arXiv:2301.01664v2 [cs.CL] 1 May 2023

基于句子Transformer的多方面可解释归纳关系预测

Zhixiang Su 1,2, Di Wang [3,4], **Chunyan Miao** 1,2,3,4, Lizhen Cui[2,5]l新加坡南洋理工大学计算机科学与工程学院[2]中国山东大学—南洋理工大学人工智能研究中心（C-FAIR）[3]南洋理工大学—英属哥伦比亚大学老年人积极生活卓越研究中心（LILY），新加坡[4]南洋理工大学—微众银行金融科技联合研究中心，新加坡 [5]中国山东大学软件学院{zhixiang002, wangdi, ascymiao}@ntu.edu.sg,clz@sdu.edu.cn

摘要

近期关于知识图谱（KG）的研究表明，借助预训练语言模型的基于路径的方法在提供归纳式且可解释的关系预测方面表现出色。本文提出关系路径覆盖率和关系路径置信度的概念，在模型训练前过滤掉不可靠的路径，以提升模型性能。此外，我们提出知识推理句子转换器（KRST），用于预测知识图谱中的归纳式关系。KRST旨在对知识图谱中提取的可靠路径进行编码，从而能够适当地对路径进行聚类并提供多方面解释。我们在三个真实世界数据集上进行了广泛实验。实验结果表明，与最先进模型相比，KRST在大多数传导式和归纳式测试用例（6例中有4例）以及少样本测试用例（12例中有11例）中取得了最佳性能。

引言

作为问答和推荐系统提供侧信息的重要工具，知识图谱（KG）已得到广泛研究（Ji 等，2021）。知识图谱通常表示为三元组 $G = \{(h_i, r_i, t_i)|i = 1, 2, 3, .., m\}$，其中包含实体 $h_i, t_i \in E_G$ 和关系 $r_i \in R_G$。由于实际应用中的知识图谱并不完整，因此需要进行知识图谱补全（KGC）以提升知识图谱的质量。知识图谱补全中最重要的任务之一是关系预测。给定目标三元组 (h, r, t)，通常通过遮蔽给定三元组中的实体 h 或 t 来构造关系预测查询，并让模型根据另一个实体和关系类型预测被遮蔽的实体。

基于嵌入的方法可能是应用最广泛的最先进模型。在实体和关系集合固定的情况下，基于嵌入的方法在知识图谱补全任务中表现相当出色。然而，现有的大多数基于嵌入的方法缺乏可解释性，也无法处理归纳式情形，因此不适合对现实世界中的动态知识图谱进行建模，因为新的实体和关系可能随时加入。归纳关系预测要求模型能够处理训练图中不存在的未见实体。

基于图神经网络的方法利用知识图谱的图连通性，因此能够预测具有足够数量已知

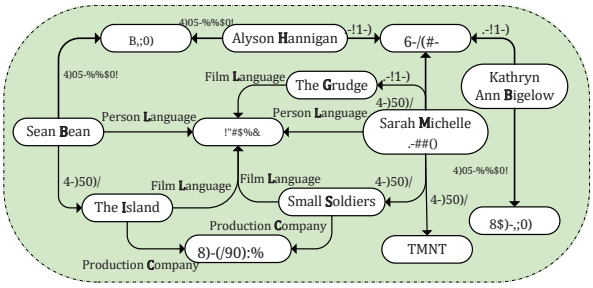


图 1：知识图谱示例。

相邻实体的新实体。近期提出的基于图神经网络的方法 GRAIL（Teru、Denis 和 Hamilton，2020年）已被证明能够进行归纳关系预测。然而，在 GRAIL 中，提取可解释规则仍有待进一步探索。迄今为止，没有证据表明 GRAIL 具有可解释性。

与基于图神经网络的方法不同，基于路径的方法通过分析头实体和尾实体之间的路径来利用图连通性。当我们只关注一条路径中的关系时，可以将它们表述为用于知识推理的霍恩规则（Horn，1951年）。因此，在基于路径的方法中，可以通过应用归纳出的规则轻松地对新实体进行建模。例如，关于图 1，对于三元组（《玩具兵团》、电影语言、英语）及其对应的路径

$$Small\ Soldiers \xrightarrow{Perform^{-1}} Sarah \xrightarrow{Person\ Language} English, \quad (1)$$

我们可以归纳出

$$(y, Perform, x) \wedge (y, PersonLanguage, z) \rightarrow (x, FilmLanguage, z). \quad (2)$$

这种基于规则的操作使基于路径的方法具有归纳性且高度可解释。BERTRL（Zha、Chen 和 Yan（2022年））采用预训练语言模型 BERT 进行评分，是此类模型中的典型代表。借助实体和关系的上下文描述，BERTRL 能够处理归纳式情形并提供单路径解释，在相关文献中取得了最先进的结果。

为进一步提升关系预测性能并追求更好的可解释性，本文提出了知识推理句子Transformer（KRST），

这是一种基于路径的新型模型，构建于句子转换器之上。
KRST 的主要创新如下：
路径提取。尽管可以轻松提取头实体和尾实体之间的路径，但提取出的路径大多可能并不可靠。此外，预训练语言模型通常更倾向于选择包含目标三元组中也出现的词语的不可靠短路径。为了评估路径的可靠性，并仅使用可靠路径进行模型训练，我们提出了关系路径覆盖率和关系路径置信度的概念（分别参见定义 4 和定义 5）。

预训练语言模型。与常用的 BERT 序列分类模型相比，采用余弦相似度的句子转换器在多项任务中表现更佳（Reimers 和 Gurevych，2019年）。因此，我们在 KRST 中采用句子转换器，将三元组和路径编码为嵌入，从而能够显式地比较嵌入，并根据不同方面对路径进行聚类。
损失函数。在 KRST 中，我们使用余弦相似度分数比较路径与三元组之间的相似度。对于负三元组，负分数不一定接近 -1。然而，常用的二元分类损失函数（例如交叉熵）通常会要求预测结果接近 1 或 -1，从而使模型过度自信。因此，我们改用余弦嵌入损失，以期更好地提升模型性能。

本文的主要贡献如下：

(i) 我们提出了两种新颖的路径提取指标，分别称为关系路径覆盖率和关系路径置信度，并提出了一种新颖的基于路径的模型 KRST。据我们所知，KRST 是首个用于知识图谱路径编码的句子转换器模型。

(ii) 我们开发了一种全面的关系预测解释方法，可从多条路径和多个角度提供解释。

(iii) 我们在三个传导式和归纳式数据集上评估了 KRST 的性能：WN18RR、FB15k-237 和 NELL-995。在大多数情况下（18 项中的 15 项），KRST 的结果显著优于最先进模型。

