

第 36 届人工智能会议(AA AI -22 )

**HEAL:**一 个 痛苦管理对话的知识图谱

**Anura** **dh** **a** **Weli** **vi** **ta,** **Pearl** **Pu**

计算机与通信科学学院

洛桑综合技术学院(Ecole’Polytechni que F’ed

de Lausanne)

瑞士

{kalpani 。welivita, pearl.pu}

@epfl.ch

摘要

现代社会 的 需 求越 来越 多地给人们造成了心 理 负 担，给 我们 的心 理 健康带 来 不 利影响。因 此 ，具有共 情 响应和 悲痛 管 理能 力 的神经会话代 理最近越 来越 受欢迎 。然 而，现 有 的 端到 端 共 情 对 话 智 能 体 经 常 产 生 通 用 的 、重 复 的 共 情 语 句 ， 如 “ 听到 这个 我 很遗憾 ”，这 未能传达特 定情 况 的特 异性。 由 于这类模 型缺乏 可控 性 ，它们还 强加 了产 生 有 毒 反应 的 风 险。利 用 知 识图谱 上 的推 理的 聊 天机器人被视 为 端到 端 模 型的 高 效 和 故障 安全解 决 方案 。然 而 ，这种 资源在情 感 困扰的 背 景 下 是有 限 的 。为 了解 决 这个 问题 ，我们介绍 了 治愈(H EA L )，这 是一个知识 图谱，基于 R ed dit 整理的 100 万个痛苦 故事及其相应 的安 慰反应开 发。它由 22K 个节 点 组 成，识 别 不 同类 型 的 压 力源、说话 人 期望 、反应 和与痛 苦 对 话 相 关 的 反 馈 类 型 ，并 在 不 同 类 型 的 节 点 之 间 形 成 10 4 K 连接。每个节点 都与 41 种情感状态中 的一种相关联。 对 H E A L 进行 的 统计 和视 觉分 析揭 示 了在 以痛 苦为 导向 的 对话中 ，说话 者和倾 听 者 之 间的 情 绪 动态，并确 定 了导致 情 绪缓解 的 有 用反应模 式。 自动 和 人类评估 实验表 明，与 基线相比，H E A L 的 反应更加多样 化、同理心 和可靠。

介 绍

现代世 界 的 需求越来越多地造成心理 负担，并给我们 的 心 理健康 带来不利影 响。苦恼是指个体在应对特 定 的个人压 力或需求 时所经历 的一种不舒服的情绪状态 ，这种情绪状 态 会 导 致 对 个 人 的 伤 害 ， 无 论 是 暂 时 的 还 是 永 久 性 的 (Ri dne r 2 004 ) 。这些压力源包括与亲人分离、人际冲突、 某些 心理健康状况(如抑郁症)、工作表现不佳，以及失眠 等睡 眠 问题。阿 尔梅达等人(2002 年)的一项研 究通过每 日 电话采访 ，对美 国全 国 1031 名成年人的样本进行了多方 面 的日 常压力源测量 ，结 果 显示他们至少经历过一种 日 常 压力源

版权所有©20 2 2 ，促进人工 智能协会(www .aaai.o rg) 。版权所有。

图 1: H E A L 部分示 意 图。红色 、 紫色 、蓝色 、绿色和黄 色节点分别代表压力源 、说话人期望 、反应和反馈类型 以及相关的情感状态。

在 40 %的 学 习 日中。人们通 常倾 向于在 日常对话中分享这 样 的经历。因此 ，嵌入 具有适当移情 响应 能力 的开放域对 话代理或聊 天机器人来解 决这种痛苦 的情况已经获得 了很 大的兴趣 (R ashkin et al . 2019; Lin et al . 2019; Maju md er 等 2020;谢、普 20 21)。

随着 复杂神经网络架构 的发展，如 transfor m er ( V as wani et al . 2 017)和预训练语言模型 ，如 B ER T ( D evl in et al . 20 17)。

(2019) 、RoBERT a(刘 等。 2019 a)和 GPT -3 (Brown 等。

2020)，在非 结构化文本上微调神经响应生成模 型已经成为 构建聊 天机器人 的常用方法之一。尽 管 它用严格 的基于规 则 的方法避 免 了大多数 限制 ，并使聊 天机器人在很大程度 上泛化 到未见过 的领域 ，但可控性 的缺 乏和黑箱性质使这 些模型不太可靠和故障安全(d 'A vil a Gar ce z 和 La mb 2 020)。 当用户 处于对错误信 息和不恰 当评论敏感 的痛苦境地 时， 这尤其成 问题。最近 的一个例子是微软 的 Tay 机器人 ，在 从 Twitt er 上 了解到种族主义和 冒犯性的信 息后 ，开始产生 非故 意 的 、无 礼 的 和种 族主 义 的 推文 ，否 认大 屠 杀 (L ee 2016)。

因此，人们对利用 知识的 兴趣越来越大(Zhu et al . 2017; 刘 等 ， 2018; H an et al . 20 15) 和 常 识 推 理 (Zho u et al . 2018;Y oung et al. 2018)

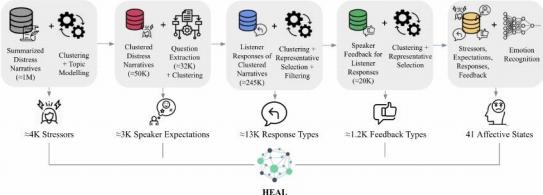


图 2:逐步开发知识图谱 HEAL 的过程。

在基于 图 的表示上生成适 当 的、有信 息量的对话 响应。与 在非结构化文本上进行训练相比 ，使用基于 图 的表示为生 成 的回 复提供 了更 多 的可控性和可解释性，从而 限制了不 适 当和不可靠 的内容。识别 知识 图谱 中 的相关主题 ，使沿 着可预测路线 引导对话流成为可 能 ，同 时也提供 了 战略性 多样化回应的能力(Liu et al. 2019b)。

