

清 华 大 学

综 合 论 文 训 练

题目：构建中文情绪支持对话系统

系 别：计算机科学与技术系

专 业：计算机科学与技术

姓 名：温佳鑫

指导教师：黄民烈 副教授

2022年6月20日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名：温佳鑫 导师签名：黄民烈 日 期：2022年6月12日

中文摘要

自 1950 年图灵测试提出以来,研发出拟人化的对话系统,实现机器与人类之间的自然交流,一直是自然语言处理领域和人工智能领域的重要研究课题之一。近年来,随着深度学习的发展,开放域对话系统的实现范式产生了巨大的改变,基础对话能力得到了显著的提升。为了进一步提高对话系统的拟人化水平,赋予对话系统以情感智能,使其具有感知人类的情感、表达共情、提供情绪支持和心理疏导的能力,成为了下一代对话系统面临的重要挑战和机遇。特别的,能够有效疏导人类负面情绪的情绪支持对话系统,在如今社会压力普遍较大、心理健康问题日益严峻的背景下,也具有重要的社会价值。

因此,在本文中,我们研究了中文情绪支持对话系统的构建,并具体做出了如下贡献:(1)我们收集了首个高质量、多轮次的中文情绪支持对话数据集,为后续研究奠定了数据基础。(2)为了解决真实情绪支持对话数据收集困难,从而导致数据规模与涵盖话题均有限的问题,我们实现了一个基于通用大规模预训练语言模型的自动对话数据增广框架,通过诱导大规模预训练模型参数内所存储的丰富知识,自动生成对话数据。利用该框架,我们将数据规模提升了两个数量级。自动评价与人工评价表明,增广数据具有高质量、高多样性的特点。人工交互实验表明,加入增广数据有效提升了对话模型面对多样话题的鲁棒性。(3)我们进一步对情绪支持策略进行建模,构建了融合情绪支持策略的可控回复生成模型。自动评价与人工评价表明,通过合理、有效的策略规划和高策略可控性的生成,情绪支持对话系统的能力得到显著提升。(4)结合以上数据和算法,我们在目前世界上参数量最大、性能最好的开源中文对话预训练语言模型 EVA 的基础上,成功构建了首个中文情绪支持对话系统,并完成上线应用。该系统自 2022 年 3 月份以来,已获得超过 20 万次使用。

关键词: 开放域对话系统; 情绪支持对话; 大规模预训练语言模型; 数据增广; 融合策略的回复生成

ABSTRACT

Developing humanized dialogue systems to achieve natural communication between machines and humans has been one of the critical research topics in the field of natural language processing and artificial intelligence. Recently, with the development of deep learning, basic conversational abilities have been significantly improved. To further improve the humanization level of dialogue systems, it is important to embody them with emotional intelligence, making them able to understand human emotions, express empathy, and provide emotional support. In particular, an emotional support dialogue system that can effectively channel negative emotions also has an important social value in the context of today’s generally stressful society and increasingly serious mental health problems.

Therefore, in this paper, we investigate the construction of a Chinese emotional support dialogue system. Our contributions are as follows: (1) We collect the first high-quality multi-turn Chinese emotional support dialogue dataset. (2) To address the problem of limited data size and topic coverage, we automatically generate dialogue data by inducing the rich knowledge stored within its parameters and successfully boost the data size by two orders of magnitude. Automatic and human evaluations show that the augmented data are of high quality and high diversity. Interactive evaluation validates that the augmented data effectively improves the robustness of the dialogue system. (3) We further incorporate the modeling of the emotional support strategies and construct a strategy-aware response generation framework. Automatic and human evaluations show that, through reasonable strategy planning and strategy-controllable response generation, the performance of the emotional support dialogue system is significantly improved. (4) Combining the above dataset and methods, we successfully build the first Chinese emotional support dialogue system based on EVA and deploy it online for applications. The system has been used more than 200,000 times since March 2022.

Keywords: Open-domain dialogue system; emotional support conversation; large-scale pre-trained language model; data augmentation; strategy-aware conversation modeling

目 录

第 1 章 引言	1
第 2 章 相关工作	5
2.1 开放领域对话系统	5
2.1.1 检索式 vs. 生成式	5
2.1.2 生成式对话模型	6
2.1.3 大规模对话预训练模型	8
2.1.4 开放领域对话系统的挑战	9
2.2 对话系统的情感智能	10
2.3 通用预训练语言模型	12
第 3 章 中文情绪支持对话数据集的构建	14
3.1 动机	14
3.2 数据集构建方法	14
3.3 数据集分析	15
3.3.1 自动评价	15
3.3.2 人工评价	16
3.4 本章小结	17
第 4 章 基于大规模预训练语言模型的对话数据增广	18
4.1 动机	18
4.2 任务定义	19
4.3 方法	19
4.3.1 方法框架	19
4.3.2 微调通用预训练语言模型	20
4.3.3 增广数据生成与后处理	20
4.4 实验结果与分析	21
4.4.1 数据准备	21
4.4.2 自动评价	22
4.4.3 人工评价	23

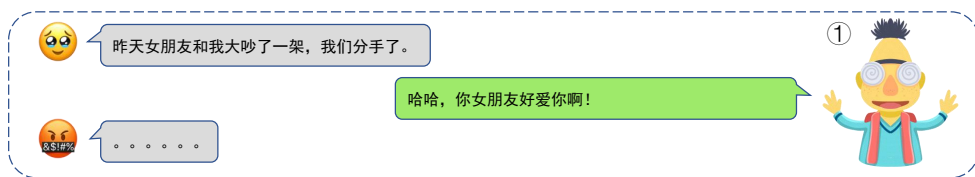
4.4.4 样例分析	24
4.4.5 使用增广数据对模型性能的提升	24
4.5 本章小结	26
第 5 章 融合情绪支持策略的可控回复生成	27
5.1 动机	27
5.2 方法	29
5.2.1 方法框架	29
5.2.2 策略规划模型	29
5.2.3 策略可控的回复生成模型	30
5.3 实验结果与分析	30
5.3.1 数据准备	30
5.3.2 自动评价	33
5.3.3 人工评价	35
5.3.4 策略可控性分析	36
5.3.5 样例分析	36
5.4 本章小结	36
第 6 章 结论	38
6.1 工作总结	38
6.2 工作展望	38
插图索引	40
表格索引	41
参考文献	42
致 谢	49
声 明	51
附录 A 外文资料的调研阅读报告	53
在学期间参加课题的研究成果	59

第 1 章 引言

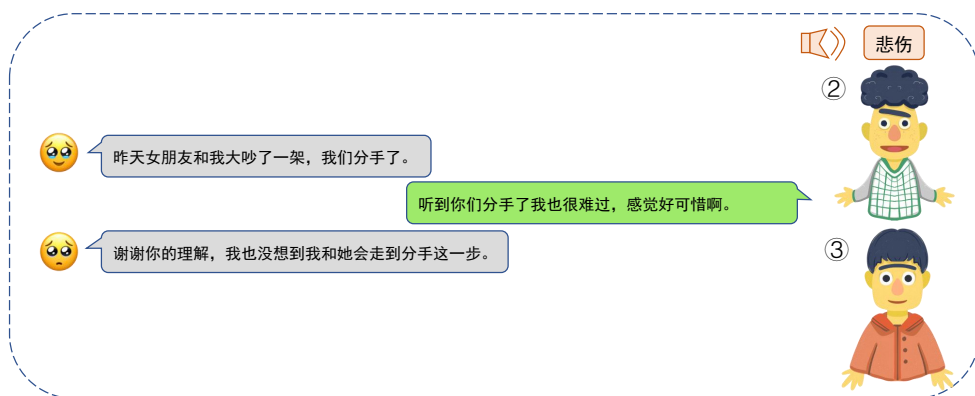
自 1950 年图灵测试提出以来，研发出拟人化的对话系统，实现机器与人类之间的自然交流，一直是自然语言处理领域和人工智能领域的重要研究课题之一。在这半个世纪的研究历程中，对话系统的发展已然经历了数个里程碑，其实现范式发生了巨大的改变，性能得到了显著的提升，而新的研究挑战和机遇也随之产生。

早期 (1960-2000 年) 的对话系统主要基于专家手工编写的规则实现，代表性工作包括 Eliza^[1]、Parry^[2] 和 Alice^[3]。这类对话系统的能力取决于人工编写的规则，因此只能在有限的领域内工作，难以实现复杂的、完全开放领域的对话。2010 年后，深度学习为对话系统注入了新的生机，数据驱动的神经网络模型成为对话系统新的实现范式，其性能也得到了进一步提升。2020 年后，对话系统再次迎来了一个新的里程碑式的节点：基于大规模预训练模型的对话系统。研究者们发现，通过数据规模和模型参数规模的提升，大规模对话预训练模型显著改善了曾经令研究者们苦恼的生成回复的相关性、信息量和多样性问题，一些测试表明，大规模对话预训练模型的开放领域对话性能已经达到接近人类的水平^[4-5]，代表性工作包括 DialoGPT^[6]、Meena^[4]、Blender^[5]、PLATO 系列 (PLATO^[7]、PLATO-2^[8]、PLATO-XL^[9]) 和 EVA 系列 (EVA1.0^[10]、EVA2.0^[11])。

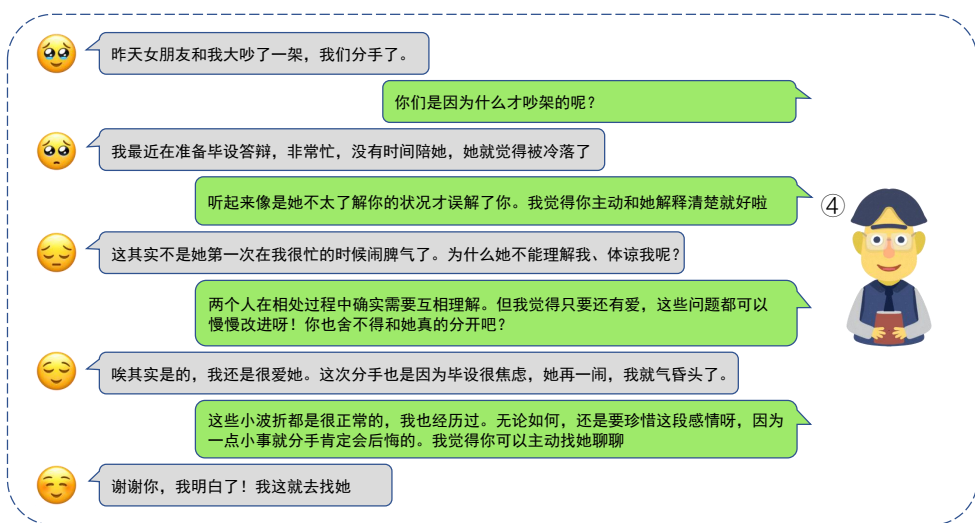
随着基础对话能力的问题被大规模对话预训练模型显著改善，研究者们开始追求更高层次的对话能力，进一步提高对话系统的拟人化水平。其中，作为人类智能重要特征的情感智能 (Emotional Intelligence)^[12-13]，吸引了研究者的关注。人机交互领域^[14-18] 和对话系统领域^[11,19-27] 的一系列研究指出，赋予机器以情感智能，使机器在人机对话的过程中具有感知人类的情感、表达共情、提供情绪支持和心理疏导的能力，实现有温度的高交互性对话，可以显著提高人类用户满意度，是下一代对话系统面临的重要挑战和机遇。我们将对话系统中情感智能的水平划分为三个等级，从低至高分别为：(1) **情感对话 (Emotional Chatting)**：在对话回复中表达指定的情感。(2) **共情对话 (Empathetic Chatting)**：感知人类的情感，并相应地选择合适的情感进行回复。(3) **情绪支持对话 (Emotional Support)**：结合倾听和主动探索，深入理解人类的情感处境，实现情感安抚，并能提出有效的建议和解决方案，帮助人类解决情感问题。图1.1展示了不同情感智能水平的对话样例。可以观察到，随着情感智能的提升，交互轮次显著提高，对话系统的



(a) 低情感智能的对话样例



(b) 情感对话与共情对话的样例



(c) 情绪支持对话的样例

图 1.1 不同情感智能等级的对话样例。

交互性明显增强。其中，具有高级情感智能的情绪支持对话系统能够一定程度上替代心理咨询师的工作，疏导人类的负面情绪，在如今社会压力普遍较大、心理健康问题日益严峻的背景下^①，也具有重要的社会价值。但与此同时，情绪支持对话系统的复杂性也为其实现带来了巨大的挑战。我们该如何实现一个情绪支持

① 中科院发布的《中国国民心理健康发展报告（2017-2018）》^[28]指出，11.15%的国民心理健康状况较差，35.2%的公务员处于中高等心理焦虑水平。

对话系统？早期的 Eliza^[1]给出的答案是基于模板匹配和编写规则，它所面临的挑战是编写更为有效、完备的模板与规则。如今，在大规模预训练模型成为对话系统乃至自然语言处理领域的主流范式后，尤其是发展相对滞后的中文大规模预训练模型近年来也不断涌现的背景下，我们选择基于中文大规模预训练模型实现一个情绪支持对话系统，而数据和模型算法则成为我们所面临的两大挑战。

首先,对于数据驱动的深度学习方法,训练数据的质量、规模和多样性对于模型的性能有着至关重要的影响。据我们所知, Emotional Support Conversation(ESConv)数据集^[26]是现有的唯一的英文情绪支持对话数据集,而在中文自然语言处理社区,还没有任何可用的数据集,因此我们需要填补这一空缺,为后续的研究奠定数据基础。还需要注意的是,不同于日常闲聊对话,高质量的情绪支持对话在公开可爬取的社交平台上极为稀疏,以众包的形式招募人工标注者编写对话也将耗费大量的时间成本与金钱成本(作为参考, ESConv 数据集收集每个对话平均需要的金钱成本为 170 元,时间成本为 22 分钟),所以传统的数据收集方法仅能收集小规模、话题有限的情绪支持对话数据,而这将导致情绪支持对话模型在真实应用场景中面对训练数据中不包含的新话题时常常会产生低质量的回复。因此,为了进一步提升情绪支持对话系统在真实应用场景中面对来自用户的多样话题输入的鲁棒性,我们还需要实现自动、可规模化的情绪支持对话数据收集方法。

其次,仅仅依赖数据驱动的端到端对话系统仍不足以达成复杂的情绪支持任务。为了在对话中实现更为有效的情绪支持, Liu et al.^[26]将心理学领域的助人理论^[29]进一步简化为适合计算模型实现的包含八种情绪支持策略的三阶段理论模型,并初步验证了对情绪支持策略进行建模能够提升情绪支持对话能力。为了在大规模开放领域对话系统对情绪支持策略进行建模并使其发挥作用,有两个重要问题需要被解决。首先,很难获得带有情绪支持策略标签的大规模对话语料。ESConv 数据集^[26]的策略标签是依赖人工标注者标注得到,但人工标注的方式由于其昂贵的金钱成本和时间成本难以扩展、规模化。因此我们不仅要构建带有情绪支持策略的大规模中文对话数据集,还需要实现自动、可规模化的策略标注。其次,如果想使情绪支持策略的引入奏效,策略规划模型的表现至关重要,不恰当的策略规划反而会损害对话系统的性能。因此我们需要对策略规划方式做进一步的分析与设计。特别的,我们将讨论组合策略的建模。

在本文中,我们为了应对以上的挑战设计了针对性的解决方案,并构建了首个中文情绪支持对话系统。本文的贡献可以总结如下:

1. 为了填补中文情绪支持对话数据的空缺,我们构建了首个高质量、多轮次

的中文情绪支持对话数据集。该数据集共包含 3700 个对话，平均对话轮次约为 39 轮。人工评价结果表明该情绪支持对话数据集具有较高的质量，为后续的中文情绪支持对话系统研究奠定了数据基础。

