HEAL:一个痛苦管理对话的知识图谱

Anuradha Welivita, Pearl Pu

计算机与通信科学学院

洛桑综合技术学院(Ecole’Polytechnique 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造成心理负担，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(HEAL)，这是一个知识图谱，基于 Reddit 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 104K 连接。每个节点都与 41 种情感状态中的一种相关联。对 HEAL 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，HEAL 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(Ridner 2004)。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

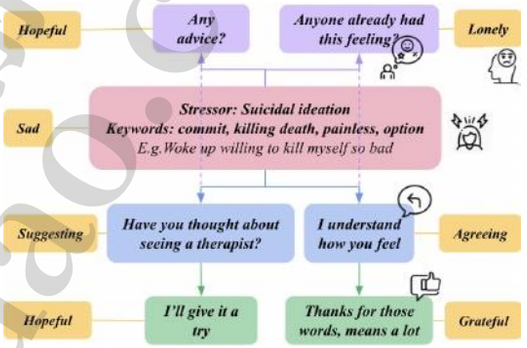


图 1:HEAL 部分示意图。红色、紫色、蓝色、绿色和黄色节点分别代表压力源、说话人期望、反应和反馈类型以及相关的情感状态。

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(Rashkin et al. 2019;Lin et al. 2019;Majumder 等 2020;谢、普 2021)。

随着复杂神经网络架构的发展，如 transformer (Vaswani et al. 2017)和预训练语言模型，如 BERT (Devlin et al. 2017)。

(2019)、RoBERTa(刘等。2019a)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 'Avila Garcez 和 Lamb 2020)。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 Twitter 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(Lee 2016)。

因此，人们对利用知识的兴趣越来越大(Zhu et al. 2017;刘等，2018;Han et al. 2015)和常识推理(Zhou et al. 2018;Young et al. 2018)

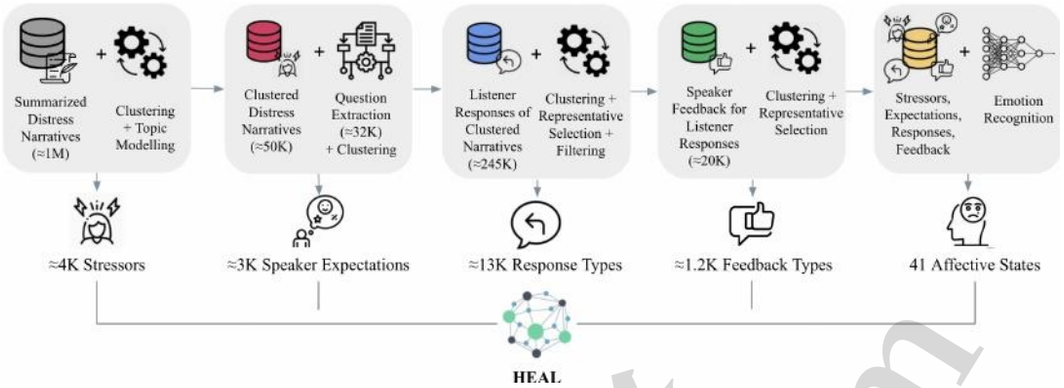


图 2:逐步开发知识图谱 HEAL 的过程。

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 Concept- Net (Speer、Chin 和 Havasi 2017)和 ATOMIC (Sap et al. 2019)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 HEAL(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。

HEAL 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 Reddit 上发帖来开始对话)，听众是这些帖子的评论者。图 1 显示了 HEAL 中典型压力源的说明。治愈(HEAL)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 Reddit 策划的 100 多万次痛苦对话中开发 HEAL。这导致识别出≈4K 的压力源，≈3K 的说话人期望，≈13K 的反应类型，≈1.2K 的反馈类型，以及相关的情感状态。最终的图由 22,037 个节点和 104,004 个不同类型节点之间的连接组成。

通过对 HEAL 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 HEAL 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 Xie 和 Pu 开发(2021);以及 Blender (Roller et al. 2021)。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，HEAL 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、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 HEAL 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。

