

订阅DeepL Pro以翻译大型文件。

欲了解更多信息， 请访问www[.DeepL.com/pro。](https://www.deepl.com/pro?cta=edit-document&pdf=1)

第三十六届AAAI人工智能会议 (AAAI-22) 。

**HEAL:** 困境管理对话的知识图谱

**Anuradha** **Welivita,** **Pearl** **Pu**

洛桑联邦理工学院计算机和通信科学学院 瑞士

*{kalpani*.welivita, pearl.pu}*@*epfl.ch

摘要

现代世界的需求越来越多地造成了心理负担，并对我们 的心理健康带来不利影响。因此，具有移情反应和困境 管理能力的神经对话代理最近得到了普及。然而，现有 的端到端移情对话代理往往会产生一般的和重复的移情 语句，如

*"我很抱歉听到这个消息"*，这不能表达对特定情况的特 殊性。由于这种模式缺乏可控性，它们也带来了产生有 毒反应的风险。聊天机器人利用knowl-

edge图进行推理，被认为是比端到端模型更有效、更安 全的解决方案。然而， 这种资源在情绪困扰的情况下是 有限的。 为了解决这个问题，我们介绍了**HEAL**，这是 一个基于100万个困扰的叙述和他们相应的安慰反应从R eddit策划出来的知识图。它由22K个节点组成，识别不 同类型的压力源、说话人的期望、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104K个连 接。每个节点都与41种情感状态中的一种相关联。在H EAL上进行的统计和视觉分析揭示了在面向困境的对话 中说话人和听众之间的情感动态，并确定了导致情感缓 解的有用反应模式。自动和人工评估经验表明，与基线 相比， HEAL的反应更多样化、更有说服力、更可靠。

简介

现代世界的需求越来越多地造成了心理负担， 给我们 的心理健康带来了不利影响 。困扰是指一个人对特定的 个人压力或需求所经历的不舒服的情绪状态 ，这种状 态会对人造成伤害 ，无论是暂时的还是永久的(Ridner 2004)。 这样的压力源包括与亲人分离，个人间的冲 突，某些心理健康状况，如抑郁症， 工作表现不佳， 和睡眠问题，如失眠。 Almeida等人 (2002)的一项研 究，通过每天的电话访谈，测量了美国全国1031个成 年人的日常压力的多个方面，发现他们至少经历了一 个日常压力因素

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org).保留所有权利。

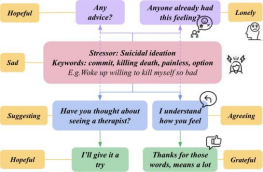


图1：*HEAL*的部分图示。红色、白色、蓝色、绿色和

黄色节点分别代表压力源、说话人的期望、反应和反

馈类型以及相关的情感状态。

在40%的研究日里。人们通常倾向于在日常对话中分享 这种经验。因此，嵌入具有适当的移情反应能力的开 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(Rashkin等人， 2019；Lin等人， 2019；Majumder等人， 2020；Xie和Pu 2021)。

随着复杂的神经网络架构的发展， 如变压器(Vaswa ni等人， 2017)和预训练的语言模型，如BERT (Devli n等人， 2019)， RoBERTa (Liu等人， 2019a)和GPT- 3 (Brown等人， 2020)， 微调非结构化文本的神经反 应生成模型已成为构建聊天机器人的常见方法之一 。 虽然它避免了严格基于规则的方法的大部分局限性， 并使聊天机器人在很大程度上可以泛化到未见过的领 域，但缺乏可控性和黑箱性质使这些模型的可靠性和 故障安全性降低(d'Avila Garcez和Lamb 2020)。当用户正在经历痛苦的情况，对错误的信息 和不恰当的评论很敏感时，这就特别有问题。最近的 一个例子是微软的Tay机器人，在得知Twitter上的种族 主义和攻击性信息后，开始产生非故意的、攻击性的 、否认大屠杀的种族推文(李2016)。

