**HEAL（治愈）:一个痛苦管理对话的知识图谱**

**摘要：**

现代社会的需求日益给我们造成心理负担，同时给我们的心理健康带来了负面的影响。因此，具有共情响应和痛苦管理能力的神经对话代理最近越来越受欢迎。然而，现有的端到端共情对话代理经常产生普通的和重复的共情语句，比如“听到这个我很遗憾”，这不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认为是端到端模型的更高效和更安全的解决方案。然而，这种资源在情感困扰的情况下是有局限的。为了解决这个问题，我们介绍了治愈(HEAL)，这是一个基于从Reddit整理的100万个痛苦故事及其相应的安慰反应所开发的知识图谱。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效反应模式。自动化评估和人类评估实验都表明，在与基线相比下，HEAL的反应更加多样化、更有同理心、更可靠。

**介绍：**

现代世界的需求日益给我们造成心理负担，同时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受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，包括暂时的或者永久性的(Ridner2004)。这些压力源包括与亲人分离、人际冲突、某些心理健康症状(如抑郁症)、工作表现不佳，以及失眠等问题。阿尔梅达等人(2002年)的一项研究，通过每日电话访谈，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，结果显示这些成年人至少经历过一种日常压力源。

在40%的研究中发现，人们通常倾向于在日常对话中分享自己的经历。因此，嵌入具有适当共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症状已经备受人们的关注(Rashkinetal.2019;Linetal.2019;Majumder等2020;谢、普2021)。随着复杂神经网络架构的发展，如transformer(Vaswanietal.2017)和预训练语言模型，如BERT(Devlinetal.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可靠性和故障安全性降低(d'AvilaGarcez和Lamb2020)。当用户处于对错误信息和不恰当评论处于敏感的痛苦境地时，这尤其是一个问题。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(Lee2016)。

因此，人们对利用知识(Zhuetal.2017;刘等，2018;Hanetal.2015)和常识推理(Zhouetal.2018;Youngetal.2018)在基于图的表示上生成适当的、有信息量的对话响应的兴趣越来越大。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(Liuetal.2019b)。

虽然存在诸如Concept-Net(Speer、Chin和Havasi2017)和ATOMIC(Sapetal.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的对话生成。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中策划的相应反应线索开发。

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;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显示了一个HEAL中典型压力源的例子。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，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，即对某个给定的痛苦情况产生同情反应。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(Rolleretal.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)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。

**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广泛注意，这得益于最近DBPedia(Aueretal.2007)和谷歌知识图谱等链接开放数据的出现。YAGO(Fabianetal.2007)、Freebase(Bollackeretal.2008)和Wikidata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Havasi2017)、ATOMIC(Sap等2019年)和ASER(Zhang等2020年)，侧重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事实和常识知识开发了开放域会话代理，这些会话代理，产生了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(Zhouetal.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。它们不具备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-反馈元组来识别有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(Rashkinetal.2019)、EmotionLines(Hsuetal.2018)和EmoContext(Chatterjeeetal.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,andLeskovec2016;Zhang和Danescu-Niculescu-Mizil2020)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，其中包含讨论真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，而且同龄人们会积极参与这些平台，以支持其他遭受精神困扰的人。

我们使用PushshiftAPI(Baumgartneretal.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，这些是Reddit用户中流行的发泄痛苦的方式。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的转折结构，并对这些对话进行了严格的数据清洗，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%的对话来派生知识图谱，并保留10%的对话，用于验证和测试下游任务。

**归纳总结：**

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Gurevych2019)。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的摘要技术来解决这个问题(Tas和Kiyani2007)。其中，抽象式摘要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Redditdis-tress叙述的样本进行了手动评分，将上述方法生成的摘要评为好和坏(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%的对话回合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，特别是在应用于大规模数据集时，因此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“聚合聚类”(MurtaghandLegendre2014)。它递归地合并最小限度增加给定联系距离的集群对。我们利用句子-BERT(ReimersandGurevych2019)生成的成对嵌入之间的余弦相似性计算了联系距离，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聚合聚类而不是其他聚类方法的原因。

**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，以困境叙述进行聚类。虽然为每个阈值计算了各种聚类质量指标，如轮廓系数(Rousseeuw1987)、Dunn指数(Misuraca、Spano和Balbi2019)和平均点到正弦距离，以选择最佳的相似性阈值，但对每个阈值处10个聚类的子集都进行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最有效

(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，在较高阈值（如0.95和0.9等）下识别的压力源过于具体，而在低于0.8以下识别的压力源过于模糊(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，发现了一些可清晰区分的压力源，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"？"的句子来提取分组后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NLTKli-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32832个期望，245707个回应和20213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3050个、13416个和1208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，以删除Reddit特有的回应(例如：请联系subredddit的版主)、由机器人产生的相应(例如：这个动作是自动执行的)和不完整的回应(例如：嘿，哇)。最终聚类结果相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。附录中包括频繁的期望、反应和反馈类型的例子。

**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BERT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-man1992；Plutchik1984)到从背景情况中得出的较大的微妙情感集(Skerry和Saxe2015)，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组的每个文本进行分类，并将该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。最终的形成的知识图谱HEAL，由22,037个节点和104,004个节点之间的连接组成。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，在HEAL中，有一些有用的反应类型，可以缓和遭受痛苦的人的负面情绪状态。

**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生成的知识图谱。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？表示同意的回答，如“我有同样的感觉，我知道这种感觉;；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，这些回答大多与说话者的积极反馈有关，如感谢他的回答，对听者表示感谢，同时也说明这是一个好的回应。

**评估HEAL在应对痛苦提示中的效用：**

我们评估了HEAL在给定的痛苦对话提示中检索适当的共情反应的能力，并将其性能与现有的最先进的共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。

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，这作为未来工作的一部分。

我们将HEAL-ranked检索的反应与两个最先进的共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。我们可以观察到，在用于衡量应答多样性的Distinct-N指标方面，HEAL排名超过了其他方法(Li等，2016)。这表明，与现有的神经反应生成模型相比，HEAL产生了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反应实例来进一步证明这点。可以看到，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HEAL排名相对于其他自动指标BLEU、METEOR和ROUGE表现较差。然而，众所周知，这些指标与人类判断的相关性很差(Liuetal.2016)，当与人共评估实验的结果进行比较时，可以很好地看到这一点，这将在下一节讨论。

**人工评价：**

我们设计了一个人工评价实验，从AmazonMechanicalTurk(AMT)招募群众工作者来评估三个模型产生的反应的共情适当性。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。工人们被指示拖拽并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。三名工作者允许在一个任务上工作，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，至少2/3的工人之间取得了一致，并产生了0.23的评价者之间的一致(Fleiss'kappa)分数，表明公平的一致。

根据结果，可以看到HEAL排名产生了最多的良好反应，在工情适宜性方面明显优于Xie和Pu的模型和Blender

**案例研究：**

在表6中，我们突出了几个精心挑选的例子，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上约100万个与痛苦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可视化分析验证了HEAL中存在有用的反应策略，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，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检索的响应以及由Xie和Pu以及Blender的最先进的共情对话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解释性，从而提高可靠性。

HEAL有一些相关的限制。它只限于识别约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，还有空间从网络上抓取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是通过将长回复拆分为单个句子来构建的，因此将其公开也无法通过使用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。