虽然存在诸如 C on cept - Net (S pe er 、C hin 和 H avasi 201 7) 和 A T O MI C (S ap et al . 20 19)等大规模知识图谱 ，但 它们主 要通过捕获事 实 知识和嵌入 具有 简单 常识推理 能力 的聊天 机器人模 型来辅助开放域对话生成。 由于它们不是为 了捕 捉共情 交流 的规 范而开发 的 ，因此该领域缺 乏 帮助遇 险管 理和共情 响应生成的语 言 资源和模 型。而且从来没有人尝 试生成 知识 图谱 ，用上 下文-反应对之 间 的关 系来表示整个 对话。为 了解决这些局 限性 ，我们 引入 了 H E A L(意 思是治 愈、 同理心和影 响 学习)，这是一个用于痛苦 管理对话 的知 识 图 谱 ， 通 过 分 析 压 力 事 件 的 叙 述 和 从 精 心 选 择 的 子 reddit 中策划 的相应回应线索开发。

H E A L 包含 5 种类型 的节点 : 1)压力源 :造成痛苦 的原 因;**2)** 期望 **:**在痛苦叙述中，说话者经常问的问题 **;3)**反应类型 **:**听众为 应对不同的压力源而给出的最常见的反应类型 **;4)**反馈类型 **:**说 话者在回应后提供的常见反馈类型 **;**和 5)情感状态 : 与每个节 点相关联 的情感状态。这里 的演讲者是那些经历痛苦情况 的人(他们通过在 R e ddit 上发帖来开始对话)，听众是这些 帖子 的评论者。 图 1 显示了 H E A L 中 典 型压力源的说 明。 治愈( HE A L)是一种与痛苦相关的话题 ，它能够准确地描述 面 向痛苦 的对话 中的潜在情境，从而使对话模 型 能够检索 到更具体 的情境 响应。此 外 ，诸如此类 的响应是 否会 导致 积极或消极的反馈， 以及它们是否会解 决 im -等信息

对 处于 困境 中 的人的 隐 式期望可 以 导致选择更适 当和有用 的回应。如 图 2 所示 ，我们遵循 了一系 列步骤 ，包括 总结、 聚类、主题建模和情绪分类 ，以从 R eddit 策划的 100 多万 次痛苦对话 中开发 HE A L。这 导致识别 出≈ 4K 的压力源， ≈ 3K 的说话人期望， ≈ 13 K 的反应类型， ≈ 1 .2 K 的反馈 类型 ，以及相关 的情感状态。最终 的图 由 22,03 7 个节 点和 104,00 4 个不 同类 型节 点之间 的连接组成。

通过对 HE A L 进行 统计和视 觉分析 ，我们 能够发现说话 者和听者之 间 的情绪动态 ，以及 导致情绪降级 的有利反应 类型。我们还测试 了 HE A L 在 下游任务 中的 效用，即对给 定 的痛苦情况产生同理 心反应。我们使用知识 图谱开发了 一个基于检 索 的模型 ，并使用自动和人工评估将其性 能与 两个最 先进的共情对话代理进行 了 比较 : 一个 由 Xi e 和 Pu 开发(202 1); 以及 B l end er (R ol l er et al . 202 1)。结 果表 明 ，以 排序方 式使用 知识图谱检 索 出的 回答在 多样性和共情适宜 性方面优于其他人产生 的 回答。通过 实例分析 ，我们还表 明 ，与神经反应生成模 型相 比 ，H E A L 方法检 索 到 的反应 更可靠。我们 的主要贡献包括: 1)开发 了大规模 的 知识图谱、 H E A L、识别不同类型的压力源、说话人的期望 、反应和 反馈类 型 以及与痛苦对话相关 的情感状态; 2)利用 统计和可 视分析 ，识别说话者和 听者之间 的情绪动态和有利 的反应 模式 ，从而 导致情绪去升级 ; 3)评估 H E A L 在获取更共情适 当、多样化和可靠的话语以应对情绪困扰方面 的效用 。1

相 关 工 作

知识图谱因其实用性而引起了自然语言处理社区的关注

1代码和数据可在 github.com/anuradha1992/HEAL 上获得。

11460

在理解自然语言输入方面。这得益于最近 DBP e di a (Au er et al . 2007)和谷歌知识图谱等链接开放数据的出现。 2 YA G O (F abi an et al . 2007) 、 F reebas e (B ol l acker et al . 2008) 和 Wiki dat a ( Vran de ˇ ci´ c 和 Kr ¨ ot zsch 2014)是其他一些基于 从网络中提取的一般 知识构建 知识图谱的例子。最近的 知识 图谱，如 conc ept - ceptN et (Speer 、Chin 和 Hav asi 2017)、 AT O MIC (S ap 等 2019 年)和 AS ER ( Zhang 等 2020 年)，专注 于表示不 同类型的 常识知识。Liu 等人(2018)和 Zhang 等人 (2020)的工作利用这些图中呈现的 fact oid 和 常识知识开发了 开放域会话代理，这些会话代理产生 了更语义和信息更丰富 的响应。

尽 管上述 资源在开发 知识感 知 的对话代理和 具有推理 能 力 的代理 时很有用( Zho u et al . 20 18 )，但通常这些图描述 了开放域 实体和关 系 ，以及建立在 它们之上 的 常识推理。 它们没有捕捉到情感推理和共情响应生成的规范。 H E A L 通过建立压力源 、说话人期望、反应 、反馈和情感状态之 间 的关 系 ，并连接提示-反应-反馈元组来识别可能导致有 利反馈 的 响应，并解 决那些 处于痛苦 中 的人 的 隐性期望， 从而扩展了上述 限制。

方 法

数 据集管理

公 开可 用 的 情 感对 话 数 据 集 ，如 Em - path et i c Di al ogu es (R as hkin et al . 20 19 ) 、E mot i on Lines ( Hsu et al . 201 8) 和 E moC ont ext ( Ch att erj ee et al . 201 9)，大 多 由在人工环境 中 创建或从 电影/ 电视字幕中精选 的开放域和日 常对话组成。 用于进行近期研究 的真实咨询 对话数据集 ( Althoff, C l ar k, and L esk ove c 201 6;Zh ang 和 D an escu - Ni cul es cu - Mi zi l 202 0) 由于伦理 原 因无法直接获得。因此 ，我们从 R ed dit 策划 了 一个新的数据集 ，其中包含讨论现实世界痛苦情况的对 话。我们选择 R ed dit 是因为它是 公开可访问 的， 同行们积 极参与这些平台，以支持其他遭受精神 困扰的人。

我们使用 Pushshi ft AP I ( B a u mg artn er et al . 2 020 ) 从精心 挑 选 的 8 个 版 块 收 集 和 处 理 对 话 线

