三思：用于情感响应生成的类人两阶段对话代理

Yushan Qian 天津大学 中国天津 yushanqian@tju.edu.cn

Bo Wang 天津大学 中国天津 bo_wang@tju.edu.cn

Shangzhao Ma 天津大学 中国天津 shangzhaoma@tju.edu.cn

Wu Bin Quesoar 有限公司 中国天津 wubin@quesoar.com

Shuo Zhang Quesoar有限公司 中国天津 s@quesoar.com

Dongming Zhao 人工智能实验室，中国移动通信集团天津有限公司，中国天津 waitman_840602@163.com

黄坤 人工智能实验室，中国移动通信集团天津有限公司，中国天津 hknzh@126.com

Yuexian Hou 天津大学，中国天津 yxhou@tju.edu.cn

arXiv:2301.04907v3 [cs.CL] 1 Oct 2024

摘要

面向类人对话系统，目前的情感对话方法使用统一神经网络联合建模情绪与语义。由于情绪与语义相互制约，这种策略往往会生成安全响应，并且需要罕见的大规模情感标注对话语料库。受人类智能对话中“三思而后行”行为的启发，我们提出了一种用于生成情感对话的两阶段对话代理。首先，使用不依赖情感标注对话语料库训练的对话模型，生成符合上下文语义的原型回复。其次，借助共情假设，通过可控情绪优化器修改第一阶段生成的原型回复。在 DailyDialog 和 EmpatheticDialogues 数据集上的实验结果表明，与对比模型相比，所提出的对话代理在情绪生成方面表现更优，并且在自动评估和人工评估中保持了语义表现。

关键词

情感对话；对话系统；人际交互

ACM参考文献格式：

Yushan Qian、Bo Wang、Shangzhao Ma、Wu Bin、Shuo Zhang、Dongming Zhao、KunHuang 和 Yuexian Hou。2023。三思：一种用于情感响应生成的类人两阶段对话代理。载于第*22*届自主智能体与多智能体系统国际会议论文集（*AAMAS 2023*），英国伦敦，*2023*年*5*月*29*日至*6*月*2*日，IFAAMAS，共10页。

第*22*届国际自主智能体与多智能体系统会议论文集（*AAMAS 2023*），*A. Ricci*、*W. Yeoh*、*N. Agmon*、*B. An*（编），*2023*年*5*月*29*日至*6*月*2*日，英国伦敦。 © 2023 国际自主智能体与多智能体系统基金会（www.ifaamas.org）。保留所有权利。

1 引言

在开放域对话生成任务中，情感对话旨在生成涉及适当情绪感知与表达的响应。大量研究 [30, 35, 40] 表明，情感对话能够显著提升用户在人机对话中的满意度。此外，构建具有人类情绪的对话系统是人工智能的终极目标之一。

在构建情感对话系统方面，除了早期由专业人员手动编写规则的方法 [45],，现有的统计方法主要基于神经网络模型 [1, 11, 21, 23, 25, 28, 46, 58–60]。采用端到端策略，这些神经网络模型能够联合生成对话回复的语义和情绪。

然而，当前端到端情感对话模型仍面临若干挑战。首先，在深度神经网络中，输入情绪信号往往会在复杂的学习过程中逐渐减弱。其次，在联合生成模型中，增强情绪的设计往往会限制生成响应的语义表现（例如，安全响应）。第三，适用于深度神经网络联合训练语义和情绪的大规模情绪标注对话语料库十分罕见。

针对上述挑战，我们提出借鉴人类智能对话行为的思路来生成情绪回应。人类在对话中回应时，同时处理情绪和语义并不能确保令人满意的结果 [7]。与语义同时产生的直觉情绪，往往难以确保回复情绪恰当。恰当的情感响应的一种来源，是在确定语义后独立进行情绪选择，即对恰当的情绪再三斟酌。在这种独立的情绪选择中，人类的一项关键策略是

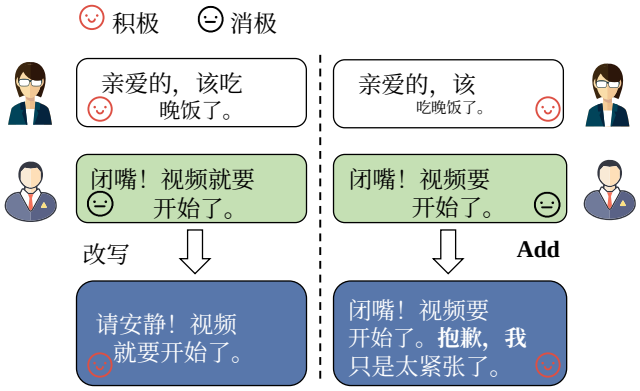

图1：人类对话中情绪调整的真实案例。人类智能行为中可以观察到“三思”策略，它通过改写表达或添加额外信息，有效提升了情绪回应的质量。粗体词元是改写或添加的部分。

确定恰当情绪的共情策略，它能使回复的情绪与上下文的情绪状态保持一致 [2, 25, 42]。如图1所示，每条回复首先都应具有恰当的语义，以回应上下文。随后，通过识别上下文的情绪状态，我们可以调整回复的情绪表达，并通过回应对方的情绪实现一定程度的共情。

因此，我们设计了一种用于情感响应生成的类人两阶段对话代理。首先，使用在未进行情绪标注的对话语料库上微调的预训练模型，生成一个语义恰当的原型回复。然后，通过对话情绪检测器识别上下文中的情绪状态。根据共情假设 [25],，生成的情绪类型与上下文中的情绪状态一致。最后，由可控情绪优化器修改原型回复，生成语义相关且情绪恰当的最终回复。

具体而言，为了在第二阶段有效地优化情绪回应，我们还借鉴了两种人类的语用策略。首先，人类会用不同的词汇选择以不同方式表达相同的信息 [39]。因此，我们通过替换原有的情绪词语或短语，而非从头构造新句子，将预期的情绪属性融入回应中，即采用“改写”策略。其次，有些隐含情绪并非通过特定词语，而是通过整句话传达。因此，我们还会通过在回应中添加额外句子的方式调整情绪，即采用“添加”策略。

总而言之，我们的贡献如下：

- 受人类智能对话行为的启发，我们提出了一种用于情感响应生成的类人两阶段对话代理。据我们所知，这是首个专为情感对话设计的两阶段模型。
- 该方法有效缓解了现有情感对话方法存在的两个问题：在复杂学习过程中情感效果减弱，以及为满足情绪需求而限制语义生成。
- 所提出的两阶段对话代理降低了对大规模情绪标注对话语料库的需求。第一阶段的原型回应生成器只需使用未经情绪标注的通用对话语料进行训练，而可控情绪对话优化器则使用非对话式且非平行的情绪标注语料进行训练。
- 所提出的方法可以作为情绪化后处理，推广应用于其他现有的端到端情绪对话生成模型。即使某些句子的情绪表达欠佳，也无需重新训练整个模型或从头构造新句子。

