基于深度学习的情感对话响应综述

杨 州[1),2),3)] 陈志豪[1),2),3)] 蔡铁城[1),2),3)] 王宇峰[1),2),3)] 廖祥文[1),2),3),4)]

[1)](福州大学计算机与大数据学院 福州 350108)

[2)](福州大学福建省网络计算与智能信息处理重点实验室 福州 350108)

[3)](数字福建金融大数据研究所 福州 350108)

[4)](鹏城实验室新型网络研究部 广东 深圳 518000)

摘 要 人机对话作为人工智能的重要领域,以其方便快捷的交互特点广泛应用于任务型和闲聊型机器人等诸多商业场景,并被视为新一代人机交互的主要形式.但情绪感知与表达能力的缺乏致使人机对话技术在复杂交互场景中难以满足人们对情感交流的强烈需求.为弥补人机对话技术中情感智能的缺失,基于深度学习的情感对话响应任务被提出且已发展为对话领域中一个重要的研究方向.本文首先回顾了基于深度学习的情感对话响应任务的发展历程,其次按照任务将情感对话响应分为可控情感对话生成、共情对话响应、情绪支持、多模态情感对话生成、新任务五类.随后本文也按照常用的结构将模型进行了归类与分析,以求更细致地阐述各种结构在情感对话响应任务中的具体用法,之后介绍了常用数据与评测指标.最后本文也对模型进行了总结,并在此基础上进一步展望了该任务未来的发展方向.

关键词 情感计算;情感分析;对话系统;文本生成;自然语言处理;深度学习

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2023.02489

A Survey of Deep Learning Based Emotional Dialogue Response

YANG Zhou[1),2),3)] CHEN Zhi-Hao[1),2),3)] CAI Tie-Cheng[1),2),3)] WANG Yu-Feng[1),2),3)]
LIAO Xiang-Wen[1),2),3),4)]

[1)](*College of Computer and Data Science, Fuzhou University, Fuzhou* 350108)

[2)](*Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, Fuzhou University, Fuzhou* 350108)

[3)](*Digital Fujian Institute of Financial Big Data, Fuzhou* 350108)

[4)](*Department Of New Networks, Peng Cheng Laboratory, Shenzhen, Guangdong* 518000)

Abstract As an important field of artificial intelligence, human-machine dialogue is regarded as the main form of the new generation of human-machine interaction. Due to its convenience, human-machine dialogue is widely used in many business scenes, such as task-based dialogue systems and chat robots. In real-life scenes, conversations are often accompanied by emotional exchanges. However, without emotion perception and expression capabilities, human-machine dialogue technology fails to complete emotional communication in such scenes. In order to make up for the lack of emotional intelligence in human-machine dialogue technology, the deep learning based emotional dialogue response task has been proposed and developed into an important research direction in the field of dialogue.

收稿日期:2022-10-26;在线发布日期:2023-07-28. 本课题得到国家自然科学基金项目(No. 61976054)资助. **杨 州**,博士研究生,主要研究领域为自然语言处理.E-mail:200310007@fzu.edu.cn. **陈志豪**,博士研究生,主要研究领域为社交网络信息传播分析、情感分析. **蔡铁城**,博士研究生,主要研究领域为情感计算,图像描述生成和自然语言处理. **王宇峰**,硕士研究生,主要研究领域为自然语言处理. **廖祥文**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、信息检索、社会计算和网络内容安全.E-mail:liaoxw@fzu.edu.cn.

In this paper, we first review the development of deep learning based emotional dialogue response task. According to different goals, we then divide the task into five subtasks: controllable emotional dialogue generation, empathetic dialogue response, emotion support, multimodal emotion dialogue generation, and new tasks. Controllable emotional dialogue generation focuses on how to generate responses with specified emotions. According to different components for handling emotions, we divide these models into emotion perception models, emotion representation models, and emotion perception representation models. Empathetic dialogue response aims to automatically perceive emotions and express empathetic responses. According to different factors influencing empathetic perception and expression, we divide this category of models into emotion factor-based models, compound factor-based models, and structural factor-based models. Emotion support is to regulate the speaker's emotion to comfort the user's feelings. Due to the different types of emotions involved, we divide this type of models into three subtypes: stimulating positive emotions, stimulating specified emotions, and reducing negative emotions. Multimodal emotion dialogue generation models focus on multiple modalities, such as images, audio, and text. Since the number of models in this category is relatively small, we do not further classify them. Subsequently, we also group the new tasks proposed in recent years into one category, including three sub-tasks: multi-person empathic dialogue, value task, and language toxicity mitigation. Based on the above categories, we further compare the models' advantages and disadvantages and thus look forward to the development trends of each subtask.

In order to further explore the deep learning based emotional dialogue response, we also organize and analyze the models according to commonly used structures. Since most models are built based on sequence-to-sequence structures, we list and analyze the improvement methods of sequence-to-sequence structures. Given the rapid development of pre-training models in recent years, some models have adopted pre-training models to enhance the effectiveness of the models. We also summarize the pre-training based models on deep learning-based emotional dialogue response task. At the same time, we also summarize more structures involved in this task, including generative adversarial networks, reinforcement learning, dual learning, and template filling. To further explore the advantages and disadvantages of the above structures, we conduct comparisons and analysis of these models.

Then, we introduce the commonly used data and evaluation metrics. Finally, we summarize the models and further discuss the future development directions of the task.

Keywords affective computing; affective analysis; dialogue system; text generation; natural language processing; deep learning

1 引 言

对话系统是一种以自然语言方式与人类用户交流的计算机程序[1],它常被应用于陪伴交流、知识问答、商品推荐等多个场景,创造了巨大的商业价值.随着人工智能不断发展,各个场景要求对话技术不仅能够完成基础功能,更能给予用户精准可靠的极致体验与优质全面的人性化服务.庞杂的场景与急迫的需求向研究者们提出了充满挑战的问题:如何提升对话质量?

1.1 情感对话响应的提出

为提升各个场景中用户的对话体验,研究人员从多个子方向探索了对话系统.如图1所示,按照构建系统的目的,学者们从任务导向型和非任务导向型两个方向进行了探索[2].任务导向型的对话系统为满足垂直业务需求而被提出[3],按照是否将各个模块独立优化被分为管道型和端到端两类[4];非任

务导向型的系统主要是针对于开放领域，又被称为聊天机器人，按照建模方式的不同被分为检索、文本生成、检索融合生成三类[4].

情感(Sentiment)①指的是观点背后一种积极或消极的感觉[5]. 而情绪(Emotion)为神经生理变化带来的精神状态，与思想、感觉、行为反应等有不同程度的联系[6-9]. 赋予对话系统情感或情绪特质能够为用户进一步带来极致体验与优质服务. 达尔文(Charles Robert Darwin)在《人与动物的感情表达》一书中认为情绪是自然的产物[10]；Mayer和Salovey[11]提出人类具有情绪感知、整合情绪、理解情绪、调节情绪的情感智能(Emotional Intelligence，EI)，人类日常对话不仅会进行信息交换，也会伴随着情绪的交换；积极的情绪能够提升学习效率[12]；适当的情绪诊断能有效判断病人心理状态以预防抑郁甚至自杀[13]；合适的情绪反馈能够提升用户满意度[14]、降低对话系统错误率[15]；贴心的情感陪护机器人能够缓解全球人口老龄化[16]带来的社会压力.

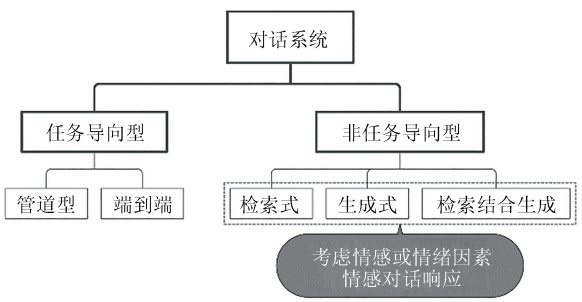

图1　对话系统的子方向

总体来说，对专业领域的细致化研究为不同对话场景提供了更为精准的解决方案，为机器人融入情感智能也会进一步提升用户的对话体验. 情感对话任务的研究主要聚焦于闲聊领域，以实现极致体验与优质服务为目标而被提出.

目前，情感对话任务可分为对话情绪感知和情感对话生成两个子任务[17]. 前者侧重于对话中情感的分类，后者侧重于生成具有情感的回复. 本文关注于如何使机器产出具有情感的回复，即情感对话生成任务. 然而，在情感对话生成任务的发展过程中，基于检索和检索结合生成式的情感对话模型已经在近年来被提出. 仅用情感对话生成一词概括该任务已不适用. 因此，本文介绍了包括检索式、生成式、检索结合生成式三种类型的方法，并将该任务统称为情感对话响应任务.

1.2　发展历史

基于深度学习[18]的情感对话响应(简称深度情感对话响应)任务属于对话任务中的情感对话支系，按照历史发展进程可将情感对话响应分为对话机器人、基于规则和统计的情感对话系统、基于深度学习的问答系统、基于深度学习的情感对话响应模型四个历史时期. 研究者们在首个时期提出了对话机器人的构想，并实现了首个对话机器人；随后的时期则在对话系统中引入了情感因素，研究了基于规则与统计的情感对话机器人，并为深度情感对话响应模型的提出打下了坚实的理论基础；之后基于深度学习的问答模型被大量研究，为深度情感对话响应模型起到了一定的借鉴作用；最后基于深度学习的情感对话响应模型被提出并演化出了结构各异的模型. 发展脉络如图2所示.

(1)对话机器人

早在1950年，图灵(Alan M. Turing)在Mind上发表文章“Computing Machineryand Intelligence”[19]，并提出“机器人能思考吗?”一问，从此构想了对话机器人的概念. 1966年，麻省理工学院的Joseph Weizenbaum首次实现了一个公开的聊天机器人Eliza[20]，它是基于关键词匹配和人工规则的机器人. 该时期对机器人的初步构想与实现为情感对话响应任务带来了里程碑式的影响.

(2)基于规则和统计的情感对话机器人

1972年，斯坦福大学的精神病医生Kenneth Colby基于Eliza构建了一个模拟精神病人的机器人Parry[21]，它是首个带有情感色彩的聊天机器人；1995年，Richard S. Wallace开发了一个采用启发式模板匹配的对话策略的机器人ALICE(Artificial Linguistic Internet Computer Entity)，它通过开源代码后进一步提升了系统的性能；2014年，微软公司推出的智能机器人小冰[22]，它通过检索技术结合情感和语义产出高质量对话回复. 时至今日，该人工智能机器人仍在不断迭代更新. 总体来说，传统的情感对话机器人针对特定场景并通过设定规则和统计特征构建对话机器人，由于其拓展成本高、规则需要依赖于专业人员设计、领域限制等缺点逐渐被研究者们遗弃.

① 情感或情绪：为中文表述的便捷，本文将Sentiment和Emotion的统称为情感，而情绪特指更细致的情感，如开心、难过、兴奋等.

(3)基于深度学习的问答系统

在深度情感对话响应任务发展过程中,研究者们借鉴了大量基于深度学习的问答(简称深度问答)任务的建模方式,例如对偶学习、引入外部知识库、检索式模型等.因此,深度问答任务某种程度上可视为深度情感对话响应任务的奠基石,并起着至关重要的铺垫作用.深度问答模型分为基于表示的模型、基于交互的模型和混合模型.2014年,Yu等人[23]利用卷积神经网络构建了基于表示的模型,该类模型将查询和答案利用神经网络独立学习后进行评分,它侧重于语义的学习和表示;2015年,Wang等人[24]使用双向长短记忆网络(Bidirectional Long Short-term Memory,BiLSTM)构建了基于匹配的模型,该类模型侧重于学习查询和回复的交互信息,如Cosine相似度等.2016年,Yin等人[25]提出了基于注意力的卷积神经网络(Attention-based CNN,ABCNN),以此构建了一个混合模型,该类模型融合了基于表示和基于交互模型的优点,能够有效捕捉交互信息和语义信息.

图2 情感对话响应历史发展脉络图

(4)基于深度学习的情感对话响应模型

基于深度学习的情感对话响应模型将人工神经网络(Artificial Neural Network,ANN,简称神经网络)作为模型的构建元素,通过将情绪融入模型使对话更加流畅地进行.为深入探究情感与对话的潜在关联,文本按照融入情绪的目的不同,将模型分为可控情感对话生成、共情对话响应、情绪支持、多模态情感对话生成四个方面.同时,为了着重强调新提出的任务,本文也额外将新任务归为额外的一类.

在深度情感对话响应任务发展的早期,研究人员着重专注于如何在对话中表达指定类型的情绪,文本将其称为可控情感对话生成.2017年,Zhou等人[26]采用序列到序列结构(Sequence To Sequence,Seq2Seq)[27]在解码器融入情绪类别以控制模型生成指定情绪类别的回复.从此开启了可控情感对话生成模型的研究热潮.

共情对话响应任务要求模型具有感知说话者情绪,并表达合适情绪的能力.因此,该类模型一方面要求模型在未给定情绪标签的情况下正确感知说话者的情绪,另一方面也需要通过将感知到的情绪合理表达于回复中,使模型具备同理心.

情绪支持任务不仅要求模型具备一定程度的同理心,同时也要求模型能够通过对话给予说话者一定的精神支持和安慰,并起到一定的心理慰藉作用.该类模型可视为情感对话响应模型在精神卫生(Mental Health)领域的一个重要应用.

以上三类模型大多都是基于文本模态进行构建的.而在现实场景中,多模态能够提供更多了解情绪的视角.为了更好地感知对话中的情绪,一些研究者也研究了多模态情感对话任务.

最近,随着情感对话响应任务的逐步发展,更多的任务被提出.例如多人共情对话、价值驱动、减轻语言毒性的任务被研究者所提出.

1.3 任务形式化定义

深度情感对话响应任务具体为:给定说话者和回复者的对话历史信息U,模型需要通过自身感知情感或用户指定情感E,产出情感恰当、语言自然的回复语句R.其中查询P特指对话历史U的最后一句话;反馈语句F为回复之后说话者的话语,部分研究者会通过反馈语句进一步优化模型.

本文列举了一个关于情感对话响应的例子，如图3所示．给定对话历史U，包括说话者的“我订到回家的机票了！”、回复者的“过节的机票是非常难定，恭喜你！”以及查询P(“太激动了，我下周就要看望我三年没有见的父母了！”)，模型需要理解对话历史的内容，并根据理解的情感或指定的情感产出合适的回复R.

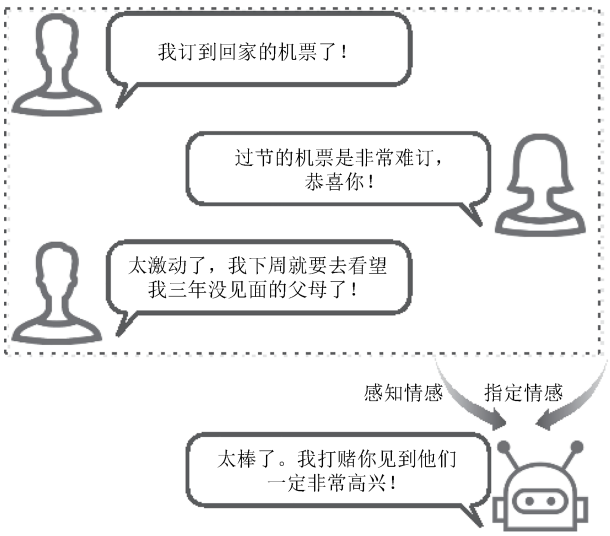

图3　情感对话响应示例图

为了清晰地说明各个子任务目标的差异，本文也对各个子任务的目标进行了说明：

(1)可控情感对话生成．在该子任务中，模型需要将用户指定的情感E_r作为输入，并产出具有指定情感E_r的回复R.

(2)共情对话响应．共情对话响应任务要求模型能够自动感知用户所表达的情感E_r，并产出情感合适且具有共情性的回复R.