相关工作

基于嵌入的方法：这类方法（例如 Com- plEx（Trouillon 等，2017）、ConvE（Dettmers 等，2018）和 TuckER（Balaevi、Allen 和 Hospedales，2019））会在潜在空间中为实体和关系生成嵌入。研究者提出了评分函数，用于训练和评估三元组。最具代表性的基于嵌入的方法是平移方法（例如 TransE（Bordes 等，2013）、TransH（Wang 等，2014）、TransR（Lin 等，2015）和 TransD（Ji 等，2015））。平移模型的核心思想是将寻找有效三元组的过程视为实体沿关系进行平移操作，定义相应的评分函数，然后最小化损失函数，以学习实体和关系的表示（Chen 等，2020年）。
基于图神经网络的方法：这类方法（例如 CompGCN（Vashishth 等人，2019年）和 R-GCN（Schlichtkrull 等人，2018年））在一个节点及其邻居之间传递消息。这些方法利用了图连通性，能够处理一种特定的归纳情形：新实体周围都是已知实体。GRAIL（Teru、Denis 和 Hamilton（2020））旨在处理包含全新实体的知识图谱。然而，正如作者在（Teru、Denis 和 Hamilton（2020））中所述，GRAIL 不具备可解释性。面对稠密知识图谱时，可达实体的数量会随着搜索深度呈指数级增长，这使得基于图神经网络的方法可能无法充分捕捉正确的长路径信息，因此在关系预测任务中的表现可能不佳。
基于路径的方法：此类方法旨在寻找查询头实体与尾实体之间的一条（或多条）逻辑推理路径。PRA（Lao、Mitchell 和 Cohen，2011年）和 AMIE（Gal´arraga 等，2013）会生成霍恩规则（Horn，1951年），其定义范围比路径更广。然而，由于现实世界知识图谱中存在噪声，这些方法的性能有限，因为它们仅适用于精确匹配。DeepPath（Xiong、Hoang 和 Wang，2017年）和 MINERVA（Das 等，2017年）通过强化学习来学习生成路径，并在成功到达目标时给予正奖励。这些方法可应用于新实体，并且具有天然的可解释性。尽管如此，可达实体数量呈指数级增长也给它们带来了挑战，导致奖励稀疏。

基于预训练语言模型的方法：预训练语言模型（PLM）（例如 BERT（Devlin 等，2018年）、GPT（Radford 等，2018年）和 XLNet（Yang 等，2019年））在各种自然语言处理任务中取得巨大成功，展现出处理上下文描述的巨大潜力。KG-BERT（Yao、Mao 和 Luo，2019年）通过微调 BERT 扩展了基于嵌入的方法。利用实体和关系的上下文描述，BERT 模型能够理解一个三元组并输出分类标签。KG-BERT 在归纳设置中表现良好，但由于嵌入难以理解，因此不具备可解释性。与 KG-BERT 不同，BERTRL（Zha、Chen 和 Yan（2022））将基于路径的方法与 BERT 相结合。具体而言，BERTRL 将三元组及其对应的路径转换为句子，并微调预训练的 BERT 来执行序列分类。由于 BERTRL 将所有路径（长度小于L）都作为输入，且不进行过滤，因此其性能可能受限。近期研究表明，句子转换器模型在常见的语义文本相似度和迁移学习任务中优于 BERT（Reimers 和 Gurevych，2019）。此外，用于序列分类的 BERT 需要同时输入两个句子并进行隐式比较，而句子转换器会分别对句子进行编码，从而能够更灵活地比较句子。本文采用句子转换器来提供多方面解释。

预备知识

本节首先介绍常用的逻辑推理路径，然后介绍归纳关系预测的定义和常见设置。

定义1（逻辑推理路径）。给定一个知识图谱 $G = \{(h_i, r_i, t_i) | i = 1、2、3、..、 m\}$、 h_i、 $t_i \in E_G$和 $r_i \in R_G$,

一种可能的逻辑推理路径 $p(h,r,t)$ 及其对应的关系路径 $R_p(h,r,t)$ 定义如下：

$$p(h,r,t) = h \xrightarrow{r_1} e_1 \xrightarrow{r_2} e_2 \xrightarrow{r_3} ... \xrightarrow{r_{n-1}} e_{n-1} \xrightarrow{r_n} t, \quad (3)$$

$$R_p(h,r,t) = (r_1, r_2, ..., r_{n-1}, r_n), \quad (4)$$

where $(h, r_1, e_1), ..., (e_{n-1}, r_n, t) \in G - \{(h,r,t)\}$.

从知识图谱推理的角度来看，关系预测任务可以视为一个逻辑归纳问题，旨在识别归纳式且可解释的逻辑推理路径。由于逻辑推理路径本质上是序列，我们可以轻松地将其转换为句子。然而，在现实世界知识图谱中，很大一部分路径要么不合逻辑，要么依赖实体，因此无法用于归纳关系预测（见定义2）。为解决这一问题，我们定义了关系路径覆盖率和关系路径置信度，以过滤掉不可靠或无意义的路径（分别见定义4和定义5）。

定义2（归纳关系预测）。给定训练图 $G_{train}(E_{G_{train}}, R_{G_{train}})$、测试图$G_{test}(E_{G_{test}}, R_{G_{test}})$和查询三元组 (h_q, r_q, t_q)，如果满足以下条件，则关系预测是归纳式的：

- $E_{G_{train}} \cap E_{G_{test}} = \emptyset$,
- $R_{G_{test}} \subseteq R_{G_{train}}, r_q \in R_{G_{train}}$.

随着不断演进的现实世界知识图谱日益涌现，处理新的关系和实体已成为一种必需。我们关注 GRAIL 引入的归纳式设定（Teru、Denis 和 Hamilton（2020）），其中包含一个训练图、一个测试图以及一系列查询三元组。训练期间只能看到训练图。由于基于图神经网络的方法和基于路径的方法都利用图连通性，因此在测试期间，测试图分别仅用于提取与查询三元组相关的邻居或路径。

方法

我们提出的架构包含三个用于关系预测的步骤，其核心思路是：将三元组及其对应的可靠路径转换为句子后，两者应具有相似的语义。具体而言，KRST 会：1）过滤为模型训练提取的不可靠逻辑推理路径；2）通过句子构建将路径和三元组转换为句子；3）计算语义相似度分数，并据此进行关系预测。

带过滤的路径提取

按照知识推理的思路，逻辑推理路径由其相对于目标三元组的支持证据决定。为了使用正、负样本训练模型，我们从知识图谱中提取正、负目标三元组对应的逻辑推理路径。然而，许多提取出的路径并不可靠。例如，对于正三元组 (Sarah, 职业, 演员) 和负三元组 (Sarah, 职业, 导演)，我们都可以提取出具有相同关系路径（性别, 性别 $^{-1}$, 职业）的路径，如下所示：

$$Sarah \xrightarrow{Gender} Female \xrightarrow{Gender^{-1}} Kathryn \xrightarrow{Profession} Actor, \quad (5)$$