因此，人们对使用知识(Zhu等人， 2017；Liu等人

，2018；Han等人， 2015)和com-

monsense推理(Zhou等人， 2018；Young等人， 2018

) 越来越感兴趣

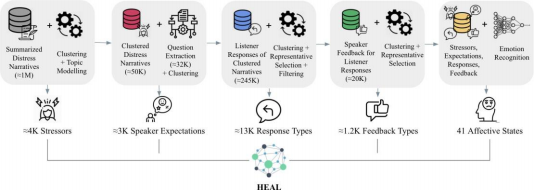


图2 ：开发知识图谱HEAL的分步过程。

在基于图形的表述上，产生适当的和信息丰富的对话 反应。与在非结构化文本上的训练相比，使用基于图 形的表述为生成的反应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，使其有可能沿着可预测的路线引导 对话流，同时也提供了战略上多样化回应的能力(Liu 等人， 2019b)。

虽然存在大规模的知识图谱，如Concept- Net (Speer, Chin, and Havasi 2017)和ATOMIC (Sap et al. 2019)，但它们主要通过捕捉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协助开放领域的 对话生成。由于它们不是为了捕捉移情交流的规范而 开发的，所以这个领域缺乏语言资源和模型来协助困 境管理和移情再回应的生成。而且，没有人尝试过生 成知识图谱来表示整个对话的上下文-

反应对之间的关系。为了解决这些限制，我们介绍了**H** **EAL** (意为治疗、移情和情感学习) ，这是一个用于 困境管理对话的知识图 ，它是通过分析从精心挑选的su breddits的压力事件的叙述和相应的反应线程而开发的

。

HEAL由五种类型的节点组成： **1**) *压力源*：造成痛 苦的原因；**2**) *期望*：在痛苦的叙述中说话者经常提出 的问题； **3**) *反应类型*：听众为解决不同的压力源而给 出的最常见的反应类型；**4**) *反馈类型*：说话者在反应 后提供的主要反馈类型；以及**5**) *情感状态*：与每个节 点相关的情感状态。这里的说话者是经历了痛苦情况 的人(那些通过在Reddit上发帖开始对话的人)，听众 是这些帖子的评论者。图1显示了HEAL中一个典型的 压力源的例子。HEAL构成了与压力有关的话题，它可 以准确地描述以压力为导向的对话中的基本背景，从 而使对话模型能够检索到更具体的背景反应 。

此外， 还有一些信息，如这种反应是否会导致

正面或负面的反馈，以及他们是否解决了IM--的问题。

对处于困境中的人的明确期望可以导致选择更合适和 有用的反应。如图2所示， 我们采取了一系列步骤，包 括总结、聚类、主题建模和情绪分类，从Reddit策划的 超过100万条困境日记中开发了HEAL。这导致识别出≈ *4K*压力源、 ≈*3K*说话人的期望、 ≈*13K*反应类型、≈*1*.2K 反馈类型以及相关的有效状态。最终的图构成了22,037 个节点和不同类型节点之间的104,004个连接。

通过对HEAL进行统计和视觉分析，我们能够发现说 话人和听话人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HEAL在下游任务中的效用 ，即对一个给定的痛苦情况产生同情的反应。我们利 用知识图谱开发了一个基于检索的模型，并利用自动 和人工评估将其性能与两个最先进的移情对话代理进 行了比较：一个由Xie和Pu (2021)开发；以及Blender (Roller等人， 2021)。 结果显示，在多样性和同理心 的适当性方面，用知识图谱以排序的方式检索出的反 应优于其他的反应。通过一个案例研究，我们还表明 ，由HEAL检索的反应比神经反应生成模型更可靠。我 们的主要贡献包括：1)开发了一个大规模的知识图谱 ，**HEAL**，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困境对话相关的有效状态； 2)使 用统计和视觉分析来识别说话人和听话人之间的情绪 动态以及导致情绪降级的有利反应模式；以及3) 评估 HEAL在检索应对情绪困境时更适合移情的、多样化和 可靠话语的有用性。 1