(3)情绪支持．情绪支持任务中，模型需要激发特定的情绪E_r来慰藉说话者．该任务旨在提升说话者的积极感受或降低说话者的负面情绪，以达到陪伴或心理健康支持的目的．

(4)多模态情感对话生成．该子任务中，模型需要根据多模态的对话历史信息(图片、语音、文本模态)来感知情绪E_r，并表达出情感恰当的回复．

此外，当对话语句仅有说话者的单条语句(即查询语句，也称为查询)时，本文将该类对话称为单轮对话，单轮对话的对话历史为查询，即$U=P$，模型仅根据查询P及情感E_r产出相应的回复．由于该任务仅有少量的模型是单轮对话的，本文则未将模型按照单轮、多轮方式进行梳理．

1.4　组织结构及说明

本文关注于基于深度学习的情感对话响应模型．本文第一章简述了该任务的发展及任务定义；第二章按照子任务对该领域模型进行了划分并给出了详尽的说明和分析；第三章按照常用的模型结构对该领域模型进行了概述；第四章阐述了常用的数据集和评价指标；第五章对模型进行了总结和展望．组织结构如图4所示．

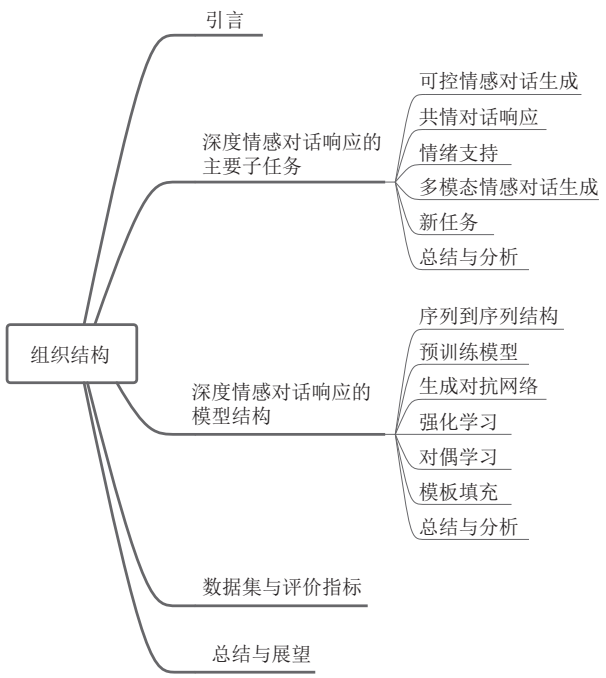

图4　深度情感对话响应任务组织结构图

2　深度情感对话响应的主要子任务

由于深度情感对话响应任务包括了不同的子任务，且构建模型的结构复杂多样，本文按照构造模型目的不同，将该任务拆分为五个方向：可控情感对话生成、共情对话响应、情绪支持、多模态情感对话生成以及新任务．其中可控情感对话生成模型是主要关注指定情感的生成模型．不同于前者主要关注于指定情感的回复生成，共情对话响应模型关注于情绪的自动感知和共情回复的生成．而情绪支持任务是共情对话响应在精神卫生领域的一种应用任务，这类模型关注于减缓说话者的负面情绪或激发其积极情绪．前两类模型仅基于文本构建模型，而现实场景往往是多模态的．因此，本文将多模态情感对话生成模型也归纳为一个重要的类别．最后，本文也介绍了近年来提出的关于情感对话的新任务．

2.1 可控情感对话生成

可控情感对话生成旨在通过指定特定的情绪，使模型生成具备该情绪的回复．该类模型将序列到序列的结构作为主体框架，并在编码器、解码器或两者一起进行改进以表达特定的情绪．根据改进序列到序列结构的部位不同，本节将模型分为情绪感知模型、情绪表示模型和情绪感知表示模型三类．因编码器能够直接感知用户的情绪，而解码器着重于情绪的表示，本文将改善编码器、解码器、两者一起的模型分别称为情绪感知模型、情绪表示模型和情绪感知表示模型．

2.1.1 情绪感知模型

该类模型着重于情绪编码器的改进．由于更改输入序列会直接影响编码器，因此本文将更改输入序列的方式也纳入情绪感知模型．Sun等人[28]在可控情感对话生成任务的早期以序列到序列结构为框架，将查询、查询情感、回复情感拼接到一起送入编码器，以促使模型获得情绪的感知能力；同时，Shen等人[29]利用了两个独立的HRED结构[30]使对话者的语言分开独立编码，并以变分自编码器融合了情绪与语义；随后Zhou等人[31]、Zhang等人[32]采用变分自编码器进一步研究了情绪与语义的融合方式；而Kong等人[33]则在Zhou等人[31]基础之上使用对抗生成网络进一步优化了模型；Peng等人[34]则发现了情绪的表达和说话的主题之间存在一定的关联，此后，杨等人[35]认为前者在聊天时仅注重一个话题导致对话无趣，因此提取了多个主题词优化模型．

总体来说，情绪感知模型的一个优势是构造模型的方式简单．通过改进编码器的结构，该类模型能够更好的感知情绪．该类模型的另一个优势是在编码过程中能够注重到词粒度的语义和情感的关联，有利于情感的准确表达．但由于回复是直接由解码器生成，在编码器中指定情绪则只能较为间接地表达情感．

2.1.2 情绪表示模型

该类模型通过将情绪融入解码器，直接指定了回复的情绪．Yuan等人[36]在使用多跳注意力[37]增强模型编码能力的同时也认为解码的首个词十分重要，采用了LST(Learning to Start)[38]生成了首个词；Zhang等人[39]随后则为每种情绪分别构建了解码器．为找到最合适的回复，该模型采用了一种内部排序机制筛选出最优的回复；之后Shantala等人[40]则将情感类别简单地随机初始化为情感向量后加入解码器，以促进情绪的表达．这一方式被后面的很多研究者所采用；随后Zhou等人[26]注意到了情绪在表达时具有逐步衰减的机制，并利用外部词库增强了情绪词的表示．更重要的是，Zhou等人也同时构造了一个用于可控情感对话生成的数据集，进一步促进了该任务的发展．随后，不同的研究者发现了情绪与回复的多种关系．Song等人[41]发现情绪不仅表达在显性的词上，也隐式地表达在不具备情绪的词上．比如具有特定语法结构的隐性情绪表达，例如查询“I bought a beautiful dress yesterday!(我昨天买了一件漂亮的连衣裙!)”和回复“Wow, you must feel walking on air!(哇，你一定觉得在空中行走!)”隐性表达了说话者羡慕和开心的情绪；最后，Firdaus等人[42]也发现了人类表达情绪时，往往掺杂着多种情绪，而且表达的情绪强度各异．因此该工作在提出了一个多情绪伴随情绪强度的数据集后，也研究发现了情绪具有混合的特性．

以上的工作都是通过手动指定回复情绪的，另外一些模型通过预训练模型引导回复情绪的产生．Zhou等人[43]利用预训练分类器为查询进行情绪分类后，提出了以情绪为导向的注意力机制，并以此引导回复生成；Zhou等人[44]则将预训练的情绪分类器标记的情绪类别放入损失函数，使模型生成更靠近该情绪的回复．

概括来说，在解码器中考虑情感因素能使情感较为直接地表达在回复中．但这种表达方式不能产生共情的回复，也并不能结合对话历史的情感表达出更为自然的情感．

2.1.3 情绪感知表示模型

该类模型注重了编码器和解码器两种层面对情绪的影响．Asghar等人[45]首先在词嵌入上拼接VAD向量[46]促使模型感知词粒度的情绪，另一方面也在损失函数上优化生成回复与标准回复的情感相似性，最后在词和句子层面提出了基于情感的束搜索算法；随后，Huang等人[47]和Sun等人[48]注重到了情绪标签对于情绪表达的影响．前者将情绪分别加入查询序列的前、后，拼接到解码器输入序列以验证不同方式对情绪表达的影响，发现拼接到解码器在整体的指标上优于前面两种方式；而Sun等人[48]则将查询情感分别放到了查询词前用于编码，随后将回复情感放到了解码的第一个词用于解码，该工作意识到了说话者和回复者在对话中可能会表达不同的情感．

此后，Colombo等人[49]注意到了情感存在于词级和话语级，并使用随机化的情感向量表示句子级

情感的同时，也使用VAD向量[46]促进了词级情感的表达；Peng等人[50]也从词级和话语级情感两方面进行了建模．该模型利用了情绪词典提取了情绪词，并利用预训练的情绪分类器获得了话语级情感．通过结合这两种情绪，模型更好地感知和表达了情感；以上的模型都在闲聊型对话中采用，Shi等人[51]则利用文本和语音结合强化学习提出了一个基于任务型的对话系统；进一步地，以上模型都建模了查询到回复的情绪一致性，而忽略了对话中查询和回复具有双向一致性．Shen等人[52]因此引入了对偶学习(Dual Learning)[53]建模了情绪和内容的双向一致性；然而，以上的模型要么基于序列到序列结构结合交叉熵损失函数优化模型，要么采用强化学习利用反馈(Reward)优化模型．前者不能很好的注重情绪词汇，后者存在搜索空间过大问题[54]．Dong等人[54]因此结合了模拟退火算法(Simulated Annealing)构建了搜索的优化方式使整个生成过程逐步收敛．情绪感知表示模型融合了两类模型的优势，但正因为在模型中既能感知对话中的情绪，又需要表达指定的情绪，若两者之间的情绪不一致，则会影响情绪在回复的表达．

2.1.4　总结与分析

总的来说，可控情感对话生成模型主要依托于序列到序列模型，通过指定特定的情感生成具有情感的回复．该子任务中，模型通过改进编码器、解码器或兼具两者以使模型表达用户所指定的情感．按照模型的数量来看，研究者热衷于改进解码器来提升情感的控制能力．其次，改进编码器和解码器两者的模型数量多余编码器．这说明了相对于改进编码器，在解码器中添加情感因素能够更加直接地控制情绪表达．同时改进解码器和编码器能够结合两者的优势，例如Dong等人[54]提出了一套注重情绪词汇的模拟退火算法，从编码器和解码器两方面改进了模型，取得了目前的最优结果．但若编码器感知的情感和指定的情感不一致，易导致情绪较难以准确控制．

这类模型大多出现在情感对话任务的早期，以研究如何将情感融入模型为目的，促使模型合理的表达情绪．该类模型开启了情感对话的先河，并为以后的研究提供了数据基础．但这些方法需要控制情绪的产生，不能灵活自由地表达情感．同时，该任务并没有特定的应用场景，导致模型的实际用途较为受限．因此，一部分学者聚焦于共情对话领域，旨在使模型能够自动感知说话者的情感，并表达出共情的回复，以更好起到陪伴和呵护的实际作用．另外一部分聚焦于情绪支持领域，试图减缓用户的消极情绪，提升积极情绪，以起到心理慰藉的实际用途．由于需指定特定情绪与应用场景受限的特点，该任务一方面会发展为特定场景下自动感知情绪的任务．另一方面，该任务也会趋于发展为特定场景下调节特定情绪的情绪控制任务．相关文献如表1所示．

表1　可控情感对话生成模型

类型	典型模型	优缺点
情绪感知模型	Sun等人[28]；Shen等人[29]；Zhou等人[31]；Zhang等人[32]；Kong等人[33]；Peng等人[34]；杨等人[35]	优点：方法简易；能融合词粒度情感和语义 缺点：间接指定情感
情绪表示模型	Zhou等人[26]；Yuan等人[36]；Zhang等人[39]；Shantala等人[40]；Song等人[41]；Firdaus等人[42]；Zhou等人[43]；Zhou等人[44]	优点：直接指定情感 缺点：不能结合对话历史的细粒度情感
情绪感知表示模型	Asghar等人[45]；Huang等人[47]；Sun等人[48]；Colombo等人[49]；Peng等人[50]；Shi等人[51]；Shen等人[52]；Dong等人[54]	优点：结合了两类的优势 缺点：易引入噪音

2.2　共情对话响应

共情指的是通过理解他人感受，并合理地做出回复的能力[55]．共情对话响应旨在正确感知说话者的感受，并根据这种感受生成合理的回复．共情对话响应任务包括了大量的生成式模型，同时也包含了少量的检索式、检索结合生成式模型．为了增强模型的共情能力，研究人员从三类因素对共情对话响应模型进行了探索．首先是情绪因素．由于情绪是共情的重要因素[56]，一些研究人员考虑了情绪这种单一因素增强模型的共情能力．随后是复合因素．情绪虽然是共情的重要因素，但却不是唯一的因素[56]．因此，一些研究者结合情绪和其他多种因素探究了共情对话响应模型．最后，以上两种模型大多是基于序列到序列结构建模的．而一些研究者也发现改善模型的结构能够提升模型的共情能力．本节将这类模型归为以结构因素而构建的模型．

2.2.1 基于情绪因素的模型

在该任务的早期,研究者以情绪标签引导模型感知对话中的情绪.由于人类在感知和表达情绪时往往会掺杂多种情绪,一些研究者模拟人类这一特性研究了混合情绪.仅以情绪标签引导模型学习情绪往往会忽略细微的情绪表达.随后,一些研究者也额外的注重了情绪词引起的细微情绪.虽然这些模型从多种角度建模了情绪,但它们忽略了查询和回复之间的情绪关联.由此,另外的一类探索查询和回复之间情绪关联的模型被提出.

(1)话语级情绪(Dialogue-Level Emotion)

共情对话任务的早期,Rashkin等人[57]提出了一个共情对话数据集,这为后来的模型奠定了数据基础.该工作中,EmoPrepend-1模型通过预训练情绪分类器为对话语句打上情感标签,并拼接到查询作为输入,促使模型生成更具有情感的回复.而MultiTRS则为查询信息进行情绪分类使模型感知情绪,并同时生成合适的回复;Xie等人[58]引入LIWC[59]为对话的每个句子构建了情感向量,并通过用门控循环单元结构(GRU)学习多个句子的情感表示使模型获得话语级的情感;Shin等人[60]采用了强化学习,将反馈语句作为反馈得分,进一步优化了回复的情绪.

(2)混合情绪(Mixed Emotions)

人类在表达情绪时,常会伴随着多种情绪.注重话语级情绪的模型仅将某种情绪作为生成回复的引导因素,而注重混合情绪的模型则模拟人类表达混合情绪的特性,以多种混合情绪引导回复生成.Lin等人[61]注意到了这种混合情绪的存在,并利用多解码器按概率结合了多种情绪.而Majumder等人[62]则认为情绪的表达依托于积极情绪组或消极情绪组,因此将多种情绪分为积极或消极组后融入了模型.而实验结果则表明Majumder等人[62]提出的模型更具备共情性.

(3)词级情绪(Word-Level Emotion)

Zhong等人[63]注重了否定词与词强度的作用,它考虑解码词权重、前一个词强度或是否为反义词(如不、非等词)带来的影响;Li等人[64]则利用词典提取出情绪词后以预测话语级情绪,同时也探讨了反馈语句的作用;而一些研究者则引入了情绪归因模型,从而更加准确地找到引起该情绪的诱发词.该研究分别从多个角度设计了探索诱发词的方法.其中Gao等人[65]按概率设计了软控门,以Gumbel-Softmax[66]作为硬控门控制情感词在回复中的表达;Kim等人[67]则使用情绪归因模型识别情绪的诱发词,并将说话者和回复者之间的交流表述为概率推理,以增加生成文本的信息量;Li等人[68]则将情绪归因的信息拼接到查询后面,输入预训练模型以捕获情绪归因信息.

(4)情绪关联

一些研究者发现查询和回复之间具有情绪和内容的关联,具体表现为一致性(Consistency)和转移性(Transition).部分研究者认为查询和回复之间的关联表现为情绪和内容具有一致性,而另外一些研究者则认为情绪在表达的过程中会不断地转移.Wei等人[69]认为在查询和回复之间的情绪和语义具有一致性,使用SSWE[70]和Word2Vec[71]表示情绪和语义,并对它们独立编码后进行融合后解码;随后Wei等人[72]在前者的基础上将融合网络改为变分自编码器,以进一步提升融合机制;Shen等人[73]则认为一致性不仅表现在查询到回复的单一方向,更具有双向的流动关系,因此该工作引入了对偶学习构建了模型.

Chen等人[74]意识到对话语句之间在内容和情绪上也具有一定的转移性,而以往整个对话仅有一个情绪标签,并不能促使模型学习到这种转移性.因此该研究提出了具有话语级情感标签的数据集,旨在学习情绪的转移;Wen等人[75]则认为个人特质会导致对话的情绪转移,并采用了五维向量表现人格特质,以此构建了一个关于情绪转移的数据集;Kim等人[76]认为情绪的转移更细致地体现在情绪和关键词上,并采用RoBERTa[77]结合沟通机制[78]检测情绪和情绪反应,同时使用Zou等人[79]的检测关键词方法提取并融合了这些特征.