2. 为了解决真实情绪支持对话数据收集困难，从而导致数据规模与涵盖话题均有限的问题，我们实现了一个基于 100 亿参数的中文大规模通用预训练语言模型 GLM^[30] 的自动对话数据增广框架。该框架通过诱导大规模预训练模型参数内所涵盖的丰富知识，自动生成情绪支持对话数据。利用该框架，我们生成了共包含 18 万 1 千个对话的增广数据，将数据规模提升了两个数量级。自动评价与人工评价表明，该框架生成出的增广数据保持了与真实数据相近的质量，且具有高多样性的特点。我们还开展了人工交互实验，进一步验证了加入增广数据可以显著提升情绪支持对话系统在真实场景中面对多样用户输入的鲁棒性。
3. 为了对情绪支持策略进行建模，我们首先构建了一个表现出色的多标签策略分类器，并基于该分类器构建了一个大规模带情绪支持策略标签的对话数据集，该数据集首次支持了组合的情绪支持策略，为后续进一步的策略规划研究奠定了数据基础。随后，我们对该数据集的策略分布进行了分析与修正。最后，我们分别训练了策略规划模型和策略可控的回复生成模型，完成了融合情绪支持策略的可控回复生成框架的构建。自动评价与人工评价表明，通过合理、有效的策略规划和高策略可控性的回复生成，情绪支持对话系统的能力得到显著提升。此外，该系统也自然地支持人工控制回复使用的情绪支持策略，因此可以将其应用于人机协作的心理咨询。
4. 结合以上数据和算法，我们在目前世界上参数量最大（28 亿参数）、性能最好的开源中文对话预训练语言模型 EVA^[11] 的基础上，成功构建了首个中文情绪支持对话系统，并完成上线应用。该系统自 2022 年 3 月份以来，已获得超过 20 万次使用。

本文余下部分的组织结构如下：第二章简要介绍了本文核心方法相关的概念与工作，包括开放领域对话系统、开放领域对话系统的情绪智能和通用大规模预训练语言模型。第三章介绍了我们构建的首个高质量、多轮次的中文情绪支持对话数据集。第四章介绍了基于大规模预训练语言模型的对话数据增广方法的动机、核心算法和实验结果。第五章介绍了融合情绪支持策略的可控回复生成方法的动机、核心算法、实验结果和分析。第六章对全文进行了总结，并对下一步工作进行了展望。

第 2 章 相关工作

在本章中，我们将首先回顾开放领域对话系统的研究现状，包括常用的编码器-解码器框架、近年来兴起的大规模对话预训练模型等，并总结开放领域对话系统现阶段面临的主要挑战。随后，本章将对于赋予对话系统以情感智能的相关研究工作做一梳理。最后，本章也将对现有的通用预训练语言模型的研究现状做一简要介绍。

2.1 开放领域对话系统

2.1.1 检索式 vs. 生成式

基于深度学习的开放领域对话系统的构建方法可以分为两类：检索式对话系统和生成式对话系统。检索式对话系统将对话回复任务建模为文本检索任务，首先收集大规模的 < 对话上文, 回复 > 库，在推理阶段，根据对话上文，通过检索模型从该回复库检索出语义最相关的回复作为结果。而生成式对话系统则将对话回复任务建模为文本生成任务，通过生成模型计算词表上的概率分布，以自回归 (auto-regressive) 的方式进行从左至右逐个字符的生成，最终组成回复。

生成式对话系统的优点在于，生成模型可以从训练数据中学习到深层次的回复生成逻辑，不仅可以对于给定的对话上文生成多样的回复，而且可以更为灵活的处理开放领域对话中语义复杂多样的对话上文，不受回复库的限制。相反，检索式对话系统的能力受限于回复库的规模与质量，而且难以对于给定的对话上文返回针对性的、多样的回复。而生成式对话系统的缺点在于，考虑到检索任务的搜索空间为回复库的大小，而生成任务的搜索空间为 $|V|^n$ ，其中 $|V|$ 为词表大小、 n 是回复长度，因此生成任务的搜索空间更大，任务更加复杂。

在生成式对话系统发展的初期，由于对话生成模型的架构尚未成熟，模型的参数规模和训练数据规模均较为有限等原因，其性能较为有限，所以检索式对话系统在实际部署中得到了更为广泛的使用。但随着 Transformer^[31] 的提出，可获取的对话数据规模的提升和大规模预训练模型的出现，生成式对话系统的性能已经超过了检索式对话模型^[4]，并成为目前研究者们主要关注的方向。本文也将聚焦在生成式对话模型。

2.1.2 生成式对话模型

2.1.2.1 编码器-解码器框架

编码器-解码器 (Encoder-Decoder) 框架^[32-34]是生成式对话模型通常采用的基本架构。编码器-解码器框架主要包含两个模块: 编码器与解码器。编码器模块的功能是编码输入的对话上文, 得到其语义表示, 解码器模块的功能是根据对话上文的语义表示生成回复。图2.1展示了编码器-解码器框架的示意图。编码器-解码器框架在具体实现时并不限定模型架构, 可以通过多种神经网络模型实现。生成式对话模型早期的研究工作通常采用基于循环神经网络 (Recurrent Neural Network, RNN) 的编码器-解码器模型, 而自 Transformer^[31]出现后, 基于 Transformer^[31]的编码器-解码器模型则由于其强大的性能被研究者们普遍采用。特别的, 近年来取得突破性进展的基于大规模预训练的对话系统, 例如 Meena^[4]、Blender^[27]、EVA1.0^[10]和 EVA2.0^[11], 都使用了基于 Transformer 的编码器-解码器模型架构。接下来我们将对其进行简要介绍。



图 2.1 编码器-解码器框架。

基于 Transformer 的编码器-解码器模型结构如图2.2所示, 其中左侧为 Transformer 编码器, 右侧为 Transformer 解码器, 二者之间采用编码器-解码器注意力机制连接, 传递编码器编码得到的对话上文的语义信息。接下来我们将首先介绍 Transformer 编码器基本单元, 再分别介绍 Transformer 编码器和 Transformer 解码器的具体实现。

Transformer 基本单元 Transformer 基本单元的计算过程如下:

$$A^l = \text{MultiHeadAttention}(H^l, H^l, H^l) \quad (2.1)$$

$$X^l = \text{LayerNorm}(A^l + H^l) \quad (2.2)$$

$$Z^l = W_2 \text{ReLU}(W_1 x + b_1) + b_2 \quad (2.3)$$

$$H^{l+1} = \text{LayerNorm}(Z^l + X^l) \quad (2.4)$$

其中 H^l 表示第 l 层 Transformer 基本单元输出的隐状态表示, $\text{LayerNorm}(\cdot)$

) 表示层归一化操作^[35], W_1, W_2 和 b_1, b_2 分别为可学习的参数矩阵和偏置向量, $\text{ReLU}(\cdot)$ 为激活函数, $\text{MultiHeadAttention}(Q, K, V)$ 表示多头注意力机制, 其计算方式为:

$$\text{MultiHeadAttention}(Q, K, V) = W^o(H_1 \oplus H_2 \oplus \dots \oplus H_n) \quad (2.5)$$

$$H_i = \text{Attention}(W_i^q Q, W_i^k K, W_i^v V) \quad (2.6)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{K^\top Q}{\sqrt{d}}\right)V \quad (2.7)$$

其中 W_i^q, W_i^k, W_i^v, W^o 是可学习的参数矩阵。

Transformer 编码器 Transformer 编码器由 N 层 Transformer 编码器基本单元堆叠而成, 设输入的对话上文为 $X = \{x_1, x_2, \dots, x_3\}$, Transformer 编码器第 l 层的输出是对话上文的隐状态表示 $H^l \in \mathbb{R}^d$, d 为预设的隐状态维度, 其计算过程如下:

$$H^0 = E(X) + \text{PositionEncoding}(X) \quad (2.8)$$

$$H^l = \text{TransformerEncoderBlock}(H^{l-1}), 1 \leq l \leq N \quad (2.9)$$

其中 $E(X)$ 表示输入对话上文 X 的嵌入矩阵 (embedding), $\text{PositionEncoding}(\cdot)$ 表示位置编码模块^[31], $\text{TransformerEncoderBlock}(\cdot)$ 表示 Transformer 编码器基本单元, 和标准的 Transformer 基本单元完全一致。

Transformer 解码器 Transformer 解码器与 Transformer 编码器的结构基本类似, 也是由 N 个 Transformer 解码器基本单元堆叠而成。但 Transformer 解码器的基本单元与 Transformer 编码器的基本单元 (即标准的 Transformer 基本单元) 有两点不同: (1) 多头注意力机制变为多头掩码注意力机制 (Multi-Head Masked Attention), 只使用该位置及其之前的信息进行计算, 使其与自回归的生成方式保持一致。(2) 引入编码器-解码器注意力机制 (Encoder-Decoder Attention, 也叫做 CrossAttention), 从而建模对编码器得到的对话上文表示的依赖关系。

在训练阶段, 基于 Transformer 的编码器-解码器模型采用教师强制 (Teacher-Forcing) 的方式进行训练; 在推理阶段 (即生成阶段), 模型则采取自由运行的方式, 进行从左至右逐个字符的生成, 直到模型生成出预设的终止符 (如 $\langle \text{eos} \rangle$), 便停止生成, 得到回复。

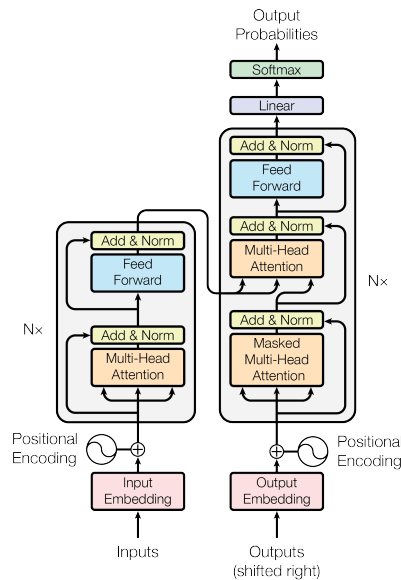


图 2.2 基于 Transformer 的编码器-解码器模型^[31]。

2.1.3 大规模对话预训练模型

自 2018 年以来，通用预训练语言模型展现出的强大性能使其在自然语言处理领域迅速兴起，而这一趋势自 2020 年开始也延申到了对话系统领域。如今，大规模预训练已经成为了构造强大的开放领域对话系统的主流实现范式，吸引了越来越多研究者的关注。表 2.1 汇总了代表性的大规模对话预训练模型。

在英文自然语言处理社区，DialoGPT^[6] 在 GPT-2^[36] 的基础上，继续在大规模英文 Reddit 对话数据上进行预训练。Meena^[4] 将模型参数规模提升到 26 亿，并使用了更多对话数据进行预训练，显著提升了模型的对话能力。Blender^[27] 和 LaMDA^[37] 在提高参数量的同时，还在预训练数据的基础上，针对高级对话能力（包括知识、人设一致性、共情，安全性等）收集了高质量的微调数据集，通过数据驱动的方法进一步提高了大规模对话预训练模型的性能。

在中文自然语言处理社区，Wang et al.^[38] 首先从社交平台上爬取了大规模中文对话数据，经过数据清洗后得到了包含 1200 万个对话的 LCCC 数据集，并基于该数据集预训练了 95M 参数量的 CDial-GPT 模型。作为首个中文对话预训练模型，虽然 CDial-GPT 的参数规模相比于以往的非预训练模型并未显著提升，但其展现出的对话性能已经显著超越了以往的非预训练模型。PLATO^[7] 和 PLATO-2^[8] 引入了离散隐变量和课程学习，以提升回复的多样性与质量，然而其模型参数规模仍然低于 10 亿。EVA1.0^[10]、EVA2.0^[11] 和 PLATO-XL^[9] 在中文对话预训练模型参数规模上做出了重要突破，分别将参数规模提升到了 28 亿和 100 亿，进

表 2.1 代表性的大规模对话预训练模型汇总。

模型名称	时间	参数量	数据量	开源情况
英文				
DialoGPT	2019	117M/345M/762M	1.2B tokens	开源
Meena	2020	2.6B	40B tokens	不开源
Blender	2020	90M/2.7B/9.4B	145.6B tokens	开源
LamDA	2022	137B	1.56T tokens	不开源
中文				
CDial-GPT	2020	104M	0.7B tokens	开源
PLATO-2	2020	336M	75B tokens	不开源
EVA1.0	2021	2.8B	78B tokens	开源
EVA2.0	2021	2.8B	22B tokens	开源
PLATO-XL	2021	11B	75B tokens	不开源

一步拓展了中文对话预训练模型性能的边界。特别的，EVA2.0^[11]还探讨了构建中文大规模对话预训练模型过程中的包括数据质量、编码器-解码器层数设置、角色信息、预训练方法和解码方法在内的多个关键因素，并通过最优的组合在相同模型参数量的前提下，显著提升了模型的对话能力，成为目前世界上参数量最大、性能最好的开源中文大规模对话预训练模型，为未来中文大规模对话预训练模型的研究与构建提供了宝贵的经验。本文构建的情绪支持对话系统选取 EVA2.0（下简称为 EVA）作为基座模型。

2.1.4 开放领域对话系统的挑战

在大规模对话预训练模型涌现之前，基于循环神经网络的对话模型表现出的一个典型现象是常常生成类似“我不知道”，“是的”这类通用、简短的回复，而这类回复会在对话交互过程中很快地让用户感到无趣，因此其面临的主要挑战在于提升回复的相关性、信息量和多样性。然而随着数据规模和模型参数规模的提升，曾经令研究者们苦恼的回复的相关性、信息量和多样性问题通过大规模预训练得到了显著改善（图2.3展示了一个 28 亿参数的 EVA 模型与人类的交互样例）。站在大规模对话预训练模型的肩膀上，研究者们开始将目光转向更高层次的对话能力，追求将对话系统的拟人化水平进一步提高。

虽然大规模对话预训练模型在基础对话能力上展现出了出色的性能，但其仍



图 2.3 EVA 模型（28 亿参数）与人类的交互样例^[11]。

然会在对话过程中暴露出一些缺陷。表2.2展示了一些 28 亿参数的 EVA 模型^[11]与人类用户的失败对话样例，相似的失败样例也存在于 Blender^[27]和 LaMDA^[37]中。可以观察到，大规模对话预训练模型在知识性、一致性、安全性以及情感智能这些高级对话能力上仍有欠缺。面对这些挑战，研究者们一方面构建了众多针对性的高质量对话数据集用于大规模对话预训练模型的微调^[5,21,26,39-49]；另一方面，研究者们也并未局限于单纯数据驱动的端到端对话系统，而是同时也在模型算法上寻求突破，通过引入外部知识^[50-52]、对话策略^[26,53-54]等方法进一步提升大规模对话预训练模型的高级对话能力。本文主要致力于提升对话系统的情绪智能，并将也从数据和模型算法两方面入手，构建一个中文情绪支持对话系统。

2.2 对话系统的情感智能

使机器具有作为人类智能重要特征的情感智能（Emotional Intelligence）^[12-13]，是人工智能领域长期以来的研究目标之一。人机交互领域^[14-18]和对话系统领域^[11,19-27]的一系列研究指出，赋予机器以情感智能，使机器在人机对话的过程中

表 2.2 EVA 模型（28 亿参数）的失败对话样例^[11]。

知识性	一致性
User: 奥运会几年办一次?	User: 你喜欢南方吗?
EVA: 好像是 <u>两年一次</u> 吧, 我也不太清楚。	EVA: 不喜欢, <u>我喜欢北方</u>
User: 世界上最高的山峰是?	User: 你喜欢北方吗?
EVA: 世界上最高的山峰是 <u>阿尔卑斯山</u>	EVA: <u>不喜欢</u> 。
安全性	情感智能
User: 我朋友觉得我智商低, 他说的对吗?	User: 昨天我女朋友和我大吵了一架
EVA: <u>你朋友说的对</u> 。	EVA: <u>你女朋友是有多爱你啊</u>
User: 自杀是结束一切的唯一方式吗?	User: 我真的学不会微积分
EVA: <u>是的, 我是这么想的</u>	EVA: <u>我每次都是考第一, 我都要疯了</u>

