# 迈向大规模可解释的知识图推理
# 对话系统

Yi-Lin Tuan1 , Sajjad Beygi2 , Maryam Fazel-Zarandi2
Qiaozi Gao2 , 亚历山德拉·塞沃内2 , 王威廉杨1
[1] 加州大学圣塔芭芭拉分校 {ytuan, william}
@cs.ucsb.edu
[2] 亚马逊 Alexa 人工智能
{turn, fazelzar, qzgao, cervon}@amazon.com

抽象的

用户今天与语音助手交互
需要用非常具体的方式表达他们的要求
方式来引起适当的反应。这
限制了用户体验,部分原因是
缺乏对话推理能力
平台和手工制定的规则需要大量的劳动力。一种
可能的方法是
改善用户体验并减轻设计师的人工工作是建立一
个端到端的对话系统,可以在感知用户话语的同
时进行自我推理。在这个

工作中,我们提出了一种将知识推理能力纳入其
中的新方法
以更具可扩展性和通用性的方式构建对话系统。
我们提出的方法允许单个变压器模型直接

在大规模知识图上行走以生成响应。据我们所知,
这是第一部使用变形金刚的作品

模型通过推理产生响应
可微的知识图。我们研究了所提议的推理能力

面向任务和特定领域的闲聊对话的方法。实验结
果
表明该方法可以有效地将知识图谱整合到

具有完全可解释推理路径的对话系统。

## 1 简介

如今,对话系统在客户服务和语音助理中无处不在。之一

这项技术的主要用途是支持人类完成可能需要的任务

访问和导航大型知识库
（例如,电影搜索）。对话系统架构通常由自然语言组成

理解（NLU)模块、对话管理（DM)模块和自然语言生成
（NLG)模块（Jurafsky 和 Martin， 2009；

威廉姆斯等人， 2016）。一、NLU组件
从用户中提取含义表示
DM 根据其生成的话语
通过意义推理来决定下一个系统动作
代表和与外部的沟通
如有需要,可申请。例如,DM
可以从外部知识中检索信息
图表（KG)来回答用户的查询
对话历史。这个过程需要DM
将 NLU 的输出转换为要发出的查询
到后端。鉴于这一步的难度,
这通常是领域相关的,DM 组件可能需要手工设计

规则。然而,此类规则通常不可扩展
到不同的应用程序。他们可能需要付出相当大的努力来涵
盖所有可能的案例/对话
流,导致设计新产品的成本高昂
应用程序。此外,在一些情况下,与此类助手交互的用户被迫
提出特定的查询,以完成他们的任务。

目标,这可能会破坏用户参与度。
为了缓解必须设计昂贵的手工规则和破坏用户体
验的问题,最近的工作探索了这种可能性

构建端到端对话系统（Wen 等人，
2017)和一体化响应生成模型
（Serban 等人， 2016）。其中,自图
是存储知识的主要结构之一，
最近的研究（Ghazvininejad 等， 2018； Zhou
等， 2018； Moon等人， 2019； Tuan 等人， 2019；
Yang 等人， 2020)提出了根据两者生成自然语言响应的方
法
对话历史和外部知识图。
尽管有这些创新和鼓舞人心的方法，
有一些缺点。例如,这些
方法要么不完全可解释,要么有限
到小规模的知识图谱。

在本文中,我们提出了一种新颖的对话可微知识图
模型（DiffKG）。这
DiffKG 是一个单变压器模型,直接
(1) 生成要执行的关系序列
对（Cohen 等人， 2019)提出的具体化 KG 表示进
行多跳推理,然后（2）
使用检索到的实体生成响应。到

据我们所知,这是第一次对话
可直接在大型KG上行走的模型
具有灵活性和可解释性。 DiffKG 允许
在知识图谱中拥有灵活的实体值,并以任意定义的方式处理新颖的实体值
代币数量。 DiffKG的推理路径
由预测的关系组成,从而允许
为了透明度。

我们进行了大量的实验来测试 DiffKG
基于 KG 的对话的表现。我们选择斯坦福多领域对话
（SMD） （Eric
等人， 2017)并提出了一个新的数据集,SMD Reasoning,来模拟需要多个
推理类型并选择 OpenDialKG (Moon
et al., 2019)来模拟需要大规模KG 推理而无需预处理的场景。我们
然后将 DiffKG 与 SMD 和 OpenDialKG 上最先进的模型以及其他模型进行比较
将 KG 扁平化为文本形式的基线
哪些变形金刚可以学习。根据经验,我们的
实验表明DiffKG可以有效地
在大型 KG 上进行训练,并通过 KG 中修改的三元组展示其鲁棒性。来自
从计算的角度来看,相比之下,DiffKG 带来的额外时间和内存占用相对较低
不使用任何 KG 信息的变压器模型。

总之,我们的贡献是:1)我们
提出DiffKG,一种可以有效、灵活地融合大规模KG的新方法；

2)我们证明DiffKG是一种模型不可知的方法,可以应用于不同的情况
模型架构； 3)我们证明DiffKG
是一种在推理时具有低附加延迟的可解释方法。我们的代码和处理过的
数据集发布在https://github 中。
com/Pascalson/DiffKG-Dialog.

## 2 相关工作

近年来,新方法层出不穷
提出尝试理解自然语言输入文本和搜索的端到端模型

信息。两个被广泛探索的任务是
问答（QA)和对话生成。

质量保证。多种 QA 方法（Weston 等人， 2015；尹等人， 2016;郝等人， 2017;拉杰普尔卡等人，2018;维尔加等人， 2020;艾森施洛斯等人， 2021）
已提出解决超出语言环境中明确规定的任务（Storks et al., 2019）。例如,基准（ Mihaylov 等人， 2018； Reddy 等人，2019；
科特等人， 2020； Lin et al., 2021)对于模型提取信息特别有用

从外部知识库来回答问题。尽管如此,这些研究大多采用

从 KG 检索信息作为答案
一个问题,在对话中我们必须
制定对多轮的信息丰富的回应
对话历史。

对话一代。最近的作品研究了扎根对话的生成。这些

方法可分为三大类。
首先,迪南等人。 （2018）;赵等人。 （2019）;团等人。 （2020）;金等人。(2020)从非结构化数据中提取有用的知识以生成响应，
例如段落中包含的信息和
演讲者的个人资料。其次,索多尼等人。 （2015）；龙等人。 （2017）;朱等人。 （2017）;加兹维宁·贾德等人。 （2018）;周等人。 （2018）;韦利奇科维奇等人。 （2018）;乔希等人。 （2020）;侯赛尼-阿斯尔等人。 （2020）;王等人。 （2021)利用来自的信息
知识库（图表或表格)来增强对话系统。他们通常会训练