总体来说,对话中的情绪是以多种方式存在的.早期的工作以情绪标签引导模型学习话语级情绪.这些工作能够感知全局的情绪,但却容易忽略细致的情绪,例如忽略表达情绪的词.同时,这些模型仅仅以某一种情绪引导回复生成,而人类在表达情绪时,往往掺杂着多种情绪.针对于第一个缺陷,一些研究额外的关注了词级情绪.这类模型虽然关注了话语级情绪和词级情绪,但它们并未考虑人类会同时表达多种情绪的习惯.针对于第二个缺陷,另外一些研究则注重了多种情绪的表达.这更加符合人类表达情绪的习惯.但这类模型并未注意细微情绪的检测和表达.此外,一些研究者发现查询和回复之间具有情绪和内容的关联性.部分研究者认为这种关联性表现为一致性,而另外一些研究者则

认为这种关联性表现为转移性．从某种程度来说，查询和回复之间同时具有两种性质．但这些模型也忽略了细微的情绪和情绪混合表达的特性．文献的分类及优缺点如表2所示．

表2　基于情绪因素的模型

类型	典型模型	优缺点
话语级情绪	Rashkin等人[57]；Xie等人[58]；Shin等人[60]	优点：融入了全局情绪 缺点：较为粗糙；忽略了混合情绪
混合情绪	Lin等人[61]；Majumder等人[62]	优点：更符合人类情感表达习惯 缺点：未注意细微情绪
词级情绪	Zhong等人[63]；Li等人[64]；Gao等人[65]；Kim等人[67]；Li等人[68]；	优点：细致易捕捉 缺点：忽略了混合情绪
情绪关联	Wei等人[69]；Wei等人[72]；Shen等人[73]；Chen等人[74]；Wen等人[75]；Kim等人[76]	优点：关注了情绪的一致性或转移性 缺点：忽略了细微情绪及混合情绪

2.2.2　基于复合因素的模型

共情的理解和表达不仅仅依赖于情绪这一单一因素，其他更多的因素也会影响共情回复的感知和表达．本节将考虑额外因素的模型归为五类说明．

（1）常识（Commonsense Knowledge）

研究人员发现常识对于对话的理解和表达具有积极作用，也提出了对于常识的三种基本理解和应用方法．Zhong等人[80]认为常识是理性的重要基础，也是理解对话的基础．该工作以情绪为关联构建了基于常识的图谱，并用于情绪的表达；Li等人[81]认为常识有助于理解对话中的隐含情绪，同样地构建了一个情感情景图，并改进了编码器和解码器以融入这些知识；Sabour等人[56]利用常识来获取更多的认知理解，该工作在ATOMIC[82]常识性推理数据集上训练了COMET[83]模型，使其具备常识性推理能力，并将这种能力结合到对话之中．

（2）意图（Intent）

有的研究者认为意图与共情对话有关联．Chen等人[84]认为以往模型仅仅用“同情”的意图表现共情性，而单一的意图不能使模型产生更相关、互动性更强的回复，因此研究者使用多意图来解决这个问题．EmpHi模型通过预测潜在的共情意图，并结合内隐和外显意图表示生成与预测意图相对应的回复．

（3）个人特质（Personality）

Zhong等人[85]认为情感特点包含在个人特质中，并将某个人说的所有话语作为他的特质．该研究使用BERT[86]构建了一个检索模型，结合CoAttention交互学习了一跳编码、二跳编码后，再对回复计算匹配得分以检索最优回复；Wen等人[75]则认为个人特质会影响情绪的转移，因此提出了一个基于个人特质的情绪转移任务；Firdaus等人[87]提出了一项新的任务，即通过对话上下文生成移情的、个性化的回复．该研究者随后提出了一个生成对抗方法，模型在产生移情回复的同时，能够保持不变的人物个性．

（4）对话主题（Topic）

说话者之间不仅与情感有关联，也与主题有联系．Li等人[88]认为传统的序列到序列方法仅可从单方向生成语句，并不符合人类的习惯，同时也注意到了主题和情绪对于对话的影响．该工作由此提出了根据主题和情绪词从两端生成的方式构建回复语句；此后，孙等人[89]对前者的方法进行了改进，该工作在前者的基础上加入了强化学习，以流畅度、主题度、情感度作为反馈（reward）以优化模型；随着预训练模型的提出，Zandie等人[90]将主题和对话动作作为标记输入，并在预训练模型进行多任务微调以使模型获得共情能力．

（5）多种因素

以上模型仅考虑了除情绪以外的单一因素．另外一些模型则考虑了多种因素来进一步提升模型的共情能力．Sharma等人[78]发现共情对话存在沟通机制（Communication Mechanism），并把这种沟通机制分为情绪反应（Emotional Reactions），解释（Interpretations），探索（Explorations）三类，随后Zheng等人[91]将沟通机制、对话动作（Dialog Act）和情绪（Emotion）视为表达共情的三个关键因素．Zheng等人[91]使用GPT-2预训练模型[92]层次化地、递进式地预测这三种机制，随后将其结合以生成回复；而Zandie等人[90]仅注意到了主题和对话动作两种因素，采用了预训练模型进行了微调，促使模型生成共情回复．由于两者在不同的数据集上验证了模

型，导致两者效果无法比较．但相对于后者，前者提出的理论更加的完整．

本节从复合因素的角度介绍了前人的工作，即从情绪与多种重要因素的层面说明了情绪及多种因素在共情对话响应任务中所起到的作用，本节介绍的文献如表3所示．由于感知和表达共情具有复杂的特点，影响共情的因素仍然未被完全探索．因此，进一步探索多种因素对共情对话响应任务的影响是未来的一个重要研究方向．

表3 基于复合因素的模型

情绪相关的特质	典型模型
常识	Sabour等人[56]；Zhong等人[80]；Li等人[81]
意图	Chen等人[84]
个性化	Wen等人[75]；Zhong等人[85]；Firdaus等人[87]
对话主题	Li等人[88]；孙等人[89]；Zandie等人[90]
多种因素	Sharma等人[78]；Zandie等人[90]；Zheng等人[91]

2.2.3 基于结构因素的模型

区别于前两节讲述的基于序列到序列架构的生成式模型，本节将介绍另外的从结构方面改进模型以提升模型共情性的方法，包括不同于常规的生成式、检索、检索融合生成、改写回复四种方法．

首先是不同于常规的生成方式．Shi等人[51]采用了基于模板填充的方式构造了基于任务型对话的模型．该工作使用了两个预训练的情绪分类器为对话生成情绪嵌入，并通过记忆模块捕捉情绪，随后通过模板填充的方式生成回复．其次是检索的方式．由于数据集的缺乏，共情对话响应任务较少出现检索式模型．直到Qiu等人[93]提出了一个基于检索式的模型．该研究者关注了对话中的隐含情绪，并结合语义构建了一个情绪转移模型．随后是检索结合生成的方式．Lu等人[94]认为检索-重排筛选回复的方式存在着以牺牲响应质量来提升共情的缺陷，因此提出了一种检索-判别-改写的方法．对于满足情绪要求的回复直接返回给用户，否则就用模型进行改写；Sharma等人[95]发现心理健康平台上的共情数据较少，为进一步提升平台的共情能力，提出了一种对于低共情回复进行改写的方式．模型首先会检测需要修改句子的位置，随后重写句子以替换掉它，以此增强句子的共情性．相关文献整理结果如表4所示．

2.2.4 总结与分析

总体来说，基于情绪因素的模型和基于复合因素的模型主要是在序列到序列结构上进行改进的，

表4 基于结构因素的模型

结构名称	典型模型	优缺点
检索式	Qiu等人[93]	优点：质量较高 缺点：依赖完备知识库
检索结合生成式	Lu等人[94]；Sharma等人[95]	优点：结合了检索和生成的优势 缺点：流水线方式导致错误传导
模板填充式	Shi等人[51]	优点：一定的灵活性和较高的质量 缺点：较难迁移

而基于结构因素的模型则使用了其他的结构对模型进行了构造．基于情绪因素的模型在共情对话响应任务的早期被提出．这类模型考虑了情绪这一单一因素对于共情回复的影响．随后，研究人员发现情绪并不是共情回复表达的单一因素．因此，基于复合因素的模型被提出．这类模型考虑了除情绪以外影响共情表达的更多因素．该类模型对于共情的理解更加全面和深入，因此带来了更优质的结果，例如，考虑了意图的模型EmpHi[84]和个性化的模型EP-GAN[87]取得了该任务的最优结果．此外，另外一些研究者也探索了新的生成回复的方式．本文将这类模型称为基于结构因素的模型．这类模型探索了更多的生成方式，提供了探索该任务的一种新视角．

在未来，该任务趋向于从以下几个方向发展：(1)作为共情的重要因素，情绪将会被更加细致的探索，例如其混合特性、一致性等．(2)由于情绪并不是共情的单一因素，更多影响共情的复合因素会被考虑进来，以更有效的建模共情对话响应模型．(3)由于目前的模型大多为生成式模型，而检索式、检索结合生成式模型尚未被广泛探索，未来基于不同结构的模型将会被进一步探索．

2.3 情绪支持

情绪支持旨在调节说话者的情绪以慰藉或降低说话者负面情绪．本节将情绪支持的模型分为三类：激发积极情绪、激发指定情绪、降低负面情绪．

(1)激发积极情绪

在该任务的早期，Hu等人[96]则首先在情绪支持任务上利用语音标签促进了积极情绪的激发；同一时期，Lubis等人[97]使用采用了积极数据对模型进行训练，采用HRED[30]并利用对话历史句子的情感预测回复情感，以促进模型生成积极回复；随后Tulika Saha等人[98]提出了一个用于激发积极情绪

的对话数据集；最近，Wang等人[99]也提出了一个用于激发积极情绪的中文数据集，并在Bao等人[100]基础上修改了情绪引导和负面情绪规则化的损失，使其具备情绪调节能力；Tulika Saha等人[101]在心理健康支持领域首次考虑了多种意图和情感对于该任务的作用. 该模型首先通过意图生成器生成回复，随后使用共情重写框架将回复改写为具有共情性的回复.

(2)激发指定情绪

Lubis等人[102]对Lubis等人[97]的工作进行了改进，对日常中更常见的情绪进行了激发，该模型采用HRED构建了多任务模型. 分别预测了情感分类、对话动作类别以及进行了回复生成任务；随后Li等人[103]进一步明确了该任务的定义，它利用了预训练模型和反馈语句进一步优化了回复语句，同时也可以激发指定的情绪.

(3)降低负面情绪

Liu等人[104]提出了一个降低说话者负面情绪的任务，意在通过对话降低说话者的负面情绪强度，以达到情绪支持作用；Peng等人[105]认为更好的情绪支持需要全局地理解原因、局部地捕捉意图，因此提出了一种全局到局部层次图网络以捕捉全局原因、局部意图和对话历史；相比于前者，Tu等人[106]更注重细粒度情绪的理解，并利用常识性推理模型COMET[83]进行了细粒度的情感理解，同时提出了一种混合响应策略融入情感信息.

以上三类任务的目的都是使用情绪支持的方法调节用户情绪. 激发指定的情绪因其依赖于人工指定激发的情绪，较于其两类模型具有一定的局限性. 而激发积极情绪和降低负面情绪实际上都是将用户的情绪从负面调整到正面，因此具备一定的关联性. 效果方面，Tulika Saha等人[101]考虑了意图因素，取得了激发积极情绪模型的最优结果；Li等人[103]采用预训练模型获得了激发指定情绪模型的最优指标；Tu等人[106]则利用推理知识获得了降低负面情绪模型的最优结果. 该子任务的模型整理结果如表5所示.

目前该任务一个难点为如何评估模型有效地慰藉说话者. 因此，该任务面临的一个挑战为如何有效地评估慰藉效果. 此外，有效的慰藉涉及到多种因素，例如意图、个性化等因素. 而目前模型考虑的因素较为有限. 因此，如何结合多种因素有效地调节说话者情绪以起到心理健康支持是未来的一个趋势.

表5　情绪支持模型

类型	典型模型
激发积极情绪	Hu等人[96]；Lubis等人[97]；Tulika Saha等人[101]
激发指定情绪	Lubis等人[102]；Li等人[103]
降低负面情绪	Liu等人[104]；Peng等人[105]；Tu等人[106]

2.4　多模态情感对话生成

人类在交流时，会有多种形式的沟通，例如视觉、听觉等. 但以上的模型都是基于文本这一种模态进行构建的. 为了更好地模拟人类在现实中实际的对话场景，研究人员研究了基于多模态的情感对话生成模型. Huber等人[107]注重到了图像和文字两种模态对于对话的影响，并采用预训练的模型识别了图片中的情感、图片场景以及面部表情，随后与文本融合后生成了具有情感的回复；Liang等人[108]认为表情、情感和人物特质(某人所说的全部话语)与共情对话比较相关. 该研究首先构建了一个关于三者的异质图，在使用HGNN[109]编码了三种特征的表示后生成了回复；随后，Liang等人[110]又对Liang等人[108]的工作加入了音频及音频之间的关联，进一步丰富了模态信息. 相比于Huber等人将多模态信息进行了简单的融合，后面两种以图构建模态间关联的方式更为精细复杂.

多模态情感对话生成任务更加接近于现实场景. 但现有的基于多模态的工作较少，这主要是由于多模态数据的限制. 现有的三个多模态模型都考虑了图片和文本两个模态的模型，而近年来Liang等人[110]的工作更多的考虑了音频模态. 在现有的工作中，后者具有更加优质的结果.

目前模型结合的模态信息有限，而现实场景中的模态更加丰富. 因此，引入更加贴近现实场景的模态信息是未来的一个趋势. 此外，由于多种模态信息的引入，如何有效的融合多种模态信息为该任务未来的另外一个趋势.

2.5　新任务

随着情感对话的发展，近来研究者提出了多人共情对话、价值任务、减轻语言毒性三类新任务.

(1)多人共情对话

由于以往的共情对话仅有两人进行交谈，而诸如讨论、会议、群聊等多人聊天场景较为常见. 因此，Zhu等人[111]提出了一个多人共情对话的任务. 此外该研究认为序列到序列模型在多轮对话具有一定的限制，同时也认为不同参与者之间会互相影响. 因此该模型采用动态图网络构建了多方对话，

随后融合了说话者的信息来探索对话的情感.

(2)价值任务

价值指的是在人们生活中想要完成的目标,它影响着人们对于行动、策略、人物和事件的选择和评估[112]. Qiu等人[112]将价值分为十个维度,并构建了一个大规模的价值数据集. 为了验证价值模型的有效性,作者首先在价值数据集上训练了一个价值模型,随后在个性化和共情任务上验证了该模型的有效性.

(3)减轻语言毒性

有毒的语言通常被描述为骚扰或攻击性语言,它降低了对话者交流和合作的可能性[113]. 预训练模型易容产生有毒语言的特点限制了其使用范围. Lahnala等人[114]认为共情和毒性有着负相关的关系,并发现共情可以减缓毒性语言的发生,随后在Sharma等人[78]构造的数据集上验证了毒性与共情的负相关性. 为检验模型的有效性,该研究首先将微调后的语言模型和未微调的语言模型的预测概率按权相加以生成回复,随后调整权重以观察生成回复的毒性是否合理的改变.

2.6 总结与分析

在任务的早期,大多数研究者注重于可控情感对话生成模型. 这类模型将指定的情绪作为前提条件,并根据对话内容和情绪生成具有指定情绪的回复. 这些研究首先在对话中融入了情感,为今后任务的发展起到了奠基作用. 但是,在实际的对话场景中,回复的情绪类别并不是可以直接获得的,而是通过人为理解得到的. 而且根据某一指定的情绪生成的回复也并不一定是共情的回复. 因此,研究人员提出了共情对话响应任务. 该任务旨在理解说话者的感受,并做出合适的回复. 为构造具有共情能力的模型,这些模型分别从情绪单一因素、复合因素、结构因素三方面进行建模. 尽管以上模型都取得了瞩目的成果,但它们都没有明确的应用场景. 另外一些研究人员聚焦于情感支持场景,提出了情绪支持任务. 该任务旨在调节用户的情绪,为用户带来更多情绪上的激励和安慰. 为了调节用户情绪,该任务的模型分别从激发积极、激发指定情绪和降低负面情绪三个方面进行了研究.

以上所有模型都有一个缺陷:这些方法大多都是基于文本这一单一模态进行的研究,而现实场景却存在多种模态. 因此,多模态情感对话生成模型被提出. 尽管多模态对话生成模型更贴近实际场景,但目前仅有三个基于多模态的模型. 这主要是由于多模态数据的不足导致的.

为更多的将深度情感对话响应模型应用于现实场景,近年来一些研究人员也探索了新的任务,比如多人共情对话、价值任务、减轻语言毒性等. 这些任务贴近于现实的应用场景,为情感对话响应任务带来了更多的发展前景.

总体来说,目前该任务发展具有两条规律. 一是从不受场景限制的任务发展为特定场景下的任务. 从可控情感对话生成任务到共情对话响应和情绪支持任务说明了这点. 二是考虑情感这一单一因素到考虑多种因素以增强模型. 各个任务的模型从早期都考虑了情感这一单一的因素. 随着研究的深入,研究者们更多地从多方面因素增强模型. 各类任务整理结果如表6所示.

表6 各类任务异同点

类型	不同点	相同点
可控情感对话生成	回复的情感是指定的	(1)基本都采用序列到序列结构构件,分为理解对话和表达回复过程; (2)各个任务大部分模型都为生成式模型; (3)都探索了情绪及其他因素对于情感理解和表达的影响,如个性化
共情对话响应	要求情感是自动识别的且回复是共情的	
情绪支持	要求回复能够缓解负面情绪或提升积极情绪	
多模态情感对话生成	能够多种模态检测情绪	

此外,为了更加清晰的表示模型之间的性能差异,本文也收集了模型在主要数据集上的指标,具体如附表所示.

3 深度情感对话响应的模型结构

由于深度情感对话响应模型的复杂和多样性,本节首先从建模方式粗略地将模型分为检索、生成、检索融合生成三种方式,随后再根据结构将模型分为序列到序列、预训练模型、生成对抗网络、强化学习、预训练模型、对偶学习五个更为细致的方面.

