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knowledge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Reddit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HEAL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

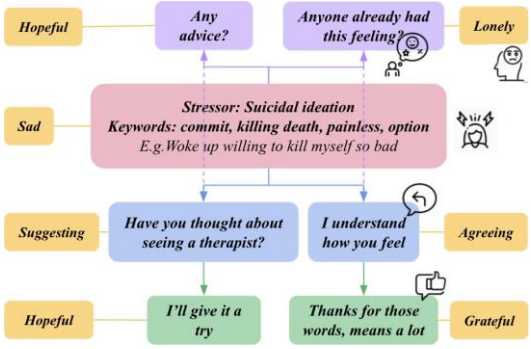


图1：*HEAL*的部分图示。红色、白色、蓝色、绿色和黄色节点分别代表压力源、说话人的期望、反应和反馈类型以及相关的情感状态。

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（Vaswani等人，2017）和预训练的语言模型，如BERT（Devlin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

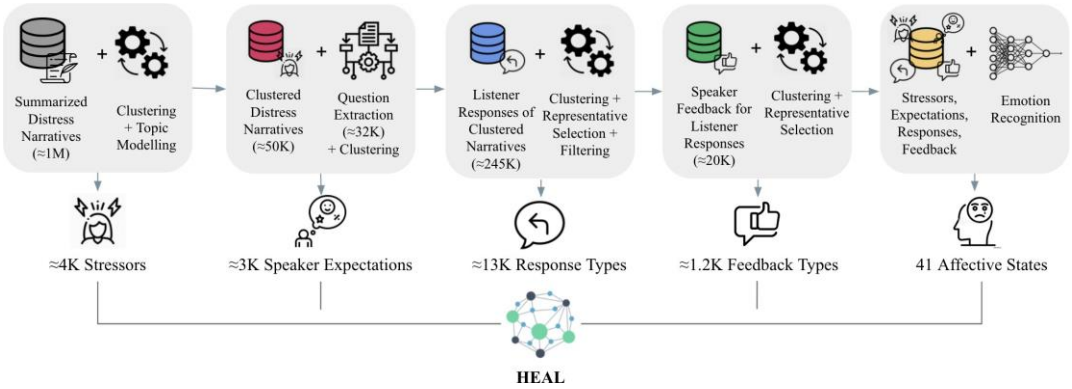


图2：开发知识图谱HEAL的分步过程。

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**HEAL**（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subreddits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。[2]YAGO（Fabian等人，2007）、Freebase（Bollacker等人，2008）和Wikidata（Vrandec˘ic´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ConceptNet（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EmpatheticDialogues（Rashkin等人，2019）、EmotionLines（Hsu等人，2018）和EmoContext（Chatterjee等人，2019），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Reddit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*；*depression*；*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Author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

[2]en.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 distress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Silhouette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	、*杀戮*、*死亡*、*无痛*、*选择*
焦虑攻击焦虑	、*焦虑*、*攻击*、*社会*、*攻击* 体重训练
	、*体重*、*吃*、*减*、*肥*
孤独	*孤独, 周围，连接，孤立，社会*
失败的大学学习	*，大学, 班级，学期，失败* 酗酒者
	喝酒, 饮料, 酒精, 醉酒, 清醒
美国选举特朗普	*, 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（Skerry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。

