

自适应特征融合的多模态实体对齐研究

郭浩¹ 李欣奕¹ 唐九阳¹ 郭延明¹ 赵翔¹

摘要 多模态数据间交互式任务的兴起对于综合利用不同模态的知识提出了更高的要求,因此融合不同模态知识的多模态知识图谱应运而生.然而,现有多模态知识图谱存在图谱知识不完整的问题,严重阻碍对信息的有效利用.缓解此问题的有效方法是通过实体对齐进行知识图谱补全.当前多模态实体对齐方法以固定权重融合多种模态信息,在融合过程中忽略不同模态信息贡献的差异性.为解决上述问题,设计一套自适应特征融合机制,根据不同模态数据质量动态融合实体结构信息和视觉信息.此外,考虑到视觉信息质量不高、知识图谱之间的结构差异也影响实体对齐的效果,本文分别设计提升视觉信息有效利用率的视觉特征处理模块以及缓和结构差异性的三元组筛选模块.在多模态实体对齐任务上的实验结果表明,提出的多模态实体对齐方法的性能优于当前最好的方法.

关键词 多模态知识图谱, 实体对齐, 预训练模型, 特征融合

引用格式 郭浩, 李欣奕, 唐九阳, 郭延明, 赵翔. 自适应特征融合的多模态实体对齐研究. 自动化学报, 2024, 50(4): 758-770

DOI 10.16383/j.aas.c210518

Adaptive Feature Fusion for Multi-modal Entity Alignment

GUO Hao¹ LI Xin-Yi¹ TANG Jiu-Yang¹ GUO Yan-Ming¹ ZHAO Xiang¹

Abstract The recent surge of interactive tasks involving multi-modal data brings a high demand for utilizing knowledge in different modalities. This facilitated the birth of multi-modal knowledge graphs, which aggregate multi-modal knowledge to meet the demands of the tasks. However, they are known to suffer from the knowledge incompleteness problem that hinders the utilization of information. To mitigate this problem, it is of great need to improve the knowledge coverage via entity alignment. Current entity alignment methods fuse multi-modal information by fixed weighting, which ignores the different contributions of individual modalities. To solve this challenge, we propose an adaptive feature fusion mechanism, that combines entity structure information and visual information via dynamic fusion according to the data quality. Besides, considering that low quality visual information and structural difference between knowledge graphs further impact the performance of entity alignment, we design a visual feature processing module to improve the effective utilization of visual information and a triple filtering module to ease structural differences. Experiments on multi-modal entity alignment indicate that our method outperforms the state-of-the-arts.

Key words Multi-modal knowledge graph, entity alignment, pre-trained model, feature fusion

Citation Guo Hao, Li Xin-Yi, Tang Jiu-Yang, Guo Yan-Ming, Zhao Xiang. Adaptive feature fusion for multi-modal entity alignment. *Acta Automatica Sinica*, 2024, 50(4): 758-770

近年来,以三元组形式表示现实世界知识或事件的知识图谱逐渐成为一种主流的结构化数据的表示方式,并广泛应用于各类人工智能的下游任务,如知识问答^[1]、信息抽取^[2]、推荐系统^[3]等.相比于传统的知识图谱,多模态知识图谱^[4-5]将多媒体信息融合到知识图谱中,从而更好地满足多种模态数据

之间的交互式任务,例如图像和视频检索^[6]、视频摘要^[7]、视觉常识推理^[8]和视觉问答^[9]等,并在近年来受到了学界及工业界的广泛关注.

现有的多模态知识图谱往往从有限的数据源构建而来,存在信息缺失、覆盖率低的问题,导致知识利用率不高.考虑到人工补全知识图谱开销大且效率低,为提高知识图谱的覆盖程度,一种可行的方法^[10-12]是自动地整合来自其他知识图谱的有用知识,而实体作为链接不同知识图谱的枢纽,对于多模态知识图谱融合至关重要.识别不同的多模态知识图谱中表达同一含义的实体的过程,称为多模态实体对齐^[5, 13].