3.1 建模方式

基于深度学习的情感对话响应任务是对话、情感分析、检索和生成的交叉学科. 它基于对话历史及情感类别,利用深度神经网络构建情感对话模型,用以产生具有合适情感倾向的回复. 图5展示了情

感对话响应任务与几个学科的关系. 从横向分析，深度情感对话响应位于情感分析与对话领域的交互处，是结合情感分析技术将情感融入对话任务的学科. 从纵向观察，深度情感对话响应可根据其构造技术的不同分为生成式、检索式、检索融合生成式三类.

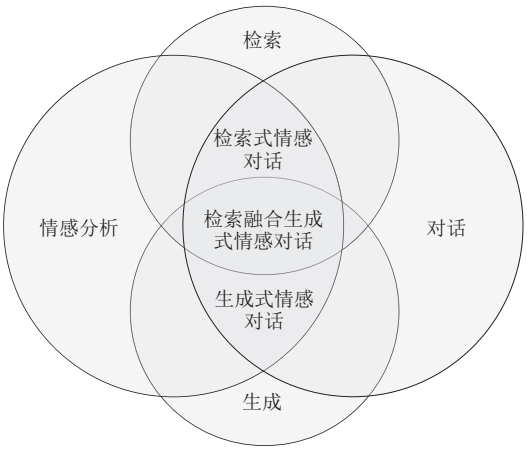

图5　情感对话响应模型交叉学科图

深度情感对话响应任务中，生成式模型主要以对话历史为前提，通过指定或感知情感后生成具有情感的回复. 该领域中主流的生成方式主要还是基于序列到序列架构的. 构建生成式模型的方式也有多种，例如改进编码器或解码器、微调预训练模型、在序列到序列架构上加入更多的模块以增强回复的情感. 由于数据的缺乏，该领域的检索式模型直至2020年才由Qiu等人[93]提出. 检索式模型首先需要构造候选的数据库，随后通过文本匹配的方式从数据集中搜索出情感和内容都适合的文本作为回复. 检索式模型产出的回复质量较依赖于候选集的质量，因此对于候选数据库有着较高的要求.

在数据质量和规模具有一定限制的条件下，一些研究者也选择检索融合生成的方式构建模型. 例如Dong等人[54]提出了一种检索-判别-改写的方式，仅仅将不符合情感要求的回复进行改写；而Sharma等人[95]则仅改写某些低共情性的子句，以达到提升共情的目的.

总体来说，生成式方法灵活性较高且结构相对简单，检索式模型在语料质量较高的情况下能产出高质量、更自然的回复. 为了兼具两种方式的优势，检索融合生成式的方式则成为一个不错的选择. 本节在介绍每种结构时，着重讲述如何改进或利用该结构，致使模型的情感能力得以提升. 随后本节将从序列到序列、预训练模型、生成对抗网络、强化学习、对偶学习、模板填充六个方面进行阐述，如图6所示.

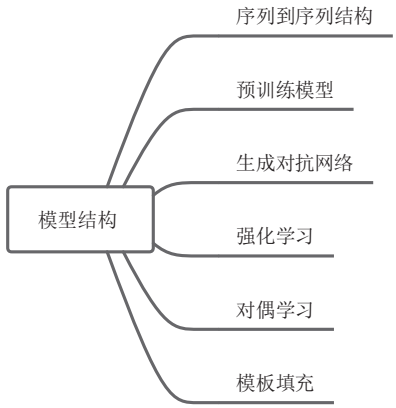

图6　情感对话响应任务常用的结构

3.1.1　序列到序列结构

序列到序列结构. 序列到序列结构(Sequence To Sequence，Seq2Seq)由Sutskever等人[27]于2014年提出，并在翻译任务中率先使用了该结构. 它首先将查询词映射为Word2Vec[71]以后，然后使用编码器学习出表示为低维语义向量的查询信息，最后通过解码器逐字解码生成回复语句. 其中编码器和解码器通常由循环神经网络(Recurrent neural network，RNN)[18]或者Transformer[115]构成. 如图7所示，序列到序列结构的编码器将内容为“我喜欢雨季”的查询语句压缩编码为语义向量C，随后解码器通过解码该语义向量逐字生成回复语句“我却爱天晴”. 由于该结构能够方便地捕捉对话信息并生成回复语句，大量的深度情感对话响应模型都是基于该结构改进得到.

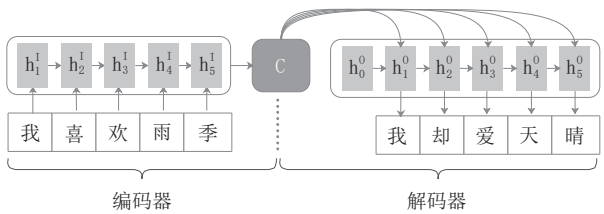

图7　序列到序列结构图

模型一般会从输入序列、编码器、解码器三个部分对模型进行改良. 本文也按照这三个方面总结各种方法，并罗列了一些典型的模型.

(1)基于输入序列的改进模型

输入的序列会直接影响编码器对于对话信息的编码. 因此，改进输入序列显得额外的重要. 一些研究者研究了将情绪或其他知识拼接到查询语句上. Huang等人[47]验证了将情绪分别拼接到查询的前、后给模型带来的效果，发现效果几乎一致；此外，Sun等人[28]也研究了将查询和回复的情感都拼接作

为输入，以促使模型获得情绪的感知能力；Li等人[68]则将情绪归因的信息拼接到查询后面送入预训练模型；Zandie等人[90]将情绪、对话动作、主题转为嵌入向量后加入了词嵌入，将具有多样信息的词嵌入作为预训练模型的输入. 总之，从内容上来看，研究者们分别通过查询情感、回复情感、VAD向量、情绪归因信息进行了试验；从位置来看，这些信息既可以拼接到查询的前后，也可以与词嵌入拼接，还可以直接加到词嵌入形成新的词嵌入. 由于各个模型的实验条件不同，并未能有工作验证这些做法的优劣.

(2)基于编码器的改进模型

Wei等人[69]利用两个编码器独立编码了情感和词向量，随后采用融合机制融合了语义与情感；Wei等人[72]对前者的工作用变分自编码器进行了增强，说明了使用变分自编码器融合语义和情感具有一定的优势；编码器能够直接编码用户的输入，对于情绪的感知有着关键作用；Zhang等人[32]也采用变分自编码器改进了编码器. 其中，EsCVAE-I的先验网络为问题的编码，后验网络为回复、查询情感、回复情感的编码，模型分别对先验和后验进行重参数化后缩小两者之间的差异，并协同查询编码进行解码并预测回复情感类别；不同于EsCVAE-I的是EsCVAE-II将查询情感加入先验网络，同时将查询加入后验网络一同编码. 实验发现后者的情绪表达能力优于前者；此后，Liang等人[108]使用图构建编码了多模态间的关系，以图结构改进了编码器；而Sabour等人[56]和Li等人[81]也通过构建图结构引入了额外的常识性知识，挖掘了对话中隐藏的情感；最近，Kim等人[76]则通过挖掘情绪和关键词进一步提升了模型对对话的感知能力.

(3)基于解码器的改进模型

解码器的作用是生成回复，因此它与情绪的表达有着直接的关联. 一部分模型通过改进解码器的内部结构使其获得更强的情感表达能力. Zhou等人[26]从三个方面改进了解码器. 该工作首先将情绪随机初始化后拼接到语义信息上输入解码器，这种方式被之后的多个可控情感对话生成模型所采用. 随后为了逐步释放情绪，研究者也改进解码器使其以情绪衰减的方式逐步解码. 最后也构建了外部情感词汇表(External Emotion Vocabulary)，在预测下一个词时预测该词是一般词汇或情绪词，这种方式也被后续的一些模型所借鉴；Song等人[41]借鉴了外部情感词汇表的做法后，也引入了外部词典，以表达更多的情感信息；Gao等人[65]利用情绪归因信息更改了解码器中交叉注意力权重机制，这种方式相对于前者更为细致；Li等人[81]也在解码器中增强了情绪词的注意力值，解码时也考虑了编码器感知到的情绪.

此外，一些研究也表明了解码的首个词较为重要，Yuan等人[36]使用了LST(Learning to start)[38]机制预测了首个词；Sun等人[48]将回复情绪标签的词嵌入作为了首个词；Li等人[64]使用了编码器感知到的情绪作为首个词. 这些研究表明了将具有情绪的向量作为首个词能够提升回复的情感.

最后，一些研究者也通过其他的方式改进了解码器. 例如Li等人[88]和孙等人[89]在解码时根据主题和关联词从两端生成回复，最后拼接起来. 再比如MoEL模型[61]为每种情感分别独立设置了解码器，最后结合这些解码器以生成回复.

总之，输入序列的改进方式有限. 而编码器负责对输入序列进行编码，直接影响着情绪和语义的感知. 由于解码器直接输出回复词，它则直接决定着回复的质量. 从现今的研究来看，解码器的修改大多集中在可控情感对话生成模型，追其原因在于此类模型着重研究指定情绪的表达. 而编码器在共情对话响应中则被重视起来，因为模型是否能正确感知情绪也间接决定着回复表达的情绪是否正确. 输入序列则在各个任务中都有所重视. 最后，按照改进模型的复杂程度来看，改进解码器的模型大于改进解码器和输入序列的模型.

3.1.2 预训练模型

自从2019年预训练模型BERT[86]被提出来后，一些工作也将预训练模型应用到了情绪对话响应任务. 研究者们多使用GPT[116]和GPT-2[92]两种预训练模型生成回复. 为了使模型具备情感，这些模型采取了三种方式：首先是微调方式. 研究者们将预训练的模型在具有情感的数据集上进行微调，以使模型具备情感. 这些模型除了进行了生成回复的微调任务，也同时进行了其他任务的微调. Zandie等人[90]进行了判断下个句子正确与否、预测情感任务上微调；Ide等人[117]进行了粗(2类)、中(6类)、细(12类)三个情绪分类任务微调；Xie等人[118]在判断下个句子正确与否、情感分类、评估句子的效阶(Valence，VAD向量[46]的V维度)和唤醒(Arousal，VAD向量的A维度). 其次是模型输入层面进行的改进. 例如Zandie等人[90]将情绪、对话动作、主题转为嵌入向量后加入了词嵌入. 随后，Muthukumar等

人[119]也从图像和视频领域引入了生成对抗网络的变体——生成对抗模仿学习(Generative Adversarial Imitation Learning，GAIL)[120]，它将GPT-2[92]作为生成器，判别器以历史对话、对话提示、标准回复多种信息作为依据判别回复的真伪；最近，Kim等人[76]利用预训练模型RoBERTa[77]检测情绪，并采用沟通机制[78]检测情绪反应以更准确感知对话中的情感信息.

直到2022年11月，OpenAI推出了ChatGPT(Chat Generative Pre-trained Transformer). 该模型是继GPT-3[121]后的一个大规模预训练模型. 基于大规模的训练数据及参数，该模型采用了人类反馈强化学习(Reinforcement Learning from Human Feedback，RLHF)[122]在评估式强化人工训练代理框架(Training an Agent Manually via Evaluative Reinforcement，TAMER)[123]上构建而成. 从功能方面，ChatGPT是一个以对话形式与人交互的模型，它能够完成流利的对话、撰写代码、翻译等工作，同时也被证明具有9岁孩子的同理心[124]. 不仅如此，ChatGPT还被运用于更专业的场景，表现出优异的结果，例如，立场检测[125]、论文写作[126]、医疗问答及评估[127]等. 虽然该模型具有强大的功能，但由于其代码不是开源的，且训练需要巨大的训练数据和计算资源，因此在研究或工业领域复现或使用该模型也具有一定的限制. 另一方面，而ChatGPT取得的优异结果也为共情对话响应任务带来了更多的挑战与机遇. 研究者可通过ChatGPT进一步提升对话质量，使机器更流畅且富有情感的与人自然交流.

总体来说，由于情感对话响应数据主要依托于人工标注的少量数据，引入大规模预训练模型能够较好的缓解可使用数据较少的问题. 现在主要的使用方式也限制于微调，且由于预训练模型参数量大导致了运行效率低下等问题，该方式也为研究人员来了一些不便. 而ChatGPT的出现带来了优异的对话效果，不仅能够流利的以自然语言方式与人进行交互，而且能够完成多种专业领域任务. 目前ChatGPT吸引了大量研究者，它为情感对话响应任务带来了新的挑战和机遇.

3.1.3　生成对抗网络

生成对抗网络(Generative Adversarial Networks，GAN)包括生成器(Generator)和判别器(Discriminator)两部分，生成器需要根据对话的内容来生成回复，判别器负责判别输入的真伪样本，并给出相应的反馈以优化判别器和生成器，该类模型的概括如图8所示. 构建基于生成对抗网络的模型需要从生成器和判别器两方面考虑.

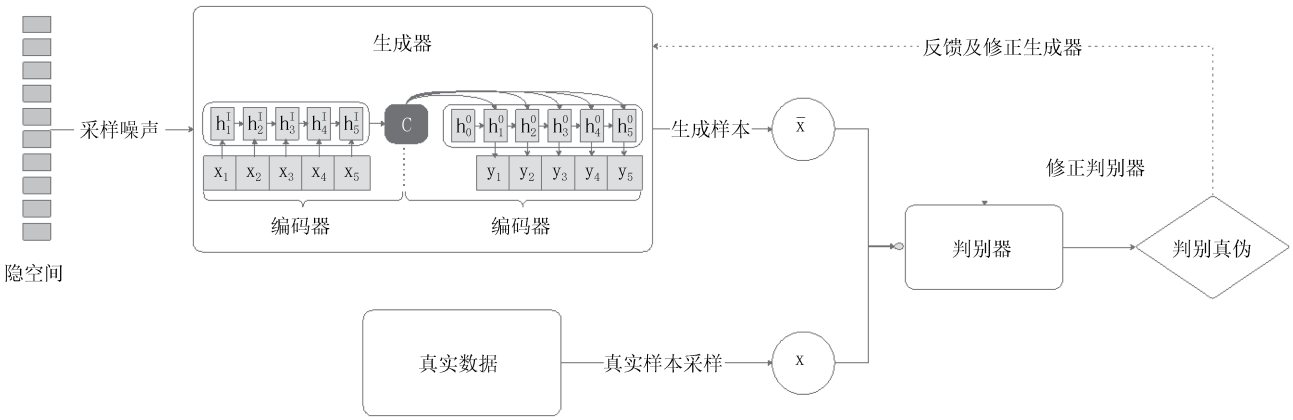

图8　Sun等人[48]提出的基于生成对抗网络模型概括图

生成器可以是一般的序列到序列结构模型，也可以是基于大规模数据训练的预训练模型，例如Muthukumar等人[119]在生成器中使用了GPT-2预训练模型[92].

不同于生成器的改进多源于对序列到序列结构的改进，判别器的不同多在于输入的差异. 典型的有，Kong等人[33]和Sun等人[48]将标准回复和生成回复提供给判别器鉴别样本真伪；Li等人[64]则增加了反馈语句，以进一步优化判别器；不同于Li等人[64]，Muthukumar等人[119]将历史对话和对话提示作为判别器的额外输入；而最近Firdaus等人[87]使用了共情、个性化、语义多种判别器对生成器进行了优化，该工作从多个角度考虑了判别的依据.

生成对抗网络的优势在于训练过程中，生成器和判别器能够不断的交互，在不断的对抗中学习. 从判别器的输入来看，多种信息的输入有利于判别

器更加准确地判断回复的真伪. 因此,全面利用多种因素构造判别器较为必要.

3.1.4 强化学习

强化学习(Reinforcement Learning, RL)模型分为智能体(Agent)和环境(Environment)两个部件,智能体负责生成回复,环境则负责调整回复的质量. 如图9所示的基于强化学习模型概括图[128],智能体通常由序列到序列结构组成,它通过动作与环境进行交互,随后环境根据智能的动作回馈相应的反馈(Reward)和下一阶段的状态,智能体通过与环境的交互得到反馈不断优化自身. 情感对话响应任务中,智能体通常采用序列到序列结构,以查询、对话历史和情绪为输入,经过不断学习后输出具有情绪的回复语句. 在该任务中,环境一般是为句子质量或者情绪的反馈函数,它能够为生成的句子或句子蕴含的情绪给定一个评判结果. 智能体根据这个评判结果迭代优化智能体.

智能体既能够局部地逐字生成回复,也能全局地替换低情感的子句生成回复. 由于该任务中,强化学习模型更大差异在于如何去有效构造反馈,因此以下着重讲解反馈的构造方式. Zhou等人[31]把正确情绪在所有预测情绪的排序位置作为反馈;Peng等人[50]在前者的基础上进行了改进,并把生成回复与标准回复的情绪差异作为反馈;Shin等人[129]和Shin等人[60]考虑了反馈语句作为反馈因素优化模型;孙等人[89]将更多因素考虑了进来,包括生成回复的流利度、主题关联度、情感关联度作为反馈;后来,Sharma等人[95]也考虑了生成与标准回复间情感差异性、相关性、非平凡性、与查询间相关性等更加复杂的因素.

