task-oriented dialog systems that consider multiple appropriate responses under the same context [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. 2020: 9604-9611.

[84] Kim S, Chang M, Lee S W. Neuralwoz: Learning to collect task-oriented dialogue via model-based simulation [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 3704-3717.

[85] Wan D, Zhang Z, Zhu Q, et al. A unified dialogue user simulator for few-shot data augmentation [C]//Findings of the Association for Computational Linguistics: EMNLP 2022. 2022: 3788-3799.

[86] He W, Dai Y, Zheng Y, et al. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection [C]//Proceedings of the AAAI conference on artificial intelligence: volume 36. 2022: 10749-10757.

[87] Wu L, Li J, Wang Y, et al. R-drop: Regularized dropout for neural networks [J]. Advances in Neural Information Processing Systems, 2021, 34: 10890-10905.

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

[89] Sun H, Bao J, Wu Y, et al. Mars: Modeling context & state representations with contrastive learning for end-to-end task-oriented dialog [C]//Findings of the Association for Computational Linguistics: ACL 2023. 2023: 11139-11160.

[90] Su Y, Shu L, Mansimov E, et al. Multi-task pre-training for plug-and-play task-oriented dialogue system [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 4661-4676.

[91] Lee Y. Improving end-to-end task-oriented dialog system with a simple auxiliary task [C]// Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 1296-1303.

[92] Cholakov R, Kolev T. Efficient task-oriented dialogue systems with response selection as an auxiliary task [C]//Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022). 2022: 12-18.

[93] Yu X, Wu Q, Qian K, et al. Krls: Improving end-to-end response generation in task oriented dialog with reinforced keywords learning [Z]. 2023.

[94] Bang N, Lee J, Koo M W. Task-optimized adapters for an end-to-end task-oriented dialogue system [J]. arXiv preprint arXiv:2305.02468, 2023.

[95] Feng Y, Yang S, Zhang S, et al. Fantastic rewards and how to tame them: A case study on reward learning for task-oriented dialogue systems [C]//The Eleventh International Conference on Learning Representations. 2022.

[96] Raghu D, Jain A, Joshi S, et al. Constraint based knowledge base distillation in end-to-end task oriented dialogs [C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 5051-5061.

[97] Madotto A, Cahyawijaya S, Winata G I, et al. Learning knowledge bases with parameters for task-oriented dialogue systems [C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 2372-2394.

[98] Huang G, Quan X, Wang Q. Autoregressive entity generation for end-to-end task-oriented dialog [C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 323-332.

[99] Xie T, Wu C H, Shi P, et al. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models [C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 602-631.

[100] Wu J, Harris I G, Zhao H. Graphmemdialog: Optimizing end-to-end task-oriented dialog systems using graph memory networks [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 36. 2022: 11504-11512.

[101] Tian X, Lin Y, Song M, et al. Q-tod: A query-driven task-oriented dialogue system [C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 7260-7271.

[102] Wan F, Shen W, Yang K, et al. Multi-grained knowledge retrieval for end-to-end task-oriented dialog [J]. arXiv preprint arXiv:2305.10149, 2023.

[103] Shi T, Li L, Lin Z, et al. Dual-feedback knowledge retrieval for task-oriented dialogue systems [J]. arXiv preprint arXiv:2310.14528, 2023.

[104] Shen W, Gao Y, Huang C, et al. Retrieval-generation alignment for end-to-end task-oriented dialogue system [J]. arXiv preprint arXiv:2310.08877, 2023.

[105] Mao K, Dou Z, Chen H, et al. Large language models know your contextual search intent: A prompting framework for conversational search [J]. arXiv preprint arXiv:2303.06573, 2023.

[106] Jiang J, Zhou K, Dong Z, et al. Structgpt: A general framework for large language model to reason over structured data [J]. arXiv preprint arXiv:2305.09645, 2023.

[107] Kim T E, Lipani A. A multi-task based neural model to simulate users in goal oriented dialogue systems [C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022: 2115-2119.