在理解自然语言输入方面。这得益于最近 DBPedia (Auer et al. 2007)和谷歌知识图谱等链接开放数据的出现。[2] YAGO (Fabian et al. 2007)、Freebase (Bollacker et al. 2008)和 Wikidata (Vrande ˇ ci´ c 和 Kr ¨ otzsch 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 concept - ceptNet (Speer、Chin 和 Havasi 2017)、ATOMIC (Sap 等 2019 年)和 ASER(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 factoid 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(Zhou et al. 2018)，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。HEAL 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- patheticDialogues (Rashkin et al. 2019)、EmotionLines (Hsu et al. 2018)和 EmoContext (Chatterjee et al. 2019)，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, Clark, and Leskovec 2016;Zhang 和 Danescu-Niculescu-Mizil 2020)由于伦理原因无法直接获得。因此，我们从 Reddit 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 Reddit 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 Pushshift API (Baumgartner et al. 2020)从精心挑选的 8 个版块收集和处理对话线索 :mentalhealthsupport;*offmychest; 悲伤的;suicidewatch;anxietyhelp;抑郁症;沮丧的;*以及抑郁症帮助，这在 Reddit 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 1,275,486 个二元对话，包含 3,396,476 个对话轮取结构(平均每个对话 2.66 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

Reddit 策划的痛苦叙事通常很长(平均每轮 84.89 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- bert (Reimers and Gurevych 2019)。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(Peng et al. 2021)。因此，我们选择了 5 种不同的摘要抽取方法:smmry 的自定义实现——Reddit TLDR 机器人(https:// smmry.com)背后的算法;以及四种不同的预训练模型- bart (Lewis et al. 2020)、GPT-2 (Radford et al. 2019)、XLNET (Yang et al. 2019)和 T5 (Raffel et al. 2020)用于建模内容重要性。我们对上述方法在 100 个 Reddit 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 SMMRY 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 token)。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 Reddit 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”(Murtagh and Legendre 2014)。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-BERT(Reimers and Gurevych 2019)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 0.95，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(Rousseeuw 1987)、Dunn 指数(Misuraca、Spano 和 Balbi 2019)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/Google 知识图谱

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，pandemic，shambolic，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-IDF 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 NLTK 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、0.75 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 1 208 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 Reddit 的响应(例如，请联系 subreddit 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 Welivita 和 Pu 提出的基于 BERT 转换器的分类器(2020)在 EmpatheticDialogues 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(Ekman 1992;Plutchik 1984)到从情境中衍生出的更大的微妙情绪集(Skerry 和 Saxe 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。

HEAL:统计分析

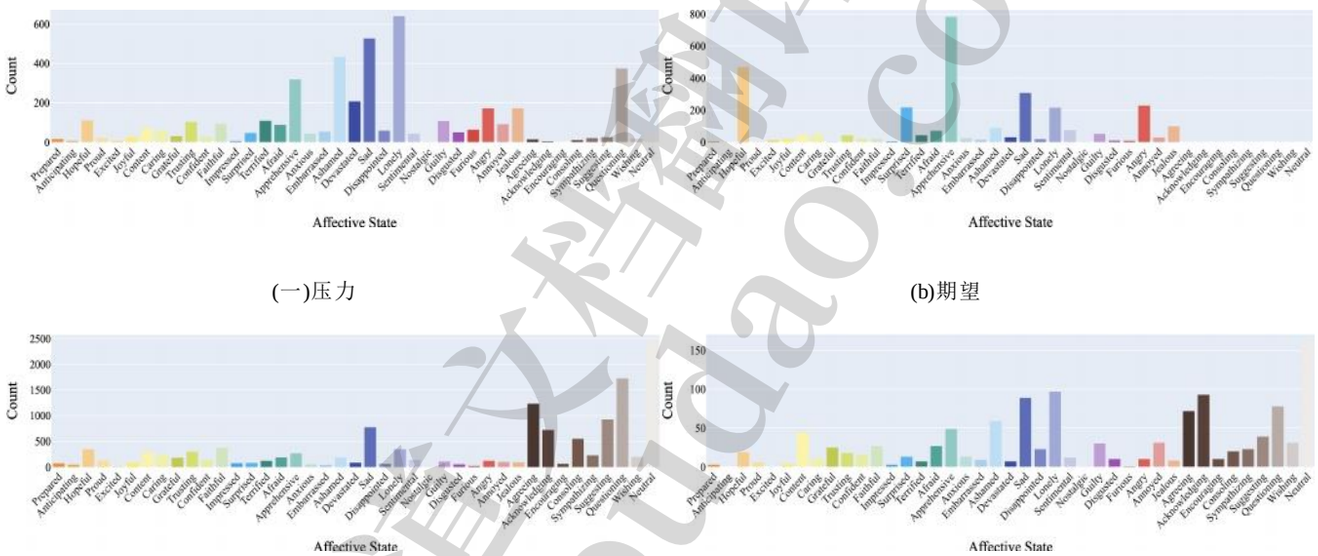
我们跟踪压力叙述的压力源标识符，从中提取出每个期望和反应，并能够在压力源和期望和反应簇之间形成联系。我们还跟踪每个反馈的对话标识符，这有助于在反馈集群和期望和响应集群之间建立联系。最终形成的知识图谱 HEAL 由 22,037 个节点和 104,004 条节点间连接构成。压力源与期望之间有 9,801 个连接，压力源与反应之间有 56,654 个连接，反应与反馈之间有 10,921 个连接，期望与反应之间有 26,628 个连接。此外，每个节点与一种情感状态相关联，形成 22,037 个连接。