索 :m ent alh e althsupp ort ; *off* *my* *c* *hest;* *悲* *伤*

*的;sui* *ci* *d* *ew* *atch* *;an* *xi* *etyh* *elp;抑郁症;沮丧的;* 以及抑郁症帮 助 ，这 在 R e ddit 用 户 中 很 受欢 迎 ，可 以发 泄 他们 的 痛 苦。我们通过匹配作者的名字 ，明确地从这些线 索中提取 出对话轮取结构 ，并对这些对话进行严格的数据清洗管 道 ，包括从 听众 的 响应中去 除脏话。通过这种方 式，我们 能够策划 1,27 5,4 86 个二元对话 ，包含 3,3 96, 476 个对话轮 取结构(平均每个对话 2.6 6 轮)。数据预 处理管道和数据集 的描述性统计包括在 附录 中。我们使用 了 80 % 的对话

派生知识图谱，并保留 10% 的对话 ，用于验证和测试 下游 任务。

摘要

R e ddit 策划 的痛苦叙事通 常很长(平均每轮 84.8 9 个令牌)， 一些超过 了某些预 先训练 的基于语 言模 型的架构 的输入令 牌长度 ，如句子- b ert (R eim ers and G ur evy c h 201 9) 。因此， 我们研 究 了各种摘要算法 ，这些算法可用于生成保 留叙事 本质的摘要。

我们研 究 了提取和抽 象 的摘要技术来解 决这个 问题(Tas 和 Kiyani 2007 )。其中 ，抽 象摘要方法主要在新闻文章等结 构化文档上进行训练和测试 ，众所周 知 ，在非 结构化文本 上表现不佳(P eng et al . 2 021)。因此，我们选择 了 5 种不同 的摘要抽取方法: sm mry 的 自 定义 实现 —— R e ddit TL D R 机 器人(https:// sm mry . co m )背后的算法 ; 以及四种不同的预训 练模 型- bart ( L e wi s et al . 202 0) 、GP T-2 ( R ad ford et al . 20 19)、 X L NE T ( Y ang et al . 201 9)和 T5 ( R a ffel et al . 20 20)用于建模 内容重要性。我们对上述方法在 100 个 R e ddit 痛苦叙述样 本中生成的摘要进行了手动评分 ，分为好、好和坏(结 果详 细见 附录)。被评为好的摘要 比例最高 的是 由 S M MR Y 算法 生成 的 。 因此 ，它 被选 择 来 总 结 冗 长 的对 话 回 合 ( 回合 ≥100 个 t oke n)。大 约 43 % 的对话回合是用这个来 总结 的。

烧 结的集群

由于手动注释 昂贵和耗 时 ，特别是在应用于大规模数据集 时，我们 决 定使用自动 聚类来识别 R e ddit 痛苦对话 中 明确 区分的压力源 、期望 、响应和反馈类 型。为此 ，我们使用 了 针 对 大 型 数 据 集 调 优 的 “ 凝 聚 聚 类 ”(M urt ag h a nd Le ge ndr e 2 014 )。它递 归地合 并最 小 限度增加给 定链接距离 的簇对。我们利用句子-B E R T(R eim ers an d G ur evy ch 20 19) 生成的成对嵌入之间 的余 弦相似性计算 了链接距离 ，因为 由此产生 的嵌入 已被证 明 具有高质量 ，并在文档级嵌入中 工作得相 当好 。在附录 中详细解释 了使用凝 聚 聚类而不是 其他聚类方法的选择。

压 力 源 的 识 别

我们 实验 了 8 个相似阈值 ，从 0.6 到 0.9 5，增量为 0.05 ， 以 聚类痛苦叙述。虽然为每个阈值计算 了各种 聚类质量指 标，如轮廓系数(R o usse eu w 19 87) 、D unn 指数( Mi sur a ca 、 Spa no 和 B albi 2 019 )和平均 点到质 心余 弦距离， 以选择最 佳 的相似性 阈值 ，但对每个 阈值处 10 个聚类的子集进行 人工检查和 聚类可视化 显示 ，这些指标对这个数据集并不 最有效

2 en.wikipedia.org/wiki/Goo gle 知识图谱

11461

|  |  |
| --- | --- |
| 压力源 | 提取关键字 |
| 自杀意念 | *犯，* *杀，* *死，* *无痛，* *选择* |
| 焦虑袭击 | *焦虑，* *焦虑，* *攻击，* *社交，* *攻击* |
| 体重增加 | *吃，* *增重，* *吃，* *减，* *胖* |
| 孤独 | *孤独，* *环绕，* *连接，* *孤立，* *社交* |
| 失败的大学 | *学习，* *大学，* *班级，* *学期，* *不及格* |
| Alchoholic | *喝酒，* *喝酒，* *喝酒，* *醉了，* *清醒了* |
| 美国大选 | *特朗普，* *总统，* *唐纳德，* *选举，* *战争* |
| Covid19 | *Covid,* *19* *岁，* *pandem* *i* *c,* *shamboli* *c，带* *来* |

表 1:在使用 TF -IDF 的痛苦叙事集群中确定 的一些应激因 素

(众所周知，上述指标仅对具有凸形聚类的数据集最有效)。 人工检查 的结 果表 明，在 0.95 和 0. 9 等较高 阈值 下识别的 压力源过于具体 ，而在 0.8 以 下识别的压力源过于模糊(在 每个阈值上通过人工检查发现的 聚类质量指标和主题包含 在附录中)。这 导致选择 了一个最佳阈值 0.85。在这个 阈值 下， 4 .93 %的痛苦叙述(总共 47,109 个叙述)被分成 4,363 个 簇。将基于 TF -I DF 的主题建模应用于这些聚类，发现 了一 些清晰 区分 的压力源 ，进一步验证 了 聚类 的有 效性。表 1 显示了在这个过程中识别出 的一些应激源。

