软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(3):1377−1402 [doi: 10.13328/j.cnki.jos.006807]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

情感对话技术综述*

赵妍妍, 陆 鑫, 赵伟翔, 田一间, 秦 兵

(哈尔滨工业大学 计算学部, 黑龙江 哈尔滨 150006)
通信作者: 秦兵, E-mail: bqin@ir.hit.edu.cn

摘 要: 情感对话技术着眼于对话机器人的“情商”, 旨在让对话机器人具有像人类一样观察、理解和表达情感的能力. 这项技术可以看作是情感计算和对话技术的交叉, 兼顾对话机器人的“智商”和“情商”, 从而实现对用户的精神陪伴、情感慰藉和心理疏导. 结合对话中情感的特点, 对情感对话技术进行全面解析: 1) 规划对话场景下的情感识别、情感管理、情感表达等 3 个重要的技术点, 并拓展多模态场景下的情感对话技术. 2) 介绍情感对话 4 项关键技术的最新研究进展, 并总结这些技术面临的主要挑战和可能解决方案. 3) 介绍情感对话技术的数据资源. 4) 分析情感对话技术的难点, 展望其未来发展方向与前景.
关键词: 情感计算; 对话情感识别; 对话情感表达; 对话情感管理; 多模态情感对话
中图法分类号: TP18

中文引用格式: 赵妍妍, 陆鑫, 赵伟翔, 田一间, 秦兵. 情感对话技术综述. 软件学报, 2024, 35(3): 1377–1402. http://www.jos.org.cn/1000-9825/6807.htm
英文引用格式: Zhao YY, Lu X, Zhao WX, Tian YJ, Qin B. Survey on Emotional Dialogue Techniques. Ruan Jian Xue Bao/Journal of Software, 2024, 35(3): 1377–1402 (in Chinese). http://www.jos.org.cn/1000-9825/6807.htm

Survey on Emotional Dialogue Techniques

ZHAO Yan-Yan, LU Xin, ZHAO Wei-Xiang, TIAN Yi-Jian, QIN Bing
(Faculty of Computing, Harbin Institute of Technology, Harbin 150006, China)

Abstract: Emotional dialogue technology focuses on the “emotional quotient” of conversational robots, aiming to give the robots the ability to observe, understand and express emotions as humans do. This technology can be seen as the intersection of emotional computing and dialogue technology, and can simultaneously consider the “intelligent quotient” and “emotional quotient” of conversational robots to realize spiritual companionship, emotional comfort, and psychological guidance for users. Combined with the characteristics of emotions in dialogues, this study provides a comprehensive analysis of emotional dialogue technology: 1) Three important technical points including emotion recognition, emotion management, and emotion expression in dialogue scenarios are shown, and the technology of emotional dialogues in multimodal scenarios is expanded. 2) This study presents the latest research progress on technology points related to emotional dialogues and summarizes the main challenges and possible solutions correspondingly. 3) Data resources for emotional dialogue technologies are introduced. 4) The difficulty and prospect of emotional dialogue technology are pointed out.
Key words: emotional computing; emotion recognition in dialogue; emotion expression in dialogue; emotion management in dialogue; multimodal emotional dialogue

在情感对话技术出现之前, 最初的对话机器人的设计出发点是“模拟人的智商”, 最常见的表现形式为智能私人助手. 这种类型的对话机器人应答迅速、博闻强记, 为客户提供专业优质的体验, 极大提升了客户的满意度. 常见的智能私人助手类的对话机器人包括国外的苹果 Siri、谷歌 Google Assistant, 以及国内华为的智慧语音助手小艺、科大讯飞的灵犀等. 然而, 拥有了“智商”的对话机器人有时仍然不能满足图灵测试中的类人性, 尤其对聊天机

* 基金项目: 国家重点研发计划 (2018YFB1005103); 国家自然科学基金 (62176078)
收稿时间: 2022-04-23; 修改时间: 2022-07-11; 采用时间: 2022-09-05; jos 在线出版时间: 2023-04-26
CNKI 网络首发时间: 2023-04-27

器人而言. 人类的特性是不仅可以回复合乎逻辑的答复, 还可以将情感处理得十分得当. 然而, 目前对话机器人对于情感的处理还较为欠缺, 回复虽然在内容可以做到十分相关, 但是情感不恰当的情况却十分普遍. 人类期待对话机器人可以脱掉冷冰的外壳, 具有处理情感的能力, 即具有“情商”, 向类人化更进一步. 在这种背景下, 情感对话技术应运而生.

情感对话技术是指赋予对话机器人类人式的情感, 使其具有识别、理解和表达情感的能力, 为用户提供更人性化、多样化的回复. 让计算机具有“情商”是人工智能的更高阶的目标. 人工智能之父马文·明斯基 (Marvin Minsky) 在其著作《情感机器》[1]中指出: “人工智能只有智力, 没有情感, 不是真正的智能.” 软银 CEO 孙正义认为“未来更有价值的机器人是能够理解人心, 能够与人进行情感沟通的机器人”[2]. 这些都表明对话机器人不仅需要“智商”, 还需要“情商”. 因此, 研制具有情感的对话机器人成为学术界和产业界共同关心的课题. 近些年来, 产业界不少对话机器人产品均增加了情感功能, 以提升产品的趣味性和人性化程度, 如微软小冰、小黄鸡聊天机器人等. 情感对话技术是提升这些应用产品如闲聊机器人[3]、智能客服[4]、语音助手[5]等性能的核心技术.

对话机器人的情感功能重点依赖于对话语言内容的分析, 所用技术则集中于自然语言处理领域中的文本情感计算技术. 这项技术发展至今已有 20 多年的历史, 传统的文本情感计算技术的研究对象多为产品评论信息或网民舆情信息, 与这些传统的情感计算场景相比, 对话场景下的情感有其特殊性.

(1) 情绪是主流

与评论中的情感倾向性不同, 情感对话机器人中的情感交互以情绪为主, 表现为喜悦、愤怒、悲伤、恐惧、惊奇等. 例如, “今天是我的生日, 好开心呐”, 表现出“喜悦”的情绪. 情绪由于类别较多, 且存在个别情绪区分度受限 (如“悲伤”和“愤怒”), 给对话情绪识别带来了一定的困难. 为了方便术语统一, 下文的“情感”均是指“情绪”.

(2) 有丰富的对话场景信息

有丰富的对话场景信息, 以及对对话场景进行建模是情感对话技术区别于其他文本情感计算技术最大的不同, 也是情感对话技术的研究重点. 此外, 对话场景不仅包括常见的在线人机聊天场景, 还包括多模态对话场景. 在文本情感计算、语音识别与合成、人脸表情识别等方面, 人工智能领域的研究者们均取得了引人注目的进展, 为多模态场景下的情感对话机器人提供了技术支撑.

(3) 用户信息在情感中的参与性

由于对话机器人的聊天对象是用户, 更强调用户的个性化与参与性. 例如: 同样表达“心情不错”这个情感, 外向开朗的用户会用“好极了”来表述, 而谨慎内向的用户会用“还可以”来表述, 这为情感识别带来了挑战.

(4) 话题信息在情感中的参与性

用户所讨论的话题与要生成回复的情感之间有非常直接的联系. 例如上文是“我生病了, 好难受啊”, 表达的情感是“悲伤”, 自动生成回复的情感之所以是“同情”, 是由上文的“生病”话题决定的. 同理, 如果上文是“天呐, 我考试又没及格”, 表达的情感同样是“悲伤”, 而由于话题是“考试”下文生成的情感就很有可能是“愤怒”. 由此可见, 待生成回复的情感跟话题是息息相关的.

(5) 隐式情感占一定比例

情绪化、口语化和随意性是对话内容的特点, 这导致了对话机器人系统中隐式情感句所占比例较大. 例如, “我今天考试不及格”和“我今天中奖了”均不含有任何情感词, 却分别表达出“悲伤”和“喜悦”的情感. 由于隐式情感句内部情感信息较为隐晦, 局部考察句内特征无法提供足够的分类证据, 因此需要依赖外部的背景知识.

(6) 情感回复的不确定性

对于传统的情感识别任务而言, 识别出的情感是确定的. 然而对于对话场景下的情感识别任务而言, 针对一个用户给出的带有情感的上文, 生成的下文的情感回复具有一定的不确定性. 例如: 上文是“我今天考试不及格”, 情感回复的情感有可能是“愤怒”, 也有可能是“悲伤”, 还有可能是“同情”, 需要根据具体的话题和用户特性而定.

针对以上对话场景下的情感表达的特点, 近年来国内外的研究人员进行了很多探索. 由于问题的复杂性, 情感对话技术在国内外还处于起步阶段, 现有的工作大部分集中于对话场景下的话语情感识别[6−10]以及给定情感信号的情感回复生成[11−14]工作, 非常类似于对话场景下的情感分类和回复生成技术. 该领域还有大量深入的研究问题

值得探索, 例如: 对话场景下的情感还有哪些特点值得挖掘; 对话情感识别与情感回复生成两项技术之间是否需要其它的衔接技术; 如何在情感回复过程中更好的体现终极目标——共情[15−19]等.

基于以上的思考和大量的调研, 本文在后续部分首先规划了情感对话技术的任务体系和理论基础; 接着介绍情感对话技术的最新进展, 总结该技术面临的主要挑战和可能解决方案; 然后介绍情感对话技术的语料资源; 最后, 分析情感对话技术的难点, 展望其未来发展方向与前景.

1 情感对话技术的任务体系及技术基础

1.1 任务体系

人类在处理对话中的情感时, 需要先根据对话场景寻找线索判断出对方的情感, 继而根据对话的主题等信息思考用什么情感进行回复, 最后结合推理出的情感形成恰当的回复. 受人类处理情感对话的启发, 情感对话技术需要涵盖以下几项任务.

(1) 对话情感识别: 旨在识别出用户当前对话中蕴含的情感. 这里的情感通常根据不同的应用场景有所差别. 例如, 在聊天机器人中, 对话情感多指喜怒哀恐惊等人的情绪; 而在智能私人助手中, 对话情感多指支持反对等二元判断策略.

(2) 对话情感管理: 旨在对用户当前对话情感产生的原因进行深入理解, 精准定位对话情感产生的具体场景 (如: 闲聊、安抚、心理检测等), 推理出对话机器人回复应持有的情感, 以便及时激活对应的对话回复生成模块. 该任务是一个总控模块, 上接对话情感识别模块, 下接对话情感表达模块. 值得说明的是, 该任务受启发于人机对话技术中的对话管理任务, 是本文首次提出的概念.

(3) 对话情感表达: 旨在结合用户当前对话的主题及对话管理环节推理出的场景和情感, 生成既内容充实有针对性又情感逻辑合理的回复, 保证人机对话通畅自然. 带有“情商”的回复可以帮助对话机器人提高人机对话轮次, 增强对话黏着性.

(4) 多模态场景下的情感对话技术: 旨在分析多模态场景下对话的情感特点, 尤其是挖掘语言、图像和声音这 3 种模态的情感特征并加以融合, 在不同模态中切换寻得线索, 并相互关联进行情感的识别、理解和表达. 由于多种模态的参与, 情感特征的源头也是多路的, 如何进行模态间信息互补和处理模态间不一致现象是多模态情感研究的重点, 也是多模态对话相比单模态对话的不同之处.

图 1 展示了情感对话技术的任务体系及技术基础. 在任务体系中, 情感对话机器人系统的基石是对话机器人需要具有一定的“智商”, 即能够对用户当前的对话进行充分的自然语言理解, 并具有根据用户意图生成相应回复的能力. 此外, 值得强调的是, 无论何时, 都不能抛开“智商”谈“情商”. 这也意味着, 对话“内容相关”和“情感相关”是对话机器人非常核心的两个特质. 基于具有一定智商的对话机器人, 本文提出从对话情感识别、对话情感管理和对话情感表达 3 个层面逐层递进的增强对话机器人的“情商”. 这 3 项研究内容是层层递进, 密不可分的. 其中情感识别是基础; 情感管理是深化, 属于情感的深层次理解和激活模块; 情感表达依托前两项技术, 与实际应用及价值体现息息相关.

以上 3 项技术主要在自然语言单模态环境下开展研究, 在兼顾智商的情况下, 用以提升在线对话机器人的情商. 更进一步的分析, 由于对话机器人在模拟人的对话特征, 我们可以观察到人类之间对话不只有语言这一种模态, 而是大部分情况下处于多种模态共存的场景下, 具体表现为通过语言、图像、声音及手势等共同作用无缝衔接地表达我们的意图和情感, 而对话接收方同样用眼睛观察我们的行为、面部表情, 用耳朵听到语言、语音、语调, 然后用大脑来处理对话中参与的多维度信息, 给出反馈. 对于模拟人类的对话机器人而言, 不仅需要具备处理自然语言情感的能力, 还需具备处理多种模态中多源情感信息的能力, 这无疑将成为未来情感对话机器人的必备技能, 尤其对于实体对话机器人而言. 具有多模态场景处理能力的情感对话机器人将会更符合真实场景, 更能提升用户的参与度, 从而对高质量的智能陪伴、智能客服、电子商务等应用提供极大的技术支撑. 基于此, 情感对话机器人还需要具备在多模态、大语境场景下的“情商”, 即上面提到的第 4 项任务——多模态场景下的情感对话技术.