图 3 显示了与压力源、期望、反应和反馈类型相关的情感状态的分布。根据统计，73.60%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(25.70%)、悲伤(10.07%)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(9.22%)、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

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

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

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 vis.js (visjs.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 HEAL 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 Reddit 对话进行测试。为了从“HEAL”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 123651 个对话提示中，60.7%与 the 显示了 0.75 或以上的相似度

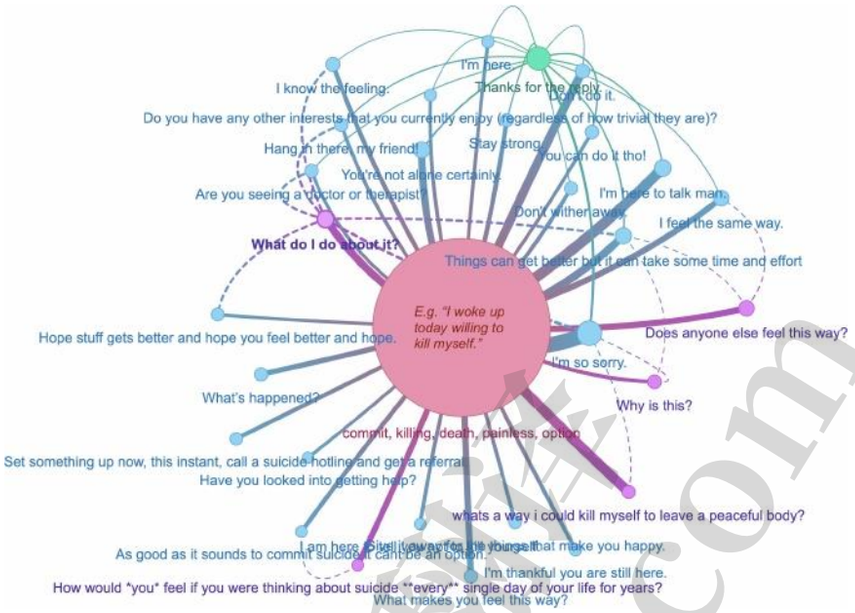


图 4:vis.js 对部分 HEAL 的可视化。压力源、期望、反应和反馈类型分别用红色、紫色、蓝色和绿色表示。只有具有显著边权值的连接才被可视化，以避免混乱。

数据集	模型	D1	D2	D3	D4	BLEU1	BLEU2	流星	胭脂	GM
Reddit	(Xie 和 Pu2021)	0.1159	0.3364	0.4818	0.5815	0.0066	0.0014	0.0277	0.0475	**0.6921**
	搅拌机	0.0686	0.2226	0.3206	0.3877	**0.0707**	**0.0150**	**0.0469**	**0.0661**	0.6047
	Heal-ranked	**0.1704**	**0.4540**	**0.6003**	**0.7100**	0.0033	0.0007	0.0252	0.0332	0.6599

表 3:Reddit 中响应遇险提示任务获得的自动评价结果 D1、D2、D3 和 D4 代表 Distinct-ngram 度量(Li et al. 2016)，GM 代表贪婪匹配分数(Rus 和 Lintean 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 healrank。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 healrank 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 Xie 和 Pu(2021)开发，另一种由 Blender (gen- erative)开发(Roller et al. 2021)。前者是基于 RoBERTa 的多轮情感参与对话生成模型(Liu et al. 2019a)。它对 open 字幕中的约 1M 对话进行预训练(Lison 等人 2019 年)，并对同理心对话进行微调(Rashkin 等人 2019 年)。后者是一个标准的基于 Seq2Seq transformer 的共情开放域聊天机器人。它在包含约 1.5 亿个评论的 Reddit 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