与一般的实体对齐方法不同^[11, 14],多模态实体对齐需要利用和融合多个模态的信息.当前主流的

收稿日期 2021-06-09 录用日期 2021-11-26

Manuscript received June 9, 2021; accepted November 26, 2021

国家自然科学基金 (62002373, 61872446, 71971212, U19B2024) 资助

Supported by National Natural Science Foundation of China (62002373, 61872446, 71971212, U19B2024)

本文责任编辑 胡清华

Recommended by Associate Editor HU Qing-Hua

1. 国防科技大学系统工程学院 长沙 410073

1. College of Systems Engineering, National University of Defense Technology, Changsha 410073

多模态实体对齐方法^[5, 13] 首先利用图卷积神经网络学习知识图谱的结构信息表示; 然后利用预训练的图片分类模型, 生成实体的视觉信息表示 (利用 VGG16^[15]、ResNet^[16] 生成多张图片向量并加和), 得到实体的视觉信息表示; 最后以特定权重将这两种模态的信息结合. 不难发现, 这类方法存在以下 3 个明显缺陷:

1) 图谱结构差异性难以处理. 不同知识图谱中对应的实体通常具有相似的邻接信息, 基于这一假设, 目前的主流实体对齐方法主要依赖知识图谱的结构信息^[14, 17-18] 来实现对齐. 然而真实世界中, 由于构建方式的不同, 不同知识图谱可能存在着较大结构差异, 这不利于找到潜在的对齐实体. 如图 1 所示, 实体 [The dark knight] 在 DBpedia 和 FreeBase 中邻接实体数量存在巨大差异, 虽然包含相同的实体 [Nolan]、[Bale], 然而在 FreeBase 还包含额外 6 个实体. 因此, DBpedia 中的实体 [Bale] 容易错误地匹配到 FreeBase 中的实体 [Gary oldman], 因为它们都是 [The dark knight] 的邻居实体且度数为 1. 真实世界中不同知识图谱的结构性差异问题比图中的示例更为严峻, 以数据集 MMKG^[5] 为例, 基于 FreeBase 抽取得到的图谱 (FB15K) 有接近 60 万的三元组, 而基于 DBpedia 抽取得到的图谱 (DB15K) 中三元组数量不足 10 万. 以实体 [Nolan] 为例, 在 FB15K 中有成百的邻居实体; 而 DB15K 中其邻居实体数量不足 10 个. 针对此类问题, 可基于链接预测生成三元组以丰富结构信息. 这虽然在一定程度上缓和了结构差异性, 但所生成的三元组的可靠性有待考量. 此外, 在三元组数量相差多倍的情况下补全难度很大.

2) 视觉信息利用差. 当前自动化构建多模态知

识图谱的方法通常基于现有知识图谱补充其他模态的信息, 为获取视觉信息, 通常利用爬虫从互联网爬取实体的相关图片以获取其视觉信息. 然而获取的结果中不可避免地存在部分相关程度较低的图片, 即噪声图片. 现有方法^[5, 13, 19] 忽略了噪声图片的影响, 使得基于视觉信息对齐实体的准确率受限. 因此, 实体的视觉信息中混有部分噪声, 进而降低了利用视觉信息进行实体对齐的准确率.

3) 多模态融合权重固定. 当前的主流多模态实体对齐方法^[5, 13] 以固定的权重结合多个模态. 这类方法假设多种模态信息对实体对齐的贡献率始终为一固定值, 并多依赖于多模态知识图谱的结构信息, 然而其忽略了不同模态信息的互补性. 此外, 由于实体相关联的实体数量以及实体在图谱中分布不同, 导致不同实体的结构信息有效性存在一定的差异, 进一步影响不同模态信息的贡献率权重. 事实上, 知识图谱中超过半数实体都是长尾实体^[20], 这些实体仅有不足 5 个相连的实体, 结构信息相对匮乏. 而实体的视觉信息却不受结构影响, 因此在结构信息匮乏的情况下应赋予视觉信息更高的权重. 总而言之, 以固定的权重结合多模态信息无法动态调节各个模态信息的贡献率权重, 导致大量长尾实体错误匹配, 进一步影响实体对齐效果.