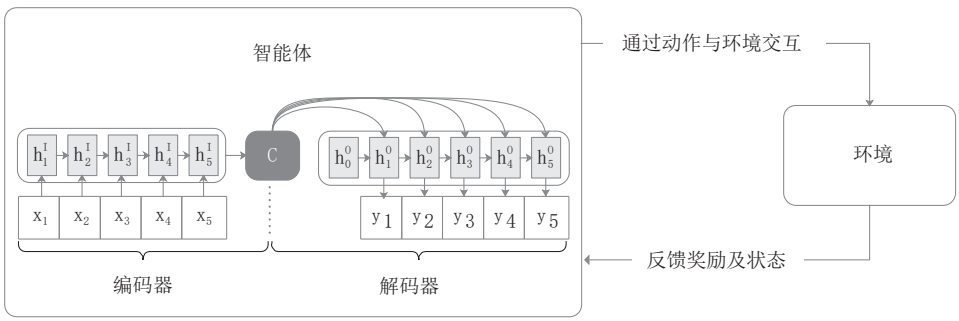

图9 Li等人[128]提出的基于强化学习模型概括图

总体来说,构造反馈的方式多种多样. 按照时间发展来看,构造反馈的方式越来越复杂,考虑的因素也更为全面,但如今未有一个模型将对话历史、标准回复、反馈等信息综合考虑进来.

3.1.5 对偶学习

对偶学习首先由He等人[53]在机器翻译任务中提出,之后Zhang等人[130]和Cui等人[131]在对话生成任务使用了对偶学习方法. 普通的模型仅考虑了查询到回复的语义和情感流动,但基于对偶学习的模型考虑了查询到回复、回复到查询双向的流动关系. 如图10所示,对偶模型通常具有两个结构:即从查询到回复的前向结构和从回复到查询的后向结构. 在情感对话响应任务中,研究人员在2020年和2021年提出了两个对偶模型. Shen等人[52]提出的CDL模型采用两个ECM模型[26]分别独立的预测双向关系,为提升模型情感,该模型把预训练情绪分类模型的概率输出也纳入考量中;Shen等人[73]提出的Dual-Emp使用了更为复杂的Transformer结构,除了把具有回复和查询的配对数据用于建模,同时也利用强化学习结合非配对数据(仅有查询或回复的数据)对模型进行优化.

从人类会反复思考的习惯来看,对偶学习的方法更加的贴近这种习惯. 将双向的关系考虑进来能够多方面的考量回复与对话历史的关系,从而生成更加优质的回复.

3.1.6 模板填充

模板填充模型常用于任务型对话系统. 在情感对话响应任务中,Shi等人[51]利用预训练的情绪分类器得到了对话词的情绪嵌入表示,随后使用记忆模块捕捉了这种细粒度的情绪,最后筛选出合适的词填入模板中以生成回复. 模板填充模型适用于任务型对话且效果较好,并且具有生成效率高的优势. 该类模型的缺点为构建优质模板较难且较难迁移场景. 总之,模板填充模型依赖于模板的质量,若能实现自动生成高质量模板,模板填充方式则会成为一种优秀的生成方法.

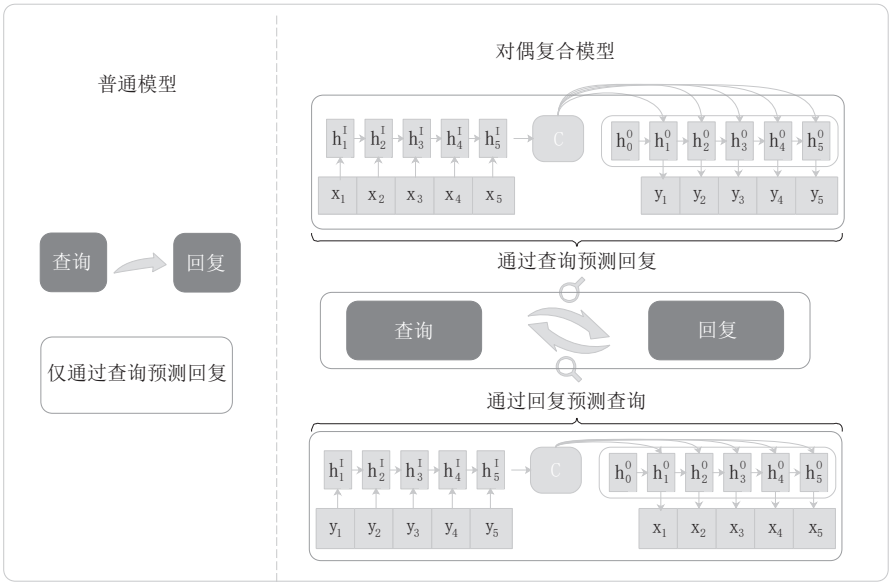

图10　Zhang等人[130]提出的基于对偶学习模型概括图

3.2　总结与分析

总体来说，由于序列到序列结构较为简单且易于改进，大多数模型都是基于序列到序列结构构建的．这类模型相比于其他模型仅能依靠损失函数优化模型参数．由于该类模型依赖于大量的训练数据，在训练数据缺失的任务中，该类模型并不具备优势．随着预训练模型的发展，基于预训练的模型也在深度情感对话响应中也被大量探索．这类模型在预训练阶段已经具备大量知识，在训练数据不足的场景中，微调预训练模型能够缓解训练数据不足的缺陷．但是这类模型参数量较大，训练模型需要较多的资源，同时训练和推理速度较慢．

基于对抗生成网络和强化学习的模型能够在序列到序列模型的基础上进一步优化模型，使模型的情绪表达能力进一步提升．但这些方法受制于判别器和环境的构造．相比于其他模型，这两类模型较难收敛．不同于普通模型仅考虑了查询到回复的单向关系，基于对偶学习的模型建模查询和回复之间双向的关联．这类模型能够以更全面的视角捕捉查询和回复之间的关系，从而生成更加可靠的回复．模板填充模型在任务型对话中被大量使用，且效果较好．但如何构造高质量的模板是一个难题．若能够以自动的方式构造优质模板，则会进一步缓解该类模型较难迁移场景的问题．各类代表模型及其优缺点如表7所示．

模型效果方面，结合了多种结构的模型更可能取得更优的效果．例如，结合了强化学习的模型普遍优于仅依靠序列到序列结构建模的模型．因此，结合多种结构发挥模型效果为未来的重要建模方式．同时，由于类似于ChatGPT等大模型的出现，结合大规模预训练模型为一个提升效果的重要方法．此外，如何利用这些模型结构结合具体的场景有效发挥实际用途也为未来的趋势．

4　数据集与评价指标

本节首先介绍了深度情感对话响应任务常用的数据集、新数据集和常用的外部知识，随后介绍了检索式和生成式模型的评价指标．

4.1　数据集

在深度情感对话响应任务的早期，面临的首要问题是数据集的缺乏．为解决该难题，研究者们首先训练一个情感分类模型，随后采用这种预训练的情感分类模型标注对话的数据，以构建情感对话响应数据集．常用的预训练情感分类模型有DeepMoji[132]、FastText[133]以及NLPCC2013①、NLPCC2014②比赛任务中的情感分类器．随着2017年NLPCC发布了情感对话响应任务，并公布了情感对话响应数据集，深度情感对话响应任务由此引起了更多的关注．此后，大家相继从电视连续剧《老友记》剧本、Twitter等渠道收集了多样化的情感对话响应数据，为该任务的发展奠定了坚实的数

① http://tcci.ccf.org.cn/conference/2013
② http://tcci.ccf.org.cn/conference/2014/

表7 不同结构的代表性模型

改进类型	典型模型	优点	缺点
输入序列	Sun等人[28];Huang等人[47];Li等人[68];Zandie等人[90]	简单有效;可控制情感	不利于情绪表达;易影响对话历史中情感的理解
编码器	Zhang等人[32];Sabour等人[56];Wei等人[72];Kim等人[76];Li等人[81];Liang等人[108]	可直接融合感知的情感和指定的情感	不能直接控制情绪表达
解码器	Zhou等人[26];Sun等人[48];Lin等人[61];Li等人[81];Li等人[88]	能直接控制情绪表达;情感表达能力较强	易忽略对话历史的情感;易表达不合理的指定情感
预训练模型	Kim等人[76];Zandie等人[90];Ide等人[117];Xie等人[118];Muthukumar等人[119]	蕴含大量预训练知识;能缓解训练数据不足的缺陷	参数量大;训练速度慢;较难充分利用预训练知识
生成对抗网络	Kong等人[33];Sun等人[48];Li等人[64];Firdaus等人[87];Muthukumar等人[119]	能结合判别器进一步优化模型;效果较好	模型不容易收敛;构造判别器较难
强化学习	Zhou等人[31];Peng等人[50];Shin等人[60];孙等人[89];Sharma等人[95];Shin等人[129]	能够结合反馈及损失共同优化模型;能考虑不可导的条件或因素;效果较好	模型不容易收敛;构造环境较难
对偶学习	Shen等人[52];Shen等人[73]	增强了对话历史与回复的一致性;建模方式更符合人类反思的习惯	易增加模型复杂度;较难掌握对话历史和回复的一致性程度
模板填充	Shi等人[51]	适用于任务型对话且效果较好;生成效率高	构建优质模板较难;较难迁移场景

据基础. 本文列举了22个考虑了多种因素的深度情感对话任务数据集,例如可控情感的数据集[134],考虑了意图的数据集[135-136]、多模态的数据集[137-138]等. 列举的数据集如表8所示. 随后本文详细介绍了该任务中最常用的4个数据集.

深度情感对话响应任务中,常用的数据集有4个,分别是DailyDialogue[134]、ED(Empathetic Dialogues)[57]、ESTC[26]、MojiTalk data[31].

DailyDialogue[134]:. 这个数据集是研究者从英语学习者的日常英语对话的网站上抓取原始数据得到的比较于其他数据,该数据是语法较为规范的多轮对话数据. 同时,该数据集的主题涉及范围较广,涵盖了从日常生活到金融主题的十个类别. 其次,它包含了丰富的情绪,并通过手动标记得到情绪标签. 其中,它包含了七个情绪类别,分别是:愤怒、厌恶、恐惧、幸福、悲伤、惊喜和其他.

ED(Empathetic Dialogues)[57]:该数据集是第一个基于众包的共情对话响应英文数据集,现在被多个共情对话响应模型所使用. 生成该数据集时,参与者根据给定的情景描述和情感标签生成对话. 相比于其他情感标签,该数据集的情感类型更为细致,具有32种常见的情感标签(详见附录),并且这些情绪标签涵盖了普遍的积极情绪和消极情绪. 此外,该数据集规模较大,共有24 850条对话,但该数据集的每条对话的对话轮次较短,最多不超过6轮.

ESTC[26]:该数据集是一个来源于微博的中文数据集. 该数据集包括了219 905条对话,常用于可控情感对话生成任务. 在早期,由于数据的缺乏,Zhou等人[26]在2018年使用NLPCC情感分类数据集训练了一个情感分类器,随后使用该分类器在STC对话数据集[139]标注形成ESTC数据集. 该研究者首先在NLPCC数据集上训练了多个分类器,在验证后选用了效果最优的基于双向的长短时记忆网络(Bi-LSTM)的分类器标注数据. 研究者为情绪标注了8种情绪类别,去掉两种不常见的类别后,剩下了愤怒、厌恶、快乐、喜欢、其他和悲伤六种类别.

MojiTalk data[31]:它是2018年Zhou等人[31]在提出MojiTalk所使用的数据集,是一个从Twitter上收集的单轮对话数据集,数据集有662 159条对话,它以64种表情符号为情绪标签. 标记数据集时,研究者获取了句子中具有至少一种表情符号的回复,如果有多种符号标签,则会选取数量最多的表情作为标签. 若出现多个数量相同的表情符号,研究者会选择出现频次最低的符号作为标签. 在数据处理的过程中,例如“!!!!”或“yessss”并不会缩短,因为该研究者认为这可能表明了情感的强度.

此外,本文也列举了一些常用的额外知识,例如,情感词典NRC VAD[140]、ConceptNet[141]等. 额外知识的数据如表9所示.

表8　深度情感对话响应任务数据集

作者	数据集	年份	语言	对话条数	适用范围	链接
Zhou等人[26]	ESTC	2021	中	219905	可控情感	https://github.com/tuxchow/ecm
Zhou等人[31]	MojiTalk data	2018	英	662159	可控情感	https://github.com/claude-zhou/MojiTalk
Rashkin等人[57]	ED	2018	英	24850	共情对话响应	https://github.com/facebookresearch/EmpatheticDialogues
Li等人[134]	DailyDialogue	2017	英	13118	可控情感	http://yanran.li/dailydialog
Welivita等人[135]	EDOS	2021	英	1 M	意图	https://github.com/anuradha1992/EDOS
Welivita等人[136]	EmpatheticIntents	2020	英	24850	意图	https://github.com/anuradha1992/EmpatheticIntents
Poria等人[137]	MELD	2019	英	1433	多模态	https://affective-meld.github.io/
Zhao等人[138]	M^3ED	2022	中	990	多模态	https://github.com/AIM3-RUC/RUCM3ED
Chen等人[74]	EmotionLines	2018	英	1000	情绪流动性	http://doraemon.iis.sinica.edu.tw/emotionlines/download.html
Firdaus等人[42]	MEIMD	2021	英	34 k	情绪混合；情绪强度	
Wen等人[75]	PELD	2021	英	6510	个性化	https://github.com/preke/PELD
Zhong等人[85]	PEC	2020	英	335 K	个性化	https://github.com/zhongpeixiang/PEC
Qiu等人[112]	PERSONA-CHAT	2022	英	21374	价值模型	https://liang-qiu.github.io/ValueNet/
Gao等人[65]	RECCON	2021	英	1126	情绪归因	https://github.com/declare-lab/RECCON
Sharma等人[78]	EPITOME	2020	英	10 k	沟通机制；减轻毒性；共情	https://github.com/behavioral-data/Empathy-Mental-Health
Zheng等人[91]	CoMAE data	2021	英	195105	三种对话因素	https://github.com/chujiezheng/CoMAE
Sharma等人[95]	PARTNER data	2021	英	-	沟通机制	https://github.com/behavioral-data/PARTNER
Saha等人[98]	MotiVAte	2021	英	4 k	激发积极情绪	
Wang等人[99]	PosEmoDial	2022	中	819391	激发积极情绪	
Li等人[103]	EmoElicitor data	2020	英	1559230	激发指定情绪	https://github.com/neuChatbotDS/emoElicitorData
Liu等人[104]	ESConv	2021	英	1300	降低负面情绪	https://github.com/thu-coai/Emotional-Support-Conversation
Lu等人[94]	SDCC	2021	中	1400	检索-判别-重写	https://github.com/luxinxyz/RDR/

表9　深度情感对话响应常用外部知识

作者	数据名称	年份	语言	数据集大小	链接
Dudău等人[59]	LIWC	2001、2007、2015	英	5000词以上	http://liwc.wpengine.com/
Warriner等人[46]	VAD	2013	英	13915个词	http://crr.ugent.be/archives/1003
Mohammad等人[140]	NRC VAD	2018	英	20000词以上	http://saifmohammad.com/WebPages/nrc-vad.html
Tang等人[70]	SSWE	2014	英	/	
Speer等人[141]	ConceptNet	2017	英	/	

4.2　评价指标

本文介绍了情感对话响应模型的常用评价指标，包括检索式和生成式两类.

4.2.1　检索式模型指标

检索式模型的目标是为多个候选回复打分，并按照排序的方式将优质的回复排在最前. 该类模型的评测方式常用的评估方式有*Precision*、*Recall*、*F*1、*MRR*、*MAP*等指标.

$$Precision=\frac{\text{模型预测相关的回复数}}{\text{模型预测回复的总数}} \tag{1}$$

$$Recall=\frac{\text{模型预测相关的回复数}}{\text{真实相关的回复总数}} \tag{2}$$

*Precision*指的是模型预测的相关正样本占正样本的比例，见公式(1). *Recall*表示预测正确的样本数占总样本数的比例，见公式(2). *F*1则是考虑了*Precision*和*Recall*两个指标的调和平均，见公式(3).

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{3}$$

随后本文介绍了两个考虑排序权重的指标，*MRR*为所有正确答案所在位置倒数的平均值，其中$position_n$为第n个预测正确文件在排序中的位置，N为测试样本的总数，见公式(4).

$$MRR=\frac{1}{n}\times\sum_{n=1}^{N}\frac{1}{position_n} \tag{4}$$

*MAP*在每条检索中将预测正确样本的得分取平均，然后再将每条检索计算的值再取平均，其中

$position_i$为第d_j条检索中第i个预测正确文件的排序位置. 其中待检索的样例中总共有D条检索,每条检索的正确样本数为d_i. AP为每条检索的平均准确数量. 详细的公式见公式(5)和(6).

$$MAP=\frac{1}{D}\times\sum_{d_j=1}^{D}AP \tag{5}$$

$$AP=\frac{1}{d_i}\times\sum_{i=1}^{d_i}\frac{1}{position_i} \tag{6}$$

4.2.2 生成式评价指标

生成式模型的评价指标包括人工评价和自动评价两方面,指标如表10所示. 总体来说,由于生成式情感对话模型较为灵活,一般采用人工评价结合自动评价的方式对模型的优劣进行评估. 其中人工综合评测一般根据其语言、情感、关联三方面综合考量,并采用二值(例如,好的回复、差的回复)或多值的方式对单个模型的结果评分,同时也可以通过Win-Tie-Loss的方式对比两个模型的优劣,即一个模型比另一个模型得分好、持平、差的数量的百分比. 对于人工评测,研究者一般会从测试集中筛选上百个回复,聘请三个或以上专业评测人员进行多值打分,并用Fleiss Kappa[142]系数比较多份数据的一致性,从而确认数据的有效性.