提示。我们可以观察到，在用于衡量应答多样性的 Distinct-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，Blender 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. 2016)，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 Amazon Mechanical Turk (AMT)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 Reddit 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	*我哥哥两年前去世了*
	到现在我还是很伤心。还是那么心痛
	多
(谢和普 2021)	*听到这个消息我很难过。*†
搅拌机	*听到这个消息我很难过。你有*
	你有什么爱好吗?‡
治疗等级	*我对你的损失感到遗*
	憾。
提示	*我似乎不能感到快乐，但我不是 de-*
	按下
(谢和普 2021)	*听到这个消息我很难*
	过。†
搅拌机	*听到这个消息我很难过。你有吗?*
治疗等级	*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 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)。

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 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 0.23 的评价者之间的一致(弗莱斯 kappa)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。

案例研究

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)。

讨论和结论

在这项工作中，我们开发了治疗，这是一个基于 Reddit 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

提示	*如果我知道怎么做，我一定会*
	做到了。也许明天我应该
	跳到火车前面:(
(谢和普 2021)	*你打算怎么做?*†
搅拌机	*我相信你能做到!坚持练习就行了*
	你总有一天会成功的!†
治疗等级	*不要这样*
	做。
提示	*你能帮我写一个通用的吗*
	遗书?
(谢和普 2021)	*我不太确定你在说什么。*†
搅拌机	*听到这个消息我很难过。我希望你能找到*
治疗等级	*有人帮你写。*†
	很抱歉让你这么难受。

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 HEAL 检索到的响应进行的自动和人工评估评分，以及 Xie、Pu 和 Blender 的最先进的共情对话模型产生的评分，显然 HEAL 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 HEAL 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 HEAL 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

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 Benton 等人(2017)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 HEAL 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 web 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 Reddit 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.;Wethington 大肠;和 Kessler, R. C. 2002。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):41-55*。

Althoff t;克拉克,k;和 Leskovec, J. 2016。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，4:463-476*。

奥氏小体,美国;商业,c;Kobilarov g;莱曼,j .;Cyganiak r;和艾夫斯，Z. 2007。Dbpedia:开放数据网络的核心。在语义网中，722-735。

费利克斯,j .;Zannettou,美国;基冈,b;乡绅,m;和布莱克本，J. 2020。Pushshift Reddit 数据集。*国际 AAAI 网络与社交媒体会议论文集，14(1):830-839*。

Benton;铜匠,g;和 Dredze, M. 2017。社交媒体健康研究的伦理研究协议。第一届 ACL 自然语言处理伦理研讨会论文集，94-102。

Bollacker k;埃文斯,c;介绍 p;,他是 t;和 Taylor, J. 2008。Freebase:一个用于结构化人类知识的协作创建的图形数据库。2008 年 ACM SIGMOD 数据管理国际会议论文集，1247-1250。

棕色、t;曼,b;莱德:;印度,m;等。2020。语言模型是少样本学习者。《神经信息处理系统的进展》，卷 33,1877 - 1901。

查特吉，a;古普塔,美国;钦纳科特拉，m.k.;Srikanth 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，93:309-317*。

d 'Avila Garcez, A.;兰姆，l.c. 2020。神经符号 AI:第三波。arXiv: 2012.05876。

Devlin, j .;Chang,硕士。李,k;和图塔诺瓦(Toutanova, K. 2019)。BERT:用于语言理解的深度双向 transformer 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)，4171-4186。

Ekman, 1992 年第 1 期。基本情绪的一种。*认知与情感，6(3-4):169-200*。

费边,m;Gjergji k;格,w;等。2007。Yago:统一 wordnet 和 wikipedia 的语义知识核心。在第 16 届国际万维网会议上，WWW，697-706。

汉族,美国;爆炸,j .;Ryu,美国;和 Lee G. G. 2015。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 Ku l.w。2018. 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;Goyal:;Ghazvininejad m;穆罕默德，a;Levy, o .;Stoyanov 诉;和 L. 2020 年的 Zettlemoyer。BART: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，7871-7880。

