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

博士学位论文

语义增强的开放域对话生成技术研究

作者姓名：欧蛟

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：工学博士

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

2023 年 6 月

Open-Domain Dialogue Generation Based on Semantic Enhancement

A dissertation submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Computer Application Technology

by

Ou Jiao

Supervisor: FENG Yang

Institute of Computing Technology, Chinese Academy of Sciences

June, 2023

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

开放域对话系统旨在满足用户的社交需求，比如情感支持、陪伴交流、归属感获取等，并最大限度地提升用户的长期参与度。目前随着深度神经网络技术的发展，数据驱动的生成式开放域对话系统逐渐受到了广泛的关注。为了构建高质量的生成式开放域对话系统，需要考虑其独特的特点，即允许对话有各种各样不同语义，但又和对话历史语义保持一致的回复。因此，对话系统首先需要具备理解复杂语义映射关系的基本能力，才能生成语义多样的回复。其中数据驱动的构建方式要求需要语义内容丰富的训练数据。此外，生成的回复还要满足和对话历史中的语义内容不能自相矛盾（即保持一致）的约束。然而，现有系统对语义信息的挖掘与利用还非常有限，使得达成上述目标仍然面临着若干挑战。

本文研究了构建系统中所面临的多个关键性挑战：（1）对话语义映射关系建模难的问题：开放域对话具有复杂的层级语义映射关系。模型自行从数据中学习这种复杂映射关系尤为困难，如何显式建模层级语义映射关系是值得探索的问题；（2）训练数据语义内容不够丰富的问题：人工构建语义内容丰富的训练数据耗时耗力。现有数据通常包含少量语义角度的回复，如何自动生成多个不同语义的回复去增强对话语义内容是非常重要的问题；（3）回复语义内容一致性约束难以满足的问题：现有系统额外训练一致性检测模型来满足回复的一致性约束。人工标注的训练数据和使用场景的数据存在分布差异，导致模型在使用时性能下降。如何缓解检测模型在数据分布差异下的性能下降使得一致性约束被更好地满足是亟待解决的问题。

针对上述三个问题，本文随着问题难度的增加逐步加深对语义信息的挖掘与利用，从基础的语义建模，到显式的语义转换，再到复杂的语义分解，分别提出相应的解决方案，并取得了如下成果：

1、基于语义建模的对话回复生成方法

针对语义映射关系建模难的问题，本文提出了一种基于语义建模的回复生成方法，通过显式建模层级语义映射关系去生成语义多样的回复。该方法沿袭了利用隐变量增强建模能力的思路，进一步将隐变量与可生成回复的语义侧面一一对应，并减少隐变量的语义表示间的重叠。最终该方法通过采样不同的隐变量值即可得到不同语义表达的回复。其中为了得到可生成回复的语义侧面的信息，该方法额外引入了多语义回复检索模块，它为训练集中每个对话上文额外扩增多个语义不同的回复。不同语义的回复对应了不同的语义侧面。模型通过监督训练将回复中蕴含的语义侧面信息与隐变量对应起来。实验结果表明该方法能够生成更多语义多样的回复。同时分析实验表明该方法能够有效地建模开放域对话的层级语义映射关系。

2、基于语义转换的对话数据增强方法

针对训练数据语义内容不够丰富的问题，本文提出了一种基于语义转换的

数据增强方法，为训练数据中每个对话上文扩增更多不同语义的回复。该方法借鉴了人类生成不同语义回复的过程。首先人关注到对话上文中的某部分内容，进而转移关注点到一个想要谈论的语义角度去生成一个回复。但随后人会去思考：如果改变当前关注的语义角度，回复会有什么不同。回答这一问题能够得到不同语义的回复。因此，针对给定的对话上文，该方法通过转换已观测回复对应的语义角度，去重新推理一个新的回复。其中为了得到可转换的语义角度，该方法基于训练数据构建了对话上文关注点与语义角度间的转移关系图，利用转移关系图保证转换的有效性。实验结果表明该方法能够有效扩增不同语义的回复，并能提升下游任务的性能，尤其是对话生成的语义多样性。

3、基于语义分解的对话一致性检测方法

针对回复语义内容一致性约束难以满足的问题，本文提出了基于语义分解的一致性检测方法，用于缓解检测模型在数据分布差异下的性能下降，进而使得一致性约束被更有效地满足。数据分布差异下的性能下降表明模型鲁棒性差。该方法借鉴了其他任务提升模型鲁棒性的思想，即构造反事实样本，并与原始样本混合去训练模型，去提升模型的鲁棒性。反事实样本为和原始样本很相似但标签相反的样本。该方法根据定义设计了“标记不一致内容-修改最少的不一致内容使标签反转”的构造方案。其中为了自动标记不一致内容，该方法首先利用语义分解从复杂对话中分解出多个独立的语义单元，再通过判断哪些语义单元相互矛盾去标记不一致内容。实验结果表明该方法能够有效提高检测模型的鲁棒性，从而更好地满足回复语义内容和对话上文中内容保持一致的约束。

综上所述，本文以系统构建中的关键要素“语义”为切入点，通过不断加深对对话语义信息的理解，不断挖掘与利用更加具体以及细粒度的语义信息，为系统构建中的多个关键性挑战设计了相应的解决方案。希望本文的研究工作能够推进这一研究领域的发展。

关键词：开放域对话系统，数据增强，对话生成，对话一致性

Abstract

Open-domain dialogue systems aim to satisfy the social needs of humans, such as emotional support, companionship, acquisition of a sense of belonging, etc., to maximize the long-term connections of users. With the recent development of deep neural networks, data-driven generative open-domain dialogue systems have gradually attracted extensive attention from researchers. In order to construct a high-quality generative open-domain dialogue system, it is necessary to consider its unique characteristic, which allows dialogues to have a variety of different-semantic responses but keep semantically consistent with the dialogue history. Therefore, dialogue systems must first comprehensively understand complex semantic mapping to generate semantically diverse responses. Accordingly, the data-driven construction approach requires dialogue training data with rich semantic content. Meanwhile, generated responses cannot contradict the semantic content of the dialogue history (i.e., keep consistency). However, the mining and utilization of semantics in existing systems are minimal, so there are still several challenges to achieving the above goals.

This thesis studies several key challenges during the construction process: (1) Difficult modeling of dialogue semantic mapping: Open-domain dialogue has complex hierarchical semantic mapping. It is particularly difficult for dialogue models to learn such a complex mapping from training data by itself. How to explicitly model the hierarchical semantic mapping is a problem worth exploring; (2) Insufficient semantic content of training data: Manually constructing training data with rich semantics is time-consuming and labor-intensive. Responses from existing datasets usually correspond to only a small number of semantic perspectives. How to automatically generate a variety of different-semantic responses to enhance the semantic content of the dialogue is a significant problem; (3) Difficulty satisfying consistency constraints on semantic content: Existing systems additionally train consistency detection models to meet the consistency constraints of responses. The training data of detection models are human-written, which has different distribution from the system-generated data. Thus, detection models perform poorly in the real-world setting. How to alleviate the performance decay of detection models under distribution shift for better satisfying the consistency constraints is an urgent problem.

Aiming at the above three problems, this thesis gradually deepens the mining and utilization of semantic information as the difficulty of the problem increases. The thesis then proposes corresponding solutions from basic semantic modeling, to explicit semantic transition, and then to complex semantic decomposition, and achieves the following results:

1. Dialogue Response Generation Based on Semantic Modelling

For difficult modeling of dialogue semantic mapping, this work proposes a response generation method based on semantic representation for explicitly modeling the hierarchical semantic mapping and generating semantically diverse responses. This method follows the idea of introducing latent variables to enhance modeling. This method further aligns latent variables with different semantic aspects and reduces the overlap between semantic representations of latent variables. To obtain all semantic aspects of the given dialogue history, this method additionally introduces a multi-semantic response retrieval module, which supplements multiple different-semantic responses for each dialogue history in the training set. Responses with different semantics correspond to different semantic aspects. Response generation models associate each semantic aspect contained in the response with different latent variables through supervised training. Experimental results show that this method can generate more semantically diverse responses, and can effectively model the hierarchical semantic mapping of open-domain dialogues.

2. Dialogue Data Augmentation Based on Semantic Transition

For insufficient semantic content of training data, this work proposes a data augmentation method based on semantic transition for augmenting more different-semantic responses. This method inspires by the human process of generating different-semantic responses. Humans first focus on a certain point in the dialogue, and then shift their attention to a semantic perspective that they want to talk about to generate a response. Afterward, humans would think about a question: the response would happen if the current semantic perspective is changed. Answering this question will infer a different response. Thus, given an observed dialogue, this method infers semantically different responses by replacing the observed semantic perspective with alternative ones. To achieve an alternative, this method constructs a shift graph based on all observed dialogues, which explicitly represents the shifts between the focuses on dialogue histories and their corresponding semantic perspective respectively. The shift graph ensures the validity of semantic transition. Experimental results show that this method can augment high-quality responses with different semantics and improve performance on downstream tasks, especially the semantic diversity on response generation.

3. Dialogue Consistency Detection Based on Semantic Decomposition

For difficulty satisfying consistency constraints on semantic content, this work proposes a consistency detection method for alleviating performance decay of detection models under distribution shift, so that the consistency constraints can be satisfied more effectively. Performance decay under distribution shift indicates the poor robustness of models. Drawing inspiration from ideas used to improve the robustness of other NLP tasks, this method constructs counterfactual samples and then merges them with

the original training samples to improve their robustness. The counterfactual samples refer to samples that are very similar to the original sample but have opposite labels. According to the definition, the method designs the construction scheme of ”labeling inconsistent content - modifying the least inconsistent content to reverse the label”. To automatically identify the inconsistent content, this method first applies semantic decomposition to decompose independent semantic units from dialogues and then marks inconsistent content by judging which semantic units are contradictory. Experimental results show that this method can effectively improve the robustness of dialogue consistency detection models, to better meet the constraint that the semantic content of the response is consistent with the content in the dialogue history.

In summary, this thesis takes ”semantics”, the key element in system construction, as the starting point, through continuously deepening the understanding of dialogue semantics, and continuously mining and utilizing more specific and fine-grained semantic information, to address multiple key challenges in system construction. Accordingly, this thesis designs corresponding solutions for more effectively building generative dialogue models. It is hoped that the techniques we explore in this thesis can promote the development of this research branch.

Key Words: Open-domain Dialogue System, Data Augmentation, Dialogue Generation, Dialogue Consistency

目 录

图目录

表目录

第 1 章　绪论

1.1　研究背景及意义

语言是人类相互交流和理解的桥梁，也是人与机器沟通的自然纽带。自 1950 年提出图灵测试以后[1]，基于自然语言的人机交互逐渐受到广泛的关注。简单来说，基于自然语言的人机交互就是人们期待有一天机器能够像人一样流畅、自然、富有感情、饱含人设的去和人对话，像人一样思考，并且人类还无法辨认对方是真实的人还是机器。随着人工智能技术的快速发展，实现基于自然语言的人机交互成为了可能，于是构建人机对话系统逐渐成为了研究热点。目前，对话系统在工业和日常生活中需求量也非常大。人机对话系统的市场规模预计将从 2021 年的 26 亿美元增长到 2024 年的 94 亿美元，复合年增长率 (CAGR) 为 29.7% [1]。近期，随着 “ChatGPT” 的火热出圈，基于自然语言的人机交互对话再次成为关注的焦点，也将进入发展的新纪元。

对话系统按照功能可以划分为两类：任务型对话系统（Task-oriented Dialogue Systems）和开放域对话系统（Open-domain Dialogue Systems）[2,3]。任务型对话系统面向特定领域解决具体的问题，旨在有效地帮助用户完成车票预订、商品咨询、餐馆查询等特定的任务。目前，任务型对话系统已经成功落地到了一些实际应用中，包括搭载在 iPhone 上的个人数字助理 Siri [2]，搭载在 Windows 系统的台式机和移动设备上的智能个人助理 Cortana [3]。此外，还有百度推出的度秘机器人，小米研发的小爱同学，以及天猫的天猫精灵等。相反，开放域对话系统并不专注于完成特定任务，而是聚焦在和用户建立长期的联系，满足用户的社交需求，比如陪伴交流、情感支持、归属感获取等[4]。开放域对话系统的一些典型代表有：微软小冰（英文名为 XiaoIce） [4]，Pandorabots 推出的 Kuki 聊天机器人 [5]，以及 Facebook 推出的 BlenderBot 聊天机器人 [6]等。

尽管大型人工智能公司已经研制出了一些开放域对话系统，但因为其开放式的目标，构建更好的开放域对话系统仍具有挑战性。首先从目标上来看，开放域对话系统旨在最大化用户的长期参与度，其不够清晰具体的特点使得很难从数学角度对它进行优化[5]。因此，需要提出不同的对话技巧作为更清晰具体的优化目标来提升用户参与度，例如生成有趣有信息量的回复[6,7]、学会提问[8]、提供情感支持[9] 等。并且它要求对话系统能够充分理解对话上下文，在正确的时间选择正确的对话技巧，并产生符合一致性要求的对话回复。其次从系统结构上来看，开放域对话系统需要具备处理开放域知识的能力，这些知识通常没有任

[1] https:// markets.businessinsider.com

[2] https://www.apple.com/siri/

[3] https://www.microsoft.com/en-us/cortana

[4] https://www.xiaoice.com/

[5] https://www.kuki.ai/

[6] https://geo-not-available.blenderbot.ai/

何预定义的任务信息，也没有特定任务的模式或标签信息。因此，近年来构建开放域对话系统的趋势变为开发完全由数据驱动的端到端模型，这些模型利用神经网络直接学习用户输入和系统输出间的映射关系[10,11]。此外从系统输入来看，大多数开放域对话系统并不以现实世界为基础，这使得这些系统无法有效地与用户环境相关的任何内容进行对话。因此，近些年研究人员开始探究如何在现实世界的知识体系中构建开放域对话系统。他们要求开放域对话系统应该额外考虑用户的情绪状态、兴趣话题、角色定位等与用户环境相关的内容去提供更合理的系统回复[12–14]。

开放域对话系统在实现方式上可以分为两大类，分别是检索式对话系统和生成式对话系统[15]。检索式对话系统[16] 首先从大量选项中选择一个候选回复集合，然后根据每个候选与对话上下文的匹配程度选出最合理的回复。由于检索式对话系统的输出源自真实的人类对话，因此它们通常是流畅且易于理解的。同时，它们也是相对安全的，因为可以提前过滤掉数据集中的有害回复。然而，检索式对话系统受到数据集规模和质量的限制，有时检索到的回复与对话上下文的相关性较弱[17]。相比检索式对话系统，生成式对话系统[3] 可以自由地生成训练数据集中不存在的回复，即它们不受限于一组预定义的话语。然而生成的回复不能保证一定正确，有时候它们缺乏和对话上下文的连贯性，并且往往是通用且无趣的回复。得益于近年来深度文本生成技术的快速发展，以及生成式方法的优良特性，即能够便捷地引入额外的信息（如结构化知识图谱、个性化角色设定、以及关注的兴趣话题等）去充分理解对话上下文内容并生成多样且有趣的回复，生成式对话系统逐渐成为了研究热点。

生成式开放域对话系统通常采用序列到序列（Sequence-to-Sequence，Seq2Seq）的神经网络模型来构建。模型基于编码器-解码器结构，以对话上文作为模型输入，模型的编码器对对话上文进行充分的理解，然后解码器逐字逐字地输出对话回复。在模型训练过程中，模型通常以最大化似然估计（Maximum Likelihood Estimation，MSE）作为训练目标。在生成式开放域对话研究中，研究课题主要分为以下三类[2]：（1）充分理解对话内容。不仅要充分理解输入文本、图像或视频等信息，还要能够去辨认说话人的人格特点、个性化信息、当前的情绪状态、关注的兴趣话题，以及所拥有的背景知识等。（2）保持一致性行为。生成的回复不应该和输入的角色信息，或者它之前产生的回复发生冲突。一致性体现在三个方面，即要符合预定义的角色信息，要保持相同的说话风格，以及要和上文中已谈论的内容保持一致。（3）提高交互性。优化对话策略以最大限度的提高用户的长期参与度，对话策略包括情感检测、话题引导、合理提问、可控回复生成等。

1.2 本文工作

本节详细介绍本文的研究目标与内容，以及针对每个研究目标所取得的研究成果。研究脉络如图 1-1所示。

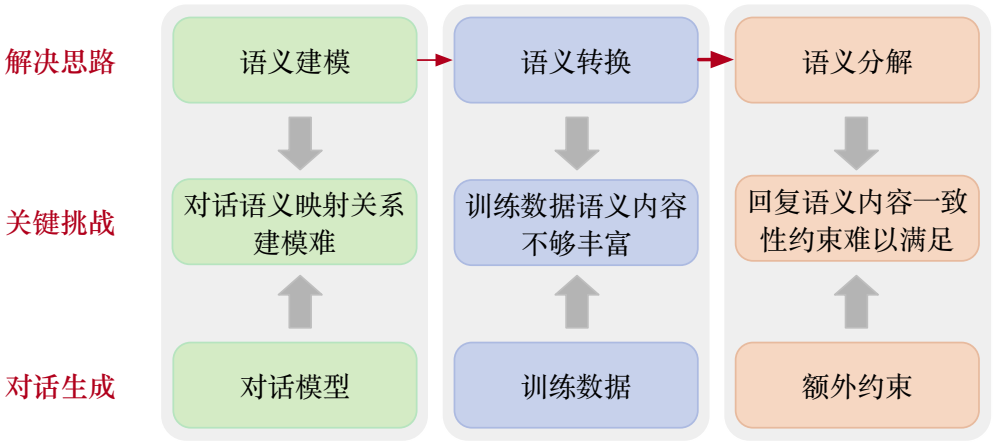


图 1-1 研究脉络图

Figure 1-1 The Framework of research flow

1.2.1　研究目标

开放域对话允许对话上文有各种各样不同语义，但又和对话历史语义保持一致的回复。因此，生成式开放域对话系统首先需要具备能够理解复杂语义映射关系的基本能力，才能生成语义多样的回复。其中，数据驱动的构建方式需要语义内容丰富的对话训练数据。同时，生成的回复还要满足和对话历史中的语义内容不能自相矛盾（即保持一致）的约束。目前生成式开放域对话系统通常以大规模形如 < 对话上文，回复 > 的对话样本作为训练数据，使用编码器-解码器结构的模型建模输入对话上文和输出回复之间的复杂语义映射，进而生成回复。进一步，系统集成一致性检测模型对生成回复进行一致性评分，选择评分最高的输出以满足回复语义内容的一致性约束。随着深度神经网络技术的快速发展，现有系统也取得了不错的成就。但由于开放域对话任务的复杂性，想要对话系统达到实际可用的水平还有很长一段路要走。本文深入分析了现有系统存在的问题以及面临的挑战，主要研究了系统在构建过程中所面临的多个关键性挑战：

1、对话语义映射关系建模难的问题：开放域对话允许对话上文有各种各样不同语义的回复，因此模型需要建模输入对话上文与多个输出回复间的复杂语义映射关系。从人类对话的角度来看，这种映射关系具有层级的特点：每个回复都谈论对话上文的一个语义侧面，而谈论相同语义侧面的回复语义相似但用词表达多样。因此，结构建模也需要采用层级建模：首先一个对话上文可以映射到多个不同可生成回复的语义侧面，每个语义侧面又进一步映射到多个不同表达的回复。现有系统通过引入隐变量来增强模型的建模能力：隐变量在监督训练中隐式地捕捉回复中蕴含的抽象共性信息，可能是说话语气或用词风格等。但这些系统没有建模对话固有的层级语义映射关系，让其自行从对话数据中学习非常困难，这导致隐变量更容易学到混合的语义表示。最终尽管采样不同的隐变量值，系统也只能生成语义相似但表达不同的回复。因此，如何显式地建模层级语义映射关系，是构建开放域对话系统中值得被探索的一个问题。

2、训练数据语义内容不够丰富的问题：开放域对话允许对话上文有各种各样不同的回复。这些回复不仅可以是语义各不相同的，相同语义的回复也可以有很多不同的表达。基于生成式的对话模型完全是数据驱动的，因此训练数据的语义内容丰富程度能够很大程度地决定生成回复的语义多样性上限。然而，由于回复所处的潜在空间无限大，人工编写大规模这样的对话数据是耗时耗力的。此外，尽管社交媒体上存在大量的对话数据，但它们包含很大比例的低质量对话样本。这些样本可能不流畅、缺乏信息量、与对话上文不连贯或者不一致。更重要的是，社交媒体中的回复数据大部分可能来自于相同的语义角度，而其他很多不同语义角度的回复往往出现频率很低，存在角度不均衡现象。因此从社交媒体上的对话数据中选择出高质量、拥有各种不同语义回复的样本同样是耗时耗力的。针对这一问题，如何自动化扩增多个不同语义角度的回复，丰富训练数据的语义内容，是目前构建开放域对话系统亟待解决的一个问题。

3、回复语义内容一致性约束难以满足的问题：为了获得用户长期的信任和好感，生成回复和给定对话上文中语义内容不能相互矛盾（即保持一致）是至关重要的。由于训练数据中可能包含多个谈论相同语义内容的样本，如询问年龄这样的事实信息，而对应的回复通常各不相同，这导致训练的系统在交互过程中多次谈论相同语义内容时会输出不同的回复，这是不可接受的。目前对话系统额外引入一致性检测模型对生成回复进行一致性评分，通过输出得分最高的回复来满足约束。一致性检测任务主要是判断生成回复包含的语义内容是否和给定对话上文中的语义内容相互矛盾。从而该任务可以很自然地被建模成自然语言推理任务（Natural Language Inference，NLI）：通过构造和对话系统训练数据领域一致的 NLI 数据，然后训练检测模型即可。但训练数据是人工构造的，而使用场景是真实对话系统产生的数据。在这种数据分布差异下模型性能出现显著下降，进而使得一致性约束难以被有效满足。因此如何缓解检测模型在数据分布差异下的性能下降是构建开放域对话系统中非常重要的一个问题。

1.2.2 研究内容

针对上述三个挑战，本文随着挑战难度的升级逐步加深对语义信息的挖掘与利用，从基础的语义建模，到显式的语义转换，再到复杂的语义分解，分别提出了相应的解决方案。本文的主要工作包括以下三个部分：

1、针对对话语义映射关系建模难的问题，本文提出一种基于语义表示的回复生成方法，用于显示建模层级语义映射关系，以生成语义多样的回复。这种语义映射关系来源于：对话上文对应多个可生成回复的语义侧面，每个语义侧面可生成若干用词表达多样的回复。为此，该方法沿袭了利用多隐变量增强建模能力的思路。其中模型通过端对端训练让每个隐变量代表不同的语义侧面，并通过训练学习每个隐变量的分布。最终模型随机采样一个隐变量分布，再进一步从分布上随机采样一个隐变量值，基于此值去生成回复。由于对话上文中可生成回复的语义侧面本身是不可知的，并且即使语义侧面已知，受限于训练数据中可观测回复所对应语义侧面的数量，模型也难以学到所有的语义侧面表示。因此，该方法

首先引入了多语义回复检索模块为每个对话上文扩增多个表示不同语义侧面的回复。其次，该方法改进了对话 Wasserstein 自编码器，将不同语义侧面与不同隐变量对齐，并保证不同隐变量分布间距离尽可能大。这样可以更加准确地建模对话上文和回复间的层级语义映射关系，从而保证最终基于采样隐变量值生成的回复不仅是用词多样的，还是语义多样的。我们将该方法在多个公开数据集上进行实验验证。实验结果表明该方法能够生成更多语义多样的回复。同时分析实验表明该方法能够有效地建模开放域对话的层级语义映射关系。

2、针对训练数据语义内容不够丰富的问题，本文提出一种基于语义转换的数据增强方法，用于为每个对话上文扩增不同语义的回复，以提高训练数据的语义内容丰富度。由数据驱动的对话系统的质量上限很大程度上由训练数据的质量来决定。如果训练数据具有人类对话固有的语义内容丰富的特点，那系统则有可能从中学到更真实的语义映射关系，从而生成更类人的回复，其中回复也会更加语义多样。为此，该数据增强方法借鉴了人类生成不同语义回复的过程。给定一个对话上文，人首先关注到对话上文中的某部分内容，进而关注点转移到一个想要谈论的语义角度去生成一个回复。但随后人会去思考一个问题：如果改变当前关注的语义角度，回复会有什么不同。人在否定过去发生的事情并进行重新推理时，会保持当前环境不变，即除了改变语义角度，其他影响回复生成的因素（如人当前的情绪状态、说话风格等）是不变的。这种在当前环境下的推理就是所谓的反事实推理，基于一定的事实基础有利于保证推理结果的质量。因此，该方法借助反事实推理，通过干预可生成回复的语义角度去推理更多语义不同的回复。首先将回复生成模型转化成可用于反事实推理的结构因果模型（Structural Causal Model，SCM）。然后利用 SCM 中的不可观测变量来建模当前环境。进一步为了得到可转换的语义角度，该方法基于训练数据构建了对话上文关注点与语义角度间的转移关系图，然后利用转移关系图去预测有效的语义角度。最终混合原始数据和增广数据去训练对话模型以提升模型的性能。实验结果表明该方法能够有效扩增不同语义的回复，并能提升下游任务的性能，尤其是对话生成的语义多样性。

3、针对回复语义内容不一致约束难以满足的问题，本文提出一种基于语义分解的一致性检测方法，用于缓解一致性检测模型在数据分布差异下的性能下降，以使得一致性约束被更有效地满足。模型在数据分布差异下出现性能下降表明模型鲁棒性差。这主要和模型倾向于探索训练数据中的伪关系有关。为此，该方法借鉴了其他任务提升模型鲁棒性的思想，即构造反事实样本，并与原始样本混合去训练模型，以缓解模型对伪关系的依赖，进而提升模型的鲁棒性。反事实样本是指和原始样本很相似但标签相反的样本。构造反事实样本要保证两件事情：反转标签和确保与原始样本尽可能相似。因此，该方法首先辨认出所有原始样本中不一致（矛盾）的内容，然后通过修改最少的不一致内容使得原始样本的标签反转。具体来说，对于不一致样本，通过删除最少的矛盾内容将样本变得一致；对于一致样本，通过添加一对相互矛盾的内容是样本变得不一致。为了辨别

出不一致的内容，该方法利用语义分解思想先将复杂对话分解成多个独立的语义内容单元，再通过判断哪些语义内容自相矛盾去标记不一致内容。抽取复杂的对话话语中所有传达的事实观点，通过辨认这些事实观点是否冲突来自动化标记出冲突内容。该方法在多个检测模型上进行试验验证。实验结果表明该方法能够缓解模型对伪关系的依赖并有效提升模型的鲁棒性，从而更好地满足回复语义内容和对话上文中内容保持一致的约束。

1.3 论文结构

本文共分为六章，其具体组织结构如下：

第一章首先介绍了开放域对话系统的研究背景以及重要研究意义，并进一步指出了目前基于生成式方法的开放域对话系统存在的问题以及面临的重大挑战，最后简要介绍了本文的主要研究目标与内容，以及相应的研究成果。

第二章主要对开放域对话系统目前的进展进行了梳理。首先介绍了开放域对话系统的主要建模方法，包括基于检索式的建模方法和基于生成式的建模方法；其次按照不同改进思路分门别类地介绍了提升开放域对话回复生成质量的相关工作；最后介绍了开放域对话系统目前主要采用的评价方法。

第三章研究了对话语义映射关系建模难的问题。首先阐述了人类对话固有的层级语义映射关系，并分析了现有建模方法所面临的主要问题；其次介绍了本研究的基础模型，对话 Wasserstein 自编码器；然后详细介绍了本研究设计的多语义 Wasserstein 自编码器：首先描述如何获取可生成回复的语义侧面的信息，其次介绍如何将隐变量和语义侧面一一对齐，以及如何减少隐变量表示间的重叠；最后介绍了实验所用的数据集、实现细节以及评价指标等，然后给出了模型在多个开放域对话数据集上实验结果与分析。

第四章研究了训练数据语义内容不够丰富的问题。首先阐述了开放域对话系统训练数据存在的问题，并指出了已有数据增强方法存在的问题；然后介绍了开放域对话系统的任务定义，以及回顾了用于反事实推理的结构因果模型（SCM）的基本概念；之后详细介绍了本文提出的数据增强方法：首先描述如何建模当前环境，其次介绍如何获得不同且有效的可生成回复的语义角度，然后介绍如何基于当前环境向量和新的语义角度去生成不同语义的回复；随后描述了数据过滤方法；最后，介绍了使用的数据集、实现细节以及评价指标等，然后给出了实验结果与实验分析。

第五章研究了回复语义内容一致性约束难以满足的问题。首先分析了一致性检测任务的重要性，以及目前的检测模型所面临的挑战；然后阐述了本文方法提出的动机，以及要面临的重大挑战；随后，详细描述了如何从对话样本中自动识别出不一致（矛盾）的内容，以及如何通过操作不一致的内容去构造相应的反事实样本；最后给出在真实使用场景下多个检测模型上的实验结果，并对所设计的方法进行了分析讨论，验证了方法的有效性。

第六章对以上几个研究工作进行了总结，并指出了本文的主要贡献以及创新点；然后进一步阐述了本文所涉及的多个研究工作之间的联系，以及展望了开放域对话系统未来值得深入探索的方向。

第 2 章　研究现状

本章详细介绍了开放域对话系统的研究现状与相关研究进展。首先回顾了开放域对话系统主流的两类建模方法：基于检索的对话建模方法和基于生成的对话建模方法。其次，本章介绍了目前提升回复生成质量的主要研究方向，以及每类方向的相关工作。最后，本章介绍了目前基于生成式方法的开放域对话系统常用的评估方法和相应的评价指标。

2.1　开放域对话系统的主流建模方法

开放域对话系统一般使用端到端的编码器-解码器模型结构实现，其核心是回复生成。对话模型以对话历史 $\boldsymbol{X}$ 作为输入，输出回复 $\boldsymbol{Y} = Y_1 Y_2 \dots Y_m$，即

$$\hat{\boldsymbol{Y}} = \arg\max_{\boldsymbol{Y} \in \Omega} P_\theta(\boldsymbol{Y}|\boldsymbol{X}), \tag{2.11}$$

其中，Ω 代表全部候选回复集合，P_θ 为可学习的候选回复打分函数。模型通过最大化搜索算法以寻找候选回复集合中得分最高的那一个作为最合适的回复。

上述公式统一解释了构建开放域对话系统的两种主流建模方法：基于检索的对话建模方法和基于生成的对话建模方法。对于检索式方法，搜索空间 Ω 通过从预定义的人类对话数据集中检索候选集合来获得，这些人类对话都是形如 < 对话上文，回复 > 的样本。$P_\theta(\boldsymbol{Y}|\boldsymbol{X})$ 则可以由文本匹配模型来实现，该模型计算每个候选回复和给定对话上文的匹配分数。对于生成式方法，搜索空间 Ω 充满整个 m 维词向量空间，即 $\boldsymbol{Y} \in \mathbf{V}^m$，其中 $\mathbf{V}$ 表示词表大小，m 表示回复序列长度。$P_\theta(\boldsymbol{Y}|\boldsymbol{X})$ 通常由自回归的 Seq2seq 模型实现，即模型在充分理解对话上文后逐字逐字地生成回复序列。

接下来，我们详细介绍两类主流建模方法的任务定义，以及相应的研究内容，面临的挑战，以及目前的研究进展。

2.1.1　基于检索的对话建模方法

基于检索的对话建模方法旨在使用任意检索算法从预定义的人类对话数据集中找出最适合于给定对话上文的回复[18,19]。图 2-1中展示了基于检索的对话建模方法的流程。在这种任务设置下，该建模方法首先利用给定对话上文检索 N 个相似的对话上文，并把这些检索出的对话上文对应的回复作为给定对话上文的候选回复集合。然后使用文本匹配模型 $P_\theta(\boldsymbol{Y}|\boldsymbol{X})$ 计算候选回复与当前对话上文的匹配分数，选择得分最高的那个作为回复 $\boldsymbol{Y}$ 输出。模型参数 θ 通常通过最小化基于间隔的成对排序损失函数（Margin-based Pairwise Ranking Loss）来优化，即

$$\mathcal{L} = \max(0, \gamma + match_\theta(\boldsymbol{Y}_-, \boldsymbol{X}) - match_\theta(\boldsymbol{Y}_+, \boldsymbol{X})), \tag{2.12}$$

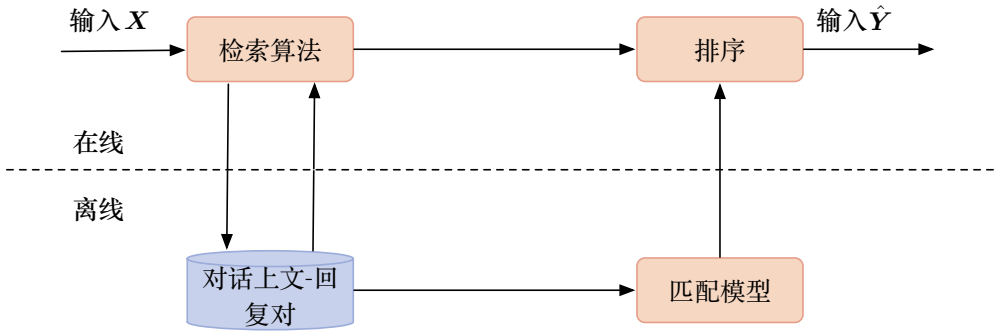


图 2-1 基于检索式建模方法的流程图

Figure 2-1 The Framework of the retrieve-based dialogue method

其中，γ 表示间隔大小，是一个超参数。$\boldsymbol{Y}_+$ 为给定对话数据集中的参考回复（正例），而 $\boldsymbol{Y}_-$ 可以通过随机从数据集中采样得到，或者通过扰动 $\boldsymbol{Y}_+$ 得到。$match_\theta(\boldsymbol{Y}, \boldsymbol{X})$ 表示可学习的匹配函数。进一步可以用似然函数来定义损失函数：

$$\begin{aligned} \mathcal{L} &= -\log P_\theta(\boldsymbol{Y}_+|\boldsymbol{X}), \\ P_\theta(\boldsymbol{Y}_+|\boldsymbol{X}) &= \frac{\exp\{match_\theta(\boldsymbol{Y}_+, \boldsymbol{X})\}}{\exp\{match_\theta(\boldsymbol{Y}_+, \boldsymbol{X})\} + \sum_{i=1}^{k} \exp\{match_\theta(\boldsymbol{Y}_-^i, \boldsymbol{X})\}}. \end{aligned} \tag{2.13}$$

Huang 等[2] 指出 $match_\theta(\boldsymbol{Y}, \boldsymbol{X})$ 可以分为两类：浅层交互网络（Shallow Interaction Network）和深层交互网络（Deep Interaction Network）。浅层交互网络是指首先独立编码对话上文 $\boldsymbol{X}$ 和回复 $\boldsymbol{Y}$ 为固定长度向量，然后对他们进行浅层交互（例如，相加、点对点相乘等），最后送入分类层输出是否匹配的标签。相比之下，深层交互网络直接利用交互网络交互 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 得到一个融合向量，然后送入分类层进行标签输出。早期的基于检索式的建模方法的工作主要采用浅层交互网络[20–22]，这类工作的重点是如何更好地学习对话上文和回复的独立向量表示。由于浅层交互网络独立编码对话上文和回复，缺乏深入交互匹配过程，因此深度交互网络逐步取代了浅层交互网络。采用深度交互网络的工作[23–27] 的研究重点就是如何设计更好地交互网络去充分融合对话上文和回复信息。目前性能最好的深度交互网络是各种各样基于编码器结构的预训练模型，比如 BERT[28] 和 RoBERTa[29] 等。预训练模型通常包含大量的参数，通过在大规模语料上训练获得良好的上下文理解能力。这一能力正是匹配模型所需要的。因此，后续研究工作多采用在预训练模型上用对话回复检索语料微调的方式[30–33] 去建模对话上文和回复间的交叉注意力或复杂交互。

2.1.2 基于生成的对话建模方法

深度神经生成模型目前已广泛用于开放域对话生成中，其中 Seq2seq 模型[10,11,34,35] 已经成为对话生成任务的首要选择。随后，一些生成式模型，例如条件自编码器（Conditional Variation Auto-Encoder, CVAE）[36–41] 和生成式对抗网络（Generative Adversarial Network, GAN）[42,43]，也逐渐被用于对话生成中。