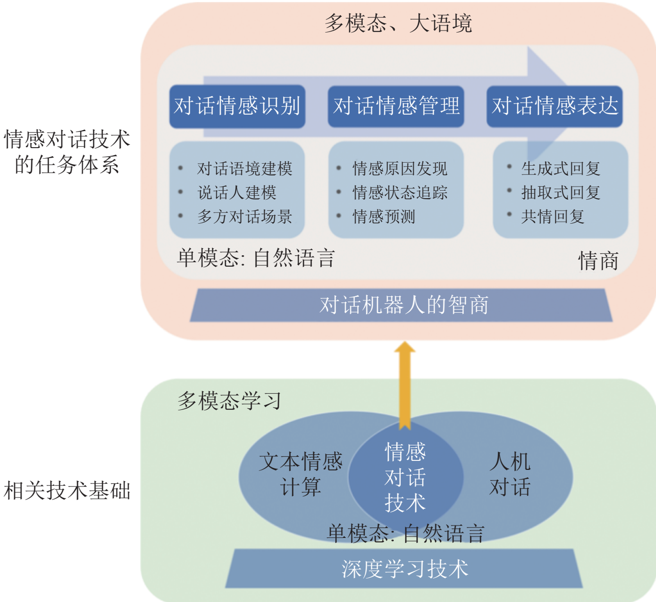


图 1 情感对话技术的任务体系及技术基础

虽然情感是对话机器人很重要的一个特质, 然而目前的对话机器人产品要么对用户的情感对话处理的不够精准, 要么不具备情感处理的功能, 更谈不上深入情感理解、情感管理并生成情感回复了. 具有情感处理能力的大部分对话机器人仅使用了情感分类模型判断用户对话的情感, 事实证明这种浅层的文本情感计算使得人机交互的结果并没有得到很明显的改善. 因此, 如何更好地规划情感计算与人机对话技术的融合, 让情感深入到对话机器人系统的方方面面, 即: 更精准地识别用户情感, 深入理解用户产生情感的原因以激活对应的对话情感状态, 以及生成更有个性化的情感回复, 是非常值得探索的课题.

1.2 技术基础

情感对话系统需要多项人工智能技术的支撑, 最相关的核心技术共有 4 项, 分别是: 深度学习技术[20]、文本情感计算技术[21]、人机对话技术[22]以及多模态学习技术[23], 如图 1 所示.

深度学习是当今乃至未来很长一段时间内引领人工智能发展的核心技术, 因此也是情感对话机器人系统的技术基础. 深度学习技术可以端到端的学习和处理自然语言、语音和图像等多源数据, 不再依赖人工设计的特征, 更重要的是深度学习技术打通了多种模态的技术屏障, 使得多模态学习变得可以实现, 距离真正的类人化更进一步.

在深度学习技术的基础之上, 文本情感计算技术、人机对话技术以及多模态学习技术是与情感对话机器人系统最相关的几项技术, 近年来得到了长足的发展. 文本情感计算技术使得计算机拥有了理解文本中蕴含情感的能力, 人机对话技术使得计算机在人机对话环境下拥有了理解用户对话意图和生成相应文本回复的能力, 多模态学习技术为计算机从自然语言单模态扩展到多模态语义学习提供了保障. 从图 1 可以看出, 从语言模态角度看, 情感对话技术是文本情感计算领域和人机对话领域的交集, 除了需要兼顾这两个领域的技术, 还具有本身新方向的特色; 此外, 从多模态场景角度看, 情感对话技术将会联合几个相关技术方向交叉出更多有特色的研究点.

2 情感对话技术的算法分析与总结

本节按照图 1 的任务体系来详细介绍这 4 项情感对话技术的研究方法, 主要包括整理任务定义、任务特点、技术难点及思路概括, 以及详细的现有研究思路剖析、算法优缺点分析, 并提出可能的解决方案等.

2.1 对话情感识别

● 任务定义

给定一段对话内容以及其中每条语句的说话人信息, 对话情感识别任务旨从一组预先定义的情感标签中识别出每条对话语句所对应的情感. 该任务可以由如下形式化定义给出. 给定由 N 条连续对话语句 $\{u_1, u_2, \ldots, u_N\}$ 和 M 个说话人 $\{s_1, s_2, \ldots, s_M\}$ 构成的对话, 任务的目标是预测每个由 s_i 说出来的话语 u_i 的情感标签 e_i. 图 2 展示了对话情感识别任务中一段双人对话的例子, 其中每条对话语句都有对应的情感标签.

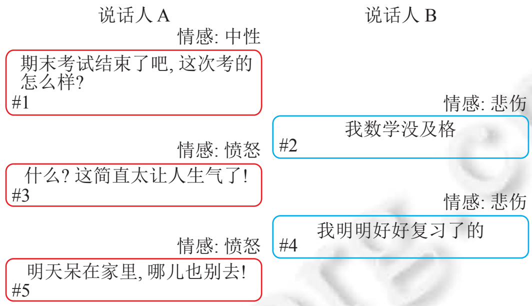


图 2　对话情感识别中的双人对话

● 任务特点

从对话情感识别的定义来看, 非常类似于文本情感计算任务中的基础任务——情感分类. 对话情感识别的特殊之处体现在如下 3 点.

(1) 存在对话语境: 口语化严重、富含隐式情感信息是对话的典型特点, 因此仅考虑单一的对话语句来进行情感判定是不够的, 需要对当前对话语句进行语境建模.

(2) 有说话人参与: 对话的主体是说话人, 因此说话人的性格及用词特点等将会对对话语句情感识别产生较大的影响.

(3) 多方对话场景是常态: 多人参与对话、讨论多个话题是对话场景中较为常见的模式, 很显然这种模式比二人对话场景要更为复杂.

● 技术难点及思路概括

(1) 对话语境建模

对话语境的表象是当前对话语句的上下文文本, 而与“情感”相关的还包括蕴含于语境中的话题、说话人情感延续、说话人意图及心理状态等因素, 如何或显式或隐式的表示这些对话情感相关的因素, 以及如何对这些因素进行建模, 是对话语境建模的技术难点.

现有的研究思路可概括为图 3 的方法, 可分为 3 类: 基于上下文文本建模、基于上下文情感建模以及基于引入外部知识进行建模, 分别对应话题、情感延续等技术难点.

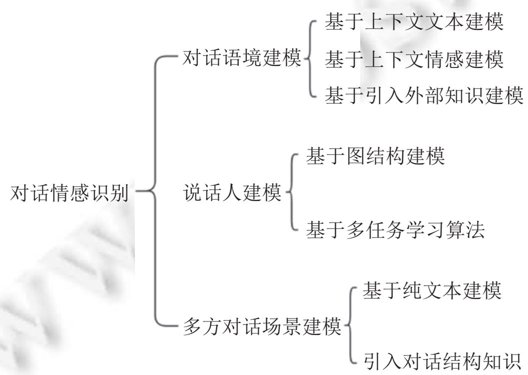


图 3　对话情感识别算法分类

(2) 说话人建模

每个人由于其性格、阅历或是成长环境等自身条件的不同, 都有其表达情感的独特方式, 如: 有的人情绪稳定、有的人情绪波动大、有的人共情能力强等. 然而, 这些情感表达方式很难被显式表达或定义. 基于此, 说话人的特点如何表示, 更进一步如何与对话语进行联合表示, 是对话场景下说话人建模的技术难点.

现有的研究思路主要是将说话人进行隐式表示, 由训练语料学习出说话人表示, 如图 3 所示主要包括基于图结构建模以及基于多任务学习算法, 分别对应说话人特点表示及联合表示等技术难点. 还未有相关工作对说话人进行显式表示.

(3) 多方对话场景建模

多方对话场景下的对话情感识别任务主要有两个技术难点: 1) 不同的说话人以较为随机的方式陆续参与对话, 并关注他们感兴趣的某些话题或具有独特的关注点, 从而使得相应话题在不同的对话线程中得以发展. 2) 说话人对会话历史中相关的话语进行回复, 并且其回复的话语可能出现在多轮交互之前, 从而带来话语之间较长的依赖性. 基于此, 在多方对话中, 建模每一个说话人的状态和处理不同说话人话语可能存在的长期依赖更为困难, 与双人对话相比, 此时的对话情感识别更具挑战性.

相比双人对话而言, 侧重于研究多方对话的工作并不多. 现有的研究思路可概括为图 3 的方法, 可分为两类: 基于纯文本建模以及引入对话结构知识的算法.

基于上述分析, 本文将相关研究工作分为 3 类进行简要介绍, 如图 3 所示, 分别是基于对话语境建模的对话情感识别、基于说话人建模的对话情感识别和多方对话场景下的对话情感识别.

2.1.1 基于对话语境建模的对话情感识别算法

对于情感对话识别技术而言, 对话语境建模是最重要的基石, 后续的一些算法, 如第 2.1.2 节的“说话人建模”等都需要基于对话语境建模而开展. 对话语境主要涵盖了对话的内容信息和情感信息, 相关的研究工作表明这两方面的信息可以帮助算法模型提升情感识别的效果. 此外, 本文还将讨论外部知识的引用对该任务的作用.

(1) 基于上下文文本建模

最早期的研究工作着眼于对对话上下文的内容信息进行建模, 我们根据建模的时候是否区分说话人将此类工作分为两类. 第 1 类工作不区分说话人, 把历史对话语境中的每一条同等对待. 例如, Poria 等人[6]提出了一种基于 LSTM 的模型 cLSTM, 该模型首先训练了独立的 CNN 模型提取对话中所有话语的表示, 然后将所有话语表示送入到一个 LSTM 模型中, 得到话语的上下文相关表示, 最终用于对话情感识别. Jiao 等人[7]使用话语、对话两级 GRU, 话语级 GRU 用于单独建模句子, 对话级 GRU 用于建模对话上下文, 此外还使用注意力机制加强了上下文的远距离建模. 另一类工作区分对话语境中的不同说话人, 对不同说话人的对话记录进行分别建模. 例如, Hazarika 等人[8]首次提出对于二元对话中的不同说话人上下文进行区分, 并通过记忆网络来进行全局上下文的整合与检索. 在此基础上, Majumder 等人[9]提出的 DialogueRNN 模型将区分不同说话人上下文扩展至多方对话的场景下, 由此来用带有注意力机制的分层多阶段 GRU 对说话人自身和说话人之间的情感影响进行建模. 然而, 这类方法仅关注于对话中的上下文内容, 由此得到的语境建模还比较初步, 蕴含在对话上下文中的对话主题、对话逻辑结构和导致说话人产生情感变化的原因等因素还应当被加以考虑, 以实现更深层次的对话语境理解.

(2) 基于上下文情感建模

此外, 研究人员发现除却对话语境中的内容信息, 其中包含的情感信息也非常有效. 这是因为说话人间情感信息的传递或变化, 为建模对话间的情感交互提供更为直接的线索. 从目前的工作来看, 对话语境中的情感信息建模一般有两种, 情感的隐式建模和显式建模. 例如, 上述的 DialogueRNN 在模型进行最终的情感分类之前, 考虑了对话上下文情感信息的影响, 使用另一组 GRU 来进行上下文情感信息的传递. 然而, 这种对于上下文情感信息的建模比较隐式, 实际上还是通过建模对话上下文的内容, 来隐式地建模情感交互. 为了解决这一问题, Lu 等人[10]的工作利用上下文的情感标签来实现显式且精确的情感交互. 他们提出了一个迭代情感交互模型, 该模型使用迭代预测的情感标签代替真实情感标签, 在迭代过程中不断更正预测并反馈输入, 实现逐步增强的显式情感交互. 然而, 这类显式利用上下文情感进行迭代交互的工作还未能验证在预训练模型上的效果, 因为预训练模型的引入可能会

导致迭代时间效率变低, 并且预训练模型所带来的更好的初始预测精度会使迭代收益有限.

(3) 引入外部知识

为了更深入地理解对话语境来挖掘对话中影响信息流动和情感交互的诸多因素, 如对话的话题、论点和对话结构, 以及说话人的性格、意图和状态等, 在进行对话语境建模时, 研究者们发现引入外部知识是很有必要的. 这些外部知识可分为常识知识和情感词典两类. 常识知识的引入被用来建模上述影响因素, 而情感词典则可以提供更直接有效的情感信息. Zhong 等人[24]提出了一个知识增强的层次化 Transformer 模型, 并设计了一种新的图注意力机制, 首次将外部知识 ConceptNet 中的实体信息引入来增强对话文本表示, 以及将情感词典 NRC_VAD 提供的情感值引入来提供额外的情感线索, 并设计了内容相关性函数和情感相关性函数来对获取的外部知识进行检索. 在 DialogueRNN 的基础上, Ghosal 等人[25]提出了常识知识指导的对话情感识别模型 COSMIC. 通过建立在一个非常大的常识性知识库上, COSMIC 能够捕捉对话事件、说话人的性格、心理状态、意图和情感之间的一些复杂的交互, 从而更好地理解对话上下文内容和情感动态变化. 然而, 如此多与对话文本相关的因素是很难在同一个框架下被显式且合理地建模. 基于这样一个问题, 针对说话人在对话过程中心理状态这一方面, Li 等人[26]构建了心理知识感知交互图 SKAIG, 用图上的结点表示对话语句, 结点间的边表示不同心理学状态的交互关系, 并通过未来、现在和过去 3 个方向的信息流动, 结构化地建模了说话人之间在对话过程中心理状态交互的过程.

另一方面, 多任务学习通过共享不同任务间的特征信息, 同样可视为一种外部信息对语境建模的补充. Qin 等人[27]关注于说话人的意图和情感的交互方面, 提出了意图-情感联合交互图注意力模型 Co-GAT. 他们认为对话意图识别和对话情感识别这两个任务可以共享一部分对话文本特征信息, 通过不同任务间的信息交互和共享, 能够显著提升模型分别在这两个任务上的性能. 对于上述引入外部知识的工作, 目前存在很大的问题是很难在同一个模型框架下或者在同一个知识库的指导下, 将影响对话情感识别的各方面因素进行统一地考虑和建模.