2 方法

2.1 预备知识

形式上，本文将对话上下文定义为两位说话者交替发出的话语，即 $C = \{U_1, U_2, ..., U_n\}$，其中 n 表示对话中的话语数量。上下文情绪集为 $E = \{e_1, e_2, ..., e_n\}$，它与对话上下文 C 相对应。我们的目标是生成下一条话语 U_{n+1}，使其与上下文连贯并包含恰当的情绪。

如图2所示，我们的模型由三部分组成：原型话语生成器 G、对话情绪检测器 D和可控情绪优化器 R。可控情绪优化器 R 包含两个模块，分别称为“改写”和“添加”。原型话语生成器 G 以对话上下文 C 为输入，并生成原型回复 U_m。对话情绪检测器 D 以对话上下文 C 为输入，并获取情绪状态集 E，该集合会动态决定回复情绪 e_{n+1}。可控情绪优化器 R 根据 e_{n+1} 对 U_m 进行优化，分别通过改写 U_m 或使用改写模块和添加模块向 U_m 中添加额外句子，并生成最终回复 R_e，即 U_{n+1}。

2.2 原型话语生成器

我们使用 DialoGPT [57] 作为原型话语生成器，以生成相关、多样且上下文连贯的回复。大规模预训练语言模型 [5, 41, 57] 近年来极大地推动了开放域对话研究的进展。预训练和微调的方法可以避免从头开始训练模型，节省计算资源，并在下游任务中取得出色的效果。

DialoGPT 采用具有 12 至 48 层的 Transformer，并使用类似 GPT-2 [41], 的层归一化结构；该模型在来自 Reddit 的 1.47 亿规模对话数据集上进行训练。上下文中的所有话语都与“<|endoftext |>”拼接成一个长句，作为输入。目标原型话语 U_m 的条件分布是条件概率序列的乘积：

$$P(U_m|C) = \prod_{i=S+1}^{S+s_m} P(t_i|t_1, t_2, ..., t_{i-1}), \quad (1)$$

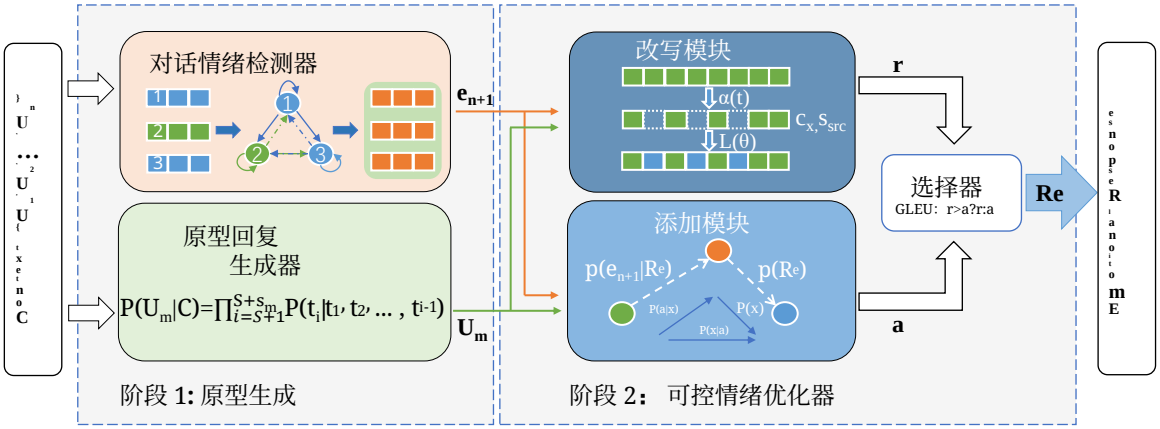

图2：所提出的两阶段对话代理的整体架构。第一阶段包括原型话语生成器和对话情绪检测器，分别用于生成原型回复 U_m 并检测作为最终回复预期情绪的上下文情感状态 e_{n+1} 。第二阶段包括改写模块、添加模块和选择器。改写模块和添加模块根据 e_{n+1}对 U_m 进行改进，选择器则根据 GLEU 分数，从改写模块和添加模块的输出中选择最终回复。

其中，$C = U_1\ U_2\ \ldots\ U_n$ n $U_i = \{t_1, t_2, \ldots, t_{s_i}\}$ t_i s_i 表示话语的数量，是话语中的每个词元，是每个话语的长度，$S = s_1 + s_2 + \cdots + s_n$ 表示所有话语中的词元数量，s_m 是 U_m 的长度。

我们采用最大互信息评分函数（MMI）和 top-k采样 [6] ，以减少无意义回复的生成。具体实现基于 Hugging Face[1] 提供的工具。我们在训练集上进行 3 个 epoch 的微调，批次大小为 8，学习率为 0.00001。训练集中的情绪标签不会被使用。在解码过程中，我们采用 top-k（k=100）采样和核采样（p=0.7） [14]。

2.3 对话情绪检测器

作为一种直观的共情假设，在两个人的情感对话中，听者通常倾向于以一种认可说话者感受的方式作出回应 [2, 25] ，并通过唤起回应者的情绪来实现一定程度的共情。在本研究的对话场景中，对话上下文（<4）中的轮次不多，因此我们采用这一共情假设，根据上下文的情绪状态来确定回复的情绪。

为此，对话情绪检测器 D 的目标是检测对话上下文中的情绪。根据共情假设， D 根据对话上下文中识别出的情绪分布，确定可控情绪优化器的预期情绪。

对话情绪检测器 D 基于 Dia- logueGCN [10],开发，它将对话中的每个话语视为图网络中的一个节点。话语之间存在有向边，话语的序列顺序决定了每条边的方向。这些有向边可以建模情绪影响，即说话者或其他人先前所说内容带来的影响。我们使用 GloVe嵌入和卷积神经网络提取话语的特征，并获得每个话语的嵌入 U_i，即每个节点的向量。每个话语之前有 b 个话语，之后有 a 个话语。每个话语节点与 $a+b+1$ 个节点（包括自身）相连。每条边的权重 a_{ij} 由节点之间的关系决定，如下所示：

$$\begin{aligned} \alpha_{ij} &= \mathrm{softmax}(U_i^T W_u [U_{i-b}, \ldots, U_{i+a}]), \\ &\text{for} \quad j = i-b, \ldots, i+a. \end{aligned} \tag{2}$$

有关 DialogueGCN 构建的更多细节，请参阅 [10]。最后，将序列编码器 sq 和说话者级编码器 sp 生成的嵌入拼接起来，并结合基于相似度的注意力机制，得到话语节点的最终嵌入。随后，我们使用全连接网络对多种情绪类别进行分类：

$$\begin{aligned} H &= [h_1, h_2, \ldots, h_n], \\ h_i &= \mathrm{softmax}([sq_i, sp_i]^T W H) H^T, \\ e_i &= \mathrm{argmax}(\mathrm{softmax}(\mathrm{FFN}(h_i))). \end{aligned} \tag{3}$$