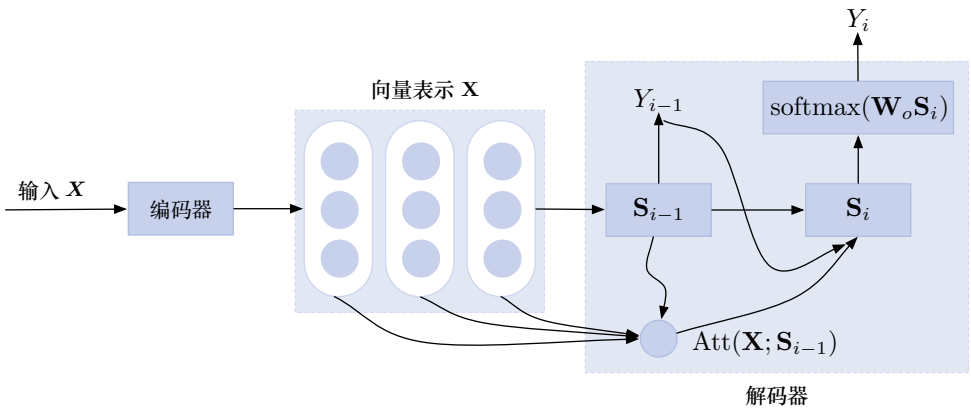


图 2-2 基于生成式建模方法的流程图

Figure 2-2 The Framework of the generated-based dialogue method

随着预训练模型的兴起，他们很快成为了主流选择[44–48]，预训练模型通过在大规模语料上训练获得了强大的文本理解与生成能力，已经在对话生成领域展现出强大的性能。生成式对话模型通常可以表示为

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{i=1}^{m} P(Y_i|\boldsymbol{Y}_{<i}; \boldsymbol{X}), \tag{2.14}$$

其中，$\boldsymbol{Y}_{<i}$ 为模型在之前时间步生成的回复序列。在 i-th 时间步，模型根据分类概率 $P(Y_i|\boldsymbol{Y}_{<i}; \boldsymbol{X})$ 进行采样得到回复的第 i 个单词。对话生成模型通常采用编码器-解码器（Encoder-Decoder）结构，图 2-2中展示了基于生成的对话建模方法的流程。首先，编码器将输入对话上文 $\boldsymbol{X}$ 转换成固定长度的语义向量，即

$$\mathbf{X} = \text{Encoder}(\boldsymbol{X}). \tag{2.15}$$

随后，解码器依次在每个时间步更新状态向量 $\mathbf{S}_i$：

$$\mathbf{S}_i = \text{Decoder}(\mathbf{S}_{i-1}, [\text{Att}(\mathbf{X}; \mathbf{S}_{i-1}); \boldsymbol{Y}_{i-1}]), \tag{2.16}$$

其中 $\text{Att}(\mathbf{X}; \mathbf{D}_{i-1})$ 是前一时刻的状态 $\mathbf{D}_{i-1}$ 对输入向量进行交叉注意力操作的结果。$\boldsymbol{Y}_{i-1}$ 为前一时刻生成的词语 Y_{i-1} 的向量表示。最后，模型根据分类概率采样获得一个生成的词语 Y_i，即

$$Y_i \sim \text{softmax}(\mathbf{W}_o\mathbf{S}_i), \tag{2.17}$$

其中 $\mathbf{W}_o$ 为解码器的权重矩阵，$\text{softmax}(\cdot)$ 为 Softmax 激活函数。

早期用于生成式对话模型的基本单元是循环神经网络结构（Recurrent Neural Network，RNN）[49]，包括长短时记忆网络（Long Short-Term Memory, LSTM）[50]和门控循环单元 (Gated Recurrent Unit, GRU)[51]。2015 年，Bahdanau 等[52] 首次提出基于 RNN 的 Seq2seq 模型。它首先将输入句子通过多层 RNN(·) 进行编码，得到一个固定维度的向量表示，然后使用另一个 RNN 模型进行解码，解码的目

标是最大化目标语言句子的生成概率。在生成目标句子的每一个词语时，利用注意力机制计算它对应输入句子的注意力应该在哪些词语上。随后，Sordoni 等[53]和 Shang 等[10]开始将这一模型用于对话生成，生成了质量不错的回复。然而，基于 RNN 结构的 Seq2Seq 模型只支持串行训练的结构，难以用于大规模快速并行训练。

随着 Transformer[54]架构的出现，它凭借着基于自注意力机制和支持并行训练的架构优势迅速在文本生成领域得到了广泛的应用，对话生成也不例外。该架构由堆叠了 N 个基本块的编码器和解码器组成。对于编码器来说，每块均由多头注意力（Multi-Head Attention）子层和前馈神经网络（Feed Forward）子层构成，子层之后都进行了残差（Add）和层正规化（Norm）操作。对于解码器，其整体结构与编码器保持不变，只是在多头注意力子层和前馈神经网络子层之间引入了用于与源端进行交互的编码器-解码器多头注意力子层。在训练过程中，为了防止探测到未来的信息，解码器的多头自注意力子层使用掩码表示（Masked Multi-Head Attention）。不像 RNN 中每一个时间步的结果都基于历史时间步的内容，Transformer 的并行特性导致无法获取时序信息，因此在输入中引入了位置编码（Positional Encoding）来弥补这一缺陷。

Transformer 支持并行训练的特性为大规模预训练模型的出现提供了条件支撑，预训练模型也给对话生成任务带来了很大的性能提升。对于对话生成而言，编码器可以是基于 Encoder 结构的预训练语言表示模型（例如 BERT[28]、RoBERTa[29]、XLNet[55] 等），解码器可以是基于 Decoder 结构的预训练语言模型（例如 GPT 系列模型[45,56,57]）。其中，编码器可以缺失，直接基于预训练语言模型进行对话生成。此外，还可以利用基于 Encoder-Decoder 结构的预训练模型进行对话生成，包括 T5[58]、BART[59] 等。利用预训练模型做对话生成主要采用的是“预训练-微调”的范式，即利用预训练模型的参数初始化生成式对话模型的参数，然后在相应的对话训练语料上微调模型。预训练模型在大规模语料上通过自监督学习任务习得了良好的上下文理解能力和文本生成能力，在此基础上，通过少量领域数据微调即可获得质量不错的回复。随着预训练模型参数规模越来越大，微调大模型的全量参数需要更多的 GPU 计算资源和领域数据。因此，模板学习（Prompt Learning）[60] 逐渐兴起，成为了新的预训练模型使用范式。模板学习旨在对输入的文本信息按照特定模板进行处理，并将任务重构成一个能够充分利用预训练语言模型知识的形式。在对话生成任务中，模板学习可通过不微调或者只微调少量预训练模型的参数，去提升回复生成的质量。

2.2 相关工作

开放域对话的一大特点为它允许对话上文有各种各样不同的回复。因此，基于生成式的开放域对话系统需要产生各种各样合理的语义丰富多样、并能和对话上文保持一致的回复。目前常用于对话生成的编码器-解码器结构最早是在机器翻译领域提出的[52]，机器翻译利用编码器将源语言编码成向量表示，再利用

解码器将源语言向量解码成目标语言。机器翻译的源语言和目标语言具有良好的语义对齐关系（一对一映射关系），因此编码器-解码器结构的模型是可以取得不错结果的。但是在开放域对话场景下，对于同一个对话上文，不同的人或者同一个人在不同情境下都会产生不一样的回复。这种语义不对齐且复杂的一对多关系使得传统的编码器-解码器模型难以充分理解对话上下文，从而产生合理且和上文保持一致的回复。更多的时候，模型只会产生乏味的通用回复，例如“我不知道。”、“我也这么觉得”等；或者只会简单地重复用户提到的内容，例如用户说到“我是学生”，模型则会产生“我也是学生”这样的回复，并不会考虑上下文语境或者考虑是否和上文保持了一致。为了生成语义丰富多样且和上文能保持一致的回复，目前的工作主要从数据、模型和约束三个方面进行改进。

2.2.1　增强数据信息

训练高质量的开放域对话生成模型需要大规模对话数据，因此自动化扩增更多的对话样本受到了很多研究人员的关注。此外，考虑到对话生成模型容易生成不流畅且质量不佳的回复，而对话检索模型从预定义的人类对话预料中检索输出回复，此类人工对话流畅、语法正确且质量高。因此，如何利用检索回复用来增强生成模型输出的回复质量也成为了研究的重要问题。进一步，对话历史中蕴含着丰富的隐含信息，如何从中挖掘它们并辅助模型提升回复生成的质量同样是很研究价值的方向。接下里我们详细介绍上述三个方向的研究工作。

数据扩增　训练对话系统需要大规模的高质量的对话数据集。对话数据允许一个对话上文拥有很多回复。具体来说，一个给定的对话上文可以有很多不同角度的回复，相同角度的回复也可以有很多不同的表达方式。然而人工收集这种高质量数据耗时耗力。数据增强是解决这一问题的常用手段，目前数据增强开始广泛用在各种自然语言处理任务上，已有许多研究者进行了综述[3,61–64]。在开放域对话生成领域，早期的数据增强方法主要通过简单扰动现有数据去得到新的样本。这类工作主要使用启发式规则[65] 或者基于改写的方法[66–71] 扰动现有数据。Du 等[65] 通过置换和翻转对话历史来增广对话选择任务的数据。Niu 等[66] 定义了一系列词语级别的扰动操作，通过强化学习去学习如何选择合适的操作用于数据增强。类似地，Cai 等[68] 也设计了一系列词语级别和句子级别的数据增强方法，分别利用掩码语言模型[28] 和反向翻译[72] 来实现。Xie 等[70] 提出一种序列级数据增强方法，通过使用解码器输出的软标签来扩增解码器的输入信息。简单扰动只带来了有限的语义修改，难以扩增出语义丰富多样的对话数据。

后来，一些工作则利用生成模型去合成数据。早期 Li 等[67] 利用条件变分自编码器 (Conditional Variational Auto-Encoder，CVAE) 作为生成器去产生更多的数据。随着预训练模型的兴起，近期更多工作[73–76] 利用大规模预训练模型去扩增数据。Chang 等[73] 和 Yang 等[74] 尝试利用预训练语言模型 GPT-2[56] 生成新的文本样本。Schick 等[75] 采用更大的预训练语言模型 GPT-XL 合成未见过的样本。这些工作首先利用对话训练数据微调预训练语言模型，然后给定对话上文

作为微调后的模型的输入，模型采样出不一样的回复。进一步， Wang 等[76] 开始尝试利用超大规模预训练语言模型 GPT-3[57] 合成数据，该工作利用少量提示让模型通过上下文学习（In-Context Learning）去合成高质量的数据，整个操作过程无需人工注释。

检索回复增强 利用检索数据增强对话生成模型可以有效地结合基于检索式方法和基于生成式方法的优势[77–79]。这些方法通常采用两阶段建模。在第一阶段中，模型利用给定的输入对话上文从预定义数据集中检索出相关对话回复[80]；在第二阶段中，这些相关的回复被用于帮助模型生成新的回复。 Song 等[81] 使用额外的编码器来编码检索到的回复集合，并在解码中应用注意力机制[82] 和拷贝模型（CopyNet）[83] 生成新的回复。类似地， Pandey 等[84] 首先使用 TF-IDF 模型从训练数据中检索相似的对话。检索到的回复作为参考示例，解码器使用这些示例来生成新的回复。此外， Wu 等[80] 同样先从训练数据中检索出相似对话，然后根据相似对话上下文和当前上下文之间的差异编辑相似对话的回复。该工作的动机是检索到的对话回复为生成提供了一个良好的起点，它是语法正确的，并且包含丰富的信息，后期编辑过程可进一步提高检索回复的相关性和连贯性。进一步， Zhang 等[85] 提出了一种对抗性学习框架，模型由类似于语言模型的生成器、排序生成器和排序鉴别器组成。该模型鼓励两个生成器生成判别排名更高的回复，而鉴别器则降低对抗性样本的权重并选择两个生成器都满意的回复。随后，Cai 等[86] 提出了一个新的框架，它通过“骨架到回复（Skeleton-to-Response）”的范式利用检索结果。首先，从检索对话中提取骨架，然后生成的骨架和原始对话上文用于通过新的回复。并且 Cai 等[87] 认为如何精确提取骨架，以及如何有效地利用检索结果增强回复生成仍具有挑战性。该工作则提出了一种新颖的框架，其中骨架提取由可解释的匹配模型完成，随后的骨架指导回复生成则由单独训练的生成器完成。

隐含信息挖掘 开放域对话的对话上文和回复之间存在复杂的一对多映射关系。如何从对话中挖掘隐含的信息去辅助对话模型更好地建模一对多映射关系，是一个很关键的问题。从人类对话角度来思考：日常对话通常会涉及一个主题和目标。实际上，主题和目标是让每个参与者参与对话的关键，因此对于开放域对话系统来说也是必不可少的。早期的工作主要关注于挖掘局部的词级别主题信息。Serban 等[88] 提出使用层级编码器-解码器结构建模对话中隐含的主题信息。该模型先利用 RNN 获取每句话语的表示，然后再利用一个 RNN 编码话语表示得到对话的表示向量，该对话向量则表示隐含的主题信息，进而模型基于该主题信息生成新的回复。 Xing 等[13] 使用 LDA 提取主题词，并将这些词编码到主题感知模型中。该模型通过联合关注对话上文和主题词来生成响应。同样的主题词使用的思路也用在了对话检索任务中[27]。此外， Chen 等[89] 构建了一个开放域对话系统 Gunrock，它对每个随机分割的文本片段进行“主题”分类，然后分配特定领域的对话模块以生成响应。和利用主题词类似的还有挖掘对话意图相关的

工作，这类工作也是挖掘高层次信息辅助对话生成，比如是否有礼貌[90]，当前情绪状态[91]，隐含的个人信息[92]，以及对话回复意图[93] 等。

开放域对话历史中通常涉及多个主题，并且主题间发生转移是很常见的现象。因此，只获取单个主题的表示是不够的，如何建模对话主题转移成为一个很有研究价值的问题。Zhang 等[94] 挖掘了对话上文中隐含的常识信息，并利用常识知识图谱显式建模对话主题转移关系。利用通用知识图谱节点间的关系来辅助建模对话主题转移的思路还体现在 Wu 等[95] 和 Wu 等[96] 的工作中。考虑到对话中的主体转移关系可能和通用知识图谱间的节点转移存在差异，因此一部分研究工作尝试从对话上文中抽取关键词代表主题，利用启发式方法构建转移关系图。Xu 等[97] 提出将有关对话转换的先验信息构建成图，并学习基于图的对话策略。随后，Zou 等[98] 指出同一主题可能包含多个关键词，因此该工作提出首先建模多关键词下的主题转移关系图，然后通过从图中适合于当前对话的关键词去生成回复。

2.2.2　增强模型表示

为了生成语义丰富多样且能和对话上文保持一致的回复，研究者们在传统的 Seq2seq 模型基础上使用更灵活的中间表示（例如，额外的隐变量）去增强模型的表示能力，旨在解决开放域对话中“一对多映射”的问题。此外，随着预训练模型的发展，研究者们开始将预训练模型用于开放域对话模型的构建中。预训练模型能够一定程度上缓解数据稀缺的问题，并且具有优秀的模型表示能力，这正是开放域对话一对多关系建模所需要的能力。因此，许多研究者进一步基于预训练模型设计新的对话生成模型，去提升回复生成的质量。除了上述两方面，还有一些工作通过设计排序模型对输出回复进行后处理，以确保回复是语义丰富多样的，并且和上文保持一致。接下来我们将介绍中间表示增强、预训练模型和回复排序三个方面的相关工作。

中间表示增强　不满足于只将输入对话上文编码成固定长度向量，一些研究工作引入隐变量去增强模型的表示能力，以及对回复生成的控制能力。Zhao 等[41] 提出了 CVAE 用于对话生成，限定隐变量服从正态分布，从而通过采样得到不同的隐变量值去生成更加多样的回复。另外，Serban 等[38] 提出在层级编码器-解码器模型（Hierarchical Recurrent Encoder-Decoder, HRED）中引入隐变量，即变分层级编码器-解码器模型模型 (Variational Hierarchical Recurrent Encoder-Decoder, VHRED)。Chen 等[99] 进一步利用记忆网络对隐变量进行存储，然后通过记忆网络的匹配结果进行回复生成。Park 等[100] 在 VHRED 模型中继续引入了一个对话层级的隐变量，用于解决对话后验坍塌的问题，并更好地保证回复和对话上文的一致性。上述方法主要采用的是连续性隐变量，可解释性不强。Zhao 等[40] 和 Gao 等[101] 使用离散型隐变量，提高了隐变量表示的可解释性。考虑到开放域对话回复的潜在空间非常复杂，单个隐变量难以进行估计。多隐变量机制逐渐代替了单隐变量机制。Gao 等[101] 提出使用多个不同的关键词代表不同的隐变

量，进行提升模型的多样性。Gu 等[102] 提出利用混合高斯分布表示多隐变量机制，进而建模对话的一对多映射关系去输出多样的回复。Zhou 等[103] 和 Zhou 等[104] 假设存在一些潜在的回复机制，每个机制都可以为单个输入对话上文生成不同的回复。这些回复机制被建模为隐变量，用于将输入转化为不同的回复机制，从而实现可控回复生成。

预训练模型应用 随着预训练模型的兴起，绝大多数对话生成模型开始采用预训练表示模型（例如 BERT[28] 模型）代替传统的 LSTM[50] 或者 Transformer[54] 模型获取良好的对话上文表示，或者利用预训练语言模型（例如 GPT 系列模型[45,56,57]）或者预训练编码器-解码器模型（例如 T5[58] 模型）直接作为基础模型，然后使用相关领域数据微调基础模型即可。Zhang 等[47] 和 Wolf 等[105] 通过在不同规模的社交媒体数据上微调 GPT-2[56] 模型，在对话生成任务上都取得了可观的进展。随后，一些工作尝试基于预训练模型进行改进，例如与外部知识结合[106,107]，融合多输入源[44,108]，引入隐变量[109]，以及对话规划[110] 等。这类工作为了适应预训练模型的数据格式，通过拼接对话上文作为模型输入，忽略了对话的层级结构。并且为这些预训练模型设计的自监督训练任务更适合于学习通用文本的理解能力和生成能力，无法匹配对话生成任务。因此，Gu 等[111] 提出 DialogBERT 模型，即采用层次化 BERT 模型的思想，学习长序列对话上下文丰富的语义。另一些工作则关注于设计更精妙的自监督任务，在不改变模型结构的基础上，帮助模型更好地学习对话的语义内容和层次结构信息。Wolf 等[105] 采用额外的无监督目标来预训练对话语言模型。Huang 等[112] 提出了一个自监督学习框架，用于从用户的对话历史记录中为个性化聊天机器人捕获更好的表示，从而确保回复的一致性。

为了更加适应对话任务场景，一些工作开始专注于大规模对话预训练模型，包括百度的 PLATO 系列模型[109,113–115]，微软的 DialoGPT[47]，谷歌的 Meena[116]，和 Facebook 的 Blender[117] 等。区别于传统的生成式预训练模型，例如 GPT 和 BART，对话预训练模型更加适合对话生成任务。明显的区别就是在于传统的生成式预训练模型的训练数据来自于百科、新闻、小说数据，而对话预训练模型使用的是对话数据进行训练。目前，对话预训练模型的结构大体分为三类：基于 Transformer 的编码器-解码器结构、基于 Transformer 的解码器结构和在 Transformer 的 Encoder 结构基础上改进的 UniLM-based 结构。对于编码器-解码器结构而言，代表模型为 Meena 和第一二代 BlenderBot。Meena 是一个采用单层 Evolved Transformer[118] 作为编码器，13 层的 Evolved Transformer 作为解码器的序列到序列模型。第一代和第二代 BlenderBot 完全采用基于原始 Transformer 的编码器-解码器结构。相比第一代，第二代在此结构基础上，引入了长时记忆模块使得模型具备了记忆较长对话历史的能力，并结合了互联网搜索模块通过调用公开可用的搜索引擎 API，使模型具备了利用即时更新的世界信息展开对话的能力。对于解码器结构而言，代表模型有 DialoGPT、CDial-GPT[119]、LaMDA[120]，以及 ChatGPT。其中，LaMDA 在采用基于 Transformer 解码器结构的基础上，将

注意力替换成了和 T5 一样的相对注意力形式。LaMDA 引入了人类反馈显著地提升了模型在生成质量和安全性上的表现，并引入了一个工具包（包括一个计算器、一个翻译器、以及一个信息检索系统）来提升模型生成回复的有根性（即是否有根据，是否符合事实）。此外，ChatGPT 基于 GPT-3.5 系列模型，采用了和 InstructGPT[121] 类似的方法，通过从人类反馈中进行强化学习（Reinforcement Learning from Human Feedback, RLHF）进行微调。在 Transformer 的 Encoder 结构基础上改进的 UniLM-based 结构而言，代表模型是 PLATO 系列模型，该系列包括 PLATO-1[109]，PLATO-2[113]，PLATO-XL[114] 和 PLATO-K[115]。这些模型一直使用的是 UniLM[122] 的模型架构，能够灵活地结合双向的上下文理解和单向的回复生成。

2.2.3　增强约束限制

为了获得语义丰富多样的回复，集束搜索或者按概率采样（包括随机采样、Top-k 采样和 Top-p 采样）经常被用于生成多个候选回复，然后用一个额外的模型排序候选回复并选择得分最高的输出。引入排序模型通常是因为一些信息无法在解码中使用（例如，输入和响应之间的互信息），或者在解码中使用成本太高（例如，大型预训练语言模型，如 BERT）。 Li 等[6] 提出使用最大互信息（Maximum Mutual Information，MMI）对候选回复进行排序，以促进生成多样的回复。 Fang 等[123] 提出了利用用户的个性信息和满意度历史重新排序以优化个性化一致性。 Challa 等[124] 提出使用“生成-过滤-排序”三阶段框架，其中首先过滤候选以消除不可接受的回复，然后进行排序以选择最佳回复。此外，为了获得和对话上文尽可能保持一致的回复，一些研究者通过使用一致性检测模型引入“自然语言推理（Natural Language Inference，NLI）”相关的知识去排序候选回复。 Welleck 等[125] 和 Song 等[126] 通过构造个性化相关的一致性检测数据，并用这些数据训练相应的检测模型。最终使用检测模型给候选回复进行一致性评分，选择得分最高的作为最终回复输出。进一步， Nie 等[127] 构建开放域对话一致性检测数据，同样通过该数据训练一个检测模型，并用该模型对开放域对话系统的回复进行排序，进而选择最优回复输出，以此来提升对话生成的一致性。

2.3　评估方法

评估开放域对话生成的质量一直是一个具有挑战性的问题。因为与任务型对话系统的评估不同，它没有明确的评价指标，例如任务完成率或任务完成需要的成本。目前评估主要分为自动评估和人工评估两种方式。

2.3.1　自动评估

由于开放域对话的开放特性，很难用一个指标评估生成回复的质量好坏。实际应用中，通常使用若干个不同的指标从不同角度来评估生成回复的质量。评估角度主要包括相关性、多样性和流畅性等。这些评估指标可以分为基于规则的指

标和基于模型的指标[128]。其中，基于规则的指标使用启发式规则去评估生成回复，而基于模型的指标则是使用特定的对话数据训练得到。

基于规则的评估指标 早期，自动评估主要采用基于规则的评估指标。

- 对于相关性评估，使用的指标有基于 n-gram 重叠率的 BLEU 系列指标，包括 BLEU[129]、METEOR[130] 和 ROUGE[131]，以及基于词嵌入向量的语义相似性指标，包括 Embedding Average、Greedy Matching 和 Vector Extrema。

BLEU 用于评估生成回复与参考回复中 n-gram 重合的比例，具体计算方式如下：

$$\text{BLEU} = W_{bp} \exp\left(\sum_{i=1}^{N} W_n \log P_n\right), \tag{2.31}$$

其中 W_{bp} 表示长度惩罚因子（Brevity Penalty），用于惩罚过短的生成回复，计算方式如下：

$$W_{bp} = \begin{cases} 1, & \text{if } l'_r > l_r \\ \exp\left(1 - \frac{l_r}{l'_r}\right), & \text{if } l'_r \leq l_r \end{cases}. \tag{2.32}$$

l'_r 表示生成回复的长度，l_r 为参考回复的长度。W_n 表示权重系数，默认采用均匀加权，即 $W_n = \frac{1}{N}$，N 的上限为 4。P_n 表示修正后的 n-gram 的准确率，计算公式为

$$P_n = \frac{\sum_{i \in n-gram} \min(h_i(r'), \max(h_i(r)))}{\sum_{i \in n-gram}(h_i(r'))}, \tag{2.33}$$

其中 $h_i(r')$ 表示 i 在生成回复中出现的次数，而 $h_i(r)$ 表示 i 在中参考回复出现的次数。METEOR 和 ROUGE 进一步针 BLEU 的固有缺陷进行改进，其中 METEOR 将词干和同义词合并到其计算中，而 ROUGE 则侧重于 n-gram 召回率而不是准确率。然而开放域对话中，相同的对话上文可能有各种各样不同表达的回复。基于 BLEU 系列指标得分低并不一定表示质量差，毕竟合理有效的参考回复的数量总是有限的。

因此，Liu 等[132] 提出了基于词嵌入向量相似性指标，用于评估生成回复与参考回复间的语义相似度。其中，Embedding Average 将回复中每个单词 w 的词向量 e_w 求平均来作为回复的特征，计算生成的回复和参考回复的特征的余弦相似度，计算公式如下：

$$\text{Average}(r', r) = \cos(\bar{e}'_r, \bar{e}_r), \tag{2.34}$$

$$\bar{e}_r = \frac{\sum_{w \in r} e_w}{|\sum_{w \in r} e_w|}. \tag{2.35}$$

Greedy Matching 寻找生成的回复和参考回复中最相似的一对单词，把这对单词的相似度近似为回复间的距离，计算公式如下：

$$\text{Greedy}(r', r) = \frac{\text{G}(r', r) + \text{G}(r, r')}{2}, \tag{2.36}$$

$$\mathrm{G}(r', r) = \frac{\sum_{w' \in r'} \max_{w \in r} \text{cosine-similarity}(e'_w, e_w)}{l'_r}. \tag{2.37}$$

由于匹配函数 $G(\cdot)$ 是非对称的，因此需要计算 r' 与 r 之间，以及 r 与 r' 之间的平均值。Vector Extrema 对回复中单词词向量的每一个维度提取最大 (小) 值作为回复向量对应维度的数值，然后计算他们的余弦相似度，计算公式如下：

$$\text{Extrema}(r', r) = \cos(e'_r, e_r), \tag{2.38}$$

$$e_{rd} = \begin{cases} \max\limits_{w \in r} e_{wd}, & \text{if } e_{wd} > |\min\limits_{w' \in r} e_{w'd}| \\ \min\limits_{w \in r} e_{wd}, & \text{if } e_{wd} \leq |\min\limits_{w' \in r} e_{w'd}| \end{cases}, \tag{2.39}$$

其中 e_{wd} 是单词词向量 e_w 第 d 维的值，max 表示如果该维度中最大的负值的绝对值大于最大的正值，则应该选择最大的负值。

- 对于多样性，基于规则的指标主要基于 n-gram 计算，包括 Distinct 指标[6]，Shannon Entropy 指标[133]，以及 Self-BLEU 指标[134]。Distinct 指标计算不同的 n-gram 数量占全部 n-gram 数量的比例，用来衡量词级别多样性，n 通常设置为 1 或 2。该指标可以在生成回复内部评估，也可以在多个生成回复之间进行评估。相应地，Distinct 指标可以被细分为 Intra-Distinct 和 Inter-Distinct 指标。计算方式如下：

$$\text{Intra-Distinct} = \frac{\text{count}(\text{distinct}_{i \in R'}(i))}{\text{count}(\text{all}_{i \in R'}(i))}, \tag{2.310}$$

其中 R' 表示生成回复 r' 所包含的所有 n-gram 的集合。类似地，

$$\text{Inter-Distinct} = \frac{\text{count}(\text{distinct}_{i \in R'_N}(i))}{\text{count}(\text{all}_{i \in R'_N}(i))}, \tag{2.311}$$

其中 R'_N 表示针对给定对话上文，生成的 N 条回复中所包含的所有 n-gram 的集合。Entropy 指标同样用作衡量词级别多样性。该指标反映了生成回复的经验 n-gram 分布的均匀程度。指标数值越大，表明生成的回复多样性就好。其计算公式如下：

$$\text{Entropy} = -\frac{1}{\sum_i \text{Freq}(i)} \sum_{i \in R'} \log \frac{\text{Freq}(i)}{\sum_i \text{Freq}(i))}, \tag{2.312}$$

其中 R' 为包含生成回复 r' 中所有 n-gram 的集合，Freq$(\cdot)$ 表示 n-gram i 出现的频率。Self-BLEU 指标计算了给定对话上文的一个生成回复相对于另一个生成回复（而不是参考回复）的 BLEU 分数。越高的 Self-BLEU 表明两个生成回复间越高的相似性和越低的多样性。

- 对于流畅性，常用的指标是困惑度（Perplexity，PPL）。困惑度衡量了概率模型与数据的拟合程度，是生成回复是否合乎语法的有力指标。其计算方式如下：

$$\text{PPL}(r') = \sqrt[l'_r]{\prod_{i=1}^{l'_r} \frac{1}{P\left(r'_i \mid r'_0, \ldots, r'_{i-1}\right)}}, \tag{2.313}$$

其中 l'_r 表示生成回复 r' 的长度，r'_i 表示生成回复的第 i 个单词。

基于模型的评估指标 与基于规则的指标相比，基于模型的指标提高了评估的有效性，但它们需要额外的训练过程。近年来，基于模型的评价指标如雨后春笋般涌现，包括 ADEM[135]、RUBER[136]、BERT-RUBER[137]、PONE[138]、MAUDE[139]、GRADE[140]、PredictiveEngage[141]、FED[142]、FlowScore[48]、DynaEval[143]，以及 IM^2[144]。不同于基于规则的指标，这类指标并非只针对单一角度进行评估，而是通常会同时针对多维评估角度，比如连贯性、流畅性、多样性和一致性等，对生成回复质量进行综合评分。

具体来说，Lowe 等[135] 提出 ADEM 指标，该指标使用循环神经网络（RNN）计算生成回复和参考回复的句子表示，然后计算两者之间的余弦相似度作为评估结果。Tao 等[136] 提出 RUBER 指标，它不依赖于人类判断的分数。RUBER 具体由一个基于参考回复的指标和一个未基于参考回复的指标组成，前者用于测量生成回复与参考回复之间的重叠比例，后者用于测量生成回复与输入对话上文之间的相关性。进一步，Ghazarian 等[137] 提出 BERT-RUBER 指标，其中使用 BERT 取代 RUBER 指标中的 RNN 单元。基于 BERT-RUBER， Lan 等[138] 提出 PONE，它使用增强的正样本和有价值的负样本来训练评分模型。之后， Sinha 等[139] 提出 MAUDE，它使用噪声对比估计（Noise Contrastive Estimation，NCE）方法进行训练，进一步提高指标和人类评估间的一致性。此外 Huang 等[140] 提出 GRADE 指标，该指标通过构建对话历史的图表示来建模对话中的话题动态转换，以评估对话的连贯性。PredictiveEngage 指标额外引入了一个话语级别的吸引力分类器，用来评估对话系统的交互能力。Mehri 等[142] 提出 FED 指标，尝试了使用对话预训练模型 DialoGPT 去评估对话的综合质量。除此之外， Li 等[48] 提出 FlowScore 指标，该指标利用对话预训练模型 DialogueFlow 评估交互式人机对话质量。 Zhang 等[143] 提出 DynaEval 指标，该指标不仅可以进行回合级生成回复的质量评估，还能够全面考虑整个交互对话的质量评估。 Jiang 等[144] 提出一个可解释的、多方面的、可控的评估指标 IM^2，该指标综合了大量针对单维评估角度最优的指标去综合评估回复质量。

2.3.2 人工评估

由于开放域对话具有的开放性和主观性特点，人工评估也更多地用于评估对话系统的好坏[4,53]。人工评估需要聘请评估人员根据预定义的评估指标，以及提供的有据可查的评估说明和评估示例去评估生成回复的质量。评估方式主要分为两种：对单个回复打分的逐点评估（Point-wise Evaluation），以及回复两两比较的成对评估（Pair-wise Evaluation）。

对于逐点评估，所有回复都将彼此独立地由评估人员进行评分。评分准则则由专业人员预定义需要评估的维度，以及相应的评估说明。举例来说，评分准则可以设置如下：

- 流畅性：回复通顺，衔接自然，无明显语法错误；
- 连贯性：回复和对话上文衔接或转换自然，是对话上文的有效延续；

- 多样性：回复内容丰富，表达多样，具有信息量；
- 一致性：回复中的语义内容和对话上文中所有的语义内容都不相互矛盾。

其中，评分范围可以设置为 0 至 2 分，0 分表示质量极差，而 2 分表示质量极好。在得到所有回复在各维度的评分后，最终对话系统在各维度的评分按照以下方式计算：首先对属于相同系统的所有样本回复在特定维度上的分数进行求和平均，然后再对多名评估人员关于该系统在该维度上的评分进行求和平均，最终的分数作为评估结果。值得注意的是，不同的工作通常会根据研究工作的目标与类型而选择不同的评估维度去评估质量。

对于成对评估，待评估系统的回复 r'_s 和各个基线系统的回复 r'_b 会两两组成一个比较对 $< r'_s, r'_b >$。和逐点评估相似，评估人员依然根据预定义的评估维度（例如流畅性、连贯性、多样性和一致性等）从多角度去比较哪个回复质量更好。与逐点评估相比不同的是，评估结果不再是具体分数，而是“优于（r'_s 更好）”、“持平（两者一样好）”和“劣于（r'_b 更好）”。最终采用类似的计算方式去得到对话系统在各维度上的评分，区别在于成对评估使用结果为“优于”、“持平”和劣于的样本的占比作为评估结果。

人工评估存在的问题是需要雇佣足够多的评估人员。目前主要采用外包模式尝试解决这一问题，即利用 Amazon Mechanical Turk（AMT）[1]平台雇佣有偿评估人员或无偿志愿者进行评估。相关组织者报告了有偿人员和无偿志愿者之间的评分差异：无偿志愿者的评估中好的（即长时交互和保持一致约束的）对话更少，而有偿人员往往对系统的评价高于无偿志愿者。此外，由于人工评估涉及多位评估人员，因此还需要使用 Fleiss 的 kappa κ 指数[145] 来计算不同人员之间的评估结果的一致性水平。

2.4　小结

本章首先介绍了目前开放域对话系统常用的两种构建方法：检索式方法和生成式方法。检索式方法的输出源自真实的人类对话，因此回复流畅且易于理解。但它受到数据集规模和质量的限制，有时检索到的回复与对话上文相关性较弱。生成式方法能够自由地生成训练数据中不存在的回复，更符合人类生成回复的过程，它不受限于一组预定义的话语。但生成的回复不能保证一定正确，有时回复缺乏和对话上下文的连贯性，并且往往是通用且无趣的回复。进一步，本章介绍了构建更好的生成式开放域系统所做的相关工作，分别包括从数据层面和从模型层面做的系列研究工作。最后，本章介绍了提升对话系统过程中最重要的一个步骤：评估对话系统；总结了主要的评估方法：自动评估和人工评估，又进一步介绍了相应地评估指标。总的来说，现有开放域对话系统相关的研究工作在训练数据准备、对话结构建模、额外约束满足方面都还存在一定的局限性。本文将在第三、四、五章中对这些问题进行详细的讨论，并给出针对性的解决方案。

[1]https://www.mturk.com/

第 3 章　基于语义建模的对话回复生成方法

3.1　引言

在人类对话中，不同的人类会产生不同语义的回复，因为他们通常会关注到不同的语义侧面。因此，构建对话系统时建模输入对话上文与多个输出回复间的复杂语义映射关系是至关重要的。复杂语义映射关系体现在层级映射的特点上，即一个对话上文首先可以映射到多个不同可生成回复的语义侧面，其次每个语义侧面又进一步映射到多个不同表达的回复。现有系统通过引入隐变量来增强模型的建模能力：隐变量在监督训练中隐式地捕捉回复中蕴含的抽象共性信息，可能是说话语气或用词风格等。但这些系统没有建模对话固有的层级语义映射关系，让其自行从对话数据中学习非常困难，这导致隐变量更容易学到混合的语义表示。最终尽管采样不同的隐变量值，系统也只能生成语义相似但表达不同的回复。因此，本章提出一种基于语义建模的对话回复生成方法，用于显式建模层级语义映射关系，以生成语义多样的回复。具体来说，该方法首先利用多语义检索模块用于保证相同对话上文具有不同语义的回复，这不仅显式地定义了不同语义侧面，并且给模型学习不同语义侧面的信息提供了监督信号。其次，该方法改进了 Wasserstein 自编码器用于保证将不同语义的回复与相应的隐变量对齐。此外，引入了额外的语义距离损失函数拉开不同隐变量间的距离，用于保证模型能够生成不同语义侧面的回复。实验结果表明本章提出的方法能够显著超过基线模型，并且能够生成更多语义多样的回复。同时分析实验表明该方法能够有效地建模开放域对话的层级语义映射关系。