为解决上述缺陷, 本文创新性地提出自适应特征融合的多模态实体对齐方法 (Adaptive feature fusion for multi-modal entity alignment, AF²M-EA). 在不失一般性的前提下, 本文从多模态知识图谱中的结构模态和视觉模态两方面出发: 一方面为解决缺陷 1), 提出三元组筛选机制, 通过无监督方法, 结合关系 PageRank 得分以及实体度, 为三元组打分, 并过滤掉无效三元组, 缓和结构差异性;

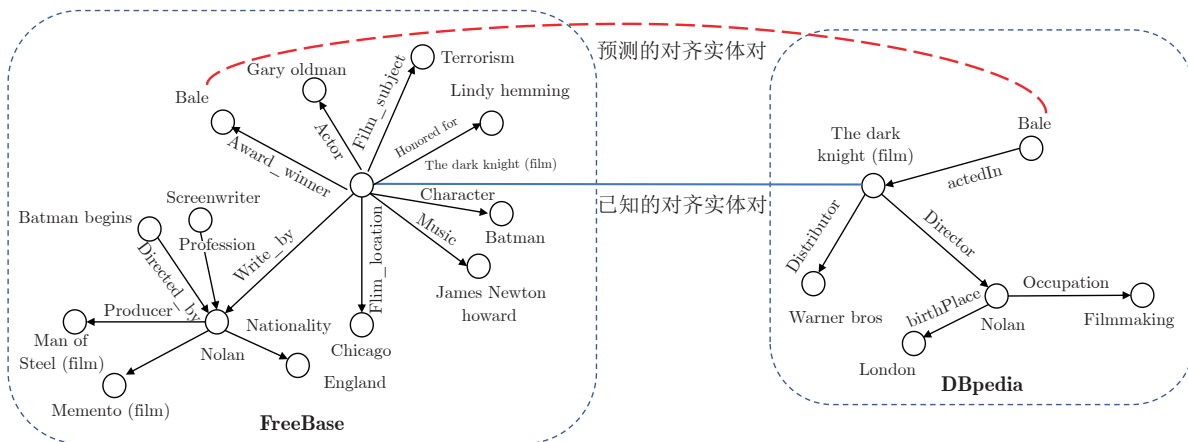


图 1 知识图谱 FreeBase 和 DBpedia 的结构差异性表现

Fig. 1 Structural differences between knowledge graphs FreeBase and DBpedia

另一方面, 针对缺陷 2), 利用图像-文本匹配模型, 计算实体-图片的相似度得分, 设置相似度阈值以过滤噪声图片, 并基于相似度赋予图片不同权重, 生成更高质量的实体视觉特征表示. 此外, 为捕获结构信息动态变化的置信度并充分利用不同模态信息的互补性以应对缺陷 3), 本文设计自适应特征融合机制, 基于实体节点的度数以及实体与种子实体之间的距离, 动态融合实体的结构信息和视觉信息. 这种机制能够有效应对长尾实体数量占比大且结构信息相对匮乏的现实问题. 本文在多模态实体对齐数据集上进行了充分的实验及分析, 表明 AF2MEA 取得了最优的实体对齐效果并证实了提出的各个模块的有效性. 本文的主要贡献可总结为以下 3 个方面:

1) 设计创新的三元组筛选模块, 基于关系 PageRank 评分和实体度生成三元组得分, 过滤三元组, 缓和不同知识图谱的结构差异性;

2) 针对视觉信息利用差的问题, 本工作基于预训练图像-文本匹配模型, 计算实体-图片的相似度得分, 过滤噪声图片, 并基于相似度得分获得更准确的实体视觉特征表示;

3) 设计自适应特征融合模块, 以可变注意力融合实体的结构特征和视觉特征, 充分利用不同模态信息之间的互补性, 进一步提升对齐效果.

本文第 1 节简要介绍相关工作; 第 2 节介绍问题定义和整体框架; 第 3 节具体介绍本文提出的多模态实体对齐模型; 第 4 节明确实验设置, 进行实验并分析结果; 第 5 节为结束语.

1 相关工作

1.1 实体对齐

实体对齐任务旨在寻找两个知识图谱中描述同一真实世界对象的实体对, 以便链接不同知识图谱. 实体对齐作为整合不同知识图谱中知识的关键步骤, 在近年来得到广泛研究.