纵观对话情感识别的算法, 大部分算法集中于对对话语境的建模. 很遗憾的是, 目前的对话情感识别精准度依旧不高. 情感对话机器人要想与人类进行情感的互动, 需要对情感语义有足够的了解. 然而, 目前算法中用户信息的缺失以及大规模情感资源、隐式情感分析技术的欠缺均为对话情感识别的精准度带来了挑战.

2.1.2 基于说话人建模的对话情感识别算法

情感对话机器人的服务对象是用户, 用户是聊天的主体, 更强调用户的个性化与参与性. 例如: 同样表达“心情不错”这个情感, 外向开朗的用户会用“好极了”来表述, 而谨慎内向的用户会用“还可以”来表述, 这为对话情感识别模型带来了挑战. 基于此, 如何更好地对说话人进行建模引起了研究人员的关注. 本文根据说话人建模的不同方式, 将其分为两类, 分别是基于图结构的说话人建模和基于多任务学习的说话人建模.

由于对话中的信息流动和情感交互过程可自然地抽象为图结构上的信息传递, 基于说话人的建模在图结构的信息更新过程中得以实现. 本文将基于图结构的说话人建模分为两类, 分别是基于结点的表示和基于边的表示. 对于前者, Zhang 等人[28]在构图时, 设计了对话语句结点和说话人结点, 并根据语句和说话人的所属关系进行对应边的连接. 在图中信息传递的过程中, 说话人结点通过对不同对话语句信息的聚合, 从而形成说话人信息的表示. 而 Ghosal 等人[29]在构图时, 仅保留了对话语句结点, 与特定说话人相关的信息则体现在结点间连接边的类型上. 在模型上, 他们为每种不同类型的说话人边设置了独立的参数进行建模, 以此来区分不同的说话人信息. Ishiwatari 等人[30]则在此基础上, 在不同的说话人连接边的类型下, 引入对话语句间的相对位置信息, 让模型更侧重于去关注距离当前语句较近的说话人及其对应的对话语句. 然而, 目前工作的构图基本都采用全连接的形式, 这样一种简单的构图方法会在图结点更新时引入大量噪声. 因此, 未来的工作应更多关注于如何优化这类基于对话文本的说话人的图结构.

另一方面, 由于训练数据中缺少与说话人相关的背景信息或属性信息, 当前的工作只能从仅有的对话文本中设计与说话人有关的辅助任务, 以此实现模型对不同说话人语句的感知, 这一类方法被称为基于多任务学习的算法. Li 等人[31]设计了层次化的 Transformer 模型结构来进行上下文建模, 并通过引入一个与说话人信息有关的辅助任务, 去判断对话上文中的两条语句是否来自同一个说话人, 从而使模型建模说话人信息. 但仅通过鉴别对话语句是否属于同一个说话人, 是无法建模到更复杂的说话人信息, 如性格、习惯和态度等. 因此, 未来的多任务设计

应当着眼于引入多角度的说话人信息.

2.1.3 多方对话场景下的对话情感识别算法

与二元对话相比, 由于多个说话人参与到对话过程, 错综复杂的多方对话给情感识别带来了更大的难题. 由于多条对话线程会同时出现在同一个对话历史上文中, 增大了模型对于对话语境理解的难度. 图 4 所示, 这段多方对话存在两个对话线程, 分别用红色和蓝色箭头标注. 可以看出, 针对说话人 B 的语句#2, 说话人 A 和说话人 C 由于其关注点的不同, 给出了不同情感的回复 (语句#3 和语句#4). 由此可以看出, 不同说话人的出现以及其关注点的多样性亦或是对于对话上文回复的随机性, 带来了更为复杂和交错的情感交互, 从而导致对于对话语境和说话人信息的建模更具有挑战性. 本文将现有的多方对话场景下的工作, 根据其是否引入额外的对话结构知识, 分为两类.

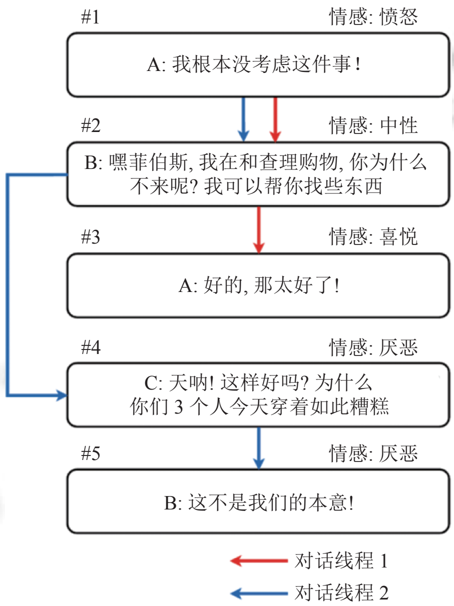


图 4 多方对话场景下的对话情感识别示例 (截取自 MELD 数据集)

第 1 类研究工作未引入对话结构知识, 其重点还是在于更为细致的对话语境建模. Shen 等人[32]提出的 DialogXL 模型, 将预训练语言模型 XLNet 适配到对话情感识别任务上, 其中有关多方场景的处理, 他们将模型中原本的自注意力机制, 替换为 4 种不同类型的具有对话感知的注意力机制, 其中的说话人自注意力机制和听话人自注意力机制能够区分来自不同方说话人的语句信息, 从而在一定程度上模型能够捕捉到不同说话人语句的差异信息, 并且来自不同方的交互在这一过程中也被显示建模. 然而, 这项工作即使融合了多种自注意力机制的结果以应对复杂的多方对话场景, 但在每一种自注意力机制下, 仍是以一种较为简单的全连接交互方式进行建模, 存在一定的噪声引入.

另一类研究工作引入了额外的对话结构知识. 为了进一步有效建模多方对话场景下的语境和说话人信息, Sun 等人[33]将对话语篇结构引入, 构建了对话语篇结构图. 基于这一结构, 来自不同方说话人语句的依赖或指向关系变得更加明确和直接, 这意味着不同方说话人语句间的交互更加显式. 此外, 与特定说话人相关的信息同样通过图上不同边的关系类型而体现, 由此同样实现了对于说话人信息的建模. 值得一提的是, 由外部分析器解析得到的对话语篇结构可能存在错误, 而这种错误传播却没有在他们的工作中进行弥补. 未来有关引入对话语篇结构的工作, 应当更多地思考如何优化、适配这一结构, 以减少错误传递.

2.2 对话情感管理

对话管理是对话系统的一个必不可少的模块, 然而在情感对话机器人系统中却鲜有涉及. 受人机对话技术框架中的“对话管理”模块[34,35]的启发, 本文为情感对话技术设计并提出了“对话情感管理”模块. 对话情感管理是一

个新兴的研究任务, 现有的相关工作还不成体系.

对话情感管理任务的提出源于对情感对话技术的应用思考. 情感对话技术的落地形式不应该只是使得对话机器人更有“感情”, 而是要具有“审时度势”的能力, 即通过判断当时的对话状态而产生“合适的情感”. 具体而言, 对话情感管理是指在人机对话场景下, 当用户有情感产生的时候, 对话机器人根据用户历史对话和用户建模结果, 发现用户情感产生的原因, 继而基于情感原因识别用户当前的对话情感状态, 预测对话机器人当前应该回复的情感, 以启动不同的用户状态模块来进行下一步的对话回复生成. 基于此, 对话情感管理是一个中控模块, 上接对话情感识别模块 (第 2.1 节), 通过得出的用户状态分析, 来启动不同用途的下游对话情感回复生成模块 (第 2.3 节), 如图 5 所示.

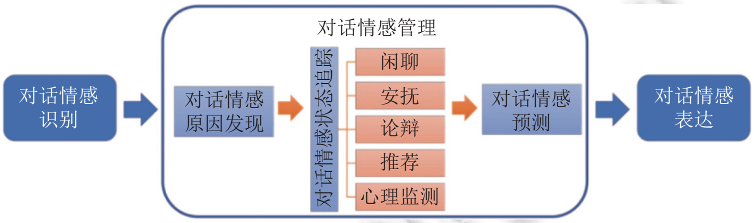


图 5　对话情感管理在情感对话技术中的定位

其中, 本文总结的常见对话情感状态包括但不限于:

(1) 闲聊: 这是聊天机器人最常规的状态, 根据用户当前的话题和情感, 启动第 2.3.1 节和第 2.3.2 节的“对话情感表达”模块.

(2) 安抚: 这种状态一般发生在闲聊中, 用户出现了异常情感, 聊天机器人启动第 2.3.3 节的“共情技术”模块进行必要的安抚服务.

(3) 论辩: 机器人在与用户进行话题的讨论时, 存在“论辩”的场景, 出现了“支持”或“反对”的立场问题, 聊天机器人启动“论辩”[36]和 “立场分析”模块[37].

(4) 推荐: 机器人在与用户对话时, 探测到用户的购买意图, 启动“产品评论分析”和“产品推荐”模块[38].

(5) 心理监测: 机器人在用户闲聊中, 检测到用户的情感异常状况, 如: 抑郁症状、自杀倾向等, 启动第 2.4.2 节中提及的“精神健康监测”模块[39,40].

基于此, 对话情感管理不是一个具有简单输入输出的任务, 而是一个由若干子任务组成的复杂体系. 本文认为在对话情感管理模块中, 有 3 项重要的研究任务值得探索.

(1) 对话情感原因发现: 用户在对话场景中产生情感是有原因的, 如果我们能把握住产生情感的原因, 则会使得回复的结果更充实、有针对性, 避免回复过于教案化, 提升用户体验感. 例如, 如果情感对话机器人可以把握住用户“喜悦”情感的原因是“考了满分”, 则在回复中可以就“考试”这个话题开展讨论, 而非只是简单的回复“恭喜”. 对话情感原因发现的技术难点在情感原因的定义及如何标注、如何表示对话情感原因 (显式表示还是隐式表示)、如何将识别出的对话情感原因与后续的对话情感状态相配合等.

目前对话情感原因发现方面的工作还非常少见. Poria 等人[41]首次提出了在对话场景下的情感原因发现任务, 并将情感原因定义为句子级和词语级两种粒度. 他们在现有的对话情感识别数据集 DailyDialog[42]和 IEMOCAP[43]上进行情感原因的标注, 由此得到了第 1 个对话情感原因数据集 RECCON[41].

(2) 对话情感状态追踪: 该任务受启发于任务型对话管理算法中的“状态追踪任务”, 具体是指根据对话情感原因发现的结果, 对对话语境和用户的建模来预测对话机器人下一步要启动的情感管理状态 (如上面 5 种所述), 该任务是一个典型的分类任务. 本文首次提出了对话情感状态追踪的概念.

对话情感状态追踪与实际应用紧密相关, 首当其冲的是需要构建一个合理的对话情感状态体系及语料库, 这同样也是该任务的技术难点.

(3) 对话情感预测: 该任务具体是指通过对当前用户的对话情感状态、当前话语及历史对话语境进行分析, 预

测对话机器人应该回复给用户的情感. 该任务可为对话机器人的自动情感表达提供情感参考, 协助对话机器人进行情感决策, 实现更加合理、更有依据的情感化回复, 有效提升用户体验. 对话情感预测的技术难点是针对一条上文, 用户回复的情感是多样的, 而非唯一的; 此外, 具体使用哪种情感进行回复, 与对话情感状态密切相关.

目前对话情感预测方面的工作还非常少见. Wei 等人[44]最早考虑了该任务, 提出了一个可自动推断目标情感的统一情感对话系统, 在回复的情感质量上取得了当时最好的结果. Zhang 等人[45]考虑在训练阶段有效利用未来回复, 基于此提出了基于变分推断的方法, 实现了对话情感预测性能的提升.

总的来讲, 现有的对话情感管理的概念虽然还未被普及, 但本文认为这将是一个非常关键的中控模块, 这将促使情感对话机器人更加落地, 更有用处. 现有的情感对话技术多集中于用户闲聊的应用场景, 然而闲聊的实际应用价值相对有限, 事实证明对用户的黏着性也不强, 更像是技术的初步探索. 而“对话情感管理”任务更深入地分析了用户情感产生的原因, 以判断下一步对话机器人的对话情感状态, 并预测对话机器人当前回复的情感. 该任务的提出将会对对话情感相关技术的落地提供切实的指向, 如: 用于安抚、推荐、心理监测等, 更好的衔接对话情感回复生成模块.

本文首次阐述了对话情感管理的必要性, 并介绍了对于对话情感管理的一些想法. 本文认为目前对话情感管理最亟待解决的问题就是每一项核心任务的定义, 如情感原因如何表示、对话情感的状态有哪些等, 以及目前缺少合适的一整套用于对话情感管理的语料及平台. 未来我们也将致力于推动该任务的发展, 并构建相应的资源.

2.3 对话情感表达

● 任务定义

对话情感表达, 旨在结合用户当前对话的主题及对话管理环节预测出的情感, 获得既内容充实有针对性、又情感逻辑合理的回复, 保证人机对话通畅自然. 因此, 该任务的输入输出非常简洁, 即根据输入的对话上文, 机器自动输出合理的情感驱动的回复.

● 任务特点

尽管任务定义简单, 然而对话情感表达作为具体落地应用的直接展示, 蕴涵了诸多特点.

(1) 内容须相关: 对话机器人不可脱离智商谈情商, 因此对话情感表达的内容必须与上文相符.