[108] Lin H C, Lubis N, Hu S, et al. Domain-independent user simulation with transformers for task-oriented dialogue systems [C]//Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2021: 445-456.

[109] Lin H C, Geishauser C, Feng S, et al. Gentus: Simulating user behaviour and language in task-oriented dialogues with generative transformers [C]//Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2022: 270-282.

[110] Lin H C, Feng S, Geishauser C, et al. Emous: Simulating user emotions in task-oriented dialogues [C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 2526-2531.

[111] Terragni S, Filipavicius M, Khau N, et al. In-context learning user simulators for task-oriented dialog systems [J]. arXiv preprint arXiv:2306.00774, 2023.

[112] Hu Z, Feng Y, Luu A T, et al. Unlocking the potential of user feedback: Leveraging large language model as user simulator to enhance dialogue system [J]. arXiv preprint arXiv:2306.09821, 2023.

[113] Zhang X, Peng B, Gao J, et al. Toward self-learning end-to-end task-oriented dialog systems [C]//Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2022: 516-530.

[114] Hu Z, Feng Y, Deng Y, et al. Enhancing large language model induced task-oriented dialogue systems through look-forward motivated goals [J]. arXiv preprint arXiv:2309.08949, 2023.

[115] Bengio S, Vinyals O, Jaitly N, et al. Scheduled sampling for sequence prediction with recurrent neural networks [J]. Advances in neural information processing systems, 2015, 28.

[116] Wen T H, Vandyke D, Mrkšić N, et al. A network-based end-to-end trainable task-oriented dialogue system [C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. 2017: 438-449.

[117] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences [C]//52nd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2014.

[118] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780.

[119] Lei W, Jin X, Kan M Y, et al. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1437-1447.

[120] Zhang Y, Ou Z, Hu M, et al. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 9207-9219.

[121] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training [J].

[122] Dong L, Yang N, Wang W, et al. Unified language model pre-training for natural language understanding and generation [J]. Advances in neural information processing systems, 2019, 32.

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

[124] Yang Y, Li Y, Quan X. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2 [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 35. 2021: 14230-14238.

[125] Sun H, Bao J, Wu Y, et al. Bort: Back and denoising reconstruction for end-to-end task-oriented dialog [C]//Findings of the Association for Computational Linguistics: NAACL 2022. 2022: 2156-2170.

[126] Zhang W, Feng Y, Meng F, et al. Bridging the gap between training and inference for neu-

ral machine translation [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4334-4343.

[127] Kim S, Yang S, Kim G, et al. Efficient dialogue state tracking by selectively overwriting memory [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 567-582.

[128] Zhang K, Shasha D. Simple fast algorithms for the editing distance between trees and related problems [J]. SIAM journal on computing, 1989, 18(6): 1245-1262.

[129] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation [C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.

[130] Wolf T, Debut L, Sanh V, et al. Transformers: State-of-the-art natural language processing [C]//Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. 2020: 38-45.

[131] Loshchilov I, Hutter F. Decoupled weight decay regularization [J]. arXiv preprint arXiv:1711.05101, 2017.

[132] Jeon H, Lee G G. Domain state tracking for a simplified dialogue system [J]. arXiv preprint arXiv:2103.06648, 2021.

[133] Lin Z, Madotto A, Winata G I, et al. Mintl: Minimalist transfer learning for task-oriented dialogue systems [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 3391-3405.

[134] Eric M, Goel R, Paul S, et al. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines [C]//Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020: 422-428.

[135] Rony M R A H, Usbeck R, Lehmann J. Dialokg: Knowledge-structure aware task-oriented dialogue generation [C]//Findings of the Association for Computational Linguistics: NAACL 2022. 2022: 2557-2571.

[136] Zhao P, Zhang H, Yu Q, et al. Retrieval-augmented generation for ai-generated content: A survey [J]. arXiv preprint arXiv:2402.19473, 2024.

[137] Guu K, Lee K, Tung Z, et al. Realm: retrieval-augmented language model pre-training [C]// Proceedings of the 37th International Conference on Machine Learning. 2020: 3929-3938.