我们使用带 L2正则化的分类交叉熵损失作为损失函数，并使用 Adam [18]作为优化器。我们按照 [28] 将情绪状态集 S 中的情绪分为负面情绪和正面情绪两组。遵循 [2, 25], ，我们假设共情回应可能会在一定程度上模仿用户的情绪。因此，最终传递给可控情绪优化器 R 的目标情绪 e_{n+1} 定义如下：

$$\begin{aligned} e_{n+1} &= \text{positive}, \\ \text{if} \quad Num_{pos}(E) &> Num_{neg}(E), \\ \text{otherwise} \quad e_{n+1} &= \text{negative}. \end{aligned} \tag{4}$$

[1]https://huggingface.co/models

其中， $E = \{e_1, e_2, \ldots\ldots, e_n\}$ 是每段对话中的情绪集合。Num_{pos} 和 Num_{neg} 分别表示 E中正面情绪和负面情绪的数量。

2.4 可控情绪优化器

可控情绪优化器 R 以原型回复U_m 和目标情绪 e_{n+1} 为输入，并生成最终回复 R_e。我们需要学习的目标定义如下：

$$P(R_e|U_m, e_{n+1}) \ \& \ \text{Stype}(R_e) = e_{n+1}, \tag{5}$$

其中，“Stype”表示情绪类型。

可控情绪优化器 R 由两个模块组成：“改写”和“添加”。改写模块通过将句子中的原始情绪符号替换为表达目标情绪的符号，来转换 U_m 的情绪属性。添加模块通过添加额外句子来调整情绪类型。我们分部分训练可控情绪优化器 R。

改写模块。改写模块由两部分组成：第一部分是删除部分，用于判断输入中的每个词元是否为情绪属性词，学习输入中的情绪部分和非情绪部分，并删除情绪部分。我们采用 Transformer 的注意力机制，提取注意力分数作为每个词元的权重 [47]:

$$\alpha(t) = \text{softmax}(QK^T), for\ t \in U_m, \tag{6}$$

其中， Q和 K 保留了 Transformer 中查询向量和键向量的原始含义。

第二部分是生成部分，用于生成具有目标情绪属性的句子。生成部分采用基于 Hugging Face [52] 的 Transformer 结构。生成部分的输入是原型回复和目标情绪。输出是符合目标情绪的句子。无需平行语料库，生成部分的训练目标是最小化以下重构损失：

$$L(\theta) = \sum_{x, s_{src} \in D} \log p(x|c_x, s_{src}; \theta), \tag{7}$$

其中， D是训练数据集。给定句子 x，改写模块模型学习在给定 c_x、 s_{src}的情况下重构 $y = x$。 c_x是 x的非情绪内容，而 s_{src} 是该句子的原始风格。

添加模块。 添加模块基于 [4] 的工作开发，通过添加带有目标情绪的额外句子来改变原句的情绪极性。利用贝叶斯定理，我们可以用模型 $p(x)$ 和模型 $p(a|x)$ 来表示模型 $p(x|a)$：

$$p(x|a) \propto p(a|x)p(x). \tag{8}$$

为了获得所需的 $p(x|a)$，以基于属性 a 生成句子，我们已有一个能够生成流畅句子的语言模型 $p(x)$。此外，我们构建了一个分类器，用于判断语言模型生成的文本 x 是否具有a 属性，即 $p(a|x)$，由此即可得到 $p(x|a)$。

添加模块的过程分为三个步骤：

1. 首先，通过语言模型进行前向传播，并利用预测 $p(a|x)$ 的属性模型计算目标属性的可能性。

表1：根据积极性和消极性对 DailyDialog 数据集中的情绪进行分组，共分为两组。

积极	消极
快乐，惊讶，其他	愤怒，厌恶、恐惧，悲伤

表2：根据积极性和消极性对 EmpatheticDialogues 数据集中的情绪进行分组。

积极	负面
自信，喜悦，感激，印象深刻，自豪，兴奋，信任-感，充满希望，忠诚，做好准备的，内容，惊讶的，关怀的	害怕的，愤怒的，恼怒的，预期-焦虑的，担忧-的，羞愧的，极度悲痛的，失-被任命的，厌恶的，尴尬的、尴尬的，狂怒的，内疚的，嫉妒的，孤独的，怀旧的，悲伤的，情感化的，恐惧的

2. 其次，进行反向传播，利用属性模型的梯度更新语言模型的内部潜在表示，以提高句子具有目标属性的可能性。

3. 第三，根据获得的新输出概率分布重新采样，以生成一个新词。

为了生成更多符合语言模型的多样化句子，我们采用两种方法，确保生成句子的语言模型尽可能接近原始语言模型：Kullback–Leibler（KL）散度和后归一化几何均值融合。关于语言模型，我们使用 GPT2。对于特定属性判别器 $p(a|x)$，我们使用现有的带有情绪标注的非对话语料库，并预训练一个分类器。

选择器。 选择器用于判断回复是来自改写模块还是添加模块，并选取相应回复作为最终输出。选择器以 GLEU [31] 作为评估回复整体效果的依据，并将其与原型回复进行比较。**（作者？）** [47] 发现，与 BLEU 分数相比，GLEU 更适合用于衡量人工评分。选择器会选取 GLEU 分数较高的回复作为最终回复。

3 实验

本节介绍数据集、基线方法和评估指标。我们通过实验将所提出的对话代理与基线方法进行比较，并讨论实验结果。

3.1 数据集

我们使用 DailyDialog [24] 和 EmpatheticDialogues [42]数据集进行实验。DailyDialog 是一个用于日常聊天场景的多轮对话数据集，共包含 12,218 段对话和 103,607 条话语。每条话语都标注了主题和情绪。情绪分为七种：愤怒、厌恶、恐惧、快乐、悲伤、惊讶和其他。参见 **(作者？)** [28], 我们将这 7 种情绪分为两组，分别包含 3 种正面情绪

和 4 种负面情绪，如表1所示。EmpatheticDialogues 是一个广泛使用的共情响应生成基准数据集，这是一个大规模多轮数据集，包含众包工作者之间的 25,000 段共情对话。EmpatheticDialogues 还为每段对话提供一个情绪标签，情绪标签来自 32 种可选情绪。按照**(作者？)** [28], 我们将这 32 种情绪分为两组，分别包含 13 种正面情绪和 19 种负面情绪，如表2所示。我们关注正面和负面情绪，因为极性层面情绪的一致性在情绪研究中更受关注，并且在应用中更为稳健。由于我们的方法和基线方法基于相同的假设，并且在评估过程中以相同方式处理，因此结果具有竞争力且令人信服。