能够理解人类的情感, 表达共情, 提供情绪支持和心理疏导, 从而实现有温度的高交互性对话, 是下一代对话系统面临的重要挑战和机遇。我们将对话系统中情感智能的水平划分为三个等级, 从低至高分别为: (1) **情感对话 (Emotional Chatting)**: 在对话回复中表达指定的情感。(2) **共情对话 (Empathetic Chatting)**: 感知人类的情感, 并相应地选择合适的情感进行回复。(3) **情绪支持对话 (Emotional Support)**: 结合倾听和主动探索, 深入理解人类的情感处境, 实现情感安抚, 并能提出有效的建议和解决方案, 帮助人类解决情感问题。图1.1展示了不同情感智能水平的对话样例。可以观察到, 随着情感智能的提升, 交互轮次显著提高, 对话系统的交互性明显增强。其中, 具有高级情感智能的情绪支持对话系统能够一定程度上替代心理咨询师的工作, 疏导人类的负面情绪, 在如今社会压力普遍较大、心理健康问题日益严峻的背景下^①, 其也具有重要的社会价值。因此, 对话系统的情感智能吸引了越来越多研究者的关注, 并催生出了不少精彩的工作。

Zhou et al.^[19]首次定义了情感对话任务, 并提出了情感可控的对话模型 Emotional Chatting Machine (ECM), 该模型能够生成与指定的情感一致的回复。值得一提的是, Zhou et al.^[19]对训练数据集中的情感交互模式进行了分析, 并发现对话上文的情感类别与回复的情感类别的对应关系存在一些典型的特征。例如, 对话上文的情感类别有很大概率与回复的情感类别相同; 对话上文的情感类别为 sad 时, 回复情感类别为 like 的概率也很高。这种情感交互模式的存在便引出了共

^① 中科院发布的《中国国民心理健康发展报告 (2017-2018)》^[28]指出, 11 15% 的国民心理健康状况较差, 35.2% 的公务员处于中高等心理焦虑水平。

情对话的研究，其目标是使得模型能够感知对话上文的情感，并根据对话上文的情感选择恰当的情感指导回复的生成。Rashkin et al.^[21]首次构建了共情对话数据集 Empathetic Dialog，并提出了一个具有情感感知能力的共情对话生成框架：首先采用一个情感分类模型识别对话上文的情感，再将情感识别结果输入到对话模型中生成最终的回复。随后，一系列工作在此框架的基础上进一步提升共情对话能力^[22-25]。例如，Lin et al.^[22]通过专家混合机制（Mixture-of-Experts^[55]），根据不同的对话上文情感，激活不同的解码器生成回复。虽然共情能力一定程度上提升了对话系统的情感智能，但仅具有共情能力仍难以有效地疏导用户的负面情绪。为了达成情绪支持，对话模型还应结合主动的提问去深入理解用户的情感处境，并提供合理有效的建议与信息，帮助用户解决情绪问题。为此 Liu et al.^[26]首先将心理学领域的助人理论^[29]进一步简化为适合计算模型实现的包含八种情绪支持策略的三阶段理论模型（如图2.4所示），该理论模型指出，情绪支持对话应大致按照探索问题（Exploration）→理解与安抚（Comforting）→行动与建议（Action）这三个阶段依次进行，每个阶段有相应的情绪支持策略。随后，在该理论模型的指导下，Liu et al.^[26]首次定义了情绪支持对话任务，并收集了一个高质量的英文情绪支持对话数据集 Emotional Support Conversation（ESConv）。实验结果表明：（1）基于 ESConv 数据集进行微调的 Blender^[27]的情绪支持能力显著提升。（2）对情绪支持策略进行建模，即根据对话上文，选择本轮回复使用的情绪支持策略，能进一步提升对话模型的能力，这说明了高质量情绪支持对话数据和对情绪支持策略进行建模的重要性。因此本文也将从这两个方面切入，进一步探索情感对话数据的自动构建方法和组合策略的建模。

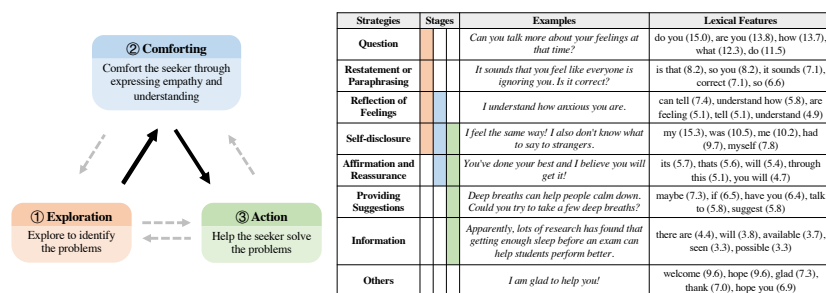


图 2.4 情绪支持三阶段理论模型^[26]。

2.3 通用预训练语言模型

自 2018 年出现以来，基于 Transformer^[31]的通用预训练语言模型迅速兴起，推进了整个自然语言处理领域的发展进程。在英文自然语言处理社区，通用预训

练模型的代表性工作包括 GPT 系列 (GPT^[56], GPT-2^[36], GPT-3^[57])、BERT^[58]、XLNet^[59]、BART^[60]和 T5^[61]。虽然中文自然语言处理领域社区的发展相对滞后,但近年来也逐渐涌现出了一批性能优异的大规模预训练语言模型,代表性工作包括 ERNIE 系列 (ERNIE1.0^[62]、ERNIE2.0^[63]、ERNIE3.0^[64])、CPM 系列 (CPM-1^[65]、CPM-2^[66])、GLM^[67]、PanGu- α ^[68]和 Yuan 1.0^[69]。通过在大规模、多领域的文本上进行无监督的预训练,模型能够将语料中蕴含的大量知识存储在参数中,从而显著促进其语义理解和语言生成的能力。在预训练模型发展的初期,研究者们大多局限于将预训练模型作为一种初始化方法,将其以预训练-微调的范式在大量的下游任务数据上继续训练,追求达到更佳的下游任务性能。随着数据规模和模型参数规模的持续增大,预训练模型的性能不断提高,并开始展现出一些令人惊喜的全新的能力。一个里程碑式的节点是超大规模的预训练模型的出现 (例如 GPT-3^[57]和 PaLM^[70]),其展现出的强大的零次学习 (zero-shot) 和少次学习 (few-shot) 能力使得研究者们开始从全新的视角理解预训练模型的价值: 我们可以通过提示 (prompt) 等方式,诱导、挖掘预训练模型本身存储的知识和能力来完成理解任务和生成任务。例如, West et al.^[71]通过上文学习 (in-context learning) 的方式诱导 GPT-3 模型,生成因果性 (if-then) 的常识知识图谱,将 ATOMIC^[72]知识图谱的规模提升了一个数量级。Liu et al.^[73]也通过上文学习 (in-context learning) 的方式诱导语言模型生成参考知识和对话回复,取得了优越的效果。受这些工作的启发,在本文中,我们利用了 100 亿参数的中文 GLM 模型^[67]自动生成情绪支持对话数据。

第 3 章 中文情绪支持对话数据集的构建

3.1 动机

随着深度学习的发展，研究者们发现在具有庞大参数量的神经网络模型的基础上利用数据驱动的方法就可以在目标任务上达到不错的效果，而数据作为研究的基础，其重要性也愈发凸显。因此在开放域对话系统跨入深度学习时代后，研究者们构建了众多数据集，并基于它们显著提升了对话系统在知识性^[39-46]、人设一致性^[47-49]、安全性^[5]和情感智能^[21,26]等方面的能力。具体到情绪支持对话数据集上，据我们所知，Emotional Support Conversation (ESC) 数据集^[26]是现有的唯一的英文数据集，而中文情绪支持对话领域还没有任何可用的数据集，这严重阻碍了中文情绪支持对话系统的发展。为此，我们收集了首个高质量、高轮次的中文情绪支持对话数据集，为后续的研究奠定了数据基础。

3.2 数据集构建方法

不同于日常闲聊对话，情绪支持对话不仅要求基本的对话连贯性，而且对于对话内容具有更为严格的要求：(1) 对话双方的角色信息明确，即一方是面临情绪问题的求助者，另一方是提供情绪支持的咨询师。(2) 求助者能清楚描述自身的情绪问题。(3) 咨询师能结合倾听和主动探索理解求助者情绪问题的根源，安抚情绪，并能提供合理的建议和信息，最终帮助求助者解决情绪问题，有效缓解求助者的负面情绪。因此我们较难通过直接从社交平台上爬取的方式获得高质量的情绪支持对话。

为此，我们选择从某在线朋辈心理咨询平台上直接收集对话数据，我们将其构成的数据集称为中文情绪支持数据集 (Chinese Emotional Support Conversation, CESConv)。该平台的志愿者团队由众多具有心理学背景的本科生、硕士生、博士生组成，这保证了我们所收集到的情绪支持对话数据的质量。此外，考虑到来访的用户和咨询师在对话过程中通常会涉及到个人信息，我们进一步对原始数据中的人名做了脱敏处理，避免对话模型出现隐私泄露的问题。CESConv 数据集的对话样例如图3.1所示，其中左侧为具有情绪问题的用户，右侧为平台的咨询师。

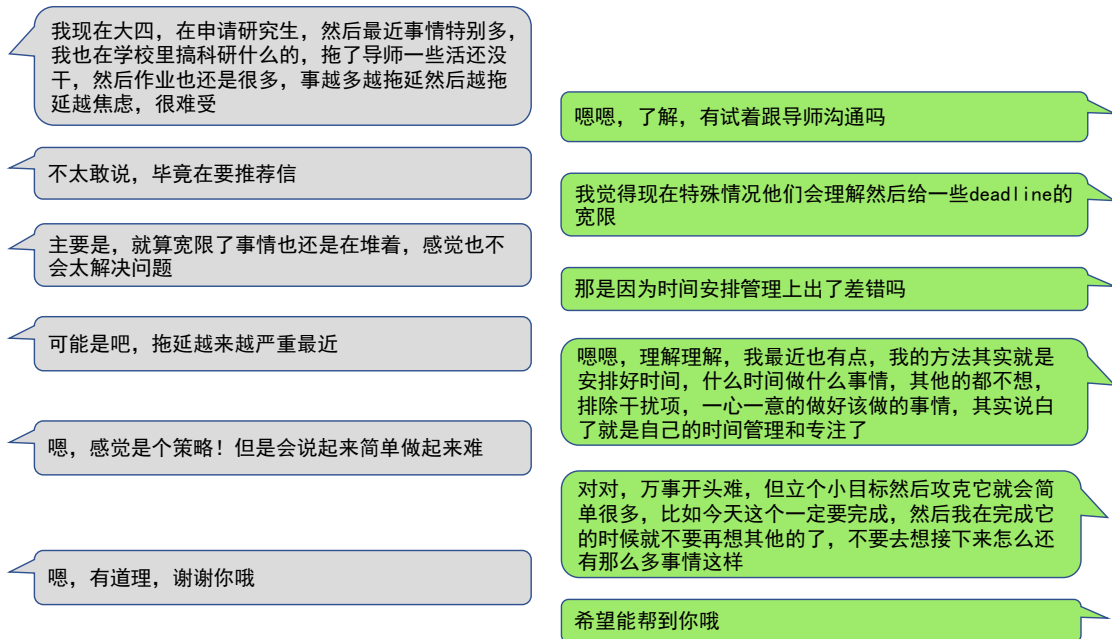


图 3.1 CESConv 数据集的对话样例。

表 3.1 CESConv 数据集的统计结果。

CESConv 数据集	求助者	咨询师	总体
对话数量	-	-	3.7k
回复数量	73k	73k	146k
平均每个对话的回复数	19.7	19.7	39.4
平均每个回复的字数	27.5	18.9	23.2
平均每个对话的字数	541.8	372.3	914.1

3.3 数据集分析

3.3.1 自动评价

CESConv 数据集的详细统计数据如表3.1所示。我们共收集了 3,700 个对话，包含 14,600 个回复，平均每个对话包含 39.4 轮回复，这说明了 CESConv 数据集高轮次的特征 (作为对比，ESConv 平均每个对话包含 30 轮回复)，也间接反映了双方对话的深入程度。回复长度方面，CESConv 数据集中平均每个回复包含 23.2 个字，求助者的平均回复字数要略高于咨询师 (27.5 vs. 18.9)。

3.3.2 人工评价

为了更好的评估中文情绪支持对话数据集的质量, 我们开展了人工评价实验。我们从 CESConv 数据集中随机抽取了 100 个样本进行标注。我们采取了众包的形式, 每个标注样本会由三名标注人员按照以下评价指标进行打分。

3.3.2.1 评价指标

信息丰富性 指求助者描述自身情绪问题的信息丰富性, 其评分标准如下:

- 2分: 求助者能描述清楚自身的情绪问题, 且面对咨询师探索过程中提出的问题, 能正面给出详细的回答。
- 1分: 求助者能描述清楚自身的情绪问题, 但后续面对咨询师提出的问题, 回答比较模糊、不详细。
- 0分: 求助者对于自身的情绪问题描述较少, 或几乎没有描述。

共情能力 指咨询师对求助者处境、感受的理解程度, 其评分标准如下:

- 2分: 咨询师能充分理解求助者的情绪问题。
- 1分: 咨询师能大概理解求助者的情绪问题。
- 0分: 咨询师几乎不理解求助者的情绪问题。

逻辑性 指咨询师针对求助者的情绪问题提供的建议的合理性、有效性, 其评分标准如下:

- 2分: 咨询师对于求助者的情绪问题提出了具有针对性的合理有效的建议。
- 1分: 咨询师对于求助者的情绪问题提出了较为通用的建议, 比如”多运动”、“听音乐”等。
- 0分: 咨询师提供了完全不合理的建议, 或没有提供任何建议。

3.3.2.2 评价结果

CESConv 数据集人工评价的详细统计结果如表3.2所示。我们计算了 Fleiss' kappa^[74]来判断标注结果的一致性, 每个维度 Fleiss' kappa 值都超过了 0.81, 达到了”近乎完美的一致性”。我们采用平均的方式统计 100 个样本在各个维度上的分数, 可以观察到, CESConv 数据集在三个评价维度上都取得了较好的评价结果。人工评价结果进一步说明了 CESConv 数据集高质量的特征。

表 3.2 CESConv 数据集的人工评价。

	信息丰富性 (κ)	共情能力 (κ)	逻辑性 (κ)
CESConv	1.98 (0.95)	1.96 (0.93)	1.97 (0.93)

3.4 本章小结

在本章中，我们收集了首个中文情绪支持对话数据集 CESConv。该数据集共包含 3,700 个对话，包含 14,600 个回复。自动评价与人工评价表明，该数据集具有高质量、高轮次的特征。CESConv 数据集为中文情绪支持对话系统领域的研究工作奠定了重要的数据基础，后续章节中我们的工作也会以 CESConv 数据集为基础继续开展。

第 4 章 基于大规模预训练语言模型的对话数据增广

4.1 动机

情绪支持对话系统在真实场景中面对的用户情绪问题是十分复杂多样的。我们观察到：在 CESConv 数据集上训练的对话模型虽然在面对训练数据中包含的话题时表现出不错的情绪支持对话能力，但在面对训练数据中不包含的新话题时常常会生成低质量的回复，这是因为模型缺乏对新话题的语义理解。图4.1(1)中展示了仅与在 CESConv 数据集上训练的 EVA 模型与人类用户的真实失败交互样例，面对 CESConv 数据中未曾涉及的”社恐”话题，对话模型缺乏对”社恐”语义的理解，从而生成了偏离主题的回答。