[138] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks [J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[139] Petroni F, Piktus A, Fan A, et al. Kilt: a benchmark for knowledge intensive language tasks [C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 2523-2544.

[140] Shi W, Min S, Yasunaga M, et al. Replug: Retrieval-augmented black-box language models [J]. arXiv preprint arXiv:2301.12652, 2023.

[141] Ram O, Levine Y, Dalmedigos I, et al. In-context retrieval-augmented language models [J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1316-1331.

[142] Ma X, Gong Y, He P, et al. Query rewriting in retrieval-augmented large language models [C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 5303-5315.

[143] Izacard G, Grave É. Leveraging passage retrieval with generative models for open domain question answering [C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 874-880.

[144] Borgeaud S, Mensch A, Hoffmann J, et al. Improving language models by retrieving from trillions of tokens [C]//International conference on machine learning. PMLR, 2022: 2206-2240.

[145] Khandelwal U, Levy O, Jurafsky D, et al. Generalization through memorization: Nearest neighbor language models [J]. arXiv preprint arXiv:1911.00172, 2019.

[146] He J, Neubig G, Berg-Kirkpatrick T. Efficient nearest neighbor language models [C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 5703-5714.

[147] Leviathan Y, Kalman M, Matias Y. Fast inference from transformers via speculative decoding [C]//International Conference on Machine Learning. PMLR, 2023: 19274-19286.

[148] He Z, Zhong Z, Cai T, et al. Rest: Retrieval-based speculative decoding [J]. arXiv preprint arXiv:2311.08252, 2023.

[149] Bang F. Gptcache: An open-source semantic cache for llm applications enabling faster answers and cost savings [C]//Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023). 2023: 212-218.

[150] Shin J, Yu H, Moon H, et al. Dialogue summaries as dialogue states (ds2), template-guided summarization for few-shot dialogue state tracking [C]//Findings of the Association for Computational Linguistics: ACL 2022. 2022: 3824-3846.

[151] Hu Y, Lee C H, Xie T, et al. In-context learning for few-shot dialogue state tracking [C]// Findings of the Association for Computational Linguistics: EMNLP 2022. 2022: 2627-2643.

[152] Izacard G, Caron M, Hosseini L, et al. Unsupervised dense information retrieval with contrastive learning [J]. arXiv preprint arXiv:2112.09118, 2021.

[153] Kim B, Ahn J, Kim G. Sequential latent knowledge selection for knowledge-grounded dialogue [C]//International Conference on Learning Representations. 2019.

[154] Zhan H, Shen L, Chen H, et al. Colv: A collaborative latent variable model for knowledge-grounded dialogue generation [C]//Proceedings of the 2021 conference on empirical methods in natural language processing. 2021: 2250-2261.

[155] Loshchilov I, Hutter F. Decoupled weight decay regularization [C]//International Conference on Learning Representations. 2018.

[156] Qin L, Liu Y, Che W, et al. Entity-consistent end-to-end task-oriented dialogue system with kb retriever [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 133-142.

[157] He Z, Wang J, Chen J. Task-oriented dialog generation with enhanced entity representation. [C]//2020.

[158] He Z, He Y, Wu Q, et al. Fg2seq: Effectively encoding knowledge for end-to-end task-oriented dialog [C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 8029-8033.

[159] Izacard G, Lewis P, Lomeli M, et al. Atlas: Few-shot learning with retrieval augmented language models [J]. Journal of Machine Learning Research, 2023, 24(251): 1-43.

[160] Lin C Y. Rouge: A package for automatic evaluation of summaries [C]//Text summarization branches out. 2004: 74-81.

[161] Sun W, Guo S, Zhang S, et al. Metaphorical user simulators for evaluating task-oriented dialogue systems [J]. ACM Transactions on Information Systems, 2023, 42(1): 1-29.

[162] Zhou D, Schärli N, Hou L, et al. Least-to-most prompting enables complex reasoning in large language models [C]//The Eleventh International Conference on Learning Representations. 2022.