期 望 、 反 应和 反 馈类 型

在对悲痛叙事进行聚类 并确 定它们各 自 的主题后 ，我们使 用简单的字符串搜索 包含“ ?”的句子 来提取 聚类后 的 悲 痛叙事 中提 出 的 问题。也提取了相应 的 回答和相关反馈。 我们使用 NL T K 库将 回 复和反馈 中 的单个句子分开 ，以便 通过聚类容 易识别独特 的 回 复和反馈类 型。这样 ，我们总 共能够收集 32 832 个期望，245 707 个 回 复和 20 213 个反 馈 。遵循上述最佳阈值选择的类似过程 ，我们分别选择 0.7 、0.7 5 和 0.7 作为 聚类期望、响应和反馈 的最佳 阈值。 这分别产生 了 3 050 、13 416 和 1 2 08 种期望 、响应和反馈 类 型，每个簇至少有两个不 同的簇元素。响应集群尤其受 到 一 个 自 动 和 人 工 验 证 的 过 程 的 影 响 ， 以 删 除 特 定于 R ed dit 的响应(例如 ，请联 系 subre ddit 的主持人)、机器人 生成的响应(例如，这个动作是自动执行的)和不完整的响 应(例如，哇)。与最终聚类 结 果相关 的 统计数据如表 2 所 示。我们 随机选择每个簇 中 的一个成 员作为簇代表。频繁 期望、响应和反馈类 型 的例子包含在附录 中。

情 感 状 态 模型

为 了将每个压力源、期望 、响应和反馈 聚类 与情感状态关 联起来，我们使用 了 W el ivit a 和 Pu 提出的基于 B E R T 转 换器的分类器

(2020 )在 Em pathet i c Di al ogu es 数据集上进行训练。其具有 显著 的 65 .88 %的分类准确率，与最 先进 的对话情感分类器 相当。该分类器能够将文本分类为 41 种情感类别 中的一种， 其中 32 种是从多个注释方案中选择 的积极和 消极情绪 ，包 括从生物反应衍生的基本情绪 (Ek man 1 992; P lut chik 19 84) 到从 情 境 中 衍 生 出 的 更大 的 微妙 情绪 集 (S kerr y 和 S a xe 2015)，其中 9 种是用于阐述中性情绪的共情反应策略。我 们使用这个分类器对属于一个簇 的每个文本进行分类 ，并 将该簇 与 出现 次数最 多 的情感状态相关联。如 果两个或两 个 以上 的情感状态出现 的 次数相等 ，我们将每个状态 的分 类器置信度加起来，并选择置信度最 高 的那个 。遵循这个 过程，我们 能够识别 出 与每个聚类相关 的最突 出 的情感状 态。

**HEA** **L** **:**统 计 分 析

我们跟踪压力叙述的压力源标识符 ，从 中提取 出每个期望 和反应 ，并 能够在压力源和期望和反应簇之间形成联 系。 我们还跟踪每个反馈 的对话标识符 ，这有助于在反馈集群 和期望和响应集群之间建立联系 。最终形成的知识图谱 H E A L 由 22,03 7 个节点和 104,0 04 条节 点间连接构成。压 力源 与期 望之 间有 9,801 个 连接 ，压力 源 与 反应之 间有 56,65 4 个连接 ，反应与反馈之间有 10,92 1 个连接 ，期望与 反应之间有 26,6 28 个连接。此 外，每个节点 与一种情感状 态相关联，形成 22,0 37 个连接。

图 3 显示了 与压力源、期望 、反应和反馈类型相关的情 感状态 的分布。根据 统计，73 .6 0 %的压力源与消极的情感 状 态 相 关 。 其 中 ， 孤 独 、 悲 伤 、 羞 愧 和 忧 虑 的 情 绪 与 44 .01 % 的压力源有关。大多数期望与忧虑(2 5 .7 0 % )、悲伤 (10 .0 7 %) 和 愤 怒 (7 .51 % ) 等 消 极情 感 状态 有关 ，也 与 希望 (15.41 % )等积极情感状态有关。

在 回答 中 ，60 .38 % 与 中 性情 感状态 有关 。其 中 ，质 疑 (12 .89 % )、同意 (9 .2 2 %)、建议 (6 . 90 % )占 比最 高。一个重要 的观察是 ，在反馈簇 中 ，可 以看 到 与压力源相 比 ，积极情 感状态增加了 7 . 17 %，中性情感状态增加了 270 .29 % 。与 反馈聚类相关 的 消极情感状态相 比 ，与压力源相关 的 消极 情感状态 下降 了 44 .77 % 。在 响应簇中 ，28 .59 %与至少一个 反馈簇相关，其中 100 %的 响应 与至少一个积极或 中性反馈 相关。在 以上 的应答簇 中 ，26 .51 %的应答簇与至少一个正 反馈簇相连 ，77 .48 %的应答簇与至少一个中性反馈相连， 这验证了存在的存在性

11462

在 HEAL 中有用的反应类型， 可以降低遭受痛苦的人的 负面情绪状态。

视 觉化和解读

我们使用 vi s .j s (vi sj s .org )，一个 图形可视化库来可视化生 成 的知识 图谱。这个库生成 的知识 图谱 的部分可视化如 图 4 所示。节点的大小对应于各自集群的大小，边的宽度对 应于不 同集群之间 的连接数量。每一种不同 的压力源、期 望、反应和反馈类 型 也 与一种情感状态相关 ，在这里没有 可视化，以避免混乱。

如关键词所示 ，中 间 的应激源节 点是包含 自 杀想法 的叙 事的代表。 图 中所示 ，有 自 杀想法 的人最 常见 的期望是 :他 应该做什么; *听者是否也有同样的感受;*他有哪些选择。在 这种情况下 ，听众最 常见 的 回答是 :同情 的回答 ，比如我很 抱歉你有这样 的感 觉 ;安慰性的 回答，如“我希望你感 觉好 点 ”; 有意义的问题，如“ 你想谈谈吗?”，你是否寻求过 帮助?、是什 么让你有这种感 觉?表示同意的回答，如“我 有同样 的感 觉 ，我 知道这种感觉 ; 一些建议， 比如拨打 自杀 热线和

*寻求转诊;*以及鼓励的回答，如“ 坚持住，朋友，坚强点 !” 通过紫色 的虚线 ，我们可 以看到一般说话者的期望和 听者 的反应之 间 的联 系。举个例子，我有 同样 的感 觉 ，有人有 这种感 觉 吗 ?“ 坚持 住 ，我 的 朋友 ，你 在看 医生 吗 ?”和 “我该怎 么做?”可 以看 出 ，这些回答大多数都 与演讲者的 积极反馈有关 ，比如感谢他 的回答 ，这表 明了对 听众 的感 激，同时也证明了这是一个好的 回应。