3.2　概述

开放域对话允许对话上文有各种各样不同语义的回复[6]。从人类对话的角度来看，开放域对话的对话上文与多个输出回复间具有层级语义映射的特点：每个回复都谈论对话上文的一个语义侧面，而谈论相同语义侧面的回复语义相似但用词表达多样[101,146,147]。图3-1给出了对话语义层级映射关系的示例说明，给定一个对话上文，它存在多个可生成回复的语义侧面，例如给定对话上文中的“吸烟”，“有害”，“健康”，针对这些不同的语义侧面，又可以对应不同表达的回复。这些回复的语义涵盖范围广并且语义不重复。由此可见，建模这种复杂的层级语义映射关系是至关重要的，即首先将一个对话上文映射到多个不同可生成回复的语义侧面，再从每个语义侧面进一步映射到多个不同表达的回复上。

现有系统主要通过引入隐变量来增强模型的建模能力[88,100,102,148]：隐变量在监督训练中隐式地捕捉回复中蕴含的抽象共性信息，比如说话语气[104]或用词风格[22,103]等。早期，研究者们主要利用单隐变量来增强对话模型的中间表示能力[38,100,104,149]，通过学习隐变量的分布，最终从分布上采样不同的值即可去

表 3-1 对话层级语义映射关系示例

Table 3-1 An example showing hierarchical semantic mapping of dialogues

对话上文：吸烟有害健康。
回复 1-1：对于戒烟你有什么好的建议吗？ 回复 1-2：我想戒烟，但是一直不成功。 回复 1-3：戒烟真的会带来长期的健康益处么？
回复 2-1：看看这篇文章，它说吸烟会导致肺癌。 回复 2-2：吸烟也会影响人体脑部神经，也就是杀死“脑细胞”。 回复 2-3：的确如此，我最好的朋友因为长期吸烟早早去世了，我好想他。
回复 3-1：我定期体检，身体没有任何疾病。 回复 3-2：我爷爷快 100 岁了，每天至少 2 根烟，身体依然非常棒！ 回复 3-3：烟让我在焦虑的生活中有了片刻安宁，没有它我早就死了。

生成不同的回复。但单隐变量机制过于简单，不具备建模层级语义映射关系的能力。随后，一些研究工作开始关注多隐变量机制[102,146,147]，认为多隐变量机制能够应对真实回复空间复杂多样的问题。但这些工作并没有建模对话固有的层级语义映射关系，而让模型自行从对话数据中学习层级结构是非常困难的，这导致隐变量更容易学到混合的语义表示。最终尽管采样不同的隐变量值，系统也倾向于生成语义相似但表达不同的回复。

因此，本章显式地建模开放域对话的层级语义映射关系，去更好地理解对话上下文内容，从而生成更加语义多样的回复。本研究沿袭了目前主流的建模对话映射关系的思路，即引入多隐变量机制。目前，多隐变量机制通常采用混合高斯模型来建模。若要达成预设的目标，只需要每个隐变量代表不同的语义值，最终采样不同的隐变量即可生成不同语义的回复。这就需要在训练模型时，将同一个对话上文的不同语义的回复分别对应到不同的隐变量上。然而，已有的工作在针对每个回复选择对应的隐变量时，仅考虑了对话上文的信息，这就导致模型没有办法将不同语义对应到正确的隐变量上。最终，每个隐变量都混合了不同语义的信息，没有学到语义可区分的隐变量。并且可生成回复的语义侧面是不可知的，并且即使语义侧面已知，受限于训练数据中可观测回复所对应语义侧面的数量，模型也难以学到所有的语义侧面表示。总的来说，为了生成语义多样的回复，需要解决三个挑战：（1）保证可生成不同回复的语义侧面信息可知并能提供监督信号；（2）保证不同语义侧面能与不同的隐变量一一对齐；（3）保证模型能够生成不同语义角度的回复。

为此，本章提出了一种基于语义建模的对话回复生成方法，用于显式建模开放域对话的层级语义映射关系。首先，为了得到关于可生成回复的语义侧面的信息和监督信号，本研究提出多语义回复检索模块，用 BM25 算法选择多个回复，

然后使用聚类算法保证回复的语义互不相同。进一步本研究提出改进 Wasserstein 自编码器用于确保混合高斯建模多隐变量时，每个隐变量都能代表不同的语义侧面，然后从不同隐变量上采样去得到语义不同的回复。最后为了让模型能够通过采样不同的隐变量就能生成不同语义的回复，则需要保证代表各个隐变量的分布尽可能区分开，减少重叠。因此本研究引入额外的语义距离损失，使得分布之间的距离尽可能大。实验结果表明，本研究提出的方法在保证回复相关性的同时，极大得提高了回复的语义多样性。同时分析实验表明该方法能够有效地建模开放域对话的层级语义映射关系。

本章组织如下，第 3.3节中介绍提升对话回复多样性的相关工作；第 3.4节中介绍回复生成任务的定义，以及本节工作的基础模型 DialogWAE；第 3.5节中介绍了本研究提出的多语义 Wasserstein 自编码器模型用于生成语义多样的回复；第 3.6节中介绍了本工作相关的实验设置，包括使用的数据集情况，对比的基线模型，使用的评价指标，和实验细节等；第 3.7节中给出了本工作的实验结果，并进行了相关对比与分析。最后对本章进行了总结。

3.3　相关工作

提升多样性　生成通用回复是开放域对话生成中的一个基本问题。为了缓解这个问题，已经从不同的角度提出了许多方法来提升回复生成的多样性。一类工作认为针对给定输入的输出可能性进行优化，神经模型将高概率分配给“安全”响应。因此，一些工作从改进训练目标的角度来解决这一问题，比如使用最大互信息作为训练目标[6]，引入特定需求场景的最大生成可能性和多样化需求场景的条件风险价值作为训练目标[7]。还有一类工作从解码策略角度来考虑这个问题，认为采用贪心解码策略只能选择概率最大的词输出，不利于输出多样的回复。因此，回复生成中更多尝试集束搜索解码策略。然而，概率靠前的 K 个词语通常非常相似，集束搜索带来的多样性提升就显得非常有限。随后，他们关注如何鼓励集束搜索产生多样性更高的输出，包括增加相似度约束[150]，增加相似度惩罚项[151]，惩罚已经生成过的词语[152]，引入注意力机制重排序[153]。还有一些工作通过引入额外的信息来提升多样性，认为更丰富的输入信息能更好的建模输入和输出间的映射关系，缓解一对多的映射问题。这一类工作主要包括引入话题[13]，引入情感倾向[9]，引入知识图谱[154,155] 等。后来，研究者们认为建模对话复杂映射关系有助于更好地充分理解对话上下文内容，从而自然而然就能够去提升回复多样性。因此，后来引入隐变量机制成为一种常用的提升多样性的手段，通过学习隐变量的分布，最终从分布上采样不同的值即可去生成不同的回复[38,100,104,149]。还有一些研究者认为单隐变量过于简单，进而提出多隐变量机制[102]。尽管这些工作建模了对话输入和输出间的映射关系，但忽略了建模对话固有的层级语义映射关系。相反，本研究关注于显式建模层级语义映射关系，去生成语义多样的回复。

多隐变量机制 本章工作借鉴了目前提升多样性的常用思路，引入多隐变量机制，包括离散隐变量[101]和连续隐变量[102]，用多个隐变量建模复杂的回复空间去生成多样的回复。早期工作[146,147]则主要利用类似的思想，多映射机制来关注对话上文的不同的语义部分并产生不同的回复。已有工作在训练过程中，仅仅利用对话上文的信息为不同回复选择需要对齐的隐变量。然而对于相同对话上文的多个回复而言，仅利用对话上文不足以将不同的回复映射到不同的隐变量上。这样造成的问题就是没有办法学到语义可区分的隐变量分布。最终在采样不同隐变量时，就难以生成语义不同的回复。相反，本研究在为每个回复选择隐变量对齐时，直接根据回复所属的聚类编号映射到有相同编号的隐变量上，实现了不同回复对应到不同隐变量上这一目标。同时，在训练过程中引入了语义距离损失函数，使得不同隐变量间的距离尽可能大。这样保证了最终采样不同隐变量时，能够更好地生成不同语义的回复。

增大语义距离 和本研究工作比较相似的工作是 Gao 等[101]的工作，该工作从不同语义类别中的关键词集合中抽取关键词，以保持隐变量的语义距离。不同的是，本研究工作将相同对话上文的不同回复与不同的隐变量一一对齐，然后最大化代表不同语义的隐变量间的距离，以此来提升语义多样性。

3.4 背景知识

任务定义 对话系统应该根据给定的上下文生成相关的、信息丰富的和多样化的对话回复。形式化来说，对于单轮对话，给定输入 $\boldsymbol{x}=\{\boldsymbol{x}_i\}_{i=1}^{N}$，对话系统学习建模回复 $\boldsymbol{y}^i$ 的概率分布 $\mathcal{P}_\phi(\boldsymbol{y}^i|\boldsymbol{x}^i)$。模型参数 ϕ 可以通过最小化以下损失来学习：

$$\mathcal{L}=-\sum_{i=1}^{N}\log P_\phi(\boldsymbol{y}^i|\boldsymbol{x}^i). \tag{3.41}$$

条件变分自编码器 由于 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 是离散文本的序列，建模输入和输出间的直接映射关系并非易事。相反，额外引入了一个连续的隐变量 $\boldsymbol{z}$ 作为中间表示则能够简化映射关系建模。因此，回复生成可以看作是一个两步骤过程，首先从隐空间 $\mathcal{Z}$ 上的分布 $P_\theta(\boldsymbol{z}|\boldsymbol{x})$ 中采样隐变量 $\boldsymbol{z}$，然后使用 $P_\phi(\boldsymbol{y}|\boldsymbol{z})$ 根据 $\boldsymbol{z}$ 解码回复 $\boldsymbol{y}$。基于该模型，回复生成的概率为

$$P_\theta(\boldsymbol{y}|\boldsymbol{x})=\int_{\boldsymbol{z}}P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})P(\boldsymbol{z}|\boldsymbol{x})d_{\boldsymbol{z}}. \tag{3.42}$$

然而，精确的对数概率很难计算，因为很难计算 $\boldsymbol{z}$ 的边缘分布。为此，可以将 $\boldsymbol{z}$ 的后验分布近似为 $Q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$，然后通过后验神经网络进行估计。使用该近似后验，回复生成的概率可以转化成计算其证据下界（Evidence Lower Bound，ELBO）：

$$\begin{aligned}\log P_\theta(\boldsymbol{y}|\boldsymbol{x})&=\log\int_{\boldsymbol{z}}P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})P(\boldsymbol{z}|\boldsymbol{x})d_{\boldsymbol{z}}\\&\geq\mathcal{L}(\boldsymbol{x},\boldsymbol{y})=\mathbf{E}_{\boldsymbol{z}\sim Q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}[\log P_\psi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})-\mathrm{KL}(Q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})||P(\boldsymbol{z}|\boldsymbol{x})],\end{aligned} \tag{3.43}$$

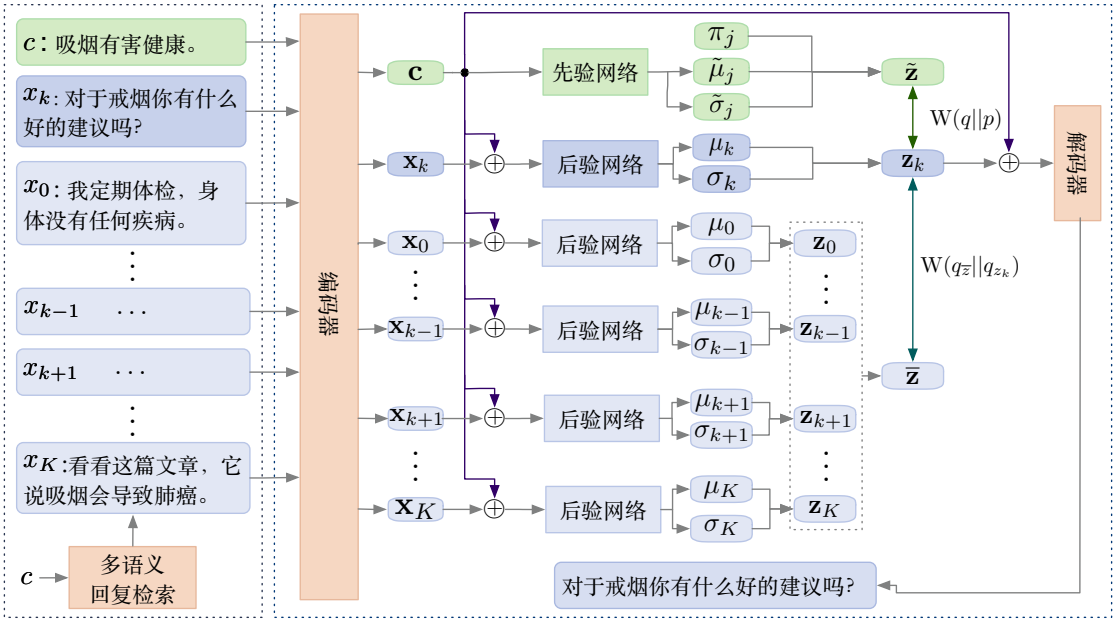


图 3-1 MS-WAE 模型架构图

Figure 3-1 The overall architecture of MS-WAE

其中，$P(\boldsymbol{z}|\boldsymbol{x})$ 表示给定 $\boldsymbol{x}$ 的 $\boldsymbol{z}$ 的先验分布，它可以用先验神经网络来建模。

传统的条件变分自编码器（Conditional Variational Auto-Encoder, CVAE）对话模型假设隐变量 $\boldsymbol{z}$ 服从简单的先验分布，例如正态分布。然而，真实回复的潜在空间非常复杂，难以用这种简单的分布进行估计。此外，传统 CVAE 模型采用 KL 散度逼近先验和后验分布的距离。然而，当先验分布和后验分布不重叠时，KL 散度就会变得无限大，使得 KL 项约束非常强，即后验分布 $Q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$ 会退化成先验一样 $P_\theta(\boldsymbol{z}|\boldsymbol{x})$ 的正态分布。这就会导致“后验坍塌[1]”问题[102]。

对话 Wasserstein 自编码器　相比传统 CVAE 模型，对话 Wasserstein 自编码器（Dialogue Wasserstein Auto-Encoder，DialogWAE）采用 Wasserstein 距离来逼近先验分布和后验分布的距离。这个距离在两个分布不重叠的时候也是连续的。因此，能够很好地解决后验坍塌的问题。并且 CVAE 希望每个 $Q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$ 都和 $P_{\boldsymbol{z}}$ 匹配，但 DialogWAE 是把 $Q_{\boldsymbol{z}}$ 和 $P_{\boldsymbol{z}}$ 匹配，相当于放开了对 $Q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$ 的约束，一定程度上也有助于缓解后验坍塌的问题。此外，为了应对真实回复空间复杂多样的问题，DialogWAE 引入多隐变量来显式建模更加复杂的潜在空间。具体来说，使用混合高斯模型来建模多个隐变量。

DialogWAE 通过在潜在空间内训练 GAN 来对隐变量的分布进行建模，然后使用神经网络对先验分布和后验分布的隐变量进行采样。具体来说，先验采样

[1]在损失计算中，KL 散度项趋近于 0，之后 CVAE 网络的 Decoder 就会忽视 Encoder 产生的后验分布而只从先验分布中采样，从而使 CVAE 网络失效，这就出现了后验坍塌的问题。造成这种问题的一个原因是：根据对话上文生成回复的条件变分自编码器 (CVAE) 结构会导致严重的数据稀疏性。即使有一个大规模的训练语料库，当以对话上文为条件时，也只存在很少的目标回复。

$\tilde{\boldsymbol{z}} \sim P_\theta(\boldsymbol{z}|\boldsymbol{x})$ 由一个先验网络生成，即

$$P(\tilde{\boldsymbol{z}}|\boldsymbol{x}) = \sum_{k=1}^{K} \upsilon_k \mathcal{N}(\tilde{\boldsymbol{z}}|\tilde{\mu}_k, \tilde{\sigma}_k^2 \boldsymbol{I}), \tag{3.44}$$

$$\begin{bmatrix} \tilde{e}_k \\ \tilde{\mu}_k \\ \tilde{\sigma}_k \end{bmatrix} = \tilde{\mathbf{W}}_k f_\theta(\boldsymbol{x}) + \tilde{b}_k. \tag{3.45}$$

$\boldsymbol{x}$ 代表对话上下文，$\boldsymbol{y}$ 代表对话回复。υ_k 表示子分布指示器，通过使用 Gumbel-Softmax 重参数化技术采样对 υ 的一个实例来计算：

$$\upsilon_k = \frac{\exp((\tilde{e}_k + \tilde{g}_k))/\tau)}{\sum_{i=1}^{K} \exp((\tilde{e}_i + \tilde{g}_i)/\tau)}, \tag{3.46}$$

其中 $\tilde{g}_i$ 表示 Gumbel 噪声。后验样本 $\boldsymbol{z} \sim Q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ 由后验网络 g_ϕ 生成，即：

$$\begin{bmatrix} \mu \\ \sigma \end{bmatrix} = \mathbf{W} g_\phi([\boldsymbol{x}, \boldsymbol{y}]) + b, \tag{3.47}$$

$$\mathbf{z} = \mu + \sigma * \epsilon, \epsilon \sim \mathcal{N}(0, \boldsymbol{I}). \tag{3.48}$$

对话 Wasserstein 自编码器的训练目标类似于 CVAE 的训练目标，即最小化先验和后验之间的差异，并最大化 $\boldsymbol{z}$ 重构回复的对数概率，如下所示：

$$\mathcal{L}(\mathbf{x}, \mathbf{c}) = -\mathrm{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{c})} \log p_\psi(\mathbf{x}|\mathbf{z}, \mathbf{c}) + \mathrm{W}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{c})||p_\theta(\mathbf{z}|\mathbf{c})). \tag{3.49}$$

3.5 多语义 Wasserstein 自编码器

本节将展示如何显式建模对话上文和多个回复间的层级语义映射关系，从而生成语义多样的对话回复。首先，为了得到关于可生成回复的语义侧面的信息和监督信号，本研究设计了多语义回复检索模块（Muti-Semantic Response Retrieval，MSRR）为每个对话上文扩增多个不同语义的回复，每个回复都代表不同的语义侧面。其次，本研究改进了对话 Wasserstein 自编码器（DialogWAE）并提出了多语义 Wasserstein 自编码器（Multi-Semantic Wasserstein Auto-Encoder，MS-WAE），通过将每个对话回复与不同的隐变量对齐，并且控制隐变量间的语义距离来实现生成语义不同的回复。图 3-1 和算法 1 展示了多语义 Wasserstein 自编码器的整体架构和详细算法流程。

3.5.1 多语义回复检索

首先通过图 3-2 中的示例和算法 1 中的回复检索过程（行 1-3）来说明如何利用已有训练数据为每个对话上文扩增多个不同语义的回复。该过程包括以下三个步骤：1) **词级别粗略检索：**为对话上文扩增候选回复。2) **语义级别过滤：**过

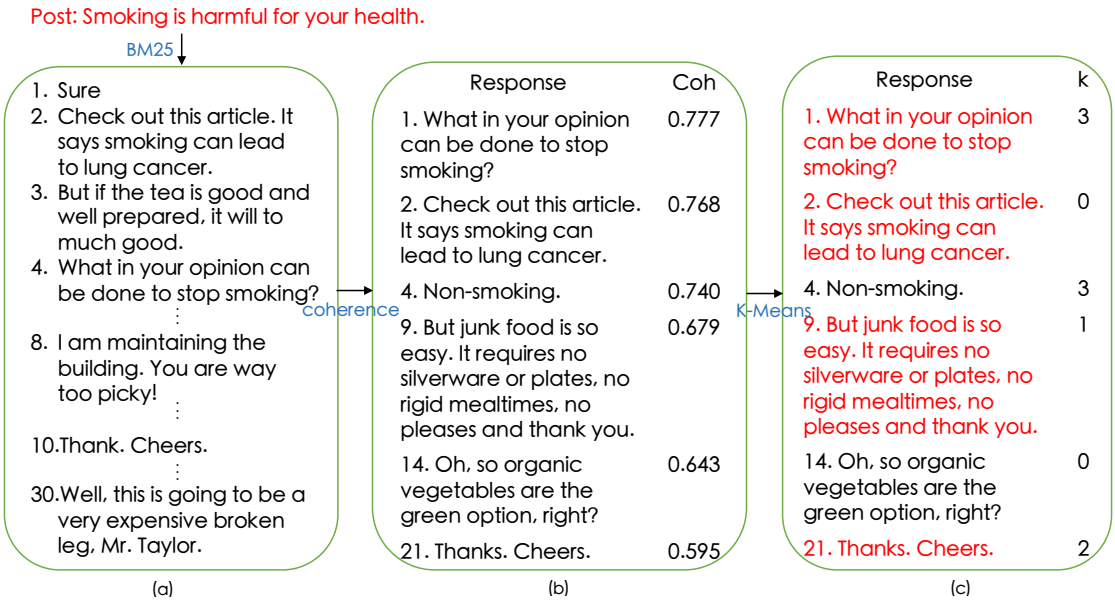


图 3-2 多语义回复检索示例

Figure 3-2 An example showing multi-semantic response retrieving

滤与对话上文不够连贯的候选回复。3) **回复聚类：**选择具有不同语义且高质量的回复。

在词级别粗略检索这一步骤，首先对于给定的对话上文，本研究通过 BM25 算法检索到与原始对话上文相似的对话上文，把这些对话上文的回复作为当前对话上文的候选回复集合（算法 1中的行 2-2）。基于的假设是，相似对话上文的回复更有可能相互共享。如图 3-2 中的 *(a)* 部分所示，在这一步中，获得了 $M = 30$ 个候选回复。然而，该候选集合存在噪音回复，即与对话上文缺乏语义连贯性，如图 3-2 中 *(a)* 部分的回复 3 和 8 所示。在语义级别过滤这一步骤，本研究使用余弦相似度来表示候选回复与对话上文间的连贯性。因此，本研究利用相似度对候选回复进行重排序，并且过滤得分低于阈值 $\tau = 0.1$ 的回复（算法 1 中的行 2）。具体来说，在计算余弦相似度时，该方法使用预训练词嵌入向量的加权平均值作为句子表示，其中每个词的权重是对应的 TF-IDF 值。中间结果如图 3-2*(b)* 所示。为了进一步选择具有不同语义的回复，本研究引入了回复聚类这一步骤。具体来说，进一步使用 K-Means 聚类算法[156] 对候选集中的回复进行聚类，最终分别选择每一类中连贯性最高的回复组成最终的多语义对话数据（算法 1 中的行 3-3）。

最终该模块扩展了原始训练数据，使得每个对话上文拥有多个语义侧面的回复。如图 3-2*(c)* 所示，标红的回复构成了后续用于训练的最终回复集合。

3.5.2　回复生成模型结构

图 3-1 的右侧为本研究模型部分的结构框架，同时算法 1中第 4- 13 行描述了本节模型的训练过程。本节模型相比于对话 Wasserstein 模型而言，进行了两

算法 1 多语义 Wasserstein 自编码器

Input: 训练集 $D = \{< \boldsymbol{x}, \boldsymbol{y} >\}^{|D|}$, M 为待检索的候选回复数量, τ 为连贯性过滤阈值, K 为聚类类别数, n_{critic} 为判别器每轮的训练次数

$\tilde{D} = \{\}$

for $\underline{\boldsymbol{x} \in D}$ **do**

  检索 M 个 $\boldsymbol{x}$ 相关的对话上文集合 P

  $R = \{$P 中所有对话上文对应的回复$\}$

  $\tilde{R} = \{$对于所有的$\boldsymbol{r} \in R, s.t.\text{coherence}(\boldsymbol{r}, \boldsymbol{x}) \geq \tau\}$

  将 $\boldsymbol{r} \in \tilde{R}$ 聚成 K 类

  $\hat{R} = \{$对于所有的$\boldsymbol{r} \in \text{Cluster}_k, s.t. \forall \boldsymbol{r}_j \in \text{Cluster}_k, \text{coherence}(\boldsymbol{r}_j, \boldsymbol{x}) \leq \text{coherence}(\boldsymbol{r}, \boldsymbol{x})\} \cup \{\boldsymbol{y}\}$

  $\tilde{D} = \tilde{D} \cup < \boldsymbol{x}, \hat{R} >$

Initialize $\{\theta_{\text{Enc}}, \theta_{\text{Post-net}}, \theta_{\text{Prior-net}}, \theta_{\text{Disc}}, \theta_{\text{Dec}}\}$

for $\underline{t < \text{max-step}}$ **do**

  从 $\tilde{D}$ 中采样 N 个样本 $\{< c, \hat{R} >\}^{|N|}$

  for $\underline{\boldsymbol{y}_k \in \hat{R}}$ **do**

    $\boldsymbol{x} = \text{Enc}(\boldsymbol{x}), \boldsymbol{y}_k = \text{Enc}(\boldsymbol{y}_k)$

    从后验网络 Post-net$(\boldsymbol{y}_k, \boldsymbol{x})$ 采样 $\boldsymbol{z}_k$

    从先验网络 Prior-net$(\boldsymbol{x}, K+1, k)$ 中采样 $\tilde{\boldsymbol{z}}$

    $\theta_{\text{Prior-net}} = \theta_{\text{Prior-net}} - lr * \frac{\partial}{\partial \theta_{\text{Prior-net}}} \mathcal{L}_{disc}$

    $\theta_{\text{Post-net}} = \theta_{\text{Post-net}} + lr * \frac{\partial}{\partial \theta_{\text{Post-net}}} \mathcal{L}_{disc}$

    $\theta_{\text{net}} = \theta_{\text{net}} - lr * \frac{\partial}{\partial \theta_{\text{net}}} \mathcal{L}_{rec}, \ s.t. \text{ net} \in \{\text{Enc, Prior-net, Post-net, Dec}\}$

    $\theta_{\text{Disc}} = \theta_{\text{Disc}} - lr * \frac{\partial}{\partial \theta_{\text{Disc}}} \mathcal{L}_{disc}$

  for $\underline{\mathbf{z}_k, k \in \{0, \ldots, K\}}$ **do**

    $\theta_{\text{Post-net}} = \theta_{\text{Post-net}} - lr * \frac{\partial}{\partial \theta_{\text{Post-net}}} \mathcal{L}_{sd}$

  for $\underline{i < n_{\text{critic}}}$ **do**

    Repeat 5

    for $\underline{x_k \in \hat{R}}$ **do**

      Repeat 6-6

      $\theta_{\text{Disc}} = \theta_{\text{Disc}} - lr * \frac{\partial}{\partial \theta_{\text{Disc}}} \mathcal{L}_{disc}$

点改进，包括先验网络和后验网络的改进，以及训练损失函数的改进。接下来，本小节将着重介绍改进部分的细节。

先验及后验网络　后验采样 $\boldsymbol{z}_k \sim Q_\phi(\boldsymbol{z}_k|\boldsymbol{y}_k, \boldsymbol{x})$ 主要由后验网络产生，即

$$\begin{bmatrix} \mu_k \\ \sigma_k \end{bmatrix} = \mathbf{W} g_\phi([\boldsymbol{y}_k, \boldsymbol{x}]) + b, \tag{3.51}$$

$$\boldsymbol{z}_k = \mu_k + \sigma_k * \epsilon, \epsilon \sim \mathcal{N}(0, \boldsymbol{I}), \tag{3.52}$$

其中 ϕ、$\mathbf{W}$ 和 b 是可训练参数，$g_\phi(\cdot)$ 是前馈神经网络。$\boldsymbol{y}_k, \boldsymbol{x}$ 分别表示输出对话回复和输入对话上文，它们都是由编码器（一个 Bi-GRU 模块）生成的固定长度向量。请注意，$k \in \{0, \ldots, K\}$ 是所选目标回复的索引，$\boldsymbol{y}_k$ 可以是参考回复，也可以是多语义回复检索模块得到的回复。

为了将不同回复代表的语义侧面分别和不同的隐变量对齐，本研究受 Gu 等[102] 启发采用高斯混合模型（Gaussian Mixed Model，GMM）作为先验分布。具体来说，先验采样 $\tilde{\boldsymbol{z}} \sim P_\theta(\tilde{\boldsymbol{z}}|\boldsymbol{x})$ 是通过两阶段采样生成的。第一阶段，根据 π_j 选择一个子分布，π_j 是第 j 个子分布的指示符。不同于 Gu 等[102] 的工作，本研究显式地根据当前选定的回复所属的聚类编号 k 去选择 GMM 的第 j-th 子分布，$j \in \{0, \ldots, K\}$(算法 1中行 3- 3)，具体选择方法如下：

$$\pi_j = \begin{cases} 1, & \text{if } \boldsymbol{y}_k \text{ belongs to cluster } j \\ 0, & \text{otherwise} \end{cases}. \tag{3.53}$$

第二阶段，从子分布中得到先验采样 $\tilde{\boldsymbol{z}}$:

$$p(\tilde{\boldsymbol{z}}|\boldsymbol{x}) = \sum_{j=0}^{K} \pi_j \mathcal{N}(\tilde{\boldsymbol{z}}|\tilde{\mu}_j, \tilde{\sigma}_j^2 \boldsymbol{I}), \tag{3.54}$$

$$\begin{bmatrix} \tilde{\mu}_j \\ \tilde{\sigma}_j \end{bmatrix} = \tilde{\mathbf{W}}_j f_\theta(\boldsymbol{x}) + \tilde{b}_j, \tag{3.55}$$

$$\tilde{\boldsymbol{z}} = \tilde{\mu}_j + \tilde{\sigma}_j * \tilde{\epsilon}, \tilde{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}), \tag{3.56}$$

其中 $f_\theta(\cdot)$ 是一个前馈神经网络。θ、$\tilde{\mathbf{W}}_j$ 和 $\tilde{b}_j$ 是可训练的参数。与 DialogWAE 相比，MS-WAE 自编码器框架通过在第一阶段采样时提供显式信号来准确选择先验子分布以匹配指定的后验分布，该信号能辅助建模可区分的多语义分布。

训练目标函数　在训练阶段，MS-WAE 从 $Q_\phi(\boldsymbol{z}_k|\boldsymbol{y}_k, \boldsymbol{x})$ 中采样隐变量 $\boldsymbol{z}_k$，然后解码器重构 $\boldsymbol{y}_k$。重构损失定义为：

$$\mathcal{L}_{rec} = -\mathrm{E}_{\boldsymbol{z}_k \sim Q_\phi(\boldsymbol{z}_k|\boldsymbol{y}_k, \boldsymbol{x})} \log P_\psi(\boldsymbol{y}_k|\boldsymbol{x}, \boldsymbol{z}_k). \tag{3.57}$$

为了拉进后验分布和先验分布的距离，MS-WAE 使用对抗式判别器 D[157] 去优化 Wasserstein 距离。判别器损失函数为

$$\mathcal{L}_{disc} = \mathrm{E}_{z_k \sim Q_\phi(z_k|\boldsymbol{x},\boldsymbol{y}_k)}\left[D(\boldsymbol{z}_k, \boldsymbol{x})\right] - \mathrm{E}_{\tilde{\boldsymbol{z}} \sim P_\theta(\tilde{\boldsymbol{z}}|\boldsymbol{x})}\left[D(\tilde{\boldsymbol{z}}, \boldsymbol{x})\right], \tag{3.58}$$

其中 D 为前馈神经网络。

进一步 MS-WAE 引入语义距离损失函数，用来控制给定对话上文 $\boldsymbol{x}$ 的每个对话回复对应的后验分布之间的语义距离。这样主要是为了减少代表不同语义侧面的子分布间的重叠。具体来说，选定 $\boldsymbol{z}_k$，MS-WAE 需要最大化 $\boldsymbol{z}_k$ 和其他 $\boldsymbol{z}_i, \{0, 1, \cdots, K\} \setminus \{k\}$ 间的距离。但为了简化计算，可以将问题转换成：最大化 $\boldsymbol{z}_k$ 和其他 $\boldsymbol{z}_i$ 的均值（即 $\overline{\boldsymbol{z}}$）的距离。具体来说，使用最大平均差异（MMD）[158] 最大化他们之间的距离，即：

$$\overline{\boldsymbol{z}} = \frac{1}{K}\sum_{i \in \mathbb{I}} \boldsymbol{z}_i, \mathbb{I} = \{0, 1, \cdots, K\} \setminus \{k\}, \tag{3.59}$$

$$\mathcal{L}_{sd} = \mathrm{E}_{\boldsymbol{z}_k, \boldsymbol{z}_k}[\mathrm{GKF}(\boldsymbol{z}_k, \boldsymbol{z}_k)] - 2\,\mathrm{E}_{\boldsymbol{z}_k, \overline{\boldsymbol{z}}}[\mathrm{GKF}(\boldsymbol{z}_k, \overline{\boldsymbol{z}})] + \mathrm{E}_{\overline{\boldsymbol{z}}, \overline{\boldsymbol{z}}}[\mathrm{GKF}(\overline{\boldsymbol{z}}, \overline{\boldsymbol{z}})], \tag{3.510}$$

其中，GFK 为高斯核函数。

总的来说，模型的训练目标函数包括：

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{disc} + \mathcal{L}_{sd}, \tag{3.511}$$

主要采用端对端的方式去优化模型参数。

在推理阶段，给定一个对话上文，模型先从先验分布中随机选择一个子分布，再从子分布上随机采样一个隐变量 $\tilde{\boldsymbol{z}}$。随后，解码器以 $\tilde{\boldsymbol{z}}$ 和对话上文 $\boldsymbol{x}$ 作为输入，去生成回复。此时，从不同子分布上采样的隐变量就具备不同的语义，从而生成的回复就能很好地保证语义多样性。

3.6 实验设置

本小节主要描述实验相关的设置，包括数据集、实验细节和对比方法，以及评估指标。

3.6.1 数据集介绍

本研究在两个广泛使用的公开对话数据集 Douban[80] 和 DailyDialog[159] 上评估了提出的方法。根据任务定义，本研究将数据处理成一种对话上文-回复对的形式。表 3-2 总结了这两个数据集的统计数据。

3.6.2 实现细节

多语义回复检索 多语义回复检索模块中的 M 和 α 分别为 30 和 0.1。K-means 中的 k 通过在验证集上试验，被确定为 4。

表 3-2 数据集统计数据

Table 3-2 Statistics of the datasets

数据集	#train	#valid	#test
Douban	894,721	15,000	15,000
DailyDialog	68,096	6,895	6,695

模型参数　Douban 和 DailyDialog 对话数据集中的词汇量分别为 20,000 和 10,000。MS-WAE 分别使用预训练的 200d Douban 词潜入向量和 GloVe 词嵌入向量[160]来初始化词嵌入矩阵。先验和后验网络分别使用两层的具有 tanh 非线性的前馈神经网络计算 μ, σ。用来采样 $\boldsymbol{z}$ 的非线性变换，以及判别器 D 的非线性变换分别是具有 ReLU 的三层前馈神经网络，其维度大小分别为 200 和 400。此外，隐变量的维度大小为 200。所有全连接层的初始权重均从均匀分布 $[-0.02, 0.02]$ 中采样得到。

模型训练　本研究在训练判别器时使用梯度惩罚，并将 λ 设置为 10。对话话语长度限制为 40，每一批数据的大小为 32。MS-WAE 在 AE 阶段使用初始学习率为 1.0 和梯度裁剪 1.0 的随机梯度下降算法（Stochastic Gradient Descent，SGD）训练生成模型，并由固定学习率为 $5 * 10^{-5}$ 和 $1 * 10^{-5}$ 的 RMSprop（Root Mean Square Propagation）进行训练 WGAN 阶段的生成器和判别器。本研究利用 PyTorch 框架来实现 MS-WAE，并使用 TITAN Xp 完成所有实验。MS-WAE 有 18,259,408 参数。MS-WAE 使用一张 GPU 卡在 Daily Dialog Corpus 上训练需要 150 分钟，在 Douban 语料上训练需要 1081 分钟。