[163] Quan J, Zhang S, Cao Q, et al. Risawoz: A large-scale multi-domain wizard-of-oz dataset with rich semantic annotations for task-oriented dialogue modeling [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 930-940.

[164] Dinan E, Roller S, Shuster K, et al. Wizard of wikipedia: Knowledge-powered conversational agents [C]//International Conference on Learning Representations. 2018.

[165] Komeili M, Shuster K, Weston J. Internet-augmented dialogue generation [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 8460-8478.

[166] User modeling for spoken dialogue system evaluation [C]//1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings. IEEE, 1997: 80-87.

[167] Human-computer dialogue simulation using hidden markov models [C]//IEEE Workshop on Automatic Speech Recognition and Understanding, 2005. IEEE, 2005: 290-295.

[168] Schatzmann J, Thomson B, Weilhammer K, et al. Agenda-based user simulation for bootstrapping a pomdp dialogue system [C]//Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers. 2007: 149-152.

[169] Schatzmann J, Young S. The hidden agenda user simulation model [J]. IEEE transactions on audio, speech, and language processing, 2009, 17(4): 733-747.

[170] Zhang S, Balog K. Evaluating conversational recommender systems via user simulation [C]// Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining. 2020: 1512-1520.

[171] Asri L E, He J, Suleman K. A sequence-to-sequence model for user simulation in spoken dialogue systems [J]. arXiv preprint arXiv:1607.00070, 2016.

[172] Gür I, Hakkani-Tür D, Tür G, et al. User modeling for task oriented dialogues [C]//2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018: 900-906.

[173] Shi W, Qian K, Wang X, et al. How to build user simulators to train rl-based dialog systems [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 1990-2000.

[174] Tseng B H, Dai Y, Kreyssig F, et al. Transferable dialogue systems and user simulators [C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 152-166.

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

[176] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways [J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.

[177] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report [J]. arXiv preprint arXiv:2303.08774, 2023.

[178] Wang X, Duan S, Yi X, et al. On the essence and prospect: An investigation of alignment approaches for big models [J]. arXiv preprint arXiv:2403.04204, 2024.

[179] Stiennon N, Ouyang L, Wu J, et al. Learning to summarize with human feedback [J]. Advances in Neural Information Processing Systems, 2020, 33: 3008-3021.

[180] Nakano R, Hilton J, Balaji S, et al. Webgpt: Browser-assisted question-answering with human feedback [J]. arXiv preprint arXiv:2112.09332, 2021.

[181] Bai Y, Kadavath S, Kundu S, et al. Constitutional ai: Harmlessness from ai feedback [J]. arXiv preprint arXiv:2212.08073, 2022.

[182] Wang Y, Kordi Y, Mishra S, et al. Self-instruct: Aligning language models with self-generated instructions [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 13484-13508.

[183] Sun Z, Shen Y, Zhou Q, et al. Principle-driven self-alignment of language models from scratch with minimal human supervision [J]. Advances in Neural Information Processing Systems, 2024, 36.

[184] Azar M G, Rowland M, Piot B, et al. A general theoretical paradigm to understand learning from human preferences [J]. arXiv preprint arXiv:2310.12036, 2023.

[185] Ethayarajh K, Xu W, Muennighoff N, et al. Kto: Model alignment as prospect theoretic optimization [J]. arXiv preprint arXiv:2402.01306, 2024.

[186] Ganguli D, Askell A, Schiefer N, et al. The capacity for moral self-correction in large language models [J]. arXiv preprint arXiv:2302.07459, 2023.

[187] Han X. In-context alignment: Chat with vanilla language models before fine-tuning [J]. arXiv preprint arXiv:2308.04275, 2023.

[188] Lin B Y, Ravichander A, Lu X, et al. The unlocking spell on base llms: Rethinking alignment via in-context learning [J]. arXiv preprint arXiv:2312.01552, 2023.