传统的实体对齐方法^[21]多依赖本体模式对齐, 利用字符串相似度或者规则挖掘等复杂的特征工程方法^[22]实现对齐, 但在大规模数据下准确率及效率显著下降. 而当前实体对齐方法^[14, 17, 23]大多依赖知识图谱向量, 因为向量表示具有简洁性、通用性以及处理大规模数据的能力. 基于不同知识图谱中等效实体具有相似的邻接结构这一假设, 即等效的实体通常具有等效的邻居实体, 这些工作具有相似框架: 首先利用基于翻译的表示学习方法 (Translating embedding, TransE)^[17, 24-25], 图卷积神经网络 (Graph convolutional network, GCN)^[9, 14]等知识

图谱表示方法编码知识图谱结构信息, 并将不同知识图谱中的元素投射到各自低维向量空间中. 接着设计映射函数, 利用已知实体对以对齐不同向量空间. 考虑到 GCN 在学习知识图谱表示上存在忽略关系类型、平均聚合相邻节点特征的缺陷, 一些方法^[26-27]利用基于注意力机制的图神经网络模型来为不同的相邻节点分配不同的权重. 文献 [28] 通过学习知识图谱的关系表示以辅助生成实体表示.

除生成并优化结构表示之外, 部分方法^[14, 26, 29]提出引入属性信息以补充结构信息. 文献 [29] 提出利用属性类型生成属性向量; 而文献 [14] 则将属性表示成最常见属性名的 One-hot 向量. 这类工作均假设图谱中存在大量属性三元组. 但文献 [30] 指出, 在大多数知识图谱中, 69% ~ 99% 的实体至少缺乏 1 个同类别实体具有的属性. 这种情况限制了此类方法的通用性.

1.2 多模态实体对齐

多数知识图谱的构建工作都倾向以结构化形式来组织和发现文本知识, 而很少关注网络上的其他类型的资源^[4, 31]. 近年来, 不同模态数据之间交互式任务大量涌现, 如图像和视频检索^[6]、视频摘要生成^[7]、视觉实体消歧^[8]和视觉问答^[9]等. 为满足跨模态数据交互式任务的需求, 知识图谱需要融合多媒体信息, 多模态知识图谱应运而生.

为提高多模态知识图谱的覆盖程度, 多模态实体对齐是关键的一步. 与实体对齐相似, 多模态实体对齐任务旨在识别不同的多模态知识图谱中表达同一含义的实体对^[13, 19]. 相关的多模态知识表示方法可用于多模态实体对齐任务, 其中基于图像的知识表示模型 (Image-embodied knowledge representation learning, IKRL)^[32]通过三元组和图像学习知识表示, 首先使用神经图像编码器为实体的所有图像构建表示, 然后通过基于注意力的方法将这些图像表示聚合到实体基于图像的集成表示中. 文献 [33] 提出一种基于多模态翻译的方法, 将知识图谱中三元组的损失函数定义为结构表示、视觉表示和语言知识表示的子损失函数的总和.

总的来说, 多模态实体对齐是一个新颖的问题, 目前直接针对该任务的研究相对较少. 其中, 文献 [5] 利用专家乘积模型 (Product of expert, PoE), 综合结构、属性和视觉特征的相似度得分以找到潜在对齐的实体. 文献 [13] 注意到欧几里得空间中知识图谱的结构表示存在失真问题, 利用双曲图卷积神经网络 (Hyperbolic graph convolutional network, HGCN) 学习实体结构特征和视觉特征, 并在双曲空间中结合不同模态特征以寻找潜在的对

齐实体. 文献 [19] 提出一种创新的多模态知识表示方法, 分别设计了多模态知识表示模块和知识融合模块, 融合实体结构特征、属性特征和视觉特征到同一个向量空间中以对齐实体. 该模型取得较好的对齐效果, 但结构设计较为复杂, 视觉特征的利用率不高.

2 问题定义与整体框架

本节主要介绍多模态实体对齐任务的定义以及本文提出的整体模型框架.

2.1 任务定义

多模态知识图谱通常包含多个模态的信息. 鉴于大多数知识图谱中属性信息的缺失^[30], 在不失一般性的前提下, 本工作关注知识图谱的结构信息和视觉信息. 给定 2 个多模态知识图谱 MG_1 和 MG_2 : $MG_1 = (E_1, R_1, T_1, I_1)$, $MG_2 = (E_2, R_2, T_2, I_2)$. 其中, E 代表实体集合; R 代表关系集合; T 代表三元组集合, 三元组表示为 $\langle E, R, E \rangle$ 的子集; I 代表实体相关联的图片集合. 种子实体对集合 $S = \{(e_i^1, e_j^2) | e_i^1 \in E_1, e_j^2 \in E_2\}$ 表示用于训练的对齐的实体对集合. 多模态实体对齐任务旨在利用种子实体对, 发现潜在对齐的实体对 $S' = \{(e_m^1, e_n^2) | e_m^1 = e_n^2; e_m^1 \in E_1, e_n^2 \in E_2\}$, 其中等号代表两个实体指代真实世界中同一实体.