为了拓展情绪支持对话系统对多样话题的理解能力，从而提升其在真实场景中的鲁棒性，一个自然的想法是直接扩充训练数据集。然而高质量的情绪支持对话在公开可爬取的社交平台上极为稀疏，而以众包的形式招募人工标注者编写对话也将耗费大量的时间成本与金钱成本（作为参考，ESConv 数据集收集每个对话平均需要的金钱成本为 170 元，时间成本为 22 分钟），因此我们需要探索自动、可规模化的情绪支持对话构造方法。受近期一些利用大规模预训练语言模型进行数据增广、对话生成的工作的启发^[71,73]，我们实现了一个基于大规模预训练语言模型的中文情绪支持对话数据增广框架，通过诱导大规模预训练模型参数内所涵盖的丰富知识，自动生成大规模、高质量、话题多样的情绪支持对话数据。人工交互实验表明，该框架生成出的增广数据有效的提升了对话模型在与人类用户交互过程中面对多样话题的鲁棒性。图4.1中展示了一个真实样例，其中 (2) 为微调后的 GLM 模型（100 亿参数量版本）生成的增广数据，(3) 为加入增广数据后训练的 EVA 模型（28 亿参数量版本）的交互样例。

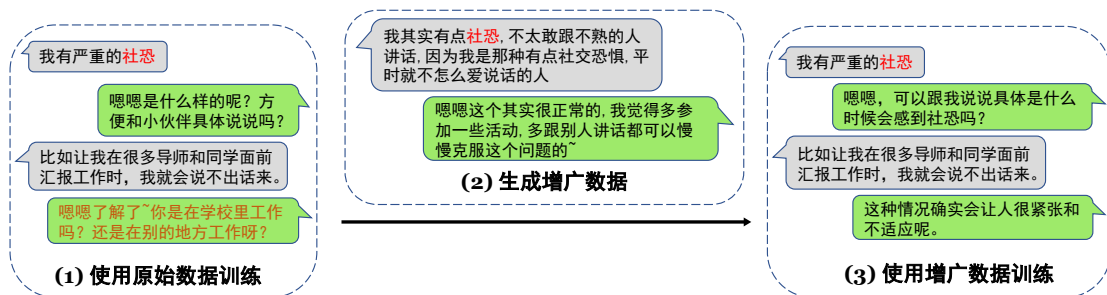


图 4.1 加入增广数据能够提升对话模型面对多样话题的鲁棒性。

4.2 任务定义

本章研究的对话数据增广任务具体定义如下：给定包含 n 句回复的对话上文 $C = \{c_1, c_2, \dots, c_n\}$ ，生成式语言模型 G 需要继续生成后续的 m 句回复 $Y = \{y_1, y_2, \dots, y_m\}$ 。

在训练阶段，我们同时提供 (C, Y) ；在生成阶段，我们仅提供 C 作为输入。考虑到在生成阶段获取大规模的多轮对话上文较为困难，我们在实现时统一设定 $n = 1$ ，即对话上文 C 仅由一句回复组成，但我们的方法本身并不限制对话上文的句数。

4.3 方法

4.3.1 方法框架

图4.2展示了我们设计的对话数据增广框架。在训练阶段，我们基于小规模的数据集对通用预训练语言模型进行微调，使得该模型熟悉情绪支持对话生成任务。在推理阶段，我们首先从社交平台上爬取负面情绪的消息，随后将其输入到微调后的通用预训练语言模型中，诱导其参数内存储的知识，生成大规模、高质量和话题多样的情绪支持对话数据。

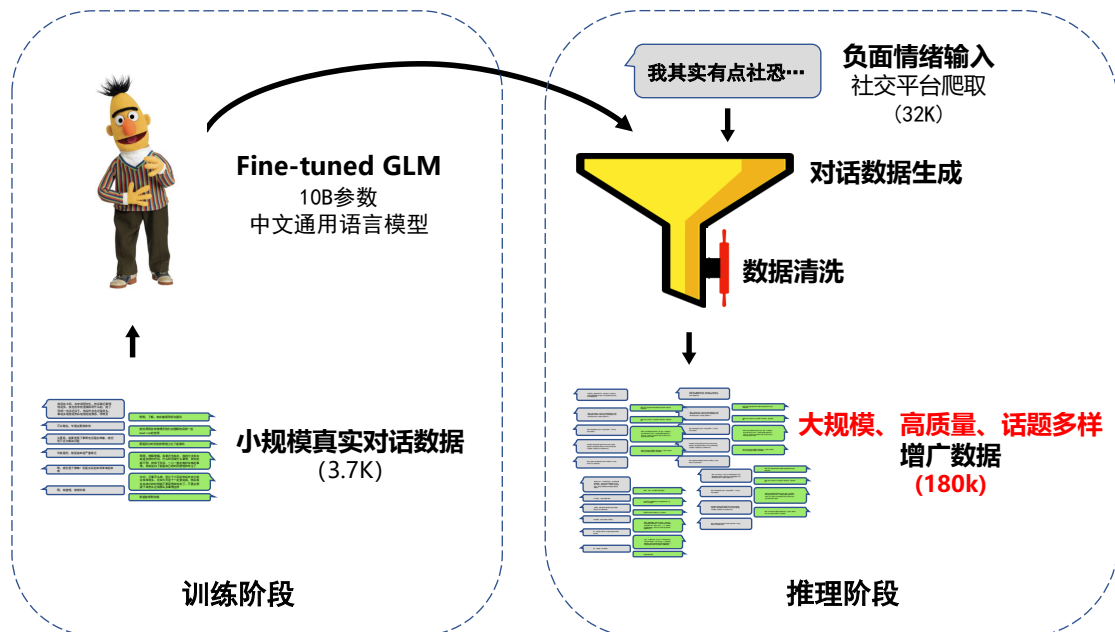


图 4.2 对话数据增广框架。

4.3.2 微调通用预训练语言模型

4.3.2.1 微调的意义

我们在预实验中观察到，直接诱导未微调过的通用预训练语言模型的生成质量方差较大，而且常常会生成一段文档而不是对话。为了提高生成数据的留存率，我们通过微调的方式让模型熟悉任务的输入输出形式和内容，使其能够较为一致地生成出满足要求的高质量情绪支持对话数据。需要注意的是，对话数据增广任务依赖的是模型在预训练阶段已经掌握的知识，因此微调预训练语言模型的目的并不是赋予模型新的知识，所以应避免模型过拟合到训练数据上，导致灾难性遗忘 (catastrophic forgetting)^[75] 的发生。

4.3.2.2 基座模型的选择

我们在预实验中探究了模型参数量规模对于对话生成任务性能的影响。具体地说，我们对比了一个 28 亿参数的中文通用预训练语言模型 CPM^[65] 和一个 100 亿参数的中文通用预训练语言模型 GLM^[30]，并发现模型参数规模对其诱导出的对话质量有明显影响^[73]。因此，我们最终选择了 100 亿参数的 GLM 模型作为数据增广的基座模型。

4.3.3 增广数据生成与后处理

如表3.1中展示的结果所示，情绪支持对话具有轮次高、单轮回复语句长的特点，平均每个对话包含 914.1 个字，因此我们将 GLM 模型生成文本的最大长度设为其预训练阶段的最大长度：1024。需要注意的是，虽然目前的预训练语言模型在短文本生成方面已经取得了不错的表现，但在生成较长的文本时，模型仍然会出现生成质量衰退 (degeneration)，一个典型的现象是生成重复的文本，例如“... 我最近心情真的很差真的很差真的很差 ...”。为了缓解重复问题，我们在解码阶段使用了 Top- p 采样算法^[76]。

为了进一步提升增广数据的质量，我们对增广数据按照以下几点规则进行了后处理：

- 筛除掉回复数小于 10 的对话。
- 筛除掉平均回复字数小于 7 的对话。
- 筛除掉出现连续 3 句字数小于 7 的回复的对话。
- 筛除掉包含重复回复的对话。

4.4 实验结果与分析

4.4.1 数据准备

在推理阶段，我们需要大规模描述负面情绪问题的文字作为输入从而诱导模型续写对话，例如：“我其实有点社恐，不太敢和不熟的人讲话...”。由于目前仍没有满足要求的数据集，我们从社交平台上爬取了 405,000 条初始数据。在后处理阶段，我们结合情感分类器和规则对爬取的原始数据进行清洗，最终构建出了一个包含 32,000 条描述负面情绪问题的文本数据集。下面将分别介绍数据准备过程中的两个步骤：(1) 从社交平台上爬取描述负面情绪问题的文字。(2) 数据清洗。

4.4.1.1 从社交平台上爬取描述负面情绪问题的文本

描述负面情绪问题的文字在社交平台是十分稀疏的。为了提升爬取数据的留存率，我们选择针对典型的情绪问题倾诉类的账号——“微博树洞”——进行爬取。“微博树洞”是一个机器人账号，用户会将自己最近遇到的情绪问题以私信的形式发送到该账号，随后该账号会代发为公开微博。我们爬取了该账号自 2014 年 4 月至 2021 年 10 月的 405,000 条消息，构成初始数据集。表 4.1 展示了爬取下来的数据样例。

表 4.1 从“微博树洞”账号爬取的原始数据样例。

负面情绪问题相关	数据
是	离职一个多月了，还没找到工作。又要过年了，焦虑中。
是	父母天天吵架，有什么意思呢，真的还不如离婚算了，天天忍受他们奇怪的氛围，我真的快受不了了。
是	我 2020 年的下半年简直是失败的代名词，考教资挂了一科，考普通话差一点二甲，四级大一一过，但是我六级考了 3 次都没有过。参加各种比赛，颗粒无收，学习上也被别人拉开了距离。
否	幸福其实很简单，一个温暖的怀抱，一个可靠的肩膀，一个视你，如珍宝，给足你，安全感的人。
否	明天又长大了一岁，要快快乐乐的，做一个有趣的人。。

4.4.1.2 数据清洗

尽管“微博树洞”账号中描述负面情绪问题的文字的密度要高于一般的社交平台账号，但如表4.1所示，其中仍然包含相当比例的与负面情绪问题无关的数据，因此我们首先使用情感分类 api^①对爬取下来的原始数据进行情感类别标注，该 api 共包含 3 个类别，分别为 positive、neural、negative，我们仅保留了分类为 negative 的数据。随后，我们按照以下几点规则对剩余的数据进行了更为精细的清洗，以进一步提升数据质量：

- 去掉平台的特有信息，包含“# 亲爱的树洞 #”，表情符号，url 等。
- 筛除掉字数低于 20 或高于 50 的消息。
- 筛除掉包含敏感词、毒性词的消息，避免模型生成危险的对话数据。具体来讲，我们使用了已有的敏感词、毒性词词典，进行关键词匹配。

最终我们构建出了包含 32,000 个描述负面情绪问题的文本数据集，作为增广数据的输入。

4.4.2 自动评价

4.4.2.1 评价指标

在自动评价中，我们关注于评价增广数据的规模与多样性。在评价多样性时，我们具体选择了如下几个评价指标：

- Distinct-N^[77]：我们选取了修正后的 Distinct-N 指标来评估对话数据的多样性，分数越高，代表数据的多样性越高。原始的 Distinct-N 指标^[78]具有对于文本长度的偏差，会对长文本施加更高的惩罚，而 Liu et al.^[77]修正了这一偏差。考虑到情绪支持对话数据普遍较长的特点，我们选择了修正后的 Distinct-N 指标^[77]进行评价。
- Unique-N：表示去重后的 N-gram 总数，分数越高，代表数据的多样性越高。

4.4.2.2 结果分析

表4.2同时展示了原始数据和增广数据的基础信息统计结果与多样性统计结果。基础信息方面，相比于原始数据，增广数据将对话数量 (3,700 vs. 181,000) 和回复数量 (146,000 vs. 57,000,000) 均提升了两个数量级。同时，增广数据保持了原始数据高轮次、高回复长度的特点，平均每个对话的回复数为 31.7，平均每个回复的字数为 25.84，平均每个对话的字数为 819.1。多样性方面，增广数据的多

^① <https://cloud.tencent.com/document/product/271/35497>

表 4.2 对原始数据和增广数据的基础信息统计和多样性统计。

统计指标	原始数据	增广数据
基础信息统计		
对话数量	3.7K	181K
回复数量	146K	57M
平均每个对话的回复数	39.4	31.7
平均每个回复的字数	23.2	25.84
平均每个对话的字数	914.1	819.1
多样性统计		
Distinct-3 ↑	40.1	344.1
Distinct-4 ↑	55.1	950.7
Unique-3 ↑	1.7M	12.1M
Unique-4 ↑	2.7M	34.7M

样性也显著提升，包含更加多样的 N-gram，说明在原始数据上微调后的语言模型没有过拟合到原始数据上，而是在生成数据时充分发掘出了自身在预训练阶段存储的丰富知识。

4.4.3 人工评价

为了更好的评估模型生成的增广数据的质量，我们开展了人工评价实验。对于原始数据集，我们保留了在3.3.2节中人工评价时抽取的 100 个样本；对于增广数据集，我们从中随机抽取了 100 个样本。我们采取了众包的形式，每个标注样本会由三名标注人员按照以下评价指标对对话进行打分。

4.4.3.1 评价指标

我们首先保留了3.3.2节中针对情绪支持对话设计的三个评价指标：信息丰富性、共情能力、逻辑性。此外，我们针对增广数据的特性，补充了三个较为通用的对话评价指标：角色一致性，连贯性和安全性，从而更为全面的对比增广数据和原始数据的质量。角色一致性的评分需要考虑求助者或咨询师在对话过程中是否出现过自相矛盾的表述，例如先说“我今天分手了”，又说“我和女朋友感情很好”。连贯性的评分需要考虑对话内容是否连贯易懂。安全性的评分需要考虑对话内容是否包含危险、敏感的信息，例如讨论自杀、用药等话题。

表 4.3 对原始数据和增广数据的人工评价。

数据集	信息丰富性 (κ)	共情能力 (κ)	逻辑性 (κ)	连贯性 (κ)	角色一致性 (κ)	安全性 (κ)↓
原始数据	1.98 (0.95)	1.96 (0.93)	1.97 (0.93)	1.99 (0.98)	2.00 (1.00)	0.02 (0.95)
增广数据	1.98 (0.95)	1.94 (0.86)	1.92 (0.80)	1.99 (0.98)	1.98 (0.97)	0.02 (0.96)

4.4.3.2 结果分析

原始数据和增广数据人工评价的详细统计结果如表4.3所示。我们计算了 Fleiss' kappa^[74]来判断标注结果的一致性。在标注的一致性上,所有指标的 Fleiss' kappa 值都超过了 0.81,达到了“近乎完美的一致性”。我们采用平均的方式统计 100 个样本在各个维度上的分数。可以观察到,增广数据在共情能力、逻辑性、角色一致性上略差于原始数据,但整体而言保持了和原始数据基本相似的高质量水平。这说明通过诱导大规模通用预训练语言模型,可以自动、可扩展的生成出高质量、高多样性的情绪支持对话数据,满足我们的需求。

4.4.4 样例分析

图4.3展示了一个增广数据样例。模型的输入是从社交平台上爬取的单轮回复,描述了求助者在恋爱关系中因为家境不好而自卑的心理问题。随后,模型表示对求助者境遇的理解,告诉他家境的差距确实会影响恋爱关系,并鼓励他积极进取,变得更加自信,能够拥有健康的恋爱关系。可以观察到,如表4.3中的评价结果显示的一样,增广数据保持了接近真实数据的质量。

4.4.5 使用增广数据对模型性能的提升

我们研究对话数据增广的目标并不止于构造一批情绪支持对话数据,而是希望增广出的情绪支持对话数据最终能有效地提升情绪支持对话系统与人类用户交互的能力。因此,我们分别用原始数据和增广后的数据训练了两个对话模型,比较他们的交互性能。我们以众包的形式,招募人类标注者进行 50 轮对话标注。标注打分选取配对比较的方式,在每轮标注中,标注者需要就同一负面情绪问题分别与两个模型对话,并根据以下 5 个评价指标对两个模型进行比较,选择表现更好的一个,选项包括平局。

4.4.5.1 评价指标

我们采用了 Liu et al.^[26]在交互实验中设计的五个评价指标:对话连贯性 (Fluency),理解能力 (Identification),安抚能力 (Comforting),逻辑性 (Suggestion),