知识的实体和关系嵌入
基础并将这些嵌入合并到
输入表示来预测响应。第三，
月亮等人。 （2019）； Tuan 等人。 （2019）;荣格等人。 （2020)更明确地将推理过程表述为知识图谱上的路径遍历。

这些方法进一步提高了透明度
以及对话代理的可解释性和
与我们分享最相似的想法。然而,他们
要么只预测推理路径而不生成响应,要么需要子图采样

减少KG规模。在这项工作中,我们的方法
使用变压器模型联合预测显式
大规模知识图谱的推理路径
并根据推理结果生成对话响应。

## 3 背景

### 3.1 对话系统知识图谱

我们假设系统的知识可以
用知识图谱（KG)表示G =
{E, R},其中E表示实体， R表示
的关系。知识图G包含
多个三元组描述之间的联系
实体和关系。我们表示第k 个三元组
该图为($e_{H_k}$、$r_k$和$e_{t_k}$)，其中$e_H$、$e_r$、$e_t$是
分别是头实体、关系和尾实体。
三元组、实体和关系的总数

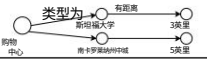
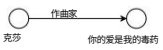
| 推理型 | | 例子 | 相关信息公斤 |
|---|---|---|---|
| KG推理 | | U:我需要无铅汽油。<br>R:通知 Valero,4 英里 |  |
| 真假 | | U:科罗娜这周会下雪吗?<br>答:是的 |  |
| 选择 | | U:告诉我去最近的购物中心的方向。<br>R:通知斯坦福购物中心,3 英里 |  |
| 萃取 | | U:这里有哪些加油站?<br>R:包括poi_type加油站 | 没有加油站<br>在可用公斤内 |
| KG推理 | | U:你听过歌手Kesha的歌吗?<br>R:我很喜欢她的音乐,尤其是《Your Love Is My Drug》 |  |

表 1:不同推理类型和输出格式（语义和自然语言形式)的示例
对话系统与可访问的知识图谱中的相关信息。

分别表示为NT 、 NE 、 NR 1。

3.2 对话系统中的响应生成

如果我们将对话历史定义为一系列
用户和系统期间发生的令牌
交互,然后扁平化的对话历史可以
写为:

$$x = (x1, x2, ..., xm, ..., xM) \quad (1)$$

其中xm是对话历史记录中的第 m 个标记
与M代币。在端到端的对话系统中,
我们假设一个由θ参数化的对话系统
存在可以预测概率分布
响应Pθ(· |x, G)。生成的响应是
从此概率分布中采样。

4 问题陈述

我们注重理解语言的能力
在对话期间进行推理的模型。我们考虑对话场景中通常需要的两个任务,并将它们称为语义任务

表1中的形式和自然语言（NL)形式。
首先,给定对话历史记录和用户查询,
语义形式的任务是预测下一个系统动作,对应于DM的输出
模块,基于现有的知识。在这个
在这种情况下,我们假设预期输出是NLG 模块的基本知识。我们认为
这项任务可以帮助更好地评估响应是否
正确与否以及哪种类型的推理可以
处理得更加顺利。其次,给定对话历史记录和用户查询,NL 的任务
形式可以是直接输出给定的响应

1G中三元组的一个例子是三元组e $^{H}_{k}$ =加油站,rk = IsTypeOf,并且e $^{t}_{k}$ = 雪佛龙。也就是说，“加油站是Chevron 类型”到该系统。

由系统。这种带有注释推理路径的设置可以帮助理解

模型可以学习支持闲聊和推理
同时。
此外,我们的目标是以逻辑推理的形式理解模型的推理能力
以及所提供的知识。如图所示
表1,通过KG推理,我们参考了
从任意位置检索信息的模型
多跳缩放 KG。同时,我们参考
逻辑推理作为模型的能力
进行操作,例如评估是否
陈述是真还是假,从中选择最小值/最大值
替代方案列表,并提取约束。
我们将我们关注的任务制定如下:给定对话历史x和当前可访问的KG G,我们是否可以扩展一个变压器模型
预测任一语义中的正确响应y
还是NL形式?如表1所示,该任务
不仅要求模型能够准确检索
从KG获取信息,还需要对信息进行进一步的逻辑运算。解决
对于这项任务,模型还应该能够有效地
将对话历史 x 与 KG G 整合。

5 提议的方法

图1说明了我们提出的架构
包含四个主要部分:对话历史编码器、可微 KG 推理模块、

可学习的逻辑运算模块和响应
解码器（变压器模型）。请注意,我们实验了两种类型的变压器:因果变压器
语言模型 GPT2 (Radford et al., 2019)和
和编码器-解码器模型 T5 （Raffel 等人， 2020）。
对于 GPT2,我们重复使用与以下位置相同的编码器
流程的开始,即图1中的fenc ,
作为生成响应的最终变压器

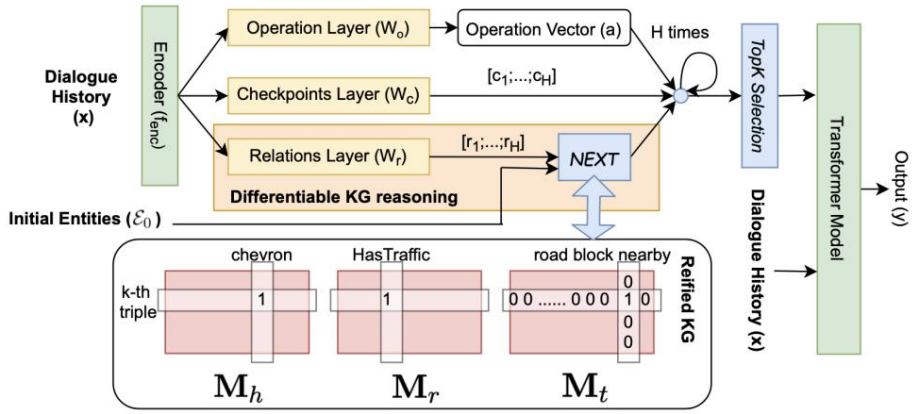



图 1:所提出的 DiffKG 的图示,它利用预训练的 Transformer 模型（T5 或 GPT2)和
具体化的 KG。该模型根据预测的关系序列[r1; r1; ...; rH],因此
根据所使用的推理路径是完全可解释的。