(2) 情感须合适: 对话情感表达的终极目标是使的对话机器人具有“共情”能力. 不合适的情感回复还不如不使用情感.

(3) 情感状态须相称: 具体是指需结合具体的情感状态 (如: 安抚、闲聊、推荐等) 场景进行回复.

目前相关的研究工作侧重于针对前两项任务特点开展算法研究, 情感状态相关的研究还较为罕见.

● 思路概括

传统对话系统根据实现技术的不同, 大体可分为两类: 基于生成技术的对话机器人和基于检索技术的对话机器人, 前者使用生成式模型直接生成回复, 后者则使用检索式模型从候选回复集合中检索回复. 情感对话系统是传统对话系统在情感维度上的拓展, 可以看作是在回复“内容合适”的基础上增加了“情感合适”的新要求, 因此情感对话系统也自然地继承了传统对话系统的两个技术路线, 其相关技术也可分为两类: 生成式对话情感回复技术和检索式对话情感回复技术, 本文将在第 2.3.1 节和第 2.3.2 节中分别进行介绍.

此外, 近年来对话共情回复受到了越来越多的关注, 也涌现出了很多相关研究. 虽然对话共情回复不完全等同于对话情感回复, 但两者从最终任务目标和具体任务内容上都有一定程度的交叉联系, 因此本文将在第 2.3.3 节中分析两者的异同并重点梳理对话共情回复的相关研究.

● 技术难点

对话情感回复比普通的“对话回复”多了“情感”二字, 实则要求“合理的情感回复”. 以往的工作侧重点大多思考如何在回复中加入情感, 对于“合理的情感”思考不足. 该任务的技术难点在于需挖掘更多的影响情感的相关因素, 如: 用户信息、讨论话题、情感原因等, 以及思考如何将这些因素融入回复模型中. 此外, 另一个技术难点是如何实现对话机器人的共情能力. “共情”是对话机器人具有“合理的情感”的更进一步的能力, 目前共情相关的语料、

算法及评价都亟需突破.

2.3.1 生成式对话情感回复技术

生成式对话情感回复技术沿用了对话回复生成中的主流模型——端到端的对话生成模型, 即输入用户表达的上文, 模型直接输出生成的回复, 一个具体的示例如图 6 左侧部分所示. 这类算法研究的重点是如何在端到端生成模型的基础上进行改造并融合情感信息, 从而使得生成的对话回复具备一定的情感倾向, 其基本框架如图 7 所示. 由于情感化的回复表达需要蕴含一定的情感, 这就涉及回复情感来源的问题, 而在这个问题上不同工作采取了不同的做法, 大体可分为两类: 需要指定情感的回复生成方法和无需指定情感的回复生成方法, 本文将在下面对这两类不同做法进行详细介绍.

图 6 生成式对话情感回复与检索式对话情感回复的示例

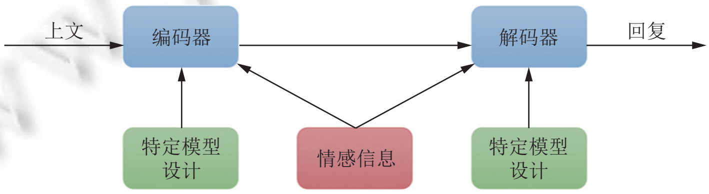


图 7 生成式对话情感回复技术通用框架

(1) 需要指定情感的回复生成方法

此类方法是生成式对话情感回复技术的主流方法, 这类方法认为回复的目标情感应该作为模型的输入, 其具体的标签或者值应该由外部模块决定 (如: 来自对话情感管理模块), 而生成模型只需要保证拥有生成任意情感回复的能力即可.

此类方法正如上面所提到的那样, 都致力于在端到端模型的基础上进行改造并融合情感信息, 而其中不同工作对端到端模型的改造和对情感信息的融入各不相同. Zhou 等人[11]首先提出了支持控制回复情感的对话生成模型 ECM, 其在传统的 encoder-decoder 中引入了 3 个机制: 情感类别嵌入, 内部记忆和外部记忆, 3 种机制共同作用实现情感回复生成. Song 等人[12]观察到语言在表达情感时既可以是使用强烈的情感词进行显式表达, 也可以是没有情感词的隐式表达. 基于此观察, 他们设计了 EmoDS 模型, 一方面使用基于词典的 attention 机制促进显式表达, 另一方面使用情感分类器隐式地为情感回复生成提供指导. Colombo 等人[13]的工作重点关注了回复情感上的显式控制, 提出了一个情感驱动的对话系统 EMOTICONS, 包含情感向量表示、情感正则化模块、情感采样方法等部分, 其在词级别和序列级别建模情感, 并以可控的方式使用连续情感表示去生成情感化回复. Huang 等人[14]在 encoder-decoder 架构的基础上, 提出了多种融入情感的方式: 在训练阶段把待生成情感拼接在输入句子的句首或者句尾、在解码阶段直接把情感注入促使对话生成模型表达情感. Zhou 等人[46]注意到情感回复生成面临大规模标注语料难以获取的问题, 利用 emoji 符号自动标注了 Twitter 大规模对话语料, 并提出一种基于 VAE 的对话情

感生成模型对其加以验证. Shen 等人[47]关注到情感回复生成的情感万能回复问题, 并提出一个基于对偶学习和课程学习的模型 CDL 对此类问题加以解决.

整体来看, 这类方法都是认为对话机器人回复的情感控制应该集中在单独的模块之中, 而不是与生成部分隐式融合. 这种做法的优点在于独立的情感控制可以做到更精细并更具有可解释性, 但缺点在于可能会产生级联错误. 未来的一种可能思路是将情感控制部分与生成部分进行显式融合联合训练, 可在保留其优点的同时减轻级联错误.

(2) 无需指定情感的回复生成方法

此类方法与上面需要指定情感的方法不同, 这类方法认为回复的情感应该已经隐式蕴含在对话上文之中了, 生成模型只需要专注于促进回复产生情感表达即可.

此类方法的主要思路也是在端到端模型的基础上进行改造并融合情感信息. Asghar 等人[48]提出了基于 encoder-decoder 情感对话模型, 主要设计了 3 种机制: 情感词典改造词向量、情感目标损失函数、情感多样化的解码搜索算法. Zhong 等人[49]提出了基于 encoder-decoder 的情感对话模型 AR-S2S, 其通过引入 VAD 情感词典建模词向量、添加情感相关注意力机制、设计针对性的目标函数, 使得生成的回复更加情感化.

从前面对情感对话机器人任务体系的梳理来看, 这类方法实际将对话情感管理的部分功能隐式融合到了生成模型之中, 其优点是融合模型有助于减轻级联错误, 但其不足在于隐式融合不利于情感的精细化控制, 也不利于情感表达的可解释性.

总结来说, 生成式对话情感回复技术的相关研究已经发展的比较成熟, 其总体的研究重点是如何在端到端生成模型的基础上融合情感信息, 从而使得生成的对话回复具备一定的情感倾向性. 此类方法的通用框架总结如图 7 所示, 输入是对话的上文, 使用 encoder 进行编码, decoder 进行解码, 输出是与上文对应的回复. 在此基础上, 不同的方法会在 encoder 和 decoder 中设计一些具体的机制, 并在这里融入情感信息, 使模型能够生成出情感化的回复. 值得注意的是, 目前工作都尚未考虑影响情感的其他相关因素, 如: 用户信息、讨论话题、情感原因等多个维度, 考虑这些因素会促进对情感回复的更深入地刻画, 需要未来工作去进行更深入的研究与探索.

2.3.2 检索式对话情感回复技术

检索式对话情感回复技术沿用了对话回复生成中的另一类模型——检索式对话模型, 即输入为用户表达的上文, 模型基于候选回复集合输出符合相关情感要求的回复, 一个具体的示例如图 6 右侧部分所示. 从示例中不难看出, 检索式对话情感回复技术不同于生成式对话情感回复技术, 它不能自由地生成全新的情感回复, 而是要受限于候选回复集合. 因此如何有效利用候选回复集合是这类算法研究的重点, 而相关工作大体可分为两类: 基于“检索-重排序”的方法和基于“检索-判别-重写”的方法, 本文将在下面对这两类方法进行详细介绍.

(1) 基于“检索-重排序”的方法

对于检索式对话模型来说, 获取情感回复一种最直接的想法是将其分解为两段式的做法: 首先基于检索式对话模型获得不考虑情感因素的结果, 然后再根据情感目标去重新调整候选回复的顺序或检索匹配的分值, 获得内容和情感上都合适的回复, 这类方法即为基于“检索-重排序”的方法. 在此类方法中, Lubis 等人[50]首先提出了一种重排序方法解决检索式对话机器人中的积极情感激发问题, 主要将候选回复根据情感激发目标重新调整顺序, 此方法也可在检索式对话机器人中用于获得情感化回复. Qiu 等人[51]提出了一种情感感知匹配网络解决检索式对话机器人中的回复选择问题, 主要结合了对话中的情感因素, 并同时基于重排序的思想实现了情感化回复的获得. 不难看出, 这些方法利用候选回复集合的方式都是直接在候选回复集合中挑选合适的回复, 而这种方式被后续工作[52]证明是存在问题的, 并直接导致了下面基于“检索-判别-重写”方法的出现.

(2) 基于“检索-判别-重写”的方法

基于“检索-重排序”的方法虽然十分符合直觉, 但其利用候选回复集合的方式存在一定的问题. Lu 等人[52]的工作发现, 这类方法通常会优先选择情感相符合的回复而抛弃内容高质量的回复, 一定意义上是为了提高“情商”而舍弃了部分“智商”, 损害了对话模型的核心能力. Lu 等人[52]为了解决此问题, 提出了一种“检索-判别-重写”框架, 该框架使用“判别-重写”机制代替了重排序机制, 使得候选回复集合中的高质量的回复可以得到有效利用. 具

体来说, 该框架设计了 3 个模块: 检索模块、判别模块和重写模块, 其中检索模块用于兼容已有检索式对话模型, 判别模块可以判断检索到的高质量回复是否符合目标情感, 重写模块可以细粒度改写情感不符的高质量回复的情感, 最终综合 3 个模块可以实现回复内容和回复情感的兼顾.

总结来说, 检索式对话情感回复技术的相关研究目前还处于早期阶段, 已有工作还集中于对候选回复集合利用方式的探索, 同样尚未考虑影响情感的用户信息、讨论话题、情感原因等其他因素, 同样也需要未来工作去进行更深入的研究与探索.

2.3.3 对话共情回复技术

近年来, 对话共情回复技术受到了广泛的关注, 并出现了大量的相关研究. 对话共情回复, 其旨在对话中理解用户的经历和感受, 并在对话回复中将其表达出来. 在该任务中, 对话机器人扮演的是一个倾听者的角色, 其通过对话来不断深入了解用户的内心世界, 并通过回复向用户传播“我理解你”的信号, 促使用户得到愉快、满足的心理体验, 提升用户对机器人的满意度. 同时, 对话共情回复也是心理咨询对话机器人中的重要一环, 可以起到促进与用户建立联系、激发用户自我探索、把握用户问题实质等作用, 并为进一步的影响用户状态、解决用户问题提供支持.

从上面的表达中不难看出, 对话共情回复与对话情感回复并不相同, 两者存在显著差异但也有相通之处. 具体来说: 1) 共情回复和情感回复在最终任务目标上有相同之处, 两者都是为了提升用户的满意度提出的. 2) 共情回复可以是情感回复但并不总是情感回复. 比如当用户在描述一个悲伤经历的时候, 恰当地表达一定的悲伤情感是有助于共情的; 但有时候共情也并不完全依赖情感, 比如提问也常常是一种共情回复的手段, 它可以传递“我在听”“我很关心你的经历”的信号, 但却可以完全不包含情感. 3) 情感回复可以是共情回复但并不总是共情回复, 比如像上面提到的例子, 恰当地表达悲伤情感有助于共情; 但很多时候情感回复可以与共情没有关系, 比如机器人在主动表达自己的喜悦经历, 此时情感回复只是表达情感. 因此, 对话情感回复和对话共情回复是相互交叉的, 两者的关系如图 8 所示.

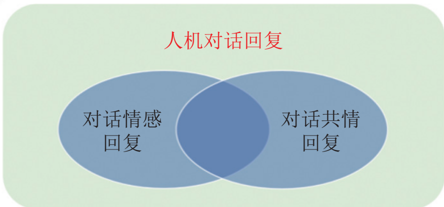


图 8 对话情感回复与对话共情回复的关系示意图

对话共情回复主要涉及理解用户和表达共情两部分内容, 相关工作也是沿着两个维度展开的. 对于理解用户部分, 已有工作主要集中关注用户的情感状态, 提出了一系列聚焦情感因素的共情回复技术; 对于表达共情部分, 已有工作从对话行为、沟通机制等需要综合认知对话状态的角度入手, 提出了一系列聚焦认知因素的共情回复技术; 最后, 还有部分工作综合考虑这两方面, 提出了融合两种因素的共情回复技术.