考虑到运行空间有限，并为了统一每段对话的轮次数，我们将原始对话切分为每段包含4轮的子对话。最终，DailyDialog 数据集的训练集/验证集/测试集对话数量分别为54,299/5,109/4,782。EmpatheticDialogues 数据集的训练集/验证集/测试集对话数量分别为18,383/2,810/3,320。

3.2 对比模型

据我们所知，这是早期探索情感对话两阶段生成的研究。鉴于共情假设，我们将本方法与相关任务中使用的一系列模型进行比较，包括通用对话、情感对话和共情对话。

Transformer [51]：标准 Transformer 模型，通过训练优化负对数似然损失。

Multi-TRS [42]：一种多任务 Transformer 模型，通过预测情绪并生成回复进行联合训练。

Mojitalk [60]：一种融入了情绪嵌入的基于编码器-解码器的条件变分自编码器模型。

MoEL [25]：一种基于 Transformer 的模型，采用情绪专属解码器，并将其输出汇总后输入最终解码器，以生成共情回复。

MIME [28]：一种基于 Transformer 的模型，利用情绪组和情绪模仿，有效融合正向和负向情绪组中的情绪，并生成共情回复。

EmpDG [21]：一种由生成器和判别器组成的交互式对抗模型。判别器需要用户反馈。此外，该模型同时利用粗粒度的对话级情绪和细粒度的词元级情绪。参照 **(作者？)** [43],，为确保测试集中的输入和输出一致，从而与其他基线方法进行公平比较，我们仅使用共情生成器。

3.3 实现细节

我们使用所有基线方法的官方代码。具体而言，EmpDG 仅使用共情生成器（Multi-TRS2、Mojitalk3、MoEL4、MIME5、EmpDG[6]）。除 Mojitalk 外，我们均使用 PyTorch 实现所有模型。所有基线方法均在 V100 GPU 上训练，批次大小为 16，并采用早停策略。对于 Adam 优化器，我们将 $\beta 1 = 0.9$ 和 $\beta 2 = 0.98$ 作为参数。自动评估中的情绪检测所用的 Senta 情绪预训练模型为“*ernie_2.0_skep_large_en*”。

[2]https://github.com/facebookresearch/EmpatheticDialogues [3] https://github.com/Claude-Zhou/MojiTalk [4] https://github.com/HLTCHKUST/MoEL [5] https://github.com/declare-lab/MIME [6] https://github.com/qtli/EmpDG

3.4 评估指标

3.4.1 自动评估。 我们在自动评估中采用以下评估指标：

BLEU：与人工回复的词语重合度得分 [34]。我们使用 BLEU-4，该指标由自然语言生成评估工具包计算得出 [7]。

多样性：Dist-n 用于衡量生成回复中不同 n元语法所占的比例 [26]。它常用于评估对话模型能否像人类一样生成多样化的回复。多样性低通常意味着模型倾向于针对不同上下文生成相似的安全响应。我们参考 [3] 的工作来计算 Dist-1 和 Dist-2 指标。

情绪准确率（准确率）：情绪准确率定义为生成回复与真实标注之间情绪极性一致的比例。在评估过程中，我们使用百度提出的情感分析情感知识增强预训练（SKEP）模型 [48] 作为情绪检测器[8]。SKEP 是一款先进的情绪检测器，在 14 项典型的中英文情感分析任务中表现出色。我们使用该模型自动检测所提出的对话代理和基线方法生成的回复中的情绪极性。

3.4.2 人工评估。我们随机抽取了100段对话，并使用我们提出的对话代理和基线方法生成回复。我们邀请了三位人工标注员，从以下四个方面评估每条回复：

内容（连贯性）：回复是否符合当前对话的上下文。采用李克特量表进行评分（1：完全不符合，3：有些符合，5：非常符合）。

情绪（情感）：回复的情绪极性是否符合上下文。1表示回复恰当，0表示回复不恰当。

情绪强度（信息性）：回复的情绪强度。0表示没有情绪，1表示强度较弱，2表示强度较强。

流畅度（流畅度）：回复是否易于阅读和理解。采用李克特量表进行评分（1：完全不符合，3：有些符合，5：非常符合）。

3.5 主要结果

DailyDialog 和 EmpatheticDialogues 数据集上的自动评估与人工评估结果分别如表3和表4所示。

在情绪生成方面，可以观察到，在两个数据集上，所提出的对话代理在自动评估的**准确率** 和人工评估的**情感** 方面均优于基线方法，这表明我们的对话代理在情绪生成方面表现出色。此外，我们的对话代理在 DailyDialog 和 EmpatheticDialogues 数据集上的**信息性** 分别取得了最佳和次佳结果。这清楚地证明，与最先进水平的端到端系统相比，我们的模型所产生的情绪效果更为显著且充分。

[7]https://github.com/Maluuba/nlg-eval [8] https://github.com/baidu/Senta

表3：DailyDialog 数据集上的自动评估和人工评估（显著提升的 p值为 < 0.05（t检验）。人工评估的 Fleiss κ 系数为 0.526，表明一致性为“中等一致性”。）

模型	自动评估				人工评估			
	BLEU-4	Dist-1	Dist-2	准确率(%)	Con	Emo	Int	Flu
Transformer	0.44	1.10	5.84	54.04	3.54	0.69	0.49	4.60
Multi-TRS	0.58	1.17	6.35	54.29	3.49	0.75	0.52	4.35
Mojitalk	0.58	4.99	22.40	53.99	3.14	0.79	0.61	4.21
MoEL	0.46	1.51	8.39	55.44	3.35	0.77	0.73	**4.67**
MIME	0.54	0.48	1.79	53.43	3.42	0.78	0.68	4.43
EmpDG	0.57	1.12	5.73	56.02	3.40	0.79	0.77	4.47
Our	**0.67**	**9.71**	**42.77**	**57.36**	**3.82**	**0.84**	**0.79**	4.45

表4：EmpatheticDialogues 数据集中的自动评估和人工评估（p值 <0.05 时结果具有显著提升（t检验）。人工评估的 Fleiss κ 系数为 0.462，表示“中等一致性”。）

模型	自动评估				人工评估			
	BLEU-4	Dist-1	Dist-2	准确率(%)	Con	Emo	Int	Flu
Transformer	0.35	0.64	2.40	62.29	3.24	0.62	0.84	**4.86**
Multi-TRS	0.35	0.73	2.77	61.36	3.30	0.76	0.98	4.80
Mojitalk	0.23	**6.99**	**33.52**	61.81	2.74	0.76	1.03	4.54
MoEL	0.34	1.15	7.28	62.47	3.44	0.77	1.04	4.82
MIME	0.37	0.89	3.93	61.87	3.32	0.80	**1.12**	**4.86**
EmpDG	**0.39**	0.75	2.50	62.62	3.33	0.81	1.05	4.78
Our	**0.39**	5.24	22.37	**65.42**	**3.65**	**0.82**	1.11	4.72

