计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：面向对话生成和心理疾病检测的心理咨询式人机对话数据集构建
作者：徐博，郝德志，于迩晨，林鸿飞，宗林林
收稿日期：2024-05-29
网络首发日期：2024-09-18
引用格式：徐博，郝德志，于迩晨，林鸿飞，宗林林．面向对话生成和心理疾病检测的心理咨询式人机对话数据集构建[J/OL]．计算机应用．
https://link.cnki.net/urlid/51.1307.TP.20240918.1019.008

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

doi:10.11772/j.issn.1001-9081.2024050705

面向对话生成和心理疾病检测的心理咨询式人机对话数据集构建

徐博[1*]，郝德志[1]，于迩晨[1]，林鸿飞[1]，宗林林[2]

(1.大连理工大学 计算机科学与技术学院，辽宁 大连 116023；
2.大连理工大学 软件学院，辽宁 大连 116023)

(*通信作者电子邮箱 xubo@dlut.edu.cn)

摘 要: 针对心理咨询式人机对话中缺乏用于建立有效对话模型的公开数据的问题，构建一种面向对话生成和心理疾病监测的心理医疗咨询对话数据集。首先通过在线医疗问诊平台获取到包含3268个医生和患者之间的多轮对话的数据集，并附有广泛的相关元数据，包括就诊医院、科室、疾病类型和患者自我陈述等。其次，提出一个知识增强的对话模型：情感感知双向自回归模型（EmBART），以增强对话模型的共情能力。最后，通过心理医疗响应生成和心理疾病检测进行了数据集可用性的实验评估。在心理医疗响应生成中，基于本数据集训练的EmBART模型在自动评估与人工评估中的各项指标上均表现出色，其中困惑度较基准模型降低了2.31；在心理疾病检测中，基于本数据集训练的CPT（Chinese Pre-trained Unbalanced Transformer）和RoBERTa（Robustly optimized BERT approach）模型具有出色的心理疾病预测能力。实验结果表明，本数据集在生成共情对话和预测心理疾病方面具有较强的实用性，为未来基于心理咨询式人机对话研究提供了数据基础。

关键词: 心理咨询对话；心理疾病检测；对话生成；共情响应；情感分析

中图分类号:TP391.1 **文献标志码：A**

Psychological counseling human-machine dialogue dataset construction for dialogue generation and mental disorder detection

XU Bo[1*], HAO Dezhi[1], YU Erchen[1], LIN Hongfei[1], ZONG Linlin[2]

(1.School of Computer Science and Technology, Dalian University of Technology, Dalian Liaoning 116023, China;
2. School of Software, Dalian University of Technology, Dalian Liaoning 116023, China)

Abstract: To address the problem of the lack of publicly available data for modeling effective dialogs in psychological counseling human-machine dialogue, a psychomedical counseling dialog dataset for dialogue generation and mental disorder detection was constructed. First, the dataset containing 3,268 multi-round conversations between doctors and patients were obtained, with a wide range of relevant metadata, including the hospitals visited, departments, types of disorders, and patients' self-representations, acquired through online medical consultation platforms. Then, a knowledge-enhanced dialog model, Empathy Bidirectional and Auto-Regressive Transformers (EmBART), was proposed to enhance the empathic capability of the dialog model. Finally, an experimental evaluation of the dataset usability was conducted through psychomedical response generation and mental illness detection. In psychomedical response generation, EmBART trained on this dataset performed excellently on all metrics in both automatic and human evaluations, with Perplexity reduced by 2.31 compared to baseline model. In mental illness detection, CPT (Chinese Pre-trained Unbalanced Transformer) and RoBERTa (Robustly optimized Bidirectional Encoder Representations from Transformers approach) trained on this dataset demonstrated outstanding mental illness prediction capabilities. Experimental results demonstrated the strong utility of this dataset in generating empathic dialogues and predicting mental illnesses, and provided a data base for future research on counseling human-machine dialogues.

收稿日期: 2024-5-29; 修回日期:2024-08-02; 录用日期: 2024-08-20。
基金项目: 辽宁省社会科学规划基金(L21CXW003)

作者简介: 徐博(1988—)，男，辽宁大连人，副教授，博士，CCF会员，主要研究方向：心理健康计算、自然语言处理；郝德志(1998—)，男，山东临沂人，硕士，主要研究方向：心理咨询式人机对话；于迩晨(2001—)，男，辽宁鞍山人，硕士研究生，主要研究方向：多模态情感计算；林鸿飞(1962—)，男，内蒙古通辽人，教授，工学博士，主要研究方向：自然语言处理；宗林林(1987—)，女，河北沧州人，副教授，工学博士，主要研究方向：多模态情感计算。

Keywords: psychological counseling dialogue; mental disorder detection; dialogue generation; natural language processing; emotion analysis

0 引言

根据世界卫生组织[1]报告，新冠疫情爆发以来，全球内焦虑症和抑郁症发病率已经上升了 25%。值得注意的是，在新型冠状病毒后疫情时代，心理障碍在医疗咨询中的占比显著上升。研究人员正致力于探索更为有效的心理疾病诊断方法。虽然心理自评量表，如患者健康问卷[2]（Patience Health Questionnaire-9, PHQ-9），被广泛用于临床心理疾病的诊断，但近年来，它在应对迅速升级的在线心理咨询需求方面正面临着严峻挑战。为有效应对这一挑战，心理咨询人机对话应运而生。

心理对话系统是辅助早期诊断心理健康问题的有效工具。通过使用这些对话系统进行在线交流，用户可以通过与机器的多次互动获取必要的信息，从而大大减轻心理学专家面对面咨询的工作量。研究表明，针对有情绪障碍的人而言[3]，由于心理疾病所带来的耻辱感和面对面交流所带来的压力，用户更倾向于向对话系统披露自己的心理困扰。