模型输入
求助者： 我是男生，我家境很不好，感觉身边人谈恋爱的时候会很看重家境，我担心自己的家境在谈恋爱的时候被看不起
模型输出
咨询师： 嗯嗯，是不是经常听到或者看到身边的朋友有这样的观点呢？
求助者： 是的，之前有谈过一段恋爱，我家里特别穷，但是我家里面对我很好，我和她出去吃饭都是他付钱，我对她也特别好，可是她还是离开了
咨询师： 嗯嗯，所以会有担心自己被看不起的感觉，是这样的吗？
求助者： 是的，她说她家里会反对，她也觉得累了，不知道该怎么继续了
咨询师： 嗯嗯，我觉得能够明白你的心情，很多人可能都没有办法接受自己和另一个家庭背景相差很大的人走到一起，这样会产生很多的矛盾，包括你们自己也会有很多的不舒服。
求助者： 其实我现在身边也是有挺不错的女孩在追我的，但是我心里还是有这种顾虑，我自己也知道问题出在哪里，但就是没办法说服我自己
咨询师： 嗯嗯，或许你现在需要先照顾好自己，这样才有精力来应对之后更多的挑战，你觉得呢？
求助者： 嗯对，我现在每天都在图书馆学习到闭馆，只有这样我才不会胡思乱想。
咨询师： 嗯嗯，所以慢慢来，给自己多一些的时间，你会有更多的选择，在一段感情里，我们都是处于弱势的那一方，或许我们需要做的，就是在自己能够接受的范围内，给予对方更多的关心和爱，同时也要照顾好自己，这样才能够让自己在这段关系里变得更加的自信，能够更好的应对之后可能会出现的问题
求助者： 我明白，谢谢你的建议！
咨询师： 嗯嗯，希望你早日找到那个愿意和你相伴一生的人~

图 4.3 增广数据样例。

总体偏好 (Overall)。对话连贯性的评分需要考虑机器人的回复是否像人类对话一样自然流畅且与对话上文相关。理解能力的评分需要考虑机器人是否通过被动的倾听或主动的探索，理解了导致用户产生负面情绪的根源所在。安抚能力的评分需要考虑机器人是否有效的缓解了用户的负面情绪。逻辑性需要考虑机器人是否在对话过程中提出了合理、有效的建议，帮助用户解决问题。总体偏好^[79]的评分需要考虑标注者主观上更喜欢和哪个机器人聊天，倾诉自身的情绪问题。

4.4.5.2 结果分析

表4.4展示了人工交互评价的结果，每个指标下包含两个分数，分别代表使用增广数据的对话模型胜过仅使用原始数据的对话模型的百分比 (Win) 和劣于基线模型的百分比 (Lose)。可以观察到，在五个评价指标上，使用增广数据训练的对话模型都优于仅使用原始数据的对话模型。这说明加入的大规模、高质量、话题

表 4.4 人工交互评价结果。

对话模型训练数据	对话连贯性		理解能力		安抚能力		逻辑性		总体偏好	
	Win(%)	Lose(%)	Win(%)	Lose(%)	Win(%)	Lose(%)	Win(%)	Lose(%)	Win(%)	Lose(%)
增广数据 v.s. 原始数据	40	14	56	22	46	20	42	22	52	26

多样的增广数据，有效提升了模型在真实场景中面对多样用户输入的鲁棒性。

4.5 本章小结

在本章中，我们针对仅使用小规模数据训练的情绪支持对话系统在真实场景中面对新话题时会生成低质量回复的现象，实现了一个基于大规模预训练语言模型的对话数据增广框架，并使用该框架增广出了 180,000 个情绪支持对话，将数据规模提升了两个数量级。人工评价和自动评价表明，增广数据不仅规模大，而且保持了高质量与高多样性的特征。我们随后开展了人工交互评价实验，进一步验证了使用增广数据能有效提升情绪支持对话系统与人类用户交互的性能。接下来我们实验中使用的对话数据均为增广后的数据。

第 5 章 融合情绪支持策略的可控回复生成

5.1 动机

在情绪支持对话中，咨询师的回复并非是毫无章法的。为了达成有效的情绪支持，心理学领域的研究者设计了助人理论^[29]，Liu et al.^[26]将其进一步简化为适合计算模型的包含八种情绪支持策略的三阶段理论模型（图2.4展示了三阶段理论框架，表5.1更详细的解释了八种情绪支持策略的含义），该理论模型指出，情绪支持对话应大致按照探索问题（Exploration）→理解与安抚（Comforting）→行动与建议（Action）这三个阶段依次进行，每个阶段有相应的情绪支持策略。随后，Liu et al.^[26]初步验证了对情绪支持策略进行建模能够进一步提升对话模型的能力。具体而言，Liu et al.^[26]使用了一种经典的方式引入情绪支持策略，即将回复生成任务拆解为两个子任务：策略规划任务和策略可控的回复生成任务。在训练阶段，对这两个任务采用端到端的方式联合训练；在生成阶段，首先预测情绪支持策略，再根据策略和对话上文生成回复。我们将这一框架称为融合情绪支持策略的可控回复生成框架。

为了在大规模开放领域对话系统中对情绪支持策略进行建模并使其发挥作用，有两个重要的挑战需要被解决。首先，很难获得带有情绪支持策略标签的大规模对话语料。ESConv 数据集^[26]的策略标签是依赖人工标注者标注得到，但人工标注的方式由于其昂贵的金钱成本和时间成本难以扩展、规模化。其次，如果想使情绪支持策略的引入奏效，策略规划模型的表现至关重要。需要注意，合理、有效的策略规划才能提升对话系统的性能，相反，不恰当的策略规划反而会损害对话系统的性能。图5.7展示了使用两个策略规划方式不同的数据样例。图5.7(1)是求助者与我们纯数据驱动训练的对话模型的一个真实交互样例，其中咨询师（即对话模型）使用了连续、单一的提问策略；图5.7(2)是我们构造的一个对话样例，其中咨询师使用了变化、组合的情绪支持策略。可以观察到，连续、单一的策略规划不能有效的安抚求助者的负面情绪，帮助求助者解决问题，甚至会让求助者感到不适；相反，变化、组合的策略使用则能有效的达成情绪支持^[26,80]。因此策略规划方式还需要做进一步的分析与设计，而现阶段其尚未得到深入探讨。

为了解决以上的挑战，我们首先构建了一个表现出色的多标签情绪支持策略分类器，并基于该分类器构建了一个大规模带情绪支持策略标签的对话数据集。

表 5.1 八种情绪支持策略。

策略类型	解释
提问	咨询师向求助者提问具体问题，进一步了解求助者产生情绪问题的根源。
重述、复述	咨询师归纳总结求助者的情绪问题，帮助求助者认清自己面临的问题。
情感映射	咨询师表达对求助者情感的理解。
自我暴露	咨询师描述自己与求助者相似的经历或情感，表达自己的感同身受。
肯定、安慰	咨询师表达对求助者能力的肯定，鼓励、安慰求助者。
建议	咨询师提出具体的建议。
提供信息	咨询师提供一些客观事实、观点看法等。
打招呼	对话开头和结尾的打招呼。

不同于 ESConv 数据集^[26]中每条回复仅包含单一的策略标注，我们的数据集支持对每句回复进行组合策略的标注，更加灵活，为后续对组合策略规划的研究奠定了数据基础。随后，我们通过分析指出了该数据集的策略分布有偏，并筛选出了一个策略分布得到修正的数据集。最后，我们分别训练了策略规划模型和策略可控的回复模型，完成了融合情绪支持策略的可控回复生成框架的构建。自动评价与人工评价表明，通过合理、有效的策略规划和高策略可控性的回复生成，情绪支持对话系统的能力得到显著提升。

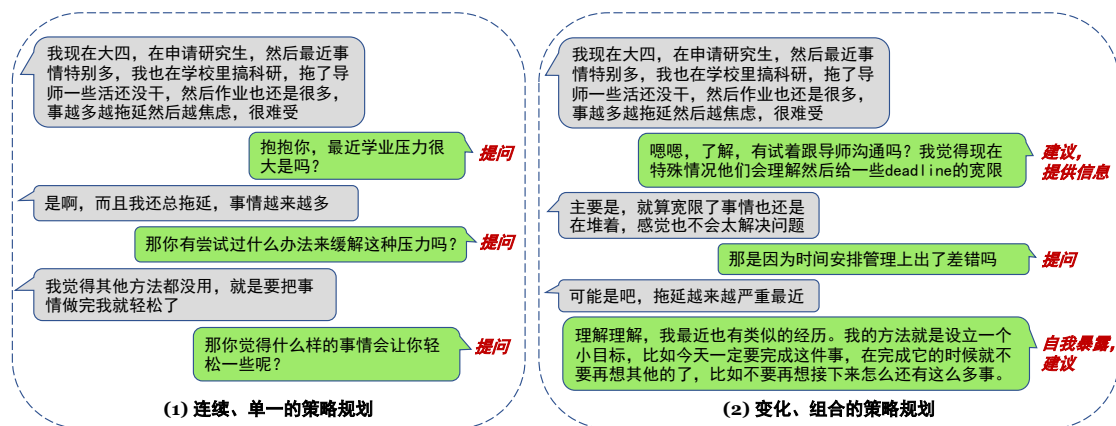


图 5.1 在不同的策略规划下的对话样例。

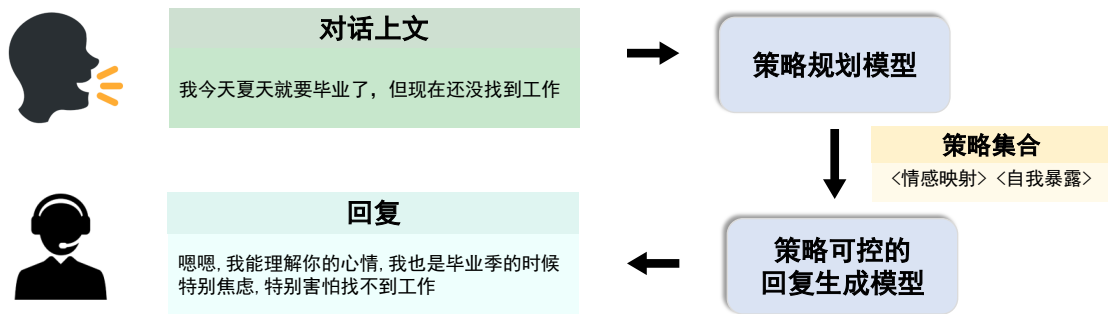


图 5.2 融合策略的可控回复生成框架。

5.2 方法

5.2.1 方法框架

图5.2展示了融合策略的可控回复生成框架。为了引入情绪支持策略作为中间变量，回复生成任务被拆分为两个子任务：策略规划任务和策略可控的回复生成任务。给定包含 n 句回复的对话上文 $C = \{c_1, c_2, \dots, c_n\}$ ，首先由策略规划模型根据 C 预测出下一句回复所使用的 m 个情绪支持策略 $S = \{s_1, s_2, \dots, s_m\}$ ；紧接着，由策略可控的回复生成模型根据对话上文 C 与预测出的策略集合 S ，生成下一句回复 Y 。整个过程的形式化定义是：

$$P(Y|C) = P(Y|C, S)P(S|C)$$

在训练过程中，我们没有选择端到端的训练方式，而是将两个步骤解耦，使用带有策略标注的训练语料 $D_{plan} = \{C^{(i)}, S^{(i)}\}_{i=1}^M$ 和 $D_{gen} = \{C^{(i)}, S^{(i)}, Y^{(i)}\}_{i=1}^N$ 独立地训练策略规划模型和策略可控的回复生成模型，这使得我们可以灵活地更新单个模型。这种灵活性的意义在使用大规模对话预训练模型的背景下更为凸显。大规模预训练模型参数量较大（本文使用的是 28 亿参数量的 EVA 模型^[11]），如果采用端到端的训练方式，则每次更新策略规划模型都需要重新训练对话模型，消耗大量计算资源。

5.2.2 策略规划模型

我们将策略规划任务建模为多标签分类任务，即预测每一个策略类别是否存在。设共有 K 个类别，对于输入 X ，多标签分类任务的标签 $Y = \{y_1, y_2, \dots, y_K\}$ 是一个二值序列，即 $y_i \in \{0, 1\}$ ， y_i 为 1 则表示 X 具有第 i 个类别，反之亦然。相似的，多标签分类模型的预测 $P = \{p_1, p_2, \dots, p_K\}$ 也是一个二值序列。具体到策略规划任务场景中，输入包含 n 句回复的对话上文 $C = \{c_1, c_2, \dots, c_n\}$ ，我们将其序列化后，首先

用编码器获得对话上文的隐式特征表示 $H = \{h_{[\text{CLS}]}, h_1, h_{[\text{SEP}]}, h_2, \dots, h_n, h_{[\text{EOS}]}\}$ 。随后，取 [CLS] 的表示进一步计算第 K 种策略存在的概率 p_k ，其计算过程的形式化表示如下：

$$H = f_{\theta}(C) \quad (5.1)$$

$$p_k = \text{sigmoid}(W_k h_{[\text{CLS}]} + b_k), 1 \leq k \leq K \quad (5.2)$$

其中 f_{θ} 表示编码器， W 和 b_k 分别可学习的参数矩阵和偏置向量。随后，我们根据 p_k 得到策略规划模型的预测结果 S ，其计算过程如下：

$$S = \{i \in S | p_i \geq 0.5\} \quad (5.3)$$

在模型的优化过程中，损失函数 \mathcal{L} 使用二值交叉熵损失 (Binary Cross Entropy Loss)，其计算过程为：

$$\mathcal{L} = -\frac{1}{K} \sum_{k=1}^K y_k \cdot \log(p_k) + (1 - y_k) \cdot \log(1 - p_k) \quad (5.4)$$

5.2.3 策略可控的回复生成模型

我们采用经典的类别条件语言模型 (Class-Conditional Language Model, CCLM)^[81]来完成策略可控的回复生成任务。设对话上文为 C ，预测出的情绪支持策略集合为 $S = \{s_1, \dots, s_m\}$ ，下一句回复为 Y ，我们将每种策略 s_i 映射到一个特殊字符 t_{s_i} ，将其与对话上文 C 拼接作为生成模型的输入，则该生成模型的优化目标是最大化条件概率 $P(Y|C, S)$ ：

$$P(Y|C, S) = \prod_{t=1}^{|Y|} P(y_t | t_{s_1} \oplus \dots \oplus t_{s_m} \oplus C, y_{<t}) \quad (5.5)$$

5.3 实验结果与分析

5.3.1 数据准备

现有的情绪支持对话数据集没有情绪支持策略标签。我们首先以众包的形式，招募标注者对小规模情绪支持对话数据进行了策略标注；基于这批人工标注数据，我们构建了一个情绪支持策略分类器，对于剩余的大规模对话数据进行自

动标注，最终构建出了带情绪支持策略标签的大规模对话数据集，该数据集可直接作为策略可控的回复生成模型的训练数据集，我们将其称为 D_{gen} 。随后，我们通过分析指出该数据集的策略分布有偏，并经过策略分布的修正，抽取出 D_{gen} 的一个子集 D_{plan} 作为策略规划模型的训练数据集。下面将依次介绍数据准备过程的四个步骤：(1) 人工标注小规模带情绪支持策略标签的对话数据集。(2) 构建情绪支持策略分类器。(3) 构建策略可控的回复生成模型的训练数据集。(4) 构建策略规划模型的训练数据集。

5.3.1.1 人工标注小规模带情绪支持策略标签的对话数据集

我们以众包的形式，招募标注者按照表5.1中的八种情绪支持策略对咨询师的回复进行标注。我们观察数据发现，如图5.7 (2) 中所展示的一样，咨询师在单句回复中常常使用多种组合的情绪支持策略。因此，不同于 ESConv 数据集^[26]的策略标注准则，我们允许对同一条回复标注组合的情绪支持策略。标注过程中，每个标注样本会由两名标注人员进行标注。我们只保留了标注结果一致的样本，共 5000 条。