给定某一实体, 寻找其在另一知识图谱中对应该实体的过程可视为排序问题. 即在某一特征空间下, 计算给定实体与另一知识图谱中所有实体的相似程度 (距离) 并给出排序, 而相似程度最高 (距离最小) 的实体可视为对齐结果.

2.2 模型框架

本工作提出的自适应特征融合的多模态实体对齐框架如图 2 所示. 首先利用图卷积神经网络学习实体的结构向量, 生成实体结构特征; 设计视觉特征处理模块, 生成实体视觉特征; 接着基于自适应特征融合机制, 结合 2 种模态的信息进行实体对齐. 此外, 为缓和知识图谱的结构差异性, 本工作设计三元组筛选机制, 融合关系评分及实体的度, 过滤部分三元组. 图 2 中 MG_1 和 MG_2 分别表示不同的多模态知识图谱; KG_1 、 KG_2 表示知识图谱; KG_1' 表示三元组筛选模块处理后的知识图谱.

3 多模态实体对齐模型

本节介绍提出的多模态实体对齐框架的各个子模块, 包括视觉特征处理模块、结构特征学习模块、三元组筛选模块以及自适应特征融合模块.

3.1 视觉特征处理模块

当前多模态知识图谱的视觉信息图片来源于互联网搜索引擎, 不可避免地存在噪声图片, 不加区分地使用这些图片信息会导致视觉信息利用率差. 而图像-文本匹配模型^[34-35]可以计算图像与文本的相似性程度. 受此启发, 为解决视觉信息利用率差的问题, 本工作设计了视觉特征处理模块, 为实体生成更精确的视觉特征以帮助实体对齐. 图 3 详细描述了实体视觉特征的生成过程. 在缺乏监督数据的情况下, 本文采用预训练的图像-文本匹配模型, 生成图片与实体相似度; 接下来设置相似度阈值过滤噪声图片; 最后基于相似度得分赋予图片相应的权重, 最终生成实体的视觉特征表示, 具体步骤如下:

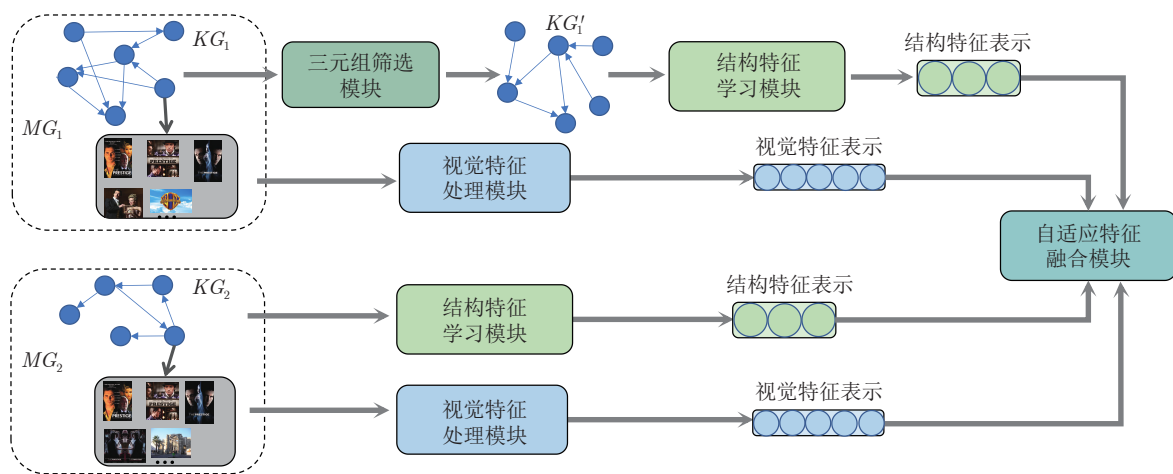


图 2 自适应特征融合的多模态实体对齐框架

Fig. 2 Multi-modal entity alignment framework based on adaptive feature fusion

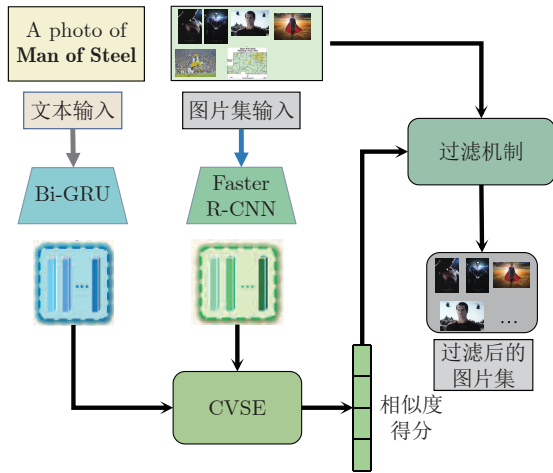


图3 视觉特征处理模块

Fig.3 Visual feature processing module

1) 计算图片-实体相似度得分. 本步骤使用预训练的文本图像匹配模型-共识感知的视觉语义嵌入模型 (Consensus-aware visual semantic embedding, CVSE)^[36] 计算实体图片集中各个图片的相似度得分. CVSE 模型将不同模态间共享的常识知识结合到图像-文本匹配任务中, 并在数据集 MS-COCO^[37] 和 Flickr30k^[38] 上进行模型训练, 取得先进的图文匹配效果. 本文基于 CVSE 模型及其训练的参数计算图片-实体相似度得分.

视觉特征处理模块的输入为实体的名称和实体相应的图片集, 见图3左侧. 首先生成实体图片集的图片嵌入 $p_i \in \mathbf{R}^{n \times 36 \times 2048}$, n 为实体对应图片集中图片的数量. 本文利用目标检测算法 Faster R-CNN^[39] 为每幅图片生成 36×2048 维的特征向量. 然后将实体名 [Entity Name] 拓展为句子 {A photo of Entity Name}, 再送入双向门控循环单元 (Bidirectional gated recurrent unit, Bi-GRU)^[40] 以生成实体的文本信息 t_i .

接着将图片嵌入 p_i 和文本信息 t_i 送入 CVSE 模型中, 本文移除 CVSE 模型的 Softmax 层, 以获取实体图像集中图片的相似度得分:

$$Sco_i^v = CVSE(p_i; t_i) \quad (1)$$

其中, CVSE 表示共识感知的视觉语义嵌入模型, 其运算结果 $Sco_i^v \in \mathbf{R}^n$ 表示图片集与文本的相似度得分.

2) 过滤噪声图片. 考虑到实体的图片集中存在部分相似度很低的图片, 影响视觉信息的精度. 鉴于此, 设置相似度阈值 α , 以过滤噪声图片:

$$set'(i) = \{j' | j' \in set(i), Sco_i^v(j') > \alpha\} \quad (2)$$

其中, $set(i)$ 代表初始图片集, $set'(i)$ 表示过滤掉噪声图片后的图片集, α 是相似度阈值超参数.

3) 实体视觉特征表示生成. 对于 $set'(i)$ 中的图片, 本文基于其相似度得分赋予权重, 为实体 e_i 生成更精确的视觉特征表示 V_i :

$$V_i = I_i' \times att_i \quad (3)$$

其中, $V_i \in \mathbf{R}^{2048}$ 表示实体 i 的视觉特征; $I_i' \in \mathbf{R}^{2048 \times n'}$ 为 ResNet 模型生成的图像特征, n' 为去除噪声后的图片数量; att_i 表示图片注意力权重:

$$att_i = \text{Softmax}(Sco_i^v) \quad (4)$$

其中, Sco_i^v 为过滤后图片集 $set'(i)$ 的相似度得分.

3.2 结构特征学习模块

本文采用图卷积神经网络 (GCN)^[41-42] 捕捉实体邻接结构信息并生成实体结构表示向量. GCN 是一种直接作用在图结构数据上的卷积网络, 通过捕捉节点周围的结构信息生成相应的节点结构向量:

$$H^{l+1} = \sigma(\hat{A}H^lW^l) \quad (5)$$

其中, H^l , H^{l+1} 分别表示 l 层和 $l+1$ 层节点的特征矩阵; W^l 表示可训练的参数; $\hat{A} = D^{1/2}\bar{A}D^{-1/2}$ 表示标准化的邻接矩阵, 其中 D 为度矩阵; $\bar{A} = A + I$, A 表示邻接矩阵, 若实体和实体之间存在关系, 则 $A_{ij} = 1$; I 表示单位矩阵. 激活函数 σ 设为 ReLU.