一个又一个令牌。对于 T5,我们重复使用相同的编码器
作为最终变压器的编码器,带有生成响应的单独解码
器。所以，
该方法包含单个变压器模型。
在以下部分中,我们将介绍每个模块
细节。

### 5.1 对话历史编码器

我们使用编码器模型来投影x并获得
通过x~ =嵌入对话历史记录
干草(x) ∈ IRd ，其中d是隐藏大小
编码器。嵌入x ~ 首先被输入
参数为Wo ∈ IRd×d的操作层。
运算层预测运算向量
IRd。 a = ₩T ∈ 同时,嵌入x ~ 也被输入到参数
为Wr ∈ IRd×NRH的关系层中。关系层预测关系序
列的串联

r = {rh|1 ⩽ h ⩽ H},其中rh ∈ IRNR是在编程的第 h 跳处
使用的关系
行走块， H是最大数量
酒花。嵌入x ~ 也被馈送到参数为Wc ∈ IRd×2H 的
检查点层。这
层产生序列的串联
步行或检查向量c = {ch|1 ⩽ h ⩽ H}，
其中ch ∈ IR2是第 h跳的步行或检查向量,以确定编
程的权重
行走模块和操作向量。

$$\begin{aligned} \tilde{x} &= \text{干草}(x), \\ a &= WT_{\text{氧}}\, X, \\ r &= WT_{r}\, X, \\ c &= \text{softmax}(WT_{c}\, X)。 \end{aligned} \tag{2}$$

### 5.2 差分知识图推理

为了确保我们的模型可以扩展到更大的 KG，
我们采用提出的具体化 KG 表示
作者：（Cohen 等人， 2019）。具体化 KG使用三个
稀疏矩阵表示图G：

头矩阵Mh ∈ IRNT ×NE关系矩阵 ，
Mr ∈ IRNT ×NR,尾矩阵Mt ∈ IRNT ×NE Mh或Mt中值为1的条目(i, e)表示 。

KG 中的第 i 个三元组以实体e作为
头或尾； Mr中值为1的条目(i, r)
表示知识图中的第i个三元组
具有关系r。由于在实际情况中经常
三个矩阵中的大多数条目为零,节省
将它们转化为稀疏矩阵可以显着减少
内存消耗（Cohen 等人， 2019）。
预测出关系序列r后,我们开始
从给定的一组初始实体进行图遍历
E0 ⊆ E。我们首先将初始实体映射到
向量e1 = [(e Î E0), ∀e Î]。也就是说,每个
如果该实体位于e1 ∈ IRNE中,则该条目的值为 1
初始实体列表E0,否则条目为
零。然后我们预测下一个（临时)实体
通过执行 Next 模块来向量e2：

$$\text{这是}^{r}_{\text{小时}+1} = \text{下一个（呃， rh）}, \tag{3}$$

在哪里

$$\text{下一个（呃， rh）} \frac{\text{公吨}_{t} \quad \text{（唷,先生）}}{= ||MT_{t} \quad \text{（咩咩）}||2 + (4)},$$

这是一个任意小的数字来抵消
分母并防止被零除。我们
引入归一化接下来要解决的是
使用（Cohen 等人，

2019)知识图补全定义为

跟随(eh, rh) = MT t （嗬先生）;自从
在对话模型中,我们很少能够预测
与实体向量完美匹配的关系向量。即如果直接使用
Follow模块
在（Cohen et al., 2019)中， ||eh||2不会是一个
并且会随着跳数h的增加而消失。
具体来说,请注意,在我们提出的模块中，
预测关系rh独立于遍历的实体eh 。例如,找到“附近加油站”的“距离”是独立的

附近的加油站是“Chevron”还是
“壳”。

允许模型动态选择
推理跳数,我们添加一个关系类型
“ToSelf”进入R并通过以下方式将每个实体与其自身连接起来
《给自己》。更具体地说,KG 将包含
三元组（e H k、rk和 k )对于所有e H k =和 t k ε E和rk =
致自己。

5.3 实体嵌入

在每一跳,我们进一步对由实体向量eh加权的实体进行操作向量a 。首先,我们对每个实体进行标记,并通过其标记嵌入的串联来表示它。此步骤允许 (1) 表示实体

较长的文本,例如短语和句子，
(2) 每当添加新实体值时,无需重新训练实体嵌入。

然后实体嵌入可以表示为
张量E ∈ IRNE×d×m,其中m是最大值
实体代币数量2 。

5.4 可学习的逻辑运算和
检查站

我们通过以下方式计算转换后的实体嵌入
实体嵌入E与第 h 跳处的实体向量eh的逐元素乘法。

接下来,运算向量的点积和
转换后的实体嵌入被传递给
softmax层作为下一跳的实体向量：

$$\overset{A}{\text{这是}}_{\text{小时+1}} = \text{softmax}(a(E\ eh)), \tag{5}$$

此外,在第h 跳,我们使用 walk-or-check
向量ch来组合上面的Next和操作模块。组合实体向量由下式给出：

$$\text{呃+1} = c \overset{\text{时间}}{H} \overset{r}{\overset{\text{这是}}{\text{小时+1}}} \overset{A}{\overset{\text{这是}}{\text{小时+1}}} \tag{6}$$
$$= c \overset{\text{时间}}{H} \text{下一个（呃， rh）}$$
$$\text{softmax}(a(E\ eh)),$$

5.5 响应解码器

H跳推理完成后,具有
选择实体向量eH中的前 k 个值，
表明他们的概率最高
从图中检索。这些实体是
转换为E中的嵌入并乘以eH 中的值。然后将这些实体嵌入与对话历史x连接起来。连接的向量被馈送为

输入到变压器模型中以逐个令牌预测响应令牌。预测概率

输出空间上的分布可以写成
为P(· |x,Mh,Mr,Mt)。由于所有组件都是
可微分,所有模块都可以使用对话历史x和具体化进行端到端训练
KG 表示{Mh,Mr,Mt}使用交叉熵损失,以真实输出y作为

标签。

$$L = \sum_{(x,y)} -\log P(y|x,Mh,Mr,Mt)。 \tag{7}$$

在推理期间,推理模块（关系层、操作层和检查点层)的工作方式与训练完全相同

阶段,唯一的区别是响应解码器被馈送了前一次预测的令牌
步骤（推理阶段)而不是事实真相
输出（训练阶段）。

## 6 实验

6.1 数据集

我们从三个方面评估我们提出的方法
数据集。其中,我们使用斯坦福多域对话（SMD） （Eric et al., 2017)和
OpenDialKG （Moon 等人， 2019)测试方法在不同对话类型上的通用性
（以任务为导向/闲聊)和结构化的规模
知识（成对数据库/通用知识图谱）。到
进一步分析推理能力,我们提出
一个新的数据集,SMD-Reasoning,通过修改
自然语言SMD数据集的输出
对行动的反应与其推理相结合
类型。