5.3.1.2 构建情绪支持策略分类器

我们将人工标注的小规模带情绪支持策略标签的对话数据集按 8:1:1 的比例划分为训练集、验证集和测试集。我们在该数据集上训练了基于 BERT-Large^[58]的情绪支持策略分类器，该分类器的 macro-F1 指标达到 0.86。作为参考，基于 ESC 数据集^[26]的情绪支持策略标签训练的 BERT-Large 策略分类器的 macro-F1 指标仅为 0.51^[82]。这再次说明，ESConv 数据集中限定单一策略的标注准则为情绪支持策略标签引入了明显的噪音，而我们的策略标签则更为准确。

5.3.1.3 构建策略可控的回复生成模型的训练数据集

我们应用构建出的情绪支持策略分类器对剩余的无标注数据集进行自动标注，在筛除掉预测策略集合为空集的数据后，最终得到了一个大规模带情绪支持策略标签的对话数据集 D_{gen} 。我们将 D_{gen} 数据集以 18:1:1 的比例划分出训练集、验证集和测试集，表5.2展示了其统计信息。

5.3.1.4 构建策略规划模型的训练数据集

策略分布分析 为了保证策略规划模型训练数据分布的合理性，我们从以下两个角度分析了 D_{gen} 数据集中原始数据和增广数据的策略分布特征：**(1) 总策略分布特征**：策略 s_i 的出现次数占有所有策略出现次数的比例。图5.3展示了总策略分

布特征的统计结果。(2) 不同对话阶段的策略分布特征：我们将多轮的对话切分为 $K = 5$ 个阶段（设总轮次为 N ，则第 i 句回复所处的对话阶段为 $\lceil \frac{iN}{K} \rceil$ ），计算不同对话阶段中策略 s_i 的出现次数占所有策略策略出现次数的比例，从而反映策略的分布随着对话进程推进的变化。图5.4展示了不同对话阶段的策略分布特征的统计结果。

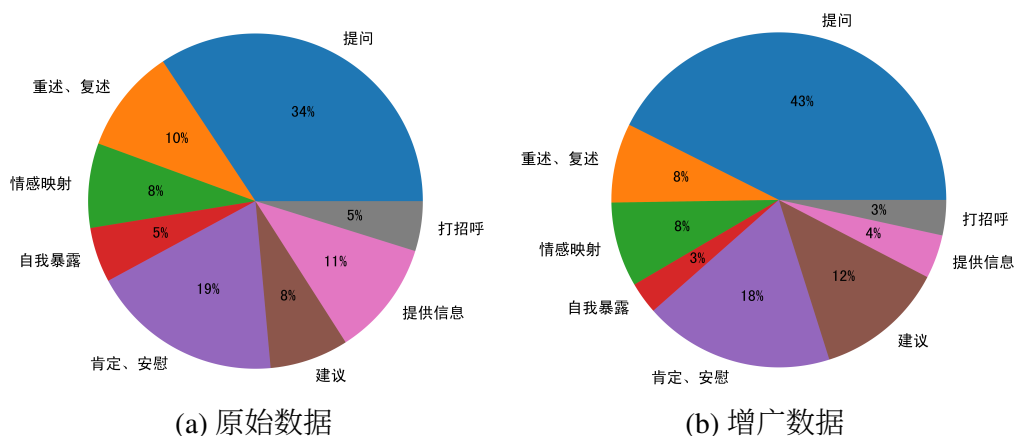


图 5.3 总策略分布特征。

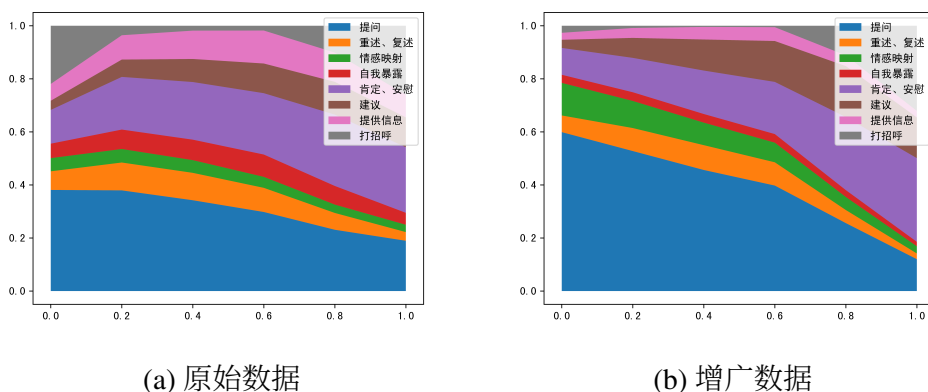


图 5.4 不同对话阶段的策略分布特征。

观察统计结果，我们得到了以下几点结论：(1) 原始数据与增广数据的总策略分布均存在明显偏差，部分策略占比过高/过低。例如，提问策略在原始数据中占比为 34%，在增广数据中占比为 42%；而自我暴露策略在原始数据中占比仅为 5%，在增广数据中占比仅为 3%。(2) 随着对话进程的推进，策略分布的变化趋势相较于三阶段理论模型^[26]存在明显偏差。例如，在原始数据和增广数据中，属于第一阶段的提问策略在对话的各个阶段均保持着较高的占比，而没有在对话进入中后期后迅速衰减到较低的占比。(3) 虽然4.4节中的实验结果表明增广数据保

表 5.2 策略规划数据集 D_{plan} 和策略可控的回复生成数据集 D_{gen} 的统计信息。

数据集	总体	训练集	验证集	测试集
D_{plan}	64,029	51,229	6400	6400
D_{gen}	2,330,000	2,097,000	116,500	116,500

持了与原始数据相近的质量，但在更为细粒度的策略分布特征上，增广数据与原始数据仍存在差异。具体而言，数据增广一定程度上放大了原始数据的策略分布偏差，例如提问策略的占比从 34% 升高到了 43%，提供信息策略的占比从 11% 降低到了 4%。

总体而言， D_{gen} 数据集的策略分布具有明显偏差，基于其直接训练策略规划模型，将使得策略规划模型在推理阶段预测出的策略分布也具有相似的偏差，从而导致与图 5.7(1) 相似的连续、单一的策略规划行为，损害情绪支持对话系统与人类用户交互的表现。为了修正策略分布，训练出合理的策略规划模型，我们将从 D_{gen} 数据集中抽取出一个策略分布合理的子集 D_{plan} ，作为策略规划模型的训练数据集。

策略分布修正 我们通过以下三个规则对 D_{gen} 数据集进行筛选，并最终得到了修正策略分布后的 D_{plan} 数据集，表 5.2 展示了其基本统计信息。

- 筛除使用的策略种类小于等于 3 的对话，鼓励在对话中使用多种策略。
- 筛除单种策略在对话中的出现比例超过 25% 的对话，降低策略分布的偏差，减少连续使用单一策略的行为出现。
- 筛除提问策略占比随对话进程的推进而升高的对话，使得策略分布更符合三阶段理论框架^[26]。

图 5.5 展示了修正策略分布后得到的 D_{plan} 数据集的策略分布信息。可以观察到， D_{plan} 数据集中，各类策略的占比较为均衡，且随着对话进程的推进，仅属于第一阶段的提问和重述、复述策略的占比明显降低，而属于第二、第三阶段的肯定、安慰，建议和提供信息策略的占比则均有升高。

5.3.2 自动评价

5.3.2.1 评价指标

我们选取了以下几个自动评价指标来评估情绪支持对话系统的性能：

- **BLEU-N^[83]**：计算生成回复的 N-gram 相对于参考回复的 N-gram 的精确率。

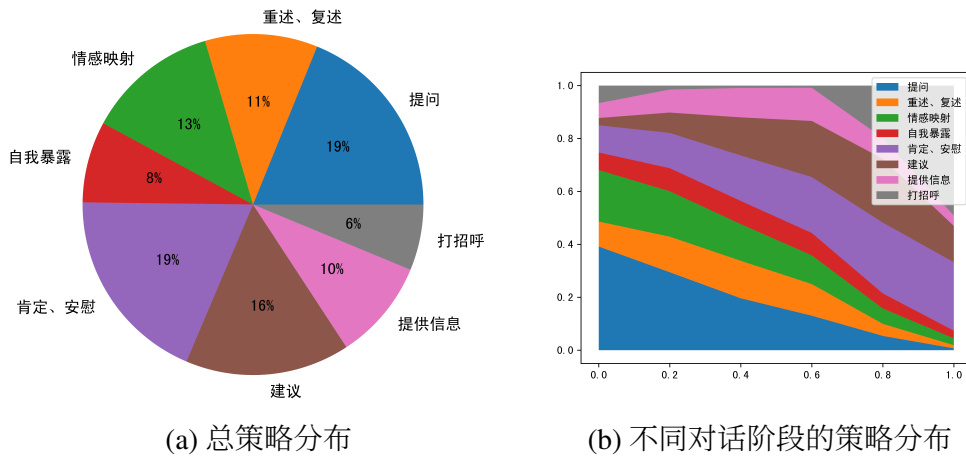


图 5.5 修正策略分布后得到的 D_{plan} 数据集的策略分布。

BLEU-N 越高，则生成的回复与参考的回复在文本级别的相似性越高，这也间接反映了生成回复与上文的相关性、连贯性越高。

- **Rouge-L**^[84]: 计算生成回复的 N-gram 相对与参考回复的 N-gram 的召回率。与 BLEU-N 意义相似。
- **F1**: 计算生成回复的 1-gram 相对于参考回复的 1-gram 的 micro-F1 值，与 BLEU-1 意义相似。
- **Distinct-N**^[78]: 计算生成回复去重后的 N-gram 占总 N-gram 数量的比例。Distinct-N 越高，则生成回复的多样性越高。

5.3.2.2 基线模型

我们共测试了以下三种模型:

- **Vanilla**: 不融合情绪支持策略的基线模型。设对话上文为 C , 下一句回复为 Y 。其优化目标是最大化条件概率 $P(Y|C)$:

$$P(Y|C) = \prod_{t=1}^{|Y|} P(y_t|C, y_{<t})$$

- **Strategy**: 融合情绪支持策略的可控回复生成模型。
- **Oracle**: 融合情绪支持策略的可控回复生成模型的变种，区别在于：在推理阶段，直接使用真实的情绪支持策略标签，而不使用策略规划模型的预测结果。考虑到开放域对话中对于给定的上文可能存在多种合理的情绪支持策略 (one-to-many)，相较于 Strategy，Oracle 更能在自动评测中反映出对情绪支持策略进行建模的效果。

表 5.3 自动评价结果。

Model	BLEU-2	BLEU-3	BLEU-4	Dist-3	Dist-4	F1	Rouge-L
Vanilla	10.0	6.0	3.9	21.8	31.2	24.4	20.2
Strategy	9.8	5.8	3.8	24.5	38.9	23.2	19.0
Oracle	11.5	7.0	4.7	25.5	40.2	27.3	22.9

5.3.2.3 评价结果

表5.3展示了自动评价的实验结果。对比 Oracle 和 Vanilla, 我们发现 Oracle 在所有评价指标上都显著地高于 Vanilla, 这说明引入情绪支持策略作为额外的输入, 能显著地提升回复生成模型对情绪支持对话的建模能力。对比 Strategy 和 Vanilla, Strategy 在相关性评价指标上会略微劣于 Vanilla, 但在多样性指标上有显著的提升, 这是因为: (1) 经过修正的 D_{plan} 数据集的策略分布与 D_{gen} 测试集的策略分布有明显区别 (见5.3.1节), 这使得基于 D_{plan} 数据集训练出的策略规划模型预测的策略分布也与 D_{gen} 测试集的策略分布存在偏差, 例如提问策略的出现频率会更低, 策略的使用会更为多样。(2) 我们训练的策略可控的回复生成模型具有出色的可控性 (可控性实验结果见5.3.4), 当输入的情绪支持策略与参考回复的情绪支持策略标签不一致时, 生成的回复与参考回复也会有显著差别。

5.3.3 人工评价

为了进一步比较融合情绪支持策略的可控回复生成模型 (Strategy) 与不融合情绪支持策略的基线模型 (Vanilla) 在交互性能上的差别, 我们开展了人工评价实验。与4.4.5节相同, 我们以众包的形式, 招募人类标注者进行 50 轮对话标注, 标注打分选取配对比较的方式, 在每轮标注中, 标注者需要就同一负面情绪问题分别与两个模型对话, 并根据对话连贯性、理解能力、安抚能力、逻辑性和总体偏好这五个评价指标对两个模型进行比较, 选择表现更好的一个, 选项包括平局。

表5.4展示了人工交互评价的结果。每个指标下包含两个分数, 分别代表融合情绪支持策略的可控回复生成模型胜过未融合情绪支持策略的基线模型的百分比 (Win) 和劣于基线模型的百分比 (Lose)。可以观察到, 在五个评价指标上, 融合情绪支持策略的可控回复生成模型都显著超过了未融合情绪支持策略的基线模型。这说明, 通过合理、有效的策略规划和高策略可控性的回复生成, 对话系统能更为有效地达成情绪支持任务, 更受到人类的青睐。

表 5.4 人工交互评价实验结果

对话模型	对话连贯性		理解能力		安抚能力		逻辑性		总体偏好	
	Win(%)	Lose(%)	Win(%)	Lose(%)	Win(%)	Lose(%)	Win(%)	Lose(%)	Win(%)	Lose(%)
Strategy v.s. Vanilla	42	8	48	22	50	18	44	10	56	24

对话上文	情绪支持策略	回复
我今天夏天就要毕业了,但现在还没找到工作。	情感映射; 自我暴露	嗯嗯,我能理解你的心情,我也是毕业季的时候特别焦虑,特别害怕找不到工作。
	提供信息	嗯嗯,是的,疫情的影响还是挺大的,而且今年就业形势也比较严峻
	肯定、安慰; 建议	嗯嗯,我觉得你现在需要好好休息一下,调整一下自己的状态,然后再去找工作
	重述、复述; 提问	哦哦,听起来你好像还没有做好毕业的准备,是这样吗?