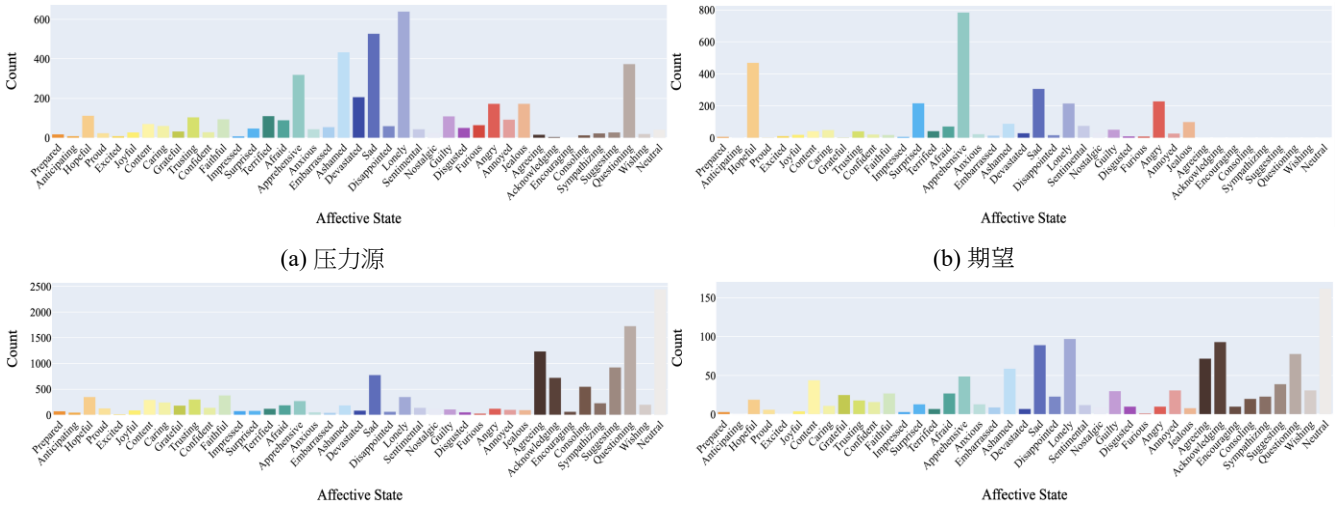
HEAL: 统计分析

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26,628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。

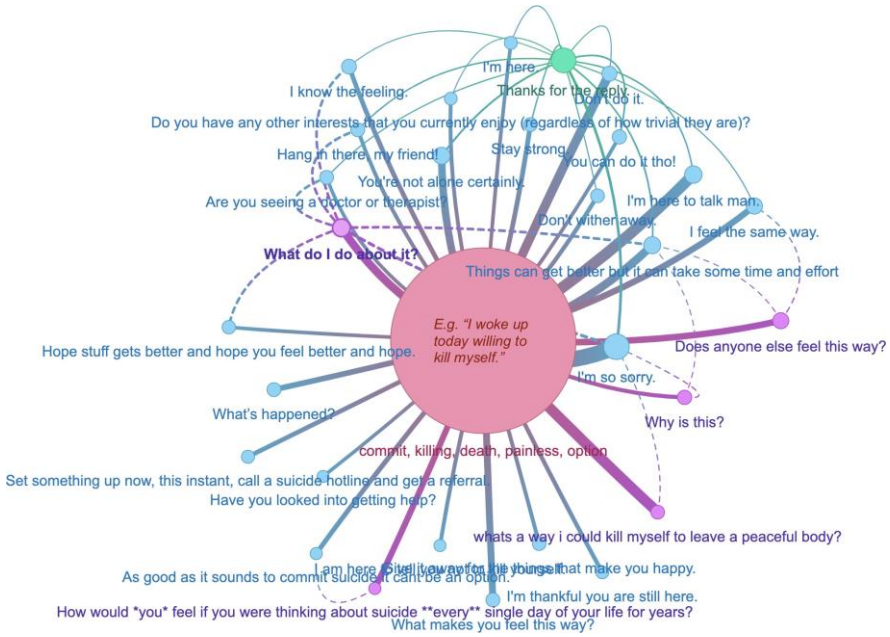


图4：通过vis.js对HEAL的部分内容进行可视化。压力源、期望、反应和反馈类型分别用红色、紫色、蓝色和绿色表示。为了避免杂乱，只有具有重要边缘权重的连接被可视化。

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)	0.1159	0.3364	0.4818	0.5815	0.0066	0.0014	0.0277	0.0475	**0.6921**
	搅拌器	0.0686	0.2226	0.3206	0.3877	**0.0707**	**0.0150**	**0.0469**	**0.0661**	0.6047
	痊愈等级的	**0.1704**	**0.4540**	**0.6003**	**0.7100**	0.0033	0.0007	0.0252	0.0332	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Blender（generative）（Roller等人，2021）。前者是一个基于RoBERTa（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Seq2Seq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Distinct-
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。†*
搅拌器	*我很抱歉听到这个消息.* *你有什么爱好，你喜欢做什么？‡*
HEAL-等级	*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。†*
搅拌器	*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？‡*
HEAL-ranked	*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Reddit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HEAL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？†*
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。†
HEAL-排名	*不要这样做。*
提示	*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。†*
搅拌器	
	我很遗憾。我希望你能找到人帮你写。†
HEAL-等级	*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（2017）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ACA-DEMIC研究者分享。

参考文献

Almeida, D. M.; Wethington, E.; and Kessler, R. C. 2002。压力事件的每日清单。一种基于访谈的方法来测量日常压力源。*评估*，9（1）。41-55.