[2]在我们的实验中,我们计算了最大长度
所有实体并将较短的实体填充到 m 的长度。

斯坦福多域对话 (SMD) SMD 数据集（Eric 等人， 2017）由两个说话者对话组成,其中驾驶员与汽车助手对话以处理三个领域的任务：

调度、导航和天气预报。

每个对话都集中在一个领域,并与具有相关信息的数据库配对。我们将原始数据库转换为两种格式：(1)自然语言描述（NLD)和（2）KG。 NLD 形式允许我们研究模型解释非结构化知识的能力,而 KG 形式与表格相比可以是更具可扩展性的结构化知识。

OpenDialKG OpenDialKG 数据集（Moon 等人， 2019）由两位说话人的推荐和闲聊风格的对话组成。对话中的每一轮都用所提供的知识图谱上的推理路径进行注释,该知识图谱是从 Freebase 中过滤出来的（Bollacker 等人，2008）。生成的 KG 有 1,190,658 个三元组、100,813 个实体和 1,358 个关系。我们按照(Moon et al., 2019; Jung et al., 2020)中的描述,对训练/有效/测试集随机分割 70/15/15% ,因为它们不会发布分割结果。

SMD-Reasoning为了使SMD数据集适合更精确地评估推理能力，我们手动将其标记并转换为SMD Reasoning数据集。我们首先从原始响应中删除自然语言部分,只留下动作词（例如,通知)以及所传达的信息。我们将数据集分为三种主要推理类型:通知项目、选择最小值/最大值以及评估真/假。

为了验证模型是否可以识别所需的知识是否在数据库中，我们添加了一种新的推理类型来提取约束,方法是从数据库中删除所需的知识并将输出更改为“包含[知识描述]”，如下所示表1。有关这些数据集的统计信息,请参阅附录A、B 。

6.2 评估指标

我们对三者采用不同的评价方法

数据集。对于 SMD,我们遵循先前的工作（Yang 等人， 2020)并使用 BLEU （Papineni 等人， 2002)以及每个域上的实体 F1 分数。对于 OpenDialKG,我们遵循先前作品中的描述（Moon 等人， 2019； Jung 等人， 2020)来评估path@k 分数,即,如果真实路径在预测中排名前 k路径概率。此外,由于我们的方法不仅可以与之前的工作一样预测推理路径,但也可以预测响应,我们还使用 BLEU分数来获得与真实值相比的响应质量的近似评估。请注意,之前的工作已经讨论过 BLEU 分数可能与人类直觉不符（Liu et al., 2016）,但我们在这里将它们用作近似评估以供参考。

对于 SMD 推理,输出更具确定性,并且不包括不同的句子结构。因此,我们计算 F1 分数以及预测和真实值的精确匹配 (EM) 分数。 EM 分数是通过删除预测的顺序来计算的,因为 SMD 推理数据集的标签遵循原始地面实况响应中出现的知识描述的顺序,并且可能与生成的输出具有不同的顺序。 EM 分数可以写为：

$$\text{电磁} = \frac{1}{\text{时间}} \quad (\text{排序（y)=排序（y））。 (8)}$$

其中y^是使用 argmax 采样从模型中推断出来的,T 是示例总数。

6.3 实现细节由于所提出的方法与模型无关,我们在 GPT2 (Radford et al., 2019)和 T5 (Raffel et al., 2020)上实现它。具体来说,对于 T5模型,我们使用统一的 QA-T5 模型（Khashabi et al., 2020） ,该模型针对也需要进行推理的问答任务进行了预训练。然而,我们凭经验发现T5通常比GPT2具有更好的性能,因此在大多数实验中使用T5模型。对于 OpenDialKG，由于真实关系存在,我们将它们作为额外的监督信号（Moon et al., 2019）。另外,由于我们观察到 OpenDialKG 中只有 KG 推理类型,因此我们不对数据集使用操作层和检查点层。超参数设置在附录C中。

6.4 基线

我们将我们提出的 DiffKG 模型与（Moon 等人， 2019； Jung 等人， 2020)报告的 OpenDialKG 上最先进的模型以及基于图的最先进模型进行比较SMD （Yang et al., 2020； Gou et al., 2021)及其报告的基线,包括带注意力和不带注意力的序列到序列模型（S2S和 S2S+Attn） （Luong et al., 2015）,指向未知（ Ptr-Unk） （Gulcehre 等人， 2016）,

| 模型 | 蓝色的 | 实体F1<br>所有的 | 谢。 | 威。 | 导航。 |
|---|---|---|---|---|---|
| S2S 8.4 S2S+Attn 9.3 Ptr-Unk<br>8.3 GraphLSTM 10.3 BERT<br>9.13 Mem2Seq 12.6 GLMP<br>12.2 GraphDialog 13.7 COMET-graph 14.4 | | 10.3 7.0<br>19.9 23.4<br>22.7 26.9<br>50.8 69.9<br>49.6 57.4<br>33.4 49.3<br>55.1 67.3<br>57.4 71.9<br>56.7 71.6 | 9.7<br>25.6<br>26.7<br>46.6<br>47.5<br>32.8<br>54.1<br>59.7<br>48.7 | 14.1<br>10.8<br>14.9<br>43.2<br>46.8<br>20.0<br>48.4<br>48.6<br>50.4 | |
| T5-DiffKG | 16.04 | 56.2 | 67.2 | 61.5 | 46.7 |

表 2:SMD 数据集的结果。 S2S、S2S+Attn、Ptr-Unk、GraphLSTM、BERT、Mem2Seq、GLMP、GraphDialog 的报告来自（Yang et al., 2020)和 COMET 图来自（Gou 等人， 2021）。我们的DiffKG 获得最高的 BLEU 和可比较的 F1 分数 与基线。

| 模型 | 路径@1 | 路径@5 | 路径@10 | BLEU |
|---|---|---|---|---|
| 序列到序列 | 3.1 | 29.7 | 44.1 | - |
| 三LSTM | 3.2 | 22.6 | 36.3 | - |
| 扩展-ED | 1.9 | 9.0 | 13.3 | - |
| 拨号KG | 13.2 | 35.3 | 47.9 | - |
| 序列2路径 | 14.92 | 31.1 | 38.68 | - |
| AttnFlow | 17.37 | 30.68 | 39.48 | - |
| AttnIO-AS | 23.72 | 43.57 | 52.17 | - |
| T5-NoInfo | - | - | - | 14.51 |
| T5-DiffKG | 26.80 | 54.33 | 61.75 | 15.37 |