相关工作

知识图谱由于其有用性， 已经引起了自然语言处理界 的注意。

1代码和数据见github.com/anuradha1992/HEAL。

在理解自然语言输入方面。最近出现的链接开放数据 ，如DBPedia (Auer等人， 2007年)和谷歌知识图谱， 推动了这一点。2YAGO (Fabian等人， 2007)、 Freeba se (Bollacker等人， 2008)和Wikidata (Vrandecˇic´和 Kro¨tzsch，2014)是其他一些建立在从网络中提取的 一般知识上的知识图谱的例子。最近的知识图谱，如C on- ceptNet (Speer、Chin和Havasi 2017)、 ATOMIC (Sap等人， 2019)和ASER (Zhang 等人， 2020) 侧重于代表不同类型的常识知识。 Liu等 人(2018)和Zhang等人(2020)的作品利用这些图中 存在的事实性和常识性知识来开发开放领域的对话代 理， 产生更多的语义和信息性反应。

尽管上述资源在开发知识感知的对话代理和具有推 理能力的对话代理中很有用(Zhou等人， 2018)，但 通常这些图解决的是开放领域的实体和关系以及建立 在它们之上的常识性推理。它们不具备情感推理和移 情再反应生成的规范。 HEAL通过建立压力源、说话人 表达、反应、反馈和情感状态之间的关系，并将提示- 反应-

反馈图元联系起来，以确定有可能导致有利反馈的反 应，并解决处于困境中的人的隐性期望，从而扩展了 上述限制。

方法论

数据集的整理

公开可用的情感对话数据集，如Em- patheticDialogues (Rashkin等人， 2019)、 EmotionLines (Hsu等人， 2018)和EmoContext (Chatterjee等人， 20 19)，大多包括在人工环境中创建的开放域和日常对 话，或从电影/电视字幕中策划。用于进行再中心研究 的真实咨询对话数据集(Althoff, Clark, and Leskovec 2016; Zhang and Danescu-Niculescu-Mizil 2020)由于伦理原因，不能直接访问。因此，我们从R eddit上策划了一个新的数据集，其中包含了讨论真实 世界的痛苦情况的对话。我们选择了Reddit，因为它是 可以公开访问的，而且同龄人会积极地参与到这样的 平台中来支持其他正在经历精神痛苦的人。

我们使用Pushshift

API (Baumgartner等人， 2020)来收集和处理来自一组 精心挑选的8个子红点的对话线程： *mentalhealthsupport*

；*offmychest*；*sad*；*suicidewatch*；*anxietyhelp*；*depressi* *on*；*depressed* ；和*depression*

*help*，这些是Reddit用户中流行的发泄痛苦的方式。我 们通过匹配Au-

thor的名字， 明确地从这些线程中提取了对话的转折结 构，并对这些对话进行了严格的数据清理程序，其中 包括从听众的回答中去除亵渎的内容。通过这个方法 ，我们能够策划出1*,*275*,*486个具有3*,*396*,*476个对话回 合的双人对话(平均每个对话2*.*66个回合)。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