为了解决在构建有效的心理医疗对话模型过程中数据匮乏的问题，本文构建了一个大规模且高质量的心理医疗对话数据集，即 PsychDialog （Psychology Dialog）。该数据集的原始数据来源于三个著名的中国健康社区平台的心理和精神科论坛：好大夫、寻医问药和拇指医生。原始数据由医疗领域专家采用标准化策略进行了细致的过滤、清理和标签注释。最终产生的数据集包含 3268 个医生与患者的对话，涵盖 11 种心理疾病。此外，该数据集还包含与这些对话相关的各种元数据，包括就诊医院、科室、11 种疾病类型以及患者自我报告的详细信息。

为了捕捉心理医疗对话中细微的情感线索，本文提出了一种情感感知双向编码与自解码模型（Empathy Bidirectional and Auto-Regressive Transformers，EmBART）。该模型通过融入情感知识来生成共情响应，从而增强模型的共情表达能力。最后，本文使用 PsychDialog 对两个关键任务进行了评估：心理医疗响应生成和心理疾病预测。实验结果表明，PsychDialog 在训练各种任务的对话模型方面非常有效。此外，本文所提出的 EmBART 模型在该数据集中建立了基准结果，为今后利用 PsychDialog 数据集开展研究奠定了基础。

本文的主要贡献有 3 个方面:

(1)构建了一个中文心理医疗对话数据集--PsychDialog，其中包含 3,268 个医患对话以及这些对话的丰富元数据。

(2)本文为所有对话数据标注了共 11 种特定的心理疾病类型，同时提出用于产生共情响应的新模型 EmBART，结合对话元数据，可以开发出个性化的心理咨询聊天机器人，满足多样化心理诉求，提高心理健康服务效率。

(3)本文使用预训练模型进行心理疾病的分类和消融实验，为未来的研究提供基线结果。

1 相关工作

医学对话研究与心理咨询人机对话密切相关，反映了该领域最相关的研究进展。最近的研究主要集中在采用基于神经网络的模型。例如，为自动诊断而开发了一个端到端医疗对话系统[4]，通过病人与机器之间的互动来获取病人的健康信息。该系统可根据患者的自我报告自主提供诊断建议。此外，Xia 等[5]利用强化学习技术提出了一种自动对话诊断系统。同样，Wei 等[6]提出了一种面向任务的自动诊断对话系统，该系统利用话语值来预测用户意图并监测用户的健康状况，从而生成医疗回复。除了这些进展之外，在医疗对话场景中还探索了下游自然语言处理任务，包括信息提取[7]、关系预测[8]和槽填充[9]。

最近，大语言模型如 ChatGPT （Chat Generative Pre-trained Transformer）等在通用领域的人机对话中取得了重大进展，极大地改变了信息获取方法。这些系统不仅在模仿人类语言生成方面表现出色，而且在进行语言分析和优化方面也很出色。大型语言模型[10]已经证明了它们在生成适应特定语境的回复方面的功效。通过对 GPT-2（Generative Pre-trained Transformer 2.0）[11]等开源大型模型进行微调，可以针对特定任务设计对话模型。然而，尽管这些模型在提高对话性能方面很有效，但将特定领域的属性因素纳入某些任务（如心理医疗对话）仍具有挑战性。心理咨询对话系统需要类似于临床心理学家的理解能力，在提供准确的心理咨询和治疗建议的同时，系统还必须具备换位思考和投入情感的能力，以避免产生机械或不近人情的反应。此外，有效的心理咨询对话系统必须具备理解复杂语义的能力，以辨别用户微妙的心理诉求。

数据的匮乏严重阻碍了对话系统的开发，尤其是在心理医疗[12]对话领域。为了缓解这一问题，以往的研究建立了一些医疗对话数据集，包括百度拇指医生数据集（Baidu Muzhi Doctor，MZ）[6]，丁香医生数据集（Dingxiang Doctor，DX）[4]与 CMDD（Chinese Medical Dialogue Dataset）[13]等。Yang[14]等人收集了两个有关 COVID-19 的医学对话数据集 COVID-EN （CovidDialog-English）和 COVID-CN（CovidDialog-Chinese）。Zeng[15]等人提出了最大的医学对话数据集 Medialog-CN （Large-scale Medical Dialogue Datasets-Chinese）和 Medialog-EN （Large-scale Medical Dialogue Datasets-English），数据集涵盖了来自在线咨询网站的 29 种疾病信息。然而，这些数据集仅包括从网页抓取的原始数据，没有人工标注，因此引起了人们对训练数据质量的担忧。与此类似，Yao 等[16]构建了一个以抑郁症聊天为主题的中文对话数据集 D^4（A Chinese Dialogue Dataset for Depression-Diagnosis-Oriented Chat），Liu 等[17]则深入研究了情感支持对话系统，以更细致的情感语义来丰富医学对话。

Gratch 等[18]收集了一个抑郁症专用数据集 DAIC-WOZ（Distress Analysis Interview Corpus – Wizard of Oz），其中包含 189 个与抑郁症相关的对话，Saha 等开发了 MotiVAte（A Large-Scale Dataset for Motivational Dialogue System）[19]，这是一个面向抑郁症的对话数据集，专为构建虚拟助手而定制。虽然这些举措促进了抑郁症相关对话的研究，但对于构建强大的心理医学对话模型而言，仍然缺乏涵盖更广泛的心理对话数据。与现有数据集相比，本文数据集涵盖了更广泛的心理对话内容，此外，该数据集还包含全面的元数据标签，有助于对各种对话相关任务进行评估。为了全面了解数据集，表 1 列出了 PsychDialog 与现有数据集的对比分析。

表 1 相关医学对话数据集的比较

数据集名称	对话数量	语句数量	疾病数量	主要领域
MZ	710	-	4	儿科
DX	527	2168	5	儿科
CMDD	2067	87005	4	儿科
COVID-EN	603	-	1	COVID
COVID-CN	1088	-	1	COVID
DAIC-WOZ	189	-	1	抑郁症
MotiVAte	4000	14809	1	抑郁症
D^4	1339	81558	1	抑郁症
MedDialog-CN	3407494	11260564	174	普通医学领域
PsychDialog(ours)	**3268**	**18869**	**11**	**心理疾病领域**