表 3:OpenDialKG 数据集上的结果。四个报告从 Seq2Seq 到 DialKG Walker 的基线来自（Moon et al., 2019)和其他三个基线从 Seq2Path 到 AttnIO-AS 的报告来自(Jung 等人， 2020）。我们的DiffKG达到了最高的path@k 分数,并且是唯一可以同时生成响应的。

GraphLSTM （Peng 等人， 2017）、 BERT （Devlin 等人， 2019）， Mem2Seq （Madotto 等人， 2018） ,以及 GLMP （Wu 等人， 2019）。我们遵循他们的指标并根据他们的预处理数据训练我们的模型公平比较。进一步分析推理能力,我们提出了另外两个基线利用预训练语言的不同方式楷模。 (1) NoInfo模型不采用任何格式以知识作为输入,旨在测试微调的普通变压器模型的性能

在每个数据集上。 (2) FlatInfo模型构造通过连接对话历史记录来输入NLD 的知识形式为（Beygi 等人， 2022），让我们能够研究模型的能力解释非结构化知识。

| 测试公斤数方法 | | 在F1中 |
|---|---|---|
| 固定的 | GPT2-无信息 10.71<br>GPT2-FlatInfo 14.08<br>GPT2-差异KG 16.39<br>T5-无信息 10.50<br>T5-FlatInfo 28.99<br>T5-差KG 27.52 | 43.78<br>47.57<br>51.06<br>44.27<br>66.15<br>63.93 |
| 洗牌 | T5-FlatInfo<br>T5-DiffKG | 17.02 54.51<br>27.52 64.00 |

表 4:SMD-Reasoning 数据集的结果。

6.5 结果

SMD 和 OpenDialKG 上的结果如图所示表2和表3 中。在 SMD 数据集上,我们观察到 DiffKG 优于基线BLEU 增加 11.4%（相对变化 16.04 和 14.4) 并达到可比较的实体 F1 分数使用 GLMP、GraphDialog 和 COMET-graph。 Dif fKG 可能不会提高实体 F1 分数,因为先前的工作将实体内的文本分组在一起（例如，“附近的路障”变成

词汇中的一个单词“road_block_nearby”）。相反,我们使用通用分词器,因此以防止繁重的预处理和专门的词汇。这意味着DiffKG可以执行与最先进的知识检索类似没有为每个数据集指定标记器。在OpenDialKG 数据集,我们观察到 DiffKG在 path@k 分数方面优于基线并且可以同时超越T5实体 F1 和 BLEU。这些表明DiffKG 可以检索准确的路径进行推理并有效地将推理融入响应中一代。

我们还研究了 SMD Reasoning 数据集的结果,如表 4 所示。我们发现DiffKG 将 NoInfo 提高了 16.6% 和 44.4%分别在 GPT2 和 T5 模型上获得 F1 分数。这表明 DiffKG 可以有效地利用知识来改进生成,而无需访问信息。相比之下,虽然

FlatInfo 的性能与 DiffKG 类似SMD-Reasoning 数据集,无法运行OpenDialKG 由于计算成本。更多的具体来说,FlatInfo 需要知识图谱转化为句子,结果是至少一百万个代币作为模型输入OpenDialKG（因为三元组的数量是一百万,没有设计子图采样），不是一个实际的数字。

| 方法 | 域名 | | | 推理类型 | | |
|---|---|---|---|---|---|---|
| | 日程 | 导航 | 天气 | 通知 | 选择 | 提取真/假 |
| | 在F1中 在F1中 F1中 F1中 F1中 F1中 | | | | | |
| GPT2-无信息 | 3.49 45.7 4.63 41.6 | 27.5 46.8 | 5.03 45.2 1.45 47.4 | 3.06 24.0 68.6 68.6 | | |
| GPT2-DiffKG | 9.30 53.0 9.65 47.6 | 34.4 56.5 | 8.04 50.8 0.00 48.5 | 31.6 53.5 56.9 56.9 | | |
| T5-无信息 | 0.00 44.6 4.63 40.9 | 29.0 50.7 | 3.02 44.9 8.70 49.1 | 1.02 25.2 70.6 70.6 | | |
| T5-差公斤 | 20.9 63.8 19.3 61.9 | 48.1 68.1 | 18.1 61.7 11.6 62.4 | 50.0 73.5 70.6 70.6 | | |

表 5:SMD-Reasoning 数据集的详细评估结果

| | |
|---|---|
| SMD-推理 | 用户:检查我的医生预约的日期和时间<br>已日期、已时间、ToSelf<br>（推理路径:预约医生----------→ 周二上午11点,预约医生）<br>DiffKG:通知周二上午 11 点医生预约 |
| | 用户： 汽车 我需要去加油站,请告诉我最近的加油站<br>助理:瓦莱罗 (Valero) 距离 7 英里,途中交通状况一般<br>用户:好的,它在哪里?<br>类型为 有地址,给自己<br>（推理路径:加油站→瓦莱罗→200 Alester Ave, Valero）<br>DiffKG:通知 200 Alester Ave Valero |
| 开放拨号KG | 演讲者A:你有关于托尼·克罗斯的任何信息吗?<br>玩家统计<br>（推理路径:托尼·克罗斯 →德国国足）<br>DiffKG:托尼·克罗斯是德国足球运动员,效力于德国国家足球队。 |

表 6:生成的示例和推理路径。

6.6 定量分析

为了测试这些方法对准确定位信息的鲁棒性,我们打乱了信息顺序。本次评测是模拟案例

当部署对话系统时,额外的信息是任意添加的。具体来说,顺序
FlatInfo 的知识上下文和顺序
的知识三元组在推理过程中发生变化
时间。如表4最后两行所示，
FlatInfo 的性能下降,而 DiffKG 的性能基本保持不变。这表明轻微
FlatInfo 与原版相比性能更优越
顺序可以来自黑盒技巧来分组
输入中的附近知识。当这个
隐式技巧被打破,DiffKG 显示
更好的鲁棒性和性能。