2en.wikipedia.org/wiki/谷歌知识图谱

来得出知识图谱，并保留10%的数据用于验证和测试

下游任务。

归纳总结

从Reddit策划的困境叙述通常很长(平均每轮84.89个

令牌)， 有些超出了某些基于预训练的语言模型架构

的输入令牌长度， 如句子-BERT (Reimers和Gurevych

2019)。因此，我们研究了各种总结算法，这些算法

可以用来生成保留叙事本质的总结。

我们研究了提取式和抽象式总结技术来解决这个问

题(Tas和Kiyani

2007)。其中，抽象式总结方法主要是在结构化文档

(如新闻文章) 上进行训练和测试的， 已知其在非结

构化文本上表现不佳(Peng等人， 2021)。因此， 我

们选择了五种不同的外部总结方法： SMMRY的定制

实现--Reddit的TLDR机器人背后的算法(https://

smmry.com)；以及四种不同的预训练模型--

BART (Lewis等人， 2020)、 GPT-

2 (Radford等人， 2019) 、XLNET (Yang等人， 2019

)和T5 (Raffel等人， 2020)用于内容重要性建模。

我们对100个Reddit dis-

tress叙述的样本进行了人工评分，将上述方法生成的

摘要评为*好*、*好*和*坏* (结果详见附录)。被评为 "*好*

*"的*摘要中，SMMRY算法的比例最高。因此，它被选

择来总结冗长的对话回合(有*≥*100个标记的回合)。

大约有43%的对话转折是用这种方法总结的。

聚合聚类法

由于人工注释成本高且耗时长， 特别是当应用于大规

模的数据集时，我们决定使用自动聚类来识别可明确

区分的压力源类型、期望、反应和来自Reddit苦恼对

话的反馈类型。为此， 我们使用了为大型数据集调整

的 "聚合聚类" (Murtagh和Legendre

2014)。它递归地合并那些最小限度地增加给定联系

距离的集群对。联系距离是使用Sentence-

BERT (Reimers和Gurevych

2019)生成的嵌入对之间的余弦相似性来计算的，因

为所产生的嵌入已被证明是高质量的，并且在文档级

嵌入中工作得非常好。在附录中详细解释了选择使用

聚类而非其他聚类方法的原因。

识别压力源

我们试验了8个相似度阈值，从0*.*6到

0*.*95，增量为0*.*05，对困境叙述进行聚类。尽管对每

个阈值都计算了各种聚类质量指标，如Silhou-

ette系数(Rousseeuw 1987)、 Dunn指数(Misuraca,

Spano, and Balbi

2019)和平均点到正弦距离，以选择最佳相似度阈值

，但在每个阈值下对10个聚类子集的人工检查和聚类

可视化重新显示，这些指标对这个数据集并不奏效

压迫者 关键词提取

|  |  |
| --- | --- |
| 自杀意念承诺  焦虑攻击焦虑  孤独  失败的大学学习  美国选举特朗普 | *、杀戮、死亡、无痛、选择*  *、焦虑、攻击、社会、攻击*体重训练 *、体重、吃、减、肥*  *孤独，周围，连接，孤立，社会* *，大学，班级，学期，失败*酗酒者 *喝酒,* *饮料,* *酒精,* *醉酒,* *清醒*  *,* *总统,* *唐纳德,* *选举,* *战争* |

Covid19 *covid,* *19,* *大流行,* *shambolic,* *带来了*

表1 ：使用TF-

IDF在苦恼叙述的集群中确定的一些稳定因素。

(众所周知， 上述指标只对具有凸形聚类的数据集效果 最好)。人工检测的结果表明，在较高的阈值 (如0.95 和0.9) 下确定的压力源过于具体，而低于0.8的压力源 则过于模糊(附录中包含了在每个阈值下通过人工检 测发现的聚类质量度量和主题)。这导致选择了一个 最佳的阈值为0*.*85。在这个阈值下， 4.93%的困境叙述 (共47 *，* 109条叙述) 被分成4 *，* 363个集群。在对这些聚类进行基于TF-

IDF的主题建模后，我们发现了一些明显可区分的压力 源， 这进一步验证了聚类的良好性。表1显示了在这个 过程中发现的一些压力源。

期望、回应和反馈类型

在对苦恼叙述进行分组并确定其再思考的主题后，我 们使用简单的字符串搜索包含"？"的句子来提取分组苦 恼叙述中提出的问题。相应的回答和相关的反馈也被 提取出来。我们使用NLTK li- brary来分离反应和反馈中的单个句子，这样就很容易 通过聚类来识别独特的反应和反馈类型。这样一来， 我们就能够收集到32 832个期望， 245 707个回应和20 213个反馈。按照上述类似的最佳阈值选择过程，我们 分别选择了0*.*7、0*.*75和0*.*7作为对期望、回应和反馈进 行聚类的最佳阈值。这样就分别产生了3 050个、 13 416个和1 208个期望、反应和反馈类型，每个集群至少有两个独 特的集群元素。特别是回应集群 ，要经过自动和人工 验证的过程，以去除Reddit特有的回应 (例如：*请联系* *subredddit的版主*)、 由机器人产生的再赞助(例如： *这个动作是自动执行的*)和半生不熟的回应(例如： *嘿*，*哇*) 。与最终聚类结果有关的统计数字显示在表2 中。我们在每个聚类中随机选择一个成员作为聚类代 表。 频繁的期望、反应和反馈类型的例子包括在附录 中。