2 数据集构建

2.1 数据收集

本文从三个在线医疗问诊平台收集原始数据：好大夫、寻医问药和拇指医生。这些平台提供在线医疗咨询服务，为患者面临的各种心理健康相关问题提供专业解决方案。获取的原始数据包含 10,000 次咨询中产生的大量问题和相应答案，还包含与这些咨询相关的大量元数据标签信息。

根据数据集初步分析显示，寻求心理健康援助的患者往往只能进行有限的几轮交谈。为了研究稳健的多轮对话模型，本文对原始数据进行了细致的筛选，保留了大约 6000 个包含两轮以上交流的对话。为了优化数据集内容，以便有效地构建对话模型，为此还收集了对话中的元数据标签信息以扩充数据集。这些元数据标签包括就诊医院、科室、疾病类型和患者自述信息。这些附加信息不仅能增强医生对病人心理状态的了解，在对话过程中提供充分的情感支持，还能丰富现有对话模型的泛化能力，使其具备专业的知识能力。

为确保数据一致性，还进行了数据结构的标准化。医疗在线平台优先保护用户隐私，在发布咨询内容前过滤了其中的敏感信息。为了进一步防范隐私风险，对收集到的数据进行了细致的筛选。具体的采用一套正则表达式规则，有针对性地提取潜在的隐私信息，包括手机号码和个人姓名等。这一严格的筛选过程确保了最终数据集不包含任何敏感的用户信息，从而大大降低了源数据的隐私泄露风险。这使得数据集泄露隐私信息的风险极低。最终数据集是合规且可公开的，不会伤害用户的隐私。图 1 展示了心理对话的构建和注释过程示意图，下面将进一步详细介绍。

2.2 数据注释

2.2.1 数据清洗

鉴于数据集来源于在线医疗问诊平台，对话内容包含大量口语表达，偶尔会出现语义模糊的情况。考虑到这种语言不规范可能会影响数据质量，在清洗过程中聘请了 12 名医学专业的研究生，在专业心理医生的指导下对数据进行细致的审查。审查过程旨在根据特定的对话场景和精神状态评估对话内容。根据中文医学主题词表制定的语言规范[20]，对偏离语言规范的对话内容进行了识别，随后将其删除。通过这种严格的质量筛选，共保留了 3268 条符合严格质量标准的对话，每条对话都包含两轮以上的回合。

2.2.2 注释策略

心理障碍的注释依据的是中文医学主题词表。每段对话的注释结果均来自对话内容和元数据信息。考虑到不同地域使用的疾病名称可能存在差异，本文将疾病名称标准化，以确保一致性。为确保疾病标签的统一性，聘请了医学专业的研究生在注释过程中对这些名称进行规范。例如，将获取的疾病类型 "烦躁、焦虑 "注释为 "焦虑症"，以建立统一的疾病标签。此外，作为注释过程的一部分，还严格筛选并删除了包含偏激词语和不规则字符的对话内容。因此，所构建的数据集包含 11 种心理疾病，即抑郁症（depression）、焦虑症(anxiety)、精神分裂症(schizophrenia)、心理障碍(mental disorder)、失眠症(insomnia)、强迫症(obsessive compulsive

disorder)、恐惧症(phobias)、多疑症(suspicious)、多动症(Attention Deficit and Hyperactive Disorder，ADHD)、自闭症(autism)和妄想症(paranoia)。这些疾病的分布如图 2 所示。

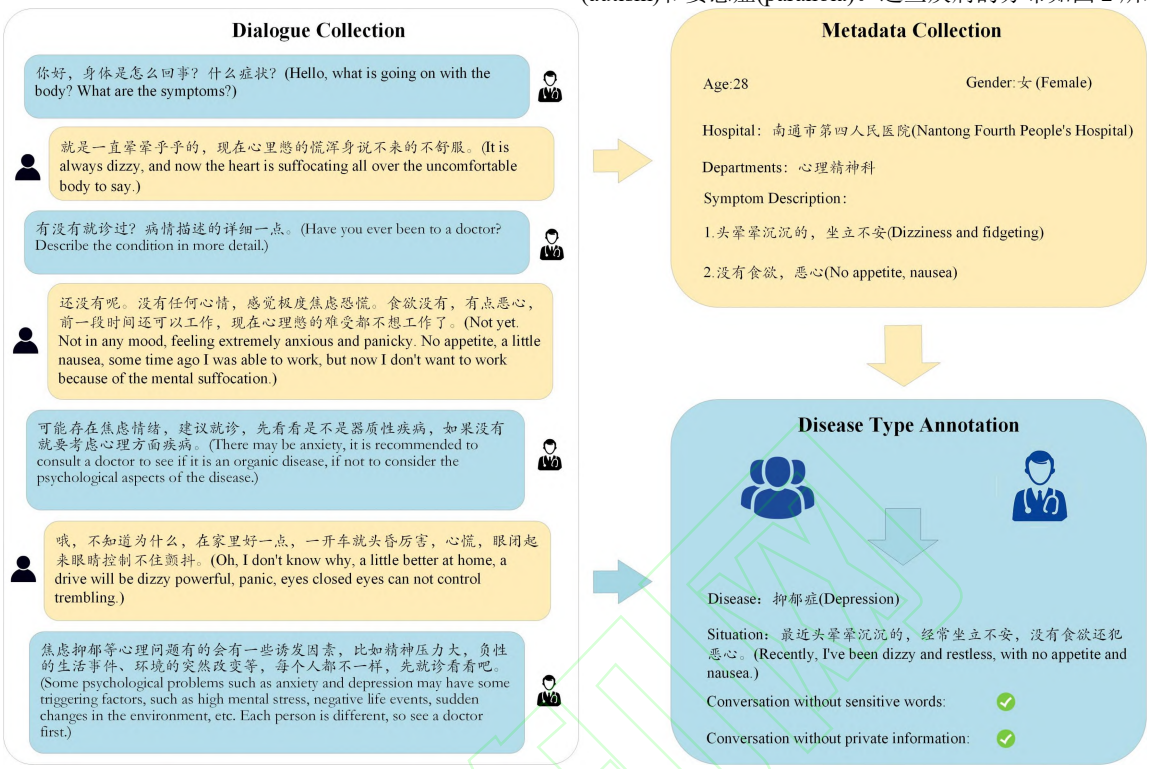

图 1 心理对话构建和注释过程示意图

Fig.1 Diagram illustrating the construction and annotation process of PsychDialog