算法1：头实体关系路径覆盖率

输入：知识图谱 G、查询三元组 (h,r,t)、关系路径长度 $|R_{p'}|$

输出：覆盖率 cover

```
// 创建一个嵌套字典
// 维度1：实体名称
// 维度2：深度
路径数 = 字典(Dictionary());
计数 = 0; 支持度 = 0
// 迭代更新字典
对于 i ∈ 范围(0, |R_{p'}|) 执行
    newNum = 路径数.copy()
    for e ∈ 路径数 do
        对 v ∈ G[e] 执行
            对于深度 ∈ 路径数[e] 执行
                新数[v][dpth + 1] = 路径数[e][dpth]
    路径数 = 新数.copy()
// 汇总路径数量
对 e ∈ 路径数执行
    如果 e == t 那么
        支持度+ = 路径数[e][|R_{p'}|]
    计数+ = 路径数[e][|R_{p'}|]
覆盖度 = 支持度/计数
返回覆盖度
```

算法2：头实体关系路径置信度

输入：知识图谱 G，查询三元组（h,r,t），逻辑推理路径 P，关系路径 $R_{p'}$

输出：置信度分数 conf

```
q =队列()
q.push((h, 0)); 计数 = 0; 支持度 = 0
// 广度优先搜索
当 q 不为空时执行
    u, l = q.弹出()
    如果 l == |R_{p'}|，则
        如果 u == t，则
            支持度 = 支持度 + 1
        计数 = 计数 + 1
        继续
    对于 v 在 G[u] 中执行
        // 检查关系路径
        如果 G[u][v]['关系'] == R_{p'}[l] 那么
            q.push((v, l + 1))
conf = 支持度/计数
返回置信度
```

$$Sarah \xrightarrow{Gender} Female \xrightarrow{Gender^{-1}} Alyson \xrightarrow{Profession} Director. \quad (6)$$

路径 (5) 和 (6) 相似，对区分职业的贡献很小。因此，我们认为关系路径为 (Gender, Gender^{-1}, Profession) 的路径是不可靠的。尽管不可靠的路径应从模型训练中排除，但由于它们与目标三元组相似，可能会获得较高的相似度分数，因此被误认为是可靠路径。例如，另一条路径是

三元组 (Sarah, 职业, 演员) 如下:

$$Sarah \xrightarrow{Perform} SmallSoldiers \xrightarrow{Company^{-1}} DreamWorks \xrightarrow{Company} TheIsland \xrightarrow{Perform^{-1}} Sean \xrightarrow{Profession} Actor. \quad (7)$$

这条路径相对路径 (5) 更可靠，因为它在预测时涉及同一家公司中的人物。然而，将路径转换为句子后，预训练语言模型会认为路径 (5) 更相似（更短，且包含目标三元组中的词语）。由于 (Sarah,Profession,Actor) 是一个正三元组，模型训练会强化这种偏差，导致路径 (5) 在用于训练后获得更高的分数。这些不可靠的路径显著限制了模型性能。为了将提取出的不可靠路径排除在训练之外，我们进行路径筛选。下面将介绍如何识别不可靠的逻辑推理路径。

定义 3（关系路径支持度）。给定三元组 (h, r, t) 和关系路径 $R_{p'} = (r_1', r_2', ..., r_{m-1}', r_m')$，$R_{p'}$ 的支持度定义如下:

$$supp_{R_{p'}}(h, r, t) = \#p(h, r, t), R_p(h, r, t) = R_{p'}, \quad (8)$$

其中，$\#p(h, r, t)$ 表示在 G 上相对于 (h, r, t) 的逻辑推理路径数量。

关系路径支持度 $supp_{R'}(h, r, t)$ 用于衡量包含与 $R_{p'}$ 相同关系、位于 h 和 t之间的路径数量。支持度分数越大，$R_{p'}$ 在 h 和 t之间的路径中就越常见。然而，关系路径支持度表示的是一个无界值。连通性更好的实体对通常具有更大的支持度。为了进一步评估支持度的比例，我们分别定义关系路径覆盖率和关系路径置信度。

定义 4（关系路径覆盖率）。给定三元组 (h, r, t) 和关系路径 $R_{p'}$，$R_{p'}$ 的头部和尾部关系路径覆盖率定义如下:

$$cover^{h}_{R_{p'}}(h, r, t) = \frac{supp_{R_{p'}}(h, r, t)}{\#p(h, r, t'), |p(h, r, t')| = |R_{p'}|}, \quad (9)$$

$$cover^{t}_{R_{p'}}(h, r, t) = \frac{supp_{R_{p'}}(h, r, t)}{\#p(h', r, t), |p(h', r, t)| = |R_{p'}|}, \quad (10)$$

其中，$|\cdot|$ 表示路径的长度。

在（9）和（10）中，分别采用以 h 为起点或以 t 为终点、长度为 $|R_{p'}|$ 的路径数作为分母。

定义 5（关系路径置信度）。给定一个三元组（h、r、t），关系路径 $R_{p'}$ 的头部和尾部关系路径置信度定义如下:

$$conf^{h}_{R_{p'}}(h, r, t) = \frac{supp_{R_{p'}}(h, r, t)}{\sum_{t' \in E_G - \{h\}} supp_{R_{p'}}(h, r, t')}, \quad (11)$$

$$conf^{t}_{R_{p'}}(h, r, t) = \frac{supp_{R_{p'}}(h, r, t)}{\sum_{h' \in E_G - \{t\}} supp_{R_{p'}}(h', r, t)}. \quad (12)$$

算法 3：路径提取

输入：知识图谱 G、查询三元组（h，r，t）、过滤阈值 α、过滤器函数 $f()$、最大搜索深度 L、最大路径数 M

输出：提取路径列表 P

```
// 初始化搜索队列和状态列表
q = 队列(); 已访问 = 列表(); 前驱 = 列表()
q.push((h, 0))
已访问[h] = 是
// 广度优先搜索
while q 不为空 do
    u, l = q.pop()
    // 检查搜索深度是否超过 L
    if l >= L then
        continue
    // 优先处理频率较低的关系
    for v in G[u] 按频率排序 G[u][v]['关系'] do
        如果 v == t那么
            如果 u == h&&G[u][v]['关系'] == r 那么
                继续
            p =生成路径(prev h, t)
            如果 f(p) ≥ α 那么
                P.添加(p)
        否则，如果已访问[v] == 0 则
            q.push((v, l + 1)); 已访问(v) = 1; 前驱(v) = u
        // 生成的路径足够时提前退出
        如果 |P| > M 则
            跳出
    如果 |P| > M 则
        中断
返回 P
```

在公式 (11) 和 (12) 中，分别采用从 h 出发或以 t 结束、关于关系路径 $R_{p'}$ 的所有可达实体总数作为分母。

关系路径覆盖率衡量具有相同源（或目标）和长度的所有路径中的比例，而关系路径置信度衡量满足目标关系路径的所有实体中的比例。为了说明关系路径置信度的有效性，我们再来看路径（5）和（7）。路径（7）的关系路径置信度分数远高于路径（5）。这是因为在关系路径（Gender,Gender^{-1},Profession）中，Gender^{-1} 可以指向多个具有不同职业的人。因此，满足路径（5）关系路径的路径数量显著多于满足路径（7）的数量，导致路径（5）的关系路径置信度分数相对较低。这两个指标在头实体情形下的实现分别见算法1和算法2。借助反向图，它们可以轻松转换为尾实体情形。