情感状态建模

为了将每个压力源、期望、反应和反馈集群与情感状 态联系起来，我们使用了Welivita和Pu提出的基于BER T变换器的分类器。

(2020)在EmpatheticDialogues数据集上进行训练。它的 分类准确率为65.88%，与最先进的对话情感分类器相 当。 该分类器能够将文本归入41个情感类别中的一个 ，其中32个是从多个注释方案中选出的积极和消极情 感，包括从生物反应中得出的基本情感(Ek- man 1992；Plutchik 1984)到从背景情况中得出的较大的微妙情感集(Sker ry和Saxe

2015)， 还有9个是用来阐述中性情感的移情反应策略 。我们用这个分类器对属于一个群组的每个文本进行 分类 ，并将该群组与出现次数最多的情感状态联系起 来。如果两个或更多的情感状态出现的次数相等 ，我们 就把每个状态的分类器的置信度加起来，选择置信度 最高的那个。 通过这个过程，我们能够确定与每个集 群相关的最突出的情感状态。

**HEAL:** 统计分析

我们跟踪了每个期望和反应所来自的痛苦叙述的压力 源标识符，并能够在压力源和期望和反应集群之间形 成联系。我们还跟踪了获得每个反馈的对话标识符， 这有助于在反馈群组和期望及反应群组之间建立联系 。最终的知识图谱， HEAL，由22*,*037个节点和104*,*004 个节点之间的连接组成。在压力源和期望之间有9*,*801 个连接，在压力源和反应之间有56*,*654个连接，在反应 和反馈之间有10*,*921个连接，在期望和反应之间有26*,*6 28个连接。此外，每个节点都与一个积极的状态相关 ，形成22*,*037个连接。

图3显示了与压力源、期望、反应和反馈类型相关的 情绪状态的分布。根据统计，73.60%的压力源与消极

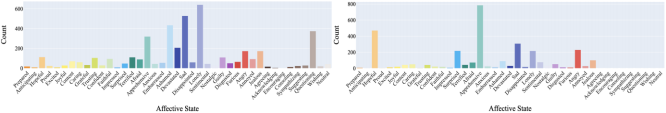
的情绪状态有关。其中，*孤独*、*悲伤*、*羞愧*和*忧虑的* 情绪与44.01%的压力源有关。大多数期望与消极的情 绪状态有关，如： *兴奋* (25.70%)、 *悲伤* (10.07%)

和*愤怒* (7.51%)，也与积极的情绪状态如： *希望* (15. 41%)有关。在这些回答中，60.38%与中性情绪状态 有关。其中*质疑* (12.89%)、 *同意* (9.22%)和*建议* ( 6.90%)比其他的更突出。一个重要的观察结果是，与 压力源相比，在反馈群中，可以看到积极情感状态增 加了7. 17%，中性情感状态增加了270.29%。与反馈群 组相关的消极情感状态与压力源相关的情感状态相比 ，下降了44.77%。在反应群中， 28.59%与至少一个反 馈群有关 ，其中100%的反应与至少一个积极或中性反 馈有关。其中，26.51%的反应与至少一个积极的反馈

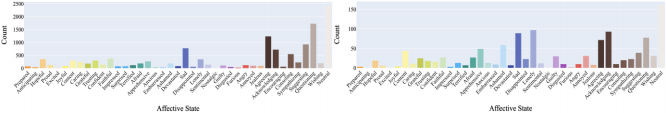
有关， 77.48%的反应与至少一个中立的反馈有关，这 就验证了反馈的存在。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 类型 | 阈值 | # ＃集群 | 最大的 集群大小 | 归根结底。 # 文件  聚集的 | 文件的百分 比  聚集的 | 剪影  系数 | Dunn-Index (余弦) | 平均余弦值 距离。 |
| 压力源 | 0.85 | 4,363 | 11,856 | 47,109 | 4.93% | 0.0554 | 0.0677 | 0.0443 |
| 期待 | 0 .7 | 3,050 | 489 | 16,316 | 49.7% | 0.3781 | 0.1008 | 0.0649 |
| 回应 | 0.75 | 13,416 | 1,025 | 78,194 | 31.82% | 0.3263 | 0.1061 | 0.0722 |
| 反馈信息 | 0.7 | 1,208 | 960 | 5,782 | 28.61% | 0.2882 | 0.1705 | 0.0895 |