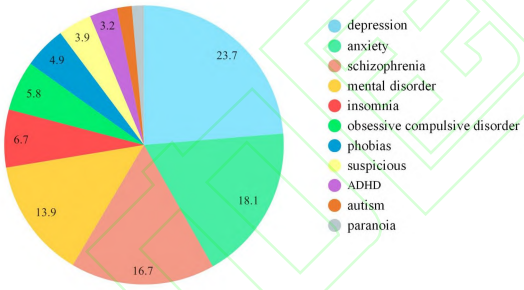

图 2 心理疾病分布比例

Fig.2 Distribution of various types of mental disorders

2.2.3 质量控制

注释过程历时三周，每周举行一次正式会议进行综合评估，旨在检查注释结果的准确性和一致性。为确保最大程度的准确性，聘请一位专业心理医生参与了有关注释结果的讨论，并在整个过程中解决新出现的问题。此外，每份注释至少经过两次审查。具体来说，注释团队分为四组，每组由三名注释员组成，负责交叉验证注释。如果注释者之间出现分歧，则由一名专家参与协助并做出最终决定，以解决分歧。为了评估注释的质量和可靠性，采用了在计算语言学中广泛用于注释方案的 Kappa 分数。在这项评估中，标注者的任务是为相同的 800 个心理医学对话注释心理疾病标签。不同注释者之间的 Kappa 系数介于 0.831 和 0.865 之间（1.0 表示完全一致），这表明注释者之间的意见高度一致，并肯定了注释结果的可靠性。

2.2.4 数据统计

统计数据如表 2 所示。平均对话为 2.9 轮，每段对话的平均长度为 250.7 个单词，与现有数据集相比，对话内容明显更长。事实证明，这种较长的对话更有利于开发针对心理咨询的对话模型，从而更全面地了解患者的症状，提供共情能力。

表 2 数据统计（-代表无）

Tab.2 Statistics of PsychDialog (- represents none)

类别	总数	医生	患者
对话历史	3268	-	-
对话平均回合数	2.9	-	-
对话平均语句数	5.8	2.8	2.9
对话平均单词数	250.7	117.0	134.0
语句平均单词数	43.6	20.3	23.3
患者自述平均单词数	272.1	-	-
心理疾病种类	11	-	-

3 评估任务

PsychDialog 提供真实心理咨询场景，反映病人向医生表达心理状态的连续过程。通过这一过程，医生可以逐步了解病人的整体健康状况，最终诊断出具体的心理疾病。鉴于对话系统具备多样化的自然语言处理任务，而 PsychDialog 同时包含对话内容和丰富元数据，因此可以为评估各种心理对话任务提供一个研究基准。本文主要侧重于评估 PsychDialog 在两个核心任务中的实用性：心理医疗响应生成和心理疾病检测。这些任务是心理咨询对话系统的基本组成部分。值得注意的是，PsychDialog 还可以作为其他任务的评估，包括对话意图检测、情绪识别、药物推荐和对话状态跟踪等，这些仍是未来探索和研究的方向。

3.1 心理医疗响应生成

心理医疗响应生成任务的目标是为患者提供专业且共情的回复，这在心理医疗对话系统中尤为重要。从本质上讲，心理医疗响应生成涉及使用序列到序列范式对语言表征进行建模。为了将这项任务形式化，考虑到从多轮对话中获得的语境内容，对潜在回复中的单词序列进行概率建模。这一形式化建模过程如式（1）：

$$p(g \mid c)=p(g_1 \mid c)\prod_{i=2}^{n} p(g_i \mid c, g_1, \ldots, g_{i-1}) \tag{1}$$

其中 c 表示多轮对话语境，而 g 表示生成回复中的后续单词。为了评估心理医学响应生成的性能，选择 Transformer[21]、CDial-GPT(Chinese Dialogue Generative Pre-trained Transformer)[22]、BERT-GPT(Bidirectional Encoder Representations from Transformers-Generative Pre-trained Transformer)[23]和 BART（Bidirectional and Auto-Regressive Transformers）[24]预训练语言模型。该评估结果可作为未来研究的基准。

3.2 心理疾病检测

心理疾病检测的目的是通过分析整个对话上下文和元数据信息来判断患者可能存在的心理问题。这一过程为真实场景下的临床诊断提供了宝贵的临床支持。在该过程中选择 BERT（Bidirectional Encoder Representations from Transformers）[25]、RoBERTa（Robustly optimized BERT approach）[26]和 CPT（Chinese Pre-trained Unbalanced Transformer）[27]作为多分类任务的模型基准。

4 情感感知双向自回归模型

本节介绍所提出的情感感知双向自回归模型。考虑到心理咨询过程中，患者往往会隐晦地表达自己的情绪，这就需要模型为他们提供足够情感支持来满足他们的情感需求。要实现这一过程，对话系统就必须了解用户敏感的情绪。这种认识有助于提供充分的情感支持，并能解释潜在的心理健康问题。为了满足这一需求，设计出 EmBART，这是一种情感感知双向自回归模型，专为心理医疗响应生成而设计。EmBART 用于生成共情对话响应，模型结构如图 3 所示。