在路径提取中，我们采用广度优先搜索，最大搜索深度为 L，每个三元组的最大路径数为 M。设置这些参数是为了避免生成数量过多的路径。此外，长度过长的路径很可能不合乎逻辑。为了确保为每个三元组生成足够数量的路径，以便有效进行模型训练，我们同步

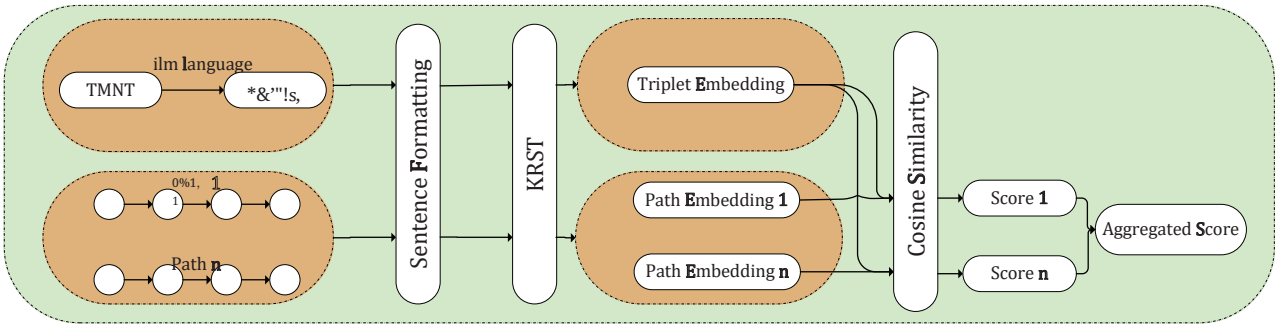


图2：KRST中关系预测的整体架构。

执行路径筛选和路径提取（见算法 3）。

句子构建

为利用句子转换器的预训练参数，KRST 将路径和三元组转换为句子。我们的主要考量如下：

实体描述的选择。各种数据集中都有较长（平均超过 20 个词）和较短的文本实体描述。较长的描述在由预训练语言模型建模时，通常会导致实体与关系之间的句子描述不平衡，因此在归纳场景中的效果较差。此外，使用较长描述转换得到的句子通常会超过预训练语言模型的最大序列长度，因此需要截断，导致内容不完整。因此，KRST 仅采用较短的描述。

逆关系。为了让 KRST 能够提供用于位置编码的序列顺序，路径中的实体和关系顺序在转换为句子后应予以保留。一种直接的方法是按照路径中实体和关系的顺序排列描述。然而，路径中会出现逆关系（例如，路径（5）中的 Gender^{-1}），而且这些关系没有对应的描述。根据经验，知识图谱中的系统性描述通常以头实体相关的描述开头，以尾实体相关的描述结尾。因此，对于逆关系，我们选择将序列中关系词语的顺序反转。此外，这样做还能保留关系与其逆关系之间的相似性，使模型更容易理解对称关系（例如朋友、配偶和队友）。

描述拼接。为了获得一个完整的句子，我们需要拼接路径中实体和关系的描述。自然语言模式能让句子更接近人类的表达方式。三元组（Sarah，职业，演员）的示例如下：

> 问题：莎拉·米歇尔·盖拉的职业是什么？正确答案是演员吗？

然而，我们的初步结果表明，将内容组织成完整句子并不会带来比仅用分号将实体和关系组合起来更好的表现，如下所示：

> 莎拉·米歇尔·盖拉；人物职业；演员

这是因为后一种方法能更好地保留预训练语言模型中路径的顺序和相对位置。因此，我们选择使用分号进行描述拼接。例如，路径（5）对应的句子表述如下：

> *Sarah Michelle Gellar; person gender; Female; gender person; Alyson Hannigan; person profession; Actor*

KRST预测

完成句子构建后，KRST便能够生成嵌入并进行关系预测。图2展示了整体模型架构。

对于每个正查询三元组，都会相应地提取多条可靠的逻辑推理路径。将三元组及其对应路径格式化为句子后，它们具有相似的语义，并通过余弦相似度进行度量：

$$cos(s_1, s_2) = \frac{s_1 \cdot s_2}{||s_1||_2 \cdot ||s_2||_2}, \tag{13}$$

其中，s_1 和 s_2 是对应的句子嵌入。

KRST 为每个三元组最多提取 $|M|$ 条路径，并将路径和三元组转换为嵌入。计算每条路径与对应三元组之间的相似度分数，并将分数最高的路径视为最合理的关系预测依据。因此，我们使用所有路径中的最高分数作为每个三元组的分数：

$$score(h, r, t) = \max_{p \in P} \{cos(s((h, r, t)), s(p))\}, \tag{14}$$

其中，P 表示与三元组（h，r，t）对应的路径集合，s（·）表示用于三元组和路径的嵌入函数（即 KRST）。

训练期间，负三元组的标签设为 -1。如前所述，常用的二元损失函数（例如交叉熵和合页损失）要求负分数接近 -1（或 0）。这在我们的场景中并不合适，因为三元组与路径之间的不匹配对不一定彼此垂直。为

表1：KRST 的超参数值

描述	值
随机种子	42
训练和验证样本	5
测试样本	50
训练集的路径提取数，验证集和测试集（M）	3
路径提取搜索深度（L）	5
支持度类型（1：覆盖率，2：置信度）。	1 2
支持度阈值（覆盖度$_{R_p}$，or conf $_{R_p}$）	10^{-5} $5*10^{-3}$
训练轮数	30
句子转换器学习率	10^{-5}

减轻损失惩罚，我们使用余弦嵌入损失函数：

$$\mathscr{L}(s_1, s_2, y) = \begin{cases} 1 - cos(s_1, s_2), & y = 1, \\ \max(0, cos(s_1, s_2) - margin), & y = -1, \end{cases} \tag{15}$$

其中，间隔 $\in (-1,1)$，$y \in \{1, -1\}$ 表示标签。公式15使正分数接近1，从而最大化相似语义。另一方面，负分数仅受限于小于间隔。

实验

我们在三种不同设置下评估 KRST 的性能：传导式、归纳式和少样本。随后，我们通过对 KRST 生成的路径嵌入进行聚类，展示多方面综合解释。我们在配备 16GB RAM 的 Tesla V100 GPU 上，使用最先进水平的句子转换器（all-mpnet-base-v2[2]）实现 KRST[1]。在不同实验中，我们不会微调 KRST 的超参数值，而是对所有数据集在传导式、归纳式和少样本设置下统一采用同一组超参数值（见表1）。

遵循 (Teru、Denis 和 Hamilton（2020）) 和 (Zha、Chen 和 Yan（2022）) 中进行的评估任务，本文衡量一个正三元组在 49 个负三元组中的排名和命中率。我们只随机生成负三元组，并将其用于训练和验证。为公平比较，我们使用 (Zha、Chen 和 Yan（2022）) 提供的负三元组进行测试。