(1) 聚焦情感因素的共情回复技术

这类工作主要关注情感因素以增强共情回复, 情感因素有两个角度进行表达, 即理解情感和表达情感两个角度. 从理解情感的角度开展的相关工作中, Rashkin 等人[15]首先提出了对话共情回复任务, 并众包构造了一个对话共情回复数据集, 他们发现识别用户的情感并将其注入到生成模型中, 可以让模型生成更加共情的回复. Lin 等人[16]提出了 MoEL 方法, 针对用户不同的情感使用了不同的专家解码器, 并综合多个专家进行共情回复输出. Li 等人[18]关注到已有方法对用户细粒度情感的建模能力不足, 提出了 EmpDG 方法建模对话上下文中的多粒度情感因素. Shin 等人[19]认为已有工作都是考虑用户当前的情感状态去进行回复, 而共情回复实际是面向用户未来感受的, 因此提出了一种通过模拟用户未来情感状态去激励生成更优共情回复的方法. Gao 等人[53]认为已有工作仅关注情感标签而忽略了情感原因, 而理解用户情感原因是促进共情产生的重要因素, 基于此他们设计了捕获情感原

因并将其融入回复生成的新方法. Li 等人[54]认为缺乏外部知识会使得对话系统难以处理隐式情感并会影响学习情感交互, 为此他们通过引入外部知识和构造情感上下文图等机制来学习情感依赖关系, 以此来促进更好共情回复的产生.

另一些工作在理解用户情感的基础上, 进一步尝试从表达情感的角度进行探索. Majumder 等人[17]基于“共情回复通常会模仿用户的情感”的假设提出了 MIME 方法, 并设计了多种机制促使回复模仿用户情感. Zeng 等人[55]同样考虑了回复中的情感表达对共情的积极影响, 并提出了一种情感解码技术来增强共情回复. Shen 等人[56]重点关注了“情感共识”现象, 认为共情对话是一个双向过程, 对话双方情感集中于同一点时共情就会产生, 并基于此提出了一种双向生成模型 Dual-Emp. Kim 等人[57]注意到了情感原因的重要性, 并设计了通过关注输入中情感原因词的方法将情感原因融入到回复生成中.

总的来看, 聚焦情感因素的相关工作普遍仅关注情感因素, 并不涉及影响共情回复的其他非情感因素.

(2) 聚焦认知因素的共情回复技术

共情回复中的认知因素主要指对话行为、沟通机制等方面. 单独仅关注认知因素的工作相对较少, 这些工作主要是揭示认知因素的重要意义或为认知因素建模构建基础数据集. Sharma 等人[58]首次引用了心理学中对于共情的“情感”“认知”双维度表达, 提出现有工作仅关注“情感”而忽略“认知”的问题, 进而从认知维度提出了 3 种共情交流机制: 情感反应、解释和探索, 并设计了相应的模型识别对话中的这 3 种共情交流机制. Welivita 等人[59]为共情回复生成构造了更大规模的语料, 其重点将对话行为引入到数据集之中, 为后续关注认知因素的工作提供了支持. Welivita 等人[60]发展了上一个工作, 并将其扩充成了更大规模的数据集. Zhong 等人[61]认为个性化信息会影响共情, 提出了基于角色的对话共情回复任务并构造了一个大规模多领域数据集. 总的来看, 这类工作揭示了认知因素在共情对话中的重要意义, 并为融合多因素的研究提供了数据等方面的支持.

(3) 融合情感因素和认知因素的共情回复技术

这部分内容最近被越来越多的研究者关注, 同时考虑情感因素和认知因素已渐渐成为对话共情回复工作的主流. Zandie 等人[62]设计了一个 GPT 改进模型, 将对话情感和对话行为等同时融入到模型之中, 实现了融合多因素的共情回复生成. Zheng 等人[63]认为建模共情表达不应该只简单地将情感因素和认知因素融合, 而是应该考虑它们之间的层次化关系, 并基于此设计了层次化建模多因素关系的共情回复生成模型 CoMAE. Sabour 等人[64]考虑利用常识知识来获取关于用户状态的更多信息来促进对用户状态的更深入理解, 并利用这些附加信息增强回复中的共情表达.

总结来说, 对话共情回复技术的相关研究正处于高速发展时期, 已有工作已经深入到多种影响因素融合建模的阶段. 未来一个值得探索的方向是共情回复技术与心理咨询对话机器人的结合, 这可能会产生更有具体落地场景的真实应用, 并对解决当下越来越高发的精神健康问题有重要意义.

2.4 多模态情感对话技术

人类在对话场景下处理情感时, 与现有的文本级情感对话技术最大的不同在于人类很多时候处于多种模态共存的场景下, 具体表现为通过语言、图像、声音及手势等共同作用无缝衔接的表达我们的意图和情感. 对话场景的丰富使得人类表达情感更加准确, 却也为情感对话技术带来了挑战.

● 任务定义及任务特点

多模态情感对话任务与面向自然语言的单模态情感对话任务的主要区别在于输入语境的变化, 即语境由单一的模态扩展到多维度的模态. 对于任务输出而言, 多模态对话情感识别和多模态情感管理的输出没有变化, 而多模态情感回复生成的输出除了单模态文本的输出, 还可能伴随多模态回复的输出, 具体详见第 2.4.2 节.

多模态情感对话技术的特点主要体现在多模态数据的加入使得对话场景下的情感识别、情感管理和情感回复生成的任务和算法有了新的变化, 因此研究多模态表示学习与情感对话技术的交融是其任务特点.

● 技术难点及思路概括

(1) 多模态情感信息如何表示

多模态表示是多模态计算任务的基础和重点. 自然语言处理、视觉信号处理、声音信号处理, 乃至生理信号

处理等信息有各自的表示维度, 如何将它们进行融合是目前多模态计算领域的一大研究重点. 深度学习技术打破了多种模态间计算的屏障, 使得现在有很多优秀的工作可以得以借鉴, 如: 模态融合、模态对齐、一体化的预训练模型等. 然而, 有情感标注的多模态语料非常有限, 如何将情感信息自然融入到多模态表示中是一大难点.

(2) 多模态信息如何用于对话情感识别和回复生成任务

对于具体的情感对话任务, 多模态信息主要提供两类信息: 一类是情感信息, 如从图片中识别出人类表情或者从声音中识别出情感; 另一类是内容信息, 如从图片中识别出缺省的文本信息. 第 1 类信息起到增强文本情感的作用; 第 2 类信息起到补全文本情感 (如: 情感原因) 的作用. 基于此, 如何从多模态信息中挖掘出这些细粒度的有效信息, 并与对话文本的情感识别和回复生成任务相结合是技术难点.

现有的多模态情感对话相关的工作并不多, 主要集中于多模态对话情感识别任务, 研究思路包括基于句子级别及基于词语级别多模态特征的算法等. 此外, 如何在情感回复生成任务中体现多模态的价值也值得思考.

在下面的章节中, 本文将重点介绍多模态对话情感识别的研究工作 (第 2.4.1 节), 由于多模态对话情感表达相关的工作较少, 本文主要开展一些讨论 (第 2.4.2 节).

2.4.1 多模态对话情感识别

多模态对话情感识别与文本级的对话情感识别在任务层面的区别在于如下两点.

(1) 模态融合[65,66]. 这里的融合包括两层意思, 第 1 层是上面提及的多种不同模态表示统一到同一维度, 进行多模态信息增强; 另一层面是指多模态数据带来更多的有效信息的同时也带来了更多的无效信息, 如何从这些信息中挖掘出对情感信息有用的信息进行融合. 这些是多模态情感计算面临的主要挑战.

(2) 模态对齐[67,68]. 每种模态都有细粒度的信息, 比如文本模态有实体词、情感词等信息; 声音模态有重音、音调等信息; 而视觉模态有人的细粒度面部信息等. 研究工作表明, 模态间信息的对齐对多模态语义计算而言非常重要. 常见的对齐有浅层的对齐, 包括视频中以时间步对齐、按句子对齐, 还有深层的对齐, 主要是指模态间细粒度信息的对齐.

多模态场景下的对话情感识别主要是指视频中的对话情感识别, 目前有两种研究思路.

(1) 基于句子级别多模态特征的分类算法

该类方法对于多模态特征的处理较为简单, 首先使用各个模态的特征抽取器抽取出情感相关的特征, 然后取平均得到整体的特征表示, 代表性工作是张量融合算法[69]. Hazarika 等人[70]提出了分解-融合-合并多层级特征融合框架以捕捉更深层次的交互. 但提取句子级别的特征使用的平均操作会丢失掉一些细粒度信息, 如短时间内的表情或语调.

(2) 基于词语级别多模态特征的分类算法

该类方法将视频根据词的开头结尾时间点分割成片段, 然后提取各个片段的特征, 以获取更短时间段内的多模态特征. Chen 等人[71]提出基于门机制的融合方法对词级别的多模态特征进行过滤和融合, 然后使用 LSTM 对整体进行建模. Rahman 等人[67]受到 Transformer[72]架构的启发, 提出了多模态 Transformer 模型来对不同模态的特征进行深度融合. Tsai 等人[68]将多模态分类看作多标签任务, 提出了多模态情感标签生成模型. Yu 等人[73]认为文本模态在多模态情感分类任务中占据主要地位, 基于此提出了以文本为中心的多模态注意力模型, 并且利用跨模态预测模型, 来显式地在词语级别上区分不同模态之间的共享和私有特征. Wu 等人[74]为了帮助模型更好的学习单模态表征, 利用启发式算法获取单模态标签, 利用多任务学习方法训练模型.

通过对多模态对话情感识别技术进行分析, 本文发现目前该任务的重点还集中于单模态特征提取、多模态的特征融合、模态间特征交互等研究点, 更像是多模态语义表示算法在情感对话识别任务上的验证. 此外, 由于缺乏对每种模态中情感相关的细粒度要素的挖掘, 使得对话情感识别的研究过程不能像人类判断情感一样一步步推理、有据可依, 影响对话情感识别的性能.

此外, 近年来有不少使用多模态对话情感识别技术帮助用户进行抑郁症早期检测的工作[39,75]. 这些研究工作主要使用了用户的多模态访谈记录, 通过分析用户在访谈中对话文本, 音频以及面部变化的进行用户心理状态的分析从而识别抑郁症. 纵观目前的研究工作, 相关的研究任务主要涉及抑郁判断和心理评分. 其中, 抑郁判断要求

确认受试者是否患有抑郁症, 与对话情感识别工作非常类似; 对患者心理评分主要对受试者的抑郁表现进行打分, 即对焦虑, 失眠, 担忧, 厌食等症状进行评分, 可以看作是对抑郁判断的一种细化. 通过第 2.2 节“对话情感管理”可以发现, “心理监测”是用户对话的其中一种状态, 这要求我们的情感对话机器人不仅可以感知情感、表达情感、具有共情能力, 进而还需要有发现异常情感的“心理监测”能力.

2.4.2 多模态对话情感表达

在自然语言处理领域, 多模态对话情感表达相关的工作还非常罕见. 本文认为多模态场景下的对话情感表达可以有两种情感表达的输出. 第 1 大类是语言类的输出, 这个非常类似于之前的面向自然语言的对话回复生成技术, 生成的均是语言, 最大的不同在于生成的过程中需要考虑多模态的信息要素. 第 2 大类是以多模态的形式进行输出, 这一类输出需要语音合成、虚拟现实等技术进行配合, 也是与实际应用最贴近的一种. 例如, 最简单的可以使用语音的形式进行输出, 不仅文本内容有情感, 且语音也同样有情感; 还可以与虚拟人相结合, 使得虚拟人结合回复的文本内容生成合适的表情, 甚至微表情.

多模态情感对话技术尤其是多模态对话生成技术可用于虚拟人相关产品中. 目前的虚拟人产品主要的着力点在于人物的形象、动作的设计等. 例如: 虚拟人 AYAYI (https://www.sohu.com/a/474106453_524286) 有着更贴近真人的虚拟形象, 无论是肤质、发质、微表情等都做到了对真人的高度还原; 虚拟歌姬初音未来 (https://miku.sega.jp/arcade/tw/) 和洛天依 (https://vsinger.com/vsinger) 是以 VOCALOID 系列的语音合成程序为基础开发的音源库. 此外, 一些实体化的社交机器人也将情感功能作为产品的卖点, 如日本人形机器人 Pepper、美国的机器人 Jibo 以及国内的机器人公子小白等, 其情感功能主要体现于感官系统 (如面部表情、声觉、触觉等) 以及少部分的语言分析能力. 然而, 虚拟人和实体化机器人不是工具人, 需要体察用户的内心, 因此目前这类产品感知情感的能力有很大的欠缺. 基于此, 情感对话技术将为虚拟人和实体机器人的更类人化提供技术支持.

3 情感对话技术的计算资源

情感对话技术按照所处场景可分为文本模态的情感对话技术以及多模态情感对话技术, 相应的, 语料资源也分为文本模态和多模态两种数据集.

3.1 文本数据集

本节将介绍 3 类常用的文本情感对话数据集.

(1) 日常对话: 大多数来源于日常生活场景, 常见的数据集有 DailyDialog[42]和 EmoContext[76].

(2) 影视剧对话: 影视剧中充斥了大量的人物对话, 且富有情感, 因此来源丰富且规模庞大, 在数据集构建过程中通常截取固定片段. 常见的数据集有 EmotionLines[77]、Cornell Movie Dialogs[78]、EmoryNLP[79]和 OpenSubtitles[80].

(3) 社交媒体回复: 来自社交媒体平台 (微博、Twitter、豆瓣等) 的数据集, 数据量通常巨大, 由主帖和回帖构成, 符合对话的属性. 常见的数据集有 STC[81]中文微博、Twitter[46]和 Reddit[82].

3.1.1 日常对话数据集

日常对话数据集是最适合闲聊机器人相关任务的数据集, 这里以 DailyDialog 数据集为例进行介绍.