评 估 **HEAL** 在 反 应 中 的 效用

对 痛苦提示的反应

我们评估 了 H E A L 在检 索特 定痛苦对话提示 的适当共情反 应 的能力 ，并 与现有 的最 先进的共情反应生成模 型进行了 比较。为此 ，我们使用 了一开始分离的 10 %的 R e ddit 对话 进行测试。为了从“ H E A L” 中检索反应，我们计算 了新 的 nar-叙事/ 提示 与知识 图谱 中不 同 聚类 的现有叙事之 间的 余 弦相似度 ，并将新 的叙事 与现有叙事 中相似度最大 的聚 类关联起来 。在测试数据集 中的 12365 1 个对话提示中， 60.7% 与 the 显示了 0 .75 或以上 的相似度

表 2:与最终 聚类 结果相关 的统计数据和 聚类质量指标(一个 聚类被认为至少有两个不同 的元素)。平均余 弦距离表示平均点 到 质心的余弦距离。轮廓 系数和邓 恩指数的值分别位于 [- 1,1]和 [0， ∞ )之间。这些值越正越好。

(一)压力

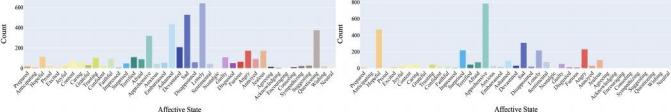
(b)期望

(c)答复

(d)反馈

图 3:HEAL 过程中与压力源、期望、反应和反馈相关的情感状态的分布

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 类型 | 阈值 | #集群 | 最大  集群规模 | 合计。 #应 该怎么办  集群 | %的 doc.s  集群 | 轮廓  系数 | Dunn-Index  (cos) | 平均余弦  距离。 |
| 压力 | 0.85 | 4363  年 | 11856  年 | 47109  年 | 4 .93% | 0.0554 | 0.0677 | 0.0443 |
| 期望 | 0。  7 | 3050  年 | 489 | 16316  年 | 49 .7% | 0.3781 | 0.1008 | 0.0649 |
| 答复 | 0.75 | 13416  年 | 1025  年 | 78194  年 | 31 .82% | 0.3263 | 0.1061 | 0.0722 |
| 反馈 | 0.7 | 1208  年 | 960 | 5782  年 | 28 .61% | 0.2882 | 0.1705 | 0.0895 |



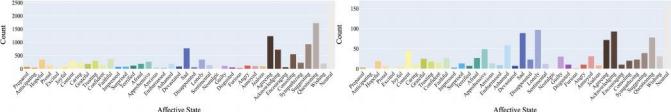


图 4:vis .js 对部分 HEAL 的可视化。压力源、期望、反应和反馈类型分别用红色、紫色、蓝色和绿色表示。只有具有显著 边权值的连接才被可视化，以避免混乱。

数据集 模型 D1 D2 D3 D4 BLEU1 BLEU2 流星 胭脂 GM

Reddit (Xie 和 Pu2021)

0.1159 0.0686

**0.1704**

0.3364 0.2226

**0.4540**

0.4818 0.3206

**0.6003**

0.5815 0.3877

**0.7100**

0.0066

**0.0707**

0.0033

0.0014

**0.0150**

0.0007

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | |
| 搅拌机  Heal-ranked |  |  |  |  |  |  | 0.0277  **0.0469**  0.0252 | 0.0475  **0.0661**  0.0332 | **0.6921**  0.6047  0.6599 |

表 3:Red dit 中响应遇 险提示任务获得的自动评价结 果 D1 、D2、 贪婪匹配分数(Rus 和 Lint ea n 2012)。

知识图谱 中覆盖 的压力源和 它们被过滤 以进行评估。然后， 我们对 与新叙事相关 的压力源相关联 的反应进行排名 ，首 先是压力源和反应之 间的边权重，然后是反应集群大 小 ， 并选择排名靠前的反应。我们称之为 he alran k。在提出的 基线中 ，与说话人期望和反馈类型 的联 系没有被考虑在 内。 但我们详细解释 了这些节 点如何有助于 改善这一基线 ，作 为未来工作的一部分。

我们将 he alrank 检 索 到 的应答 与两种最先进的共情应答 生成模型进行比较，一种由 Xi e 和 Pu (20 21)开发 ，另一种 由 B l end er (ge n - er at ive)开发(R ol l er et al . 202 1)。前者是基 于 R oB E R T a 的 多 轮 情 感 参 与 对 话 生 成 模 型 (Liu et al . 2019 a)。它对 op en 字幕中的约 1M 对话进行预训练(Li son 等人 2019 年 )， 并对 同 理 心对话 进行微调 (R ashkin 等 人 2019 年)。后者是一个标准 的基于 S eq2S eq tra nsfor m er 的 共 情 开 放域 聊 天 机 器 人 。 它 在包 含 约 1. 5 亿 个 评 论 的 R ed dit 讨论上进行了预训练 ，并在几个规模较 小但重 点集 中的数据集上进行了微调。

自 动评价

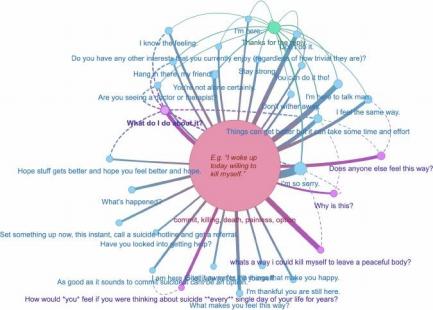
表 3 包括根据上述 Reddit 对话模型产生的响应计算的自动 指标

D3 和 D4 代表 Di stin ct -ngra m 度量(L i et al. 2016) ，GM 代表

提 示 。 我 们 可 以 观 察 到 ， 在 用 于 衡 量 应 答 多 样 性 的 Di stinct - N 指标方面，治疗等级优于其他方法(Li 等， 2016)。 这表明 ，与现有 的神经反应生成模 型相 比 ，治疗在产生更 多样化的反应方面的 效用。我们在表 4 中通过展示三个模 型对几个 与痛苦相关 的提示产生 的一些示例响应来进一步 证明这一 点。可以看到 ，B l ender 和谢和普的模型都对两个 完全不 同 的提示产生 了重 复 的通用反应 ，而从 治疗 中检索 到的反应更 多样化 ，并针对特 定 的情况(更多 的例子包含在 附录 中)。我们还观察 到治疗等级相对于其他 自动指标蓝、 流星和胭脂表现较差。然而 ，已 知这些指标与人类判 断的 相关性较差(Liu et al . 20 16)，当 与人类评估实验的 结果进行 比较时，可以很好地看 到这一点，下 一节将讨论。