Althoff, T.; Clark, K.; and Leskovec, J. 2016。咨询对话的大规模分析。自然语言处理在心理健康中的应用。*Transac- tions of the Association for Computational Linguistics*, 4: 463-476.

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007。Dbpedia:一个开放数据网络的核心。In *The semantic web*, 722-735.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020。The Pushshift Reddit Dataset.*国际AAAI网络和社会媒体会议的论文集*，14(1)。830-839.

Benton, A.; Coppersmith, G.; and Dredze, M. 2017。社交媒体健康研究的道德研究协议。在*第一届ACL自然语言处理中的伦理问题研讨会论文集中*，94-102。

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J.
2008.Freebase：一个合作创建的用于构建人类知识的图形数据库。In *Proceedings of the 2008 ACM SIGMOD international conference on Management- ment of data*, 1247-1250.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; et al. 2020。语言模型是少数人的学习者。在《*神经信息处理系统的进展*》中，第33卷，1877-1901。

Chatterjee, A.; Gupta, U.; Chinnakotla, M. K.; Srikanth, R.; Galley, M.; and Agrawal, P. 2019。利用深度学习和大数据理解文本中的情感。*Co mputers in Hu-man Behavior*, 93: 309-317.

d'Avila Garcez, A.; and Lamb, L. C. 2020.神经符号人工智能：第三次浪潮。
arXiv:2012.05876。

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019.BERT：用于语言理解的深度双向变换器的预训练。In *Proceedings of the 2019 Con- ference of the North American Chapter of the Association for Computational Linguistics:Human Language Technolo- gies, Volume 1 (Long and Short Papers)*, 4171-4186.

Ekman, P. 1992.关于基本情感的论证。*Cognition & emotion*, 6(3-4):169-200.

Fabian, M.; Gjergji, K.; Gerhard, W.; et al. 2007。Yago:统一词网和维基百科的语义知识核心。In *16th International World Wide Web Confer- ence, WWW*, 697-706.

Han, S.; Bang, J.; Ryu, S.; and Lee, G. G. 2015.利用知识库为自然语言对话倾听者生成回应。In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Di-alogue*, 129-133.

Hsu, C.-C.; Chen, S.-Y.; Kuo, C.-C.; Huang, T.-H.; and Ku, L.-W.2018.EmotionLines:一个多方对话的情感语料库。在*第11届国际语言学会会议上。*

语言资源与评估全国会议（LREC 2018）。

李，D.
2016。泰。微软就种族主义聊天机器人的惨败发表道歉声明。*BBC新闻*。

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020。BART:用于自然语言生成、翻译和编译的去噪序列对序列预训练。在*计算语言学协会第58届年会的论文集中*，7871-7880。

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016。一个促进多样性的神经关系模型的目标函数。在*Compu- tational Linguistics协会北美分会的2016年会议上。人类语言技术*，110-119。

Lin, Z.; Madotto, A.; Shin, J.; Xu, P.; and Fung, P. 2019.MoEL：移情听众的混合体。In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing（EMNLP-IJCNLP）*, 121-132.

Lison, P.; Tiedemann, J.; Kouylekov, M.; et al. 2019。开放字幕2018。在大型、嘈杂的平行语料库中对句子排列进行统计重构。In *LREC 2018, Eleventh International Conference on Language Resources and Eval- uation*.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016。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。In *Pro- ceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2122-2132.

Liu, S.; Chen, H.; Ren, Z.; Feng, Y.; Liu, Q.; and Yin, D. 2018。神经对话生成的知识扩散。In *Proceedings of the 56th Annual Meeting of the Associa- tion for Computational Linguistics（第一卷：长篇论文）*，1489-1498。

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019a。Roberta:一个稳健优化的伯特预训练方法。*arXiv预印本arXiv:1907.11692*。

Liu, Z.; Niu, Z.-Y.; Wu, H.; and Wang, H. 2019b.Knowledge Aware Conversation Generation with Explainable Rea- soning over Augmented Graphs.In *Proceedings of the 2019 Conference on Empirical Methods in Natural Lan-guage Processing and the 9th International Joint Confer- ence on Natural Language Processing（EMNLP-IJCNLP）*, 1782-1792.