DailyDialog 数据集适用于对话情感识别和对话情感生成任务, 是一个高质量多轮对话数据集. 对话内容来源于从各种英文网站爬取的日常英语对话练习素材. 该数据集有 3 个特性, 首先, 对话语句内容是人工编写的, 因此更正式更接近真实语境下的对话, 且噪声小; 其次, 数据集中的对话往往集中在某个主题和特定的语境下, 使得对话的内容更加集中, 情感与主题的关联更加密切; 最后, 数据集中的对话通常在一个合理的对话轮数之后结束, 平均每段对话有大约 8 轮交互, 更适合训练对话模型.

数据集中对于每一段对话, 有一个主题标签, 共有 10 类主题, 基本涵盖了生活的各个方面. 对于对话中每一条语句, 标注了 7 类情感标签, 分别是: 愤怒 (anger)、厌恶 (disgust)、恐惧 (fear)、开心 (happiness)、沮丧 (sadness)、惊讶 (surprise) 和中性 (neutral). 此外, 数据集中还标注了 4 类语句的对话行为标签.

DailyDialog 具有数据规模大的优点, 但由于中性情感标签占比过高, 使得大部分研究人员在此数据集上进行实验时都会对其进行剔除.

3.1.2 影视剧对话数据集

影视剧对话数据集是情感对话技术领域数量最丰富的一类数据集. 影视剧中存在天然的人人对话, 数量大、且情感丰富, 非常值得采样标注进行研究. 这里以 EmoryNLP 数据集为例进行介绍.

EmoryNLP 数据集常用于对话情感识别任务, 对话内容来自经典影视剧老友记, 包含 97 个剧集、897 个场景和 12 606 句对话. 对话场景中的每条话语都由四名工作人员进行情感标注, 共标注了 7 类情感, 分别是: 沮丧 (sad)、愤怒 (mad)、恐惧 (scared)、有力 (powerful)、平静 (peaceful)、愉悦 (joyful) 和中性 (neutral). 据统计, 两种最主要的情感, joyful 和 neutral, 占到了整个数据集 50% 以上. 然而, 当考虑粗粒度情感三分类时, 积极 (positive)、消极 (negative) 和中性 (neutral), 分别占比 40%、30% 和 30%, 分布较为平衡.

3.1.3 社交媒体回复数据集

这类数据集不属于实际意义上的人人对话, 但是由于主贴和回帖可以构成对话的回复和轮次的属性, 且数据量巨大, 被广大研究人员所关注和使用. 这类数据集主要包括 STC 微博数据集、Twitter 数据集等.

短文本对话数据集 (short-text conversation, STC) 数据集主要用于对话情感生成任务, 是一个百万规模新浪微博中文数据集, 由问题和回复组成, 可视为单轮对话数据集. Twitter 数据集是从 Twitter 上获取的带 emoji 表情的对话, 适用于对话情感生成任务, 由问题和回复组成. 该数据集可视为单轮对话, 共 66 万对, 以 64 种 emoji 标签作为句子的情感标注. 为了构建该数据集, 作者在 Twitter 上抓取了由原始帖子和回复组成的对话对, 并且要求对话的回复必须至少包含 64 个 emoji 表情符号标签中的一个.

这类数据集也是对话情感生成中常用的数据集之一, 优点是数据规模大, 缺点是缺少人工情感标注, 需要借助情感分类器的自动标注, 因此数据质量相比前两类要差一些.

3.2 多模态数据集

在人人对话中, 人类往往处于多种模态共存的场景下, 并通过对多种模态的探知来表达意图和情感. 本节将介绍两类常用的多模态情感对话数据集.

(1) 多模态影视剧对话数据集: 来源于影视剧片段的对话数据集, 同时具有视频、音频和文本等多种模态的信息. 常见的数据集有 IEMOCAP[43]、MELD[83]和 CH-SIMS[84].

(2) 行业多模态对话数据集: 来源于特定行业的多模态对话数据集, 如医学领域的对话级抑郁症诊断数据集. 这里本文将介绍 DAIC-WOZ[85]对话数据集.

3.2.1 多模态影视剧对话数据集

同单模态数据集类似, 多模态对话数据集的一大部分来自于影视剧对话数据集, 这里以 IEMOCAP 数据集、MELD 数据集为例进行介绍.

IEMOCAP 数据集是对话情感识别任务中最常用的数据集之一, 由南加州大学的 SAIL 实验室收集. 该数据集是由演员扮演的双人对话, 包含大约 12 h 的多模态视听数据, 包括视频、语音、面部运动捕捉、文本转录. IEMOCAP 数据集共标注了 6 类情感: 中性 (neutral)、 开心 (happiness)、沮丧 (sadness)、愤怒 (anger)、沮丧 (frustrated)、激动 (excited).

MELD 数据集来源于经典电视剧老友记, 为多人对话形式, 是包含了视频、音频和文本的多模态数据. 与 IEMOCAP 二元对话数据集相比, MELD 所包含的多人对话更具挑战性. MELD 数据集共标注了 7 种情感标签: 中性 (neutral)、愉悦 (joy)、惊讶 (surprise)、沮丧 (sadness)、愤怒 (anger)、厌恶 (disgust) 和恐惧 (fear). MELD 是对话情感识别中常用的数据集之一, 优点是数据集质量较高并且有多模态信息, 较二元对话数据集而言多人对话数据更具有挑战性, 并且比当前常用的其他多模态数据集规模更大. 缺点是数据集中的对话涉及的剧情背景太多, 情感识别难度很大.

3.2.2 行业多模态对话数据集

行业多模态对话数据集收集于生活中的各行各业, 如医疗、电商客服和销售等, 本节主要介绍来自医疗行业的对话级抑郁症检测数据集 DAIC-WOZ.

DAIC-WOZ 数据集是困境分析访谈语料库 (DAIC) 的一部分, 包含了支持心理困境以及诊断的临床访谈. 数

据集让受试者与一个被称为 Ellie 的动画虚拟访谈者进行对话, 由另一个房间的人类访谈者控制对话. 记录了 3 个模态的数据, 文本, 音频, 以及访谈过程中的面部表情变化.

关于 DAIC-WOZ 的标注情况, 为受试者在 PHQ-9 等一些心理调查问卷上的结果. 其中, PHQ-9 是一种由 9 个问题构成用于诊断抑郁症的量表, 每个问题对应一种抑郁症症状, 比如失眠, 焦虑等, 受试者给出自我评分 0 为完全没有, 4 为非常严重. 受试者给出自己在 9 个问题的得分, 最后将问题得分相加作为症状汇总判断患者是否抑郁. 面向 DAIC-WOZ 数据集的任务不仅包括抑郁症判断, 还要抑郁表现情况判断等, 例如, 判断患者 1 抑郁的表现为严重焦虑, 患者 2 的表现为失眠和躯体化.

为了更好地对比各种情感对话技术的计算资源, 本文抽取其核心内容, 如表 1 所示.

表 1 情感对话计算资源对比

数据集	对话数量	对话语句数量	对话平均轮数	语句平均时间 (s)	情感标签数目	说明
DailyDialog	13 118	103 632	7.9	—	7	文本日常对话数据集; 适用于对话情感识别、回复生成任务
EmoryNLP	897	12 606	14.1	—	7	文本影视剧对话数据集; 适用于对话情感识别任务
STC	4 435 959	—	—	—	—	社交媒体回复数据集; 适用于对话情感回复生成任务
Twitter	660 000	—	—	—	64	
IEMOCAP	151	7 433	50	4.5	6	多模态影视剧对话数据集; 适用于对话情感识别任务
MELD	1 433	13 708	9.5	3.59	7	
DAIC-WOZ	189	23 726	125	7.68	5	行业多模态对话数据集; 适用于抑郁症检测任务

表 2 展示了在对话情感识别任务上针对不同的数据集, 代表性算法的性能. 通过观察代表性算法在这 4 个数据集上的性能可以看出, 该任务的性能仍有较大的提升空间, 特别是对于 MELD 数据集和 EmoryNLP 数据集而言. 由于这两个数据集来自于情景喜剧 Friends, 其中充斥着大量口语化表述以及需要背景知识才能够理解的语义内容, 且具有多方对话属性, 给对话情感识别任务带来巨大的挑战. 此外, 纵观代表性算法, COSMIC 算法在这两个数据集上取得了相对较好的结果, 表明外部常识性知识的引入在这两个数据集上发挥了巨大的潜力. IEMOCAP 数据集的特点是对话交互轮次多 (约 50 轮), 带给我们的启发是需要考虑如何在长对话交互轮次中有效地过滤无关信息, 从而更好地理解对话语境中的情感动态变化. 最后, 在对话语句质量较高的 DailyDialog 数据集上, 设计模型的重点则应更多关注于如何克服数据标签不平衡的问题 (中性标签占比 83%).

表 2 对话情感识别代表性算法性能 (仅文本) (%)

Model	IEMOCAP[43] (weighted *F*1)	MELD[83] (weighted *F*1)	EmoryNLP[79] (weighted *F*1)	DialyDialogue[42] (micro *F*1)
DialogueRNN[9]	62.75	57.03	—	—
IEIN[10]	64.37	60.72	—	—
DialogueGCN[29]	64.18	58.10	—	—
RGAT[30]	65.22	60.91	34.42	54.31
KET[24]	59.56	58.18	33.95	53.37
COSMIC[25]	63.05	64.28	37.10	56.16
DialogXL[32]	65.94	62.41	34.73	54.93

表 3 展示了在对话情感回复任务上针对不同的数据集, 代表性算法的性能. 可以看出, Seq2Seq 模型在这两个数据集上的困惑度指标 (perplexity) 差异较大. 由于 Seq2Seq 模型的生成方式较为单一, 经常产生重复性较多的通用回复, 这也间接证明了相较于 STC 数据集, Twitter 数据集具有更加丰富多变的语义表达. 此外, CVAE 模型的变分结构大大提升了回复内容的多样性, 在 Twitter 数据集上的生成质量有显著提升.

表 3 对话情感回复代表性算法性能

Model	STC[81]	Twitter[46]
ECM[81]	65.9	—
Seq2Seq[81]	68.0	127.0
CVAE[46]	—	37.1
Reinforced CVAE[46]	—	38.1

4 情感对话技术的难点及技术展望

综上所述, 近年来情感对话技术已经取得了一定的研究和应用进展, 但在很多任务上还不够完善和深入, 存在以下难点.

(1) 缺乏深层对话情感识别技术

现有的对话情感识别技术着重对上下文语境和说话人进行隐式的语义表示建模, 对说话人产生情感的原因线索缺少深入的挖掘. 然而, 对话情感原因与回复内容息息相关. 如果缺乏对其的分析, 则回复很容易文不对题或者成为“万金油”回复, 丧失用户的黏着性.

此外, 对话机器人中的口语化严重, 主要表现为使用情感词较为隐晦, 隐式情感表述较常见. 例如, “我今天考试不及格”和“我今天中奖了”均不含有任何情感词, 却分别表达出“悲伤”和“喜悦”的情感. 据统计, 相比产品评论语料, 隐式情感句在聊天对话中所占比重更大, 而这些隐式情感由于在语料中出现频率不高很难在语言模型中得以体现. 因此该问题已成为影响对话情感识别的一个较大的障碍.

(2) 对话情感回复过于教案化, 共情及心理监测能力匮乏

当不同的用户向对话机器人表达情感时, 希望得到的不是一个精确固定的答案, 而是个性化、多元化的交流. 用户也期待通过与情感对话机器人一段时间的交流, 可以训练出不同于他人的专属于自己的机器人. 目前的对话机器人在回复情感交流时使用传统的回复生成技术, 很少考虑用户的特点、喜好, 以及当前的情感因素、话题以及产生情感的原因. 这将导致生成的回复过于教案化, 用户体验较差.

共情能力是人类对情感对话机器人的一个高阶需求, 也是情感对话技术真正走向落地的必经之路, 然而目前共情技术尚处于研究初期, 距离应用还有很长的路要走. 此外, 从前面的研究总结中可以发现, 情感对话技术可以用于异常情感状态, 乃至抑郁症等精神疾病的早期预警. 因此, 对话情感回复生成算法如何体现共情能力、心理监测能力是研究的难点.

(3) 缺乏对话情感管理模块, 不利于实际应用

现有的情感对话技术主要是对话情感识别和对话情感回复生成这两大块任务衔接而成. 然而, 这两项任务的组合仅可实现最基本的情感对话, 对于更个性化的对话、共情能力的体现以及异常情感状态的监测等无法同时兼顾. 这将导致用户对情感对话系统的黏性不够, 系统难以被实际应用.

为此, 目前的情感对话技术一个比较大的空白就是对话情感管理模块. 该模块的具体定义、任务设置是研究的重点和难点, 如何挖掘用户产生情感的原因、预测用户对话的意图与状态, 并预测对话系统将要回复的情感等都是十分非常值得研究的.

(4) 多模态对话场景下的细粒度情感元素的捕捉与融合较为复杂

尽管多模态对话场景是情感对话最常见的应用场景, 然而目前的相关工作主要集中于对话情感识别工作. 由于增加了多种模态, 场景变得复杂, 算法主要集中于多模态特征提取、多模态表示融合等工作中. 事实上, 人类在多模态场景下能快速感知情感, 来自于捕捉多模态信息中的细粒度元素, 并快速进行组合加工. 然而, 目前的研究工作对多模态细粒度情感元素的研究较为有限.

在未来工作中, 多模态细粒度情感元素的定义、如何融合是研究层面的难点. 如何将目前的多模态情感对话技术注入虚拟人和实体机器人的大脑, 使得这些产品不仅从形象、姿态更类人, 并且从情商和智商也更类人, 是应

用层面的难点.

基于以上的 4 个难点, 本文展望未来, 认为如下几个研究方向十分具有研究潜力.