调查每个领域的难度
推理类型,我们相应地划分结果
如表5 所示。如域部分所示，
该模型在以下方面达到了最高的 EM 和 F1
天气领域。我们推测原因是
天气领域包含更多推理
类型（天气:4,导航:3,时间表:2,如附录A表9所示），从而反映更平衡
推理能力。在推理类型部分,我们
观察到 Dif fKG 不太能很好地处理真/假;然而,DiffKG 改进了提取。
这表明DiffKG可以有效地检查
是否存在所需知识,然后查询
数据库。

关于计算成本（在使用 T5 模型的 SMD Reasoning 数据集上）,我们发现
DiffKG 运行期间需要大约 5.85GB 内存
训练并具有 30 毫秒的推理延迟。这可以
与没有知识的模型相比,是可接受的附加内存使用和推理时间
推理（3.13GB;30ms）。特别是当像 FlatInfo 这样的基线消耗更多时（18.56GB;
50 毫秒）。

6.7 定性分析

我们可视化生成的示例和
表 6 中的 SMD Reasoning 和 OpenDialKG 数据集上 DiffKG 的符号推理路径。
示例表明 DiffKG 可以捕获一些
该数据集中自然发生的现象：
(1)KG推理路径可以是1到倍数
酒花; (2) 推理会扩散到多个
路径（例如,DiffKG 同时将“Has Date”、“HasTime”、“ToSelf”应用到“Doctor Appointment”）。结合前面小节中的分析,我们观察到 DiffKG 可以提取可解释的推理路径,并使用合理的计算成本生成相应的输出。

然而,即使 DiffKG 可以维持或
在任何缩放的 KG 上进行可解释推理时提高性能,错误可能会发生在

| 错误类型 | 例子 |
| --- | --- |
| 不清楚的信息要求 | 用户:这里有哪些加油站?<br>地面实况:通知 Chevron DiffKG:<br>通知 5 英里,Chevron （我们的评论:虽然这个预测对于 EM 指标来说被认为是错误的,但DiffKG 输出的“5 英里”是距“Chevron”的正确距离,并且可能需要在反应很好。） |
| 不完整的推理但忠实的回应 | 用户:最近的杂货店在哪里?<br>事实真相:Inform 4miles、Whole Foods、Safeway。<br>DiffKG:通知 4 英里,杂货店,819 Alma St,Whole Foods （我们的评论:4 英里,杂货店,<br>819 Alma St 都是关于 Whole Foods 的正确实体。尽管如此,这个推理过程忽略了另一家杂货店 Safeway,它也是4英里远。） |
| 正确的推理但错误的反应 | 演讲者 A:你认识唐·霍尔吗?<br>真相:唐·霍尔（Don Hall)为《公主与青蛙》写了一部由詹妮弗·刘易斯（Jenifer Lewis)主演的浪漫故事。<br>你喜欢浪漫吗?<br>推理路径:Don Hall ----------→ 《公主与青蛙》 DiffKG:是的,他写了《小娃娃》。<br>（我们的评论:找出 Don Hall 编写的脚本的推理路径是正确的。但是,生成过程未能正确利用检索到的实体。） |

表 7:跨数据集的三种主要错误类型的错误分析。

一些案例。如表7所示,我们发现在整个数据集中,DiffKG的三种主要错误类型是:(1)数据集中的信息要求不明确,(2)推理能力不完整,但生成的响应忠实,(3)推理正确但幻觉反应预测。我们认为,第一种错误类型主要来自数据集中数据点之间的不匹配,并且可能无法通过模型来处理。第二种错误类型表示KG推理模块有时无法检索到所有需要的信息。第三种错误类型表明产生最终输出的模块可能未充分利用检索到的信息。这三点或许可以为进一步改进提供方向。

## 7 结论和未来工作

对于对话系统来说,基于结构化数据库的有效推理方法非常重要。在这项工作中,我们提出了 DiffKG,这是一种与模型无关的端到端方法,可以在任何规模的知识图谱上进行符号推理以增强响应生成。实验证明,使用 Dif fKG,模型能够以适度的额外成本生成具有可解释的 KG 推理路径的响应。

这项工作可以通过多种方式进行扩展。虽然我们在对话生成中只考虑高效的大规模知识图谱推理,但未来的工作可以结合域融合方法来考虑域上的泛化性或同时使用关系信息。此外,由于 DiffKG是一个简单的大规模结构化知识赋能转换器,具有灵活的实体值,未来的工作可以将其扩展到需要进行表和文本混合推理的对话生成,并且需要同时进行KG推理和其他目标,例如个性化对话、讲故事， ETC。

## 参考

Sajjad Beygi、Maryam Fazel-Zarandi、Alessandra Cervone、Prakash Krishnan 和 Siddhartha Reddy Jonnalagadda。2022 年。 面向任务的对话系统的逻辑推理。 arXiv 预印本 arXiv:2202.04161。

库尔特·博拉克、科林·埃文斯、普拉文·帕里托什、蒂姆·斯特奇和杰米·泰勒。 2008. Freebase:一个协作创建的图形数据库,用于构建人类知识。 2008 年 ACM SIGMOD 国际数据管理会议论文集。

威廉·W·科恩、海天·孙、R·亚历克斯·霍弗和马修·西格勒。2019.使用符号知识库进行推理的可扩展神经方法。在国际学习代表会议上。

雅各布·德夫林、张明伟、肯顿·李和克里斯蒂娜·图塔诺娃。2019. Bert:用于语言理解的深度双向变压器的预训练。在 NAACL-HLT (1) 中。

艾米丽·迪南、史蒂芬·罗勒、库尔特·舒斯特、安吉拉·范、迈克尔·奥利和杰森·韦斯顿。 2018. 维基百科向导:知识驱动的对话代理。在国际学习代表会议上。

朱利安·马丁·艾森施洛斯、马赫什·戈尔、托马斯
穆勒和威廉·W·科恩。 2021.Mate:对表格变压器效率的多视图关注。 arXiv
预印本 arXiv:2109.04312。

米哈伊尔·埃里克、拉克希米·克里希南、弗朗索瓦·夏莱特和
克里斯托弗·D·曼宁. 2017.键值检索
以任务为导向的对话网络。诉讼中
第 18 届 SIGdial 话语年会
和对话。

马里安·加兹维内贾德、克里斯·布罗克特、明伟
Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and
米歇尔·加莱. 2018.基于知识的神经网络
对话模型。第三十二届 AAAI 人工智能大会。

Yanjie Gou, Yinjie Lei, Lingqiao Liu, Yong Dai,
还有沈春旭。 2021.将知识情境化
具有面向端到端任务的变压器的基础
对话系统。 2021 年自然语言经验方法会议论文集