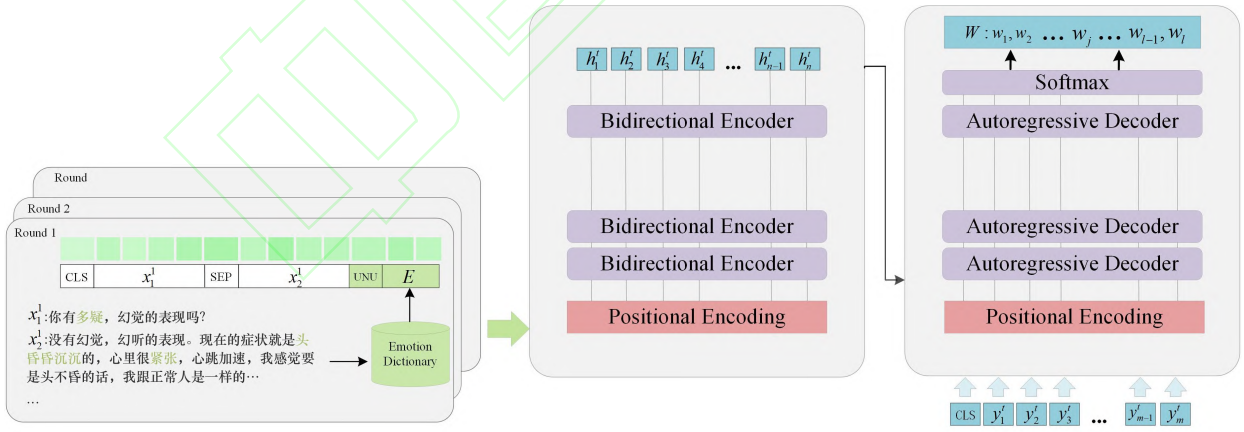

图 3 EmBART 的模型结构

Fig.3 Model architecture of EmBART

EmBART 的骨干结构依赖于 BART[22]，BART 采用了基于 Transformer 的神经网络架构，专为文本生成任务定制。尽管架构简单，但 BART 在生成连贯、准确的文本方面表现出了出色能力。本文选择 BART 作为 EmBART 的基础模型。相比于自回归 Transformer，BART 结合了双向编码器和自回归解码器的优点，既能同时考虑上下文中的前后信息，又能生成连贯且上下文相关的输出。EmBART 将外部情感知识整合到双向自回归的 Transformer 架构中，从而增强模型的情感感知能力。EmBART 的结构是一个双向编码器和一个从左到右的解码器。双向编码器用于辨别正在进行的 n 轮对话中第 i 轮的情感状态。在每一轮对话中，使用中文情感词汇本体表（http://ir.dlut.edu.cn/）来识别当前话语中的情感词汇。为

了将对话语境和情感知识进行编码，将每轮对话的最后一句话作为预训练解码器的输入，从而将当前语句通过特殊的分隔符与相应的情感词连接起来用于训练多个情绪向量表示出词语的潜在情绪，其表示形式如式（2）：

$$[CLS],x_1^1,x_2^1,...,x_j^1[SEP]x_1^i,x_2^i,...,x_j^i[UNU]e_1,e_2,...,e_k \quad (2)$$

在输入序列的开始位置，[CLS]标记句子的开始位置，而[SEP]标记句子的结束位置，从而将上一个语句与下一个语句分隔开来，此外[UNU]标记情感词开始的指示符。x_j^i 表示第 i 个句子中第 j 个单词。该公式描述了将扁平化标记序列输入到编码器的方法。该序列不仅将对话上下文进行编码，还将对话语篇中蕴含的情感信息进行编码。为有效捕捉整个对话序列中不同语句的情绪状态，采用了一种利用正弦和余弦函数的位置编码方法，具体如式（3）与式（4）：

$$P_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \quad (3)$$

$$P_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \quad (4)$$

其中，pos 表示语句的位置，而 d_{model} 指的是表征的维度。值得注意的是，这种位置编码方法包含了对话中每个情感词的位置，而无需为模型添加额外参数。通过将情感词语的位置整合到位置编码中，对话模型能够更有效地识别情感变化和微妙的情感细微差别。将扁平化序列与位置编码相加并输入到编码器，就得到了隐藏状态 h_t^i。从左到右解码器根据 H_n^t 和 Y 预测每个生成词 w_j 的概率，其中词的数量由最大长度参数 l 决定，具体如式（5）与式（6）：

$$Q_j^t = EmBART\_decoder(H_n^t, Y) \quad (5)$$

$$P(w_j^t) = softmax(Q_j^t) \quad (6)$$

最后，采用交叉熵作为解码器损失函数，测量输出分布 $P(w_j^t)$ 与真实标准分布 Y_j 之间的差异。所提出的 EmBART 模型是将情感知识纳入心理医疗对话模型构建的初步尝试，它在 PsychDialog 上建立了基准结果，为未来的模型优化奠定了基础。

5 实验

5.1 心理医疗响应生成

5.1.1 实验设置

为了训练心理医学对话响应生成模型，将数据集按 8:1:1 的比例分为三个子集：训练集、验证集和测试集。基线模型包括 Transformer、CDial-GPT、BERT-GPT 和 BART。对于 EmBART，将双向编码器和自回归解码器都配置为 6 层，并将表示的最大维度设置为 768。本文采用对话生成任务中常用的四种评估指标：困惑度（Perplexity）衡量生成回复的语言质量，BLEU-*n*[28]（其中 *n* 设为 2）评估基于 n-gram 匹配生成的回复与地面真实回复之间的相似度，Entropy-4 和 Dist-*n*[29]（其中 *n* 设为 1 和 2）用于评估生成回复的词汇多样性，METEOR 用于评估生成文本于参考答案之间的相似程度。通过这些指标，可以根据对话生成质量的不同方面对模型性能进行全面评估和比较。

5.1.2 自动评估结果

表 3 列出了模型在数据集中心理医疗响应生成的性能。与其他模型相比，EmBART 在所有指标上都表现出色，优于原始的 BART 模型，其困惑度较基准模型降低了 2.31。这表明加入情感信息对提高生成文本的多样性和信息量具有重大影响。此外 CDial-GPT 在 BLEU 和 METEOR 指标方面也表现出了卓越的性能，凸显了与其他端到端模型相比，通过使用大规模社交媒体对话数据集对该模型进行微调所获得的优势。未来的工作中，我们将进一步探索 EmBART 模型的参数空间，并分析其对各项性能指标的影响。