数据集

为了评估 KRST 的传导式和归纳式性能，我们使用了 Zha、Chen 和 Yan（2022）采用的全部三个数据集，这些数据集由 Teru、Denis 和 Hamilton（2020年）引入[3]。这些数据集常用于各种归纳式方法，分别是 WN18RR、FB15k-237，以及 NELL-995。在归纳式设定下，训练实体与测试实体没有重叠。

表2：数据集统计数据

数据集	G	$\lvert R_G\rvert$	$\lvert E_G\rvert$	#三元组
WN18RR	训练集	9	2746	6670
	训练集-2000	9	1970	2002
	训练集-1000	9	1362	1001
	传导式测试集	7	962	638
	归纳式测试集	8	922	1991
FB15k-237	训练集	180	1594	5223
	训练集-2000	180	1280	2008
	训练集-1000	180	923	1027
	传导式测试集	102	550	492
	归纳式测试集	142	1093	2404
NELL-995	训练集	88	2564	10063
	训练集-2000	88	1346	2011
	训练集-1000	88	893	1020
	传导式测试集	60	1936	968
	归纳式测试集	79	2086	6621

对于少样本评估，我们使用（Zha、Chen 和 Yan，2022年）中提供的相应少样本传导式数据集和少样本归纳式数据集。本文所用所有数据集的统计数据见表2。

传导式与归纳式关系预测

在传导式情形下，我们会为所有训练、验证和测试三元组从训练图中提取路径。然而，在归纳式情形下，由于测试所用的实体未出现在训练图中（见定义 2），因此无法从训练图中获取测试三元组的路径。相反，我们使用（Teru、Denis 和 Hamilton（2020））提供的归纳图来提取路径。路径提取的详细配置见表1。

我们将 KRST 与最先进的归纳式方法、最先进的基于嵌入的方法以及最先进的强化学习方法进行性能比较。表3展示了传导式与归纳式关系预测的结果。与最先进的方法相比，KRST 在平均倒数排名（5 项中有 5 项）和命中率@1（6 项中有 4 项）指标上取得了最佳表现。具体而言，KRST 在 WN18RR 的传导式情形下（平均倒数排名为 +0.216，命中率@1 为 +18.0%）、WN18RR 的归纳式情形下（平均倒数排名为 +0.103，命中率@1 为 +6.6%）以及 FB15k-237 的归纳式情形下（平均倒数排名为 +0.111，命中率@1 为 +5.9%）均取得了显著提升。

在 KRST 方法中，大多数最佳结果由采用关系路径置信度的 KRST 取得（ 12 种情况中有 8 种）。仅在 12 种情况中的 1 种，采用无过滤器的 KRST 表现最佳。因此，我们通过实证表明，过滤（尤其是关系路径置信度过滤）能够提升模型性能。关系路径覆盖率表现相对较差，是因为较长的路径通常会导致可达实体数量呈指数级增长，使关系路径覆盖率过滤倾向于选择较短的路径。然而，短路径未必可靠。因此，KRST 的性能受到了未经关系路径覆盖率过滤的非可靠路径的限制。

[1] https://github.com/ZhixiangSu/KRST[2]
https://huggingface.co/sentence-transformers/all-mpnet-base-v2[3]
https://github.com/kkteru/grail

表3：传导式与归纳式关系预测结果

		传导式			归纳式		
		WN18RR	FB15k-237	NELL-995	WN18RR	FB15k-237	NELL-995
MRR	RuleN	0.669	0.674	0.736	0.780	0.462	0.710
	GRAIL	0.676	0.597	0.727	0.799	0.469	0.675
	MINERVA	0.656	0.572	0.592	-	-	-
	TuckER	0.646	0.682	0.800	-	-	-
	KG-BERT	-	-	-	0.547	0.500	0.419
	BERTRL	0.683	0.695	0.781	0.792	0.605	**0.808**
	KRST（无过滤）	0.881	0.671	0.730	0.883	0.713	0.753
	KRST（覆盖率）	0.897	0.709	**0.803**	**0.902**	0.704	0.696
	KRST（置信度）	**0.899**	**0.720**	0.800	0.890	**0.716**	0.769
命中率@1	RuleN	0.646	0.603	0.636	0.745	0.415	0.638
	GRAIL	0.644	0.494	0.615	0.769	0.390	0.554
	MINERVA	0.632	0.534	0.553	-	-	-
	TuckER	0.600	0.615	**0.729**	-	-	-
	KG-BERT	-	-	-	0.436	0.341	0.244
	BERTRL	0.655	0.620	0.686	0.755	0.541	**0.715**
	KRST（无过滤）	0.807	0.576	0.618	0.803	**0.602**	0.633
	KRST（覆盖率）	0.831	0.624	0.692	**0.835**	0.573	0.554
	KRST（置信度）	**0.835**	**0.639**	0.694	0.809	0.600	0.649

表4：少样本传导式与归纳式关系预测结果

		传导式						归纳式					
		WN18RR		FB15k-237		NELL-995		WN18RR		FB15k-237		NELL-995	
		1000	2000	1000	2000	1000	2000	1000	2000	1000	2000	1000	2000
MRR	RuleN	0.567	0.625	0.434	0.577	0.453	0.609	0.681	0.773	0.236	0.383	0.334	0.495
	GRAIL	0.588	0.673	0.375	0.453	0.292	0.436	0.652	0.799	0.380	0.432	0.458	0.462
	MINERVA	0.125	0.268	0.198	0.364	0.182	0.322	-	-	-	-	-	-
	TuckER	0.258	0.448	0.457	0.601	0.436	0.577	-	-	-	-	-	-
	KG-BERT	-	-	-	-	-	-	0.471	0.525	0.431	0.460	0.406	0.406
	BERTRL	0.662	0.673	0.618	0.667	0.648	0.693	0.765	0.777	0.526	0.565	0.736	**0.744**
	KRST（置信度）	**0.871**	**0.882**	**0.696**	**0.701**	**0.743**	**0.781**	**0.886**	**0.878**	**0.679**	**0.680**	**0.745**	0.738
命中率@1	RuleN	0.548	0.605	0.374	0.508	0.365	0.501	0.649	0.737	0.207	0.344	0.282	0.418
	GRAIL	0.489	0.633	0.267	0.352	0.198	0.342	0.516	0.769	0.273	0.351	0.295	0.298
	MINERVA	0.106	0.248	0.170	0.324	0.152	0.284	-	-	-	-	-	-
	TuckER	0.230	0.415	0.407	0.529	0.392	0.520	-	-	-	-	-	-
	KG-BERT	-	-	-	-	-	-	0.364	0.404	0.288	0.317	0.236	0.236
	BERTRL	0.621	0.637	0.517	0.583	0.526	0.582	0.713	0.731	0.441	0.493	0.622	0.628
	KRST（置信度）	**0.790**	**0.810**	**0.611**	**0.602**	**0.628**	**0.678**	**0.811**	**0.793**	**0.537**	**0.524**	**0.637**	**0.629**