由于不同知识图谱的实体结构向量并不在同一空间中, 因此需要利用已知实体对集合 S 将不同知识图谱中的实体映射到同一空间中. 具体的训练目标为最小化下述损失函数:

$$L = \sum_{(e_1, e_2) \in S} \sum_{(e'_1, e'_2) \in S'} (|h_{e_1} - h_{e_2}| - |h_{e'_1} - h_{e'_2}| + \gamma)_+ \quad (6)$$

其中, $(x)_+ = \max\{0, x\}$; S' 代表负样本集合, 基于已知的种子实体对 (e_1, e_2) , 以随机实体替换 e_1 或者 e_2 生成. h_e 代表实体 e 的结构向量, $|h_{e_1} - h_{e_2}|$ 代表实体 e_1 和 e_2 之间的曼哈顿距离; 超参数 γ 代表正负例样本分隔的距离.

3.3 三元组筛选模块

知识图谱的结构特征以三元组形式表示: (h, r, t) , 其中, h 代表头实体, t 代表尾实体, r 代表关系. 不同知识图谱三元组的数量差异较大, 导致基于结构信息进行实体对齐的效果大打折扣. 为缓和不同知识图谱的结构差异性, 本工作设计三元组筛选模块, 评估三元组重要性, 并基于重要性得分过滤部分无效三元组. 筛选流程如图4所示, 其中三元组重要性得分结合关系 r 的 PageRank 得分, 以及实体 h 和 t 的度.

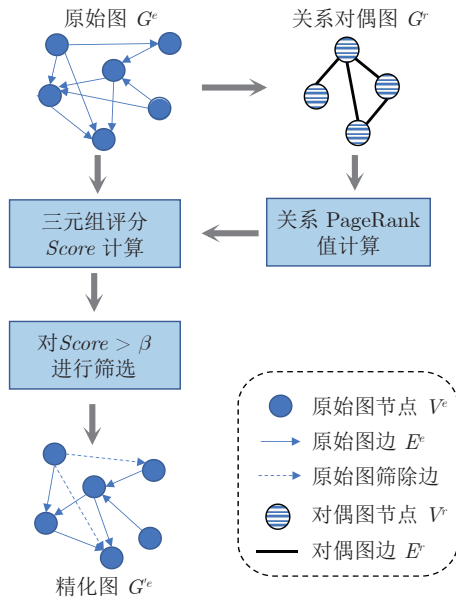


图 4 三元组筛选模块

Fig. 4 Triples filtering module

1) 关系 PageRank 评分计算. 首先构建以关系为节点、实体为边的关系-实体图, 也称知识图谱的关系对偶图^[43]. 定义知识图谱为 $G^e = (V^e, E^e)$, 其中 V^e 为实体集合, E^e 为关系集合. 而关系对偶图 $G^r = (V^r, E^r)$ 以关系为节点, 若两个不同的关系由同一个头实体 (尾实体) 连接, 则这两个关系节点间存在一条边. V^r 为关系节点的集合, E^r 为边的集合.

基于上述生成的关系对偶图, 本文使用 PageRank^[44] 算法计算关系得分. PageRank 算法是图结构数据上链接分析的代表性算法, 属于无监督学习方法. 其基本思想是在有向图上定义一个随机游走模型, 描述随机游走者沿着有向图随机访问各个结点的行为. 在一定条件下, 极限情况访问每个结点的概率收敛到平稳分布, 这时各个结点的平稳概率值就是其 PageRank 值, 表示结点的重要度. 受该算法的启发, 基于知识图谱关系对偶图, 计算关系的 PageRank 值以表示关系的重要性:

$$PR(r) = \sum_{v \in B_r} \frac{PR(v)}{L(v)} \quad (7)$$

其中, $PR(r)$ 为关系的 PageRank 评分; 关系 $v \in B_r$, B_r 表示关系 r 的邻居关系集合; $L(v)$ 代表与关系 v 连接的关系数量 (即关系节点的度数).