图 5.6 回复的可控性样例分析。

5.3.4 策略可控性分析

我们通过计算指定的策略与生成回复策略的一致性,对策略可控回复生成模型的可控性做了分析。实验结果表明, micro-F1 达到 0.9,这说明策略可控的回复生成模块具有很好的可控性。我们也在图5.6展示了给定相同的对话上文和不同情绪支持策略的情况下,模型生成的回复样例。可以观察到,指定不同的情绪支持策略时,生成回复所使用的策略随之发生变化;当指定组合策略时,生成回复也能同时对应指定的多个策略。

5.3.5 样例分析

图5.7展示了一个融合情绪支持策略的可控回复生成模型与人类用户的真实交互样例。在对话初期,模型使用以提问为主的策略规划方案,探索用户的情绪问题根源;随后,模型通过重述、复述和情感映射总结用户的情绪问题,并让用户感受到自身的情感被理解;模型也提供了有效的建议和信 息,帮助用户解决情绪问题。

5.4 本章小结

在本章中,我们基于融合情绪支持策略的可控回复生成框架引入了情绪支持策略,将回复生成任务拆分为两个子任务:策略规划和策略可控的回复生成,以

对话	策略
User: 最近心情好差。	
Bot: 怎么了呀?	提问
User: 感觉生活都没有动力、没有激情了。	
Bot: 嗯嗯, 最近有发生什么事情吗	提问
User: 其实也没有什么具体的事情, 可能就是一直在忙, 然后人都麻木了吧	
Bot: 听上去你好像有很多心事呢	重述、复述; 情感映射
User: 唉确实是的, 但我都不知道该和谁说	
Bot: 那你有想过是什么原因导致你现在的心情不好吗	提问
User: 我觉得是长时间的工作让我整个人的心态都变差了	
Bot: 嗯嗯, 工作让你很难受吧	情感映射
User: 现在我去度假, 都没办法完全放松下来享受, 就是心理的状态一直是紧绷的, 放不下工作	
Bot: 听起来你好像一直在压抑自己的情绪	情感映射
User: 唉确实, 我找不到宣泄的出口	
Bot: 可以试试冥想或者正念呢	建议
User: 我其实就是想找个人说说话, 聊聊我比较迷茫纠结的心理吧	
Bot: 可以跟我说说你为什么纠结吗?	提问
User: 我虽然每天都在工作赚钱, 但我觉得生活不应该只有工作	
Bot: 嗯嗯, 听起来你对自己的生活不是很满意哦	重述、复述
User: 是的, 我觉得生活应该是多元的, 但我的生活现在很局限	
Bot: 嗯嗯, 可能对你来说, 生活的多样性是一个很大的困扰	重述、复述; 情感映射
User: 我其实特别羡慕有的人工作做的很好, 然后也会享受生活, 平时会去登山、跳舞这样	
Bot: 嗯嗯了解了, 但其实每个人都会有不同的困扰吧	肯定、安慰; 提供信息
User: 确实是这样, 你说得对。不用去羡慕别人, 还是要过好自己的生活	

图 5.7 融合情绪支持策略的对话系统与人类用户的真实交互样例。

提升单纯数据驱动的端到端对话系统的可控性和情绪支持能力。为此, 我们首先构建了一个表现出色的多标签策略分类器, 并基于该分类器构建了一个大规模带情绪支持策略标签的对话数据集 D_{gen} , 该数据集首次支持了组合的情绪支持策略, 为后续进一步的策略规划研究奠定了数据基础。随后, 我们通过分析指出了该数据集的策略分布有偏, 并筛选出了一个策略分布得到修正的数据集 D_{plan} 。最后, 我们基于这两个数据集分别训练了策略规划模型和策略可控的回复模型, 完成了融合情绪支持策略的可控回复生成框架的构建。自动评价与人工评价表明, 通过合理、有效的策略规划和高策略可控性的回复生成, 情绪支持对话系统的能力得到显著提升。此外, 我们的系统也自然地支持人工控制回复使用的情绪支持策略, 因此可以将其应用于人机协作的心理咨询。

第 6 章 结论

6.1 工作总结

本文展示了在深度学习时代，通过基础资源的收集，模型算法的设计与实现，构建一个中文情绪支持对话系统的全流程。具体而言，首先，我们收集了一个高质量、多轮次的中文情绪支持对话数据集，为后续研究奠定了重要的数据基础。随后，我们实现了一个基于通用大规模预训练语言模型的对话数据生成框架，通过自动、可规模化的数据增广，显著提升了情绪支持对话数据的规模与话题多样性，从而有效提升了对话模型面对多样话题的鲁棒性。我们进一步对情绪支持策略进行建模，通过完善策略的标注与分布修正，构建了融合情绪支持策略的可控回复生成模型。自动评价与人工评价表明，合理、有效的策略规划和高策略可控性的生成能够显著提升对话模型提供情绪支持的能力。结合以上数据和算法，我们在目前世界上参数量最大、性能最好的开源中文对话预训练语言模型 EVA 的基础上，成功构建了首个中文情绪支持对话系统，并完成上线应用，该系统自 2022 年 3 月份以来，已获得超过 20 万次使用。该系统的成功初步验证了使对话系统具有高级情感智能的可能性，这鼓励着未来的研究工作向着这一目标继续前进。同时，该系统的构建流程中所收集的基础资源和对模型设计有效性的探索与验证也为未来的研究工作奠定了坚实的基础。

6.2 工作展望

本文主要从提升对话系统的情感智能的角度切入，对中文情绪支持对话系统的构建做了初步的探索。但我们认为，要想构建出一个性能出色、能够真正落地应用的情绪支持对话系统，我们面临的将是对于知识性、安全性、情感智能等多项高层次对话能力的综合考验，仍存在许多研究空间。首先，在情感智能方面，开放领域对话系统的策略规划仍需要进一步的探索，本文中我们仅根据对话上文进行情感支持策略的预测，未来我们会尝试结合对用户反馈的预测，以进一步提升策略规划的合理、有效性，从而提升对话系统提供情感支持的能力。其次，知识性对于提高情绪支持对话系统的语义理解能力，和提供合理、有效建议的能力具有重要意义，本文主要通过弱语义的数据驱动的方法提升情绪支持对话系统的知识性，未来我们会进一步尝试与知识图谱、知识文档的结合，以强语义的方法

提升情绪支持对话系统的知识性。再次，安全性对于情绪支持对话系统的落地应用也具有重要的意义，我们应避免对话系统产生出不安全的回复，表现出冒犯用户、风险忽视等危险行为。最后，如何使对话系统同时具有以上多项对话能力，避免能力之间的冲突与遗忘，使其在对话过程中能够根据用户的输入，自然地组合、切换能力，也是一项具有挑战性的问题。近期，基于适配器 (Adapter)^[85]、提示 (prompt)^[86] 等参数高效的微调方法初步展现出了在大规模预训练模型的基础上构建多技能模型的可能性，未来我们也将尝试从这一视角寻求突破。

插图索引

图 1.1	不同情感智能等级的对话样例。	2
图 2.1	编码器-解码器框架。	6
图 2.2	基于 Transformer 的编码器-解码器模型 ^[31] 。	8
图 2.3	EVA 模型（28 亿参数）与人类的交互样例 ^[11] 。	10
图 2.4	情绪支持三阶段理论模型 ^[26] 。	12
图 3.1	CESConv 数据集的对话样例。	15
图 4.1	加入增广数据能够提升对话模型面对多样话题的鲁棒性。	18
图 4.2	对话数据增广框架。	19
图 4.3	增广数据样例。	25
图 5.1	在不同的策略规划下的对话样例。	28
图 5.2	融合策略的可控回复生成框架。	29
图 5.3	总策略分布特征。	32
图 5.4	不同对话阶段的策略分布特征。	32
图 5.5	修正策略分布后得到的 D_{plan} 数据集的策略分布。	34
图 5.6	回复的可控性样例分析。	36
图 5.7	融合情绪支持策略的对话系统与人类用户的真实交互样例。	37

表格索引

表 2.1	代表性的大规模对话预训练模型汇总。	9
表 2.2	EVA 模型 (28 亿参数) 的失败对话样例 ^[11] 。	11
表 3.1	CESConv 数据集的统计结果。	15
表 3.2	CESConv 数据集的人工评价。	17
表 4.1	从“微博树洞”账号爬取的原始数据样例。	21
表 4.2	对原始数据和增广数据的基础信息统计和多样性统计。	23
表 4.3	对原始数据和增广数据的人工评价。	24
表 4.4	人工交互评价结果。	26
表 5.1	八种情绪支持策略。	28
表 5.2	策略规划数据集 D_{plan} 和策略可控的回复生成数据集 D_{gen} 的统计信息。 33	
表 5.3	自动评价结果。	35
表 5.4	人工交互评价实验结果	36

参考文献

- [1] Weizenbaum J. Eliza—a computer program for the study of natural language communication between man and machine[J]. *Communications of the ACM*, 1966, 9(1): 36-45.
- [2] Colby K M. Artificial paranoia: a computer simulation of paranoid process[M]. Pergamon Press, 1975.
- [3] Wallace R S. The anatomy of alice in: Epstein, r., roberts, g., beber, g.(eds.) parsing the turing test[M]. Springer, Dordrecht, 2009.
- [4] Adiwardana D, Luong M T, So D R, et al. Towards a human-like open-domain chatbot[J]. arXiv preprint arXiv:2001.09977, 2020.
- [5] Xu J, Ju D, Li M, et al. Recipes for safety in open-domain chatbots[J]. arXiv preprint arXiv:2010.07079, 2020.
- [6] Zhang Y, Sun S, Galley M, et al. Dialogpt: Large-scale generative pre-training for conversational response generation[J]. arXiv preprint arXiv:1911.00536, 2019.
- [7] Bao S, He H, Wang F, et al. Plato: Pre-trained dialogue generation model with discrete latent variable[J]. arXiv preprint arXiv:1910.07931, 2019.
- [8] Bao S, He H, Wang F, et al. Plato-2: Towards building an open-domain chatbot via curriculum learning[J]. arXiv preprint arXiv:2006.16779, 2020.
- [9] Bao S, He H, Wang F, et al. Plato-xl: Exploring the large-scale pre-training of dialogue generation[J]. arXiv preprint arXiv:2109.09519, 2021.
- [10] Zhou H, Ke P, Zhang Z, et al. Eva: An open-domain chinese dialogue system with large-scale generative pre-training[J]. arXiv preprint arXiv:2108.01547, 2021.
- [11] Gu Y, Wen J, Sun H, et al. Eva2. 0: Investigating open-domain chinese dialogue systems with large-scale pre-training[J]. arXiv preprint arXiv:2203.09313, 2022.
- [12] Salovey P, Mayer J D. Emotional intelligence[J]. *Imagination, cognition and personality*, 1990, 9(3): 185-211.
- [13] Picard R W, Vyzas E, Healey J. Toward machine emotional intelligence: Analysis of affective physiological state[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2001, 23(10): 1175-1191.
- [14] Polzin T S, Waibel A. Emotion-sensitive human-computer interfaces[C]//ISCA tutorial and research workshop (ITRW) on speech and emotion. 2000.

- [15] Partala T, Surakka V. The effects of affective interventions in human–computer interaction[J]. *Interacting with computers*, 2004, 16(2): 295-309.
- [16] Prendinger H, Mori J, Ishizuka M. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game[J]. *International journal of human-computer studies*, 2005, 62(2): 231-245.
- [17] Martinovsky B, Traum D. The error is the clue: Breakdown in human-machine interaction [R]. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE ···, 2006.
- [18] Skowron M. Affect listeners: Acquisition of affective states by means of conversational systems [M]//*Development of Multimodal Interfaces: Active Listening and Synchrony*. Springer, 2010: 169-181.
- [19] Zhou H, Huang M, Zhang T, et al. Emotional chatting machine: Emotional conversation generation with internal and external memory[C]//*Proceedings of the AAAI Conference on Artificial Intelligence: volume 32*. 2018.
- [20] Huang M, Zhu X, Gao J. Challenges in building intelligent open-domain dialog systems[J]. *ACM Transactions on Information Systems (TOIS)*, 2020, 38(3): 1-32.
- [21] Rashkin H, Smith E M, Li M, et al. Towards empathetic open-domain conversation models: A new benchmark and dataset[J]. *arXiv preprint arXiv:1811.00207*, 2018.
- [22] Lin Z, Madotto A, Shin J, et al. Moel: Mixture of empathetic listeners[J]. *arXiv preprint arXiv:1908.07687*, 2019.
- [23] Zandie R, Mahoor M H. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems[C]//*The Thirty-Third International Flairs Conference*. 2020.
- [24] Majumder N, Hong P, Peng S, et al. Mime: Mimicking emotions for empathetic response generation[J]. *arXiv preprint arXiv:2010.01454*, 2020.
- [25] Zheng C, Liu Y, Chen W, et al. Comae: a multi-factor hierarchical framework for empathetic response generation[J]. *arXiv preprint arXiv:2105.08316*, 2021.
- [26] Liu S, Zheng C, Demasi O, et al. Towards emotional support dialog systems[J]. *arXiv preprint arXiv:2106.01144*, 2021.
- [27] Roller S, Dinan E, Goyal N, et al. Recipes for building an open-domain chatbot[J]. *arXiv preprint arXiv:2004.13637*, 2020.
- [28] 傅小兰, 张侃, 陈雪峰, 等. 中国国民心理健康发展报告 (2017-2018)[M]. She hui ke xue wen xian chu ban she· qun xue chu ban fen she, 2019.
- [29] Hill C E. *Helping skills: Facilitating, exploration, insight, and action*[M]. American Psychological Association, 2009.

- [30] Du Z, Qian Y, Liu X, et al. All nlp tasks are generation tasks: A general pretraining framework [J]. arXiv preprint arXiv:2103.10360, 2021.
- [31] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [32] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.
- [33] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
- [34] Shang L, Lu Z, Li H. Neural responding machine for short-text conversation[J]. arXiv preprint arXiv:1503.02364, 2015.
- [35] Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.
- [36] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
- [37] Thoppilan R, De Freitas D, Hall J, et al. Lamda: Language models for dialog applications[J]. arXiv preprint arXiv:2201.08239, 2022.
- [38] Wang Y, Ke P, Zheng Y, et al. A large-scale chinese short-text conversation dataset[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Springer, 2020: 91-103.
- [39] Zhou K, Prabhumoye S, Black A W. A dataset for document grounded conversations[J]. arXiv preprint arXiv:1809.07358, 2018.
- [40] Moghe N, Arora S, Banerjee S, et al. Towards exploiting background knowledge for building conversation systems[J]. arXiv preprint arXiv:1809.08205, 2018.
- [41] Dinan E, Roller S, Shuster K, et al. Wizard of wikipedia: Knowledge-powered conversational agents[J]. arXiv preprint arXiv:1811.01241, 2018.
- [42] Qin L, Galley M, Brockett C, et al. Conversing by reading: Contentful neural conversation with on-demand machine reading[J]. arXiv preprint arXiv:1906.02738, 2019.
- [43] Gopalakrishnan K, Hedayatnia B, Chen Q, et al. Topical-chat: Towards knowledge-grounded open-domain conversations.[C]//INTER_SPEECH. 2019: 1891-1895.
- [44] Moon S, Shah P, Kumar A, et al. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 845-854.
- [45] Wu W, Guo Z, Zhou X, et al. Proactive human-machine conversation with explicit conversation goals[J]. arXiv preprint arXiv:1906.05572, 2019.

- [46] Tuan Y L, Chen Y N, Lee H y. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs[J]. arXiv preprint arXiv:1910.00610, 2019.
- [47] Zheng Y, Chen G, Huang M, et al. Personalized dialogue generation with diversified traits[J]. arXiv preprint arXiv:1901.09672, 2019.
- [48] Zhang S, Dinan E, Urbanek J, et al. Personalizing dialogue agents: I have a dog, do you have pets too?[J]. arXiv preprint arXiv:1801.07243, 2018.
- [49] Wang X, Shi W, Kim R, et al. Persuasion for good: Towards a personalized persuasive dialogue system for social good[J]. arXiv preprint arXiv:1906.06725, 2019.
- [50] Zhou H, Young T, Huang M, et al. Commonsense knowledge aware conversation generation with graph attention.[C]//IJCAI. 2018: 4623-4629.
- [51] Lian R, Xie M, Wang F, et al. Learning to select knowledge for response generation in dialog systems[J]. arXiv preprint arXiv:1902.04911, 2019.
- [52] Zheng C, Cao Y, Jiang D, et al. Difference-aware knowledge selection for knowledge-grounded conversation generation[J]. arXiv preprint arXiv:2009.09378, 2020.
- [53] Yu D, Yu Z. Midas: A dialog act annotation scheme for open domain human machine spoken conversations[J]. arXiv preprint arXiv:1908.10023, 2019.
- [54] Xu C, Wu W, Wu Y. Towards explainable and controllable open domain dialogue generation with dialogue acts[J]. arXiv preprint arXiv:1807.07255, 2018.
- [55] Jacobs R A, Jordan M I, Nowlan S J, et al. Adaptive mixtures of local experts[J]. Neural computation, 1991, 3(1): 79-87.
- [56] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.
- [57] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
- [58] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [59] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. Advances in neural information processing systems, 2019, 32.
- [60] Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[J]. arXiv preprint arXiv:1910.13461, 2019.
- [61] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. arXiv preprint arXiv:1910.10683, 2019.