3.6.3　对比方法

为了验证本方法的有效性，本研究提出的方法和以下相关方法进行了对比：

- **Seq2Seq-attn**[82]：具有注意力机制的标准 Seq2Seq 架构；
- **DCVAE**[101]：基于 CVAE 结构的带有离散隐变量机制的方法，通过优化离散隐变量间的语义距离来提升多样性。其中聚类数量 $K = 10$；
- **MMPMS** [147]：具有多重映射机制的 Seq2seq 模型，可以通过建模对话上文与回复间的语义映射来提升多样性。映射模块的数量为 5；
- **DialogWAE**[102]：使用混合高斯模型的基于 WAE 结构的方法，混合高斯模型的子分布数量为 5。

3.6.4　评价指标

自动评估　本研究跟随 DialogWAE[102] 使用了以下自动评价指标：

- **BLEU**[129]：分别报告了 BLEU 的准确率（Precision）、召回率（Recall）和 F1 值，用于衡量生成的回复与给定的参考回复间 n-grams 的重叠比率；

- **BOW Embedding**[161]：分别报告了 Embedding Average、Vector Extrema 和 Greedy Matching 三个指标，用来衡量生成的回复与给定的参考回复间的文本相似性。其中，Embedding Average 将回复中每个单词的词向量求平均来作为回复的特征，计算生成的回复和参考回复的特征的余弦相似度。Vector Extrema 对回复中单词词向量的每一个维度提取最大 (小) 值作为回复向量对应维度的数值，然后计算他们的余弦相似度。Greedy Matching 寻找生成的回复和真实回复中最相似的一对单词，把这对单词的相似度近似为回复间的距离。
- **Distinct**[6]：在生成的回复中不一样的 n-grams (n=1,2) 种类与所有 n-grams 的比率，它衡量 n-gram 的多样性。每个测试对话上文抽取 3 个回复，将在抽取的回复内部和抽取的回复之间进行评估。相应地，Dist-n 被细分为 **Intra-Dist** 和 **Inter-Dist**。**Intra-Dist** 计算在采样回复内部计算该比率，**Inter-Dist** 在 3 个采样回复之间计算该比率。

人工评估 本研究跟随 Ke 等[37] 采用以下指标用于人工评估生成的回复，并雇佣 3 位评估人员来评估。

- **Informativeness**：用于评估生成的回复是否提供了有意义的信息；
- **Appropriateness**：用于评估生成的回复是否是合理，并且符合当前对话的逻辑；
- **Semantic Diversity**：用于评估给定对话上文生成的 3 个回复表达了几种不同的语义。

其中，Informativeness 和 Appropriateness 用于评估每一条生成的回复，Semantic Diversity 用于评估给定对话上文对应生成的一组回复。评级范围为 0 到 2，其中 0 表示最差，2 表示最好。

3.7 实验结果与分析

本小节对实验结果进行展示和分析，包括对话生成模型在 Douban 和 DailyDialog 数据集上的性能。本小节还进一步分析了模型各个模块的有效性，以及聚类类别数量 K 对模型性能的影响。此外本小节还验证了模型建模层级语义映射关系的能力。

3.7.1 评估模型性能

表 3-3和 3-4分别列出了关于相关性和多样性自动评估的结果。 如表中所示，MS-WAE 在多样性指标上，相比对比方法，取得了明显提升。同时，在相关性指标上，MS-WAE 相比对比方法得到了持平的结果。这表明 MS-WAE 在提升多样性的同时也保证了回复的质量，说明 MS-WAE 是有效的。进一步观察发现：(1) 相比所有对比方法，MS-WAE 获得了更好的性能，这表明利用多语义检索模块确保每个对话上文都有多个回复，并且不同回复具备不同语义是有效的。(2) 相比 DialogWAE，MS-WAE 获得了更高的 Intra-dist 和 Inter-dist 指标，这表明从混

表 3-3 自动评估相关性结果

Table 3-3 Results of automatic evaluation for the relevant metrics

模型	BLEU			BOW Embedding		
	Recall	Precision	F1	Average	Extrema	Greedy
Douban Corpus						
Seq2seq-attn	0.165	0.165	0.165	0.372	0.221	0.544
DCVAE	0.141	0.103	0.120	0.378	0.220	0.285
MMPMS	0.236	0.119	0.158	0.402	0.213	0.331
DialogWAE	**0.360**	0.218	0.272	0.537	**0.341**	**0.700**
MS-WAE	0.356	**0.221**	**0.273**	**0.556**	0.320	0.566
Daily Dialog Corpus						
Seq2seq-attn	0.195	0.195	0.195	0.874	0.508	0.706
DCVAE	0.274	0.241	0.257	0.897	0.509	0.758
MMPMS	0.301	0.230	0.261	0.915	0.506	0.758
DialogWAE	0.341	**0.245**	**0.285**	0.926	0.600	**0.803**
MS-WAE	**0.348**	0.222	0.271	**0.933**	**0.615**	0.625

合分布中抽样限制了多样性。并且 MS-WAE 获得了和 Dialog WAE 持平的 BLEU 分数和 BOW Embedding 分数，这表明 MS-WAE 在提升多样性的同时还能保证相关性。考虑到 BLEU 评估生成的回复和参考回复之间共享的词语的比例，所以 BLEU 指标并不太适合于对话生成任务。BOW Embedding 指标只计算生成的回复和参考回复之间的文本相似度，仅关注几个固定的语义侧面。当存在其他不同语义但和参考回复不太相关的合理回复时，则无法通过该指标进行合理评估。多样性的评估指标评价的是词级别的多样性，而不是语义级别的多样性。因此，本实验进一步通过人工评估来衡量生成回复的质量，尤其是语义多样性。

对于人工评估，评估结果如表 3-5 中所示。实验结果表明 MS-WAE 在所有评估指标上都要优于基线方法，这表明 MS-WAE 能够生成更有信息量、合理且语义多样的回复。进一步发现，MS-WAE 在两个数据集上的语义多样性都显著优于基线方法，这表明多语义建模有助于生成更多语义多样化的回复，并证明了 MS-WAE 针对多语义建模的有效性。信息量和适当性的结果也优于基线方法，因为从多语义分布中抽样可能会产生具有特定语义方面的回复，这比混合分布产生的回复更能具备信息量，也更容易产生合理的回复。图 3-8 给出了直观的对比。观察以上结果，可以得出结论，MS-WAE 可以提高语义多样性并保持质量。

3.7.2　实验分析

消融性实验　接下来，本研究进行以下消融测试来验证 MS-WAE 每个组件的效果：（1）移除语义距离损失函数，仅使用重构损失函数和判别器损失函数来训练模型（-sd_loss）；（2）在去掉语义损失函数的基础上，进一步移除不同回复与不

表 3-4 自动评估多样性结果

Table 3-4 Results of automatic evaluation for the diversity metrics

模型	Intra-dist		Inter-dist	
	dist-1	dist-2	dist-1	dist-2
Douban Corpus				
Seq2seq-attn	0.849	0.847	0.084	0.084
DCVAE	0.539	0.646	0.090	0.128
MMPMS	0.736	0.860	0.256	0.389
DialogWAE	0.701	0.769	0.345	0.541
MS-WAE	**0.872**	**0.925**	**0.554**	**0.879**
Daily Dialog Corpus				
Seq2seq-attn	0.916	0.969	0.091	0.096
DCVAE	0.857	0.943	0.155	0.207
MMPMS	0.883	0.968	0.304	0.438
DialogWAE	0.869	0.956	0.455	0.773
MS-WAE	**0.920**	**0.984**	**0.578**	**0.909**

同隐变量一一对齐的操作，仅使用对话上文为回复选择隐变量（-MSDM）；（3）在去掉语义损失函数的基础上，进一步移除多语义检索模块，仅使用原始对话数据训练模型（-MSRR）。实验结果如表格 3-6和 3-7 所示。经过观察发现：（1）移除语义距离损失函数（-sd_loss）后，MS-WAE 在相关性指标（BLEU、BOW Embedding）上的分数几乎没有任何变化，而在多样性指标（Intra-dist，Inter-dist）上的分数有了很明显的下降。这表明本研究提出的语义距离损失函数能够更好的提升模型的多样性，同时也能保持不错的相关性。（2）移除隐变量和不同语义侧面的一一对齐操作（-MSDM）后，MS-WAE 同样在相关性指标（BLEU、BOW Embedding）上的分数几乎没有任何变化，而在多样性指标（Intra-dist，Inter-dist）上的分数有了很明显的下降。这表明将隐变量和不同语义侧面一一对齐能够使得模型更好地生成不同语义的回复，但同时也会略微牺牲一些相关性。（3）移除多语义回复检索模块（-MSRR）后，MS-WAE 在多样性指标（Intra-dist，Inter-dist）上的分数下降得非常明显。这也同样表明本研究提出引入多语义回复检索模块对于提升模型语义多样性是有效的。总的来说，实验结果表明本节设计的所有这些模块对于提高回复生成的语义多样性都是必不可少的。

聚类类别数量 K 的影响 为了探究超参数聚类类别数量 K 对 MS-WAE 模型性能的影响，本实验分别在不同 K 值设置下去训练 MS-WAE，然后观察 MS-WAE 在验证集上的性能表现。具体来说，本实验设定 $K \in \{1, 7\}$。图 3-3展示了性能随着 K 值变化的曲线，性能由相关性指标（BLEU、BOW Embedding）和多样性指标（Intra-dist1,2, Inter-dist1,2）来体现。实验结果表明大多数情况下，MS-WAE

表 3-5 人工评估结果

Table 3-5 Results of human evaluation metrics

模型	Informativeness	Appropriateness	Semantic Diversity
	Douban Corpus		
Seq2seq-attn	0.380 ± 0.012 (8.00%)	0.300 ± 0.003 (30.0%)	0.000 ± 0.000 (0.00%)
DCVAE	0.644 ± 0.144 (35.0%)	0.380 ± 0.170 (39.2%)	0.286 ± 0.054 (8.00%)
MMPMS	0.708 ± 0.180 (36.5%)	0.432 ± 0.035 (42.4%)	0.760 ± 0.180 (24.0%)
DialogWAE	0.982 ± 0.059 (48.6%)	0.460 ± 0.071 (42.6%)	1.020 ± 0.143 (36.0%)
MS-WAE	**1.150** ± 0.120 **(60.4%)**	**0.544** ± 0.085 **(47.8%)**	**1.560** ± 0.126 **(82.0%)**
	Daily Dialog Corpus		
Seq2seq-attn	0.320 ± 0.019 (16.0%)	0.224 ± 0.003 (38.0%)	0.000 ± 0.000 (0.00%)
DCVAE	0.516 ± 0.043 (25.8%)	0.246 ± 0.089 (46.3%)	0.420 ± 0.100 (14.0%)
MMPMS	0.570 ± 0.047 (30.8%)	0.272 ± 0.043 (50.0%)	0.570 ± 0.078 (32.0%)
DialogWAE	0.914 ± 0.015 (51.0%)	0.292 ± 0.026 (53.6%)	0.714 ± 0.110 (36.0%)
MS-WAE	**1.078** ± 0.073 **(63.2%)**	**0.358** ± 0.063 **(58.0%)**	**1.480** ± 0.108 **(88.0%)**

表 3-6 相关性指标上的消融性实验结果

Table 3-6 Evaluation results of the ablation studies on the relevant metrics

模型	BLEU			BOW Embedding		
	Recall	Precision	F1	Average	Extrema	Greedy
			Douban Corpus			
MS-WAE	0.356	0.221	0.273	**0.556**	0.320	0.566
-sd loss	0.356	**0.258**	**0.299**	0.551	**0.337**	0.563
-MSDM	**0.359**	0.243	0.290	0.541	0.330	**0.683**
-MSRR	0.352	0.230	0.278	0.514	0.297	0.551
			Daily Dialog Corpus			
MS-WAE	0.348	0.222	0.271	**0.933**	0.615	0.625
-sd loss	**0.353**	0.223	0.273	**0.933**	0.614	0.624
-MSDM	0.344	**0.237**	**0.280**	0.925	0.585	**0.806**
-MSRR	0.344	0.234	0.278	0.931	**0.618**	0.591

的相关性随着 K 值的增加而增加。一旦 K 达到某个阈值，性能会发生下降。结果也表明最佳的 K 值在 4 左右。这可能是因为随着 K 值增加，给定对话上文拥有语义更加丰富的回复集合，有助于 MS-WAE 建模到多种语义可区别的隐变量，从而提升回复生成的质量。然而，当 K 增加到一定值时，不同类别的语义可能开始发生重合，那么 MS-WAE 在训练时，要求不同隐变量间的语义距离尽可能大可能无法正确训练 MS-WAE，从而导致回复生成性能下降。

表 3-7 多样性指标上的消融性实验结果

Table 3-7 Evaluation results of the ablation studies on the diversity metrics.

模型	Intra-dist		Inter-dist	
	dist-1	dist-2	dist-1	dist-2
Douban Corpus				
MS-WAE	**0.872**	**0.925**	**0.554**	**0.879**
-sd loss	0.826	0.913	0.532	0.863
-MSDM	0.830	0.897	0.532	0.817
-MSRR	0.750	0.823	0.505	0.830
Daily Dialog Corpus				
MS-WAE	**0.920**	**0.984**	**0.578**	**0.909**
-sd loss	0.909	0.979	0.558	0.885
-MSDM	0.814	0.948	0.562	0.879
-MSRR	0.889	0.973	0.532	0.875

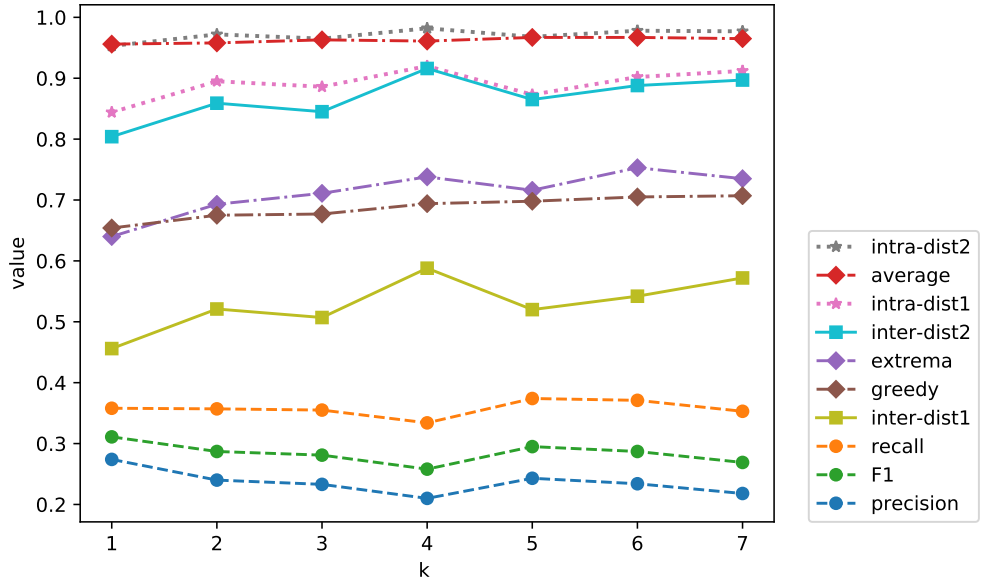
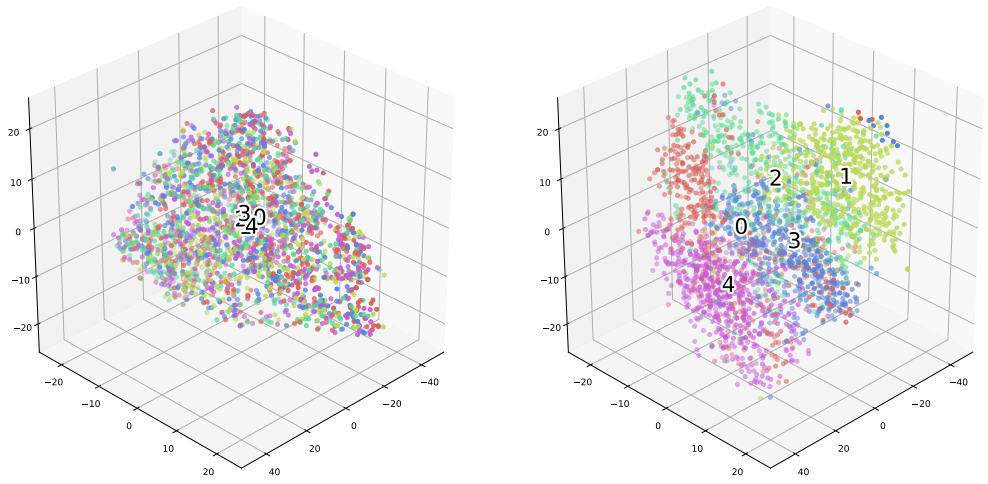


图 3-3 不同 K 值下模型的性能

Figure 3-3 Performance with respect to the number of clusters in K-means.

层级语义映射关系建模分析 本实验进一步分析 MS-WAE 能否建模出对话上文和回复间的层级语义映射关系。为此，本实验从 Douban 测试集中随机选择一个对话上文，可视化 MS-WAE 针对该对话上文学习的多语义分布，额外选择 DialogWAE 模型针对该对话上文建模的分布作为对比。具体来说，本实验分别从 DialogWAE 和 MS-WAE 的先验分布中的每个子分布里随机采样 500 个隐变量 $\tilde{\boldsymbol{z}}$ 并使用 t-SNE[162] 对隐变量进行可视化。可视化结果如图 3-4 所示。图 3-4b表明 MS-WAE 学到的分布里相同语义的隐变量值是相互聚集的，而不同语义的隐



(a) DialogWAE　　**(b) MS-WAE**

图 3-4 隐变量的 t-SNE 可视化展示

Figure 3-4 t-SNE visualization of latent variables

变量值之间是可区分的。相反，图 3-4a表明 DialogWAE 学到的分布里所有的隐变量值都是混合在一起的，这也就说明 DialogWAE 学习的不同隐变量之间是语义不可区分的。总的来说，可视化结果验证了 MS-WAE 能够建模对话上文和回复间的层级语义映射关系。同时也证明了将给定对话上文的不同回复与不同的子分布一一对应，并且使得子分布间的距离尽可能的大的改进是有效的。

案例分析　为了分析 MS-WAE 生成回复的质量，并和基线方法生成的回复进行对比，本实验从两个数据集中分别随机选择一条对话样本，然后图 3-8 分别展示了 MS-WAE 和基线方法 DialogWAE 生成的回复。对于 post-1 来说，案例表明 MS-WAE 能够捕获不同的语义去生成回复。回复 1 与语义侧面“坏事”相关联，回复 4 强调“微笑虽然是坏事”。回复 3 和 5 则侧重于“许久不联系的人”的语义侧面。相反，尽管 DialogWAE 生成的回复可以捕获一些语义侧面，但它倾向于从一个全局视角进行生成回复，这可能和 DialogWAE 倾向于建模混合分布有关。对于 post-2 来说，案例表明 MS-WAE 的回复 1 和回复 5 与语义侧面“来这儿”相对应，回复 2 和回复 4 又与语义侧面“爬山”有关联。此外，回复 3 则侧重于“经常”这一语义侧面。同样地，DialogWAE 生成的回复关联的语义侧面则相对单一，比如回复 1、2、3、4 都侧重于关注“爬上”这一语义侧面，只有回复 5 关联到了“经常”。总的来说，示例能够表明本研究提出的方法 MS-WAE 相比基线方法 DialogWAE 确实能够关注到更多不同可生成回复的语义侧面，进而去生成更多语义多样的回复。

表 3-8 多语义回复示例

Table 3-8 Real Cases of multi-semantic responses

Post-1: 许久不联系的人跟你说话都没好事，勒紧裤腰带，强颜欢笑吧
DialogWAE
Response-1:: 脾气不好啊 Response-2：什么时候打的电话 Response-3：我承认你很聪明 Response-4：回想一下这个也没什么啊，只是想通了 Response-5：对我的热情热情啊
MS-WAE
Response-1：哈哈想死 Response-2：哈哈能不联系的人就不要联系，突然之间联系你准没好事的。这是真理呀 Response-3：你就这么想啊 Response-4：每个人在一起都会开心的。即使不开心了，等过一段时间彼此冷静下来就好了吧 Response-5：怎么又联系你了
Post-2：不错！我经常来这儿爬山
DialogWAE
Response-1：说实话，还不错。我感觉我应该去看电影的，但我现在又不是很想去看电影，怎么办? Response-2：温度还不错呢 Response-3：我知道，但我不认为现在去一个新的地方是个好主意 Response-4：喔，今天天气不错，适合去公园玩 Response-5：你有时间吗
MS-WAE
Response-1：你可以先做 5 路公交，然后步行到这儿 Response-2：你计划什么时候去旅行 Response-3：好的，王先生。那我想确认下明天早上 9 点的会议您还能准时参加吗? Response-4：哇，这需要很长的时间呢 Response-5：我很喜欢在这里工作

3.8　小结

本章提出了一个基于语义表示的对话回复生成方法，用于显式建模层级语义映射关系，进而提升回复生成任务的语义多样性。具体来说，本章设计了一个多语义 Wasserstein 自编码器，即 MS-WAE。首先该方法从数据集中检索 k 个有效回复扩充原始对话数据，确保训练数据中每个对话上文都有多个语义回复侧面。其次该方法改进了对话 Wasserstein 自编码器（DialogWAE）。相比 DialogWAE，该方法优化了混合高斯子分布选取的方式，不再像前人工作一样随机选择，而是根据回复所属的聚类编码自动对齐到对应编号的子分布上。进一步，本方法额外引入了基于 MMD 的语义距离损失函数使得子分布间尽可能分得开。实验结果表明本节方法能够有效地显式建模层级语义映射关系，并有效地提升回复生成的语义多样性。

第 4 章　基于语义转换的的对话数据增强方法

4.1　引言

开放域对话系统的构建需要高质量的对话训练数据，因为训练数据的质量很大程度上决定了模型性能的上限。对话数据可以针对给定的对话上文提供多种不同的回复，尤其是具有不同语义的回复。然而，在大多数情况下，收集这样高质量的数据集需要耗费大量人力和时间。因为这需要标注人员人工编写大量各种各样的有效回复。尽管这类数据可以从社交网络爬取，但这样收集的数据集充满噪音和无意义的回复数据。从中挑选出符合要求的充足的高质量数据也是非常耗时耗力的。近期一些工作尝试使用数据增强技术来扩增数据。现有方法仅仅考虑了扩增原始回复词语级别或者句子级别的可替代表达，扩增的数据和原始数据仅存在有限的语义差异。因此，如何有效地扩增出不同语义的回复是该领域的一个重要问题。针对这一问题，本文提出了一种反事实数据增强方法，通过反事实推理自动扩增不同语义的高质量回复。具体来说，针对观察到的对话，该方法的反事实生成模型首先通过转换观测到的可生成回复的语义角度，然后重新推理新的不同语义的回复。此外，该方法的数据选择方法可以过滤掉有害的增强回复。实验结果表明，该方法能够为给定的对话上文扩增出不同语义的高质量回复，并且显著地提升了下游任务的模型性能，尤其是明显地提升了对话生成任务中回复的语义多样性。

4.2　概述

开放域对话系统因其潜在的应用价值已经受到了广泛的关注[2,3,15,163]。一般来说，训练开放域对话系统需要高质量的对话训练数据。对话数据允许给定的对话上文有各种各样的回复[164]。具体而言，对于给定的对话上文，可以存在许多具有不同语义的回复，每种语义信息的回复也可以具有不同的表达方式[67]。然而，人工收集高质量的这类数据集通常是费时费力的。因为这需要标注人员去人工写大量各种各样的有效回复。尽管这类数据可以从社交网络爬取，但这样收集的数据集充满噪音和无意义的回复数据。从中挑选出符合要求的充足的高质量数据也是非常耗时耗力的。

解决这一问题的可行方案是使用数据增强技术。目前已经有一些研究工作[65–70,72] 被用于开放域对话中。现有研究工作主要分为两类：对现有数据稍作修改，以及创建合成数据。具体来说，第一类工作包括置换和翻转对话历史扩增数据[65]，或者利用语义不变的扰动、掩码语言模型[28] 和回译[72] 来分别进行词语级别和句子级别的数据扩增[66,68,70]。第二类工作包括利用对话数据训练一个生成模型去产生更多的数据[67]，以及利用检索方法检索和已有对话数据相似的样本组成新的对话数据[69]。然而这些扩增的数据和原始数据相比都只有有限

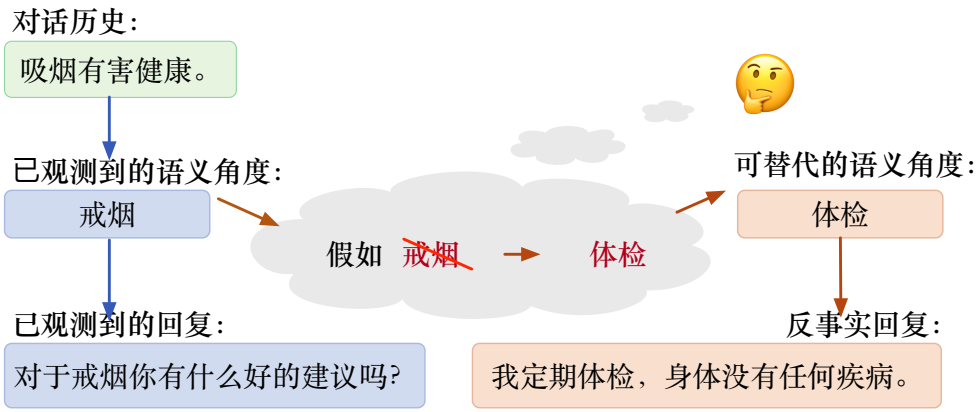


图 4-1 不同语义回复的生成过程示例

Figure 4-1 An example of response generation with different semantics

的语义不同，因为都只进行了受限的修改。这些已有的方法仅仅考虑了扩增原始回复词语级别或者句子级别的可替代表达，并没有扩增更多的不同语义角度的回复。

因此，本研究的目标是为给定的对话上文扩增更多不同语义的回复。本文借鉴了人类生成不同语义回复的过程。给定一个对话上文，人首先关注到对话上文中的某部分内容，进而关注点转移到一个想要谈论的语义角度去生成一个回复。但随后人会去思考一个问题：如果改变当前关注的语义角度，回复会有什么不同。回答这个问题会推断出不同的回复。图 4-1给出一个示例，如果将已观测到的语义角度“戒烟”转换成“体检”，重新推理得到的回复也会语义完全不同。人在否定过去发生的事情并进行重新推理时，会保持当前环境不变，即除了改变语义角度，其他影响回复生成的因素（如人当前的情绪状态、说话风格等）是不变的。这种在当前环境下的推理就是所谓的反事实推理[165]，基于一定的事实基础有利于保证推理结果的质量[166]。

受其启发，本文提出了基于语义转换的反事实数据增强方法（Counterfactual Data Augmentation via Semantic Transition, CAST），用于为给定的对话上文生成反事实回复。CAST 把反事实生成模型解释为一个可进行反事实推理的结构因果模型（Structural Causal Model, SCM）[165]，用于描述在当前环境下的生成过程。具体来说，当前环境由 SCM 中的不可观测变量建模，该变量捕捉了所有不可观测但影响回复生成的相关因素，即当前并不关注但会影响回复生成的因素（如说话人当前的情绪状态和说话风格等）。反事实回复通过在当前环境下干预 SCM 中的语义角度，即把已观测到的语义角度转换为其他有效的语义角度去生成。为了获得不同且有效的可生成回复的语义角度，该方法首先基于所有观测到的训练数据构建一个上文关注点-可生成回复的语义角度转移图，它显式地建模了人类在对话上文上的关注焦点与其相应的语义角度之间的转移关系。然后该方法从给定的对话上文上随机选择一个关注焦点，并将该关注焦点在转移图中的一跳邻居节点，即所有观测数据中从关注焦点合理转移到的语义角度，作为候选集

合。进一步从候选集合中预测出一个有效的回复角度。在得到所有重新推理的反事实回复后，该方法额外使用数据选择方法去过滤增强数据。最后，该方法将观测数据与过滤后的增强数据混合，作为下游任务的训练数据。

在开放域对话数据集 Weibo[69] 上的实验结果表明，该方法能够扩增出具有不同语义的高质量回复。并且使用本文丰富语义内容后的训练数据能够有效地提升对话生成的语义多样性。此外，使用这种更加类人的训练数据不仅能够提升对话生成任务，还能提升对话检索任务的整体质量。

本章组织如下，第 4.3节中介绍本工作的相关工作；第 4.4节中介绍本文工作相关的背景知识；第 4.5节中介绍本文提出的基于视角转换的反事实数据增强方法；第 4.6节中介绍本工作相关的实验设置，包括使用的数据集情况，对比的基线模型，使用的评价指标，和实验细节等；第 4.7给出了本工作的实验结果，并进行了相关对比与分析。最后对本章进行了总结。

4.3　相关工作

数据增强　数据增强技术目前已广泛用于各种 NLP 任务中，并由 Ni 等[3], Shorten 等[61], Wen 等[62], Feng 等[63], Chen 等[64] 进行了文献综述。总的来说，数据增强方法要么对现有数据稍作修改，要么创建合成数据。对于对现有数据稍作修改，这类工作主要使用启发式规则[65] 或基于改写的方法[66,68,70,71]。对于创建合成数据，这类工作正在利用生成模型进行数据扩充。此前，Li 等[67] 采用条件变分自动编码器 (CVAE) 作为生成器来输出更多的训练数据。此外，Zhang 等[69] 利用数据检索的思路扩增数据：以已有对话数据为基准，分别从非平行语料中检索到和给定对话上文和回复最相似的句子，将其构造出新的对话数据。目前更多的研究[73–76,167] 使用大规模预训练模型生成更多数据，比如 GPT-2[56] 和 BART[59]。然而，这些现有方法的目标是用多样化的词产生观察到回复的可替代表达，而不是构建更多的语义上不同的回复。

不同语义的扩增　Gangal 等[168] 利用外部知识源，包括 COMET[169] 和语料库检索[170] 来增加对话评估过程中的语义多样的参考回复。其中，COMET 为常识知识模型，它学习了在自然语言中生成丰富和多样的常识描述，能够为数据增强提供常识层面的扩增角度。语料检索库则为已有的对话数据，利用 BM25 算法从中检索出更多和对话上文连贯的回复作为扩增数据，以此去丰富参考回复的语义内容。这两种方法都只预定义了有限的扩增角度。相比之下，本文工作通过构建转移关系图能够获得更丰富的可生成回复的语义角度。

反事实推理　反事实推理目前已经在许多 NLP 任务中都带来了可喜的成果，包括问答系统[171,172]、机器翻译[173] 和故事生成[174–176] 等。这些工作都是借助于反事实推理技术去产生更多不一样的高质量且更加真实的数据。对于开放域对话任务，Zhu 等[166] 尝试使用反事实推理进行回复生成。该方法通过反事实离

线策略训练模型去探索更多潜在的回复。相比之下，本文工作侧重于反事实数据增强，能够扩增出更多语义不同的回复。并且本文丰富语义内容后的训练数据可用于提高多个下游任务的整体性能，不仅仅针对回复生成任务。

图构建 一些研究工作[97,98]同样通过构建图来建模概念转换以产生对话回复，使用图旨在生成更连贯和可控的对话。具体来说，Xu 等[97]提出将有关对话转换的先验信息构建成图，并学习基于图的对话策略。Zou 等[98]指出同一主题可能包含多个关键词，该工作提出首先建模多关键词下的主题转移关系图，然后通过从图中适合于当前对话的关键词去生成回复。相比之下，本文方法构建了一个转移关系图来预测有效的其他可生成回复的语义角度，这些角度被用作增强具有不同语义的回复。由于使用目的不同，本文的图构建方法与这些现有工作有很大的不同。

4.4 背景知识

本节主要描述开放域对话系统相关的任务定义，并回顾用于反事实推理的结构因果模型的相关概念。

4.4.1 任务定义

回复选择 给定数据集 $\mathcal{D} = \{(\boldsymbol{x}^i, \boldsymbol{y}^i, l^i)_{i=1}^N\}$，基于检索的对话模型学习一个映射函数，用于从一组负例回复中正确识别出正例回复。具体来说，映射函数 $P_\theta(l^i|\boldsymbol{x}^i, \boldsymbol{y}^i)$ 预测回复 $\boldsymbol{y}^i$ 是否匹配对话上文 $\boldsymbol{x}^i$，即是否是一个合理回复。$l^i \in \{0, 1\}$ 表示匹配标签，如果 $l^i = 1$ 表明 $\boldsymbol{y}^i$ 是 $\boldsymbol{x}^i$ 的一个合理回复，$l^i = 0$ 则相反。模型参数 θ 通过最小化损失函数来学习，损失函数表示为

$$\mathcal{L}_{sel} = -\sum_{i=1}^{N}[l^i \log P_\theta(l^i = 1|\boldsymbol{x}^i, \boldsymbol{y}^i) + (1 - l^i) \log P_\theta(l^i = 0|\boldsymbol{x}^i, \boldsymbol{y}^i)]. \tag{4.41}$$

通常，训练过程中的负例回复是从数据集中 $\mathcal{D}$ 随机选择的。

回复生成 给定数据集 $\mathcal{D} = \{(\boldsymbol{x}^i, \boldsymbol{y}^i)_{i=1}^N\}$，基于生成的对话模型在给定对话历史 $\boldsymbol{x}^i$ 时，学习建模回复 $\boldsymbol{y}^i$ 的概率分布 $\mathcal{P}_\phi(\boldsymbol{y}^i|\boldsymbol{x}^i)$。模型参数 ϕ 可以通过最小化以下损失来学习：

$$\mathcal{L}_{gen} = -\sum_{i=1}^{N} \log P_\phi(\boldsymbol{y}^i|\boldsymbol{x}^i). \tag{4.42}$$

然而，一个对话数据集允许每个对话历史有多个语义不同的回复，收集这样的对话数据集通常非常昂贵。因为它需要标注人员编写各种各样的有效回复。尽管可以从社交网络中抓取这样的数据集，但它会包含许多嘈杂且无意义的回复。挑选出足够多的符合要求的高质量对话也是很昂贵的。因此，反事实数据增强旨

在进一步增强数据集 $\mathcal{D}$ 中 $\boldsymbol{x}^i$ 的不同语义回复 $\tilde{\boldsymbol{y}}^i$，而无需手动收集新数据。在以下部分中，为简单起见，将省略上标 i。

4.4.2　结构因果模型

定义　结构因果模型（SCM）包含若干个可观测变量 $\mathbf{V} = \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m\}$ 和若干服从分布 $P(\mathbf{U})$ 的独立不可观测的随机变量 $\mathbf{U} = \{\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m\}$。这些变量之间通过一系列函数 $\mathbf{F} = \{f_1, \ldots, f_m\}$ 进行连接。具体来说，$\forall i$，$\boldsymbol{V}_i$ 由以下变量因果决定：一系列父变量 $\mathbf{PA}_i$ 和相应的不可观测变量 $\boldsymbol{U}_i$，即 $\boldsymbol{V}_i = f_i(\mathbf{PA}_i, \boldsymbol{U}_i)$。其中 $\mathbf{PA}_i \subseteq \mathbf{V} \setminus \boldsymbol{V}_i$ 在因果有向无环图（Directed Acyclic Graph，DAG）中[177]。

对于反事实生成模型而言，它可以被解释为一个结构因果模型（SCM），包含三个可观测变量对话历史 $\boldsymbol{X}$，语义角度 $\boldsymbol{Z}$ 和回复 $\boldsymbol{Y}$。反事实生成 SCM 将反事实生成模型学习预测的条件分布 $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ 转换成学习一个确定性预测函数 $\boldsymbol{Y} = f(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{U})$。其中 $\boldsymbol{U}$ 捕捉了当前环境中所有不可观测但又对反事实生成有影响的因素，比如人类的说话风格、当前的情感倾向或者所拥有的背景知识等。函数 f 由训练好的反事实生成模型表示。总体而言，SCM 可以根据已知函数 f 和不可观测变量的后验推断反事实回复 $P(\boldsymbol{U}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{Y} = \boldsymbol{y})$。

干预　在观察可观测变量 $\boldsymbol{V}_i$ 会发生什么之前，对其父变量 $\boldsymbol{V}_j, \boldsymbol{V}_j \in \mathbf{PA}_i$ 进行干预，其中 SCM 中的干预是通过修改 SCM 和改变观测值来采取的一种行动。具体来说，SCM 中变量 $\boldsymbol{V}_j$ 设置为 $\boldsymbol{V}_j = \boldsymbol{v}'_j$，并且切断 $\boldsymbol{V}_j$ 所有父节点到 $\boldsymbol{V}_j$ 之间的边。对于反事实生成 SCM，如图 4-2中的 SCM，由于 $\boldsymbol{Z}$ 没有父节点，因此干预只将语义角度 $\boldsymbol{Z}$ 的观测值 $\boldsymbol{z}$ 转换成不同的值 $\tilde{\boldsymbol{z}}$。