能受到未被关系路径覆盖率过滤掉的不可靠路径的限制。

少样本关系预测

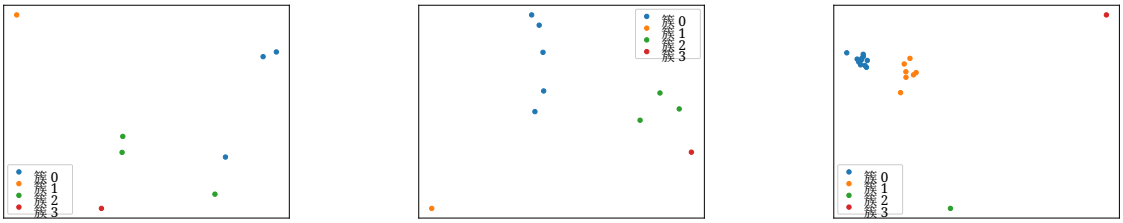
在少样本设置中，仅提供整个数据集的子集用于训练，因此我们采用与整个数据集相同的路径提取方法。具体而言，对于传导式情形，训练、验证和测试路径均从整个训练图中提取。对于归纳式情形，归纳式训练所用的路径从整个训练图中提取，而归纳式验证和归纳式测试所用的路径则均从归纳图中提取。

由于采用关系路径置信度的 KRST 在整个数据集上表现最佳，因此我们将其用于所有少样本设置。如表4所示，采用关系路径置信度的 KRST 在 12 种情况下有 11 种优于最先进的方法。MRR 和命中率@1 的传导式与归纳式平均提升分别为 0.119 和 0.082（+3.1% 和 +5.0%）。此外，少样本样本数为 1000 和 2000 时的性能差距较小（6 种情况下有 5 种的 MRR 差距小于 0.015）。这些结果表明 KRST 具有很强的泛化能力，只需较少的训练样本即可达到相当的性能。

多方面解释

如前所述，KRST 能够为关系预测结果提供多方面解释。这是因为 KRST 为每条路径生成一个嵌入，使我们能够对路径之间的关系进行定量分析。尽管这些路径应当与给定的查询三元组相似，但它们彼此之间不一定相似。通过将它们分组成不同的簇，我们可以提供全面的多方面解释。

图 3 展示了对三个不同三元组应用线性判别分析（LDA）后，在降维空间中的聚类结果。簇是使用 K 均值算法生成的。如图 3(a) 所示



(a) 三元组（大西洋城，电影类型，剧情）的簇。

(b) 三元组（Kelsey Grammer，人物语言，英语语言）。

(c) 三元组（加利福尼亚州，位置包含，卡尔斯巴德）的簇。

图 3：多方面解释的聚类结果。

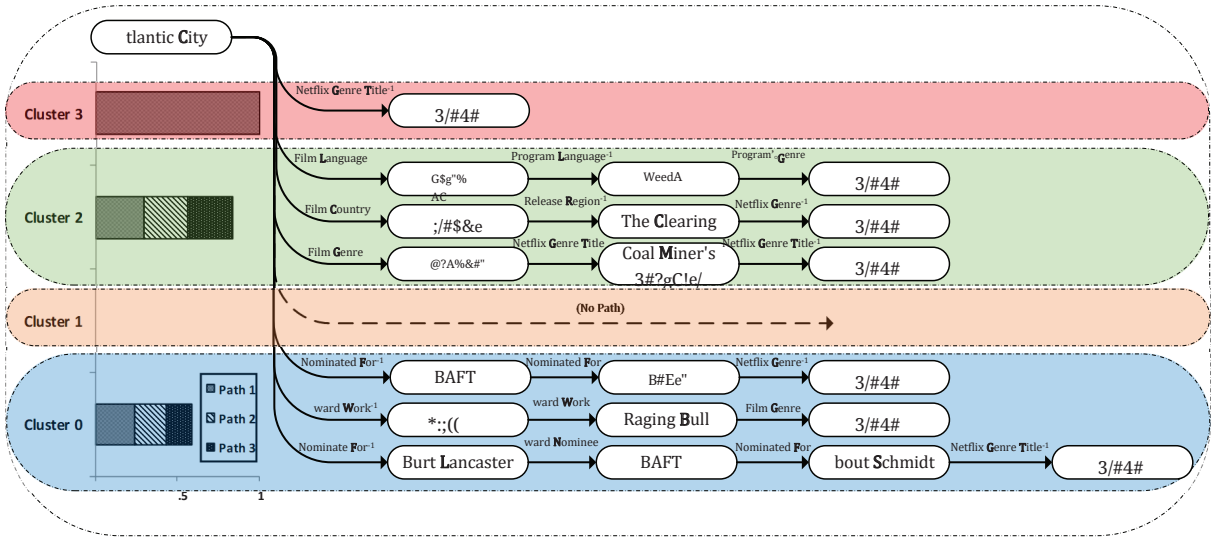


图4：按 si 聚类的路径　　多方面解释的相似度分数　　三元组（大西洋城，电影类型，剧情）。

图 4 中，KRST 基于相似度分数（经最小-最大缩放预处理）成功为三元组（大西洋城，电影类型，剧情）提供了多方面解释。归入各簇的路径如图4所示，并在图 3(a) 中以相同颜色呈现。如图所示，簇 3 中唯一的路径展示了 Netflix 提供的外部知识所给出的解释：大西洋城是一部可在 Netflix 上观看的剧情片。该解释直观且令人信服，因此被认为最为可靠（分数为 1.0）。对于簇 2 中的路径，解释借助具有相同属性（如语言和国家）的相似影片给出，这与人类的类比推理相似。簇 0 则通过三条路径，利用与大西洋城相关的获奖或提名信息来解释目标三元组，这是一项重要且独特的侧信息。KRST 认为簇 2 和簇 0 的说服力相对较弱，因此分别赋予较低的平均分数（0.837 和 0.593）。我们还输入了一条空路径作为对照，该路径单独构成簇 1。KRST 正确地为其赋予了最低相似度分数 0。总之，基于图——在图4中，我们可以轻松地看出 Netflix 平台、相似影片和奖项等相应方面的不同解释。补充表6和表5展示了更多用于多方面解释的聚类路径示例。

结论

本文提出了一种名为 KRST 的新型架构，在大多数传导式与归纳式关系预测任务中（18项任务中的15项）优于最先进模型。此外，我们对 KRST 生成的嵌入进行聚类，并提供全面的多方面解释。尽管如此，KRST 是一种基于 BERT 的模型，需要相对较大的内存使用量和计算资源。今后，我们计划通过提出一种参数高效模型来解决这一计算密集型问题。

致谢

本研究部分由中国山东大学人工智能联合研究中心（C-FAIR）以及 NTU-WeBank联合

金融科技研究中心（奖项编号：NWJ-2020-010）、新加坡南洋理工大学提供支持。本研究还部分获得了国家重点研发计划（编号：2021YFF0900800）、国家自然科学基金委员会（编号：91846205）以及山东省重点研发计划（重大科技创新工程）（编号：2021CXGC010108）的支持。