表 3 心理响应生成指标对比

Tab.3 Comparison of Psychological Response Generation Indicators

模型	Perplexity	Dist-1	Dist-2	BLEU-2	Entropy-4	METEOR
Transformer	26.45	0.181	0.384	0.014	4.152	0.019
CDial-GPT	25.55	0.116	0.501	0.856	6.213	1.145
BERT-GPT	27.09	0.125	0.571	0.017	5.316	0.075
BART	26.12	0.051	0.268	0.033	7.814	0.074
EmBART(ours)	**23.24**	**0.196**	**0.601**	**0.904**	**7.959**	**1.215**

为了进一步比较这些模型，图 4 展示了由不同模型生成的两个回答示例，其中加粗文字代表情感词语。值得注意的是，结果中包含了由 ChatGPT 生成的回复，这些回复是通过两种不同的交互界面获得的：(1)只提供给大语言模型最后一轮对话数据，让 ChatGPT 按照格式生成所需要的内容。(2)提供多轮对话，以便 ChatGPT 能够接收更多的上下文信息，这些信息被用作对比样本。对比两个案例下多个模型的回答来看：一方面，通过采样生成示例的对比发现，CDial-GPT、BERT-GPT 和 BART 所提供的回答较为简短，存在绝对、生硬等问题，不利于与患者进行交流。另一方面，Transformer、chatGPT 和 EmBART 的回复提供了情感支持和诊断建议，与

Transformer 相比，EmBART 和 ChatGPT 表现出更加委婉和易于接受的语气。此外，通过在生成的回复中统计情感词语（图中加粗字体），EmBART 的用词比其他模型更具有共情能力。这表明 EmBART 在利用情感信息来增强对话模型方面具有卓越的能力。对于未来的研究，在 PsychDialog 上训练的模型可以通过整合多轮心理医学对话和特定语境的情感知识来进一步提高生成回复的质量。

Patient:	玩手机很长时间了，以前很正常，就是最近不跟家里人说话，最好的朋友也**不说话**，叫他出去散步，只是**不说话**不到人多的地方去，他这是什么病啊？不肯去医院，特别是**精神病**院我都不敢提。 Play with the phone for a long time, before very normal, is recently not talking to the family, best friends do not talk, called him out for a walk, just do not talk to go to a crowded place, he is what disease ah? Not willing to go to the hospital, especially the psychiatric hospital I do not dare to mention.
Doctor:	如果孩子不是**闹情绪**，就是心理有问题，可以和他说带他去看看心理医生。 If the child is either in a mood or has a psychological problem, you can talk to him about taking him to a psychiatrist.
Transformer:	心理障碍的可能性大一些，建议带孩子去医院看看。 Psychological disorders are more likely, it is recommended to take the child to the hospital.
CDial-GPT:	不要太**担心**了，就是有点心理障碍。 Don't worry too much about it, it's just a bit of a mental block.
BERT-GPT:	应该是心理方面的问题，尽早带孩子去医院治疗。 It should be a psychological problem, take your child to the hospital as soon as possible.
BART:	带孩子去医院看精神心理科。 Take your child to a hospital to see a psychiatric psychiatrist.
chatGPT:	孩子的症状需要**及时**就医，否则可能会越来越**严重**。建议先到普通医院就诊，如果需要再到精神心理科做进一步检查。 The child's symptoms need prompt medical attention or they may get worse. It is recommended to visit a general hospital first and then the psychiatric-psychological department for further examination if needed.
EmBART:	这样下去可能**不行**，**必要**时给孩子看看心理医生，或者电话咨询我。 This may not work, see a psychologist for your child if necessary, or call me for advice.
Patient:	为什么到下午就**头晕**的厉害，感觉自己在消瘦，口干、没味口、有时感觉身上**燥热**、有时又怕冷、有时又手脚冰冷，老**担心**自己有大病。 Why do I feel dizzy in the afternoon, feel like I'm losing weight, have a dry mouth, no taste, sometimes I feel hot, sometimes I'm cold, sometimes my hands and feet are cold, and I'm always worried that I have a serious illness.
Doctor:	你这是**焦虑**的表现。 You are a sign of anxiety.
Patient:	谢谢美女医生。以前也会有**焦虑**，这次感觉**严重**些，这次不舒服担心血有问题，想去查血，怕是白血病，但武汉又没解封，不知是不是我**担心**想太多，引起身体这些**不适**。 Thank you beautiful doctor. Previously there will be anxiety, this time feel more serious, this time uncomfortable worry about blood problems, want to check the blood, afraid of leukemia, but Wuhan and not unblocked, I do not know if I worry about thinking too much, causing these discomforts in the body.
Doctor:	你这是**焦虑**伴躯体**不适**感。 You have anxiety with somatic discomfort.
Transformer:	能缓解**焦虑**不用担心，先调整心态面对治疗。 Can relieve anxiety do not worry, first adjust the mindset to face treatment.
CDial-GPT:	是处于一种**焦虑**状态，太感性了引起的**情绪**不适。 It is an emotional discomfort caused by being in a state of anxiety and being too emotional.
BERT-GPT:	**焦虑**的症状明显一点，不要太过劳累。 The symptoms of anxiety are a little more obvious, do not overwork yourself.
BART:	可能是**焦虑**了，不用太过担心，要及时就诊。 It may be anxiety. Don't worry too much about it, but seek medical attention.
chatGPT:	您好，您的症状可能是由于**焦虑**引起的，所以建议您先放松心态，不要过于**紧张担忧**。如果您担心身体出现了问题，可以在当地的医院进行相关检查，不一定要去武汉。 Hello, your symptoms may be caused by anxiety, so we recommend that you first relax your mind and do not worry too much. If you are worried about a physical problem, you can have a relevant examination at a local hospital, not necessarily in Wuhan.
EmBART:	对于**焦虑**的治疗时间必须得长，一般两个月**左右**，不用**担心**，如果你的**焦虑**症状持续的话，建议在当地**正规**医院精神科诊断。 For the treatment of anxiety must be long, generally about two months, do not worry, if your anxiety symptoms persist, it is recommended that the local regular hospital psychiatric diagnosis.