表 5：消融分析。

模型	DailyDialog				EmpatheticDialogues			
	BLEU-4	Dist-1	Dist-2	准确率(%)	BLEU-4	Dist-1	Dist-2	准确率(%)
Our	**0.67**	**9.71**	**42.77**	**57.36**	**0.39**	**5.24**	**22.37**	65.42
不添加	0.59	7.48	33.79	56.04	0.29	4.07	19.63	65.33
不改写	0.49	7.93	35.74	55.88	0.35	4.95	20.27	**65.78**
不使用 DED	0.63	8.77	39.27	56.84	—	—	—	—

这是因为在两个阶段的过程中，回复的情绪效果在第二阶段被单独确定并加以优化，不受第一阶段语义生成的影响。

在语义生成性能方面，所提出的对话代理在**BLEU-4** 和**连贯性**指标上达到最高水平。在**Dist-1** 和**Dist-2**方面，我们的对话代理同样取得了中等水平的得分。这些结果证实，所提出的对话代理在保持适当语义的同时，显著提升了情感表达。我们还注意到，所提出的对话代理在**流畅度**方面的得分一般，这可能是由于“添加”策略增加了句子长度，从而影响了阅读难度。鉴于所有模型的得分都高于 4，这只是一个小的局限性，我们将在未来对此进行探索。

就所比较的基线方法的表现而言，MoEL 的**BLEU-4**得分较低，而情绪准确率得分较高，这表明该现有模型在追求情绪特征时丢失了语义信息。此外，除 Mojitalk 外，所有基线方法的多样性指标得分都较低，这说明生成的响应中存在大量安全响应。在 DailyDialog 数据集上，MoEL、Mojitalk、MIME 和 EmpDG 的**连贯性**得分也较低，低于 Transformer。这可能主要是由于语义和情绪相互制约，从而缩小了输出空间。

由于评估的指标共有 8 项（自动评估和人工评估），少于 30 项，因此我们选择使用 t 检验。具体而言，我们使用 dcdd"scipy.stats" 软件包中的“ttest_ind”来计算我们方法与基线方法的指标值。所有 p 值均小于 0.05，接受该假设，这意味着结果具有统计显著性。

在图3中，我们针对100个人工评估样本，比较了我们提出的对话智能体生成的原型回复和精炼后的回复在情绪正确性和显著性方面的表现。红色和蓝色柱分别表示正确（即与上下文情绪连贯）和不正确的情绪。柱子的长度表示情绪的显著性。我们可以注意到，精炼后的回复中红色柱子的数量更多、长度也更长，这说明

精炼后的回复提升了原型回复中情绪的正确性和显著性。

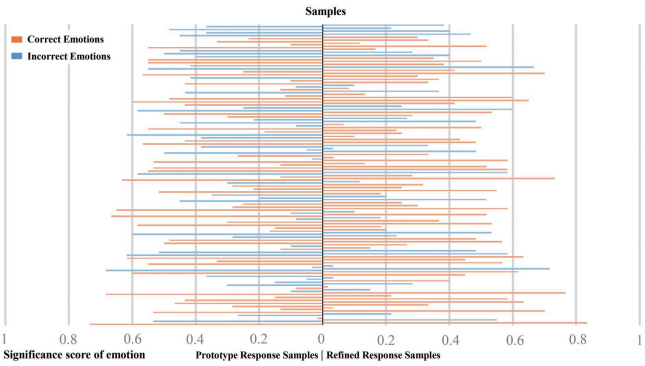

图3：从情绪的正确性和显著性方面比较原型回复与精炼后的回复。左列和右列分别表示原型回复和精炼后的回复。红色和蓝色列分别表示正确的情绪（即与上下文情绪连贯）和不正确的情绪。列的长度表示情绪的显著性。

3.6 人工 A/B 测试

我们进行了人工 A/B 测试，结果如表 7 所示。我们随机抽取了 100 对回复，并要求 3 名标注员根据对话上下文选择更优的回复。如果两者都好或都不好，则判为平局。标注者间一致性通过 Fleiss κ 系数进行测量。每个示例需要评判六次。每 100 个示例的费用约为 43.12 美元。我们可以观察到，与其他模型生成的回复相比，所提出的对话代理生成的回复更受标注员青睐，这表明具有恰当的情绪和多样性的回复对用户更具吸引力。

3.7 消融分析

为了验证我们提出的对话代理的有效性，我们还进行了消融研究。**1) 不添加**：移除可控情绪优化器中的添加模块。我们仅考虑通过“改写”来优化原型回复；**2) 不添加改写**：移除可控情绪优化器中的改写模块，并考虑通过“添加”来优化原型回复；**3) 不添加 DED**：移除对话情绪检测器，并以单个话语的情绪识别取而代之。由于 EmpatheticDialogues 数据集的情绪标注是对话级的，因此此实验仅在 DailyDialog 数据集上进行。

如表 5 所示，我们可以观察到，移除添加模块或改写模块都会导致大多数指标下降。这表明，结合“改写”和“添加”策略，有助于生成恰当的回复，使其符合显性与隐性表达中的人类语言特征。然而，第二阶段的选择器仍有改进空间，例如，可以根据原型回复和情绪预先选择要使用的模块。如何通过“改写”或“添加”模块更有效、合理地优化原型回复，是一个值得探索的问题。此外，对话情绪检测器在情感响应生成中也发挥着重要作用，其效果优于将上下文拼接成长句或识别单个话语。

3.8 案例研究

我们从表 6 中的所有模型生成的回复里抽取了一些样本。可以观察到，其他基线方法生成的回复包含情绪表达，但整体而言语义不够恰当。尽管 Transformer 生成的回复流畅，但往往与上下文不符。相比之下，我们提出的对话代理生成的回复不仅保留了上下文语义，同时也表达了丰富且恰当的情绪。例如，所提出的对话代理将“It is a problem. I am not sure how to solve it.”连贯地改写为“It is a problem. I am not sure how to solve it. I feel very sorry.”