人 的评价

我们 设计 了一 个 人工 评价 实 验 ，从 A ma z on M e ch ani c al Tu rk ( A M T ) 招募众包 工人来 评估三个 模 型产生 的反应 的 共情 适 宜 性。我 们 从 R e ddit 测 试数 据 集 中 随 机 选择 了 200 个对话 ， 由众工作者进行评估。工人们被指示拖拽和



|  |  |
| --- | --- |
| 提示  (谢和普 2021)  搅拌机  治疗等级 | *我似乎不能感到快乐，但我不是de-*  *按下*  *听到这个消息我很难*  *过。*†  *听到这个消息* *我很难过。* *你有吗?*  *有什么爱好能让你感觉更好?*$  *最重要的是，* *如果你感到失落也没关系* *按下。* |

表 6:在应对危难提示时显示毒性的反应示例(上标:†)。

通过开发一个简单的检索模型来应对遇险提示 。根据对 H E A L 检 索 到 的 响应进 行 的 自动和 人工 评估评 分 ， 以及 Xi e 、Pu 和 B l end er 的最 先进 的共情对话模型产生的评分， 显然 H E A L 能够产生更 多样化和共情更适当 的 响应。人工 检查这些模 型产生的反应进一步揭示 了使用纯粹生成模型 来解 决痛苦 的危 险， 以及 H E AL 如何通过战 略性地识别与 给定提示相关的特定压力源来避免不适 当 的反应。

当评估对求救提示的反应时 ，我们建立的检索模型使 用了简单的重新排序 。通过将边缘权重与说话人期望和 反馈相 结 合 ，可 以开 发 出更 复杂 的 重新排 序方法 。来 自 知识图谱的信息可以用于增强神经响应生成模型 ，也可 以为这些模型引入更多的可控性和可解释性 ，从而增加 可靠性。

有一些 与 H E A L 相关 的局限性。它仅 限于识别 ≈ 4K 的 压力源 。但是 ，新的提示可 能涉及无数其他的压力源 ，这 些都没有在 知识 图谱 中涵盖。然而 ，从 网络上抓取更 多的 数据来扩充 知识 图谱是有 空 间的 ，这将使 它能够 处理更广 泛的压力源和期望。

道 德的声明

|  |  |
| --- | --- |
| 提示  (谢和普 2021) 搅拌机  治疗等级 | *你能帮我写一个通用的吗*  *遗书?*  *我不太确定你在说什么。*†  *听到这个消息我很难过。* *我希望你能找到*  *有人帮你写。*†  *很抱歉让你这么难受。* |

表 5:三种模型产生 的反应共情恰 当性的人工评价 结果。

将模 型产生 的反应分成好的 、好 的和坏 的 区域 ，这取 决于 这些反应对给 定提示 的共情恰 当程度 。三名工人被 允许完 成一项任务 ，结 果是根据多数投票来计算 的。包括用户界 面在 内的人类 实验 的细节包括在 附录 中。实验 结 果如表 5 所示 。在 85 %的 时 间里 ，工人之间至少获得 了 2/ 3 的一致， 并产生 了 0.2 3 的评价者之间 的一致(弗莱斯 kapp a)分数， 表明公平一致。

根据 结 果，可 以看 到 治疗-排名产生 了最多 的 良好反应， 在移情适宜性方面明 显优于谢和普 的模 型和搅拌机。

在表 6 中 ，我们 突 出了 几个精 心挑选的例子，其中神经反 应生成方法在 响应与痛苦相关的提示 时假 定存在危 险。可 以看到 ，治疗疗法的使用可 以通过 战 略性地识别 与提示相 关的特 定压力源 并适当地选择应对措施来避 免这种毒性(更 多例子见附录)。

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 好 | 好吧 | 坏 |
| (谢和普 2021) | 10 .48% | 27 .42% | 62 . 10% |
| 搅拌机 | 51 .75% | 21 .05% | 27 . 19% |
| Heal-ranked | **61.19%** | **31.34%** | **7.46%** |

表 4: 生成模型在响应不 同提示时重复产生 的响应(重复 的响 应用相同的上标符号标记)。

*到现在我还是很伤心。还是那么心痛* *多*

*听到这个消息我很难过。*†

(谢和普 2021) 搅拌机

*听到这个消息我很难过。你有*

*你有什么爱好吗?*$

*我对你的损失感到遗*

治疗等级

*憾。*

案 例 研 究

讨 论 和 结 论

在这项工作中 ，我们开发了 治疗 ，这是一个基于 R ed dit 上 约 100 万个 与痛苦相关 的对话得出 的知识 图谱。它在将每 个节 点与 41 种情感状态中 的一种联 系起来的 同时 ，在不同 类 型的压力源 、说话者期望、反应和反馈类型之 间形成了 联 系。对 知识 图谱进行 的 统计和可视化分析证 实 了在 治疗 中存在有用 的反应策 略 ，以 降低遭 受痛苦 的人 的 负面情绪 状态。我们进一步评估 了治疗的 效用

|  |  |
| --- | --- |
| (谢和普 2021)  搅拌机  治疗等级 | *做到了。也许明天我应该* *跳到火车前面:(*  *你打算怎么做?*†  *我相信你能做到!* *坚持练习就行* *了*  *你总有一天会成功的!*†  *不要这样*  *做。* |

虽然在这项工作 中使用 的数据是 公开 的 ，但不应该 因为其 中包含高度敏感 的信 息而被破坏。因此 ，根据 B ent on 等人 (201 7)在健康研究中使用社交媒体数据的指南， 本文仅引 用了数据集 的复述节选。由于 H E A L 是通过将长响应拆分 为单个句子来构建的 ，因此将其 公开将无法通过使用逐字 post 文本进行 we b 搜 索来恢 复用户名。只有 与压力源相关 的痛苦叙述 的嵌入将被共享 ，以使基于检索 的模 型 的开发 成为可能。使用 匿名用户名 的 R ed dit 对话可 以根据要求在 特殊条款下与其他学术研究人 员共享。

参 考文献

阿尔梅达 ，d .m .; W ethingt on 大肠 ;和 K essl er, R . C . 2 002 。 每日压力事件清单: 测量每 日压力源的一种访谈方法。*评估，* *9(1):4* *1-55。*

Altho ff t ; 克拉克,k; 和 L es ko ve c, J . 2 016 。咨询对话的大规 模分析 : 自然语言处理在心理 健康方面的一种。 *计算语言* *学协会汇刊，4:4* *63* *-47* *6。*