表10 深度情感对话生成式模型常用评价指标

	语言					情感			关联性	
	语法	流利度	多样性	平凡性	文本长度	准确性	有无情感	情感强度	问题关联	回复关联
人工评测	√	√	√	√	×	√	√	√	√	√
自动评测	×	PPL	Dist-1;Dist-2	×	Avg Length	Accuary;混淆矩阵	×	×	×	BLEU;Embedding;

自动评测方面,研究人员常用以下三方面衡量模型的语言质量,流利度方面使用困惑度(Perplexity, PPL)[143],即预测句子出现的概率,其中M为生成回复的词个数,P代表概率,由此可见句子出现的概率越大,困惑度越低,回复更优质. 困惑度详见公式(7).

$$Perplexity=\sqrt[M]{\frac{1}{P(y_1,y_2,y_3,\cdots,y_M)}} \tag{7}$$

多样性方面则主要采用*Dist*-1,*Dist*-2:该指标主要计算模型生成的回复在单粒度、双粒度方面词的不重复率,得分越高则代表多样性越好. 其中n代表的是粒度,分子代表了生成回复中该粒度的词组不重复的个数,分母代表了生成回复中所有该粒度词的个数,其中N为词组的个数. 详见公式(8).

$$Dist\text{-}n=\frac{Distinct(\sum_{i=1}^{N}n-gram)}{\sum_{i=1}^{N}n-gram} \tag{8}$$

文本平均长度(Average Length)[30, 144]:若模型生成的回复更长,文本的质量较高,故模型常用生成回复的平均长度作为生成质量衡量的标准之一.

对于情感方面,模型常使用准确度(Accuracy, acc)判断回复情感的情感准确性,即预测分类正确的样本数占总样本数的比例,详见公式(9).

$$Accuracy=\frac{\text{模型预测正确的回复数}}{\text{模型预测回复的总数}} \tag{9}$$

同样地,为判断指定情感与生成回复的情感之间的差异性,可以构建混淆矩阵计算指定情感下各种生成情感的概率,以此可更加详细地观察指定情感与生成情感之间的关系.

衡量生成回复与标准回复的关联性有多种方法,常用的有用词衡量其相似度的*BLEU*指标[145],分子是同时在标准和生成回复中的n-gram数量,分母是生成回复的n-gram数量,旨在衡量生成回复与标准回复的相似程度. 详见公式(10).

$$BLEU=\frac{\sum_{n\text{-}gram\in R,R_{true}}n\text{-}gram}{\sum_{n'\text{-}gram\in R}n'\text{-}gram} \tag{10}$$

*BLEU*指标明显的缺点是不能考虑近义词的影响,因此,Liu等人[146]提出了Embedding的三种方式衡量生成与标准回复之间的语义相关性. 如公式(11)-(13)所示为表示贪婪匹配法(Greedy Matching, GM),它首先将词用词嵌入表示,随后计算生成回复的每个词与标准回复中相似度最大词之间的cosine值并累加,然后除以标准回复词的长度,之后按同样的方式计算标准回复与生成回复之间的值并取两者平均值,其中$Emb(w_i)$代表将词w_i用词嵌入表示,R,R_{true}分别代表生成与标准回复. $|R|$为待计算文本的词的个数.

$$sim(w_i,w_j)=cosine(Emb(w_i),Emb(w_j)) \tag{11}$$

$$G(R,R_{true})=\frac{\sum_{w_i\in R,w_j\in R_{true}}max(sim(w_i,w_j))}{|R|} \tag{12}$$

$$GM=\frac{G(R,R_{true})+G(R_{true},R)}{2} \tag{13}$$

平均匹配法(Embedding Average，EA)将生成回复与标准回复的所有词嵌入取平均后再计算cosine值．详见公式(14)-(15)．

$$\bar{e}(R)=\frac{1}{|R|}\times\sum_{w_i\in R}Emb(w_i) \tag{14}$$

$$EA=cosine(\bar{e}(R),\bar{e}(R_{true})) \tag{15}$$

向量极值法(Vector Extrema，VE)是在词嵌入的每个维度筛选距离零最远距离的值作为该维度的值，构成句子向量后计算其cosine相似度．其中e^d代表了句子向量在d维度上的取值，$e^d_{w_i}$代表词w_i在d维度上的取值，$e(R)$则代表了句子的极值向量．M为计算极值向量句子中词的个数．详见公式(16)-(18)．

$$e^d(R)=e^d_{w_i} \tag{16}$$

$$|e^d_{w_i}|>|e^d_{w_j}|;\ i\neq j;\ 1<i,j<M \tag{17}$$

$$VE=cosine(e(R),e(R_{true})) \tag{18}$$

5　总结与展望

本节首先从任务发展、模型结构和情感特性三方面总结了深度情感对话响应模型，随后在此基础上阐述了深度情感对话响应任务所面临的挑战，最后对该任务的未来进行了展望．

5.1　总结

本节首先从时间发展的角度介绍了任务的发展过程，随后根据常见的结构对模型进行了总结，最后总结了模型所关注的情感特性．

5.1.1　任务发展

从时间发展来看，深度情感对话响应任务从2017年NLPCC发布情感对话任务开始，逐步得到研究者的重视．随后情感对话响应模型经历了一个由简单到复杂的过程．在任务的早期，可控情感对话生成模型仅关注于生成指定情绪的回复；后来，更多的研究者关注于使模型具备共情性，即让模型自身感知说话者的感受并生成具有共情性的回复；之后一些研究者倾向于将共情的模型应用于具体的场景，一个重要的场景就是情绪支持．情绪支持旨在为寻求帮助的人员降低负面情绪或提升情绪的积极度；以上的模型都是基于文本单模态的，为了增强模型的情绪识别能力和场景的适用性，一些研究者提出了基于多模态的方法．这些方法从图片、文本等模态多角度感知了情感，以提升模型的感知力；近年来，一些新的任务针对于具体的场景被提出，例如多人共情对话、价值任务、减轻语言毒性等任务．总之，从时间发展的角度看，情感对话响应任务从一个有一定限制和约束的控制型生成任务发展为了一个能够有实际落地场景、更具应用前景的任务．

5.1.2　模型结构

一方面，深度情感对话响应模型可分为检索式、生成式、检索融合生成式三类．检索式模型在深度情感对话响应方面还未被广泛探索，本文收集到的典型模型仅有Qiu等人[93]在2020年提出的一个模型．生成式模型多采用序列到序列结构，前期的序列到序列结构多采用长短记忆网络、门控循环单元结构等，对于长句子常采用HRED[30]编码．随着Transformer[115]的广泛使用，以往的结构也逐渐被Transformer结构所取代．由于检索和生成式模型都具有各自的优势，部分研究者也提出了一些新的方式促使模型更具备情感，例如通过检索-判别-改写的方式、或通过改写低共情子句的方式．总体来说，检索和生成式方法并不孤立，两种方式的结合方法有待于进一步探究．

另一方面，常用于情感对话响应模型的结构有强化学习、生成对抗网络、预训练模型、对偶方法、变分自编码器．本节也从另一个角度对模型进行了总结．

(1)强化学习．基于强化学习的方法与其他方法最大的不同是它具有环境，可通过环境进一步调整生成模型．早期的方法仅将情感因素考虑到环境中，而随后模型考虑了例如流利度、主题关联度等更复杂的因素．

(2)生成对抗网络．基于生成对抗网络的方法同样也因其有判别器而区别于其他的模型，它的判别器可随同生成器一同学习，生成器和判别器能够在相互博弈中找出更优解．早期模型的判别器仅依靠生成与标准回复判别真伪，后来也结合对话历史、回复的反馈语句等信息使判别器有更多的判别依据．

(3)预训练模型．预训练模型一般在大数据上训练至收敛后，再将其运用在对话数据上微调．早期时，模型在微调时除了预测回复任务以外，仅将情感分类作为其他任务协同训练．随着时间的发展，研究者们加入了更多的任务以微调模型，例如加入粗、细粒度的情感分类，下一个句子是否正确的分类任务．

(4)对偶学习. 对偶式方法不仅考虑了查询到回复的关系,也考虑了回复到查询的关联,以强化查询和回复之间的情感和内容的一致性. 从多方面因素来看,该方式从更全面的角度探寻了回复的生成,因此具备一定的研究价值和前景.

(5)变分自编码器. 情感对话响应任务中,变分自编码器一般分为先验网络和后验网络,先验网络编码查询,后验网络编码回复信息,并在训练过程中不断缩小两者的差异. 训练过程中,重参数化技巧能够增强模型的扰动. 由于其中的随机因子,该方式能够提升回复的多样性和减少通用回复的生成.

5.1.3 情感特性

概览深度情感对话响应模型可以发现,情感自身具有粒度性、混合性、关联性、多模态、等特质. 从更加全面的方面,情感与其他特质具有复杂的关联性,具体表现为复合性. 由于情感具有自身和外部的复杂性,探究情感的特质对于情感对话响应的发展具有关键作用. 本节总结了常见的用于建模的情绪特性.

(1)粒度性. 情感在模型中可以通过词或话语的粒度体现出来. 深度情感对话响应中,情绪表现在词层面往往是显性且局部的,例如“高兴”、“悲伤”、“激动”等词,通过该类词能够较为明显的发现情绪,而现有的研究也通过类似引入额外词典、额外知识、情绪归因等方法检测这种情绪. 相对于词粒度的情绪,话语级的情感更具有隐性和全局性的特点. 一句话的情感可以通过特定的结构表现出来,这体现了情绪的隐性. 当语句中出现多种情绪词、否定词时,通常需要理解整句话的含义才能把握说话者的情感,这体现了话语级情感的全局性.

(2)混合性. 多个研究者在标注数据或研究模型时发现,对话语句中往往存在不止一种情绪. 一些研究也表明这些情绪通过不同的强度蕴含在对话语句中,如何挖掘与表达这种混合情绪仍然是一个难题.

(3)关联性. 说话者和回复者语言之间具有情绪和内容的关联. 一些研究人员倾向于认为这种关联表现为内容和情绪的一致性,而另外一些研究人员则认为情绪在不断地转变,这种转变受到多方面因素影响,其中之一就是个人的特质. 而今,更多的研究显示,研究者倾向于认为情绪在对话中是不断地转变的.

(4)多模态. 情感作为人类表达信息的重要方式,存在着多种传递途径. 现有的工作大多是基于文本这一单一模态进行研究的,但也有少量模型基于多模态内容进行了研究. 关于多模态的研究方面,一些研究人员从图像和文本两方面对情感对话响应任务进行了研究,而其他一些人员则通过语音和文字两种模型对该任务进行了探索. 但总体来说,感知及表达多模态的情绪对于深入探究情绪具有重要意义.

(5)复合性. 情绪并不是影响对话的单一因素. 为了更全面地探究对话,一些研究人员探究了情绪与多方面因素. 这些典型的因素有常识、意图、个人特质、对话主题、价值观等. 在加入了这些因素后,模型能够更好地感知和表达情绪. 探索这种复合性,以及在对话中融入这种复合性已证明为一种有效的方式.

5.2 面临的挑战

随着深度情感对话响应任务的发展,越来越多的优秀模型被提出,但该任务仍然面临许多挑战. 本节将从多方面概述该任务面临的挑战.

(1)情绪方面. 单从情绪本身来说,情绪未有一个准确、完整、细致的定义. 研究者通过观察数据,构建模型去猜想验证情绪在对话中的性质. 但对于情绪更深入的特质,在该任务中并未被清晰明确的探索. 而由于其复杂性,大量的模型仅能建模情绪的某一个较为细小的特性,而不是全部的特质. 由于情绪在对话中的重要性、情绪本身的复杂性,较为准确的研究及建模情绪则是一个十分艰巨的挑战;从另一方面来说,情感和多种特质一起影响着对话的质量. 例如个性、意图等. 由于情感与这些特质的关联性较为复杂,如何发现、引入、融合这些特质则是另一个挑战.

(2)数据层面. 研究一个方向首要的条件是具备可用的研究数据,数据的质量也会严重影响模型的质量. 现今大部分情感对话响应的模型都是生成式的,一部分原因是因为生成式模型的数据集较为充足. 而同样可以发现,检索式、检索融合生成式模型未被大量研究的一个限制性条件就是数据的缺乏. 因此,构造高质量的大规模的数据是情感对话响应任务发展的一个挑战.

(3)模型层面. 从最基本的构建方式来看,模型分为检索式、生成式、检索结合生成式三种类型. 目前的研究大多集中于生成式模型,基于检索式、检索结合生成式两类模型较少. 而检索式和生成式具有各自的优势. 而部分研究工作也表明结合两者能够

构建出更具有优势的模型．但如何结合两者则是一个具有挑战的问题．

5.3　未来展望

(1)从心理学或人类行为中模拟情感的感知和表达特点并将其用于情感对话响应任务．目前深度情感对话响应任务中，感知和表达情感仅在于较为粗浅的阶段．相对于该任务，心理学中关于情感的研究则更为成熟．同时，人类在感知和表达情绪的时候也会有自身内在的规律和模式．因此，借鉴心理学研究和人类日常行为习惯以模拟情绪的特性能够进一步增强对话系统对于情绪的把控能力．

(2)结合情绪和个性化特质构造更符合人类习惯的对话模型．情感虽然是人机对话的一个重要因素，但并不是唯一的因素，即人类在表达语言时会受多种特质的影响．其中一个重要的特质就是个性化．由于个体的差异，不同个体针对于同一环境会得到不一样的情绪感受并做出不同的回复．而机器给出千篇一律的回复也会使人乏味．因此，模拟人类这一特性融合个性化和情绪是未来的一个重要研究方向．

(3)融合生成与检索式方法构造深度情感对话响应模型．在深度情感对话响应任务中，仅有少量的工作采用检索式方法构造模型．相比于生成式方法，检索式方法更容易产生流利自然的回复．而相对于检索式方法，生成式方法更容易产生个性化的、更具情感的回复．因此，结合检索和生成方法的优势能够进一步提升模型的质量．由于该方式具有以上优越性，因此具有一定的发展前景．

(4)探究任务型情感对话任务．在以上调研中，情感对话任务大多都是基于闲聊场景的．但在实际的场景中，基于任务的一些场景融入情感后也能够带来对话质量的提升，进而进一步提升机器的服务质量．例如在机票预定的场景中，客户表达自己很孤独，想去一个缓解孤独的地方．具有共情的机器人能够理解这种情绪，并根据这种情绪找到适合的旅游景点并推荐给用户．如果机器人缺乏这种情绪理解能力，则不会进行顺畅的交流．因此，在一些特定的任务型场景中，融入情绪十分必要．

(5)结合大规模预训练模型提升情感对话响应任务．最近OpenAI公司提出了大规模预训练模型ChatGPT，该模型被应用于对话、撰写论文、书写代码等多方面任务，且表现出优异的性能，甚至相关研究者认为其已经具有9岁孩子的共情能力．为了进一步提升深度情感对话响应任务，在深度情感对话响应任务中研发和结合类似于ChatGPT的大规模预训练模型变得十分必要．

(6)利用大规模预训练模型探索可靠的自动评价指标模型．目前，现有的模型通常采用人工评价指标和自动评价指标．自动评价的可靠性低于人工评价，其优势是成本低且不要专业的评估人员．人工评价准确度较高但评测数量少．例如，目前大多数工作仅能筛选上千条样例中100条左右数据进行评估．当回复偏差较大时会出现评测失真问题．因此，若能够通过自动评价指标准确评测回复质量，则会节省研究的成本和提升模型评估的效率．目前大规模预训练模型取得了令人瞩目的结果．但该任务中大多数研究仅将其运用于生成优质回复，而非训练优质的评价模型．因此，利用大规模预训练模型训练优质的评价模型是一个值得期待的发展方向．

(7)探究多模态情感对话响应任务．尽管目前基于多模态情感对话的模型较少，但由于多模态更符合人类的表达习惯，且结合多种模态更易于在现实场景中准确识别和表达情绪．因此，基于多模态的情感对话响应任务是一个具有前景的方向．但由于多模态情感对话响应数据的缺失，该方向发展仍然较为缓慢．因此，构造更多高质量的多模态数据和发展多模态情感对话响应任务是一个具有前景的方向．

(8)利用数据增强技术构造高质量数据集．当前深度情感对话响应任务面临一个重要的挑战是数据的缺乏．当前检索式、检索结合生成式模型以及基于多模态的模型较少，其中一个重要原因就是数据的缺乏．目前数据增强技术已经在图像领域取得了不错的成果．而在深度情感对话响应领域却较少被探索．因此，利用数据增强技术构建大规模高质量数据集是一个值得努力的方向．

6　结束语

人类在对话交流中不仅伴随着信息的传递，也会伴随着情感的交流．研究表明在对话系统中引入情感能够有效提升用户满意度，减少系统故障机率．因此，加快情感对话响应任务的研究进度具有重要意义．本文专注于情感对话响应任务，首先从历史发展、社会影响及研究意义的角度阐述了任务的发展历程，其次按照子任务对情感对话响应模型

进行了分类梳理,随后根据模型结构进一步总结了该任务的建模特点,之后归纳了常用的数据集与评价指标.本文最后总结了深度情感对话响应领域的模型,并以此为基础对该领域进行了展望.总体来说,本文为深度情感对话响应任务构建了一幅清晰的图纸,希望能给相关的研究人员提供一定的帮助.