Majumder, N.; Hong, P.; Peng, S.; Lu, J.; Ghosal, D.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020。MIME: MIM-icking Emotions for Empathetic Response Generation.在*2020年自然语言处理中的经验方法会议上*，8968-8979。

Misuraca, M.; Spano, M.; and Balbi, S. 2019.BMS：用于文档聚类验证的改进的邓恩指数。

统计学通讯-理论与方法, 48(20): 5036-5049.

Murtagh, F.; and Legendre, P. 2014.Ward的分层ag环亚娱乐平台聚类方法：哪些算法实现了Ward的标准？*Journal of classification*, 31(3):274-295.

Peng, Y.-H.; Jang, J.; Bigham, J. P.; and Pavel, A. 2021.全说出来。改善非视觉演示可及性的反馈。在*2021年CHI计算系统中人的因素会议的论文集中*，1-12。

Plutchik, R. 1984.情感。一个一般的心理演化理论。*情感的方法*，1984：197-219。

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019。语言模型是无监督的多任务学习者。*OpenAI博客*，1（8）。9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020。用统一的文本到文本转换器探索转移学习的极限。*机器学习研究杂志*，21（140）。1-67.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019.实现移情的开放域对话模型。一个新的基准和数据集。In *Proceedings of the 57th Annual Meeting of the Association for Computational Lin-guistics*, 5370-5381.

Reimers, N.; and Gurevych, I. 2019.Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing（EMNLP-IJCNLP）*，3982-3992。

Ridner, S. H. 2004.心理困扰：概念分析。*高级护理学杂志*, 45(5):536-545.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.-L.; and Weston, J. 2021。构建开放域聊天机器人的配方。在*计算语言学协会欧洲分会第16次会议的论文集中。主卷*，300-325。

Rousseeuw, P. J. 1987.Silhouettes:对聚类分析的解释和验证的图形帮助。*计算和应用数学杂志*, 20: 53-65.

Rus, V.; and Lintean, M. 2012.使用词与词之间的相似度指标对自然语言学生输入的贪婪和优化评估的比较。在*第七届使用NLP构建教育应用研讨会上*，157-162。

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019。Atomic:一个用于if-then reasoning的机器常识图集。在*AAAI人工智能会议论文集*，第33卷，3027-3035。

Skerry, A. E.; and Saxe, R. 2015.情感的神经表征是围绕抽象的事件特征组织的。*Cur-rent biology*，25（15）。1945-1954.

Speer, R.; Chin, J.; and Havasi, C. 2017。Conceptnet 5.5：一个开放的多语言通用知识图。在*AAAI人工智能会议的Pro- ceedings中，* 第31卷。

Tas, O.; and Kiyani, F. 2007.自动文本汇总的调查.*PressAcademia Procedia*, 5(1):205-213.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017。注意力是你所需要的一切。In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Process- ing Systems*, volume 30.

Vrandec˘ic´, D.; and Kro¨tzsch, M. 2014.维基数据：一个自由协作的知识库。*ACM通讯》*，57（10）。78-85.

Welivita, A.; and Pu, P. 2020。人类社会对话中移情反应意图的分类法。In *Pro- ceedings of the 28th International Conference on Computa- tional Linguistics*, 4886-4899.

Xie, Y.; and Pu, P. 2021.用大规模的对话数据集生成移情反应。在*第25届计算自然语言学习会议论文集（即将出版）*。

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; and Le, Q. V. 2019.XLNet:用于语言理解的广义自律性预训练。In *Advances in Neural Information Processing Systems*, volume 32.

Young, T.; Cambria, E.; Chaturvedi, I.; Zhou, H.; Biswas, S.; and Huang, M. 2018。用常识性知识增强端到端对话系统。在*AAAI人工智能会议的论文集中，* 第32卷。

Zhang, H.; Liu, X.; Pan, H.; Song, Y.; and Leung, C. 2020.ASER:
一个大规模的偶发事件知识图谱.*2020年网络会议论文集*, 201-211.

Zhang, J.; and Danescu-Niculescu-Mizil, C. 2020。平衡咨询对话中的目标。向前推进或向后看。I n *Proceedings of the 58th Annual Meeting of the Association for Computational Lin-guistics*, 5276-5289.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X.2018.使用图注意的共知知识感知对话生成。In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 4623-4629.

Zhu, W.; Mo, K.; Zhang, Y.; Zhu, Z.; Peng, X.; and Yang, Q. 2017.灵活的端到端对话系统的知识基础对话。*arXiv预印本arXiv:1709.04264*。