奥氏小体,美国 ;商业,c; Kobi l arov g; 莱曼,j .; Cygani ak r;和艾夫 斯， Z . 2007 。Dbp edi a:开放数据 网络的核心。在语义网 中， 722-735。

费利克斯,j .;Zan nett ou,美国 ; 基冈,b; 乡绅,m; 和布莱 克本 ，J . 2020 。Pus hshi ft R e ddit 数据集。*国际* *AAAI* *网络与社交媒* *体会议论文集，* *14(1)* *:83* *0-83* *9。*

B ent on; 铜匠,g;和 Dr ed ze, M . 2 017。社 交媒体健康研 究的伦 理研 究协议。第一届 A CL 自然语 言处理伦理研讨会论文集， 94-102。

B ol l ack er k; 埃文斯 ,c; 介 绍 p; ,他 是 t; 和 Ta yl or, J . 2008 。 F ree bas e: 一个用于结构化人类知识的协作创建的图形数据 库。 2008 年 AC M S I G M O D 数据管理国际会议论文集， 1247- 125 0。

棕色 、t;曼,b; 莱德 :; 印度,m; 等。 202 0。语言模型是少样本 学 习者。《神经信息处理 系统 的进展》 ，卷 3 3,187 7 - 1901。

查特吉 ，a;古普塔,美国; 钦纳科特拉 ，m .k .; S rikanth r; 厨房,m; 和阿格拉 瓦尔 ，2019 年 p。用深度 学习和大数据理解文本 中的情感。 *人类行为中的计算机，93:* *309-31* *7。*

d 'A vil a Ga rc ez, A.; 兰姆 ，l .c. 20 20。神经符 号 AI: 第三波。 arXiv: 2012.0587 6。

D evl in, j .; Ch ang, 硕 士 。李 ,k; 和 图 塔 诺 瓦 (To ut anov a, K. 2019) 。B ER T: 用于语言理解 的深度双向 transfor m er 预训练。 在计算语言 学协会 2019 年北美分会会议论文集: 人类语 言 技术，第 1 卷(长短论文)，4 171 - 4186。

Ekman, 1 992 年第 1 期。基本情绪的一种。*认知与情感，6(3-* *4):169-2* *00。*

费边, m; Gj ergj i k; 格, w ; 等。 2 00 7。Y a g o: 统一 w or dn et 和 wi kip edi a 的语 义 知识核 心 。在第 16 届 国 际 万 维 网会议 上， WW W , 697 -70 6。

汉族,美国 ;爆炸,j .; Ryu,美国 ;和 L ee G. G . 2 015。利用知识库 为 自然语 言对话倾 听 agent 生成 响应。《 话语 与对话特别 兴趣小组第 16 届年会论文集》 ，129 - 133。

许,c c。陈、 S .- Y; 郭、 c c; 黄、郭宏源 ;以及 K u l . w 。20 18. 情感线 :多方对话 的 实证研 究。发表于第十一届 国际互联网 会议论文集

*语言资源与评价学术会议(语言资源与评价2018)。*

Lee, D . 2016 。Tay:微软就种族主义聊天机器人惨败致歉。

*BBC* *新闻。*

刘易斯, m; 刘,y; G oy al :; Gh az vininej ad m; 穆罕默德， a; L e vy, o .; S t oya no v 诉 ; 和 L . 202 0 年的 Zett l e moy er。B AR T: 去噪 序 列到序 列预训练 ，用于 自然语 言生成、翻译和理解。计 算语言学协会第 58 届年会论文集 ，787 1 -7 88 0。

李,j .;厨房,m; Bro ckett c; 高,j .;和 多兰，B . 20 16。神经会话模 型的多样性促进目标函数。在 2016 年计算语言 学协会北美 分会会议论文集:人类语言技术，110 - 119。

林,z; 马托(A. ); 胫骨,j .; 徐,p;冯德伦， 2019 年 p。M oE L: 感同 身 受 的听众。在 20 19 年自然语言处理经验方法大会暨第 九届自然语言处理国际联合会议( E M N LP -IJ C N LP )论文集 中， 121- 13 2。

Li son p; 蒂,j .; Kou yl eko v m; 等。 2019。Op en 副标题 2018: 在 大 型、有 噪 声 的平行语料库中对句子对齐进行 统计重评分。 语言 资源 与评价 2018 ，第十一届 国 际语 言资源 与评价会议。

刘、 C .- W; 劳,r; S erb an i ;Nos e w orthy m; Ch arl in l ; 和 Pine a u, J . 201 6。 如何不评估 你 的对话 系 统 : 对对话 响 应生成 的无监 督评估指标的实证研 究。《 2 016 年 自 然语 言 处理经验方 法会议论文集》 ，21 22 - 213 2。

刘,美国; 陈,h; 任,z; 冯,y; 刘问:; 和尹丹 。2018。神经对话生成 的知识扩散 。在计算语 言学协会第 56 届年会论文集(第 1 卷:长论文)中， 1489 - 149 8。

刘 ,y; 奥 特 ,m; Go yal :; 杜 ,j .; 乔 希 ,m; 陈 ,d; Lev y, o .; 刘 易 斯,m;Z ett l em oye r l ; 和斯托亚诺夫， 2019 a。R ob ert a: 一种鲁 棒优化的 bert 预训练方法 。*arXi* *v* *预印本* *arXi* *v:190* *7.11* *692。*

刘,z; 妞妞 , Z .- Y。; 吴, h; 和王， H . 20 19 b。增广 图上基于可 解释推理 的 知识感 知对话生成。在 20 1 9 年自然语言处理 经 验 方 法 会 议 论 文 集 和 第 九 届 自 然 语 言 处 理 国 际 联 合 会 议(EM N L P -I JCN L P )中， 17 82 - 1 79 2。

M aju md er: ; 在 香 港 ,p; 彭 , 美 国 ; 陆 ,j .; Ghos al d; G el - bu kh, A.; Mih al ce a r; 和 P ori a, S . 2 020。哑剧 :模仿情感 以产生 同理 心反应。《 2020 年 自然语 言 处理经验 方法会议论文 集》 (EMN L P ) ， 8968- 897 9。