图 4 两组不同模型对话响应生成示例

Fig.4 Two examples of generated responses by different models based on PsychDialog

5.1.3 人工评估

除自动评估之外，还进行了人工评估，该过程从数据集中随机选取了 50 个对话，并聘请五位评估员从流畅度、情感丰富度和句子准确性等方面对每个模型生成的回复进行评估。评估人员将模型生成的回答与每段对话的标准答案进行比较并打分，最终将评估人员的平均打分作为评估结果。使用的指标包括：(1)流畅性，表示语法的正确性和流畅性；(2)情感丰富度，衡量句子中是否有情感术语；(3)句子准确性，评估在上下文中回答是否恰当。指标范围均为 0-3，分数越高代表对应性能越好。如表 4 所示，人工评估结果表明，EmBART 生成的回复在流畅度、情感丰富度和句子准确性方面都优于其他模型。这些结果与自动评估的结果一致。

表 4 人工评估结果

Tab.4 Human evaluation results on PsychDialog

模型	流畅性	情感丰富度	准确性
Transformer	2.02	1.15	2.11
CDial-GPT	2.21	1.22	2.15
BERT-GPT	2.11	1.04	2.07
BART	2.07	1.13	2.03
EmBART	**2.39**	**2.03**	**2.27**

5.2 心理疾病预测实验

5.2.1 实验设置

本章节利用数据集中包含对话内容和患者自我描述的内容评估了心理疾病检测的有效性。为了确保模型训练过程的平衡并优化心理疾病预测的性能，将每种心理疾病的数据样本按照 5:1 的比例分为训练集和测试集。基线模型选用 BERT、RoBERTa、CPT 预训练模型。除此之外还进行消融研究，以调查纳入患者自我描述信息对心理疾病预测效果的影响，患者自我描述信息包含大量与疾病综合症相关的情绪表达，没有患者自我描述的模型在训练过程中完全依赖于对话上下文，而使用患者自我描述训练的模型则将自我描述内容和对话上下文同时作为模型的输入。在实验过程中，使用 AdamW 来学习权重参数，并将学习率设置为 5e-5。其他超参数根据所使用的预训练模型的默认设置进行配置。为了评估训练好的检测模型，采用了广泛认可的指标，包括准确率（Accuracy）、微精确率（Micro-Precision，Micro-P.）、微召回率（Micro-Recall，Micro-R.）、微 F1 指标（Micro-F1）、加权精确率（Weighted-Precision，Weighted-P.）、加权召回率（Weighted-Recall，Weighted-R.）与加权 F1 指标（Weighted-F1）。

5.2.2 实验结果

表 5 展示了基于 PsychDialog 的三种分类模型的性能。值得注意的是，引入患者的自我描述内容后，每个模型的性能都有显著提高。特别是 CPT 和 RoBERTa 在各种指标上都优于 BERT。这些结果强调了纳入患者自我描述信息（尤其是情绪细节）对于提高模型准确性的价值。然而，心理疾病预测任务错综复杂，这是由几个特定任务因素造成的，包括患者所表达情绪的复杂性、通过对话诊断心理疾病的不确定性以及所用语言的模糊性。这些复杂性给准确检测带来了巨大挑战。未来的研究应侧重于全面解决与心理疾病相关的不确定性和多样性问题，同时也应考虑到各种心理疾病在症状上的细微差别。例如，抑郁症和焦虑症在精神状态和恶化的身体健康方面有相似之处，它们之间的区别往往依赖于对话的细微差别和特定词汇，这些细微差别可能会增加疾病预测的难度。

表 5 心理疾病预测结果，* 表示使用患者自我报告内容后的分类结果

Tab.5 Performance of mental disorder detection，* Indicates categorization results after the use of patient self-reported content

模型	Accracy	Micro-P.	Micro-R.	Micro-F1	Weighted-P.	Weighted-R.	Weighted-F1
BERT	0.414	0.404	0.395	0.394	0.411	0.414	0.409
RoBERTa	0.454	0.443	0.374	0.386	0.451	0.454	0.442
CPT	0.431	0.397	0.402	0.457	0.441	0.457	0.445
BERT*	0.503	0.489	0.489	0.484	0.498	0.503	0.499
RoBERTa*	0.536	0.507	0.472	0.473	**0.520**	**0.526**	**0.517**
CPT*	**0.564**	**0.531**	**0.528**	**0.501**	0.511	0.524	0.515

5.2.3 更多讨论

实验结果有力地证明了 PsychDialog 数据集在心理医疗响应生成和心理疾病预测方面的实用性。在心理医疗响应生成任务中，本文提出的 EmBART 模型与其他模型相比表现出更优越的性能。图 4 中提供的示例说明 EmBART 在生成更流畅、信息更丰富的回复方面的能力，展示了它在提供情感支持方面的共情能力，这种情感支持模拟了心理咨询过程中病人与医生之间的真实互动。在这些互动中强调情感支持，可以帮助机器理解患者症状的根本原因，有助于做出更准确的诊断，并有可能减少精神障碍的发生。鉴于情感支持在引导心理咨询对话向更深层次发展方面具有重要影响，因此有可能设计出优先考虑共情的对话模型，生成既包含指导性内容又包含细微情感支持的心理医疗响应内容。

在心理疾病任务中，CPT 和 RoBERTa 模型的表现优于 BERT。然而，尽管本文注释依赖于核心症状和疾病，但数据集在心理对话内容和语言模糊性方面存在相当大的不确定性。这种固有的复杂性提高了准确检测心理疾病的难度。该数据集的心理疾病种类繁多，使其有别于其他以疾病为导向的对话，强调了模型需要深入了解用户的情绪，以揭示其内心的想法和感受，因此，数据集可作为心理疾病预测领域未来研究的基准数据集，它的多标签属性和心理咨询对话的复杂性为完善识别心理疾病模型提供了理想的基础。

6 结论

本文介绍了心理医疗对话数据集--PsychDialog，一个具有多标签属性的中文数据集，数据来源于三个在线医疗问诊平台。该数据集包含 11 种心理疾病的对话数据，并整合了各种对话元数据，包括医院就诊、科室和患者自述。此外，本文提出了 EmBART，旨在通过整合情感因素来增强共情响应。对 PsychDialog 的评估证明了在心理医疗响应生成和心理疾病预测任务方面的有效性并为开发专业的心理咨询对话系统奠定了基础。未来的工作将偏重于获取大量高质量关于

心理疾病方面的数据，并致力于全面解决与心理疾病相关的不确定性和多样性问题，进一步提高心理医疗响应生成与心理疾病监测性能。

参考文献

[1] WORLD HEALTH ORGANIZATION. Responding to community spread of COVID-19: interim guidance[R]. Geneva, World Health Organization, 2020.