表2 ：与最终聚类结果有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 表示点到中心点的平均余弦距离。 Silhouette系数和Dunn指数的值分别位于[- 1, 1]和[0, *∞*]之间。这些值越正越好。



(a) 压力源 (b) 期望



(c) 回应 (d) 反馈

图3 ：与HEAL中的压力源、期望、反应和反馈有关的情绪分布状态。

在HEAL中， 有一些有用的反应类型，可以缓和受困者 的负面情绪状态。

视觉化和解释

我们使用vis.js (visjs.org)，一个图形可视化库来可视 化所产生的知识图。图4显示了该库生成的知识图的部 分可视化。节点的大小与各自集群的大小相对应，边 的宽度与不同集群之间的连接数相对应。每个不同的 压力源、期望、再反应和反馈类型也都与一个情感状 态相关联，为了避免杂乱，这里没有将其可视化。

正如关键词所表示的，中间的压力源节点代表了含 有*自杀想法*的叙述。如图所示 ，一个有自杀念头的人 最常见的期望是： *他应该怎么做*；*听众是否有同样的* *感觉*；以及*他有哪些选择*。在这种情况下，倾听者最 常见的反应是：同情的反应，如*我很抱歉你有这种感* *觉*；安慰的反应，如*我希望你感觉好些*；有意义的问 题，如*你想谈谈吗？* *你有没有寻求帮助？*

*获得推荐*；以及鼓励性的回答，如*坚持下去，我的朋* *友*，*保持坚强！*。通过紫色的虚线，我们可以看到常 见的说话人期望和听众反应之间的联系。例如， "*我有*

*同样的感觉* "与 "*还有人有这种感觉吗？"*而 "*坚持住，我的朋友* *"*和 "*你在看医生或治疗师吗？"*则与

"*我该怎么做？"相*联系。可以看出，这些反应大多与说 话者的积极反馈有关，如*感谢*你*的回答*，对听者表示 感谢，同时也说明这是一个好的反应。

评估**HEAL**在应对困境提示中的效用

我们评估了HEAL为给定的痛苦对话提示检索适当的移 情反应的能力，并将其性能与现有的最先进的移情反 应生成模型进行比较。为此，我们使用了一开始就分 开的10%的Reddit对话进行测试。为了从HEAL中检索 反应，我们计算了新的叙述/提示和属于知识图谱中不 同集群的现有叙述之间的余弦相似度，并将新的叙述 与现有叙述中相似度最高的集群相关联。在测试数据 集中的123,651个对话提示中， 60.7%的对话提示与现有 叙述的相似度为0*.*75或以上。

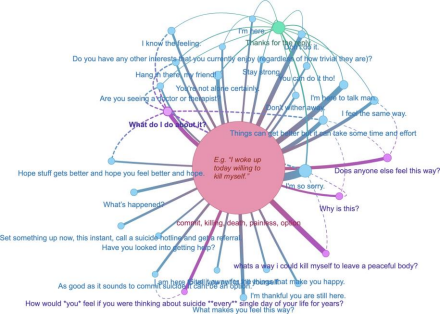


图4 ：通过vis.js对HEAL的部分内容进行可视化。压力源、期望、反应和反馈类型分别用红色、紫色、蓝色和绿色表 示。 为了避免杂乱， 只有具有重要边缘权重的连接被可视化。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 模型 | D1 | D2 | D3 | D4 | BLEU1 | BLEU2 | 梅特-  鲁格(METEOR  ROUGE | | 全球性的 |
| Reddit | (Xie and Pu 2021) 搅拌器  痊愈等级的 | 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 | 0.0277  **0.0469**  0.0252 | 0.0475  **0.0661**  0.0332 | **0.6921**  0.6047  0.6599 |