反事实推理　给定一个 SCM 并观察变量 $\boldsymbol{V}_i = \boldsymbol{v}_i$，反事实推理回答了以下问题：如果 $\boldsymbol{V}_i$ 的父变量 $\boldsymbol{V}_j$ 在保持当前环境不变的情况下被干预，变量 $\boldsymbol{V}_i$ 会发生什么变化？因此，生成反事实回复需要思考一个问题：如果将可生成回复的语义角度 $\boldsymbol{Z}$ 设置为不同的值 $\tilde{\boldsymbol{z}}$，而不是已观测到的值 $\boldsymbol{z}$，即对其进行干预，回复 $\boldsymbol{Y}$ 会发生怎样的变化。

总的来说，为了产生反事实回复，需要遵循以下三个步骤：

- 外展（Abduction）：预测 SCM 的当前环境，即计算后验分布 $P(\boldsymbol{U}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{Y} = \boldsymbol{y})$，然后从中采样 $\boldsymbol{u}$ 表示当前环境；
- 行动（Action）：执行干预，即用不同的 $\tilde{\boldsymbol{z}}$ 替换已观测到的 $\boldsymbol{z}$；
- 预测（Prediction）：给定后验采样 $\boldsymbol{u}$，去生成反事实回复 $\tilde{\boldsymbol{y}}$。

4.5　对话数据增强方法

本节的目标是以给定的对话样本 $(\boldsymbol{x}, \boldsymbol{y})$ 为输入，去扩增高质量的且具有和 $\boldsymbol{y}$ 不同语义的回复。为此， 4.5.1节介绍了一种称为通过视角转换进行反事实生成

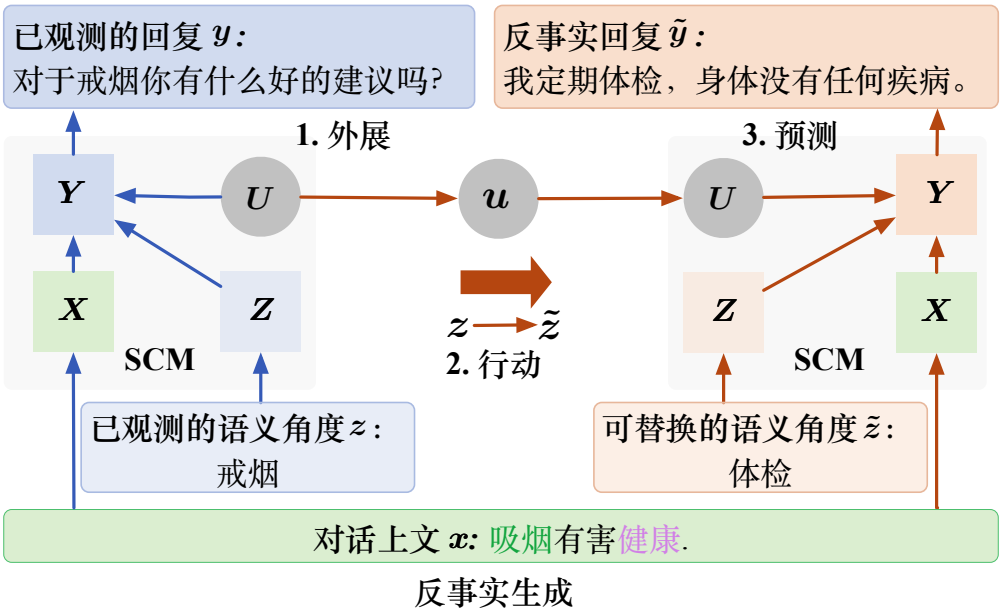


图 4-2 反事实生成的三个步骤

Figure 4-2 The three-step procedure of counterfactual generation

的技术，用于干预已观测到的回复视角，以在当前环境下扩增回复。 4.5.2节描述了如何训练 4.5.1节中涉及的模型，包括语义角度预测器和反事实回复生成器。4.5.3节设计了一种名为双向困惑选择的数据选择方法来选择高质量且有趣的扩增数据。

4.5.1 基于语义转换的回复生成

本研究主要关注于单轮对话。给定一个对话样本 $(\boldsymbol{x}, \boldsymbol{y})$，按照图 4-2中所示的三个步骤使用 SCM 生成一个反事实回复 $\tilde{\boldsymbol{y}}$。

1、外展 该步骤用于根据已观测到的样本 $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y})$ 去估计不可观测变量（更多关于 $\boldsymbol{z}$ 的细节将在“2、行动”小节中介绍）。具体来说，当生成回复 $\boldsymbol{y}$ 的第 t 个单词时，反事实生成模型输出的词表分类概率为 $P(Y_t|\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Z}=\boldsymbol{z}, \boldsymbol{Y}_{<t}=\boldsymbol{y}_{<t})$，其中 $\boldsymbol{y}_{<t}$ 为之前时间步已生成的词序列。根据 Oberst 等[178] 的工作，不可观测变量 $\boldsymbol{U}_t$ 可以用 Gumbel 随机噪音来估计。因此，对该分类概率使用 Gumbel-Max Trick[179]，即：

$$\begin{aligned} p_{tk} &= P(Y_t = k|\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Z}=\boldsymbol{z}, \boldsymbol{Y}_{<t}=\boldsymbol{y}_{<t}), \\ y_t &= \arg\max_{k=1,\ldots,|V|} (\log p_{tk} + u_{tk}), \end{aligned} \tag{4.51}$$

其中 $u_{tk} \sim \text{Gumbel}(0, 1)$，$|V|$ 表示词表大小。

随后，反事实生成 SCM 被转化成 Gumbel-Max SCM[178]。不可观测变量能够通过从这些 Gumbel 随机变量的后验分布中采样。一种直观推理后验的方法[180]是利用位移 Gumbel 变量 $g_{tk} = \log p_{tk} + u_{tk}$ 的性质：最大值服从标准 Gumbel

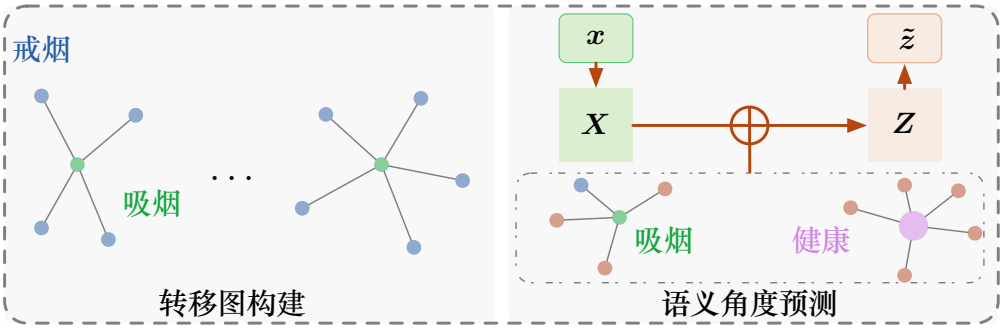


图 4-3 行动步骤流程

Figure 4-3 The process of action step

分布，并且和位移 Gumbel 变量的最大值独立。因此，对于 $y_t = k^*$（$*$ 表示观测到的词语），采样 $g^*_{tk} \sim \text{Gumbel}(0,1)$；对于剩下的 g_{tk}，从位移 Gumbel 分布 $\text{Gumbel}(\log p_{tk}, 1)$ 中采样。然后，u_{tk} 的采样通过从 g_{tk} 中减去 $\log p_{tk}$ 即可获得。最终 $\boldsymbol{u}_t = [u_{t1}, \ldots, u_{t|V|}]$ 表示当前环境，将被用于推理反事实回复。

2、行动　该步骤用于把已观测到的语义角度 $\boldsymbol{z}$ 转换成不同且有效的语义角度 $\tilde{\boldsymbol{z}}$，流程如图 4-3所示。然而有两个问题亟需解决：如何表示语义角度和如何预测一个不一样且有效的语义角度。通过观察发现，人类对话的不同语义回复可以通过以下过程得到：人类首先自然地关注到给定对话上文的一个焦点，比如“吸烟”，然后会无意识地转移这个关注焦点到另一个焦点，比如“戒烟”。本研究把对话上文的关注焦点和转移的关注点分别称为上文关注点和可生成回复的语义角度。当人类有不同的上文关注点时，如图 4-2中的“健康”，从该上文关注点转移到的所有想谈论到的语义角度都是不同的。此外，即使人类有相同的上文关注点“健康”，只要有不同的转移关系也能得到不同的语义角度。

进一步为了获得有效的语义角度，建模从对话上文关注点到可生成回复的语义角度间的有效转移尤为重要。本文基于所有已观测到的对话样本构造上文关注点-语义角度转移关系图。其中头节点和尾节点分别是对话上文关注点和语义角度，通过观察发现二者均可以由关键词来表示，比如图 4-2中的“戒烟”。而边代表已观测到的对话上文关注点和语义角度间的真实转移关系。受 Xu 等[97]和 Zou 等[98]启发，本研究把给定对话上文关注点的一跳节点，即所有在已观测到的对话数据中从给定对话上文关注点合理转移来的语义角度，作为候选集合，然后从该候选集合中预测一个即可作为有效的语义角度。这主要是基于一个假设：如果对话上文和观测数据中某个样本的对话上文相似，并且两个样本具有相同的对话上文关注点，那么可生成回复的语义角度则可以共享。

为了构建转移关系图 $\mathcal{G}$，主要包括两个步骤：节点构建和边构建。对于节点构建，本文首先利用基于规则的关键词抽取方法[181]分别从已有对话数据集 $\boldsymbol{D}$ 中对话上文和回复中抽取有价值的关键词。进一步，为了从 $\boldsymbol{x}$ 的所有关键词中识别出上文关注点 $\boldsymbol{c}$，本文使用来自未来信息（即参考回复）的指导来选择语义

上最接近 $\boldsymbol{y}$ 的关键词。而为了识别语义角度 $\boldsymbol{z}$，本文选择与 $\boldsymbol{c}$ 语义最接近的关键词。更具体来讲，本文使用通过 BERT[28] 计算的的句子嵌入向量之间的余弦相似度作为语义接近程度的度量，其中每个句子嵌入向量都是通过取每个词嵌入向量的平均值来表示的。对于边构建，本文通过连接 $\boldsymbol{c}$ 和 $\boldsymbol{z}$ 来构建一条边。通过这种方式能够显式地表示观测数据 $\mathcal{D}$ 中的所有有效的转移关系。

一旦转移关系图构建完成，本文按照以下方式预测 $\tilde{\boldsymbol{z}}$：

$$\tilde{\boldsymbol{z}} = \arg\max_{\tilde{\boldsymbol{z}}} P(\boldsymbol{Z}|\boldsymbol{C}=\tilde{\boldsymbol{c}}, \boldsymbol{X}=\boldsymbol{x}, \boldsymbol{N}=\mathcal{N}(\tilde{\boldsymbol{c}})), \tag{4.52}$$

这需要一个已训练的语义角度预测器来实现，其中 $\tilde{\boldsymbol{c}}$ 可以是对话上中的任意一个对话上文关注点，$\mathcal{N}(\tilde{\boldsymbol{c}})$ 表示该对话上文的一跳邻居节点，即有效可转移到的语义角度。

3、预测 该步骤在给定的后验采样 $\boldsymbol{u}_t = [u_{t1}, \ldots, u_{t|V|}]$ 下生成反事实回复。具体来说，当生成反事实回复的第 t 个词语时，反事实生成模型按照以下公式计算分类概率分布，

$$\begin{aligned} \tilde{p}_{tk} &= P(Y_t = k|\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Z}=\tilde{\boldsymbol{z}}, \boldsymbol{Y}_{<t}=\tilde{\boldsymbol{y}}_{<t}), \\ \tilde{y}_t &= \arg\max_{k=1,\ldots,|V|} (\log \tilde{p}_{tk} + u_{tk}), \end{aligned} \tag{4.53}$$

其中 $\tilde{\boldsymbol{z}}$ 是预测出的语义角度，$\boldsymbol{y}_{<t}$ 为之前时间步已生成的词序列。

总的来说，通过语义转换的反事实回复生成可以用作开放域对话领域有效的数据增强方法，以扩增具有更广泛语义覆盖的回复。整个数据增强流程由算法 2展示。该算法将已观测到的样本 $(\boldsymbol{x}, \boldsymbol{y})$ 作为输入，并循环遍历 $\boldsymbol{x}$ 的每个关键字作为不同的上文关注点 $\tilde{\boldsymbol{c}}$。对于每个 $\tilde{\boldsymbol{c}}$，为了采样多个对应的语义角度，该算法将候选集合 $\mathcal{N}(\tilde{\boldsymbol{c}})$ 平均划分为 K 个子集，即 $\mathcal{N}_1(\tilde{\boldsymbol{c}}), \ldots, \mathcal{N}_K(\tilde{\boldsymbol{c}})\}$，采用嵌套循环设置。在每次迭代中，该算法预测一个不同的 $\tilde{\boldsymbol{z}}$ 用于语义转换以输出反事实样本 $(\boldsymbol{x}, \tilde{\boldsymbol{y}})$。

4.5.2 相关模型训练

CAST 依赖于语义角度预测模型和反事实回复生成模型，它们极大地影响了增强数据的质量。受 Yang 等[74] 和 Schick 等[75] 工作的启发，本文选择大规模预训练编码器-解码器模型 BART[59] 作为主干模型。

语义角度预测模型 本文在数据集 $\mathcal{D}$ 上微调 BART 学习 $P(\boldsymbol{Z}|\boldsymbol{C}, \boldsymbol{X}, \boldsymbol{N})$。具体来说，输入是由对话上文 $\boldsymbol{X}$，对话上文关注点 $\boldsymbol{C}$ 和可生成回复的语义角度的候选集合 $\boldsymbol{N}$ 拼接而成的文本序列。输出是需要预测的可转换的语义角度 $\boldsymbol{Z}$。通过最大化以下目标函数来训练模型，即

$$\mathcal{L}_p = -\sum_{t=1}^{|Z|} \log P(Z_t|[\boldsymbol{C}, \boldsymbol{X}, \boldsymbol{N}], \boldsymbol{Z}_{<t}), \tag{4.54}$$

算法 2 反事实数据增强

Input: $(\boldsymbol{x}, \boldsymbol{y})$: 观测对话样本
C: $\boldsymbol{x}$ 包含的全部关键词 $\{\tilde{\boldsymbol{c}}_1, \ldots, \tilde{\boldsymbol{c}}_{|C|}\}$
$\mathcal{G}$: 转移关系图
Output: 反事实对话样本 $(\boldsymbol{x}, \tilde{\boldsymbol{y}})$

给定已观测到的可生成回复的语义角度 $\boldsymbol{z}$
for $i \leftarrow 1$ **to** $|C|$ **do**
    从 $\mathcal{G}$ 中获取 $\tilde{\boldsymbol{c}}_i$ 的一跳邻居节点 $\mathcal{N}(\tilde{\boldsymbol{c}}_i)$
    从 $\mathcal{N}(\tilde{\boldsymbol{c}}_i)$ 中移除 $\boldsymbol{z}$
    将 $\mathcal{N}(\tilde{\boldsymbol{c}}_i)$ 等分为 $\{\mathcal{N}_1(\tilde{\boldsymbol{c}}_i), \ldots, \mathcal{N}_K(\tilde{\boldsymbol{c}}_i)\}$
    for $j \leftarrow 1$ **to** K **do**
        $\tilde{\boldsymbol{y}} \leftarrow \text{Trans}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \tilde{\boldsymbol{c}}_i, \mathcal{N}_j(\tilde{\boldsymbol{c}}_i))$
Function Trans$(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \tilde{\boldsymbol{c}}, \mathcal{N}(\tilde{\boldsymbol{c}}))$**:**
    从 $P(\boldsymbol{U}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ 中推理代表当前环境的后验采样 $\boldsymbol{u}$
    从 $P(\boldsymbol{Z}|\boldsymbol{x}, \tilde{\boldsymbol{c}}, \mathcal{N}(\tilde{\boldsymbol{c}}))$ 中预测有效且不同的语义角度 $\tilde{\boldsymbol{z}}$
    在当前环境 $\boldsymbol{u}$ 下根据 $P(\boldsymbol{Y}|\boldsymbol{x}, \tilde{\boldsymbol{z}})$ 推理出反事实回复 $\tilde{\boldsymbol{y}}$
    return $\tilde{\boldsymbol{y}}$

其中中括号 $[\cdot, \cdot, \cdot]$ 表示各个输入文本通过标记 [SEP] 拼接。候选集合 $\boldsymbol{N}$ 使用逗号拼接。$\boldsymbol{Z}_{<t}$ 表示之前时间步生成的语义角度的前缀。$|Z|$ 表示 $\boldsymbol{Z}$ 的长度。

反事实回复生成模型　本文在数据集 D 上微调 BART 学习 $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$。具体来说，反事实回复生成模型的输入包含对话上文 $\boldsymbol{X}$ 和语义角度 $\boldsymbol{Z}$，模型被训练生成回复 $\boldsymbol{Y}$。相似地，通过最大化以下目标函数来训练模型，即

$$\mathcal{L}_g = -\sum_{t=1}^{|Y|} \log P(Y_t|[\boldsymbol{X}, \boldsymbol{Z}], \boldsymbol{Y}_{<t}), \tag{4.55}$$

其中 $|Y|$ 表示可观测回复 $\boldsymbol{Y}$ 的长度。

4.5.3　双向困惑度数据过滤

过滤掉低质量的增强样本可以提高下游任务的性能[182]。现有方法[69,183,184]主要使用已观测到的对话数据训练一个模型，然后让模型挑选出它认为质量最好的样本。然而，这些模型只见过有限的对话样本，因此它们可能无法从反事实生成的样本中识别出有效但未见过的样本。受到 Lee 等[185] 的启发，本研究尝试借助大规模对话预训练语言模型 DialoFlow[48] 而非通用预训练模型，利用其强大的迁移学习能力去判断样本的质量。由于大规模对话预训练语言模型已经见过大量的对话样本，所以它可以像“专家”一样通过困惑度 (Perplexity，PPL) 分数来识别有效但未见过的样本。但 PPL 分数难以过滤掉那些包含有通用回复

的样本。受 Li 等[6] 的启发，本研究进一步引入后向 PPL 来对生成的若干回复进行重排序，以便对那些有效且有趣的样本进行优先选择。

具体来说，本研究在 $\boldsymbol{D}$ 上分别微调 DialoFlow 去学习 $\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X})$ 和 $\boldsymbol{P}(\boldsymbol{X}|\boldsymbol{Y})$，用来计算前向和后向 PPL 分数。一旦获得了所有样本的前向 PPL 分数，则需要找到一个将有效样本与无效样本分开的最佳阈值 η。受 Lee 等[185] 的启发，本研究利用验证集寻找最优单阈值参数 η。首先将验证集中观测到的样本视为有效样本，然后通过用随机采样的回复替换有效样本的回复的方式来构造无效样本。最终找到一个数值使得前向 PPL 分数小于该值的样本中，有效样本的准确率和召回率是综合最优的，这个数值就是所寻找的最优阈值参数 η。此外，本研究根据反向 PPL 分数对有效样本中每个对话上文的回复进行重新排序。考虑到后向 PPL 分数越高，回复越有可能是通用回复[6]，因此再按照从低到高的顺序选择样本，直到获得所需数量的增强样本。

4.6 实验设置

本小节主要描述实验相关的设置，包括数据集、实验模型和对比方法、实现细节以及评估方案。

4.6.1 数据集介绍

本研究在中文微博语料库 Weibo[69] 上进行了实验。具体来说，数据集 $\boldsymbol{D}$ 包含训练集、验证集和测试集，它们分别具有 300K、5K 和 10K 条对话样本。为了构建转移关系图，本研究使用 YAKE[181] 自动提取训练数据中每个话语中最重要的关键词，YAKE 主要依赖于文本统计特征进行自动提取。关键词仅限于名词、形容词和动词。关键词节点和边的数量分别为 77,439 和 202,266。此外随机抽取了 200 个回复样本，并雇佣了三位评估人员来评估代表对话上文关注点的关键词和代表可生成回复语义角度的关键词是否合理。评估人员认为大约 86% 关键词对是合理的。训练阶段和扩增阶段候选语义角度的平均数量分别为 102 和 124。在获得增强数据后，本研究类似地评估了回复是否与给定的语义角度共享相似的核心语义。评估人员认为大约 96.5% 的回复是满足要求的。

4.6.2 实现细节

基于语义转换的数据增强方法 对于转移关系图构建，本研究利用文本表示工具 bert-as-service[186]，通过将可变长度的文本序列映射到固定长度的向量来获取文本嵌入表示。该方法的语义角度预测模型和反事实回复生成模型分别使用公式 4.54和公式 4.55微调 BART-large 模型[187]。训练模型的 epoch 次数为 10 次，批处理大小为 64，学习率为 $1e^{-5}$。其他模型超参数设置和 Shao 等[187] 的工作保持一致。模型输入和输出的最大序列长度设置为 512。因此，语义角度转换模型的候选集合最大被限制为 100。如果候选集合规模大于 100，则从集合中随机抽取 100 个候选语义角度。此外，还过滤掉了那些候选集合小于 5 的样本。对于数

据选择，本研究通过在数据集 D 上微调对话预训练模型 DialoFlow[48] 来实现评分函数。训练模型的迭代次数为 2 次，批处理大小为 64，学习率为 $1e^{-5}$。最佳阈值 η 为 10。

在数据增强阶段，本研究同样将候选集合的大小限定在 5 到 100 之间。因此，整个候选集被划分为 K 个子集，并设置每个子集的候选大小为

$$N_{\tilde{c}} = \max(\min(\frac{|\mathcal{N}(\tilde{c})|}{K}, 100), 5), \tag{4.61}$$

其中 K 初始化为 20。进一步更新

$$K = \frac{|\mathcal{N}(\tilde{c})|}{N_{\tilde{c}}}. \tag{4.62}$$

语义角度预测模型通过集束搜索（Beam Search）预测。反事实回复生成模型从后验 Gumbel 噪声中采样反事实回复，温度系数设置为 0.5。

基于检索的对话模型　基于检索的模型通过微调预训练 BERT-base[28] 模型构建。训练模型的迭代次数为 2 次，批处理大小为 64，学习率为 $1e^{-5}$，并且模型输入的最大序列长度为 512，本研究采用最后一个检查点进行评估。

基于生成的对话模型　基于生成的模型是通过微调 BART-large[187] 模型构建。训练模型的迭代次数为 5 次，批处理大小为 64，学习率为 $1e^{-5}$，并且模型输入的最大序列长度为 512。在推理阶段，本研究使用 top-k 采样（k=10），最大解码长度设置为 50，并且采用最后一个检查点进行评估。

训练与评估　本研究使用 4 个 GPU 训练基于检索的对话模型、使用 8 个 GPU 训练基于生成的对话模型、使用 8 个 GPU 训练回复视角预测模型和使用 8 个 GPU 训练反事实生成模型。并且使用的是 Nvidia Tesla V100 GPU。基于检索的对话模型、基于生成的对话模型、语义角度预测模型和反事实回复生成模型的训练时间分别约为 2 小时、4 小时、4 小时和 5 小时。在数据增强阶段，预测可转换的语义角度需要 55 分钟，为所有样本生成反事实回复需要 1 小时。计算前向和后向 PPL 分数分别需要 40 分钟。

4.6.3　对比方法

本研究的方法 CAST 与一系列对比方法进行比较：

- **Observed**：它只使用观察到的数据来微调对话模型；
- **Augmented**：它只使用本研究的扩增数据来微调对话模型；
- **Back-Trans**[72]，它通过谷歌翻译回译对话回复；
- **MLM**[68]，它在 D 上微调 BERT-large 模型以替换对话回复中的一些词语。替代概率为 0.15。

- **DL**[69]，以已观测的对话上文-回复样本为基准，从非平行语料中分别找到和对话上文与回复最相似的句子，并将它们构建成新的对话上文-回复样本。扩增的数据进一步被他们的排序模块过滤。
- **BM25**[168]，利用 BM25 算法检索 top-k 个与已观测对话上文相似的对话上文，将检索到的对话上文对应的回复作为已观测对话上文的扩增回复。
- **BART**[59]，微调一系列 BART-large 模型，以对话上文作为输入去生成更多的回复。不同的模型采用不同的解码策略，其中包括贪心搜索，温度系数为 0.5 的随机采样, 和 top-k 采样 (k=10,25)。这些模型分别被表示为 **BART-gree**, **BART-samp**, **BART-k10**, and **BART-k25**。

其中由 BM25 和 BART 方法扩增的数据对由本研究提出的数据选择方法过滤。

4.6.4 评价指标

在实验中，本研究分别使用了自动指标和人工评价这两种方式来衡量对话模型的性能。

自动评价 以下指标用于自动评估基于检索的对话模型。

- **MAP**（Mean Average Precision）：测试样本的平均精度 (Average Precision，AP) 的平均值。AP 为找到参考文献的排名的平均精度分数；
- $\boldsymbol{R}_{10}@\boldsymbol{k}$：当总共给定 10 个候选回复时，前 k 个选择的回复（k=1,2,5）中出现正确回复的百分比。

以下指标用于评估基于生成的对话模型。

- **BLEU**：生成的回复和参考回复之间的重叠的 n-grams (n<4) 比例；
- **Dist-n**：在生成的响应中不一样的 n-grams (n=1,2) 种类与所有 n-grams 的比率，它衡量 n-gram 的多样性。每个测试对话上文抽取 3 个回复，将在抽取的回复内部和抽取的回复之间进行评估。相应地，Dist-n 被细分为 **Intra-Dist** 和 **Inter-Dist**。**Intra-Dist** 计算在采样回复内部计算该比率，**Inter-Dist** 在 3 个采样回复之间计算该比率。
- $\mathbf{BS}_f$: **BERTScore**[188] 的 F1 值，它衡量 3 个采样回复中每 2 个回复之间的语义相似性。分数越低表明语义更多样。

本研究还使用 **Dist-n** 和 $\mathbf{BS}_f$ 来自动评估扩增数据的质量，它们评估了扩增的反事实回复间的多样性。此外，还引入了以下指标来评估可观测样本回复与生成回复之间的多样性。

- **Novelty-n**：扩增回复中新出现的 n-grams (n=1,2) 的比率。相似地，Novelty-n 被细分为 **Intra-Novelty** 和 **Inter-Novelty**。**Intra-Novelty** 在每个扩增回复内部计算该比率，即在扩增回复中出现但不在已观测回复中出现的 n-gram 比率。**Inter-Novelty** 在 3 个扩增回复中计算该比率。
- $\mathbf{BS}_{fo}$: **BERTScore** 的 F1 值，它衡量扩增回复与其对应的已观测回复之间的语义相似性。

表 4-1 扩增数据质量的自动评估结果

Table 4-1 Automatic evaluation of the quality of augmented data

方法	Intra-Dist		Inter-Dist		BS_f	Intra-Novelty		Inter-Novelty		BS_{fo}
BART-gree	93.34	98.37	64.83	81.81	66.46	84.42	95.54	60.54	80.38	58.12
BART-samp	94.14	98.79	70.85	89.27	63.60	84.24	95.99	65.84	87.74	58.11
BART-k10	93.08	98.63	70.60	90.07	63.15	85.00	96.23	67.36	89.11	58.08
BART-k25	93.74	98.77	74.63	91.98	61.61	85.76	96.43	71.01	90.90	57.83
CAST	**94.64**	**98.90**	**79.91**	**94.79**	**59.59**	**85.84**	**96.63**	**74.47**	**92.98**	**57.31**
Observed	94.05	98.90	-	-	-	-	-	-	-	-

人工评价　人工评估扩增数据和基于生成的对话模型输出的回复的质量，主要雇佣 3 位评估人员来完成。使用如下指标进行评估：

- **流畅性 (Fluency)**: 回复的通顺，无明显语法错误
- **连贯性 (Coherence)**：回复和对话上文衔接或转换自然，是对话上文的有效延续；
- **有趣性 (Interesting)**：回复具有信息量、非通用回复；
- **丰富性（Richness)**：给定对话上文的多个回复由不同的语义角度生成。

评级范围为 0 到 2，其中 0 表示最差，2 表示最好。

4.7　实验结果与分析

本小节对实验结果进行展示和分析，包括扩增数据质量的评估结果，以及对下游任务的评估结果。此外，本小节还对提出了数据增强方法进行进一步分析。

4.7.1　评估扩增数据

本研究首先评估扩增数据的质量。首先分别从各个数据增强方法产生 $900K$ 条数据用于自动评估。然后进一步选择 600 个样本，其中包含 200 个随机选择的对话上文，每个对话上文有 3 个相应的回复，去进行人工评估。评估人员之间的一致性是通过 Fleiss 的 kappa κ [145] 指数测量的。Fluency、Coherence、Interesting 和 Richness 的 κ 值分别为 0.67（中等一致性）、0.46（中等一致性）、0.64（中等一致性）和 0.69 （中等一致性）。

实验结果如表 4-1和 4-2所示。实验结果表明本研究扩增数据的质量超过了所有对比方法产生的数据的质量。进一步观察到：（1）本研究扩增的数据和已观测到的数据具有相似的分数，这表明本研究扩增的数据是高质量的。此外，在图 4-4中展示了一些扩增数据的示例，并呈现了不同语义回复的生成过程。（2）本研究扩增的数据在指标 BS_f, BS_{fo} 和 Richness 上获得了更高的分数，这表明本研究提出的方法能够扩增出更多不同语义角度的回复。其中，CAST 相比 BART-samp 仅增加了语义转换，CAST 获得了更好的结果表明干预语义角度是有效的。

表 4-2 扩增数据质量的人工评估结果

Table 4-2 Manual evaluation of the quality of augmented data

方法	Fluency	Coherence	Interesting	Richness
BART-gree	1.921	1.507	1.222	0.611
BART-samp	1.833	1.383	1.500	0.926
BART-k10	1.853	1.461	1.506	0.983
BART-k25	1.813	1.333	1.560	1.182
CAST	**1.953**	**1.653**	**1.707**	**1.660**
Observed	1.941	1.744	1.740	-

上文关注点	语义角度	反事实回复
对话上文: I am *sleepless* because of *coughing*. (咳嗽睡不着)		
sleepless (睡不着)	Sleep (睡)	I *slept* badly. (我是没睡好。)
coughing (咳嗽)	doctor (医生)	Have you seen the *doctor* already? (去看医生了吗?)
	cold (感冒)	Honey, do you have a *cold* too? (亲爱的，你是不是也感冒了？)
对话上文: I have a *stomachache* every day. (最近每天胃痛唉)		
stomachache (胃痛)	spicy (辣的)	You can't eat *spicy* food. (不能吃辣的)
	check (检查)	You need to *check* your body. (去检查下吧)
	serious (严重)	What happened? Why is it so *serious*? (搞什么那么严重)

图 4-4 不同语义回复的生成过程的真实示例展示

Figure 4-4 Real cases showing the generation process of responses with different semantics

(3) 更进一步，与其他对比方法 (BART-gree, BART-samp, and BART-k25) 相比，BART-k10 在所有指标上都取得了相对较好的分数。这表明 top-k 采样 (k=10) 优于其他解码策略 (贪心搜索, 温度系数为 0.5 的随机采样, 和 top-k 采样 (k=25))。因此，top-k 采样（k=10）可用于后续基于生成的对话模型中。

4.7.2 评估对话模型

本研究进一步评估了提出的增强数据对基于检索和基于生成的对话模型带来的性能提升，用于验证提升对话训练数据的语义内容丰富程度，即准备更加类人的训练数据，能够提升下游任务的整体性能，不单单是对话生成的语义多样性。为了和对比方法进行公平比较，本研究为每个方法都选择 300K 增广数据，在 5K 测试数据进行自动评估，并且在 600 条样本上进行人工评估。评估人员之间的一致性是通过 Fleiss 的 kappa κ[145] 指数测量的。Fluency、Coherence、

表 4-3 基于检索的对话模型上不同数据增强方法的自动评估结果

Table 4-3 Automatic evaluation on data augmentation methods for retrieve-based models

方法	MAP	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
Observed	80.21	69.72	82.05	94.96
Augmented	76.67	65.14	78.16	92.46
MLM	80.22	69.76	82.05	94.90
Back-Trans	80.26	69.75	82.21	94.99
DL	80.47	70.05	82.41	95.03
BM25	80.07	69.68	81.62	94.82
BART-gree	80.37	70.03	82.17	94.75
BART-samp	80.42	70.17	82.03	94.88
BART-k10	80.38	70.06	82.15	94.79
BART-k25	80.53	70.30	82.21	94.91
CAST	**81.08**	**71.08**	**82.86**	**95.14**

表 4-4 基于生成的对话模型上不同的数据增强方法的自动评估结果

Table 4-4 Automatic evaluation on data-augmented methods for generation-based models

方法	BLEU	Intra-Dist-1,2		Inter-Dist-1,2		BS_f
Observed	2.22	91.11	98.21	73.83	93.18	60.54
Augmented	1.85	92.29	98.16	77.86	93.28	59.76
MLM	2.16	91.19	98.25	74.41	93.37	60.50
Back-Trans	2.21	91.26	98.26	74.66	93.49	60.45
DL	2.23	92.09	98.35	75.02	93.42	60.35
BM25	1.68	91.55	98.14	76.51	92.02	60.17
BART-gree	**3.54**	91.54	98.02	64.79	80.87	67.18
BART-samp	2.86	92.12	98.42	69.81	88.91	63.51
BART-k10	2.72	91.71	98.53	70.51	90.02	63.45
BART-k25	2.70	91.93	98.45	71.29	90.46	62.81
CAST	2.11	**93.39**	**98.67**	**78.03**	**93.62**	**59.64**

Interesting 和 Richness 的 κ 值分别为 0.71（高度一致性）、0.59（中等一致性）、0.48（中等一致性）和 0.32 （中等一致性）。

数据增强方法在基于检索的模型上评估的结果在表格 4-3中展示，而在基于生成的模型上的自动评估和人工评估结果分别在表格 4-4和 4-5中展示。实验结果表明本研究提出的方法在两个对话模型上在几乎所有的评价指标上都超过了基准模型。这说明了丰富对话训练数据的语义内容对提升下游任务的整体质量都是有效的。进一步观察发现：（1）和 BART 系列模型相比，CAST 获得了更高的分数，尤其是 BART-samp。这表明了干预回复角度能够有效地提高下游任务的性能。（2）CAST 在 BLEU 上获得了更低的分数，这可能是因为扩增更多语义

表 4-5 基于生成的对话模型上不同的数据增强方法的人工评估结果

Table 4-5 Manual evaluation on data-augmented methods for generation-based models

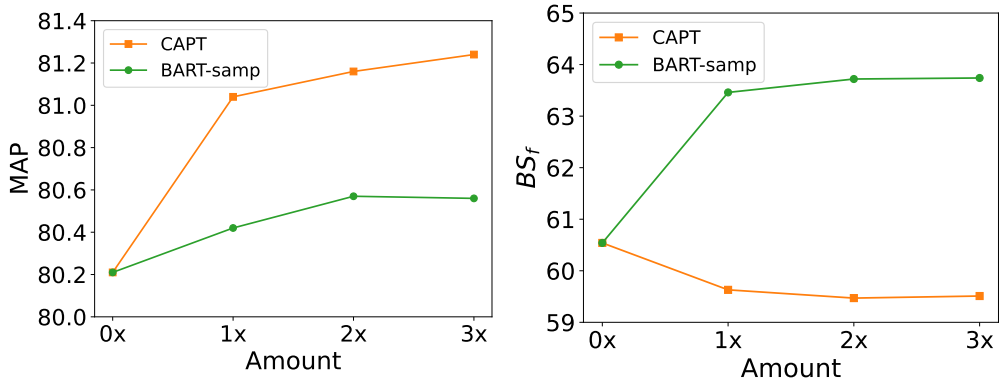
方法	Fluency	Coherence	Interesting	Richness
Observed	1.806	1.377	1.645	1.075
Augmented	1.848	1.363	1.652	1.320
MLM	1.813	1.438	1.653	1.095
Back-Trans	1.791	1.443	1.657	1.115
DL	1.823	1.462	1.665	1.135
BM25	1.803	1.155	1.650	1.185
BART-gree	1.841	1.453	1.508	0.895
BART-samp	1.822	1.448	1.582	0.910
BART-k10	1.835	1.480	1.584	0.925
BART-k25	1.812	1.425	1.623	0.935
CAST	**1.867**	**1.492**	**1.677**	**1.355**