致 谢 我们在此对给出宝贵建议和支持的各位同行表示感谢!

参 考 文 献

[1] Bui T H. Multimodal dialogue management-state of the art. Enschede, Netherlands: Centre for Telematics and Information Technology, University of Twente, CTIT Technical Report Series: 06-01, 2006

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

[3] Che Wan-Xiang, Zhang Wei-nan. A Survey of Human-Computer Dialogue Systems. Artificial Intelligence VIEW, 2018, 000(001): 76-82 (in Chinese)
(车万翔, 张伟男. 人机对话系统综述. 人工智能, 2018, 000(001): 76-82)

[4] Chen Chen, Zhu Qing-Qing, Yan Ru, et al. Survey on deep learning based open domain dialogue system. Chinese Journal of Computers, 2019, 42(7): 1439-1466 (in Chinese)
(陈晨, 朱晴晴, 严睿等. 基于深度学习的开放领域对话系统研究综述. 计算机学报, 2019, 42(7): 1439-1466)

[5] Kim E, Klinger R. A survey on sentiment and emotion analysis for computational literary studies. arXiv preprint arXiv: 1808.03137, 2018

[6] Damasio A R. Emotion in the perspective of an integrated nervous system. Brain Research Reviews, 1998, 26(2-3): 83-86

[7] Panksepp J. Affective neuroscience: The foundations of human and animal emotions. Oxford, England: Oxford university press, 2004

[8] Ekman P E, Davidson R J. The nature of emotion: Fundamental questions. Oxford, UK: Oxford University Press, 1994

[9] Cabanac M. What is emotion? Behavioural processes, 2002, 60(2): 69-83

[10] Darwin C. The expression of the emotions in man and animals. Chicago, USA: University of Chicago press, 1872

[11] Mayer J, Salovey P. Emotional development and emotional intelligence: Implications for educators. New York, USA: Basic Books, 1997

[12] Kort B, Reilly R, Picard R W. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion//Proceedings of the IEEE International Conference on Advanced Learning Technologies. Wisconsin, USA, 2001: 43-46

[13] Taylor S, Jaques N, Nosakhare E, et al. Personalized multitask learning for predicting tomorrow's mood, stress, and health. IEEE Transactions on Affective Computing, 2017, 11(2): 200-213

[14] Prendinger H, Mori J, Ishizuka M. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. International Journal of Human-Computer Studies, 2005, 62(2): 231-245

[15] Martinovsky B, Traum D. The error is the clue: Breakdown in human-machine interaction. Los Angeles, USA: University Of Southern California Marina Del Rey Ca Inst For Creative Technologies, ISCA Tutorial and Research Workshop on Error Handling in Dialogue Systems: ADA459168, 2006

[16] McNicoll G. World population ageing 1950-2050. Population and development Review, 2002, 28(4): 814-816

[17] Zhuang Yin, Liu Zhen, Liu Ting-Ting, et al. Survey of affective-based dialogue system. Journal of Frontiers of Computer Science and Technology, 2021, 15(5): 825-837 (in Chinese)
(庄寅, 刘箴, 刘婷婷等. 文本情感对话系统研究综述. 计算机科学与探索, 2021, 15(5): 825-837)

[18] LeCun Y, Bengio Y, Hinton G. Deep learning. nature, 2015, 521(7553): 436-444

[19] Turing A M. Computing machinery and intelligence. Creative Computing, 1980, 6(1): 44-53

[20] Weizenbaum J. Eliza—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 1966, 9(1): 36-45

[21] Colby K M, Le Magnen J. Parrying. The Behavioral and Brain Sciences, 1981, 4(4): 550-560

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

[23] Yu L, Hermann K M, Blunsom P, et al. Deep learning for answer sentence selection. arXiv preprint arXiv: 1412.1632, 2014

[24] Wang D, Nyberg E. A long short-term memory model for answer sentence selection in question answering//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015: 707-712

[25] Yin W, Schütze H, Xiang B, et al. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. Transactions of the Association for Computational Linguistics, 2016, 4: 259-272

[26] Zhou H, Huang M, Zhang T, et al. Emotional chatting machine: Emotional conversation generation with internal and external memory//Proceedings of the 32nd AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. Louisiana, USA, 2018: 730-738

[27] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks//Proceedings of the Advances in Neural Information Processing Systems. Quebec, Canada, 2014: 3104-3112

[28] Sun X, Peng X, Ding S. Emotional human-machine conversation generation based on long short-term memory. Cognitive Computation, 2018, 10(3): 389-397

[29] Shen X, Su H, Li Y, et al. A conditional variational framework for dialog generation. arXiv preprint arXiv:1705.00316, 2017

[30] Serban I, Sordoni A, Bengio Y, et al. Building end-to-end dialogue systems using generative hierarchical neural network models//Proceedings of the AAAI Conference on Artificial Intelligence. Arizona, USA, 2016: 3776-3783

[31] Zhou X, Wang W Y. Mojitalk: Generating emotional responses at scale//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 1128-1137

[32] Zhang R, Wang Z. Learning to converse emotionally like humans: A conditional variational approach//Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing. Hohhot, China, 2018: 98-109

[33] Kong X, Li B, Neubig G, et al. An adversarial approach to high-quality, sentiment-controlled neural dialogue generation. arXiv preprint arXiv:1901.07129, 2019

[34] Peng Y, Fang Y, Xie Z, et al. Topic-enhanced emotional conversation generation with attention mechanism. Knowledge-Based Systems, 2019, 163: 429-437

[35] Yang Feng-Rui, Huo Na, Zhang Xu-Hong, et al. Topic-expanded emotional conversation generation based on attention mechanism. Journal of Computer Applications, 2021, 41(4): 1078-1083 (in Chinese)
(杨丰瑞，霍娜，张许红等. 计算机应用，2021，41(4)：1078-1083)

[36] Yuan J, Zhao H, Zhao Y, et al. Babbling-the hit-scir system for emotional conversation generation//Proceedings of the National CCF Conference on Natural Language Processing and Chinese Computing. Dalian, China, 2017: 632-641

[37] Tang D, Qin B, Liu T. Aspect level sentiment classification with deep memory network. arXiv preprint arXiv: 1605.08900, 2016

[38] Zhu Q, Zhang W, Zhou L, et al. Learning to start for sequence to sequence architecture. arXiv preprint arXiv: 1608.05554, 2016

[39] Zhang R, Wang Z, Mai D. Building emotional conversation systems using multi-task seq2seq learning//Proceedings of the Natural Language Processing and Chinese Computing. Hohhot, China, 2018: 612-621

[40] Shantala R, Kyselov G, Kyselova A. Neural dialogue system with emotion embeddings//Proceedings of the IEEE First International Conference on System Analysis Intelligent Computing (SAIC). Kyiv, Ukraine, 2018: 1-4

[41] Song Z, Zheng X, Liu L, et al. Generating responses with a specific emotion in dialog//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 3685-3695

[42] Firdaus M, Chauhan H, Ekbal A, et al. More the merrier: Towards multi-emotion and intensity controllable response generation//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2021: 12821-12829

[43] Zhou Z, Lan M, Wu Y. A neural generation-based conversation model using fine-grained emotion-guide attention//Proceedings of the International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro, Brazil, 2018: 1-8

[44] Zhou G, Fang Y, Peng Y, et al. Neural conversation generation with auxiliary emotional supervised models. ACM Transactions on Asian Language Information Processing, 2019, 19(2): 1-17

[45] Asghar N, Poupart P, Hoey J, et al. Affective neural response generation//Proceedings of the European Conference on Information Retrieval. Grenoble, France, 2018: 154-166

[46] Warriner A B, Kuperman V, Brysbaert M. Norms of valence, arousal, and dominance for 13, 915 english lemmas. Behavior Research Methods, 2013, 45: 1191-1207

[47] Huang C, Zaane O, Trabelsi A, et al. Automatic dialogue generation with expressed emotions//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, USA, 2018: 49-54

[48] Sun X, Chen X, Pei Z, et al. Emotional human machine conversation generation based on seqgan//Proceedings of the First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia). Beijing, China, 2018: 1-6

[49] Colombo P, Witon W, Modi A, et al. Affect-driven dialog generation. arXiv preprint arXiv:1904.02793, 2019

[50] Peng D, Zhou M, Liu C, et al. Human-machine dialogue modelling with the fusion of word-and sentence-level emotions. Knowledge-Based Systems, 2020, 192: 105319

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

[52] Shen L, Feng Y. Cdl: Curriculum dual learning for emotion-controllable response generation. arXiv preprint arXiv: 2005.00329, 2020

[53] He D, Xia Y, Qin T, et al. Dual learning for machine translation//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 820-828

[54] Dong C, Huang C, Zaane O, et al. Simulated annealing for emotional dialogue systems//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. Queensland, Australia, 2021: 2984-2988

[55] Keskin S C. From what isn't empathy to empathic learning process. Procedia-Social and Behavioral Sciences, 2014, 116: 4932-4938

[56] Sabour S, Zheng C, Huang M. Cem: Commonsense-aware empathetic response generation//Proceedings of the AAAI Conference on Artificial Intelligence. British Columbia, Canada, 2022: 11229-11237

[57] Rashkin H, Smith E M, Li M, et al. Towards empathetic open-

domain conversation models: A new benchmark and dataset//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 5370-5381

[58] Xie Y, Svikhnushina E, Pu P. A multi-turn emotionally engaging dialog model. arXiv preprint arXiv:1908.07816, 2019

[59] Dudău D P, Sava F A. Performing multilingual analysis with linguistic inquiry and word count 2015 (liwc2015): An equivalence study of four languages. Frontiers in Psychology, 2021, 12: 570568

[60] Shin J, Xu P, Madotto A, et al. Generating empathetic responses by looking ahead the user's sentiment//Proceedings of the ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain, 2020: 7989-7993

[61] Lin Z, Madotto A, Shin J, et al. Moel: Mixture of empathetic listeners//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 121-132

[62] Majumder N, Hong P, Peng S, et al. Mime: Mimicking emotions for empathetic response generation. arXiv preprint arXiv:2010.01454, 2020

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

[64] Li Q, Chen H, Ren Z, et al. Empdg: Multi-resolution interactive empathetic dialogue generation//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain, 2020: 4454-4466

[65] Gao J, Liu Y, Deng H, et al. Improving empathetic response generation by recognizing emotion cause in conversations//Proceedings of the Findings of the Association for Computational Linguistics. Punta Cana, Dominican Republic, 2021: 807-819

[66] Jang E, Gu S, Poole B. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144, 2016

[67] Kim H, Kim B, Kim G. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. arXiv preprint arXiv:2109.08828, 2021

[68] Li Y, Li K, Ning H, et al. Towards an online empathetic chatbot with emotion causes//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 2041-2045

[69] Wei W, Liu J, Mao X, et al. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China, 2019: 1401-1410

[70] Tang D, Wei F, Yang N, et al. Learning sentiment-specific word embedding for twitter sentiment classification//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, 2014: 1555-1565

[71] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781, 2013

[72] Wei W, Liu J, Mao X, et al. Target-guided emotion-aware chat machine. ACM Transactions on Information Systems (TOIS), 2021, 39(4): 1-24

[73] Shen L, Zhang J, Ou J, et al. Constructing emotion consensus and utilizing unpaired data for empathetic dialogue generation. arXiv preprint arXiv:2109.07779, 2021

[74] Chen S-Y, Hsu C-C, Kuo C-C, et al. Emotionlines: An emotion corpus of multi-party conversations. arXiv preprint arXiv:1802.08379, 2018

[75] Zhiyuan W, Jiannong C, Ruosong Y, et al. Automatically select emotion for response via personality-affected emotion transition. arXiv preprint arXiv:2106.15846, 2021

[76] Kim W, Ahn Y, Kim D, et al. Emp-rft: Empathetic response generation via recognizing feature transitions between utterances. arXiv preprint arXiv:2205.03112, 2022

[77] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv: 1907.11692, 2019

[78] Sharma A, Miner A S, Atkins D C, et al. A computational approach to understanding empathy expressed in text-based mental health support. arXiv preprint arXiv:2009.08441, 2020

[79] Zou Y, Liu Z, Hu X, et al. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems. arXiv preprint arXiv:2109.04084, 2021

[80] Zhong P, Wang D, Li P, et al. Care: Commonsense-aware emotional response generation with latent concepts. arXiv preprint arXiv:2012.08377, 2020

[81] Li Q, Li P, Ren Z, et al. Knowledge bridging for empathetic dialogue generation//Proceedings of the AAAI Conference on Artificial Intelligence. British Columbia, Canada, 2022: 10993-11001

[82] Sap M, Le Bras R, Allaway E, et al. Atomic: An atlas of machine commonsense for if-then reasoning//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 3027-3035

[83] Bosselut A, Rashkin H, Sap M, et al. Comet: Commonsense transformers for automatic knowledge graph construction. arXiv preprint arXiv:1906.05317, 2019

[84] Chen M Y, Li S, Yang Y. Emphi: Generating empathetic responses with human-like intents. arXiv preprint arXiv: 2204.12191, 2022

[85] Zhong P, Zhang C, Wang H, et al. Towards persona-based empathetic conversational models//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2020: 6556-6566

[86] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018

[87] Firdaus M, Thangavelu N, Ekbal A, et al. I enjoy writing and playing, do you: A personalized and emotion grounded dialogue agent using generative adversarial network. IEEE Transactions

on Affective Computing，2022，1：1-1

[88] Li J，Sun X. A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels，Belgium，2018：678-683

[89] Sun Xiao，Li Jia，Wei Xing. Emotional editing constraint conversation generation based on reinforcement learning. Journal of Automatica Sinica，2019：1-16 (in Chinese)
(孙晓，李佳，卫星. 基于强化学习的情感编辑约束对话内容生成. 自动化学报，2019：1-16)

[90] Zandie R，Mahoor M H. Emptransfo：A multi-head transformer architecture for creating empathetic dialog systems. arXiv preprint arXiv：2003.02958，2020

[91] Zheng C，Liu Y，Chen W，et al. Comae：A multi-factor hierarchical framework for empathetic response generation. arXiv preprint arXiv：2105.08316，2021

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

[93] Qiu L，Shiu Y，Lin P，et al. What if bots feel moods? Towards controllable retrieval-based dialogue systems with emotion-aware transition networks//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Xi′an，China，2020：1161-1170

[94] Lu X，Tian Y，Zhao Y，et al. Retrieve，discriminate and rewrite：A simple and effective framework for obtaining affective response in retrieval-based chatbots//Proceedings of the Findings of the Association for Computational Linguistics. 2021：1956-1969

[95] Sharma A，Lin I W，Miner A S，et al. Towards facilitating empathic conversations in online mental health support：A reinforcement learning approach//Proceedings of the Web Conference. Ljubljana，Slovenia，2021：194-205

[96] Hu T，Xu A，Liu Z，et al. Touch your heart：A tone-aware chatbot for customer care on social media//Proceedings of the CHI Conference on Human Factors in Computing Systems. Quebec，Canada，2018：1-12

[97] Lubis N，Sakti S，Yoshino K，et al. Eliciting positive emotion through affect-sensitive dialogue response generation：A neural network approach//Proceedings of the AAAI Conference on Artificial Intelligence. Louisiana，USA，2018：5293-5300

[98] Saha T，Chopra S，Saha S，et al. A large-scale dataset for motivational dialogue system：An application of natural language generation to mental health//Proceedings of the International Joint Conference on Neural Networks (IJCNN). Shenzhen，China，2021：1-8

[99] Wang S，Xu X，Wu W，et al. Towards multi-turn empathetic dialogs with positive emotion elicitation. arXiv preprint arXiv：2204.10509，2022

[100] Bao S，He H，Wang F，et al. Plato-2：Towards building an open-domain chatbot via curriculum learning. arXiv preprint arXiv：2006.16779，2020

[101] Saha T，Gakhreja V，Das A S，et al. Towards motivational and empathetic response generation in online mental health support//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid，Spain，2022：2650-2656

[102] Lubis N，Sakti S，Yoshino K，et al. Dialogue model and response generation for emotion improvement elicitation//Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS). Vancouver，Canada，2019：1-11

[103] Li S，Feng S，Wang D，et al. Emoelicitor：An open domain response generation model with user emotional reaction awareness//Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama，Japan，2020：3637-3643