致 谢

匆匆三年，硕士求学生涯即将画上句点。站在毕业的门槛前，回首这段时光，心潮澎湃，感慨万千。在这段漫长而又短暂的旅程中，得到了无数的帮助和支持，不论是家人、老师还是同窗好友，他们的关怀与鼓励，都是我前行路上最温暖的庇护。此刻，愿我能停下脚步，向曾与我同行的每一个人，表达我最诚挚的谢意。在即将踏上新征程之际，我愿以这些文字，为你们的支持与陪伴，献上最深沉的感恩。

首先，我要由衷地感谢我的导师冯洋老师。感谢冯洋老师给予我这个宝贵的机会，录取我进入研究组，开启了我这段丰富而充实的求学经历。冯老师不仅在学术上给予我认真的指导，还在日常生活中呵护备至，让我感受到无微不至的关怀和温暖。她的引导和关心使我能够迅速适应并投入到课题组的科研工作中。冯老师在实验室管理方面严格要求，制定了严谨的规章制度，营造了浓厚的学术氛围和优越舒适的科研环境。她以自己深厚的学识和勤奋认真的科研态度激励着我，时刻提醒我要持之以恒、不断进取。无论是科研上的挫折还是生活中的困难，冯老师总是倾听并给予充分的理解和帮助，让我感到莫大的鼓舞和支持。在科研方面，冯老师总是教导我们要从思考问题和动机的角度出发，在深入细节前先进行全局把握，建立好正确的科研思维方式，这对我科研方式的转变起到十分积极的作用。在我读研期间，冯老师不仅是我的导师，更是我的良师益友。她细致地指导我，鼓励我困难勉励，使我如沐春风，不断进步。再次衷心感谢冯洋老师在我求学路上的悉心栽培和无私奉献！

其次，我要感谢大四时引领我入门科研的赵海老师和张倬胜老师，在我第一次投稿的时候，赵海老师在论文写作方面对我进行了十分细节的指导，使我的论文得到了充分的打磨。此外，赵海老师不仅学术水平很高，还博览群书，在和赵海老师谈笑风生的过程中，我拓宽了视野。张倬胜老师则手把手地带领我入门了一次完整的科研，从问题调研，思路建立，工程实现到论文撰写和投稿流程，都离不开张老师的帮助，很感激有这么一段和这么优秀的两位老师合作的经历！

感谢实验室秘书程一老师和裴晓雪老师在三年来对我的科研和生活的帮助。在这三年中，两位老师帮助我解决了各种零零碎碎大大小小的事项，解答了我一个个细节的疑问，小到日常报销，大到毕业流程，两位老师认真负责的工作态度极大地缓解了我在科研之外的压力，让我能够更专注于科研本身。尤其是毕业前的这段时间，裴老师帮助我安排好各项事务以及进行在各环节进行充分的提醒，缓解了我在繁杂的毕业流程中焦虑的心情。

感谢课题组的各位同学，很感激能有机会和大家一起度过精彩充实的这段时光，大家优秀的品质与突出的能力时刻提醒着我要朝着更高的目标奔跑，在科研和项目方面，大家也给我提供了很多的帮助，让我能够顺利跨过一个一个原本以为将要跨不过去的坎。感谢李秀星师兄对我的帮助和关心，秀星师兄一直给予