不同的样本数据使得对话模型产生的回复越不像参考回复，从而生成回复和参考回复间的词重叠率就会下降。（3）但 CAST 在指标 BS_f 和 Richness 获得了更高的分数，这表明了丰富对话训练数据的语义内容很大程度上能够直接提升对话生成的语义多样性。

4.7.3 实验分析

进一步，本研究探索了增广回复数量对下游任务的影响以及 CAST 每个组成部分的重要性。此外，本研究还探索了目前大规模对话生成模型在对话生成任务上的表现，并与全量微调模型进行了对比。

增广回复数量影响 本研究选择 0x、1x、2x、3x 的训练样本量来评估提供更多不同语义的回复对下游任务的影响，并将 CAPT 与基线模型（即 BART-samp）进行比较。请值得注意的是，3x 表示选择了 $3*300K$ 的扩增样本。考虑到按顺序选择的样本具有不同的有趣程度，本实验通过统一选择 900K 个增广样本并从中随机选择来消除有趣程度的影响。结果如图 4-5所示。经过观察发现：(1) BART-samp 上的 MAP 分数在 2x 时达到峰值，然后下降，并且在 BS_f 分数从 0x 增加到 3x 也是不断增加的。这可能是因为 BART-samp 仅输出具有词语级别多样的可替代表达，具备有限的语义变化。因此，大量增加相似的样本反而会对模型训练产生负面影响。(2) 相反，CAST 上的 MAP 分数从 0x 增加到 3x 是在不断增加，而 BS_f 分数在 0x 增加到 3x 时并没有增加。这表明本研究提出的方法可以扩增具有不同语义的回复，通过提高原始训练数据的语义内容丰富度去进一步提高下游任务的整体性能。



(a) 基于检索的对话模型　　(b) 基于生成的对话模型

图 4-5 不同数量的增广数据对对话模型性能的影响

Figure 4-5 Performance changes on dialogue models with different-amount augmented data

表 4-6 在基于检索的对话模型上关于 CAST 不同组件的消融实验

Table 4-6 Ablation study on different components of CAST on retrieve-based dialogue models

方法	MAP	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
CAST	81.08	71.08	82.86	95.14
-Predictor	80.63	70.33	82.65	94.96
-Candidate	80.22	69.90	82.01	94.62
-Selection	80.41	69.92	82.44	95.08
-Dial PLM	80.52	70.19	82.39	94.98
-Back PPL	80.68	70.41	82.51	95.07
-Gumbel	80.83	70.62	82.76	95.02

消融性实验　本研究进行以下消融测试来验证每个组件的效果：（1）移除语义角度预测模型，仅从候选集合中随机选择一个关键词作为语义角度（-Predictor）；（2）只将对话上文和上文关注点作为语义角度预测模型的输入，没有 1-hop 邻居，即观测数据中出现过的从对话上文关注点合理转移而来的语义角度作为候选集合（-Candidate）；（3）不使用数据选择方法过滤增广数据（-Selection）；（4）用没有见过足够多对话样本的通用预训练语言模型 GPT2 来替换大规模对话预训练语言模型 DialoFlow（-Dial PLM）；（5）仅使用前向 PPL 分数过滤掉低质量样本，不采用后向 PPL 分数进行排序选择（-Back PPL）；（6）不在当前环境下生成回复，即回复生成时不考虑后验 Gumbel 噪声（-Gumbel）。消融实验结果如表 4-6 所示。经过观察发现，移除每个组件都会带来不同程度的性能下降。这证明了 CAST 中设计所有这些组件都是有必要的。

大模型提示方法 vs. 全量微调方法　本研究进一步对比了基于大模型（Large Language Model）提示学习的开放域对话生成和基于全量微调的开放域对话生成。本实验选择开源大模型 Vicuna-13B，通过设置系统消息的方式告诉模型的

表 4-7 全量微调方法和大模型提示方法对比

Table 4-7 Comparison of full fine-tuning method and prompt learning of LLM

方法	BLEU	Intra-Dist-1,2		Inter-Dist-1,2		BS_f
$Vicuna_{p=0.5}$	1.05	81.23	96.07	44.62	62.17	79.54
$Vicuna_{p=0.75}$	0.98	81.38	96.06	52.65	76.17	71.34
$Vicuna_{p=1}$	0.95	81.93	96.20	57.19	82.73	66.99
$Vicuna_{\tau=0.5}$	1.01	81.39	96.12	54.02	78.46	69.89
$Vicuna_{\tau=0.75}$	0.86	81.80	96.08	57.87	84.05	65.93
$Vicuna_{\tau=1}$	0.74	81.81	96.10	61.49	88.11	62.14
Obersved	2.22	91.11	98.21	73.83	93.18	60.54
CAST	**2.11**	**93.39**	**98.67**	**78.03**	**93.62**	**59.64**

任务是和用户聊天，并提供对话上文去引导模型生成回复。本实验探究了解码过程中温度系数（temperature，τ）和 Top-p 中 p 的取值对生成结果的影响，其中温度系数和 p 分别取值 0.5, 0.75, 1。温度系数和 p 值越小，生成的回复越确定保守；取值越大，生成的回复越随机多样。此外，本实验选择全量微调方法 Observe 和 Ours 进行对比。实验结果在 4-7中给出，结果表明基于大模型提示学习的方法在所有指标上都没有超过基于全量微调的方法。进一步观察发现：（1）基于大模型提示学习方法生成的回复的 BLEU 值明显低于全量微调方法的 BLEU 值。这主要是因为 BLEU 计算生成回复与参考回复之间的词重叠率，而大模型的训练数据和测试数据分布完全不一样，大模型在没有用领域数据微调时生成的回复和参考回复词重叠率低使正常现象。（2）基于大模型提示学习方法在 Intra-Dist 指标上同样不如全量微调方法。这可能是因为大模型在经过和人类对齐后更倾向于生成冗长的回复。过长的回复会导致 Intra-Dist 值偏低。（3）在 Inter-Dist 指标上，基于大模型提示学习方法同样不如全量微调方法。这可能是因为目前的大模型的定位是智能 AI 助手，用于解答用户的问题，并非是模仿人类的社交机器人。因此在开放域对话场景中，大模型在经过和人类对齐后针对日常生活场景比如用户询问工作、家人、宠物、以及娱乐方式等，会更倾向于强调 AI 助手没有工作、家人，不能养宠物等，而不是多样地回答。这可能极大地限制了开放域对话场景下的多样性。此外，大模型在和人类对齐后，会偏好生成某一类回复，这可能也会限制多样性。（4）同样地，在 BERTScore 指标上给予大模型提示学习方法效果不理想，可能也是基于上述原因。

4.8 小结

本章提出了一种基于语义转换的反事实数据增强方法 CAST，以针对给定的对话上文扩增更多具有不同语义的回复，用于丰富已有训练数据的语义内容。具体来说，CAST 借助反事实推理，通过干预观察到的生成回复的语义角度来生成反事实回复，通过转换成可替换的语义角度去生成不同语义的回复。首先，当前

环境由 Gumbel-Max SCM 中的后验 Gumbel 噪音来建模。其次，为了获得有效的语义角度，该方法基于所有观测对话数据构建一个转移关系图，它显式地构建了人类在对话上文中的关注点与其相应的可生成回复的语义角度间的转移关系。通过从给定的对话上文中随机选择一个关注点，并将该关注点在所有观测数据中合理转移到的语义角度，作为候选集合。随后从候选集合中预测出一个有效的语义角度。最终利用得到的有效的语义角度去扩增不同语义的回复。在得到所有反事实回复后，进一步使用数据选择模块去过滤增广数据。最后，通过将已有训练数据与所有的扩增数据混合，作为下游任务的训练数据。实验结果表明，本研究提出的方法可以扩增具有不同语义的高质量回复，并且使用本文丰富语义内容后的训练数据能够有效地提升对话生成的语义多样性。此外，使用这种更加类人的训练数据不仅能够提升对话生成任务，还能提升对话检索任务的整体质量。

第 5 章　基于语义分解的对话一致性检测方法

5.1　引言

开放域对话系统生成的回复要满足和对话上文中的语义内容不能自相矛盾（即保持一致）的约束。现有系统主要利用一致性检测模型对生成回复进行一致性评分，通过输出得分最高的回复来满足约束。目前一致性检测模型主要通过人工编写的相关领域数据来训练，但这样训练的模型检测对话系统产生的数据时出现了性能下降。在这种数据分布差异下出现的性能下降表明模型鲁棒性不够好。目前其他 NLP 领域提升鲁棒性的常用思路是构造反事实样本。然而如何针对对话一致性检测任务构造反事实样本还未被探索。因此该研究提出一种适用于对话一致性检测任务的反事实样本构造方法。该方法首先辨认出所有不一致样本中相互矛盾的内容。然后对于不一致样本，通过删除其中相互矛盾的内容去构造对应的反事实样本；对于一致样本，通过添加相互矛盾的内容去构造对应的反事实样本。本研究在两个广泛使用的对话一致性检测模型上进行了实验，并和其他方法进行了对比，结果表明本研究提出的方法能够有效地提升检测模型在数据分布差异下的鲁棒性。

5.2　概述

开放域对话系统[113,189]在生成流畅且有信息量的回复方面已取得显著进步，但在一致性方面仍存在不足[127]。为了获取用户对对话系统的长期信任和好感，生成回复和给定对话上文中语义内容不能相互矛盾（即保持一致）是至关重要的。因此对话一致性检测任务（Dialogue Contradiction Detection）[127,190,191]逐渐受到了研究者们的广泛关注。目前对话系统通过额外引入一致性检测模型对生成回复进行一致性评分，然后输出得分最高的回复来满足一致性约束。已有研究主要把一致性检测任务建模成自然语言推理（Natural Language Inference，NLI）任务[192]，然后人工编写领域相关的 NLI 数据去训练检测模型。然而当这样训练的模型用于真实场景时，即检测对话系统产生的数据是否包含相互矛盾内容时，出现了性能下降[127]。这种数据分布差异下的性能下降表明模型鲁棒性差。许多研究[193–195]已经证实模型鲁棒性差和模型倾向于探索训练数据中的伪关系有关。所谓的伪关系是指训练集中某些内容片段和标签之间的高频共现关系。举例来说，假如训练集中“冲突”标签和一对以“我是”开头的句子经常共现（如图 5-1 A1 中“我是莎拉”和 A2 中“我是不会画太多头像的”），测试时模型只要看到对话中有一对以“我是”开头的句子就会输出“不一致”，而不管是否真的包含相互矛盾内容。

目前其他 NLP 领域提升鲁棒性的可行思路是构造反事实样本[196]。反事实样本是指和原始样本尽可能相似但标签相反的样本。如图 5-1所示，当该样本移

A1：我是莎拉！我喜欢画画。
B1：我也喜欢，尤其是画头像。
A2：③我是不会画太多头像的①④因为不擅长。②我最近专注于风景图。哈哈，你呢？
B2：我曾经也画过风景画，包括城市建筑和海洋风景。
A3：我可以做这些尽管①我最擅长画头像。②③④我以有偿画头像为生。

图 5-1 不一致对话示例

Figure 5-1 An example of contradictory dialogue

除掉红色内容“因为不擅长”和“我以有偿画头像为生”，则转换成了对应的反事实样本。当混合反事实样本与原始样本训练模型时，伪关系如“不一致”标签和一对以“我是”开头的句子间的高频共现关系将被消除，从而可以缓解模型对伪关系的依赖。然而目前如何针对对话冲突检测任务去构造反事实样本还未被探索。

因此，本研究提出一种反事实样本构造方法用于提升检测模型在数据分布差异下的鲁棒性（Robustness Improvement via Counterfactual Samples，RICS），进而更好地满足语义内容一致性约束。从反事实样本的定义来看，构造反事实样本首先需要反转标签，即使得原始不一致的样本不再有相互矛盾的内容，使原始一致的样本包含相互矛盾的内容；其次要保证和原始样本尽可能相似，即进行最小修改即可。为此，本研究首先辨认出所有不一致样本中相互矛盾的内容（如图 5-1中成对的相互矛盾的内容，由 ①②③④ 标记所示）。为了自动标记出完整且简洁的相互矛盾的内容，本研究提出将对话话语分解成多个独立的语义内容单元，通过辨别出相互矛盾的语义内容单元来标记矛盾的内容。然后对于不一致的样本，通过删除原始样本中能使标签反转的最少相互矛盾的内容去得到反事实样本；对于一致的样本，通过向原始样本中添加一对相互矛盾的内容去得到反事实样本。最终，混合反事实样本和原始样本去训练现有的冲突检测模型。该方法能自动化地构造反事实样本，能够简单有效地提升模型在真实场景下的鲁棒性。

在开放域对话一致性检测数据集 DECODE[127] 上的实验结果表明，本研究提出的方法能够有效构造反事实样本去提升检测模型在真实使用场景下的性能，即提升模型在数据分布差异下的鲁棒性。此外，分析实验也进一步表明本研究构造的反事实样本能够缓解一致性检测模型对伪关系的依赖，更多关注到真实矛盾内容去做决策。

本章组织如下，第 5.3节中介绍对话一致性检测任务的相关工作；第 5.4节中介绍对话一致性检测任务定义，以及代表性检测模型；第 5.5节中介绍了本研究提出的基于话语分解的反事实样本构造方法；第 5.6节中介绍了本工作相关的实验设置，包括使用的数据集情况，对比的基线模型，使用的评价指标，和实验细节等；第 5.7节中给出了本工作的实验结果，并进行了相关对比与分析。最后对本章进行了总结。

5.3　相关工作

对话一致性检测　为了对话系统产生的回复是否存在冲突，许多研究者都进行了探索。早期工作主要关注对话回复是否和提供的个性化信息冲突。 Zhang 等[14] 和 Zheng 等[197] 分别提出包含个性化信息的数据集 PersonaChat 和 PersonalDialog，旨在帮助对话系统具备一定的人设。 Welleck 等[125] 和 Song 等[126] 分别构造了相应的一致性检测数据。他们使用人工构造的数据训练检测模型，然后再用检测模型自动输出每个回复的一致性得分，选择高评分的回复以保证对话系统的人设一致性。随后， Nie 等[127] 构造了开放域对话领域的冲突检测数据集 DECODE，该数据集包括人工编写的对话，其中标注人员在对话中的某个时刻故意编写与他们之前所说的内容相矛盾的话语。此外，他们还收集了第一个真实场景下通用对话系统产生的冲突数据作为测试集。 Qin 等[198] 进一步构造了任务型对话领域的冲突检测数据集。所有这些工作都关注人工构造特定领域的对话一致性检测数据，而本研究关注提升现有一致性检测模型在数据分布差异下的鲁棒性。

使用场景下检测模型性能提升　Jin 等[199] 指出对话话语中频繁出现的指代和省略问题会阻碍一致性检测模型充分理解对话话语的上下文内容，从而导致检测错误。因此，他们提出通过改写不完整的话语来消除指代和省略现象，进而提升检测模型在真实场景下的性能。相反，本工作通过构造反事实样本缓解模型对伪关系的依赖，去提升模型在真实场景下的性能。本研究和此工作是相互独立和不可互相替代的。实验证实了在此方法基础上，本研究提出的方法仍能提升模型在真实场景下的性能。

对话冲突诱导　还有研究者关注如何诱导对话系统产生冲突[190,191]。由于对话系统更多时候是在不断谈论不同的事情，并不会时常产生冲突。因此为了量化对话系统的一致性，需要诱导系统去谈论已经谈论过的事情。Li 等[190] 和 Honovich 等[191] 采用相似的思路，即借助自动化问题生成（Question Generation，QG）随时在对话系统交互过程中针对特定对话系统提及的实体提问，诱导模型去谈论重复的事情。不同的是，本工作假设待检测的数据已给定，去提升已知的对话一致性检测模型在真实场景下的鲁棒性。

反事实样本构造　目前已有一些 NLP 任务探究了如何自动化构造反事实样本，包括情感分析[200,201]、性别偏见[202,203] 和视觉问答[204,205] 等。这些任务的构造思路同样是找到对模型决策影响最大的内容，通过修改这些内容去得到反事实样本。区别在于这些内容是文本中的某个词语或图片中的某个物体，它们在文本或图片中很容易标记出来。相反，影响冲突检测任务决策的是包含了多个可能不连续出现的词语或者短语片段的冲突信息，这种不连续出现使得自动化标记冲

突极为困难。因此，本工作设计了一种专门适用于对话一致性检测任务的冲突内容辨认的方法。

冲突内容辨认 冲突内容辨认方法的设计灵感来自于 Welleck 等[125] 人的工作。该工作主要利用事实三元组标注句子间是否冲突。具体来说，对于给定的句对，首先人工标注各自所传达的事实三元组，然后通过判断事实三元组是否相互冲突来判断句对是否相互冲突。三元组相互冲突能够推出它们所关联的句子相互冲突，这表明冲突主要发生在事实观点之间。受其启发，本文提出了一种可行的冲突内容辨认方法，即先从话语中抽取出事实观点三元组，再辨认哪些相互冲突。不同之处在于，该工作人工标注三元组，且三元组中实体与关系都是预定义的；而本文工作自动化抽取三元组和判断三元组是否相互冲突，且考虑到对话的多样性，并没有预定义三元组中的实体与关系。

5.4 背景知识

本小节首先描述对话一致性检测任务的定义，然后介绍目前广泛使用的对话一致性检测模型。

5.4.1 任务定义

对话一致性检测旨在通过建模输入对话和输出间的映射来判断对话是否包含相互矛盾的语义内容。任务的输入是一组话语 $\boldsymbol{X} = \{\boldsymbol{U}_0, \ldots, \boldsymbol{U}_i, \ldots, \boldsymbol{U}_{m-1}, \boldsymbol{U}_m\}$，表示一个对话或对话片段。输出为 $\boldsymbol{Y}$，用于表明最后一句话语 $\boldsymbol{U}_m$ 是否和对话上文 $\{\boldsymbol{U}_0, \ldots, \boldsymbol{U}_i, \ldots, \boldsymbol{U}_{m-1}\}$ 中任何信息相互矛盾。其中 $\boldsymbol{Y}$ 为 1 或 0，分别代表矛盾或不矛盾。该任务需要学习映射 $P_\theta(\boldsymbol{Y}|\boldsymbol{X})$，$\theta$ 表示模型参数。

5.4.2 对话一致性检测模型

目前广泛使用的对话一致性检测模型[127] 主要采用以下两种方法，非结构化方法（Unstructured Approach）和基于话语的结构化方法（Structured Utterance-based Approach）来学习预测 $P_\theta(\boldsymbol{Y}|\boldsymbol{X})$。

非结构化方法 非结构化方法把所有对话上文 $\{\boldsymbol{U}_0, \ldots, \boldsymbol{U}_i, \ldots, \boldsymbol{U}_{m-1}\}$ 拼接成单一文本和最后一句话语 $\boldsymbol{U}_m$ 一起作为输入 $\boldsymbol{X}$，把对话是否不一致的标签作为输出 $\boldsymbol{Y}$，其中每句话语前插入特殊符号用于标记不同的说话人。然后训练模型学习处理 X 得到隐层表示 $\mathbf{H}$, 通过分类器预测得到 $\boldsymbol{U}_m$ 和对话上文相互矛盾的概率，即

$$P_\theta(\boldsymbol{Y}|\boldsymbol{X}) = \text{Classifier}(\mathbf{H}), \tag{5.41}$$

$$\mathbf{H} = \text{Encoder}([\boldsymbol{U}_0, \ldots, \boldsymbol{U}_i, \ldots, \boldsymbol{U}_{m-1}], \boldsymbol{U}_m). \tag{5.42}$$

基于话语的结构化方法 不同于非结构化方法，基于话语的结构化方法采用依次判断最后一句话语和对话上文中每句话语是否相互矛盾的方式来预测标签。由

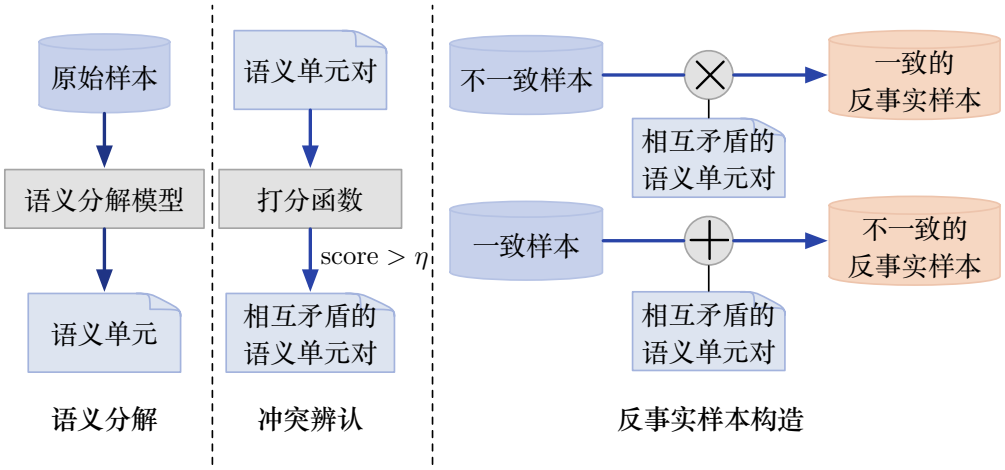


图 5-2 基于语义分解的反事实样本构造框架图

Figure 5-2 Framework of counterfactuals construction based on semantic decomposition

于矛盾发生在相同说话人之间，因此仅考虑对话上文中和 $\boldsymbol{U}_m$ 同属一个说话人的那些话语即可。这些话语形成集合 $\mathcal{U}$。从而模型的输入 $\boldsymbol{X}$ 变为 $\boldsymbol{U}$ 中的一句话语 $\boldsymbol{U}_i$ 和 $\boldsymbol{U}_m$ 拼接形成的单一文本，输出 $\boldsymbol{Y}$ 变为 $\boldsymbol{U}_i$ 和 $\boldsymbol{U}_m$ 是否相互矛盾的标签。然后训练模型去预测 $\boldsymbol{U}_m$ 和 $\boldsymbol{U}_i$ 相互矛盾的概率，即

$$P_\theta(\boldsymbol{Y}|\boldsymbol{X}) = \text{Classifier}(\mathbf{H}), \tag{5.43}$$

$$\mathbf{H} = \text{Encoder}(\boldsymbol{U}_i, \boldsymbol{U}_m). \tag{5.44}$$

因此，在推理阶段，给定待检测对话，模型首先计算每一个 $P_\theta(\boldsymbol{Y}|[\boldsymbol{U}_i, \boldsymbol{U}_m]), 0 \leq i < |\mathcal{U}|$，最终对话不一致的概率通过计算 $\max\{P_\theta(\boldsymbol{Y}|[\boldsymbol{U}_i, \boldsymbol{U}_m])\}$ 得到。

Encoder 和 Classifier 可以采用任意模型结构。在训练阶段，模型参数 θ 通过优化负对数似然损失函数（NLL）学习，即

$$\mathcal{L} = -\sum[\boldsymbol{Y}\log_\theta(\boldsymbol{Y}=1|\boldsymbol{X}) + (1-\boldsymbol{Y})\log P_\theta(\boldsymbol{Y}=0|\boldsymbol{X})]. \tag{5.45}$$

此外，模型通过比较预测的对话矛盾概率和阈值 η 来检测不一致，其中 η 为 0.5。

5.5　基于语义分解的反事实样本构造

本节旨在构造和给定样本 $(\boldsymbol{X}, \boldsymbol{Y})$ 相对应的反事实样本 $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$。构造反事实样本的重点是标记相互矛盾的内容。一种简单的方式是标记相互矛盾的话语。但话语通常会表达多个信息，而矛盾发生在表达某个信息时，如图 5-1 A2 和 A3 中“我不擅长画头像”和“我擅长画头像”。以话语为单位标记相互矛盾的内容无法保证构造样本时修改最少的内容。因此需要以信息为单位标记更简洁的不一致内容。然而话语在表达多个语义信息时，如果后表达信息中某个内容在前面信息中已出现，就容易出现省略。如 A2 所示，“我不擅长画头像”中“我”和“画头像”在前一个信息“我是不会画太多头像的”中已出现，就发生了省略。这表明一个表达完整的信息可能由句中多个没有连续出现的词或者短语表达，这

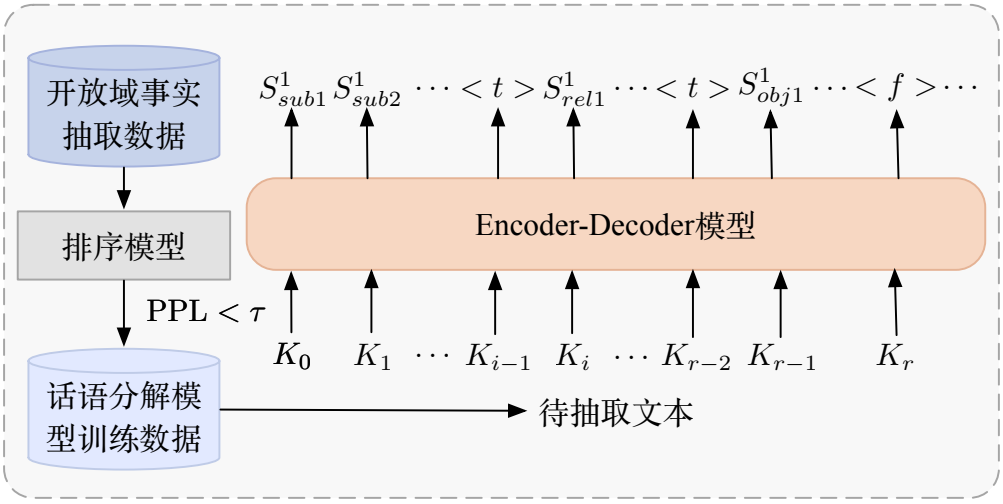


图 5-3 语义分解模型训练

Figure 5-3 Training of semantic decomposition

使得标记出完整的矛盾信息很有挑战。受 Welleck 等[125] 启发，发现发生相互矛盾的信息主要是传达的诸如事实信息的语义单元。如果能从话语中分解出每一个表达完整的语义单元，然后从中判断哪些语义单元相互矛盾即可标记完整且简洁的不一致内容。为此， 5.5.1中提出语义分解方法去抽取话语中的语义单元。5.5.2中介绍了如何辨认相互矛盾的语义单元。 5.5.3中提出了通过删除或添加冲突的语义单元的方式去构造反事实样本。构造方法流程如图 5-2所示。

5.5.1 语义分解

本节旨在抽取给定样本 $\boldsymbol{X}$ 的每句话语 $\boldsymbol{U}_i$ 中表达完整的语义单元。表达完整的语义单元由关系三元组（主体；关系；客体）来定义[125]。举例来说，图 5-1 中 A2 传达的语义单元包括:（我；是不会画；太多头像的），（我；不擅长；画头像），（我；最近专注于；风景图）。因此，本研究实现一个端对端的语义分解模型用于抽取语义单元。模型结构及训练流程如图 5-3所示。

语义分解模型是一个序列到序列（Seq2Seq）的生成模型。模型输入为话语 $\boldsymbol{U}_i$，输出为语义单元序列 $\boldsymbol{S}_i$，即 “我 <t> 是不会画 <t> 太多头像的 <f>…”。<t> 和 <f> 分别用于分隔三元组中的成分和不同的三元组。此外，采用 Transformer 编码器-解码器结构。在编码阶段，模型通过 Transformer Encoder 将话语 $\boldsymbol{U}_i$ 编码成隐层向量表示 $\mathbf{H}_i^e$，即

$$\mathbf{H}_i^e = \text{Encoder}(\boldsymbol{U}_i). \tag{5.51}$$

在第 j 个解码时间步，Transformer Decoder 预测序列下一个词语在词表上的概率分布，即

$$\boldsymbol{H}_{i,j}^d = \text{Decoder}(e(\hat{\boldsymbol{S}}_{i,j-1}), \mathbf{H}_i^e, \mathbf{H}_{i,0:j-1}^d), \tag{5.52}$$

$$P(\hat{\boldsymbol{S}}_{i,j}|U_i, \hat{\boldsymbol{S}}_{i,0:j-1}) = \text{softmax}(\boldsymbol{H}_{i,j}^d \cdot \boldsymbol{O}), \tag{5.53}$$

其中 $e(\cdot)$ 表示词语的嵌入向量表示，$\boldsymbol{H}_{i,j}^{d}$ 为 Decoder 第 j 时间步的隐层向量表示，$e(\hat{\boldsymbol{S}}_{i,j-1})$ 为前一时间步输出的词语。$\boldsymbol{O}$ 为输出变换矩阵。最终通过以下方式得到 $e(\hat{\boldsymbol{S}}_{i,j-1})$：

$$\hat{\boldsymbol{S}}_{i,j} = \arg\max_{\hat{\boldsymbol{S}}_{i,j}}(\log P(\hat{\boldsymbol{S}}_{i,j}|\boldsymbol{U}_i, \hat{\boldsymbol{S}}_{i,0:j-1})). \tag{5.54}$$

为了训练语义分解模型，需要大量的话语-语义单元平行语料。然而该领域目前没有相应的语料。理想情况是人工标注这样的训练语料，但是耗时耗力。考虑到对话中语义单元中提及的关系是无法全部预定义的，本文借助大规模开放域知识抽取数据 IMoJIE[206] 来训练语义分解模型。但开放域知识抽取数据和对话数据的数据分布并不匹配，直接使用也会导致性能不佳。为了获得与该领域相似的数据，本研究采用一种检测域外数据的通用技术。首先用原始数据中的所有话语来微调预训练语言模型 GPT[56]，然后用该模型计算事实抽取数据中每条待抽取文本 $\boldsymbol{K}$ 的困惑度（Perplexity，PPL），即

$$\mathrm{PPL}(\boldsymbol{K}) = \sqrt[r]{\prod_{i=1}^{r} \frac{1}{\mathrm{P}\left(K_i \mid K_0, \ldots, K_{i-1}\right)}}, \tag{5.55}$$

其中 K_i 为 $\boldsymbol{K}$ 中的第 i 个词语，r 为样本长度。随后将这些文本按照它们的 PPL 从低到高排序，选择 PPL 小于阈值 τ 的文本对应的事实抽取数据。为了得到最优的 τ，首先从给定对话数据和开放域事实抽取数据各自的验证集中选择等量的话语和待抽取文本，然后计算它们的 PPL。随后选择一个数值，使得上述数据中 PPL 小于该数值的所有数据里话语的正确率和召回率综合最优，该数值即为最优的 τ。

最终，将 $\boldsymbol{X}$ 中每句 $\boldsymbol{U}_i$ 依次送入到语义分解模型中，得到对应的语义单元序列 $\boldsymbol{S}_i$。考虑到后续操作均针对自然语言文本进行，将 $\boldsymbol{S}_i$ 转化成自然语言表达的语义单元。由于语义单元中的关系短语是无固定类别的，所以可以直接按顺序拼接三元组中的成分去表示，如“我是不会画太多头像的。”。如果抽取的三元组成分缺失，则过滤该语义单元。因此，话语 $\boldsymbol{U}_i$ 传达的语义单元表示为 $\boldsymbol{T}_i = \{\boldsymbol{T}_{i,0}, \ldots, \boldsymbol{T}_{i,j}, \ldots, \boldsymbol{T}_{i,n}\}$, $\boldsymbol{T}_{i,j}$ 为 $\boldsymbol{U}_i$ 的第 j 个自然语言表达的语义单元。从而 $\boldsymbol{X}$ 包含的语义单元为 $\{\boldsymbol{T}_i\}_{i=0}^{m}$。

5.5.2　矛盾语义单元辨认

接下来标记不一致样本 $\boldsymbol{X}$ 中哪些语义单元是相互矛盾的。考虑到检测的是最后一句话语 $\boldsymbol{U}_m$ 是否和对话上文中内容相互矛盾，并且矛盾发生在相同说话人之间，因此仅判断 $\boldsymbol{U}_m$ 中语义单元与上文同属相同说话人话语中的语义单元是否相互矛盾即可。为此，本文用 MultiNLI 数据集[207] 微调预训练模型 RoBERTa[29] 用作矛盾打分函数 $f(\cdot,\cdot)$。该数据包含一对句子和它们是否相互矛盾的标签，其中原始标签“蕴含”和“中立”转换成“不矛盾”，原始标签“矛盾”保持不变。理想情况是用对话领域的语义单元一致性检测数据，然而目前没有相应的数据，所以跟随 Li 等[190] 使用大规模通用 NLI 数据。

	原始样本：	反事实样本：
冲突样本	U_0：我是莎拉！我喜欢画画。 U_1：我也喜欢画画，尤其是画头像。 U_2：我是不会画太多头像的因为不擅长。我最近专注于风景图。哈哈，你呢？ U_3：我曾经也画过风景画，包括城市建筑和海洋风景。 U_4：我可以做这些尽管我最擅长画头像。我以有偿画头像为生。	U_0：我是莎拉！我喜欢画画。 U_1：我也喜欢画画，尤其是画头像。 $\widetilde{U}_2$：我是不会画太多头像的。我最近专注于风景图。 U_3：我曾经也画过风景画，包括城市建筑和海洋风景。 $\widetilde{U}_4$：我可以做这些。我最擅长画头像。
非冲突样本	**原始样本：**	**反事实样本：**
	U_0：你好。 U_1：我刚结束我的汽车销售工作。 U_2：我也刚下班。我在一家律师事务所做文员。我很喜欢我的工作！ U_3：我也喜欢我的工作，但也有令人糟心的时候。	U_0：你好。 $\widetilde{U}_1$：我刚结束我的汽车销售工作。我晚上通常和我的宠物狗一起散步。 U_2：我也刚下班。我在一家律师事务所做文员。我很喜欢我的工作！ $\widetilde{U}_3$：我从来没养过宠物。我也喜欢我的工作。我的工作也有令人糟心的时候。

图 5-4 构造的反事实样本示例

Figure 5-4 Examples of Constructed Counterfactual Sample

给定 $\boldsymbol{X}$，依次拼接 $\boldsymbol{U}_m$ 的每个语义单元 $\boldsymbol{T}_{m,k}$ 和上文同属相同说话人的每句话语 $\boldsymbol{U}_i$ 的每个语义单元 $\boldsymbol{T}_{i,j}$ 并送入模型，由模型输出矛盾的概率 $f(\boldsymbol{T}_{m,k},\boldsymbol{T}_{i,j})$，$(0 \le i \le m-1, 0 \le k,j \le n)$。如果概率大于阈值 $\tau = 0.5$ 就认为是不一致的。最终辨认出样本中所有的矛盾语义单元 $<\boldsymbol{T}_{m,k},\boldsymbol{T}_{i,j}>$，将其组成集合 $\mathcal{C}$。

5.5.3 反事实样本构造

本节将介绍如何通过删除或添加相互矛盾的语义单元去构造反事实样本。构造反事实样本需要保证最小修改原则，因为这样才能消除样本中尽可能多的伪关系，并且给模型提供尽可能准确的真实矛盾内容[201]。此外，样本还需要保证句子表达完整，因为明显的内容缺失会给模型训练带来噪音[208]。接下来，分别介绍如何构造不一致和一致样本对应的反事实样本。图 5-4以非结构化方法的训练样本为例，给出构造的反事实样本示例。

给定不一致样本 $(\boldsymbol{X}, \boldsymbol{Y}=1)$，删除样本中最少的矛盾语义单元使标签反转即可构造对应的反事实样本。第一步确定要删除的语义单元集合。具体来说，首先找到 $\boldsymbol{X}$ 的相互矛盾的语义单元对集合 $\mathcal{C}$ 中出现次数最多的语义单元 $\boldsymbol{T}_{i,j},(0 \le i \le m, 0 \le j \le n)$，然后从 $\mathcal{C}$ 中移除包含 $\boldsymbol{T}_{i,j}$ 的矛盾语义单元对 $<\cdot,\boldsymbol{T}_{i,j}>$ 或 $<\boldsymbol{T}_{i,j},\cdot>$。重复上述过程，直至 $\mathcal{C}$ 变空为止。其中每一步找到的语义单元 $\boldsymbol{T}_{i,j}$ 组成集合 $\tilde{\mathcal{C}}$。第二步从 $\boldsymbol{X}$ 中删除 $\tilde{\mathcal{C}}$ 中所有 $\boldsymbol{T}_{i,j}$。首先确定每个 $\boldsymbol{T}_{i,j}$ 所属的话语 $\boldsymbol{U}_i,(0 \le i \le m)$。直接从 $\boldsymbol{U}_i$ 中删除 $\boldsymbol{T}_{i,j}$，需要删除 $\boldsymbol{T}_{i,j}$ 中每个单词，这会导致移除后的话语表达不完整。因为 $\boldsymbol{U}_i$ 表达某个语义单元时若其中内容已在前文出现可能就会被省略，比如图 5-4 $\boldsymbol{U}_2$ 中“我不擅长画头像”的“我”和“画头像”被省略，从 $\boldsymbol{U}_2$ 中删除该语义单元会导致前文出现相应内容缺失。因此，将 $\boldsymbol{U}_i$ 用 $\boldsymbol{T}_i = \{\boldsymbol{T}_{i,0},\ldots,\boldsymbol{T}_{i,j},\ldots,\boldsymbol{T}_{i,n}\}$