4 相关工作

对话情绪识别。与独立句子的情绪识别不同，对话中的情绪应结合上下文进行识别。常用的上下文通常包括历史话语 [38], 历史情绪 [29] 以及说话者之间的相互影响 [13]。为对上下文建模，可以使用 GRU 对话语和说话者分别进行建模 [13] , 也可以进行交互式建模 [12] 。 [10]使用 GCN 解决现有基于 GRU 的方法中的上下文传播问题。常识知识 [9], 心理学知识 [19],以及情绪认知理论 [15]也被用于增强对话情绪识别。

情感对话。情感对话旨在生成情绪回应，主要有两种策略。一种策略是预先指定目标情绪 [23, 46, 59, 60]。这种策略的优点是生成的情绪灵活且可控，缺点是需要大规模的情绪标注对话语料库。另一种策略是利用对话上下文自行学习情绪 [1], , 这与共情对话相近 [8, 21, 22, 25, 28, 44] , 其假设是听者能够推断说话者的情绪 [42]。这种策略的优点是能够利用现有的大规模对话语料库，缺点是生成回应的情绪难以控制。此外，近期出现了一项颇具前景的任务——情感支持对话 [27] , 它为有需要的人提供了宝贵的帮助 [36, 37, 50]。

可控文本生成。可控文本生成旨在生成具有可控风格的文本。风格被定义为属于特定类别或标签的词元 [39]。典型流程包括从头开始训练大规模条件生成模型、基于预训练语言模型进行微调，以及替换关键 n 元组以调整生成句子的风格 [4]。情绪作为一种风格，其可控生成具有重要的实际意义。情绪可控文本生成是指在不改变语境意图的情况下，重新定义文本，使其包含特定情绪 [16, 17, 32, 47, 53]。

表 6：我们提出的对话代理和基线方法生成的回复示例。相应的上下文未出现在训练集中。粗体词元为经过润色的部分。

上下文	真实标注回复	模型	生成的回复
		DailyDialog	
不，亲爱的，你昨晚为什么没去补习学校？ 爸爸，我现在不想谈这件事。 亲爱的，如果你不想去补习学校， 你应该告诉我原因。	对不起，爸爸。但我会 我宁愿待在学校，也不愿去 补习学校。	Transformer	我觉得我做不到。
		Multi-TRS	Yeah, 我知道, 但我确定我会的。
		Mojitalk	嗯，能做到这一点就很好。
		Moel	我不知道。我想去上学。
		MIME	没错。但我不知道。
		EmpDG	对不起。我迫不及待想离开互联网。
		Ours	**对不起，**爸爸。我不能**再**那样做了。我太忙了。
我可以让 Suzy 住在我家。 我不想那样对你，Karen。 为什么不呢?	你自己的孩子会得到 Suzy 感冒了。我觉得最好 还是找个靠谱的 保姆来照看她。	Transformer	Yes, 我想我会很乐意的。
		Multi-TRS	对不起，我想我必须给她一个女孩。
		Mojitalk	嗯，我感冒了。
		Moel	我不知道。我愿意。你呢?
		MIME	我不确定。我不确定。
		EmpDG	对不起，但我一点办法也没有。
		Ours	对不起。我真的做不到。**我没什么可负的**。 **我只是想充分利用我所拥有的一切，尽我所能做到最好**。
		EmpatheticDialogues	
为什么麦当劳总是弄错我的订单！ 我说了不要生菜！ 向他们的管理层报告。 有时我觉得管理层根本不在乎！	就试试看吧。	Transformer	听到这个消息我很难过。希望他们能赢。
		Multi-TRS	希望你下次能做得更好。
		Mojitalk	这主意真不错！我相信你能得到他的家人，他也遇到过一些问题。
		Moel	希望一切顺利！
		MIME	听到这个消息我很遗憾。希望你能找到一份新工作。
		EmpDG	我不知道自己在做什么。
		Ours	这是个问题。我不确定该如何解决。**我感到非常抱歉**。
我曾经去探望过一名孤儿，那是一种如此超现实的时刻。这让我意识到自己应该多么感恩能拥有一个家庭。 那一定是一次强烈而充满情绪的经历 属于你的时刻。 是的，确实如此。我本来是想去 教这个孩子一些东西，但最后我却从他身上学到了更多东西。	Uau, 真是个超现实的时刻。 我一直相信孩子们拥有一种 有很多东西可以教给 我们。	Transformer	我相信他会成为一个好人。
		Multi-TRS	我很高兴你能找到一个新的。
		Mojitalk	我也是。原来的孩子现在确实多了很多。
		Moel	我很高兴你站在正确的一边?
		MIME	太好了。我很高兴你为自己感到开心。
		EmpDG	我相信你能做到。
		Ours	我很高兴你有过**如此美好的**经历。

表 7：人工 A/B 测试结果。DailyDialog 和 EmpatheticDialogues 的 Fleiss' κ系数分别为 0.612 和 0.496，分别表示"高度一致"和"中等一致性"。

模型	DailyDialog			EmpatheticDialogues		
	Win	Lose	Tie	Win	Lose	Tie
我们与 Transformer 对比	45.0%	34.7%	20.3%	56.3%	27.3%	16.3%
我们的方法 vs Multi-TRS	46.0%	31.0%	23.0%	52.3%	28.0%	19.7%
我们的与 Mojitalk	56.3%	26.0%	17.7%	53.3%	26.0%	20.7%
我们的方法与 MoEL 对比	44.7%	31.0%	24.3%	46.7%	19.3%	14.0%
我们的方法与 MIME 对比	49.0%	31.7%	19.3%	51.0%	25.7%	23.3%
我们 vs EmpDG	42.0%	32.0%	26.0%	48.0%	31.7%	26.3%

所提出的对话代理与现有方法之间的区别如下：

(1) 据我们所知，我们是首个专门研究情感对话领域两阶段情感响应生成范式的团队。

(2) 我们利用动态识别的对话上下文情绪来改进响应。然而，当前的改写方法并未考虑情绪的动态获取。

5 结论

本文设计了一种用于情感对话领域的两阶段对话代理，可生成与内容相关且富有情感的回应。该对话代理在第一阶段生成语义连贯的原型回复，并在第二阶段对其进行情感润色。大量自动评估和人工评估表明，所提出的对话代理能够生成语义恰当的高质量情绪回应，并降低对大型情感标注对话语料库的需求。