[2] KROENKE K, SPITZER R L. The PHQ-9: a new depression diagnostic and severity measure[J]. Psychiatric annals, 2002, 32(9): 509-515.

[3] HART J, GRATCH J, MARSELLA S. How virtual reality training can win friends and influence people[M]//Fundamental issues in defense training and simulation. Boca Raton, FL: CRC Press, 2017: 235-249.

[4] XU L, ZHOU Q, GONG K, et al. End-to-end knowledge-routed relational dialogue system for automatic diagnosis[C]//Proceedings of the 33rd AAAI conference on artificial intelligence, Palo Alto, CA: AAAI Press, 2019: 7346-7353.

[5] XIA Y, ZHOU J, SHI Z, et al. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis[C]//Proceedings of the 34rd AAAI conference on artificial intelligence, Palo Alto, CA: AAAI Press, 2020: 1062-1069.

[6] WEI Z, LIU Q, PENG B, et al. Task-oriented dialogue system for automatic diagnosis[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA: Association for Computational Linguistics, 2018: 201-207.

[7] ZHANG Y, JIANG Z, ZHANG T, et al. Mie: a medical information extractor towards medical dialogues[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 6460-6469.

[8] DU N, WANG M, TRAN L, et al. Learning to infer entities, properties and their relations from clinical conversations[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: Association for Computational Linguistics, 2019: 4979-4990.

[9] SHI X, HU H, CHE W, et al. Understanding medical conversations with scattered keyword attention and weak supervision from responses[C]//Proceedings of the 34rd AAAI Conference on artificial intelligence, Palo Alto, CA: AAAI Press, 2020: 8838-8845.

[10] ZHANG Y, SUN S, GALLEY M, et al. Dialogpt: large-scale generative pre-training for conversational response generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Stroudsburg, PA: Association for Computational Linguistics, 2020: 270-278.

[11] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[12] NI J, YOUNG T, PANDELEA V, et al. Recent advances in deep learning based dialogue systems: A systematic survey[J]. Artificial intelligence review, 2023, 56(4): 3055-3155.

[13] LIN X, HE X, CHEN Q, et al. Enhancing dialogue symptom diagnosis with global attention and symptom graph[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2019: 5033-5042.

[14] YANG W, ZENG G, TAN B, et al. On the generation of medical dialogues for COVID-19[EB/OL]. [2020-05-11]. https://arxiv.org/pdf/2005.05442

[15] ZENG G, YANG W, JU Z, et al. Meddialog: large-scale medical dialogue datasets[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2020: 9241-9250.

[16] YAO B, SHI C, ZOU L, et al. D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2022: 2438-2459.

[17] LIU S, ZHENG C, DEMASI O, et al. Towards emotional support dialog systems[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 2021: 3469-3483.

[18] GRATCH J, ARTSTEIN R, LUCAS G M, et al. The distress analysis interview corpus of human and computer interviews[C]//Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14). Paris: European Language Resources Association (ELRA), 2014: 3123-3128.

[19] SAHA T, CHOPRA S, SAHA S, et al. A large-scale dataset for motivational dialogue system: An application of natural language generation to mental health[C]//Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE, 2021: 1-8.

[20] 李丹亚,胡铁军,诸文雁,等.中文医学主题词表检索系统[J].中华医学图书馆杂志,2001, 10(4):1-2+9. (LI D Y, HU T J, ZHU W Y, et al. Retrieval system for the Chinese medical subject headings[J]. Chinese Journal of Medical Library, 2001, 4: 1-9.)

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[22] WANG Y, KE P, ZHENG Y, et al. A large-scale chinese short-text conversation dataset[C]//Proceedings of the 9th CCF Conference on Natural Language Processing and Chinese Computing. Cham: Springer, 2020: 91-103.

[23] WU Q, LI L, ZHOU H, et al. Importance-aware learning for neural headline editing[C]//Proceedings of the 34rd AAAI Conference on Artificial Intelligence, Palo Alto, CA: AAAI Press, 2020: 9282-9289.

[24] LEWIS M, LIU Y, GOYAL N, et al. Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 7871-7880.

[25] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg, PA: Association for Computational Linguistics, 2019: 4171-4186.

[26] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[EB/OL]. [2019-07-26]. https://arxiv.org/pdf/1907.11692

[27] SHAO Y, GENG Z, LIU Y, et al. CPT: A pre-trained unbalanced transformer for both Chinese language understanding and generation[J]. Science China Information Sciences, 2024, 67(5): 152102.

[28] 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. Stroudsburg, PA: Association for Computational Linguistics, 2002: 311-318.

[29] LI J, GALLEY M, BROCKETT C, et al. A diversity-promoting objective function for neural conversation models[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2016: 110-119.

This work is supported by Liaoning Provincial Social Science Planning Fund (L21CXW003)

XU Bo, born in 1988, Ph. D., associate professor. His research interests include mental health computing and natural language processing.

HAO Dezhi, born in 1998, M. S.. His research interests include mental health counseling dialogue system.

YU Erchen, born in 2001, M. S. candidate. His research interests include multimodal affective computing.

LIN Hongfei, born in 1962, Ph.D, professor. His research interests include natural language processing.

ZONG Linlin, born in 1987, Ph. D., associate professor. Her research interests include multimodal affective computing.