表示，从中移除 $\boldsymbol{T}_{i,j}$ 即可。移除后的话语 $\tilde{\boldsymbol{U}}$ 为其余语义单元 $\boldsymbol{T}_i$ $\boldsymbol{T}_{i,j}$ 拼接而成的文本序列。最终，反事实样本为 $(\tilde{\boldsymbol{X}} = \{\boldsymbol{U}_0, \ldots, \tilde{\boldsymbol{U}}_i, \ldots, \tilde{\boldsymbol{U}}_m\}, \tilde{\boldsymbol{Y}} = 0)$。

给定一致样本 $(\boldsymbol{X}, \boldsymbol{Y} = 0)$，向样本中仅插入一对相互矛盾的语义单元即可得到对应的反事实样本。第一步确定要插入的相互矛盾的语义单元对。首先混合所有不一致样本的矛盾语义单元对集合 $\boldsymbol{C}$，然后从中随机采样一对 $< \boldsymbol{T}_{s,m,k}, \boldsymbol{T}_{s,i,j} >$，其中 $\boldsymbol{T}_{s,m,k}$ 为第 s 个不一致样本中第 m 句话语的第 k 个语义单元。第二步将 $\boldsymbol{T}_{s,m,k}$ 和 $\boldsymbol{T}_{s,i,j}$ 插入到 $\boldsymbol{X}$ 中。为了保证随机性，首先从 $\boldsymbol{X}$ 中随机选择一句 $\boldsymbol{U}_k$，$(0 \leq k \leq m-1)$，然后将 $\boldsymbol{T}_{s,m,k}$ 和 $\boldsymbol{T}_{s,i,j}$ 分别插入到 $\boldsymbol{U}_m$ 和 $\boldsymbol{U}_k$ 中。$\boldsymbol{U}_m$ 和 $\boldsymbol{U}_k$ 同样用 $\boldsymbol{T}_m$ 和 $\boldsymbol{T}_k$ 来表示，以 $\boldsymbol{U}_k$ 为例，随机从 $\{\boldsymbol{T}_{k,0}, \ldots, \boldsymbol{T}_{k,j}, \ldots, \boldsymbol{T}_{k,n}\}$ 的 $n+2$ 个间隙中选择一个位置插入 $\boldsymbol{T}_{s,i,j}$。相应地，添加后的话语 $\tilde{\boldsymbol{U}}_k$ 为所有语义单元 $\boldsymbol{T}_k \cup \boldsymbol{T}_{s,i,j}$ 按顺序拼接而成的文本序列。最终，反事实样本为 $(\tilde{\boldsymbol{X}} = \{\boldsymbol{U}_0, \ldots, \tilde{\boldsymbol{U}}_k, \ldots, \tilde{\boldsymbol{U}}_m\}, \tilde{\boldsymbol{Y}} = 1)$。

5.6　实验设置

本小节主要描述实验相关的设置，包括数据集、实验细节和对比方法，以及评估方案。

5.6.1　数据集介绍

本研究选择开放域对话一致性检测数据集 DECODE[127] 来进行实验。DECODE 是开放域支持检测生成回复是否与对话上文中语义内容相互矛盾唯一的数据。为了验证本研究提出的方法能否提升模型在真实使用场景下的性能，即数据分布差异下的鲁棒性，本文选择 DECODE human-bot 测试集。该测试集是通过雇佣标注人员和对话系统交互，并诱导对话系统产生冲突内容收集的。Human-bot 测试集和人工编写的训练集有很大分布差异，模型在 human-bot 测试集上的性能表现能用来衡量模型在数据分布差异下的鲁棒性[209]。

DECODE 包含的训练集、验证集和测试集分别为 27184、4026 和 764 条，其中训练数据包含 19699 对冲突的事实观点。过滤后训练话语分解模型的数据有 52768 条。训练打分函数的数据有 392703 条。为了评估是否能够抽取出话语中所有语义单元，并且抽出的语义单元是否完整，随机选择 200 条话语并雇佣 3 位标注人员去评估，分别有 93% 的话语和 87.5% 的语义单元被所有标注人员同时接受。类似地进一步评估了辨认相互矛盾的语义单元对的正确率，其中 92.5% 被所有标注人员同时认为正确。此外，本文采用同样的方式评估了构造的反事实样本是否反转了原始样本标签，其中 91.5% 的样本被接受。

5.6.2　实现细节

本研究的实验都是基于 Huggingface 的 Transformers 库实现的。检测模型使用 RoBERTa 预训练模型，分别实现 base 和 large 两个版本。语义分解模型基于 T5-large[58] 实现，过滤语义分解模型的训练数据的模型基于 GPT2-large 实现，打分函数基于 RoBERTa-large 实现。所有实验的输入长度限制为 512。对于 large 模

型，学习率 $lr = 1e^{-5}$, 而 base 模型的学习率 $lr = 3e^{-5}$。所有实验的批次大小为 64。基于 large 的实验使用 8 张 Nvidia V100 GPU 训练，基于 base 的实验使用 4 张。训练冲突检测模型、话语分解模型、数据过滤模型和打分函数的迭代次数分别为为 5，10，3 和 5。

5.6.3 对比方法

为了验证本研究提出的方法的有效性和通用性，在两种对话一致性检测模型和两种基础模型结构上进行了测试，并和以下方法进行了对比：

- 基础模型（Baseline）[127]：仅使用原始数据训练模型；
- 话语改写（UR）[199]：先针对原始数据的指代和省略部分进行改写，然后使用改写后的数据训练模型。在测试阶段，对测试数据进行相同的改写操作；
- 回译（BackTrans）[210]：通过谷歌翻译回译对话数据，中间语言为法语；
- 基于重要项删除的构造方法（RM-CT）[201]：利用移除每个单词前后输出概率变化的显著程度来判断该单词对决策的重要性，然后通过删除对决策影响最大的那些单词去构造反事实样本；
- 基于话语修改的构造方法（MANI-U）：随机采样其他话语替换不一致样本中的相互矛盾的话语使样本变得一致；随机往一致样本中添加一对相互矛盾话语使样本变得不一致；
- 基于局部修改的构造方法（MANI-E）：对于不一致样本，确定要删除的相互矛盾的语义单元后，仅删除语义单元三元组中相互矛盾的成分，比如关系；对于一致样本，确定要添加的相互矛盾的语义单元对后，仅添加语义单元三元组中相互矛盾的成分；
- 本文提出的方法（RICS）：首先基于语义分解辨认相互矛盾的语义单元。通过删除不一致样本中的相互矛盾的语义单元使样本变得一致；随机往一致样本中添加一对相互矛盾的语义单元使样本变得不一致；
- 话语改写 + 本文提出的方法（UR+RICS）：首先对原始数据进行改写，然后再基于改写后的数据构造反事实样本。

5.6.4 评价指标

本文采用以下指标进行评估。其中假设预测矛盾且实际矛盾的样本为 True Positive（TP）；预测矛盾实际不矛盾的样本为 False Positive（FP）；预测不矛盾实际也不矛盾的样本为 True Negative（TN）；预测不矛盾实际矛盾的样本为 False Negative（FN）。

- Accuracy：预测正确的样本比例；

$$\text{Accuracy} = \frac{TP + TF}{TP + FP + TN + FN} \tag{5.61}$$

- Precision: 预测为冲突的样本是真实冲突的占比；

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.62}$$

表 5-1 基于 RoBERTa-base 的模型鲁棒性对比

Table 5-1 Robustness comparison of models based on RoBERTa-base

模型	方法	Accuracy	Recall	Precision	F1
Unstructured Approach	Baseline	66.11	44.24	78.61	56.62
	UR	67.67	48.17	78.97	59.84
	BackTrans	66.88	46.59	78.41	58.46
	RM-CT	65.96	44.34	78.24	56.52
	MANI-U	69.76	54.71	78.27	64.41
	MANI-E	67.54	52.88	74.81	61.96
	RICS	**70.64**	**59.64**	**79.22**	**68.54**
	UR+RICS	73.42	62.30	80.13	70.11
Structured Utterance-based Approach	Baseline	79.06	75.65	81.18	78.32
	UR	79.58	78.74	80.71	79.71
	BackTrans	79.71	77.48	81.09	79.25
	RM-CT	79.05	75.23	81.02	78.02
	MANI-U	79.97	76.67	81.32	78.93
	MANI-E	79.18	75.91	81.23	78.48
	RICS	**80.34**	**80.41**	**81.25**	**80.83**
	UR+RICS	81.03	81.42	80.78	81.10

- Recall：真实冲突中被实际预测出来的占比；

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.63}$$

- F1-score：综合衡量预测出冲突的能力。

$$\text{F1-score} = \frac{2}{1/\text{Precision} + 1/\text{Recall}} \tag{5.64}$$

5.7　实验结果与分析

本小节对实验结果进行展示和分析，包括检测模型在数据分布差异下的鲁棒性评估结果，探究模型的决策依据，以及验证本研究提出的方法能否提升检测模型作为排序模型的性能，进而更好地保证生成式对话系统的一致性。此外，本小节还进一步进行了消融性实验分析。

5.7.1　评估鲁棒性

本研究首先测试提出的方法是否能够提升模型在真实使用场景下的性能，即在数据分布差异下的鲁棒性。为了公平比较，本研究对所有方法都扩增原始数据 1x 的数据。实验结果在表 5-1和 5-2中给出，结果显示本研究提出的方法在多种实验设置下在几乎所有指标上都超过了对比方法。这表明本研究提出的方法能够有效地提升模型在真实使用场景下的性能，进而表明本方法能够提升模型

表 5-2 基于 RoBERTa-large 的模型鲁棒性对比

Table 5-2 Robustness comparison of models based on RoBERTa-large

模型	方法	Accuracy	Recall	Precision	F1
Unstructured Approach	Baseline	70.54	47.12	88.67	60.53
	UR	72.38	51.57	88.34	65.12
	BackTrans	71.73	49.48	89.15	63.64
	RM-CT	70.68	49.21	86.24	62.66
	MANI-U	73.17	54.97	86.42	67.19
	MANI-E	70.55	53.66	81.03	64.56
	RICS	**75.55**	**58.59**	**89.43**	**70.79**
	UR+RICS	78.79	62.56	92.64	74.69
Structured Utterance-based Approach	Baseline	84.81	79.58	88.88	83.97
	UR	85.18	81.43	88.51	84.83
	BackTrans	84.82	82.46	86.53	84.45
	RM-CT	84.16	80.10	87.18	83.49
	MANI-U	84.49	81.56	87.65	84.49
	MANI-E	84.29	79.84	87.64	83.56
	RICS	**85.29**	**82.64**	**88.19**	**85.32**
	UR+RICS	86.26	82.98	88.78	85.78

在数据分布差异下的鲁棒性。进一步观察到：（1）本研究提出的方法获得了更高的召回率（Recall），并且保持了准确率（Precision）。这表明研究方法能够帮助模型识别出更多真实的不一致样本。（2）相比以话语为单位标记矛盾内容的 MANI-U，以语义单元为标记单位的 RICS 带来了更大的提升，这表明标记简洁的矛盾内容是有效的。相反，MANI-E 提供不完整的相互矛盾内容（即只删除或者添加语义单元三元组中具体相互矛盾的成分），相比 RICS 获得了更差的性能，这表明标记完整的矛盾内容是有效的。总的来说，这表明使用语义分解方法标记简洁且完整的矛盾内容是有效的。（3）然而其他领域（如情感识别领域）的构造方法 RM-CT 用于该任务几乎没有带来任何提升，这验证了设计专门适用于对话一致性检测任务的矛盾内容辨认方法是很有必要的。（4）相比话语改写方法 UR，进一步使用本研究方法的 UR+RICS 带来了更大的提升，这表明本研究提出的方法和话语改写方式是相互正交的：在话语改写的基础上，本研究方法仍然是有效的。（5）和传统数据增强方法 BackTrans 比较，RICS 获得了更高的分数，这表明良好设计的反事实样本增强比常规的数据增强更能有效提升模型在数据分布差异下的鲁棒性。（6）相比 Unstructured Approach，本研究提出的方法给 Structured Utterance-based Approach 带来的提升相对小一些。这可能是因为 Structured Utterance-based Approach 引入的归纳结构偏置已经部分缓解了模型对伪关系的依赖[127]。因此本文方法只能带来相对小的收益。（7）本研究提出的方法给 RoBERTa-large 和 RoBERTa-base 结构的模型都带来了提升，这表明本研究

表 5-3 全量微调方法和大模型提示方法对比

Table 5-3 Comparison of full fine-tuning method and prompt learning of LLM method

模型	方法	Accuracy	Recall	Precision	F1
Unstructured Approach	Zero-shot	54.97	42.41	56.64	48.51
	Few-shot	56.15	87.69	53.77	66.66
	Baseline	70.54	47.12	88.67	60.53
	RICS	**75.55**	**58.59**	**89.43**	**70.79**
Structured Approach	Zero-shot	53.14	39.25	51.93	44.65
	Few-shot	59.03	70.41	57.35	63.22
	Baseline	84.81	79.58	88.88	83.97
	RICS	**85.29**	**82.64**	**88.19**	**85.32**

方法对不同参数量的一致性检测模型都是有效的。

此外本研究还进一步对比了基于大模型（Large Language Model）提示学习的一致性检测和基于全量微调的一致性检测。本实验选择开源大模型 Vicuna-13B，通过设计不同的提示（Prompt）去指示模型去完成一致性检测任务。具体来说，设计了以下两种提示：（1）零样本学习：任务是判断回复和对话上文中的内容是否相互矛盾。只给出 “Contradiction” 或 “Non-Contradiction” 作为答案（Zero-shot）；（2）小样本学习：任务是判断回复和对话上文中的内容是否相互矛盾。参考下面两个例子给出 “Contradiction” 或 “Non-Contradiction” 作为答案。1 - Input: ＜对话上文＞，＜回复＞，Output：Contradiction。2 - Input: ＜对话上文＞，＜回复＞，Output：Non-Contradiction（Few-shot）。进一步选择以 RoBERTa-large 为骨架模型的 Baseline 和 RICS 方法作为基于全量微调的一致性检测的代表方法。本实验分别在 Unstructured Approach 和 Structured Approach 上进行测试。实验结果在表格 5-3中给出，结果显示基于大模型提示学习的一致性检测的性能（Zero-shot 和 Few-Shot）目前还无法超越利用领域内数据微调传统预训练模型的方法。这可能是因为以下几方面原因：（1）大模型对于具体任务理解不是特别清楚，会导致结果错误；（2）未经微调的大模型无法与特定任务形式对齐，严格输出结构化信息，导致在后处理步骤中会出现误判；（3）不够精良的提示设计可能还无法有效地激发大模型在特定任务的表现。进一步观察到，相比 Zero-shot，设计提示时提供少量的参考样例（Few-shot）能够明显地提高一致性检测的效果，例如 54.97 vs. 56.15（Accuracy）和 48.51 vs. 66.66（F1）。因此，设计更加有效的提示能够极大提升大模型进行一致性检测的性能。

5.7.2　探究模型的决策依据

本实验测试使用本研究构造的反事实样本能否帮助检测模型消除对伪关系的依赖，转而通过辨认出真实的矛盾内容去做决策。本实验选择基于 RoBERTa-large 模型的 Unstructured Approach 进行实验，并选择 Baseline 和 RICS 对比。随

测试样本	移除内容	Baseline	RICS
U0：我喜欢各种音乐，我是很多种类的忠实粉丝，但乡村音乐是我的最爱。（I like all kinds of music, I am a huge fan of many kinds, but country is my favorite.） U1：你最喜欢的乡村音乐歌手是谁？（What are you favorite country artists?） U2：我不是乡村音乐的忠实粉丝，但我确实喜欢约翰尼·卡什。 你呢？（I'm not a huge fan of country, but I do like johnny cash. What about you?）	我是很多种类的忠实粉丝（I am a huge fan of many kinds）& 我不是乡村音乐的忠实粉丝（I'm not a huge fan of country）	0.683	0.002
	乡村音乐是我的最爱（country is my favorite）& 我不是乡村音乐的忠实粉丝（I'm not a huge fan of country）	0.002	0.976
U0：不，我不吃肉是出于对生命的尊重。（Nope, I don't eat meat out of respect for sentient life.） U1：你吃什么食物？（What kinds of food do you eat?） U2：比萨，肉饼，任何她能做的灵魂食物我都吃（Pizza, meat loaf, any soul food she can cook up I eat.） U3：我以为你不吃肉。（I thought you did not eat meat.） U4：不，这意味着我不吃动物的部位或动物制品（Nope, it means I eat no animal parts or product.）	不（Nope）& 不（Nope）	0.329	$1.18e^{-5}$
	肉饼我吃（meat loaf I eat）& 我不吃动物的部位或动物制品（I eat no animal parts or product）	0.021	0.996

图 5-5 探究模型的决策依据

Figure 5-5 Exploration on the decision-making basis of models

后从测试集中选择 2 条测试样本，并且跟随 Yang 等[201] 采用探究模型对高频共现内容和真实矛盾内容的敏感性的方法来测试。具体来说，对于给定的测试样本，每次移除高频共现内容或真实矛盾内容中的一方信息，计算模型在移除前后预测概率的变化差值。最终取两次结果的均值代表敏感性。敏感程度越高，表明模型更容易根据该内容做决策。采用每次移除一方信息而非直接移除成对信息的方式计算是因为模型在影响决策的任意一方信息缺失时就已经表现出极大的敏感性，无法展示是否对另一方信息同样敏感。

实验结果在图 5-5中给出，其中 “移除内容” 中第一行是高频共现内容，第二行是真实矛盾内容。“Baseline” 和 “RICS” 列的数值分别表示对应模型对高频共现内容和真实矛盾内容的敏感性。对于样本 1 而言，移除高频共现内容，Baseline 的预测概率变化差值为 0.683，而 RICS 的预测概率几乎不变。相反，移除真实矛盾内容，RICS 的变化差值为 0.976，而 Baseline 几乎不受影响。对于样本 2 而言，移除高频共现内容之后，Baseline 的预测概率变化差值明显大于 RICS（0.329 vs. $1.18e^{-5}$）。而移除真实矛盾内容后，RICS 的预测概率变化幅度非常之大（0.996）；相比而言，Baseline 的预测概率几乎没有受到什么影响（0.021）。因此，实验结果验证了使用本文构造的反事实样本能够消除模型对伪关系的依赖，使得模型能通过辨认出真实矛盾内容去做决策。

表 5-4 传统对话生成模型一致性评估

Table 5-4 Consistency Evaluation for Conventional Dialogue Generation Model

排序模型	Beam Search	Top-k	Top-p
None	29.97%	29.34%	26.83%
Baseline	25.26%	17.28%	17.27%
RICS	18.84%	8.11%	8.76%

表 5-5 大规模对话生成模型一致性评估

Table 5-5 Consistency Evaluation for Large Dialogue Generation Model

排序模型	Top-p ($p = 0.5$)	Top-p ($p = 0.75$)	Top-p ($p = 1$)
None	14.14%	13.21%	13.87%
Baseline	10.21%	7.72%	6.41%
RICS	7.46%	4.97%	3.14%

5.7.3　评估对话生成一致性

本实验评估提升一致性检测模型在真实使用场景下的性能后，使用检测模型对生成回复进行一致性重排序能否更好地满足一致的约束。本实验跟随 Nie 等[127] 选择 BlenderBot（BST 2.7B）作为对话生成模型。具体来说，BlenderBot 输出 N 个回复，然后检测模型计算生成回复的一致性得分，选择得分最高的回复作为输出。最后评估所有样本中输出回复与对话上文中语义内容相互矛盾的比率。为了公平比较，本实验选择 3 种排序模型进行对比：（1）不使用一致性排序（None）：BlenderBot 仅输出 1 个回复，直接评估所有样本中出现不一致回复的比例；（2）使用 Baseline 进行一致性排序（Baseline）：BlenderBot 输出 10 个回复，使用一致性检测模型 Baseline 计算回复的一致性得分；（3）使用 RICS 进行一致性排序（RICS）：BlenderBot 输出 10 个回复，使用一致性检测模型 RICS 计算回复的一致性得分。此外，对话上文数据来自于 DECODE 的 Human-bot 测试集。选择该数据的原因是该数据集中的对话上文已经被标注人员诱导去生成不一致的回复，能够暴露出模型的不一致问题，从而更好地评估本研究提出的方法对对话生成模型满足一致性约束的影响。BlenderBot 针对给定对话上文解码生成回复时，考虑了三种解码方式：集束搜索（Beam Search）、Top-k 采样、以及 Top-p 采样。对于集束搜索，束大小（beam size）为 10；对于 Top-k 采样，$k = 40$；对于 Top-p 采样，$p = 0.9$。本实验的所有一致性检测模型采用 Unstructured Approach，模型的基础模型为 RoBERTa-large 模型。

实验结果如表格 5-4所示，结果显示本研究提出的方法在所有解码方式下都能明显地降低生成回复的不一致比率。进一步观察到：（1）相比不使用一致性排序模型时（None），使用一致性检测模型 Baseline 使得对话生成模型 BlenderBot 输出的不一致回复比率有所下降，即 29.97% vs. 25.26%（Beam Search）、29.34%

表 5-6 消融实验结果

Table 5-6 Ablation experiment results

模型	Accuracy	Recall	Precision	F1
RICS	75.55	58.59	89.43	70.79
w/o MinChange	72.12	51.05	88.23	64.67
w/o Complete	73.69	57.06	84.15	68.01

vs. 17.28%（Top-k）、26.83% vs. 17.27%（Top-p）。（2）相比 Baseline，提升鲁棒性后的一致性检测模型 RICS 进一步使得 BlenderBot 输出中不一致的回复明显减少，即 25.26% vs. 18.84%（Beam Search）、17.28% vs. 8.11%（Top-k），17.27% vs. 8.76%（Top-p）。这表明直接使用人工标注训练语料训练的检测模型（Baseline）在真实使用场景下检测结果的确不够准确。提升检测模型在真实使用场景下的鲁棒性，能够为生成回复输出更准确的一致性得分，从而更好地使生成式对话系统满足一致性约束。

进一步，本实验还选择了 Vicuna-13B 作为对话生成模型。Vicuna-13B 在生成回复时的解码方式使用 Top-p 采样，本实验分别针对 $p = 0.5$，$p = 0.75$ 和 $p = 1$ 进行测试。实验结果如表格 5-5所示，结果显示：（1）目前的大模型相比传统对话生成模型输出不一致回复的比例有了明显下降（26.83% vs. 13.87%），但是不可完全避免。（2）使用一致性检测模型（Baseline 和 RICS）对回复进行一致性排序仍然能够提升对话模型的一致性，这表明集成一致性检测模型使大规模对话模型更好地满足一致性约束依然是有效的思路。（3）相比 Baseline，RICS 仍然能够进一步使得 Vicuna-13B 模型输出的不一致回复有明显减少。这表明本文提出的方法具有良好的通用性。

5.7.4 消融性实验

为了进一步验证反事实样本构造过程中保证的原则的有效性，本实验进行了以下消融测试：(1) 在现有构造方案基础上，对于不一致样本和一致样本，再分别随机删除或添加 15% 的语义单元，用于验证保证最小修改原则的有效性（w/o MinChange）；（2）构造反事实样本时，对于需要修改的话语不采用语义单元表示，直接从原始话语中删除或添加相互矛盾的内容，用于验证保证话语完整性是否有效（w/o Complete）。所有实验都是基于 RoBERTa-large 结构的 Unstructured Approach 上完成的。实验结果在表格 5-6中给出，观察结果发现不保证任何一项原则都会带来性能下降。不保证最小修改原则（w/o MinChange）时，检测模型性能在 Accuracy 指标上下降了 3.43 个点，在 F1 指标上下降了 6.12 个点。同样，不保证话语完整性（w/o Complete）时，检测模型在 Accuracy 指标上下降了 1.86 个点，在 F1 指标上下降了 2.78 个点。这验证了反事实样本构造过程中保证最小修改原则和话语完整性原则是不可或缺的。

5.8　小结

本章提出了一种适用于对话一致性检测任务的反事实样本构方法。构造的反事实样本用于缓解一致性检测模型对伪关系的依赖，提升模型在真实使用场景下的性能，也即提升模型在数据分布差异下的鲁棒性。具体来说，本方法首先采用语义分解方法将原始对话中的复杂话语分解成多个独立的诸如事实信息的语义单元，通过判断哪些语义单元相互矛盾去标记矛盾内容。然后对于不一致样本，通过删除能使样本标签反转的最少的相互矛盾的语义单元去得到对应的反事实样本；对于一致样本，向样本中随机添加一对相互矛盾的语义单元去得到对应的反事实样本。最终，混合反事实样本和原始样本去训练现有的对话一致性检测模型。实验结果表明本文提出的方法能够有效地提升检测模型在真实使用场景下的性能，即提升检测模型在数据分布差异下的鲁棒性。相应地，这也表明能够使得生成回复更好地满足与对话上文中的语义内容间的一致性约束。此外，分析实验还验证本研究构造的反事实样本能够有效地消除模型对伪关系的依赖，使得模型能通过辨认出真实矛盾内容去做决策。

第 6 章　总结与展望

6.1　研究工作总结

开放域对话系统是自然语言处理领域非常值得关注的研究方向之一。随着大量真实人类对话数据被收集以及可用的计算资源逐渐增加，基于深度神经网络技术的开放域对话系统取得了显著的进展。尤其是深度文本生成技术的快速发展使得生成式开放域对话已具备一定的人机交互能力，展现出足够的智能和广阔的应用前景。开放域对话具有独特的开放性特点，即允许对话有各种各样从不同角度生成的回复，尤其是语义内容各不相同的回复。因此，生成式开放域对话系统需要能够从语义内容丰富的训练数据中充分理解复杂的语义结构，进而才能生成语义多样的回复。同时，生成的回复还要满足和对话历史中的语义内容不能自相矛盾（即保持一致）的约束。但由于开放域对话生成任务的复杂性，目前距离实际可用的水平还有很远一段距离。本文针对系统在构建过程中所面临的多个关键性挑战展开了研究。具体来说，本文以关键要素“语义”作为切入点，不断加深对语义信息的理解，并不断深入对语义的挖掘与利用，为多个关键性挑战都提供了相应的解决方案，并提升了生成式开放域对话系统回复生成的质量。本文主要完成了三个研究工作：（1）研究建模对话的语义映射关系，提出了基于语义表示的对话回复生成方法；（2）研究增强训练数据的语义内容，提出了基于语义转换的对话数据增强方法；（3）研究满足回复语义内容的一致性约束，提出了基于语义分解的对话一致性检测方法。本文的主要研究内容和贡献总结如下：

1、对于对话语义映射关系建模难的问题，本文提出了一种基于语义表示的对话回复生成方法，用于显示建模层级语义映射关系，以生成语义多样的回复。首先，该方法设计了多语义检索模块，为每个对话上文从预定义的对话数据集中检索出 k 个有效的回复来扩充回复集合。这不仅显式地定义了不同语义侧面，并且给模型学习不同语义侧面的信息提供了监督信号。其次，该方法改进了对话 Wasserstein 自编码器，通过优化对齐方式确保可生成回复的不同语义侧面都能对应到不同的隐变量上，并且增加额外的语义距离损失函数使得每个隐变量的分布距离尽可能大。这样保证了每个对话上文首先映射到多个代表不同语义侧面的隐变量上，每个隐变量服从的高斯分布又建模了回复语义相似但用词多样的表达。最终在推理阶段，模型首先随机选择代表不同语义侧面的隐变量分布，再从分布上采样获取不同的隐变量值，基于该值即可生成语义内容多样的回复。实验结果表明该方法能够生成更多语义多样的回复。同时分析实验表明该方法能够有效地建模开放域对话的层级语义映射关系。

2、对于训练数据语义不够丰富的问题，本文提出了一种基于语义转换的数据增强方法，为给定的对话上文扩增更多具有不同语义的回复。具体来说，该方法借助反事实推理，通过干预当前环境下观察到的用来生成回复的语义角度来生成反事实回复，通过提供不同的语义角度去生成不同语义的回复。首先，该方

法将反事实回复生成模型转换成可进行反事实推理的结构因果模型，利用其中的不可观测变量来模拟当前推理环境。其次，该方法基于所有观测对话数据构建一个转移关系图去获得有效的可生成回复的语义角度。因为它明确地表示了人类在对话上文中的关注点与其相应可生成回复的语义角度之间的转移关系。通过从给定的对话上文中随机选择一个关注点，并将该关注焦点在所有观测数据中合理转移到的语义角度，作为候选集合。随后从候选集合中预测即可保证得到的可转换的语义角度的有效性。最终该方法根据新的语义角度在当前环境下重新推理去扩增出不同语义的回复。在得到所有反事实增广回复后，该方法设计了双向困惑度数据选择模块进一步去选择出高质量且有趣的数据。最后，该方法将观测到的数据与所有增强的数据混合，作为下游任务的训练数据。实验结果表明，该方法可以扩增具有不同语义的高质量回复，并且能够有效地提升对话生成的语义多样性。此外，使用这种更加类人的训练数据也能够提升开放域对话下游任务的整体质量。

3、对于回复语义内容一致性约束难以满足的问题，本文提出了一种基于语义分解的对话一致性检测方法。构造的反事实样本用于缓解检测模型对伪关系的依赖，提升模型在真实使用场景下的性能，也即提升模型在数据分布差异下的鲁棒。该方法构造反事实样本的思路为：先自动化标记相互矛盾的内容，再通过修改尽可能少的矛盾内容反转原始样本的标签。具体来说，该方法首先采用语义分解方法将原始对话中的复杂话语分解成多个独立的诸如事实信息的语义单元，通过判断哪些语义单元相互矛盾去标记矛盾内容。然后对于不一致样本，通过删除能使样本标签反转的最少的相互矛盾的语义单元去得到对应的反事实样本；对于一致样本，向样本中随机添加一对相互矛盾的语义单元去得到对应的反事实样本。其中该方法实现了一个语义分解模型抽取语义单元。由于并没有对话领域的语义单元抽取语料，本研究利用了开放域知识抽取的语料。进一步发现，由于开放域知识抽取数据和开放与对话数据存在数据分布差异，因此该方法额外引入了数据排序模型优先选择分布接近对话数据的开放域数知识抽取据作为检测模型的训练数据。最终，混合反事实样本和原始样本去训练现有的冲突一致性模型。实验结果表明本文提出的方法能够有效地提升检测模型在真实使用场景下的性能，即提升检测模型在数据分布差异下的鲁棒性。相应地，这也表明能够使得生成回复更好地满足与对话上文中的语义内容间的一致性约束。

6.2 未来工作展望

本文针对生成式开放域对话系统所面临的三个关键问题展开了研究，并提出了相应的解决方案。但生成式开放域对话系统仍存在很多亟待解决的问题。本文立足于文中工作，结合目前 ChatGPT 相关模型的发展，提出如下几点展望：

开放域对话生成的幻觉问题 深度神经网络主要是从海量数据中学习相关模式。有些事实知识频繁出现，且上下文相对固定，那预测的词语概率就很尖锐，模

型就容易输出正确的事实知识；相反，一些事实不经常出现，模型没有习得相关模式，并且上下文结构相对松散，那预测的词语概率就会相对平滑，就容易产生不确定的输出，从而就出现了幻觉问题。幻觉问题分为内部幻觉 (Intrinsic Hallucination) 和外部幻觉 (Extrinsic Hallucination) 两大类。所谓内部幻觉，是指模型生成的回复与对话历史或与自身已生成回复相矛盾，出现了不一致现象。因此，如何避免开放域对话系统产生不一致的回复，以及如何简单高效地检测不一致现象都是值得探索的课题。所谓外部幻觉，是指开放域对话系统生成了输入中未提及的内容，但无法找出相关证据，也不能断言未提及内容是错误的。尽管外部幻觉提高了对话生成的多样性和信息量，但可能会提供给用户错误的信息，长期以往会降低用户对它的信任度和好感度。因此，如何提高生成式开放域对话系统的事实可靠性是一个亟待解决的问题。其中造成模型产生幻觉的首要因素是训练数据，那么如何高效简便地构造高质量的数据集仍然是值得关注的问题。其次，对话一对多的特性也会造成幻觉问题，如何更好地进行可控文本生成也值得继续深入研究。另外，模型幻觉的评估仍是一个悬而未决的难题，截至目前也没有标准指标。因此，评估开放域对话的事实可靠性具有重要的研究价值。

对话生成模型的个性设定　开放域对话系统具备一致的个性化能够更好的获取用户的信任和好感，并提高用户的参与度。目前的个性化一致性主要体现在用户的基本信息上，例如姓名、性别、年龄、星座、以及住址等信息上。现在所关注的个性化定义是非常简单的，并不符合人类对话中的个性化一致性。人类对话中一致性更多地体现在对事件的关注点，对相应关注点的态度上。此外还和用户自身的人格特点或者情绪状态有关。因此未来的研究方向分为以下几类：（1）任务定义：如何定义更准确的个性化对话任务，使得对话系统具有更贴近真实人类的个性化一致性；（2）数据构建：如何快速有效地构建高质量的标注数据；（3）个性化建模：给定对话上文和用户的历史言论，从中挖掘出用户的个性化信息，并能够产生保持个性化一致的回复；总的来说，构建个性化一致的开放域对话系统从任务定义、数据收集、模型设计、数据评估方面都需要进一步探索。

开放域对话生成的评估问题　开放域对话系统的目标在于最大化用户的长期参与度，其不够清晰具体的特点使得很难评估生成的回复的质量。目前的评估方式主要是采用一系列评估指标来评估回复的特定方面，比如相关性、流畅性、多样性等。具体就是通过自动指标或人工标注的方法对回复的不同方面进行打分。自动评估和人工评估主要采用的是静态评估的方法，即给定对话上文，对生成的回复进行评估。但静态评估并不能真实反映开放域对话系统在现实世界中的能力，因此交互式对话评估逐渐受到关注。交互式对话评估可以采用人-机器评估和机器-机器评估两种方式。对于人-机器评估，需要人和机器进行交互对话，并收集对话数据进行评估。这一过程非常耗时耗力，如何提高效率并减少人力消耗是值得研究的问题。对于机器-机器评估，可能需要机器具备一定的策略诱导待评估机器犯错，例如评估一致性时需要机器具备诱导对方重复谈论同一个问题的能

力，但如何实现这些策略还有待探索。此外，目前对话评估指标多集中在对话系统的基础能力上，比如相关性和流畅性等。随着对话系统的发展，对于一些高阶能力，比如共情能力、可靠程度、安全性等，如何对这些高阶能力进行评估也是一个有趣的研究课题。

参考文献

[1] Turing A M. Computing machinery and intelligence [M]. Springer, 2009.

[2] Huang M, Zhu X, Gao J. Challenges in building intelligent open-domain dialog systems [J]. ACM Transactions on Information Systems (TOIS), 2020, 38(3): 1-32.

[3] Ni J, Young T, Pandelea V, et al. Recent advances in deep learning based dialogue systems: A systematic survey [J]. Artificial intelligence review, 2022: 1-101.

[4] Ritter A, Cherry C, Dolan W B. Data-driven response generation in social media [C]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011: 583-593.

[5] Gao J, Galley M, Li L. Neural approaches to conversational ai [C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 1371-1374.

[6] Li J, Galley M, Brockett C, et al. A diversity-promoting objective function for neural conversation models [C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 110-119.

[7] Zhang H, Lan Y, Guo J, et al. Tailored sequence to sequence models to different conversation scenarios [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, 2018: 1479-1488.

[8] Ling Y, Cai F, Chen H, et al. Leveraging context for neural question generation in open-domain dialogue systems [C]//Proceedings of The Web Conference 2020. 2020: 2486-2492.

[9] Zhou H, Huang M, Zhang T, et al. Emotional chatting machine: Emotional conversation generation with internal and external memory [C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[10] Shang L, Lu Z, Li H. Neural responding machine for short-text conversation [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015: 1577-1586.