Misuraca m;斯帕诺,m;和巴尔比， S. 2019 。BMS:用于文 档聚类验证的实证研究。

11466

*统计理论与方法，* *48(20):5* *036-* *5049。*

Murt agh f; 和 勒 让 德 (Leg endr e) ， 2014 年 P . 。 W ard’s hi erarchi cal a ggl omer at ive clust ering m ethod: 哪些算法实现了 W ard’s crit eri on?*分类学报，31(3):274-29* *5。*彭, 中州。张成 泽,j .;比翰， J . P .;帕维 尔，A. 2021 年。一语道破 :改善非视觉 呈现可及性的反馈。在 2021 年 CHI 计算机 系统人因会议论

文集中， 1-12。

普鲁契 克，R . 198 4。 《情绪:一般的 心理进化 理论》。 *《情*

*绪的途径》，* *1984:19* *7-219。*

雷德福; 吴,j .;孩子,r;烹调的菜肴,d;A mod ei d; 和 Sut skever, I . 2019。语 言模 型是无监 督的 多任务学 习者。*Op* *enAI* *博客，* *1(8):9。*

R affel c; S ha z eer: ; 罗伯茨， A.; 李,k; 纳 ,美国 ; Mat e na m; 周,y; 李,w; 和 Liu, P . J . 2 020。用 统一的 T ext -to- Te xt T rans for m er 探 索迁移 学 习 的 局 限性 。*机器学习研究学报，21(1* *40):* *1-* *67。*

R ashkin h; 史密斯 ，e.m .; 李,m; 和 B our ea u, y .l 。2019.移情开 放域对话模 型 :一个新的基准和数据集 。在第 57 届计算语 言学协会年会论文集中 ，5370 -538 1。

雷蒙 :; 和古雷维奇， I . 2 01 9。S e nt e nc e- b ert : 使用孪生 b ert 网络的句子嵌入。 2 01 9 年自然语言处理经 验方法会议论 文集和第九届自然语言处理国际 联合会议论文集， 39 82 - 3992。

Ridner, S. H. 2004。《心理困扰 :概念分析》。 *先进护理杂* *志，45(5):536-* *545。*

辊、美 国; Dina n 大肠; Goy al :; 朱， D.; 威廉森; 刘,y; 徐,j .; Ott , m; 史密斯 ; 布里奥 ; 韦斯顿( W est on) ， 20 21 年。构建开放域聊 天机器人的秘诀。发表于计算语言学协会 欧洲分会第 16 届 会议论文集:主卷， 300-3 25。

卢梭 ，P . J . 1 987 。剪影 :对 聚类分析的解释和验证 的 图形辅 助。 *计算与应用数学学报，2* *0:5* *3-65。*

俄文,诉 ; 和 Lint ean, M . 201 2。基于词 到词相似性度量的贪 婪和最优 自然语 言 学生 输入评 估 的 比较 。在第 七届使 用 NLP 构建教 育应用研讨会论文集中 ，157- 16 2。

萨 普 ， m; 勒 · 布 拉 斯 ， r;Al l aw ay 大 肠 ; Bhagav atul a c; 劳 里:; R ashkin h;屋顶,b;史密斯， n .a.;和 Choi Y. 20 19 。原子 :机 器常识图集，用于 if-then 推理。《 A A AI 人工智能会议论文 集》，第 33 卷， 3027- 3035。

S kerry , A. E .; 和萨 克斯 ，R . 20 15。情绪 的神经表征是围绕 抽象事件特征组织的。 *当代生物学，25* *(15)* *:1945* *-195* *4。*

斯皮 尔,r;下巴,j .; 和哈 瓦西 ，C . 201 7。概念网 5 .5: 机器 常识 开放 的多语 言通用 知识 图谱 。AAAI 人工智能会议论文集， 第 31 卷。

助教,o .;和Kiyani, F. 2007 。综述自动文本摘要。*新闻学术* *进展，* *5(1):205-213。*

V as wa ni , a; S ha ze er: ; P ar m ar: ;Usz kor eit , j .; 琼斯， L .; 戈麦斯， A. N.; 路 易斯安那州凯泽; 以及波洛苏欣(P ol osukhin, I . 20 17)。 注 意 力 就 是 你 所 需 要 的 。 在 盖 恩 ， 我 ; Lux bur g, U. V.; B en gi o,美国 ; 瓦拉赫,h; 费格斯 r; Vi sh wan ath an,美国 ; 和加 内特， R, eds。， 《 神经信 息处理 系统进展 》，第 30 卷。

vande⁺ ci´c, d;和 Kr ¨ ot zsch, M . 2014 。Wiki dat a:一个免费的 协作知识库。 *ACM* *通讯，57(10):78-85。*

W el ivit a, a; 和 Pu, P . 20 20。人类社会对话 中 同理 心反应 意 图 的分类。第 28 届计算 语 言 学 国 际会议论 文集 ，488 6 - 4899。

谢,y; 和濮 存 昕 ，P . 2 02 1。 用大规模 对话数据 集生成感 同 身受的回应。第 25 届计算 自 然语 言 学 习 会议论文 集(即 将发表)。

杨,z; 戴,z;杨,y; C a rbo nel l j .; 萨拉赫丁诺夫 ，r · r; 以及勒 ，Q. V. 20 19。XL N et : 用于语 言理解的广义自 回归预训练。神经 信息处理系统进展， 第 32 卷。

年轻 、t; 威 尔士,大肠 ;查 图 尔维迪, 即 ; 周,h; B i sw as 美国 ; 黄， M . 201 8。用常识 知识增强端到端对话 系统 。A A AI 人工智 能会议论文集 ，第 32 卷。张,h;刘,x; 锅,h;歌,y; 和梁朝伟 ，C . 2020。AS ER : 大规模 的不测 知识图谱 。*2020* *年网络会议论* *文集，201-2* *11。*

张,j .; 和 Da nes cu - Ni cul es cu - Mi zi l , C . 20 20。咨询对话 中的 平衡 目标: 向前推进还是 向后看。计算语 言学协会第 58 届 年会论文集， 527 6 -5 289。

周,h;年轻、t; 黄 M .; 赵,h;徐,j .;和朱 X . 20 18。基于图注 意力 的 常识 知识感 知对话生成。《第二十七届国 际人工 智 能联 合会议论文集》， IJCA I- 18 , 4623 - 462 9。

朱,w; 密苏里州,k;张,y; 朱,z; 彭,x; 和杨 Q. 201 7。面 向 知识基 础 对 话 的 灵 活 端 到 端 对 话 系 统 。 *arXi* *v* *预* *印* *本* *arXi* *v:17* *09.04* *264。*

11467