[104] Liu S，Zheng C，Demasi O，et al. Towards emotional support dialog systems. arXiv preprint arXiv：2106.01144，2021

[105] Peng W，Hu Y，Xing L，et al. Control globally，understand locally：A global-to-local hierarchical graph network for emotional support conversation. arXiv preprint arXiv：2204.12749，2022

[106] Tu Q，Li Y，Cui J，et al. Misc：A mixed strategy-aware model integrating comet for emotional support conversation. arXiv preprint arXiv：2203.13560，2022

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

[108] Liang Y，Meng F，Zhang Y，et al. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. arXiv preprint arXiv：2012.04882，2020

[109] Shao Z，Xu Y，Wei W，et al. Heterogeneous graph neural network with multi-view representation learning. arXiv preprint arXiv：2108.13650，2021

[110] Liang Y，Meng F，Zhang Y，et al. Emotional conversation generation with heterogeneous graph neural network. Artificial Intelligence，2022，308：103714

[111] Zhu L Y，Zhang Z，Wang J，et al. Multi-party empathetic dialogue generation：A new task for dialog systems//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin，Ireland，2022：298-307

[112] Qiu L，Zhao Y，Li J，et al. Valuenet：A new dataset for human value driven dialogue system. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)，2022，44(5)：2468-2484

[113] Wulczyn E，Thain N，Dixon L. Ex machina：Personal attacks seen at scale//Proceedings of the 26th International Conference on World Wide Web. Perth，Australia，2017：1391-1399

[114] Lahnala A，Welch C，Neuendorf B，et al. Mitigating toxic degeneration with empathetic data：Exploring the relationship between toxicity and empathy. arXiv preprint arXiv：2205.07233，2022

[115] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. California，USA，2017：6000-6010

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

[117] Ide T, Kawahara D. Multi-task learning of generation and classification for emotion-aware dialogue response generation. arXiv preprint arXiv:2105.11696, 2021

[118] Xie B, Park C H. Empathetic robot with transformer-based dialogue agent//Proceedings of the 18th International Conference on Ubiquitous Robots (UR). Gangwon-do, Korea, 2021: 290-295

[119] Muthukumar P, Muthukumar K, Muthirayan D, et al. Generative adversarial imitation learning for empathy-based ai. arXiv preprint arXiv:2105.13328, 2021

[120] Ho J, Ermon S. Generative adversarial imitation learning//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 4572-4580

[121] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners//Proceedings of the Advances in Neural Information Processing Systems. Online, 2020: 1877-1901

[122] Knox W B, Stone P. Augmenting reinforcement learning with human feedback//Proceedings of the ICML Workshop on New Developments in Imitation Learning. Bellevue, USA, 2011: 3

[123] Knox W B, Stone P. Interactively shaping agents via human reinforcement: The tamer framework//Proceedings of the 5th International Conference on Knowledge capture. California, USA, 2009: 9-16

[124] Kosinski M. Theory of mind may have spontaneously emerged in large language models. arXiv preprint arXiv: 2302.02083, 2023

[125] Zhang B, Ding D, Jing L. How would stance detection techniques evolve after the launch of chatgpt? arXiv preprint arXiv:2212.14548, 2022

[126] Tenhundfeld N L. Two birds with one stone: Writing a paper entitled "chatgpt as a tool for studying human-ai interaction in the wild" with chatgpt. Eurasia Journal of Mathematics, Science and Technology Education, 2023, 19(7): em2286

[127] Yeo Y H, Samaan J S, Ng W H, et al. Assessing the performance of chatgpt in answering questions regarding cirrhosis and hepatocellular carcinoma. medRxiv, 2023: 2023.2002.2006.23285449

[128] Li J, Monroe W, Ritter A, et al. Deep reinforcement learning for dialogue generation. arXiv preprint arXiv:1606.01541, 2016

[129] Shin J, Xu P, Madotto A, et al. Happybot: Generating empathetic dialogue responses by improving user experience look-ahead. arXiv preprint arXiv:1906.08487, 2019

[130] Zhang H, Lan Y, Guo J, et al. Reinforcing coherence for sequence to sequence model in dialogue generation//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 4567-4573

[131] Cui S, Lian R, Jiang D, et al. Dal: Dual adversarial learning for dialogue generation. arXiv preprint arXiv:1906.09556, 2019

[132] Felbo B, Mislove A, Søgaard A, et al. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv: 1708.00524, 2017

[133] Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain, 2017: 427-431

[134] Li Y, Su H, Shen X, et al. Dailydialog: A manually labelled multi-turn dialogue dataset//Proceedings of the 8th International Joint Conference on Natural Language Processing. Taipei, China, 2017: 986-995

[135] Welivita A, Xie Y, Pu P. A large-scale dataset for empathetic response generation//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2021: 1251-1264

[136] Welivita A, Pu P. A taxonomy of empathetic response intents in human social conversations. arXiv preprint arXiv: 2012.04080, 2020

[137] Poria S, Hazarika D, Majumder N, et al. Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508, 2018

[138] Zhao J, Zhang T, Hu J, et al. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. arXiv preprint arXiv: 2205.10237, 2022

[139] Shang L, Lu Z, Li H. Neural responding machine for short-text conversation. arXiv preprint arXiv:1503.02364, 2015

[140] Mohammad S. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 174-184

[141] Speer R, Chin J, Havasi C C. 5.5: An open multilingual graph of general knowledge//Proceedings of the 31st AAAI Conference on Artificial Intelligence. California, USA, 2017: 4444-4451

[142] Fleiss J L. Measuring nominal scale agreement among many raters. Psychological Bulletin, 1971, 76(5): 378

[143] Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model//Proceedings of the Advances in Neural Information Processing Systems. 2000: 1137-1155

[144] Mou L, Song Y, Yan R, et al. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. arXiv preprint arXiv: 1607.00970, 2016

[145] Papineni K, Roukos S, Ward T, et al. Bleu: A method for automatic evaluation of machine translation//Proceedings of the 40th Aannual Meeting of the Association for Computational Linguistics. Pennsylvania, USA, 2002: 311-318

[146] 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. arXiv preprint arXiv: 1603.08023, 2016

附录.

(1) 32种情绪类别：Prepared、Anticipating、Hopeful、Proud、Excited、Joyful、Content、Caring、Grateful、Trusting、Confident、Faithful、Impressed、Surprised、Terrified、Afraid、Apprehensive、Anxious、Embarrassed、Ashamed、Devastated、Sad、Disappointed、Lonely、Sentimental、Nostalgic、Guilty、Disgusted、Furious、Angry、Annoyed、Jealous

(2)附表说明：表1至表2囊括了可控情感对话生成的模型，剩下表3至表6则为共情对话响应、情绪支持、多模态情感对话生成模型的指标.

(3)附表指标说明：B、B-1、B-2、B-3、B-4指的是BLEU、BLEU-1、BLEU-2、BLEU-3、BLEU-4；EA、GM、VE指的是使用Embedding的三种评价方式：Greedy Matching、Embedding Average、Vector Extrema；D-1、D-2指的是Dist-1、Dist-2；其中表4中，由于研究者在处理EmpatheticDialogue数据集时的方式差异，加粗指标之间的值具有可比性，相应地，而未加粗的指标之间具有可比性.

附表1　模型在2017 NLPCC Corpus(ESTC)上的指标

	作者及模型	PPL	B-1	B-2	B-3	B-4	EA	GM	VE	Acc	D-1	D-2
感知模型	Zhang等人(EsCVAE-II,2018)[32]	46.1	-	-	0.114	-	-	-	-	0.675	-	-
	Zhang等人(EsCVAE-II,2018)[32]	44.3	-	-	0.139	-	-	-	-	0.69	-	-
	Peng等人(TE-ECG,2019)[34]	-	-	-	-	-	-	-	-	0.59	0.06	0.206
	杨等人(2021)[35]	-	-	-	-	-	-	-	-	-	0.057	0.202
表示模型	Zhou等人(ECM,2018)[26]	65.9	0.1023	0.0332	-	1.68	0.624	0.434	0.409	0.773	0.05	0.173
	Zhang等人(MECS,2018)[39]	169.45	-	-	-	-	-	-	-	0.9658	-	-
	Song等人(EmoDS,2019)[41]	-	0.0976	0.0282	-	1.73	0.634	0.451	0.435	0.81	0.05	0.174
	Zhou等人(ECGM,2018)[43]	-	-	-	-	0.5007	-	-	-	0.6583	0.0515	0.2242
	Zhou等人(nCG-ESM,2019)[44]	-	-	-	-	-	-	-	-	-	0.106	0.44
	Shen等人(CDL,2020)[52]	-	0.1254	0.037	-	-	0.642	0.438	0.457	0.823	0.065	0.221
	Dong等人(2021)[54]	-	0.1426	0.0412	-	-	0.786	0.556	0.611	0.942	0.053	0.239

附表2　模型在Mojitalk data、MEIMD、Cornell Movie Dialogs Corpus数据集上的指标

	作者及模型	Mojitalk data		MEIMD				Cornell Movie Dialogs Corpus			
		PPL	Acc	PPL	EA	GM	VE	PPL	D-1	D-2	B
感知模型	Kong等人(CVAE-CGAN,2019)[33]	69.54	78.8	-	-	-	-	-	-	-	-
	Zhou等人(MojiTalk,2018)[31]	38.3	42.1	-	-	-	-	-	-	-	-
表示模型	Zhou等人(ECM,2018)[26]	-	-	74.6	0.519	0.375	0.381	-	0.031	0.1412	0.0099
	Firdaus等人(MEI-DG,2021)[42]	-	-	71.2	0.552	0.443	0.428	-	-	-	-
	Song等人(EmoDS,2019)[41]	-	-	74.1	0.526	0.389	0.387	-	-	-	-
	Shantala等人(2018)[40]	-	-	-	-	-	-	4.85	-	-	-
	Colombo等人(EmotiCons,2019)[49]	-	-	74.3	0.523	0.381	0.385	-	0.0406	0.203	0.014

附表3　模型在EmoElicitor Corpus、MELD、DailyDialog上的指标

		EmoElicitor Corpus				MELD				DailyDialog			
	作者及模型	D-1	D-2	B	Acc 5%	PPL	B	D-1	D-2	PPL	B	D-1	D-2
	Zhou等人(ECM,2018)[26]	0.3	1.085	13.57	48	-	-	-	-	-	-	-	-
	Zhou等人(MojiTalk,2018)[31]	4.16	13.18	15.04	50.1	-	-	-	-	-	-	-	-
激发指定情绪	Li等人(EmoElicitor,2020)[103]	4.08	21.97	18.54	51.7	-	-	-	-	-	-	-	-
激发积极情绪	Lubis等人(Emo-HRED,2018)[97]	-	-	-	-	116.14	1.87	1.23	3.77	81.48	0.54	0.12	0.23
多模态	Liang等人(2020)[108]	-	-	-	-	96.9	2.8	4.84	9.15	76.59	0.75	0.22	0.47
	Liang等人(2022)[110]	-	-	-	-	96.34	3.08	5.08	9.58				

附表4 模型在EmpatheticDialogue的指标

	作者及模型	PPL	Acc	D-1	D-2	B	B-4	EA	GM	VE
话语级情绪	Rashkin等人(EmoPrepend-1,2018)[57]	38.3	33.28	**0.46**	**2.08**	-	-	-	-	-
	Rashkin等人(MultiTRS,2018)[57]	33.35	33.28	1.06	4.29	2.92	-	0.938	0.786	0.541
	Shin等人(2022)[60]-Forward	-	-	2	0.123	6.292	-	0.961	0.947	0.901
	Shin等人(2022)[60]-Improvement	-	-	2.6	0.17	6.317	-	0.966	0.952	0.919
	Shin等人(2022)[60]-Simulation	-	-	1.1	0.08	6.192	-	0.966	0.949	0.919
混合情绪	Lin等人(MoEL,2019)[61]	33.58	32	1.38/**0.56**	4.66/**2.70**	2.9	1.61	0.945	0.793	0.537
	Majumder等人(MIME,2020)[62]	37.53	-	0.51	2.68	2.98	1.578	0.946	0.794	0.536
词级情绪	Li等人(EmpDG,2020)[64]	34.18	34.31	1.81/**0.47**	6.94/**2.10**	-	1.506	0.935	0.788	0.532
	Gao等人(2021)[65] -hard	-	42.3	1.8	9	-	1.734	-	-	-
	Gao等人(2021)[65] soft	-	42.4	1.7	8.4	-	1.774	-	-	-
情绪关联	Shen等人(Dual-Emp,2021)[73]	31.01	37.53	**1.08**	**3.23**	2.91	-	0.957	0.796	0.545
	Kim等人(Emp-RFT,2022)[76]	14.71	42.08	**3.21**	**4.48**	-	-	-	-	-
常识	Li等人(KEMP,2022)[81]	36.89	39.31	0.55	2.29	-	-	-	-	-
	Sabour等人(CEM,2022)[56]	36.11	39.11	0.66	2.99	-	-	-	-	-
意图	Chen等人(EmpHi,2022)[84]	-	-	**1.1188**	**5.3332**	-	-	-	-	-

附表5 模型在CARE data和ESConv数据集上的指标

	CARE data (Reddit)				CARE data(Twitter)				ESConv			
作者及模型	PPL	D-1	D-2	Acc	PPL	D-1	D-2	Acc	PPL	B-2	D-1	D-2
Zhou等人(ECM,2018)[26]	65.6	0.0044	0.0506	0.5893	91.3	0.0056	0.063	0.5619	-	-	-	-
Song等人(EmoDS,2019)[41]	76.6	0.003	0.0455	0.6186	113.5	0.003	0.045	0.595	-	-	-	-
Zhong等人(CARE,2020)[80]	70.4	0.0049	0.046	0.684	100.1	0.0064	0.0775	0.6693	-	-	-	-
Lin等人(MoEL,2019)[61]	-	-	-	-	-	-	-		62.93	5.02	2.71	14.92
Majumder等人(MIME,2020)[62]	-	-	-	-	-	-	-		43.27	4.82	2.56	12.33
Peng等人(GLHG,2022)[105]	-	-	-	-	-	-	-		15.67	7.57	3.5	21.61
Tu等人(MISC,2022)[106]	-	-	-	-	-	-	-		16.16	7.31	4.41	19.71

附表6 模型在ESTC、NLPCC2017+OpenSubtitles、MC-HRED data上的指标

		ESTC				STC+ OpenSubtitles		MC-HRED data
类型	作者及模型	D-1	D-2	B-1	B-2	PPL	Acc	PPL
情绪关联	Wei等人(EACM,2021)[69]	0.0819	0.284	0.2305	0.1518	-	-	-
	Wei等人(TG-EACM,2020)[72]	0.0839	0.407	0.2708	0.1832	-	-	-
	Zhou等人(ECM,2018)[26]	0.0603	0.207	0.2316	0.1525	-	-	-
对话主题	Li等人(E-SCBA,2018)[88]	-	-	-	-	65.6	0.771	-
	孙等人(2019)[89]	-	-	-	-	61.533	0.87	-
激发积极情绪	Lubis等人(Emo-HRED,2018)[97]	-	-	-	-	-	-	42.6
激发指定情绪	Lubis等人(MC-HRED,2019)[102] unigram-Action	-	-	-	-	-	-	49.74
	Lubis等人(MC-HRED,2019)[102] bigram-Action	-	-	-	-	-	-	49.62
	Lubis等人(MC-HRED,2019)[102] trigram-Action	-	-	-	-	-	-	49.78

YANG Zhou, Ph. D. candidate, His research interests focus on natural language processing.

CHENG Zhi-Hao, Ph. D. candidate, His research interests include social network information diffusion analysis and sentiment analysis.

CAI Tie-Cheng, M. S, candidate. His research interests include affective computing, image captioning and natural language processing.

WANG Yu-Feng, Ph. D. candidate, His research interests focus on natural language processing.

LIAO Xiang-Wen, Ph. D, professor. His research interests include natural language processing, information retrieval, social computing and network content security.

Background

Human-machine dialogue technology is widely applied in various business fields, such as chatbots, question-answering systems, and product recommendation systems. Human-machine dialogue technology has provided significant economic returns in many scenarios. However, due to the involvement of emotional communication in dialogues, the lack of emotional intelligence in dialogue technology hinders the provision of better emotional communication and care. This limitation restricts the further development of human-machine dialogue technology. Meanwhile, in recent years, deep learning technology and its application in the field of dialogue have experienced rapid development. In order to compensate for the lack of emotion in dialogue technology using deep learning techniques, the task of deep learning based emotional dialogue response has been proposed and rapidly developed. We summarize the deep learning based emotional dialogue response and divide it into five subtasks: controllable emotion dialogue generation, empathetic dialogue response, emotion support, multimodal dialogue generation, and new tasks. Subsequently, we also review and analyze the models based on commonly used model structures. Furthermore, we introduce commonly used datasets and evaluation metrics. Based on the above summary and analysis, we further propose the development trends of this task.

This work was supported by National Natural Science Foundation of China (No. 61976054)