我非常正面的鼓励，让我即使在失意的时候也能够认同自身的价值，在我的论文投稿前很细心地帮我审阅论文，也很关心我们大家的生活，也会在未来规划方面提供很多有用的建议。在参与项目的过程中，师兄也承担了很多前方后方的工作，极大减少了我们的压力。感谢丁春发、杨郑鑫、谷舒豪、邵晨泽、单勇、李绩成、李泽康、郭登级、张绍磊、张倬诚、伍烜甫、田畅师兄和刘舒曼、申磊、欧蛟、谢婉莹师姐，他们在我的科研方面提供了很多帮助，在我的找工作期间提供了非常多有价值的信息与建议，他们的认真刻苦和对自己的严格要求也为我树立了榜样，激励着我前行。感谢同级的房庆凯、马铮睿、黄浪林、桂尚彤、赵彤钰、郭雯钰、卫李赋凌同学，依稀记得第一次到北京的那个雨夜，在计算所楼下第一次见到了庆凯和铮睿，在他们的帮助下，我很快就克服了孤独感，适应了这座新的城市和新的环境。感谢研一疫情封控期间大家的相互陪伴，每一个一起度过的生日聚会、每一首一起在操场上唱过的歌、每一节坐在一起上过的课，都成为值得回味的经历，成为一生中重要的宝藏。也真心地祝愿一起毕业的同学们，下一站能够迈向更远的远方。感谢李熠阳同学，身为本科同班同学而如今又是科研的同行，与你的交流让我产生了很多有意思的科研想法，也让我在孤独无助的时候，能够感受到力量，得空回上海的一次交流使我受益匪浅，祝你在新的起点也能够实现更高的目标。感谢我最亲密的高中同学们，一年一到两次的集体出游让我一直感受到，传递自高中以来的，属于我们大家的快乐，即便日后大家走上了不同的职业发展道路，不在一个城市，我们也依旧是能够一起跨年的最好的小伙伴们。感谢研究组的师弟师妹们，他们是杨哲、郭守涛、张珂豪、鄢子文、卜梦煜、周奚、雨田、温卓凡和乔康裕师弟，很感谢能和大家一起讨论科研，一起参加组里面的春游秋游和年会等活动，你们也是课题组未来发展的希望，祝你们未来能够披荆斩棘，创造更多的辉煌。感谢物理所软物质实验室 SM04 组的小伙伴们，大家的聪明才智给我留下了深刻的印象，和大家一起玩很开心，也拓宽了我在本学科以外的视野。

感谢一直支持我的家人们，无论经历过多少的困难，距离有多远，你们永远给予我无限的包容和理解，在物质和心灵层面都对我关怀备至，你们是我的坚强后盾。你们的支持让我能投入足够多的精力在科研上，无需担心其他的事物。

感谢一直陪伴着我的孙运海同学，感谢你在我每次生病的时候陪着我一起去看病，宽慰我的不安；感谢你一直以来与我分享各自的喜悦，一直聆听我难过时的倾诉，引导我逐步构建起更为坚强的精神世界，学会表达自我，适当的时候与自我和解；感谢大大小小的生活日常，一起运动一起旅行，一起做那些最重要的小事。人的成长就如一条行走在陆地的鱼，要忍着痛往前行，别人也许只会爱我的光，而你能够在这不安的世纪中，为我保留哭的权利。

在本文的最后，感谢在百忙之中评阅论文并提出宝贵意见的老师，以及在完成论文的过程中给予我关心和帮助的所有人。衷心祝愿大家后面的日子能够平安喜乐，健康幸福。

2024 年 6 月

作者简历及攻读学位期间发表的学术论文与其他相关学术成果

作者简历:

2017 年 09 月——2021 年 06 月，在上海交通大学电子信息与电气工程学院计算机科学与工程系获得学士学位。

2021 年 09 月——2024 年 06 月，在中国科学院计算技术研究所攻读硕士学位。

已发表（或正式接受）的学术论文:

(1) **Longxiang Liu**, Xiuxing Li, Yang Feng. TA&AT: Enhancing Task-Oriented Dialog with Turn-Level Auxiliary Tasks and Action-Tree Based Scheduled Sampling. AAAI 2024.

(2) **Longxiang Liu**, Zhuosheng Zhang, Hai Zhao, Xi Zhou, Xiang Zhou. Filling the Gap of Utterance-aware and Speaker-aware Representation for Multi-turn Dialogue. AAAI 2021.

申请或已获得的专利:

(1) 冯洋，刘龙祥，李秀星. 基于大模型的多模态实体链接方法. 申请号: 202311439852.6

参加的研究项目及获奖情况:

(1) 2022 年 10 月-2024 年 3 月: 科技创新 2030-新一代人工智能重大项目: 人机协同智能系统软硬件技术研究，人机行为与情景常识的大规模知识处理与推理，课题编号：2018AAA0102502，项目成员

(2) 2023-2024 中国科学院大学三好学生