参考文献

Balaevi, I.; Allen, C.; and Hospedales, T. M. 2019. TuckER：用于知识图谱补全的张量分解。arXiv 预印本 arXiv:1901.09590。Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. 用于建模多关系数据的嵌入翻译。载于《神经信息处理系统进展》论文集，1–9。Chen, Z.; Wang, Y.; Zhao, B.; Cheng, J.; Zhao, X.; and Duan, Z. 2020. 知识图谱补全：综述。《IEEE Access》，8：192435–192456。Das, R.; Dhuliawala, S.; Zaheer, M.; Vilnis, L.; Durugkar, I.; Krishnamurthy, A.; Smola, A.; and McCallum, A. 2017. 去散步并找到答案：使用强化学习对知识库中的路径进行推理。arXiv 预印本 arXiv:1711.05851。Dettmers, T.; Minervini, P.; Stenetorp, P.; and Riedel, S. 2018. 卷积二维知识图谱嵌入。载于 AAAI 人工智能会议论文集，1811–1819。Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT：用于语言理解的深度双向变换器预训练。arXiv 预印本 arXiv:1810.04805。Gal´arraga, L. A.; Teflioudi, C.; Hose, K.; and Suchanek, F. 2013. AMIE：在本体知识库中基于不完整证据进行关联规则挖掘。载于国际万维网会议论文集，413–422。Horn, A. 1951年. 关于代数直和所满足的句子¹。《符号逻辑杂志》，14–21。Ji, G.; He, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. 通过动态映射矩阵进行知识图谱嵌入。载于计算语言学协会年会暨自然语言处理国际联合会议论文集，687–696。Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; and Philip, S. Y. 2021. 知识图谱综述：表示、获取与应用。《IEEE神经网络与学习系统汇刊》，494–514。Lao, N.; Mitchell, T.; and Cohen, W. 2011. 大规模知识库中的随机游走推理与学习。载于自然语言处理实证方法会议论文集，529–539。Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015. 学习用于知识图谱补全的实体与关系嵌入。载于 AAAI 人工智能会议论文集，2181–2187。Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. 通过生成式预训练提升语言理解。Reimers, N.; and Gurevych, I. 2019. Sentence-BERT：使用孪生 BERT 网络生成句子嵌入。arXiv预印本 arXiv:1908.10084。Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Berg, R. v. d.; Titov, I.; and Welling, M. 2018. 使用图卷积网络对关系数据建模。载于欧洲语义网会议论文集，593–607。Teru, K.; Denis, E.; and Hamilton, W. 2020. 通过子图推理进行归纳式关系预测。载于国际机器学习会议论文集，9448–9457。Trouillon, T.; Dance, C. R.; Welbl, J.; Riedel, S.; Gaussier, .; and Bouchard, G. 2017. 通过复张量分解进行知识图谱补全。arXiv预印本 arXiv:1702.06879。Vashishth, S.; Sanyal, S.; Nitin, V.; and Talukdar, P. 2019. 基于组合的多关系图卷积网络。arXiv预印本 arXiv:1911.03082。Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. 通过在超平面上进行平移来嵌入知识图谱。载于AAAI人工智能会议论文集。Xiong, W.; Hoang, T.; and Wang, W. Y. 2017. DeepPath：一种用于知识图谱推理的强化学习方法。arXiv预印本 arXiv:1707.06690。Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet：面向语言理解的广义自回归预训练。载于神经信息处理系统进展论文集，1–11。Yao, L.; Mao, C.; and Luo, Y. 2019. KG-BERT：用于知识图谱补全的 BERT。arXiv预印本 arXiv:1909.03193。Zha, H.; Chen, Z.; and Yan, X. 2022. 基于 BERT 的归纳式关系预测。载于AAAI人工智能会议论文集，5923–5931。

附录

更多关于多方面解释的内容

我们提供了更多关于多方面解释的案例研究，见补充表5和6。不同簇中的路径以水平线分隔。补充表5和6中所示的聚类路径可视化结果，分别见提交论文中的图3(b)和图3(c)。

表 5：三元组（凯尔西·格拉默，人物语言，英语）的聚类路径及其置信度分数

三元组	Path	分数
凯尔西·格拉默 $\xrightarrow{\text{人物语言}}$ 英语	凯尔西·格拉默 $\xrightarrow{\text{结合类型}}$ 婚姻 $\xrightarrow{\text{配偶}}$ Diane Sawyer $\xrightarrow{\text{领域研究}}$ 英语	0.577
	凯尔西·格拉默 $\xrightarrow{\text{奖项提名人}}$ Peter Casey $\xrightarrow{\text{人物职业}}$ 编剧 $\xrightarrow{\text{职业人物}}$ Diane Sawyer $\xrightarrow{\text{领域研究}}$ 英语	0.580
	凯尔西·格拉默 $\xrightarrow{\text{奖项提名人}}$ David Angell $\xrightarrow{\text{人物职业}}$ 编剧 $\xrightarrow{\text{职业人物}}$ Diane Sawyer $\xrightarrow{\text{领域研究}}$ 英语	0.561
	凯尔西·格拉默 $\xrightarrow{\text{人物性别}}$ 男性 $\xrightarrow{\text{性别人物}}$ 哲学 $\xrightarrow{\text{领域研究}}$ 英语	0.653
	凯尔西·格拉默 $\xrightarrow{\text{人物宗教}}$ 天主教 $\xrightarrow{\text{宗教人物}}$ 布列塔尼·斯诺 $\xrightarrow{\text{配音表演语言}}$ 英语	1.000
	凯尔西·格拉默 $\xrightarrow{\text{奖项提名}}$ SAGA $\xrightarrow{\text{奖项提名人}}$ 威尔·阿奈特 $\xrightarrow{\text{配音表演语言}}$ 英语	0.484
	凯尔西·格拉默 $\xrightarrow{\text{奖项提名人}}$ 杰伊·科根 $\xrightarrow{\text{电视节目}}$ 《辛普森一家》 $\xrightarrow{\text{电视节目语言}}$ 英语	0.774
	凯尔西·格拉默 $\xrightarrow{\text{奖项提名}}$ 托尼奖最佳音乐剧男主角奖 $\xrightarrow{\text{奖项提名人}}$ 蒂姆·克里 $\xrightarrow{personlanguage}$ 英语	0.780
	凯尔西·格拉默 $\xrightarrow{\text{学生教育毕业生}}$ 茱莉亚学院 $\xrightarrow{\text{教育学生}}$ 蒂姆·布雷克·尼尔森 $\xrightarrow{\text{演员电影表演电影}}$ 《细细的红线》 $\xrightarrow{\text{电影语言}}$ 英语	0.572
	（无路径）	0.000