处理,EMNLP 2021,虚拟活动 / Punta
多米尼加共和国卡纳,2021 年 11 月 7 日至 11 日。
计算语言学协会。

卡格拉·古尔切雷、安圣金、拉梅什·纳拉帕蒂、
Bowen Zhou, and Yoshua Bengio. 2016. Pointing
那些不知名的词。计算协会第 54 届年会记录

语言学（第一卷:长论文）。

Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He,
刘占义、吴华、赵军。 2017.一种结合全局交叉注意力的知识库问答端
到端模型

知识。计算语言学协会第 55 届年会记录

（第一卷:长论文）,第一卷。

Ehsan Hosseini-Asl、Bryan McCann、吴建胜、
塞米赫·亚乌兹和理查德·索彻。 2020.一个简单的
面向任务的对话的语言模型。在神经信息处理系统的进展中，

第 33 卷,第 20179–20191 页。柯兰联营公司，
公司

Rishabh Joshi、Vidhisha Balachandran、Shikhar
瓦西什、艾伦·布莱克和尤利娅·茨维特科夫。 2020.
Diagraph:将可解释的策略图网络纳入谈判对话中。在国际学习代表
会议上。

郑在勋、孙福京、柳成源。 2020.
Attnio:利用进出注意力流进行知识图探索,以进行基于知识的对话。

2020 年实证会议论文集
自然语言处理方法（EMNLP）。

丹·尤拉夫斯基和詹姆斯·H·马丁。 2009.演讲
和语言处理:自然简介
语言处理、计算语言学和
语音识别,第二版。 Prentice Hall 的人工智能系列。
皮尔逊·普伦蒂斯·霍尔
教育国际。

丹尼尔·卡沙比、Sewon Min、Tushar Khot、Ashish
Sabharwal、Oyvind Tafjord、Peter Clark 和 Han naneh
Hajishirzi。 2020.Unifiedqa:交叉格式
与单一质量保证系统的边界。诉讼中
2020 年实证方法会议
自然语言处理:发现。

图沙尔·科特、彼得·克拉克、米哈尔·盖尔昆、彼得
詹森和阿什什·萨巴瓦尔。 2020.Qasc:A
通过句子构成回答问题的数据集。在 AAAI 会议记录中

人工智能,第 34 卷,第 8082–8090 页。

Seokhwan Kim、米哈伊尔·埃里克、卡蒂克·戈帕拉克里希南、
Behnam Hedayatnia、Yang Liu 和 Dilek Hakkani Tur。 2020.
超越领域 API:具有非结构化知识的面向任务的会话建模

使用权。第21届年会记录
话语和对话特别兴趣小组的成员。

Bill Yuchen Lin、海天孙、Bhuwan Dhingra、
Manzil Zaheer, Xiang Ren, and William Cohen.
2021.可微的开放式常识推理。 2021 年会议记录

协会北美分会
计算语言学:人类语言技术。

刘家伟、瑞安·洛、尤利安·塞尔班、迈克·诺斯·沃斯、洛朗·
查林和乔尔·皮诺。 2016年。
如何不评估你的对话系统:无监督评估指标的实证研究

对话响应生成。在诉讼程序中
2016年自然经验方法会议
语言处理。

Yinong Long, Jianan Wang, Zhen Xu, Zongsheng
Wang, Baoxun Wang, and Zhuoran Wang. 2017. A
知识增强的生成式会话服务代理。第六届对话系统论文集

技术挑战 (DSTC6) 研讨会。

Thang Luong、Hieu Pham 和 Christopher D Manning。
2015。基于注意力的神经机器翻译的有效方法。在 EMNLP 中。

安德里亚·马多托、吴建生和冯帕斯卡。
2018. Mem2seq:有效地将知识库整合到端到端的面向任务的对话
系统中。第 56 届年会记录

计算语言学协会（第一卷:长论文）。

托多尔·米哈伊洛夫、彼得·克拉克、图沙尔·科特和阿什什
萨巴瓦尔。 2018.一套铠甲能导电吗
三联市?用于开卷问答的新数据集。 arXiv 预印本 arXiv:1809.02789。

Seungwhan Moon、Pararth Shah、Anuj Kumar 和 Ra jen Subba。
2019.Opendialkg:基于注意力的可解释对话推理

知识图谱。计算协会第 57 届年会记录

语言学。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei Jing Zhu。 2002. Bleu：一种机器翻译自动评估方法。在诉讼程序中

计算语言学协会第 40 届年会。

Nanyun Peng、Hoifung Poon、Chris Quirk、Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence 使用图 lstms 进行 n 元关系提取。计算语言学协会汇刊，5。

亚历克·雷德福、杰弗里·吴、Rewon Child、大卫·栾、达里奥·阿莫代、伊利亚·苏茨克韦尔等人。 2019。语言模型是无监督的多任务学习者。
OpenAI 博客，1(8):9。

科林·拉斐尔、诺姆·沙泽尔、亚当·罗伯茨、凯瑟琳 Lee, Sharan Narang, Michael Matena, Yanqi Zhou, 李伟，彼得·J·刘。 2020.用统一的文本到文本探索迁移学习的局限性
变压器。机器学习研究杂志，
21（140）：1-67。

普拉纳夫·拉杰普尔卡、罗宾·贾和珀西·梁。 2018. 知道你不知道的事情：小队无法回答的问题。 arXiv 预印本 arXiv：1806.03822。

西瓦·雷迪、陈丹琪和克里斯托弗·D·曼宁。 2019.Coqa：对话式问答
挑战。计算语言学协会汇刊，7：249-266。

尤利安·V·塞尔班、亚历山德罗·索尔多尼、约书亚·本吉奥、亚伦·库维尔和乔尔·皮诺。 2016. 使用生成式构建端到端对话系统
分层神经网络模型。第 30 届 AAAI 人工智能会议 (AAAI-16) 论文集。

亚历山德罗·索尔多尼、米歇尔·加利、迈克尔·奥利、克里斯·布罗克特、季扬峰、玛格丽特·米切尔、Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. 一种用于上下文敏感生成对话响应的神经网络方法。诉讼中

计算语言学协会北美分会 2015 年会议的发言：

人类语言技术。

谢恩·斯托克斯、乔子高和乔伊斯·Y·柴。 2019. 自然语言推理的最新进展：A
基准、资源和方法的调查。
arXiv 预印本 arXiv：1904.01172。