未来，为了改进所提出的对话代理，我们将探索对显性与隐性表达的预测，并灵活增强对话系统中除情绪之外的其他特定特征，例如现有对话模型的领域适配和风格适配。

参考文献

[1] Nabiha Asghar, P. Poupart, J. Hoey, Xin Jiang, and Lili Mou. 情感神经响应生成。*ArXiv*，abs/1709.03968，2018。[2] Laurie Carr, Marco Iacoboni, Marie-Charlotte Dubeau, John C. Mazziotta, and Gian Luigi Lenzi. 人类共情的神经机制：从模仿神经系统到边缘区的中继。《美国国家科学院院刊》，100(9):5497–5502，2003。[3] Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 通过多重映射和后验映射选择生成多个多样化响应。发表于第二十八届国际人工智能联合会议论文集，*IJCAI-19*，第4918–4924页。国际人工智能联合会议组织，2019年7月。[4] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 即插即用语言模型：一种可控文本生成的简单方法，2019。[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT：用于语言理解的深度双向变换器预训练。发表于*2019*年计算语言学协会北美分会年会：人类语言技术会议论文集，第*1*卷（长篇与短篇论文），第4171–4186页，明尼阿波利斯，明尼苏达州，2019年6月。计算语言学协会。[6] Angela Fan, Mike Lewis, and Yann Dauphin. 分层神经故事生成。*arXiv*预印本 *arXiv:1805.04833*，2018。[7] Nico H Frijda. 情绪、认知结构与行动倾向。《认知与情绪》，1(2):115–143，1987年。[8] Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 通过识别对话中的情绪原因改进共情响应生成。发表于 Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih（编），计算语言学协会会议论文集：*EMNLP 2021*，线上会议 / 多米尼加共和国蓬塔卡纳，*2021*年*11*月*16*日至*20*日，第807–819页。计算语言学协会，2021。[9] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. COSMIC：用于对话情绪识别的常识知识。发表于计算语言学协会会议论文集：*EMNLP 2020*，第2470–2481页，线上，2020年11月。计算语言学协会。[10] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN：用于对话情绪识别的图卷积神经网络。发表于*2019*年自然语言处理实证方法会议与第九届国际自然语言处理联合会议论文集（*EMNLP-IJCNLP*），第154–164页，中国香港，2019年11月。计算语言学协会。[11] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. Affect-LM：用于可定制情感文本生成的神经语言模型。发表于计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），第634–642页，加拿大温哥华，2017年7月。计算语言学协会。[12] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. ICON：用于多模态情绪检测的交互式对话记忆网络。发表于*2018*年自然语言处理实证方法会议论文集，第2594–2604页，比利时布鲁塞尔，2018年10月至11月。计算语言学协会。[13] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 用于双人对话视频情绪识别的对话记忆网络。发表于*2018*年计算语言学协会北美分会年会：人类语言技术会议论文集，第*1*卷（长篇论文），第2122–2132页，新奥尔良，路易斯安那州，2018年6月。计算语言学协会。[14] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 神经文本退化的奇特案例。*arXiv*预印本 *arXiv:1904.09751*，2019。[15] Dou Hu, Lingwei Wei, and Xiaoyong Huai. DialogueCRN：用于对话情绪识别的上下文推理网络。发表于 Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli（编），计算语言学协会第*59*届年会与第十一届国际自然语言处理联合会议论文集，*ACL/IJCNLP 2021*（第*1*卷：长篇论文），线上会议，*2021*年*8*月*1*日至*6*日，第7042–7052页。计算语言学协会，2021。[16] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 朝着可控文本生成迈进。发表于 Doina Precup and Yee Whye Teh（编），第三十四届国际机器学习会议论文集，机器学习研究论文集第70卷，第1587–1596页。PMLR，2017年8月6日至11日。[17] N. Keskar, Bryan McCann, L. Varshney, Caiming Xiong, and R. Socher. Ctrl：一种用于可控生成的条件变换器语言模型。*ArXiv*，abs/1909.05858，2019。[18] Diederik P. Kingma and Jimmy Ba. Adam：一种随机优化方法，2017。[19] Jiangnan Li、Zheng Lin、Peng Fu 和 Weiping Wang。过去、现在与未来：通过心理学知识的结构建模进行对话情绪识别。载于 Marie-Francine Moens、Xuanjing Huang、Lucia Specia 和 Scott Wen-tau Yih 编，计算语言学协会会议论文集：*EMNLP 2021*，线上会议／多米尼加共和国蓬塔卡纳，*2021*年*11*月*16*日至*20*日，第1204–1214页。计算语言学协会，2021年。[20] Jiwei Li、Michel Galley、Chris Brockett、Jianfeng Gao 和 Bill Dolan。促进多样性的神经对话模型目标函数。载于*2016*年计算语言学协会北美分会年会论文集：人类语言技术，第110–119页，圣迭戈，加利福尼亚州，2016年6月。计算语言学协会。[21] Qintong Li、Hongshen Chen、Zhaochun Ren、Pengjie Ren、Zhaopeng Tu 和 Zhumin Chen。EmpDG：多分辨率交互式共情对话生成。载于 Donia Scott、Núria Bel 和 Chengqing Zong 编，第*28*届国际计算语言学会议论文集（*COLING 2020*），西班牙巴塞罗那（线上），*2020*年*12*月*8*日至*13*日，第4454–4466页。国际计算语言学委员会，2020年。[22] Qintong Li、Piji Li、Zhaochun Ren、Pengjie Ren 和 Zhumin Chen。共情对话生成中的知识桥接。2022年。[23] Shifeng Li、Shi Feng、Daling Wang、Kaisong Song、Yifei Zhang 和 Weichao Wang。Emoelicitor：一种具备用户情绪反应感知能力的开放域响应生成模型。载于 Christian Bessiere 编，第二十九届国际人工智能联合会议论文集，*IJCAI-20*，第3637–3643页。国际人工智能联合会议组织，2020年7月。主会场论文。[24] Yanran Li、Hui Su、Xiaoyu Shen、Wenjie Li、Ziqiang Cao 和 Shuzi Niu。DailyDialog：人工标注的多轮对话数据集。载于第八届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），第986–995页，中国台湾台北，2017年11月。亚洲自然语言处理联盟。[25] Zhaojiang Lin、Andrea Madotto、Jamin Shin、Peng Xu 和 Pascale Fung。MoEL：共情倾听者混合模型。载于*2019*年自然语言处理实证方法会议与第九届国际自然语言处理联合会议论文集（*EMNLP-IJCNLP*），第121–132页，中国香港，2019年11月。计算语言学协会。[26] Chia-Wei Liu、Ryan Lowe、Iulian Serban、Mike Noseworthy、Laurent Charlin 和 Joelle Pineau。如何不评估对话系统：对用于对话响应生成的无监督评估指标进行实证研究。载于*2016*年自然语言处理实证方法会议论文集，第2122–2132页，奥斯汀，得克萨斯州，2016年11月。计算语言学协会。[27] Siyang Liu、Chujie Zheng、Orianna Demasi、Sahand Sabour、Yu Li、Zhou Yu、Yong Jiang 和 Minlie Huang。迈向情感支持对话系统。载于 Chengqing Zong、Fei Xia、Wenjie Li 和 Roberto Navigli 编，计算语言学协会第*59*届年会暨第十一届国际自然语言处理联合会议论文集，*ACL/IJCNLP 2021*（第*1*卷：长篇论文），线上会议，*2021*年*8*月*1*日至*6*日，第3469–3483页。计算语言学协会，2021年。[28] Navonil Majumder、Pengfei Hong、Shanshan Peng、Jiankun Lu、Deepanway Ghosal、Alexander Gelbukh、Rada Mihalcea 和 Soujanya Poria。MIME：模仿情绪以生成共情响应。载于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），第8968–8979页，线上会议，2020年11月。计算语言学协会。[29] Navonil Majumder、Soujanya Poria、Devamanyu Hazarika、Rada Mihalcea、Alexander Gelbukh 和 Erik Cambria。DialogueRNN：用于对话情绪检测的注意力循环神经网络。*AAAI*人工智能会议论文集，33(01):6818–6825，2019年7月。[30] Bilyana Martinovski 和 David R. Traum。人机交互中的故障：错误就是线索。载于*ISCA*对话系统中的错误处理教程与研究研讨会，2003年8月。2003年。[31] Courtney Napoles、Keisuke Sakaguchi、Matt Post 和 Joel Tetreault。无需调优的GLEU，2016年。[32] Kartikey Pant、Yash Verma 和 R. Mamidi。SentiInc：在无平行数据的情况下，将情感信息纳入情感迁移。信息检索进展，12036：312–319，2020年。[33] Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第*40*届年会论文集，第311–318页，美国宾夕法尼亚州费城，2002年7月。计算语言学协会。[34] Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第*40*届年会论文集，*2002*年*7*月*6*日至*12*日，美国宾夕法尼亚州费城，第311–318页。ACL，2002年。[35] Timo Partala 和 V. Surakka。情感干预对人机交互的影响。*Interact. Comput.*，16：295–309，2004年。[36] Wei Peng、Yue Hu、Luxi Xing、Yuqiang Xie、Yajing Sun 和 Yunpeng Li。全局控制，局部理解：一种全局到局部层次图网络，用于