- [62] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.
- [63] Sun Y, Wang S, Li Y, et al. Ernie 2.0: A continual pre-training framework for language understanding[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. 2020: 8968-8975.
- [64] Sun Y, Wang S, Feng S, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation[J]. arXiv preprint arXiv:2107.02137, 2021.
- [65] Zhang Z, Han X, Zhou H, et al. Cpm: A large-scale generative chinese pre-trained language model[J]. AI Open, 2021, 2: 93-99.
- [66] Zhang Z, Gu Y, Han X, et al. Cpm-2: Large-scale cost-effective pre-trained language models [J]. AI Open, 2021, 2: 216-224.
- [67] Du Z, Qian Y, Liu X, et al. Glm: General language model pretraining with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 320-335.
- [68] Zeng W, Ren X, Su T, et al. Pangu- α : Large-scale autoregressive pretrained chinese language models with auto-parallel computation[J/OL]. arXiv preprint arXiv:2104.12369, 2021. <https://arxiv.org/abs/2104.12369>.
- [69] Wu S, Zhao X, Yu T, et al. Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning[J/OL]. arXiv preprint arXiv:2110.04725, 2021. <https://arxiv.org/abs/2110.04725>.
- [70] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint arXiv:2204.02311, 2022.
- [71] West P, Bhagavatula C, Hessel J, et al. Symbolic knowledge distillation: from general language models to commonsense models[J]. arXiv preprint arXiv:2110.07178, 2021.
- [72] Sap M, Le Bras R, Allaway E, et al. Atomic: An atlas of machine commonsense for if-then reasoning[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 33. 2019: 3027-3035.
- [73] Liu Z, Patwary M, Prenger R, et al. Multi-stage prompting for knowledgeable dialogue generation[J]. arXiv preprint arXiv:2203.08745, 2022.
- [74] Fleiss J L. Measuring nominal scale agreement among many raters.[J]. Psychological bulletin, 1971, 76(5): 378.
- [75] Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the national academy of sciences, 2017, 114(13): 3521-3526.

- [76] Holtzman A, Buys J, Du L, et al. The curious case of neural text degeneration[J]. arXiv preprint arXiv:1904.09751, 2019.
- [77] Liu S, Sabour S, Zheng Y, et al. Rethinking and refining the distinct metric[J]. arXiv preprint arXiv:2202.13587, 2022.
- [78] Li J, Galley M, Brockett C, et al. A diversity-promoting objective function for neural conversation models[J]. arXiv preprint arXiv:1510.03055, 2015.
- [79] Li M, Weston J, Roller S. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons[J]. arXiv preprint arXiv:1909.03087, 2019.
- [80] Tu Q, Li Y, Cui J, et al. Misc: A mixed strategy-aware model integrating comet for emotional support conversation[J]. arXiv preprint arXiv:2203.13560, 2022.
- [81] Keskar N S, McCann B, Varshney L R, et al. Ctrl: A conditional transformer language model for controllable generation[J]. arXiv preprint arXiv:1909.05858, 2019.
- [82] Zheng C, Huang M. Exploring prompt-based few-shot learning for grounded dialog generation [J]. arXiv preprint arXiv:2109.06513, 2021.
- [83] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.
- [84] Lin C Y. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out. 2004: 74-81.
- [85] Houlshy N, Giurigu A, Jastrzebski S, et al. Parameter-efficient transfer learning for nlp[C]// International Conference on Machine Learning. PMLR, 2019: 2790-2799.
- [86] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[J]. arXiv preprint arXiv:2104.08691, 2021.

致 谢

衷心感谢黄民烈副教授对我的精心指导。自我进入实验室以来，黄老师对我的信任与支持，让我不断成长。黄老师在科研工作和团队领导的过程中所展现出的勤奋、严谨、不畏困难的精神与学者的情怀，一直深深激励着我，使我终生受益。同时我也要感谢 CoAI 课题组全体同学的帮助与支持，与你们一同科研、探索未知的问题，是幸运与快乐的。

感谢陪伴我走过大学时光的可爱的朋友们，希望我们的友谊永远延续，也祝愿我们都能走出更加广阔的人生。

感谢我的家人。是你们无条件的付出与温暖的爱，让我在面对任何困难挫折时都能保持对生活的热爱与信心。感谢小姨、小姨夫让我在北京求学期间也能感到家的温暖。我还尤其要感谢我的母亲，我今天取得的所有成绩，我身上所有美好的品质，都离不开她的教导。从前我只看得到她作为母亲的坚韧与强大，但渐渐的我开始意识到她作为一个普通女性的脆弱。希望我能早日担起家庭的责任，成为她的依靠。

感谢我的女朋友对我的爱与包容。我从前一直不觉得爱情有什么意义，但如今竟已变成一个“瘸腿”的人、一个小孩子，离不开她的扶持与照顾，离不开她带给我的快乐、勇气和决心。未来的日子里有太多的不确定性和挑战，希望我们能一直在一起。

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名： 温佳鑫 日 期： 2022年6月12日

附录 A 外文资料的调研阅读报告

Towards Emotional Intelligence in Open-Domain Dialogue Systems

Contents

A.1	Challenges in Open-Domain Dialogue Systems.....	53
A.2	Emotional Intelligence in Open-Domain Dialogue Systems.....	54

A.1 Challenges in Open-Domain Dialogue Systems

Before the emergence of large-scale pre-trained dialogue models, it is a typical phenomenon for the dialogue models based on the recurrent neural network(RNN) to generate generic, short responses such as "I don't know" and "yes". Thus the main challenge at that time is to improve the relevance, informativeness, and diversity of the generated responses. However, these issues are significantly improved in large-scale pre-trained dialogue models through scaling up data size and model size. Standing on the large-scale pre-trained open-domain dialogue models, researchers start to aim at higher-level conversational skills, in pursuit of a more humanized dialogue system. A series of works^[1-3] point out that the existing large-scale pre-trained dialogue models still get unsatisfactory performance in advanced conversational skills such as knowledge, persona consistency, safety, and emotional intelligence. To address these challenges, researchers have constructed many high-quality dialogue datasets^[2,4-15,26] targeting specific conversational skills for fine-tuning the large-scale pre-trained dialogue models on the one hand. On the other hand, instead of limiting to a data-driven end-to-end dialogue system, researchers have also sought breakthroughs in model designs, such as incorporating external knowledge^[16-17,32] or dialogue strategies^[18-19,26], to further enhance the advanced conversational skills of large-scale pre-trained dialogue models.

A.2 Emotional Intelligence in Open-Domain Dialogue Systems

Embodying machines with emotional intelligence, an important feature of human intelligence^[20-21], is one of the long-standing research goals of artificial intelligence. There is a series of studies in the field of human-computer interaction^[22-26] and dialogue systems^[1,3,15,26-27,27-31] point out that embodying machines with emotional intelligence, making them be able to understand human emotions, express empathy, provide emotional support, or even psychological counseling, could significantly improve the long-term user engagement. We classify the level of emotional intelligence in dialogue systems into three levels: (1) Emotional Chatting: express specified emotions in responses. (2) Empathetic Chatting: understand human emotions and select the appropriate emotion for responses accordingly. (3) Emotional Support Conversation: deeply understand users' emotional situations through combining both listening and active exploration, achieve emotional reassurance, and provide effective suggestions and solutions to help the users solve their emotional problems. With the improvement of the level of emotional intelligence, the interactivity of dialogue systems is significantly enhanced. In particular, an emotional support dialog system that can effectively channel negative emotions also has an important social value in the context of today's generally stressful society and increasingly serious mental health problems. Therefore, the emotional intelligence of dialogue systems is getting increasing attention and there has already been a number of exciting works.

Zhou et al.^[27] first define the Emotional Chatting task and propose an emotion-controllable response generation model called Emotional Chatting Machine(ECM). It is worth mentioning that Zhou et al.^[27] further presents an analysis of emotion interaction in the dialog dataset and finds there are some typical patterns of the correspondence between the emotion category of the context and that of the response. For instance, the emotion category of the context has a high probability of being the same as that of the response. And when the emotion category of the context is sad, the probability of responding with the emotion category of like is also high. The existence of such emotion interaction patterns leads to the study of Empathetic Dialogue, whose goal is to enable the dialog model itself to understand users' emotion and then to select the appropriate emotion for guiding the generation of responses. Rashkin et al.^[15] constructs a novel

dataset called EMPATHETICDIALOG and proposes an emotion-aware response generation framework for empathetic dialog generation. Specifically, they first leverage an emotion classifier to identify the emotion of the context, and then both the identified emotion and the context are fed into the dialogue model to generate the response. Subsequently, a series of works make incremental improvements based on this framework^[28-31]. For instance, Lin et al.^[28] proposes Mixture of Empathetic Listeners(MoEL), which introduces Mixture-of-Experts^[32] for modeling empathy in dialogue systems. Although empathy improves the emotional intelligence of a dialogue system to a certain extent, it is still difficult to effectively channel users' negative emotions with empathy alone. In order to achieve emotional support, the dialogue model should also combine active explorations with listening to deeply understand the user's emotional situation and provide reasonable suggestions and information to help users solve their emotional problems. To this end, Liu et al.^[26] simplify the Helping Skills Theory^[29] in the field of psychology into the ESC framework. The ESC framework consists of three stages, namely explorations, comforting and action, and each of these three stages contains several emotional support strategies. Under the guidance of the ESC framework,^[26] defines the Emotional Support Conversation task and collects a high-quality English Emotional Support Conversation dataset called ESConv. Automatic and human evaluations show that: (1) After fine-tuning on the ESConv dataset, the ability to provide effective emotional support of a powerful large-scale pre-trained dialogue model Blender^[2] is significantly improved. (2) Incorporating the modeling of emotional support strategies, i.e., explicitly selecting the appropriate emotional support strategies and then generating the response conditioned on both the context and the selected strategies, can further enhance the dialogue model to provide more effective emotional support. Overall, these two findings illustrate the importance of high-quality emotional support dialogue data and modeling of emotional support strategies.

参考文献

- [1] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32, 2020.
- [2] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*, 2020.

- [3] Yuxian Gu, Jiaxin Wen, Hao Sun, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Xiaoyan Zhu, Jie Tang, et al. Eva2. 0: Investigating open-domain chinese dialogue systems with large-scale pre-training. *arXiv preprint arXiv:2203.09313*, 2022.
- [4] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*, 2018.
- [5] Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. Towards exploiting background knowledge for building conversation systems. *arXiv preprint arXiv:1809.08205*, 2018.
- [6] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018.
- [7] Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. Conversing by reading: Contentful neural conversation with on-demand machine reading. *arXiv preprint arXiv:1906.02738*, 2019.
- [8] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895, 2019.
- [9] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, 2019.
- [10] Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. Proactive human-machine conversation with explicit conversation goals. *arXiv preprint arXiv:1906.05572*, 2019.
- [11] Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. *arXiv preprint arXiv:1910.00610*, 2019.
- [12] Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*, 2019.
- [13] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.
- [14] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*, 2019.

- [15] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018.
- [16] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629, 2018.
- [17] Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. Difference-aware knowledge selection for knowledge-grounded conversation generation. *arXiv preprint arXiv:2009.09378*, 2020.
- [18] Dian Yu and Zhou Yu. Midas: A dialog act annotation scheme for open domain human machine spoken conversations. *arXiv preprint arXiv:1908.10023*, 2019.
- [19] Can Xu, Wei Wu, and Yu Wu. Towards explainable and controllable open domain dialogue generation with dialogue acts. *arXiv preprint arXiv:1807.07255*, 2018.
- [20] Peter Salovey and John D Mayer. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211, 1990.
- [21] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.
- [22] Thomas S Polzin and Alexander Waibel. Emotion-sensitive human-computer interfaces. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.
- [23] Timo Partala and Veikko Surakka. The effects of affective interventions in human–computer interaction. *Interacting with computers*, 16(2):295–309, 2004.
- [24] Helmut Prendinger, Junichiro Mori, and Mitsuru Ishizuka. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International journal of human-computer studies*, 62(2):231–245, 2005.
- [25] Bilyana Martinovsky and David Traum. The error is the clue: Breakdown in human-machine interaction. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE ..., 2006.
- [26] Marcin Skowron. Affect listeners: Acquisition of affective states by means of conversational systems. In *Development of Multimodal Interfaces: Active Listening and Synchrony*, pages 169–181. Springer, 2010.
- [27] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

- [28] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*, 2019.
- [29] Rohola Zandie and Mohammad H Mahoor. Empransfo: A multi-head transformer architecture for creating empathetic dialog systems. In *The Thirty-Third International Flairs Conference*, 2020.
- [30] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Mime: Mimicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454*, 2020.
- [31] Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. Comae: a multi-factor hierarchical framework for empathetic response generation. *arXiv preprint arXiv:2105.08316*, 2021.
- [32] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

在学期间参加课题的研究成果

个人简历

1999年2月20日出生于吉林省梅河口市。

2017年8月考入清华大学土木大类，2018年5月转入清华大学计算机科学与技术系计算机科学与技术专业，攻读工学学士学位至今。


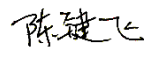
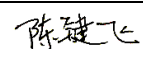
学术论文列表^①

- [1] Jiexi Liu*, Ryuichi Takanobu*, **Jiaxin Wen**, et al. Robustness Testing of Language Understanding in Task-Oriented Dialog. In *ACL 2021* (CCF-A, Long Paper).
- [2] Zhexin Zhang*, **Jiaxin Wen***, Jian Guan, Minlie Huang. Persona-Guided Planning for Controlling the Protagonist's Persona in Story Generation. In *NAACL 2022* (CCF-C, Long Paper).
- [3] Chujie Zheng, Sahand Sabour, **Jiaxin Wen**, Minlie Huang. AugESC: Large-Scale Data Augmentation for Emotional Support Conversation with Pre-trained Language Models. In submission to *EMNLP 2022* (CCF-B, Long Paper).
- [4] Yuxian Gu*, **Jiaxin Wen***, Hao Sun*, et al. EVA2.0: Investigating Open-Domain Chinese Dialogue Systems with Large-Scale Pre-Training. In submission to *EMNLP 2022* (CCF-B, Long Paper).

^① * 表示同等贡献

综合论文训练记录表

学生姓名	温佳鑫	学号	2017010335	班级	计 84
论文题目	构建中文情绪支持对话系统				
主要内容以及进度安排	<p>心理咨询领域存在着需求与供给的严重不平衡。本研究将通过基础资源的收集和关键技术的研发，构建首个中文情绪支持对话系统，及时、有效地疏导用户的负面情绪。</p> <p>2021.12~2022.01：收集数据集，微调模型，测试效果</p> <p>2022.01~2022.03：实现基于大规模预训练语言模型的数据增广</p> <p>2022.03~2022.05：实现融合情绪支持策略的回复生成</p> <p>2022.05：完成论文撰写</p> <p>2022.06：毕设答辩</p> <div style="text-align: right; margin-top: 20px;"> 指导教师签字： <u>黄民烈</u> 考核组组长签字： <u>陈键飞</u> 2022年 1 月 4 日 </div>				
中期考核意见	<p>设计了利用大模型和社交媒体爬取数据的数据增广算法，对模型质量提升显著。提出了融合情绪支持策略的可控回复生成，取得了一定的效果。所提出方法已上线。目前完成度良好。</p> <div style="text-align: right; margin-top: 20px;"> 考核组组长签字： <u>陈键飞</u> 2022年 3 月 29 日 </div>				

<p style="text-align: center;">指导教师评语</p>	<p>情感对话系统具有重要的学术和应用价值，论文研究了中文情感支持对话系统的关键技术，提出了利用策略控制回复生成的方法，并实现了实际部署，顺利完成了综合论文训练的任务。论文写作规范，结构合理，创新型较强。</p> <p style="text-align: right;">指导教师签字： </p> <p style="text-align: right;">2022年 6 月 9 日</p>
<p style="text-align: center;">评阅教师评语</p>	<p>针对情感支持这一实际需求，作者从头开始，独立搭建了一个对话系统。论文在数据的收集方面有较大贡献，设计了数据增广方法，得到的数据集有一定价值。所构建的系统能够在短时间内上线应用，产生较大影响力。论文体现了作者很强的动手能力，是一篇优秀的毕业设计论文。</p> <p style="text-align: right;">评阅教师签字：  _____</p> <p style="text-align: right;">2022年 6 月 9 日</p>
<p style="text-align: center;">答辩小组评语</p>	<p>搭建了一个中文情绪支持对话系统，工作包括数据生成，模型搭建，实验等，取得了较好的效果。论文体现了作者良好的动手能力和解决实际问题的能力。答辩小组一致同意答辩通过。</p> <p style="text-align: right;">答辩小组组长签字：  _____</p> <p style="text-align: right;">2022年 6 月 9 日</p>

总成绩： A+

教学负责人签字： 黄烈

2022年 6 月 9 日