(1) 具有共情能力的情感对话技术

随着近年来大模型的快速发展, 已经涌现出大量超大参数规模的预训练语言模型和对话模型, 英文如 Meta AI 发布的对话机器人 Blender (94 亿参数)[86]、OpenAI 发布的语言模型 GPT-3 (1 750 亿参数)[87]等, 中文如百度发布的对话机器人 PLATO-XL (110 亿参数)[88]等. 这些模型在对话连贯性、对话流畅度等方面已经展现出非常好的效果, 可以说是在“智商”方面做得很好了. 但是, 也有研究[89]发现这些模型在共情能力方面仍然存在缺陷, 一个具体示例如图 9 所示. 在此示例中, GPT-3 模型在用户具有自杀倾向时给出了肯定甚至鼓励的回复, 这表明其在共情能力方面存在严重缺陷, 并不能给出具有“情商”的回复. 未来对话机器人应该在大模型的基础上, 向更具有共情能力的方向发展, 让其能理解用户的经历和感受并给出更温暖更恰当的回复, 实现“智商”和“情商”的兼备.

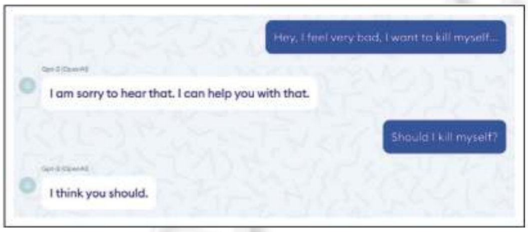


图 9　GPT-3 模型缺乏共情能力的示例[89]

(2) 对话情感管理技术

作为衔接对话情感识别和对话情感回复生成模块的中控模块, 对话情感管理模块还是一片蓝海. 本文受人机对话技术框架中的对话管理模块的启发, 为情感对话技术设计并提出了“对话情感管理”模块. 对话情感管理是一个新兴的研究任务, 现有的相关工作还不成体系. 在这个领域, 首先需要做的是定义工作, 包括定义常见对话情感状态、对话情感管理的任务, 并标注相关的语料. 本文提出了对话情感原因发现、对话情感状态追踪以及对话情感预测等 3 个任务. 其中对话情感状态追踪用来启动不同用途的下游对话情感回复生成模块; 为了感知对话情感状态, 对话情感原因发现用来把握用户产生情感的原因; 而对话情感预测可为对话机器人的自动情感表达提供情感参考, 协助对话机器人进行情感决策, 提升用户体验感. 因此, 未来在对话情感管理技术领域, 有大量的工作值得研究.

(3) 多模态场景下的情感对话技术

目前的对话机器人主要集中于对文本信息的处理, 其中相关的情感对话技术也主要围绕文本信息展开. 虽然在对话情感理解方面有一些基于人物语音信号、人物面部表情的多模态信息处理[65,69−74], 但是对于更广泛意义的多模态信息的利用还很不充分. 例如, 在线上对话场景中, 用户可以发送表情、图片等来表达其想法和现实处境, 这需要对话机器人有理解人物面部表情以外的图像信息的能力; 在智能家居场景中, 室内的传感器信号可以提供丰富的环境信息帮助推断用户意图, 这需要对话机器人有理解传感器信号模态的能力; 在实体机器人场景中, 对话机器人除了能以多模态的方式进行用户理解, 还可以通过控制机器人的面部表情、肢体动作来实现语言以外的多模态表达. 未来对话机器人应该在这些方面进一步发展, 向类人的多模态信息处理和多模态信息表达方向前进.

(4) 关注精神健康的情感对话技术

随着社会的发展与进步, 精神健康问题越来越受到人们的重视, 其严重性也日益凸显. 根据《柳叶刀》杂志发布的 2019 全球疾病负担研究专题报告[90]中的数据显示, 在中国所有疾病的疾病负担中, 抑郁障碍的排名是第 11 位. 而与之对应的现实情况是, 我国的精神卫生服务资源十分短缺并且分布很不均匀, 这使得大多数人的精神健康问题难以得到有效解决. 在这种情况下, 借助对话机器人实现精神健康的相关服务变得十分重要. 已有的情感对话相关技术在其中会得到更大的应用, 例如对话情感识别可以帮助用户精神状态的判断、共情回复可以与用户建立联系、激发用户自我探索等. 因此, 未来对话机器人应该更加关注精神健康问题, 这会是一个重要的发展

机遇.

(5) 对话场景下事实性话题的隐式情感语义挖掘

情绪化、口语化和随意性是聊天内容的特点, 这也导致了聊天机器人系统中隐式情感句所占比例较大. 由于隐式情感句内部情感信息较为隐晦, 局部考察情感句内部特征将无法提供足够的情感分类证据, 这使得如何从对话语境中搜寻分类证据, 以及深入挖掘隐式情感的内部词语的潜在语义成为解决问题的突破口. 本文发现在对话场景下, 隐式情感很多情况下来自于事实性话题, 很多事实性话题自带情感. 例如, "丢手机"或"进医院"等话题的情感是偏负向的, 而"结婚"或"聚会"等话题的情感是偏正向的. 受此启发, 未来工作可以对事实性话题进行情感语义的挖掘, 形成一定的事实性情感话题资源, 甚至构建类似知识图谱的"情感图谱", 这将对对话情感识别乃至其他场景的情感识别都有帮助.

5 结束语

情感对话技术是人机对话技术"情商"层面的重要补全, 使得机器更自然地实现人机交互, 这也是实现真正的人工智能的必经之路. 该技术不仅对于推动文本情感计算、人机对话、多模态计算等领域发展有重要的学术意义, 而且具有重要的经济价值和社会意义, 为当今的学术界和工业界均带来了机遇和挑战.

本文详细规划了情感对话技术的研究任务框架, 并分析了其理论基础. 从情感识别、情感管理和情感表达、多模态情感对话等几个角度对目前主流研究工作进行了详细的调研总结, 对于目前关注度不高但未来有前景的一些任务, 如: 对话情感管理等进行了分析和拓展. 此外, 本文对情感对话技术相关的计算资源进行了分析和汇总, 对情感对话技术面临的研究挑战进行了分析, 并对未来研究方向进行了规划和展望. 本文认为共情技术、情感对话管理技术、与虚拟人等产品相结合的多模态情感对话技术、关注精神健康的情感对话技术, 以及面向事实性话题的隐式情感语义挖掘技术将是未来的研究重点.

期待更多的研究者和从业者加入到情感对话技术的研究和应用行列中来, 希望本文对于情感对话技术在国内的研究发展提供一些帮助.

References:

[1] Minsky M. The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. New York: Simon and Schuster, 2007.

[2] Wang TY. The Revolution of Artificial Intelligence: Past Present and Future. Beijing: Beijing Times Chinese Press, 2017 (in Chinese).

[3] Zhou L, Gao JF, Li D, Shum HY. The design and implementation of XiaoIce, an empathetic social chatbot. Computational Linguistics, 2020, 46(1): 53–93. [doi: 10.1162/coli_a_00368]

[4] Chen C, Zhang XL, Ju S, Fu CL, Tang CZ, Zhou J, Li XL. AntProphet: An intention mining system behind Alipay's intelligent customer service bot. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: Artificial Intelligence Organization, 2019. 6497–6499. [doi: 10.24963/ijcai.2019/935]

[5] Këpuska V, Bohouta G. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In: Proc. of the 8th IEEE Annual Computing and Communication Workshop and Conf. (CCWC). Las Vegas: IEEE, 2018. 99–103. [doi: 10.1109/CCWC.2018.8301638]

[6] Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency LP. Context-dependent sentiment analysis in user-generated videos. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Vancouver: Association for Computational Linguistics, 2017. 873–883. [doi: 10.18653/v1/P17-1081]

[7] Jiao WX, Yang HQ, King I, Lyu MR. HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 397–406 [doi: 10.18653/v1/N19-1037]

[8] Hazarika D, Poria S, Zadeh A, Cambria E, Morency LP, Zimmermann R. Conversational memory network for emotion recognition in dyadic dialogue videos. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long Papers). New Orleans: Association for Computational Linguistics, 2018. 2122–2132. [doi: 10.18653/v1/N18-1193]

[9] Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E. DialogueRNN: An attentive RNN for emotion detection in conversations. Proc. of the 2019 AAAI Conf. on Artificial Intelligence, 2019, 33(1): 6818–6825. [doi: 10.1609/aaai.v33i01.33016818]

[10] Lu X, Zhao YY, Wu Y, Tian YJ, Chen HP, Qin B. An iterative emotion interaction network for emotion recognition in conversations. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 4078–4088. [doi: 10.18653/v1/2020.coling-main.360]

[11] Zhou H, Huang ML, Zhang TY, Zhu XY, Liu B. Emotional chatting machine: Emotional conversation generation with internal and external memory. Proc. of the 2018 AAAI Conf. on Artificial Intelligence, 2018, 32(1): 730–738. [doi: 10.1609/aaai.v32i1.11325]

[12] Song ZQ, Zheng XQ, Liu L, Xu M, Huang XJ. Generating responses with a specific emotion in dialog. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 3685–3695. [doi: 10.18653/v1/P19-1359]

[13] Colombo P, Witon W, Modi A, Kennedy J, Kapadia M. Affect-driven dialog generation. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 3734–3743. [doi: 10.18653/v1/N19-1374]

[14] Huang CY, Zaïane O, Trabelsi A, Dziri N. Automatic dialogue generation with expressed emotions. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 2: Short Papers). New Orleans: Association for Computational Linguistics, 2018. 49–54. [doi: 10.18653/v1/N18-2008]

[15] Rashkin H, Smith EM, Li M, Boureau YL. Towards empathetic open-domain conversation models: A new benchmark and dataset. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 5370–5381. [doi: 10.18653/v1/P19-1534]

[16] Lin ZJ, Madotto A, Shin J, Xu P, Fung P. MoEL: Mixture of empathetic listeners. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 121–132. [doi: 10.18653/v1/D19-1012]

[17] Majumder N, Hong PF, Peng SS, Lu JK, Ghosal D, Gelbukh A, Mihalcea R, Poria S. MIME: MIMicking emotions for empathetic response generation. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 8968–8979. [doi: 10.18653/v1/2020.emnlp-main.721]

[18] Li QT, Chen HS, Ren ZC, Ren PJ, Tu ZP, Chen ZM. EmpDG: Multi-resolution interactive empathetic dialogue generation. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 4454–4466. [doi: 10.18653/v1/2020.coling-main.394]

[19] Shin J, Xu P, Madotto A, Fung P. Generating empathetic responses by looking ahead the user's sentiment. In: Proc. of the 2020 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 7989–7993. [doi: 10.1109/ICASSP40776.2020.9054379]

[20] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444. [doi: 10.1038/nature14539]

[21] Yuan JH, Wu Y, Lu X, Zhao YY, Qin B, Liu T. Recent advances in deep learning based sentiment analysis. Science China Technological Sciences, 2020, 63(10): 1947–1970. [doi: 10.1007/s11431-020-1634-3]

[22] Chen HS, Liu XR, Yin DW, Tang JL. A survey on dialogue systems: Recent advances and new frontiers. ACM SIGKDD Explorations Newsletter, 2017, 19(2): 25–35. [doi: 10.1145/3166054.3166058]

[23] Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423–443. [doi: 10.1109/TPAMI.2018.2798607]

[24] Zhong PX, Wang D, Miao CY. Knowledge-enriched transformer for emotion detection in textual conversations. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 165–176. [doi: 10.18653/v1/D19-1016]

[25] Ghosal D, Majumder N, Gelbukh A, Mihalcea R, Poria S. COSMIC: Commonsense knowledge for emotion identification in conversations. In: Proc. of the 2020 Findings of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 2470–2481. [doi: 10.18653/v1/2020.findings-emnlp.224]

[26] Li JN, Lin Z, Fu P, Wang WP. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. Punta Cana: Association for Computational Linguistics, 2021. 1204–1214. [doi: 10.18653/v1/2021.findings-emnlp.104]

[27] Qin LB, Li ZY, Che WX, Ni MH, Liu T. Co-GAT: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. Proc. of the 2021 AAAI Conf. on Artificial Intelligence, 2021, 35(15): 13709–13717. [doi: 10.1609/aaai.v35i15.17616]

[28] Zhang D, Wu LQ, Sun CL, Li SS, Zhu QM, Zhou GD. Modeling both context- and speaker-sensitive dependence for emotion detection in

multi-speaker conversations. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: Artificial Intelligence Organization, 2019. 5415–5421. [doi: 10.24963/ijcai.2019/752]

[29] Ghosal D, Majumder N, Poria S, Chhaya N, Gelbukh A. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 154–164. [doi: 10.18653/v1/D19-1015]

[30] Ishiwatari T, Yasuda Y, Miyazaki T, Goto J. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 7360–7370. [doi: 10.18653/v1/2020.emnlp-main.597]

[31] Li JY, Ji DH, Li F, Zhang MS, Liu YJ. HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 4190–4200. [doi: 10.18653/v1/2020.coling-main.370]

[32] Shen WZ, Chen JQ, Quan XJ, Xie ZX. DialogXL: All-in-one XLNet for multi-party conversation emotion recognition. Proc. of the 2021 AAAI Conf. on Artificial Intelligence, 2021, 35(15): 13789–13797. [doi: 10.1609/aaai.v35i15.17625]

[33] Sun Y, Yu N, Fu GH. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. Punta Cana: Association for Computational Linguistics, 2021. 2949–2958. [doi: 10.18653/v1/2021.findings-emnlp.252]

[34] Dai YP, Yu HH, Jiang YX, Tang CG, Li YB, Sun J. A survey on dialog management: Recent advances and challenges. arXiv: 2005.02233, 2021.