情感支持对话，2022年。[37] Wei Peng、Ziyuan Qin、Yue Hu、Yuqiang Xie 和 Yunpeng Li。Fado：用于情感支持对话的反馈感知双重控制网络。Knowledge-Based Systems，第110340页，2023年。[38] Soujanya Poria、Erik Cambria、Devamanyu Hazarika、Navonil Majumder、Amir Zadeh 和 Louis-Philippe Morency。用户生成视频中的上下文相关情感分析。载于计算语言学协会第55届年会论文集（第1卷：长篇论文），第873–883页，加拿大温哥华，2017年7月。计算语言学协会。[39] Shrimai Prabhumoye、Alan W Black 和 Ruslan Salakhutdinov。探索可控文本生成技术。载于第28届国际计算语言学会议论文集，第1–14页，西班牙巴塞罗那（线上），2020年12月。国际计算语言学委员会。[40] Helmut Prendinger、Junichiro Mori 和 Mitsuru Ishizuka。利用人体生理反应评估数学游戏中虚拟问答主持人的细微表现力。International Journal of Human-Computer Studies，62(2)：231–245，2005年。角色与机器人的细微表现力。[41] Alec Radford、Jeff Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。语言模型是无监督的多任务学习者。2019年。[42] Hannah Rashkin、Eric Michael Smith、Margaret Li 和 Y-Lan Boureau。迈向富有同理心的开放域对话模型：一种新的基准与数据集。载于计算语言学协会第57届年会论文集，第5370–5381页，意大利佛罗伦萨，2019年7月。计算语言学协会。[43] Sahand Sabour、Chujie Zheng 和 Minlie Huang。CEM：常识感知的共情响应生成。CoRR，abs/2109.05739，2021年。[44] Sahand Sabour、Chujie Zheng 和 Minlie Huang。CEM：常识感知的共情响应生成。CoRR，abs/2109.05739，2021年。[45] M. Skowron。情感倾听者：通过对话系统获取情感状态。载于COST 2102 培训学校，2009年。[46] Zhenqiao Song、Xiaoqing Zheng、Lu Liu、Mu Xu 和 Xuanjing Huang。生成带有特定情绪的对话回复。载于计算语言学协会第57届年会论文集，第3685–3695页，意大利佛罗伦萨，2019年7月。计算语言学协会。[47] A. Sudhakar、Bhargav Upadhyay 和 A. Maheswaran。转换“删除、检索、生成”方法，实现可控文本风格迁移。ArXiv，abs/1908 09368，2019年。。[48] Hao Tian、Can Gao、Xinyan Xiao、Hao Liu、Bolei He、Hua Wu、Haifeng Wang 和 Feng Wu。SKEP：通过情感知识增强预训练进行情感分析。载于计算语言学协会第58届年会论文集，第4067–4076页，线上会议，2020年7月。计算语言学协会。[49] Hao Tian、Can Gao、Xinyan Xiao、Hao Liu、Bolei He、Hua Wu、Haifeng Wang 和 Feng Wu。SKEP：通过情感知识增强预训练进行情感分析。载于计算语言学协会第58届年会论文集，第4067–4076页，线上会议，2020年7月。计算语言学协会。[50] Quan Tu、Yanran Li、Jianwei Cui、Bin Wang、Ji-Rong Wen 和 Rui Yan。MISC：一种整合COMET的混合策略感知模型，用于情感支持对话。CoRR，abs/2203.13560，2022年。[51] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Ł ukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的一切。载于 I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编，《神经信息处理系统进展》第30卷。Curran Associates, Inc.，2017年。[52] Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Remi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。Transformers：自然语言处理领域的最先进技术。载于2020年自然语言处理实证方法会议论文集：系统演示，第38–45页，线上会议，2020年10月。计算语言学协会。[53] Xing Wu、Tao Zhang、Liangjun Zang、Jizhong Han 和 Songlin Hu。掩码与填充：将掩码语言模型应用于情感迁移。载于IJCAI，2019年。[54] Jingjing Xu、X. Sun、Qi Zeng、Xuancheng Ren、X. Zhang、Houfeng Wang 和 W. Li。无配对的情感到情感翻译：一种循环强化学习方法。载于ACL，2018年。[55] Yi Zhang、Jingjing Xu、Pengcheng Yang 和 X. Sun。利用情感记忆，在无平行数据的情况下进行情感修改。载于EMNLP，2018年。[56] Yizhe Zhang、Michel Galley、Jianfeng Gao、Zhe Gan、Xiujun Li、Chris Brockett 和 Bill Dolan。通过对抗性信息最大化生成信息丰富且多样的对话回复。载于 S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett 编，《神经信息处理系统进展》第31卷。Curran Associates, Inc.，2018年。[57] Yizhe Zhang、Siqi Sun、Michel Galley、Yen-Chun Chen、Chris Brockett、Xiang Gao、Ji anfeng Gao、Jingjing Liu 和 Bill Dolan。DIALOGPT：用于对话响应生成的大规模生成式预训练。载于计算语言学协会第58届年会论文集

计算语言学协会会议：系统演示，第270–278页，线上举行，2020年7月。计算语言学协会。[58] Peixiang Zhong、D. Wang 和 C. Miao。采用偏置注意力和加权交叉熵损失的富含情感的神经对话模型。发表于*AAAI*，2019年。[59] Hao Zhou、Minlie Huang、T. Zhang、Xiaoyan Zhu 和 Bing Liu。情感聊天机器：利用内外部记忆进行情感对话生成。发表于*AAAI*，2018年。[60] Xianda Zhou 和 William Yang Wang。Mojitalk：大规模生成情绪回应。发表于计算语言学协会第*56*届年会论文集（第*1*卷：长篇论文），第1128–1137页，澳大利亚墨尔本，2018年7月。计算语言学协会。