[11] Sordoni A, Galley M, Auli M, et al. A neural network approach to context-sensitive generation of conversational responses [C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 196-205.

[12] Huber B, McDuff D, Brockett C, et al. Emotional dialogue generation using image-grounded language models [C]//Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018: 1-12.

[13] Xing C, Wu W, Wu Y, et al. Topic aware neural response generation [C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[14] Zhang S, Dinan E, Urbanek J, et al. Personalizing dialogue agents: I have a dog, do you have pets too? [J].

[15] Chen H, Liu X, Yin D, et al. A survey on dialogue systems: Recent advances and new frontiers [J]. Acm Sigkdd Explorations Newsletter, 2017, 19(2): 25-35.

[16] Tao C, Feng J, Yan R, et al. A survey on response selection for retrieval-based dialogues. [C]//IJCAI. 2021: 4619-4626.

[17] Zhu Q, Cui L, Zhang W, et al. Retrieval-enhanced adversarial training for neural response generation [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3763-3773.

[18] Jafarpour S, Burges C J, Ritter A. Filter, rank, and transfer the knowledge: Learning to chat [J]. Advances in Ranking, 2010, 10: 2329-9290.

[19] Leuski A, Traum D. Npceditor: Creating virtual human dialogue using information retrieval techniques [J]. Ai Magazine, 2011, 32(2): 42-56.

[20] Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data [C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2013: 2333-2338.

[21] Yan R, Song Y, Wu H. Learning to respond with deep neural networks for retrieval-based human-computer conversation system [C]//Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. 2016: 55-64.

[22] Zhou X, Dong D, Wu H, et al. Multi-view response selection for human-computer conversation [C]//Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 372-381.

[23] 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). Vancouver, Canada: Association for Computational Linguistics, 2017: 496-505.

[24] Zhou X, Li L, Dong D, et al. Multi-turn response selection for chatbots with deep attention matching network [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1118-1127.

[25] Yang L, Qiu M, Qu C, et al. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems [C]//The 41st international acm sigir conference on research & development in information retrieval. 2018: 245-254.

[26] Zhang Z, Li J, Zhu P, et al. Modeling multi-turn conversation with deep utterance aggregation [C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 3740-3752.

[27] Wu Y, Li Z, Wu W, et al. Response selection with topic clues for retrieval-based chatbots [J]. Neurocomputing, 2018, 316: 251-261.

[28] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186.

[29] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach [J].

[30] Gu J C, Li T, Liu Q, et al. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots [C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 2041-2044.

[31] Whang T, Lee D, Oh D, et al. Do response selection models really know what' s next? utterance manipulation strategies for multi-turn response selection [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 35. 2021: 14041-14049.

[32] Sevegnani K, Howcroft D M, Konstas I, et al. Otters: One-turn topic transitions for open-domain dialogue [C]//Proceedings of the 59th Annual Meeting of the Association for Com-

putational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 2492-2504.

[33] Feng J, Tao C, Liu C, et al. How to represent context better? an empirical study on context modeling for multi-turn response selection [C]//Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 7285-7298.

[34] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [J]. Advances in neural information processing systems, 2014, 27.

[35] Vinyals O, Le Q. A neural conversational model [Z]. 2015.

[36] Du J, Li W, He Y, et al. Variational autoregressive decoder for neural response generation [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3154-3163.

[37] Ke P, Guan J, Huang M, et al. Generating informative responses with controlled sentence function [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1499-1508.

[38] Serban I, Sordoni A, Lowe R, et al. A hierarchical latent variable encoder-decoder model for generating dialogues [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 31. 2017.

[39] Shen X, Su H, Niu S, et al. Improving variational encoder-decoders in dialogue generation [C]//Proceedings of the AAAI conference on artificial intelligence: volume 32. 2018.

[40] Zhao T, Lee K, Eskenazi M. Unsupervised discrete sentence representation learning for interpretable neural dialog generation [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1098-1107.

[41] Zhao T, Zhao R, Eskenazi M. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 654-664.

[42] Li J, Monroe W, Shi T, et al. Adversarial learning for neural dialogue generation [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2157-2169.

[43] Xu J, Ren X, Lin J, et al. Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation [C]//Proceedings of the 2018 conference on empirical methods in natural language processing. 2018: 3940-3949.

[44] Golovanov S, Kurbanov R, Nikolenko S, et al. Large-scale transfer learning for natural language generation [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 6053-6058.

[45] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training [J]. 2018.

[46] Wolf T, Sanh V, Chaumond J, et al. Transfertransfo: A transfer learning approach for neural network based conversational agents [J].

[47] Zhang Y, Sun S, Galley M, et al. Dialogpt: Large-scale generative pre-training for conversational response generation [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020: 270-278.

[48] 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.

[49] Bidirectional recurrent neural networks [J]. IEEE transactions on Signal Processing, 1997, 45(11): 2673-2681.

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

[51] Cho K, van Merriënboer B, Gülçehre Ç, et al. Learning phrase representations using rnn encoder–decoder for statistical machine translation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1724-1734.

[52] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv:1409.0473, 2014.

[53] Sordoni A, Galley M, Auli M, et al. A neural network approach to context-sensitive generation of conversational responses [C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 196-205.

[54] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Guyon I, Luxburg U V, Bengio S, et al. Advances in Neural Information Processing Systems. 2017.

[55] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding [J]. Advances in neural information processing systems, 2019, 32.

[56] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019, 1(8): 9.

[57] 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.

[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 [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020: 7871-7880.

[60] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing [J]. ACM Computing Surveys, 2023, 55(9): 1-35.

[61] Shorten C, Khoshgoftaar T M. A survey on image data augmentation for deep learning [J]. Journal of big data, 2019, 6(1): 1-48.

[62] Wen Q, Sun L, Yang F, et al. Time series data augmentation for deep learning: A survey [C]//Zhou Z H. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. International Joint Conferences on Artificial Intelligence Organization, 2021: 4653-4660.

[63] Feng S Y, Gangal V, Wei J, et al. A survey of data augmentation approaches for NLP [C]// Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021: 968-988.

[64] Chen J, Tam D, Raffel C, et al. An empirical survey of data augmentation for limited data learning in nlp [J]. ArXiv, 2021, abs/2106.07499.

[65] Du W, Black A. Data augmentation for neural online chats response selection [C]// Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on

Search-Oriented Conversational AI. Brussels, Belgium: Association for Computational Linguistics, 2018.

[66] Niu T, Bansal M. Automatically learning data augmentation policies for dialogue tasks [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). Hong Kong, China: Association for Computational Linguistics, 2019.

[67] Li J, Qiu L, Tang B, et al. Insufficient data can also rock! learning to converse using smaller data with augmentation [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019.

[68] Cai H, Chen H, Song Y, et al. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 6334-6343.

[69] Zhang R, Zheng Y, Shao J, et al. Dialogue distillation: Open-domain dialogue augmentation using unpaired data [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 3449-3460.

[70] Xie S, Lv A, Xia Y, et al. Target-side input augmentation for sequence to sequence generation [C]//International Conference on Learning Representations. 2022.

[71] Cao Y, Bi W, Fang M, et al. A model-agnostic data manipulation method for persona-based dialogue generation [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

[72] Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016.

[73] Chang E, Shen X, Zhu D, et al. Neural data-to-text generation with LM-based text augmentation [C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, 2021.

[74] Yang Y, Malaviya C, Fernandez J, et al. Generative data augmentation for commonsense reasoning [C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 1008-1025.

[75] Schick T, Schütze H. Generating datasets with pretrained language models [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 6943-6951.

[76] Wang Z, Yu A W, Firat O, et al. Towards zero-label language learning [J]. arXiv preprint arXiv:2109.09193, 2021.

[77] Weston J, Dinan E, Miller A. Retrieve and refine: Improved sequence generation models for dialogue [C]//Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI. Brussels, Belgium: Association for Computational Linguistics, 2018: 87-92.

[78] Yang L, Hu J, Qiu M, et al. A hybrid retrieval-generation neural conversation model [C]// Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 1341-1350.

[79] Zhou L, Gao J, Li D, et al. The design and implementation of xiaoice, an empathetic social chatbot [J]. Computational Linguistics, 2020, 46(1): 53-93.

[80] Wu Y, Wei F, Huang S, et al. Response generation by context-aware prototype editing [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 7281-7288.

[81] Song Y, Yan R, Li C T, et al. An ensemble of retrieval-based and generation-based human-computer conversation systems. [J]. 2018.

[82] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [C]//3rd International Conference on Learning Representations. 2015.

[83] Gu J, Lu Z, Li H, et al. Incorporating copying mechanism in sequence-to-sequence learning [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 1631-1640.

[84] Pandey G, Contractor D, Kumar V, et al. Exemplar encoder-decoder for neural conversation generation [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1329-1338.

[85] Zhang J, Tao C, Xu Z, et al. Ensemblegan: Adversarial learning for retrieval-generation ensemble model on short-text conversation [C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 435-444.

[86] Cai D, Wang Y, Bi W, et al. Skeleton-to-response: Dialogue generation guided by retrieval memory [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 1219-1228.

[87] Cai D, Wang Y, Bi W, et al. Retrieval-guided dialogue response generation via a matching-to-generation framework [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). Hong Kong, China: Association for Computational Linguistics, 2019: 1866-1875.

[88] Serban I, Sordoni A, Lowe R, et al. A hierarchical latent variable encoder-decoder model for generating dialogues [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1).

[89] Chen C Y, Yu D, Wen W, et al. Gunrock: Building a human-like social bot by leveraging large scale real user data [J].

[90] Niu T, Bansal M. Polite dialogue generation without parallel data [J]. Transactions of the Association for Computational Linguistics, 2018, 6: 373-389.

[91] Zhong P, Wang D, Miao C. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 33. 2019: 7492-7500.

[92] Song H, Zhang W N, Cui Y, et al. Exploiting persona information for diverse generation of conversational responses [J].

[93] Gupta P, Bigham J P, Tsvetkov Y, et al. Controlling dialogue generation with semantic exemplars [C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 3018-3029.

[94] Zhang H, Liu Z, Xiong C, et al. Grounded conversation generation as guided traverses in commonsense knowledge graphs [C]//Proceedings of the 58th Annual Meeting of the Asso-

ciation for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 2031-2043.

[95] Wu S, Li Y, Zhang D, et al. Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact [C]//Bessiere C. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. International Joint Conferences on Artificial Intelligence Organization, 2020: 3766-3772.

[96] Wu W, Guo Z, Zhou X, et al. Proactive human-machine conversation with explicit conversation goal [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 3794-3804.

[97] Xu J, Wang H, Niu Z Y, et al. Conversational graph grounded policy learning for open-domain conversation generation [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020: 1835-1845.

[98] Zou Y, Liu Z, Hu X, et al. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021: 2215-2226.

[99] Chen H, Ren Z, Tang J, et al. Hierarchical variational memory network for dialogue generation [C]//Proceedings of the 2018 World Wide Web Conference. 2018: 1653-1662.

[100] Park Y, Cho J, Kim G. A hierarchical latent structure for variational conversation modeling [C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana, 2018: 1792-1801.

[101] Gao J, Bi W, Liu X, et al. A discrete CVAE for response generation on short-text conversation [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: 1898-1908.

[102] Gu X, Cho K, Ha J W, et al. DialogWAE: Multimodal response generation with conditional wasserstein auto-encoder [C]//International Conference on Learning Representations. 2019.

[103] Zhou G, Luo P, Cao R, et al. Mechanism-aware neural machine for dialogue response generation [C]//Proceedings of the AAAI conference on artificial intelligence: volume 31. 2017.

[104] Zhou G, Luo P, Xiao Y, et al. Elastic responding machine for dialog generation with dynamically mechanism selecting [C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[105] Wolf T, Sanh V, Chaumond J, et al. Transfertransfo: A transfer learning approach for neural network based conversational agents [J]. arXiv preprint arXiv:1901.08149, 2019.

[106] Zhao X, Wu W, Tao C, et al. Low-resource knowledge-grounded dialogue generation [C]// International Conference on Learning Representations.

[107] Zhao X, Wu W, Xu C, et al. Knowledge-grounded dialogue generation with pre-trained language models [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 3377-3390.

[108] Cao Y, Bi W, Fang M, et al. Pretrained language models for dialogue generation with multiple input sources [C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 909-917.

[109] 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.

[110] Dziri N, Madotto A, Zaïane O, et al. Neural path hunter: Reducing hallucination in dialogue systems via path grounding [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 2197-2214.

[111] Gu X, Yoo K M, Ha J W. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 35. 2021: 12911-12919.

[112] Huang Z, Dou Z, Zhu Y, et al. MCP: Self-supervised pre-training for personalized chatbots with multi-level contrastive sampling [C]//Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 1030-1042.

[113] Bao S, He H, Wang F, et al. Plato-2: Towards building an open-domain chatbot via curriculum learning [C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 2513-2525.

[114] Bao S, He H, Wang F, et al. Plato-xl: Exploring the large-scale pre-training of dialogue generation [C]//Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022. 2022: 107-118.

[115] Bao S, He H, Xu J, et al. PLATO-K: internal and external knowledge enhanced dialogue generation [J]. CoRR, 2022, abs/2211.00910.

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

[117] 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.

[118] So D, Le Q, Liang C. The evolved transformer [C]//International conference on machine learning. PMLR, 2019: 5877-5886.

[119] Wang Y, Ke P, Zheng Y, et al. A large-scale chinese short-text conversation dataset [C]// Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I 9. Springer, 2020: 91-103.

[120] Thoppilan R, De Freitas D, Hall J, et al. Lamda: Language models for dialog applications [J]. arXiv preprint arXiv:2201.08239, 2022.

[121] 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.

[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] Fang H, Cheng H, Clark E, et al. Sounding board–university of washington's alexa prize submission [J].

[124] Challa A, Upasani K, Balakrishnan A, et al. Generate, filter, and rank: Grammaticality classification for production-ready nlg systems [C]//Proceedings of the 2019 Conference of the

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers). 2019: 214-225.

[125] Welleck S, Weston J, Szlam A, et al. Dialogue natural language inference [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3731-3741.

[126] Song H, Wang Y, Zhang W N, et al. Profile consistency identification for open-domain dialogue agents [J].

[127] Nie Y, Williamson M, Bansal M, et al. I like fish, especially dolphins: Addressing contradictions in dialogue modeling [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: 1699-1713.

[128] Yeh Y T, Eskenazi M, Mehri S. A comprehensive assessment of dialog evaluation metrics [C]//The First Workshop on Evaluations and Assessments of Neural Conversation Systems. 2021: 15-33.

[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 on association for computational linguistics. 2002: 311-318.

[130] Banerjee S, Lavie A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments [C]//Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005: 65-72.

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

[132] Liu C W, Lowe R, Serban I V, et al. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 2122-2132.

[133] Dušek O, Novikova J, Rieser V. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge [J]. Computer Speech & Language, 2020, 59: 123-156.

[134] Zhu Y, Lu S, Zheng L, et al. Texygen: A benchmarking platform for text generation models [C]//The 41st international ACM SIGIR conference on research & development in information retrieval. 2018: 1097-1100.

[135] Lowe R, Noseworthy M, Serban I V, et al. Towards an automatic Turing test: Learning to evaluate dialogue responses [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 1116-1126.

[136] Tao C, Mou L, Zhao D, et al. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems [C]//Proceedings of the AAAI conference on artificial intelligence: volume 32. 2018.

[137] Ghazarian S, Wei J, Galstyan A, et al. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings [C]//Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation. 2019: 82-89.

[138] Lan T, Mao X L, Wei W, et al. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems [J]. ACM Transactions on Information Systems (TOIS), 2020, 39(1): 1-37.

[139] Sinha K, Parthasarathi P, Wang J, et al. Learning an unreferenced metric for online dialogue

evaluation [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2430-2441.

[140] Huang L, Ye Z, Qin J, et al. Grade: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 9230-9240.

[141] Ghazarian S, Weischedel R, Galstyan A, et al. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. 2020: 7789-7796.

[142] Mehri S, Eskenazi M. Unsupervised evaluation of interactive dialog with dialogpt [C]// Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2020: 225-235.

[143] Zhang C, Chen Y, D' Haro L F, et al. Dynaeval: Unifying turn and dialogue level evaluation [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: 5676-5689.

[144] Jiang Z, Ye G, Rao D, et al. Im^ 2: an interpretable and multi-category integrated metric framework for automatic dialogue evaluation [C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 11091-11103.

[145] Randolph J J. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. [J]. Online submission, 2005.

[146] Tao C, Gao S, Mingyue, et al. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism [C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. 2018: 4418-4424.

[147] Chen C, Peng J, Wang F, et al. Generating multiple diverse responses with multi-mapping and posterior mapping selection [C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. 2019: 4918-4924.

[148] Zhao T, Zhao R, Eskénazi M. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders [J]. CoRR, 2017.

[149] Zhao T, Zhao R, Eskenazi M. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 654-664.

[150] Baheti A, Ritter A, Li J, et al. Generating more interesting responses in neural conversation models with distributional constraints [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 3970-3980.

[151] Li J, Monroe W, Jurafsky D. A simple, fast diverse decoding algorithm for neural generation [J]. ArXiv, 2016, abs/1611.08562.

[152] Li J, Monroe W, Shi T, et al. Adversarial learning for neural dialogue generation [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 2157-2169.

[153] Shao L, Gouws S, Britz D, et al. Generating long and diverse responses with neural conversation models [J]. 2016.

[154] Moon S, Shah P, Kumar A, et al. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs [C]//Proceedings of the 57th Annual Meeting

of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 845-854.

[155] Tuan Y L, Chen Y N, Lee H y. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs [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). Hong Kong, China: Association for Computational Linguistics, 2019: 1855-1865.

[156] MacQueen J, et al. Some methods for classification and analysis of multivariate observations [C]//1967.

[157] Zhao J, Kim Y, Zhang K, et al. Adversarially regularized autoencoders [C]//Proceedings of the 35th International Conference on Machine Learning. 2018: 5902-5911.

[158] Tolstikhin I, Bousquet O, Gelly S, et al. Wasserstein auto-encoders [J]. arXiv preprint arXiv:1711.01558, 2017.

[159] Li Y, Su H, Shen X, et al. DailyDialog: A manually labelled multi-turn dialogue dataset [C]// Proceedings of the Eighth International Joint Conference on Natural Language Processing. 2017: 986-995.

[160] Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

[161] Liu C W, Lowe R, Serban I, et al. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 2122-2132.

[162] Maaten L v d, Hinton G. Visualizing data using t-sne [J]. Journal of machine learning research, 2008: 2579-2605.

[163] Fu T, Gao S, Zhao X, et al. Learning towards conversational ai: A survey [J]. AI Open, 2022: 14-28.

[164] Hou Y, Liu Y, Che W, et al. Sequence-to-sequence data augmentation for dialogue language understanding [C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018: 1234-1245.

[165] Pearl J, et al. Models, reasoning and inference [J]. Cambridge, UK: CambridgeUniversity-Press, 2000, 19: 2.

[166] Zhu Q, Zhang W N, Liu T, et al. Counterfactual off-policy training for neural dialogue generation [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 3438-3448.

[167] Zheng C, Sabour S, Wen J, et al. Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models [J]. CoRR, 2022.

[168] Gangal V, Jhamtani H, Hovy E, et al. Improving automated evaluation of open domain dialog via diverse reference augmentation [C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021: 4079-4090.

[169] Bosselut A, Rashkin H, Sap M, et al. COMET: Commonsense transformers for automatic knowledge graph construction [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2019: 4762-4779.

[170] Robertson S E, Walker S, Jones S, et al. Okapi at trec-3. [C]//Harman D K. NIST Special Publication: 500-225 TREC. National Institute of Standards and Technology (NIST), 1994: 109-126.

[171] Paranjape B, Lamm M, Tenney I. Retrieval-guided counterfactual generation for QA [C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022: 1670-1686.

[172] Yu S, Zhang H, Niu Y, et al. COSY: COunterfactual SYntax for cross-lingual understanding [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). Association for Computational Linguistics, 2021: 577-589.

[173] Liu Q, Kusner M, Blunsom P. Counterfactual data augmentation for neural machine translation [C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021: 187-197.

[174] Qin L, Bosselut A, Holtzman A, et al. Counterfactual story reasoning and generation [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). Association for Computational Linguistics, 2019: 5043-5053.

[175] Hao C, Pang L, Lan Y, et al. Sketch and customize: A counterfactual story generator [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 12955-12962.

[176] Chen J, Gan C, Cheng S, et al. Unsupervised editing for counterfactual stories [J]. Proceedings of the AAAI Conference on Artificial Intelligence: 10473-10481.

[177] Buesing L, Weber T, Zwols Y, et al. Woulda, coulda, shoulda: Counterfactually-guided policy search [C]//7th International Conference on Learning Representations, ICLR 2019. 2019.

[178] Oberst M, Sontag D. Counterfactual off-policy evaluation with gumbel-max structural causal models [C]//International Conference on Machine Learning. PMLR, 2019: 4881-4890.

[179] Luce R D. Individual choice behavior: A theoretical analysis [M]. New York, NY, USA: Wiley, 1959.

[180] Maddison C J, Tarlow D, Minka T. A* sampling [C]//Ghahramani Z, Welling M, Cortes C, et al. Advances in Neural Information Processing Systems: volume 27. Curran Associates, Inc., 2014.

[181] Campos R, Mangaravite V, Pasquali A, et al. Yake! keyword extraction from single documents using multiple local features [J]. Information Sciences, 2020: 257-289.

[182] Bras R L, Swayamdipta S, Bhagavatula C, et al. Adversarial filters of dataset biases [C]// Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020: 1078-1088.

[183] Axelrod A, He X, Gao J. Domain adaptation via pseudo in-domain data selection [C]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011: 355-362.

[184] Xie Q, Dai Z, Hovy E, et al. Unsupervised data augmentation for consistency training [C]// Larochelle H, Ranzato M, Hadsell R, et al. Advances in Neural Information Processing Systems: volume 33. Curran Associates, Inc., 2020: 6256-6268.

[185] Lee N, Bang Y, Madotto A, et al. Towards few-shot fact-checking via perplexity [C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 1971-1981.

[186] Xiao H. bert-as-service [EB/OL]. 2018. https://github.com/hanxiao/bert-as-service.

[187] Shao Y, Geng Z, Liu Y, et al. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation [J]. arXiv preprint arXiv:2109.05729, 2021.

[188] Zhang T, Kishore V, Wu F, et al. Bertscore: Evaluating text generation with bert [C]// International Conference on Learning Representations. 2020.

[189] Smith E M, Williamson M, Shuster K, et al. Can you put it all together: Evaluating conversational agents' ability to blend skills [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 2021-2030.

[190] Li Z, Zhang J, Fei Z, et al. Addressing inquiries about history: An efficient and practical framework for evaluating open-domain chatbot consistency [C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 1057-1067.

[191] Honovich O, Choshen L, Aharoni R, et al. Q2:: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 7856-7870.

[192] Bowman S, Angeli G, Potts C, et al. A large annotated corpus for learning natural language inference [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 632-642.

[193] Agrawal A, Batra D, Parikh D, et al. Don't just assume; look and answer: Overcoming priors for visual question answering [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4971-4980.

[194] Slack D, Hilgard S, Jia E, et al. Fooling lime and shap: Adversarial attacks on post hoc explanation methods [C]//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020: 180-186.

[195] Gardner M, Artzi Y, Basmov V, et al. Evaluating models' local decision boundaries via contrast sets [C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 1307-1323.

[196] Kaushik D, Hovy E, Lipton Z. Learning the difference that makes a difference with counterfactually-augmented data [C]//International Conference on Learning Representations.

[197] Zheng Y, Chen G, Huang M, et al. Personalized dialogue generation with diversified traits [J]. arXiv preprint arXiv:1901.09672, 2019.

[198] Qin L, Xie T, Huang S, et al. Don't be contradicted with anything! ci-tod: Towards benchmarking consistency for task-oriented dialogue system [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 2357-2367.

[199] Jin D, Liu S, Liu Y, et al. Improving bot response contradiction detection via utterance rewriting [C]//Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2022: 605-614.

[200] Wang Z, Culotta A. Robustness to spurious correlations in text classification via automatically generated counterfactuals [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 35. 2021: 14024-14031.

[201] Yang L, Li J, Cunningham P, et al. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis [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: 306-316.

[202] Zmigrod R, Mielke S J, Wallach H, et al. Counterfactual data augmentation for mitigating

gender stereotypes in languages with rich morphology [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1651-1661.

[203] Maudslay R H, Gonen H, Cotterell R, et al. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution [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). Hong Kong, China: Association for Computational Linguistics, 2019: 5267-5275.

[204] Agarwal V, Shetty R, Fritz M. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9690-9698.

[205] Chen L, Yan X, Xiao J, et al. Counterfactual samples synthesizing for robust visual question answering [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10800-10809.

[206] Kolluru K, Aggarwal S, Rathore V, et al. Imojie: Iterative memory-based joint open information extraction [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5871-5886.

[207] Williams A, Nangia N, Bowman S. A broad-coverage challenge corpus for sentence understanding through inference [C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 1112-1122.

[208] Liu Q, Kusner M, Blunsom P. Counterfactual data augmentation for neural machine translation [C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 187-197.

[209] Wang X, Wang H, Yang D. Measure and improve robustness in nlp models: A survey [C]// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022: 4569-4586.

[210] Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 86-96.

致 谢

行文至此，感慨万千，曾经以为迈不过去的博士生涯也即将迎来终点。在此毕业之际，我借此机会衷心地向一直以来所有支持、帮助和关心过我的人表达真诚的感谢！

首先我要由衷地感谢我的博士生导师冯洋老师。2019 年春天，我因为研究工作在原课题组进展不顺利，想要寻找更相关的研究团队继续进行研究工作。非常感谢冯老师当时能够接受我进入课题组，给了我一个机会能够学习并掌握到最前沿的自然语言处理相关知识，并在最前沿的研究领域中进行探索，在最好的课题组氛围中进行自己的博士课题。冯老师渊博深厚的专业知识、勤奋认真的科研态度，以及严谨认真的治学精神对我产生了深远的影响，激励着我不断前进，树立更加远大的科研目标。在科研上，冯老师为课题组创造了浓厚的科研氛围，无论多忙都会定期与每位同学讨论新的想法。在每一次讨论新的科研想法的过程中，冯老师总是提前做足功课，对科研想法和研究方案进行深入思考与分析，听取我的思考方式，并且总是能一针见血地指出想法和方案中的不足之处，及时纠正我在科研道路上的前进方向。此外，冯老师对待科研极为认真，要求同学们对于概念的理解、名称的使用，思维逻辑的链条等都要非常的准确与清晰。随着长时间潜移默化的影响，我在科研上也变得凡事力求准确，并且思维逻辑要缜密清晰。在生活中，冯老师给予大家充分的人文关怀，提倡大家科研之余要保持健康的生活习惯，帮助大家解决生活中所遇到的困难。当我在科研受挫时，冯老师总是能在第一时间给予我充分的关爱与理解，并鼓励我以更好的心态去面对挫折。感谢冯老师给我提供了优越舒适的科研环境、全方位的指导和极大的信任，让我可以在学术道路上自由地探索，让我学会从整体思考研究方向和研究内容，让我学会如何做科研，更学会如何做有价值的科研。在此，谨向冯老师致以我最诚挚的谢意和最衷心的感谢。

其次我要由衷地感谢我的硕士生导师方金云老师。2015 年我进入方老师课题组开始攻读硕士学位，并在 2017 年硕转博继续攻读博士学位。感谢方老师之前能让我在自己感兴趣的方向上自由探索。之前课题组里的研究方向和自然语言处理并没有太大关联，方老师也给了我足够的自由度去进行研究与探索。在我进展缓慢的时候，方老师给了我足够的鼓励与支持，也在帮我考虑应该如何去解决。方老师为了让我在科研道路上更好地前行，也支持我转去和我研究方向更相关的团队继续学习。非常感谢方老师对我的理解，以及给了我这样一次机会。此外，在方老师课题组期间做了大量的横向项目，在方老师的指导下，我的人际沟通能力与团队合作能力有了非常大的提升。而且方老师为人和善，平易近人，方老师的性格也对我产生了深远的影响，要做一个让人舒服的人。衷心感谢方老师在工作态度和为人处世方面对我的谆谆教导，使我受益良多。在此向方老师也致以我由衷的谢意。

感谢在腾讯微信模式识别中心三年校企合作期间张金超师兄和孟凡东师兄对我的帮助。金超师兄博览群书、风趣幽默，对研究有极好的品味和逻辑，教会我如何去做有价值的研究。无论金超师兄工作多忙，每次想找他讨论问题或者寻求帮助，师兄总是很耐心地和我讨论，或者帮我分析问题。还记得 2020 年投 ACL 会议时，金超师兄通宵帮我逐次逐句的修改论文，让我非常感动。在科研上，金超师兄缜密的逻辑思维能力，对问题的洞察能力给我留下了深刻的印象，也给我树立了一个学习榜样。金超师兄时常和我讨论科研问题，帮我修改论文，分析审稿意见，在这个过程中我学到了很多。另外，也非常感谢凡东师兄给予了我很多学术研究上的指导，指导我如何更好的做研究和写论文。在论文被拒之后，凡东师兄对我的鼓励和安慰也让我及时的调整好状态，继续前行。三年的校企合作经历让我受益匪浅。

感谢实验室秘书刘琳老师和程一老师这期间对我科研和生活的帮助。两位老师帮助我解决了不计其数的各种事项，小到费用报销，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，让我能够专注于科研之中。还要感谢教育处李琳、李丹、周世佳、冯钢和李慧等几位老师在日常学习生活和毕业的各个事项上给予我的帮助和关心。各位老师在工作中所展现出的兢兢业业、不辞劳苦的精神将会一直激励和鼓舞我。

感谢课题组的各位同学，很感激能有机会与大家度过这一段时光，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。感谢李秀星师兄对我的帮助与关心。虽然秀星师兄来组里时间不长，但师兄在科研论文上以及找工作上给我提供了非常多的帮助与指导。在生活上，师兄时常关心我们，经常能吃到师兄分享的各种零食。感谢对话团队的小伙伴们，包括秀星师兄、刘舒曼、申磊、李泽康、刘龙祥、赵彤钰、张珂豪、雨田，以及实习生田畅，每次和大家讨论都在不断的扩展我的研究思路，以及对问题有新的看法。大家对待科研认真的态度，不断产出的成果都在不断督促我继续努力。尤其是要感谢舒曼，舒曼一直以来耐心地帮我解决了科研和生活上的各种问题。无私地把自己整理的笔记和踩过的坑都告诉我。舒曼顽强的生命力和做事的执行力也是我学习的榜样。非常感谢舒曼在我艰难的读博生涯中和我一起前行，一起努力，一起毕业。感谢同年毕业的师弟们谷舒豪、邵晨泽、伍 xuan 甫，你们在科研上认真严谨的态度以及在生活中乐观向上的精神将一直激励着我。还要感谢已经毕业的张文师兄、薛海洋、李京谕，在我刚来组里时给我提供了很多的帮助。感谢王树根和张倬诚两位师弟，帮我熟悉课题组服务器上的环境，耐心地为我一次次的解答关于服务器使用上的问题。感谢李绩成，在一起组织秋游活动以及数据标注的时候给我提供了非常多的帮助。感谢谷舒豪、邵晨泽、伍 xuan 甫、杨郑鑫、单勇、郭登级、张绍磊、马铮睿、房庆凯、刘龙祥、黄浪林、桂尚彤、郭雯钰、谢婉莹、卫李赋凌，在科研之余和大家一起玩桌游让我非常开心。感谢郭守涛、杨哲、周 yan、鄢子文、卜梦煜，你们是课题组未来发展的希望，相信你们一定能帮助课题组不断创造新的辉煌！

感谢我的闺蜜们，邓斯桐同学、高安琪同学和费陶同学。我们从小一起长大，一起读小学、读初中高中，再到读大学，再到你们工作我在读书。尽管我们身处不同的城市，但是我们的感情一直都没有变过。你们是我最信任的人，陪我玩，陪我疯，陪我闹。当我分享我的快乐给你们的时候，你们陪我一起哈哈大笑；当我分享我的悲伤给你们的时候，你们陪我一直伤心难过。不管我做什么事情，你们都是在理解我、鼓励我、支持我，想尽办法帮助我，你们是我一路走来强大的精神支柱。感谢你们，祝我们友谊长存！

感谢我的好朋友梁海宇同学。我们相识于 2011 年，想来也已经 12 年了。我们在大学期间一起去吃饭，一起去上嵌入式课，一起做实验，再到一起毕业。你对待生活洒脱的态度，以及自信热情享受生活的样子深深影响着我。你总是能处理好生活中的各种事情，对生活充满好奇心，不断探索、不断体验，让我深受鼓舞。时常也能收到你精心为我准备的各种节日礼物，以及从家里带来的特产，非常的暖心！感谢梁海宇同学一直以来对我的关心！感谢刘欣 yue 妹妹总是能在我状态不佳的时候开导我，安慰我，鼓励我，让我重拾信心。每次和你聊天都能了解到很多新奇的事情，也能看到你热爱生活，勇于探索，闪闪发光的样子，给了我非常大的触动，也让我非常敬佩！还要感谢张梦菲和潘茂，感谢一路以来的陪伴，感谢给我的安慰，关心，鼓励，与肯定！

然后，我要感谢一直支持我的家人们我的爷爷丁荣坤，我的妈妈丁吉艳、我的二姨周晓萍、我的爸爸欧立平、我的二姨夫张守翼、我的弟弟欧玮桢和我的姐姐张潇澜（张帆）。感谢你们一直以来对我无微不至的关心、支持、理解与帮助。感谢你们在我人生低谷的时候始终相信着、鼓励着我。无论何时何地，你们都会给予我最大的帮助，是我最坚强的后盾，让我可以不断上向也可以掉落下来。你们无条件的信任与包容是我是我不断前进的巨大动力！！

感谢我的先生李功波，感谢这些年来对我的关心、鼓励、支持、包容、信任与尊重。你总是在我难过的时候陪在我身边，让我很有安全感；也总是在我脑洞大开和你说一些奇奇怪怪的观点想法时，陪我一起胡说八道；在我对你发脾气的时候也还是在安抚我，让我学会了表达愤怒而不是愤怒的表达。你谦虚谨慎、为人温和的性格潜移默化地影响着我。你对待工作极其认真，对待生活非常投入，让我非常的敬佩。和你在一起，我成为了自己想要成为的那个人。过去的岁月里，我们彼此见证了对方的成长，有欢笑有泪水。未来的日子里，我们携手并肩，共创美好生活！

最后，感谢所有对我的论文提出宝贵意见的老师，以及在完成论文过程中给予我关心和帮助的所有人。

2023 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:

2011 年 09 月——2015 年 07 月，在西安电子科技大学软件学院获得学士学位。

2015 年 09 月——2017 年 07 月，在中国科学院计算技术研究所攻读硕士学位。

2017 年 09 月——2023 年 07 月，在中国科学院计算技术研究所攻读博士学位。

已发表（或正式接受）的学术论文:

(1) Jiao Ou, Jinchao Zhang, Yang Feng, Jie Zhou. Counterfactual Data Augmentation via Perspective Transition for Open-domain Dialogues. The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), December 7–11, 2022, 1635–1648, 2022, Abu Dhabi. (CCF B 类)

(2) Jiao Ou, Yang Feng. Better Learning and Fusing Multi-Granularity Context Representations for Relevant Response Generation. The International Joint Conference on Neural Networks (IJCNN 2021), July 18-22, 2021, 1-8, 2021, online. (CCF C 类)

(3) 欧蛟，冯洋. 基于反事实样本增强的鲁棒对话冲突检测研究. 中文信息学报. 2023. (CCF B 类)

申请或已获得的专利:

(1) 欧蛟，张金超，冯洋，孟凡东. 一种回复信息确定方法、装置、存储介质及电子设备. 申请号：2020104439899

(2) 欧蛟，张金超，冯洋，孟凡东. 对话模型的训练方法、装置、计算机设备及存储介质. 申请号：2020104501940

参加的研究项目及获奖情况:

(1) 2019 年 10 月 - 2021 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) 2019 年 10 月 - 2022 年 9 月：企业委托项目：机器翻译及对话系统的前沿技术与行业应用，项目成员

(3) 2016 年 4 月 - 2017 年 2 月：国家重点研发计划：全空间信息系统与智能设

施管理，多粒度时空对象组织与管理，编号：2016YFB0502300，项目成员

(4) 2018 年，中国科学院计算技术研究所，易方达金融科技博士生奖

(5) 2017 年，中国科学院大学，三好学生