李,j .;厨房,m;Brockett c;高,j .;和多兰，B. 2016。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。MoEL:感同身受的听众。在 2019 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(EMNLP-IJCNLP)论文集中，121-132。

Lison p;蒂,j .;Kouylekov m;等。2019。Open 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;Serban i;Noseworthy m;Charlin l;和 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。《2016 年自然语言处理经验方法会议论文集》，2122-2132。

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，1489-1498。

刘,y;奥特,m;Goyal:;杜,j .;乔希,m;陈,d;Levy, o .;刘易斯,m;Zettlemoyer l;和斯托亚诺夫，2019a。Roberta:一种鲁棒优化的 bert 预训练方法。*arXiv 预印本 arXiv:1907.11692*。

刘,z;妞妞,Z.-Y。;吴,h;和王，H. 2019b。增广图上基于可解释推理的知识感知对话生成。在 2019 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(EMNLP-IJCNLP)中，1782-1792。

Majumder:;在香港,p;彭,美国;陆,j .;Ghosal d;Gel- bukh, A.;Mihalcea r;和 Poria, S. 2020。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》(EMNLP)，8968-8979。

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。

统计理论与方法，48(20):5036-5049。

Murtagh f;和勒让德(Legendre)， 2014 年 P.。Ward’s hierarchical agglomerative clustering method:哪些算法实现了 Ward’s criterion?*分类学报，31(3):274-295*。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. 1984。《情绪:一般的心理进化理论》。*《情绪的途径》，1984:197-219*。

雷德福;吴,j .;孩子,r;烹调的菜肴,d;Amodei d;和 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:;罗伯茨， A.;李,k;纳,美国;Matena m;周,y;李,w;和 Liu, P. J. 2020。用统一的 Text-to- Text Transformer 探索迁移学习的局限性。*机器学习研究学报，21(140):1-67*。

Rashkin h;史密斯，e.m.;李,m;和 Boureau, 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，5370-5381。

雷蒙:;和古雷维奇，I. 2019。Sentence- bert:使用孪生 bert 网络的句子嵌入。2019 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，3982-3992。

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;Dinan 大肠;Goyal:;朱，D.;威廉森;刘,y;徐,j .;Ott, m;史密斯;布里奥;韦斯顿(Weston)， 2021 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，300-325。

卢梭，P. J. 1987。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，20:53-65*。

俄文,诉;和 Lintean, M. 2012。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，157-162。

萨普，m;勒 • 布拉斯，r;Allaway 大肠;Bhagavatula c;劳里:;Rashkin h;屋顶,b;史密斯，n.a.;和 Choi Y. 2019。原子:机器常识图集，用于 if-then 推理。《AAAI 人工智能会议论文集》，第 33 卷，3027-3035。

Skerry, A. E.;和萨克斯，R. 2015。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):1945-1954*。

斯皮尔,r;下巴,j .;和哈瓦西，C. 2017。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*。

Vaswani, a;Shazeer:;Parmar:;Uszkoreit, j .;琼斯， L.;戈麦斯，A. N.;路易斯安那州凯泽;以及波洛苏欣(Polosukhin, I. 2017)。注意力就是你所需要的。在盖恩， 我;Luxburg, U. V.;Bengio,美国;瓦拉赫,h;费格斯 r;Vishwanathan,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¨otzsch, M. 2014。Wikidata:一个免费的协作知识库。*ACM通讯，57(10):78-85*。

Welivita, a;和 Pu, P. 2020。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，4886-4899。

谢,y;和濮存昕，P. 2021。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;Carbonell j .;萨拉赫丁诺夫，r • r;以及勒，Q. V. 2019。XLNet: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;Biswas 美国;黄，M. 2018。用常识知识增强端到端对话系统。AAAI 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。ASER:大规模的不测知识图谱。*2020 年网络会议论文集，201-211*。

张,j .;和 Danescu-Niculescu-Mizil, C. 2020。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，5276-5289。

周,h;年轻、t;黄 M.;赵,h;徐,j .;和朱 X. 2018。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》，IJCAI-18，4623-4629。

朱,w;密苏里州,k;张,y;朱,z;彭,x;和杨 Q. 2017。面向知识基础对话的灵活端到端对话系统。*arXiv 预印本 arXiv:1709.04264*。