表 6：三元组（加利福尼亚州，位置包含，卡尔斯巴德）的聚类路径及置信度分数

三元组	Path	分数
加利福尼亚州 —地点包含→ 卡尔斯巴德	加利福尼亚州 —州地点书目→ 卡尔斯巴德	0.390
	加利福尼亚州 —地点包含→ 圣迭戈县 —地点包含→ 卡尔斯巴德	1.000
	加利福尼亚州 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.801
	加利福尼亚州 —地点包含→ 山景城 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.639
	加利福尼亚州 —地点包含→ 圣克拉拉 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.649
	加利福尼亚州 —地点包含→ 拉霍亚 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.575
	加利福尼亚州 —地点包含→ 圣佩德罗 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.621
	加利福尼亚州 —居住地 人物→ Sharon Osbourne —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.617
	加利福尼亚州 —州位置 书目→ 桑尼维尔 —地点相邻→ 山景城 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.310
	加利福尼亚州 —政府职位→ 国会议员 —政府政治人物→ 詹姆斯·麦迪逊 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.327
	加利福尼亚州 —宗教统计区域→ 基督教 —宗教人物→ 威尔·史密斯 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.549
	加利福尼亚州 —居住地 人物→ 维维卡·A·福克斯 —奖项得主→ 威尔·史密斯 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.449
	加利福尼亚州 —邮寄总部组织→ 索尼影视娱乐 —奖项提名人→ Jay-Z —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.344
	加利福尼亚州 —代表的选区→ *104*届美国国会 —政府政治人物→ 约翰·克里 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.356
	加利福尼亚州 —邮寄总部组织→ 加利福尼亚州立大学 —教育学生毕业生→ 凯文·科斯特纳 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.361
	加利福尼亚州 —邮寄总部组织→ 皮克斯 —获提名→ 《怪兽电力公司》 —仪式地点→ 卡尔斯巴德	0.356
	电影发行地区 —仪式地点→ 法国 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	
	加利福尼亚州 —埋葬地点→ 克拉克·盖博 —人物性别→ 男性 —性别人物→ 杰夫·莫斯 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.562
	加利福尼亚州 —埋葬地点→ 塞缪尔·戈德温 —人物性别→ 男性 —性别人物→ 杰夫·莫斯 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.552
	加利福尼亚州 —居住地人物→ 弗兰克·盖里 —人物性别→ 男性 —性别人物→ 杰夫·莫斯 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.491
	加利福尼亚州 —邮寄总部组织→ 基因泰克 —服务语言→ 英语 —语言人物→ Aidan Gillen —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.296

我们采用一种自动机 种方法是，如果提取了 20 条路径，我们就不会加入表示为“（无路径）”的空路径，用于聚类（如 源代码中所反映的），本表展示的就是这种情况。

arXiv:2301.01664v2 [cs.CL] 1 May 2023

附录

关于多方面解释的更多内容

我们提供了更多多方面解释的案例研究，见补充表1和2。不同簇中的路径用水平线分隔。补充表1和2中聚类路径的可视化结果分别见投稿论文中的图3(b)和图3(c)。

补充表1：三元组（凯尔西·格拉默，人物语言，英语）对应的聚类路径及其置信度分数

三元组	Path	分数
凯尔西·格拉默 —人物语言→ 英语	凯尔西·格拉默 —结合类型→ 婚姻 —配偶→ Diane Sawyer —领域研究→ 英语	0.577
	凯尔西·格拉默 —奖项提名人→ Peter Casey —人物职业→ 编剧 —职业人物→ Diane Sawyer —领域研究→ 英语	0.580
	凯尔西·格拉默 —奖项提名人→ David Angell —人物职业→ 编剧 —职业人物→ Diane Sawyer —领域研究→ 英语	0.561
	凯尔西·格拉默 —人物性别→ 男性 —性别人物→ 哲学 —领域研究→ 英语	0.653
	凯尔西·格拉默 —人物 宗教→ 天主教 —宗教 人物→ 布列塔尼·斯诺 —配音表演 语言→ 英语	1.000
	凯尔西·格拉默 —奖项提名→ SAGA —奖项提名人→ 威尔·阿奈特 —配音表演语言→ 英语	0.484
	凯尔西·格拉默 —奖项提名人→ 杰伊·科根 —电视节目→ 《辛普森一家》 —电视节目语言→ 英语	0.774
	凯尔西·格拉默 —奖项提名→ 托尼奖最佳音乐剧男主角奖 —奖项提名人→ 蒂姆·克里 —*personlanguage*→ 英语	0.780
	凯尔西·格拉默 —学生教育毕业生→ 茱莉亚学院 —教育学生→ 蒂姆·布雷克·尼尔森 —演员电影表演电影→ 《细细的红线》 —电影语言→ 英语	0.572
	（无路径）	0.000

补充表2：三元组（加利福尼亚州，位置包含，卡尔斯巴德）的聚类路径及置信度分数

三元组	Path	分数
加利福尼亚州 —地点包含→ 卡尔斯巴德	加利福尼亚州 —州地点书目→ 卡尔斯巴德	0.390
	加利福尼亚州 —地点包含→ 圣迭戈县 —地点包含→ 卡尔斯巴德	1.000
	加利福尼亚州 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.801
	加利福尼亚州 —地点包含→ 山景城 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.639
	加利福尼亚州 —地点包含→ 圣克拉拉 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.649
	加利福尼亚州 —地点包含→ 拉霍亚 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.575
	加利福尼亚州 —地点包含→ 圣佩德罗 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.621
	加利福尼亚州 —居住地 人物→ Sharon Osbourne —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.617
	加利福尼亚州 —州位置 书目→ 桑尼维尔 —地点相邻→ 山景城 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.310
	加利福尼亚州 —政府职位→ 国会议员 —政府政治人物→ 詹姆斯·麦迪逊 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.327
	加利福尼亚州 —宗教统计区域→ 基督教 —宗教人物→ 威尔·史密斯 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.549
	加利福尼亚州 —居住地 人物→ 维维卡·A·福克斯 —获奖者→ 威尔·史密斯 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.449
	加利福尼亚州 —邮寄总部组织→ 索尼影视娱乐 —奖项提名人→ Jay-Z —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.344
	加利福尼亚州 —代表的选区→ *104*届美国国会 —政府政治人物→ 约翰·克里 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.356
	加利福尼亚州 —邮寄总部组织→ 加利福尼亚州立大学 —教育学生毕业→ 凯文·科斯特纳 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.361
	加利福尼亚州 —邮寄总部组织→ 皮克斯 —获得提名→ 《怪兽电力公司》 —电影发行地区→ 法国 —仪式地点→ 婚姻 —仪式地点→ 卡尔斯巴德	0.356
	加利福尼亚州 —埋葬地点→ 克拉克·盖博 —人物性别→ 男性 —性别人物→ 杰夫·莫斯 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.562
	加利福尼亚州 —埋葬地点→ 塞缪尔·戈德温 —人物性别→ 男性 —性别人物→ 杰夫·莫斯 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.552
	加利福尼亚州 —居住地人物→ 弗兰克·盖里 —人物性别→ 男性 —性别人物→ 杰夫·莫斯 —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.491
	加利福尼亚州 —邮寄总部组织→ 基因泰克 —服务语言→ 英语 —语言人物→ Aidan Gillen —结合类型→ 婚姻 —仪式地点→ 卡尔斯巴德	0.296

我们采用一种自动化方法：如果提取出 20 条路径，就不会将空路径（记作“(no path)”）加入聚类（如源代码所示）；本表展示的正是这种情况。