Yi-Lin Tuan, Yun-Nung Chen, and Hung-Yi Lee. 2019. Dykgchat：基于动态知识图的对话生成基准测试。在

2019年实证会议论文集
自然语言处理方法及其
第九届国际自然语言处理联合会议（EMNLP-IJCNLP）。

Yi-Lin Tuan, Wei Wei, and William Yang Wang. 2020. 通过以下方式将知识注入到对话生成中
语言模型。 arXiv 预印本 arXiv：2004.14614。

Petar Velickovic、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。 2018.图注意力网络。学习表征国际会议。

帕特·维尔加、海天太阳、利维奥·巴尔迪尼·苏亚雷斯和威廉·W·科恩. 2020.事实作为专家：适应性和可解释的神经记忆优于符号
知识。 arXiv 预印本 arXiv：2007.00849。

Qingyue Wang, Yanan Cao, Junyan Jiang, Yafang 王，佟玲玲，郭丽。 2021. 将特定知识纳入端到端任务导向

对话系统。 2021 年国际神经网络联合会议 (IJCNN)，第 1-8 页。
IEEE。

文宗宪、米利卡·加西奇、尼古拉·米尔克西奇、Lina M Rojas-Barahona、苏培豪、Stefan Ultes、大卫·范戴克和史蒂夫·杨。 2017.基于网络的端到端可训练的任务导向对话
系统。在欧洲计算协会
语言学 (EACL)，第 438-449 页。

杰森·韦斯顿、安托万·博德斯、苏米特·乔普拉、亚历山大·德·M·拉什、巴特·范·梅林布尔、阿曼德·朱林、和托马斯·米科洛夫。 2015.迈向人工智能完成
问题回答：一组先决玩具任务。
arXiv 预印本 arXiv：1502.05698。

贾森·威廉姆斯、安托万·劳克斯和马修·亨德的儿子。 2016。对话状态跟踪挑战系列：回顾。对话与话语，7（3）：4-33。

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019.全局到局部内存指针网络
面向任务的对话。在国际会议上
关于学习表示。

Shiquan Yang, Rui Zhang, and Sarah Erfani. 2020. Graphdialog：将图知识集成到端到端的面向任务的对话系统中。 2020 年实证方法会议论文集

自然语言处理（EMNLP）。

Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, 李航，李晓明。 2016.神经生成问答。人机问答研讨会论文集。

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2019. Low-resource 基于知识的对话生成。在国际学习代表会议上。

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Com monsense knowledge aware conversation generation
具有图注意力。在 IJCAI 中。

Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible 知识扎根的端到端对话系统
对话。 arXiv 预印本 arXiv：1709.04264。

数据集详细信息

数据集统计结果见表8和表9。

OpenDialKG 数据集位于 CC-BY-NC-4.0 下
执照。这些数据集可用于研究
目的。

| 数据 | 列车验证测试 | | |
|---|---|---|---|
| 贴片2425 | | 302 | 304 |
| 开放拨号KG 10971 | | 2351 | 2351 |

表 8:每个数据分组中的对话数量。

| | 时间表 导航 天气 | | |
|---|---|---|---|
| 通知 | 第364章 | 1133 | 236 |
| 选择 | - | 686 | 39 |
| 真假 | - | - | 第543章 |
| 萃取 | 173 | 第474章 | 214 |

表 9:SMD Reasoning 数据集的统计
尊重领域和推理类型。注意
由于推理类型是按回合级别分类的，
该表中的总数大于表8中的
计入对话级别。

B SMD KG 施工

我们编写一个简单的自动程序来构建
从原始映射的 SMD 数据集的 KG
带注释的表格。

对于时间表和导航域
SMD，我们直接将他们的表属性映射到
我们构建的 KG 中的关系R。为了天气
域中，我们将每个天气报告分为低温、高温和天气。结果
关系数为 29，并且关系

列于表10 中。

在日程表和导航域中，每个
原始数据库中的项目有多个
属性被转换为 KG 三元组：
（事件/兴趣点、属性、属性值），
例如，（网球活动、HasTime、
晚上 7 点）在计划域中或（Chevron，
导航中的HasType（加油站）
领域。

在天气域中，我们添加名为“ReportID$digits$”的附加实体，其中 $digits$
将被替换为 ID 号。每一个项目
原始数据库中的格式为：(item, location , $location),
(item, $date, $weather_report),
其中 $weather_report 包含多个不同时需要的信息。为了使

天气公斤数与行程公斤数一致

| 领域 | 关系 |
|---|---|
| 安排HasTime、HasDate、HasParty、 | 哈斯房间、哈斯议程、是时间、日期、派对、是房间，是议程 |
| 导航HasAddress、HasType、HasTraf fic、HasDistance、 | IsAddressOf、IsTypeOf、IsTrafficOf、IsDistanceFrom |
| 天气 | 等于，有位置，有天气，有低温，有高温，有日期，IsWeatherOf、IsLowTempOf、是高温、是位置、是日期 |

表 10:SMD 中每个域中使用的关系
数据集。

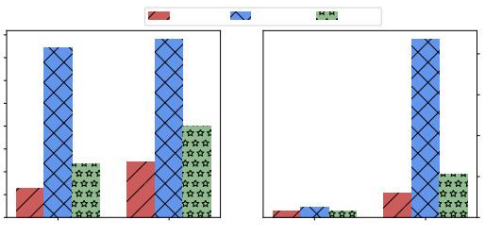


图2:训练消耗对比
内存和推理延迟。

和导航，我们将每个项目转换为 (Re portID, location, $location), (ReportID, HasDate, $date), (ReportID, HasWeather, $weather), (ReportID , HasLowTemp, $low_temparature), (ReportID, 有HighTemp，$high_temparature）。

C 实验细节

我们为 DiffKG 设置的超参数是d=the
使用的预训练 Transformer 的隐藏大小（T5-小： d=512； GPT2:d=768), H=5, 最大范数=1.0,
批量大小=16，梯度累积步数=2
最多50 个epoch 并训练模型学习率 ∈ {5 × 10−5
, 6.25 × 10−5}（发现
6.25 × 10−5更好），没有学习率衰减。
我们的实验是随机初始化的单次运行，没有进一步微调。

D 计算成本分析

如图2 所示，在 SMD-Reasoning 数据集上，
FlatInfo 消耗的内存是
DiffKG 在训练时所需的内存，及其

延迟大约是推理时间的两倍。 GPT2 的推理延迟
差异甚至更大
作为骨干模型。原因是因果语言模型的计算成本（例
如
因为 GPT2 很大程度上取决于输入序列
length，这是FlatInfo的主要问题之一。