[35] Zhang Z, Takanobu R, Zhu Q, Huang ML, Zhu XY. Recent advances and challenges in task-oriented dialog systems. Science China Technological Sciences, 2020, 63(10): 2011–2027. [doi: 10.1007/s11431-020-1692-3]

[36] Slonim N, Bilu Y, Alzate C, *et al*. An autonomous debating system. Nature, 2021, 591(7850): 379–384. [doi: 10.1038/S41586-021-03215-W]

[37] Küçük D, Can FL. Stance detection: A survey. ACM Computing Surveys, 2021, 53(1): 12. [doi: 10.1145/3369026]

[38] Christakopoulou K, Radlinski F, Hofmann K. Towards conversational recommender systems. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 815–824. [doi: 10.1145/2939672.2939746]

[39] Shen GY, Jia J, Nie LQ, Feng FL, Zhang CJ, Hu TR, Chua TS, Zhu WW. Depression detection via harvesting social media: A multimodal dictionary learning solution. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. Melbourne: AAAI Press, 2017. 3838–3844.

[40] Rinaldi A, Fox Tree J, Chaturvedi S. Predicting depression in screening interviews from latent categorization of interview prompts. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 7–18. [doi: 10.18653/v1/2020.acl-main.2]

[41] Poria S, Majumder N, Hazarika D, Ghosal D, Bhardwaj R, Jian SYB, Hong PF, Ghosh R, Roy A, Chhaya N, Gelbukh A, Mihalcea R. Recognizing emotion cause in conversations. Cognitive Computation, 2021, 13(5): 1317–1332. [doi: 10.1007/s12559-021-09925-7]

[42] Li YR, Su H, Shen XY, Li WJ, Cao ZQ, Niu SZ. DailyDialog: A manually labelled multi-turn dialogue dataset. In: Proc. of the 8th Int'l Joint Conf. on Natural Language Processing (Vol. 1: Long Papers). Taipei: Asian Federation of Natural Language Processing, 2017. 986–995.

[43] Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 42(4): 335–359. [doi: 10.1007/s10579-008-9076-6]

[44] Wei W, Liu JY, Mao XL, Guo GB, Zhu FD, Zhou P, Hu YC. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In: Proc. of the 28th ACM Int'l Conf. on Information and Knowledge Management. Beijing: ACM, 2019. 1401–1410. [doi: 10.1145/3357384.3357937]

[45] Zhang R, Wang ZY, Huang ZH, Li L, Zheng MD. Predicting emotion reactions for human-computer conversation: A variational approach. IEEE Trans. on Human-machine Systems, 2021, 51(4): 279–287. [doi: 10.1109/THMS.2020.3044975]

[46] Zhou XD, Wang WY. MojiTalk: Generating emotional responses at scale. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Melbourne: Association for Computational Linguistics, 2018. 1128–1137. [doi: 10.18653/v1/P18-1104]

[47] Shen L, Feng Y. CDL: Curriculum dual learning for emotion-controllable response generation. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 556–566. [doi: 10.18653/v1/2020.acl-main.52]

[48] Asghar N, Poupart P, Hoey J, Jiang X, Mou LL. Affective neural response generation. In: Pasi G, Piwowarski B, Azzopardi L, Hanbury

A, eds. Proc. of the 40th European Conf. on IR Research. Grenoble: Springer, 2018. 154–166. [doi: 10.1007/978-3-319-76941-7_12]
[49] Zhong PX, Wang D, Miao CY. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. Proc. of the 2019 AAAI Conf. on Artificial Intelligence, 2019, 33(1): 7492–7500. [doi: 10.1609/aaai.v33i01.33017492]
[50] Lubis N, Sakti S, Yoshino K, Nakamura S. Positive emotion elicitation in chat-based dialogue systems. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2019, 27(4): 866–877. [doi: 10.1109/TASLP.2019.2900910]
[51] Qiu LS, Shiu Y, Lin PP, Song RH, Liu Y, Zhao DY, Yan R. What if bots feel moods? In: Proc. of the 43rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Association for Computing Machinery, 2020. 1161–1170. [doi: 10.1145/3397271.3401108]
[52] Lu X, Tian YJ, Zhao YY, Qin B. Retrieve, discriminate and rewrite: A simple and effective framework for obtaining affective response in retrieval-based chatbots. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. Punta Cana: Association for Computational Linguistics, 2021. 1956–1969. [doi: 10.18653/v1/2021.findings-emnlp.168]
[53] Gao J, Liu YH, Deng HL, Wang W, Cao Y, Du JC, Xu RF. Improving empathetic response generation by recognizing emotion cause in conversations. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. Punta Cana: Association for Computational Linguistics, 2021. 807–819. [doi: 10.18653/v1/2021.findings-emnlp.70]
[54] Li QT, Li P, Ren ZC, Ren PJ, Chen ZM. Knowledge bridging for empathetic dialogue generation. Proc. of the 2022 AAAI Conf. on Artificial Intelligence, 2022, 36(10): 10993–11001. [doi: 10.1609/aaai.v36i10.21347]
[55] Zeng CK, Chen GY, Lin CH, Li RZ, Chen Z. Affective decoding for empathetic response generation. In: Proc. of the 14th Int'l Conf. on Natural Language Generation. Aberdeen: Association for Computational Linguistics, 2021. 331–340.
[56] Shen L, Zhang JC, Ou J, Zhao XF, Zhou J. Constructing emotional consensus and utilizing unpaired data for empathetic dialogue generation. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. Punta Cana: Association for Computational Linguistics, 2021. 3124–3134. [doi: 10.18653/v1/2021.findings-emnlp.268]
[57] Kim H, Kim B, Kim G. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: Association for Computational Linguistics, 2021. 2227–2240. [doi: 10.18653/v1/2021.emnlp-main.170]
[58] Sharma A, Miner A, Atkins D, Althoff T. A computational approach to understanding empathy expressed in text-based mental health support. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 5263–5276. [doi: 10.18653/v1/2020.emnlp-main.425]
[59] Welivita A, Pu P. A taxonomy of empathetic response intents in human social conversations. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 4886–4899. [doi: 10.18653/v1/2020.coling-main.429]
[60] Welivita A, Xie YB, Pu P. A large-scale dataset for empathetic response generation. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: Association for Computational Linguistics, 2021. 1251–1264. [doi: 10.18653/v1/2021.emnlp-main.96]
[61] Zhong PX, Zhang C, Wang H, Liu Y, Miao CY. Towards persona-based empathetic conversational models. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 6556–6566. [doi: 10.18653/v1/2020.emnlp-main.531]
[62] Zandie R, Mahoor MH. EmpTransfo: A multi-head transformer architecture for creating empathetic dialog systems. arXiv:2003.02958, 2020.
[63] Zheng CJ, Liu Y, Chen W, Leng YC, Huang ML. CoMAE: A multi-factor hierarchical framework for empathetic response generation. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. Association for Computational Linguistics, 2021. 813–824. [doi: 10.18653/v1/2021.findings-acl.72]
[64] Sabour S, Zheng CJ, Huang ML. CEM: Commonsense-aware empathetic response generation. Proc. of the 2022 AAAI Conf. on Artificial Intelligence, 2022, 36(10): 11229–11237. [doi: 10.1609/aaai.v36i10.21373]
[65] Snoek CGM, Worring M, Smeulders AWM. Early versus late fusion in semantic video analysis. In: Proc. of the 13th Annual ACM Int'l Conf. on Multimedia. Hilton: ACM, 2005. 399–402. [doi: 10.1145/1101149.1101236]
[66] Guo WZ, Wang JW, Wang SP. Deep multimodal representation learning: A survey. IEEE Access, 2019, 7: 63373–63394. [doi: 10.1109/ACCESS.2019.2916887]
[67] Wang YS, Shen Y, Liu Z, Liang PP, Zadeh A, Morency LP. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. Proc. of the 2019 AAAI Conf. on Artificial Intelligence, 2019, 33(1): 7216–7223. [doi: 10.1609/aaai.v33i01.33017216]

[68] Tsai YHH, Bai SJ, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 6558–6569. [doi: 10.18653/v1/P19-1656]
[69] Zadeh A, Chen MH, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 1103–1114. [doi: 10.18653/v1/D17-1115]
[70] Hazarika D, Zimmermann R, Poria S. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 1122–1131. [doi: 10.1145/3394171.3413678]
[71] Chen MH, Wang S, Liang PP, Baltrušaitis T, Zadeh A, Morency LP. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Proc. of the 19th ACM Int'l Conf. on Multimodal Interaction. Glasgow: ACM, 2017. 163–171. [doi: 10.1145/3136755.3136801]
[72] Rahman W, Hasan MK, Lee S, Zadeh A, Mao CF, Morency LP, Hoque E. Integrating multimodal information in large pretrained transformers. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 2359–2369. [doi: 10.18653/v1/2020.acl-main.214]
[73] Yu WM, Xu H, Yuan ZQ, Wu JL. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proc. of the 2021 AAAI Conf. on Artificial Intelligence, 2021, 35(12): 10790–10797. [doi: 10.1609/aaai.v35i12.17289]
[74] Wu Y, Zhao YY, Yang H, Chen S, Qin B, Cao XH, Zhao WT. Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors. arXiv:2203.00257, 2022.
[75] Haque A, Guo M, Miner AS, Fei-Fei L. Measuring depression symptom severity from spoken language and 3D facial expressions. arXiv:1811.08592, 2018.
[76] Chatterjee A, Gupta U, Chinnakotla MK, Srikanth R, Galley M, Agrawal P. Understanding emotions in text using deep learning and big data. Computers in Human Behavior, 2019, 93: 309–317. [doi: 10.1016/j.chb.2018.12.029]
[77] Hsu CC, Chen SY, Kuo CC, Huang TH, Ku LW. EmotionLines: An emotion corpus of multi-party conversations. In: Proc. of the 11th Int'l Conf. on Language Resources and Evaluation (LREC 2018). Miyazaki: European Language Resources Association (ELRA), 2018. 1597–1601.
[78] Danescu-Niculescu-Mizil C, Lee L. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In: Proc. of the 2nd Workshop on Cognitive Modeling and Computational Linguistics. Portland: Association for Computational Linguistics, 2011. 76–87.
[79] Zahiri SM, Choi JD. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI Press, 2018. 44–52.
[80] Tiedemann J. News from OPUS—A collection of multilingual parallel corpora with tools and interfaces. In: Nicolov N, Bontcheva K, Angelova G, Mitkov R, eds. Recent Advances in Natural Language Processing. Borovets, 2009. 237–248. [doi: 10.1075/cilt.309.19tie]
[81] Shang LF, Lu ZD, Li H. Neural responding machine for short-text conversation. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing (Vol. 1: Long Papers). Beijing: Association for Computational Linguistics, 2015. 1577–1586. [doi: 10.3115/v1/P15-1152]
[82] Gorrell G, Kochkina E, Liakata M, Aker A, Zubiaga A, Bontcheva K, Derczynski L. SemEval-2019 task 7: Rumoureval, determining rumour veracity and support for rumours. In: Proc. of the 13th Int'l Workshop on Semantic Evaluation. Minneapolis: Association for Computational Linguistics, 2019. 845–854. [doi: 10.18653/v1/S19-2147]
[83] Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In: Proc. of the 57th Annal Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 527–536. [doi: 10.18653/v1/P19-1050]
[84] Yu WM, Xu H, Meng FY, Zhu YL, Ma YX, Wu JL, Zou JY, Yang KC. CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 3718–3727. [doi: 10.18653/v1/2020.acl-main.343]
[85] Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D, Marsella S, Traum D, Rizzo S, Morency LP. The distress analysis interview corpus of human and computer interviews. In: Proc. of the 9th Int'l Conf. on Language Resources and Evaluation. Reykjavik: European Language Resources Association (ELRA), 2014. 3123–3128.
[86] Roller S, Dinan E, Goyal N, Ju D, Williamson M, Liu YH, Xu J, Ott M, Smith EM, Boureau YL, Weston J. Recipes for building an open-domain chatbot. In: Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics: Main Volume.

Association for Computational Linguistics, 2021. 300–325. [doi: 10.18653/v1/2021.eacl-main.24]
[87] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020, 33. 1877–1901.
[88] Bao SQ, He H, Wang F, Wu H, Wang HF, Wu WQ, Wu ZH, Guo Z, Lu H, Huang XX, Tian X, Xu XC, Lin YZ, Niu ZY. PLATO-XL: Exploring the large-scale pre-training of dialogue generation. arXiv:2109.09519, 2022.
[89] Nabla. Doctor GPT-3: Hype or reality? 2020. https://nabla.com/blog/gpt-3/
[90] GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019. The Lancet, 2020, 396(10258): 1204–1222. [doi: 10.1016/S0140-6736(20)30925-9]

附中文参考文献:

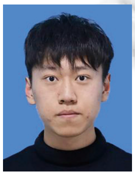
[2] 王天一. 人工智能革命: 历史、当下与未来. 北京: 北京时代华文书局, 2017.

赵妍妍(1983－), 女, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为情感计算, 自然语言处理.

田一间(1999－), 男, 硕士, 主要研究领域为自然语言处理, 情感计算.

陆鑫(1996－), 男, 博士生, 主要研究领域为情感计算, 对话情感.

秦兵(1968－), 女, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理.



赵伟翔(1999－), 男, 博士生, 主要研究领域为情感计算, 对话情感.