表3：对Reddit中的窘迫提示做出反应的任务获得的自动评估结果。 D1、D2、D3和D4代表Distinct- ngram指标(Li等人， 2016)， GM代表Greedy Matching得分(Rus和Lintean，2012)。

我们对知识图谱中涵盖的压力源进行了评估，并将其 归档。然后，我们对与新叙述相关的压力源的反应进 行排序， 首先根据压力源和反应之间的边缘权重，然 后根据反应集群的大小，选择排名靠前的反应。我们 称其为**HEAL-**

**ranked**。在这个基线建议中，与说话人期望和反馈类 型的联系没有被考虑在内。但是，我们将详细解释这 些节点如何对改进这个基线作为未来工作的一部分作 出贡献。

我们将HEAL-

ranked检索的反应与两个最先进的移情反应生成模型进 行比较，一个是Xie和Pu (2021)开发的，另一个是Bl ender (gen-

erative)(Roller等人， 2021)。前者是一个基于RoBE RTa (Liu等人，2019a)的多轮情感参与对话生成模型 。它在OpenSubtitles (Lison等人， 2019年)的≈*1M*条对 话上进行预训练 ，并在EmpatheticDialogues (Rashkin 等人， 2019年)上进行微调 。后者是一个标准的基于S eq2Seq转化器的移情开放域聊天机器人。它在包含*15* 亿条评论的Reddit讨论中进行了预训练，并在几个较小 但重点突出的数据集上进行了微调。

自动评估

表3包括对上述模型为Reddit对话产生的再赞助所计算的 自动指标

提示。我们可以观察到，在用来衡量反应多样性的Dis

tinct-

N指标方面， HEAL排名超过了其他的指标 (Li等人，

2016)。 这表明HEAL在产生比现有的神经再反应生

成模型更多样化的反应方面的效用。我们在表4中进

一步证明了这一点，显示了三个模型对几个与苦恼有

关的提示所产生的一些反应实例。可以看出 ，Blender

和Xie和Pu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HEAL检索到的反应则更加多样化

，并且在顶部针对给定的情况(更多例子包括在附录

中)。我们还观察到， HEAL-

排名在其他自动计量学BLEU、METEOR和ROUGE方

面表现不佳。然而，众所周知，这些计量学与人类判

断的相关性很差(Liu等人，2016)，当与人类评价实

验的结果相比较时，可以很好地看到这一点，这将在

下一节讨论。

人的评价

我们设计了一个人类评估实验，从Amazon Mechanical

Turk (AMT)招募人群工作者，以评估三个模型所产

生的反应的移情适当性。我们从Reddit测试数据集中

随机选择了200个对话，由群众工作者进行评估。工

人们被指示拖动和

|  |  |  |
| --- | --- | --- |
| 提示  *我哥哥两年前去世了，我仍然很* *伤心。它仍然是如此的痛苦*  (Xie and Pu 2021) 我 *很遗憾听到这个消息。t* | | |
| 搅拌器  HEAL-等级 | *我很抱歉听到这个消息.*  *你有什么爱好，你喜欢做什么？‡* *我为你的损失感到遗憾。* | |
| 提示  *我似乎感觉不到快乐，但我并没* *有被压迫。*  (Xie and Pu 2021) 我 *很遗憾听到这个消息。t* | | |
| 搅拌器  HEAL-ranked | | *我很抱歉听到这个消息.*  *你有什么爱好可以帮助你感觉更好吗* *？‡*  *最重要的是，如果你感觉到------* *就可以了。*  *按规定。* |

表4 ：生成模型产生的反应在回应不同的提示时是重复 的(重复的反应用相同的上标符号标记) 。

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 良好 | 好的 | 坏的 |
| (Xie and Pu 2021) | 10.48% | 27.42% | 62. 10% |
| 搅拌器 | 51.75% | 21.05% | 27. 19% |
| 痊愈等级的 | **61.19%** | **31.34%** | **7.46%** |

表5 ：三个模型所产生的反应的移情适当性的人类评价 结果。

将模型产生的反应分为*好*、*好*、*坏*三个区域，取决于 这些反应对给定提示的适合程度。三个人被允许在一 个任务上工作，结果是根据多数人的投票来计算的。

包括用户界面在内的人类实验的具体细节包含在附录 中。 实验结果在表5中表示。在85%的时间里，工作人 员之间至少取得了2/3的一致意见，并且产生了0*.*23的 测评者之间的一致意见(Fleiss'kappa)分数，表明了 公平的一致。根据结果，可以看出， HEAL-

排名产生了最多的*良好*反应，在以下方面优于Xie和Pu 的模型和Blender

的同理心适当性有很大的差距。

案例研究

在表6中，我们强调了几个偷梁换柱的例子，在这些例 子中，神经反应生成方法在回应与痛苦有关的提示时 提出了危险性。可以看出，使用HEAL可以通过战略性 地识别与提示有关的特定压力源并适当地选择反应来 避免这种毒性 (更多例子见附录)。

讨论和结论

在这项工作中，我们开发了HEAL，这是一个利用Redd it上的≈100万个与苦恼有关的对话而形成的知识图。它 在不同类型的压力源、说话人的期望、反应和反馈类 型之间形成联系，同时将每个节点与41种情感状态之 一联系起来。对知识图谱进行的统计和视觉分析验证 了HEAL中有用的反应策略的存在，这些策略可以减轻

人们遭受痛苦的负面情绪状态 。我们进一步评估了HEA L的效用

提示

*如果我知道怎么做，我就会完全*

*做到这一点。也许明天我应该跳到火*

*车前面去:(*

(Xie and Pu 2021) *你打算怎么做？t*

搅拌器

*我相信你能做到！只要继续练习*

*，总有一天你会成功的。* *t*

HEAL-排名

提示

*不要这样做。*

*你能帮我写一份普通的遗书吗？*

(Xie and Pu 2021) 我 *不知道你在说什么。t*

搅拌器

*我很遗憾。我希望你能找到人帮*

HEAL-等级

*你写。* *t*

*我很抱歉，* *这对你来说是如此艰难。*

表6：显示对痛苦提示作出反应的毒性的例子 (以上标

*†*表示)。

通过开发一个简单的re-

trieval模型，在对痛苦提示做出反应时， HEAL可以提

供更多的信息。根据对HEAL检索的反应和由Xie和Pu

以及Blender的最先进的移情转换模型产生的反应所计

算的自动和人工评估分数，很明显HEAL能够产生更

多的、在情感上更合适的反应。对这些模型产生的反

应的人工检查进一步揭示了使用纯粹的生成模型来解

决困境的危险，以及HEAL如何通过战略性地识别与

特定提示相关的特定压力源来避免不恰当的反应。

在评估HEAL对窘迫提示的反应时，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

话人的表达和反馈结合起来，可以开发出更复杂的重

新排名方法。来自知识图谱的信息可用于增强神经反

应生成模型，并为这些模型引入更多的可控性和可预

测性，从而提高可靠性。

HEAL有一些相关的限制。它只限于识别≈4K的压力

源。但是，新的提示可能涉及许多其他的压力源，而

这些压力源在知识图谱中并没有涵盖。然而，还有空

间用从网络上刮来的更多数据来增强知识图谱，这将

使它能够处理更广泛的压力源和期望。

道德声明

虽然这项工作中使用的数据是公开的，但不应该削弱

它包含高度敏感的信息。因此，根据Benton等人(201

7)关于在健康研究中使用社交媒体数据的指导方针

，在本文中， 我们只引用了数据集的转述摘录。由于

HEAL是通过将长回复分割成独立的句子来构建的，

因此公开它将无法通过网络搜索来恢复用户名和帖子

文本。只有与压力源相关的痛苦叙述的嵌入才会被分

享，以使基于检索的模型的开发成为可能。带有匿名

用户名的Reddit对话可以在特殊条件下应要求与其他A

CA-DEMIC研究者分享