2) 三元组评分机制. 对三元组的筛选, 一方面要过滤掉冗余或无效的关系; 另一方面要保护知识图谱的结构特征. 由于结构信息缺乏的长尾实体仅有少量相关三元组, 若基于关系重要性评分直接过滤一种关系可能会加剧长尾实体的结构信息匮乏问

题. 为此, 本工作结合关系的 PageRank 评分和头尾实体的度, 设计三元组评分函数:

$$Score(h, r, t) = \frac{PR(r)}{\ln d^h \times \ln d^t} \quad (8)$$

其中, d^h 和 d^t 分别表示头实体和尾实体的度, 即实体相关联的边的数量. 基于三元组评分 $Score$, 并设置阈值 β , 保留 $Score(h, r, t) > \beta$ 的三元组, 以精化知识图谱. 值得注意的是, 阈值 β 的取值由筛选的三元组数量决定.

3.4 自适应特征融合模块

多模态知识图谱包含至少 2 个模态的信息, 多模态实体对齐需要融合不同模态的信息. 已有的方法将不同的嵌入合并到一个统一的表示空间中^[45], 这需要额外的训练来统一表示不相关的特征. 更可取的策略是首先计算不同模态特征在其特定空间内的相似度, 然后组合各个模态特征的相似度得分以寻找匹配的实体对^[14, 46].

形式上, 给定结构特征向量表示 S , 视觉特征表示 V . 计算每个实体对 (e_1, e_2) 中实体之间的相似度得分, 然后利用该相似度得分来预测潜在的对齐实体. 为计算总体相似度, 当前方法首先计算 e_1 和 e_2 之间的视觉特征向量相似度得分 $Sim^v(e_1, e_2)$ 和结构特征向量的特征相似度得分 $Sim^s(e_1, e_2)$. 相似度得分一般用向量的余弦相似度或曼哈顿距离表示. 接下来, 以固定权重结合上述相似度得分:

$$Sim(e_1, e_2) = Sim^s(e_1, e_2) \times Att^s + Sim^v(e_1, e_2) \times Att^v \quad (9)$$

其中, Att^s 和 Att^v 分别代表结构信息和视觉信息的贡献率权重; $Sim(e_1, e_2)$ 表示最终的实体相似度得分.

不同模态的特征从不同视角表征实体, 具有一定相关性和互补性^[47-49]. 当前多模态实体对齐方法以固定的权重结合结构信息和视觉信息, 认为多种模态信息对实体对齐的贡献率始终为一定值, 忽略了不同实体之间结构信息的有效性差异. 基于度感知的长尾实体对齐方法^[10] 首次提出动态赋予不同特征重要性权重的方法, 设计了基于度感知的联合注意力网络, 提升了长尾实体的对齐准确率. 这证明实体结构信息的有效性与实体度的数量呈正相关, 并且不同知识图谱中对等的实体通常具有对等的邻居实体, 实体与种子实体关联的密切程度与其结构特征的有效性也呈正相关. 而实体的视觉信息的有效性不受此类影响, 对于结构信息匮乏的实体, 应更多地信任视觉信息.

基于此, 为捕捉不同模态信息的贡献率动态变

化, 本工作基于实体度的数量, 并进一步结合实体与种子实体关联的密切程度, 设计自适应特征融合机制:

$$Att^s = \frac{K}{1 + b \times e^{-a \times (degree + N^{hop})}} \quad (10)$$

$$Att^v = 1 - Att^s \quad (11)$$

其中, K , b , a 均为超参数, $degree$ 表示该实体的度数, N^{hop} 表示实体与种子实体关联密切程度:

$$N^{hop} = n^{1-hop} \times w_1 + \lg(n^{2-hop} \times w_2) \quad (12)$$

其中, n^{1-hop} 和 n^{2-hop} 分别表示距离种子实体 1 跳和 2 跳的实体数量; w_1 和 w_2 为超参数.

4 实验

本节首先介绍实验的基本设置, 包括参数设置、数据集、对比方法以及评价指标. 接着展示在多模态实体对齐任务上的实验结果, 并进行消融分析以验证各个模块的有效性. 此外, 对各个模块进行分析, 验证设计的合理性及有效性.

4.1 数据集和评价指标

在实验中, 我们使用文献 [5] 构建的多模态实体对齐数据集 MMKG. 数据集 MMKG 从知识库 FreeBase、DBpedia 和 Yago 中抽取得到, 包含两对多模态数据集 FB15K-DB15K 和 FB15K-Yago15K. 表 1 描述了数据集的详细信息. SameAs 表示等效实体. 在实验中, 等效实体以一定比例划分, 分别用于模型训练和测试.

表 1 多模态知识图谱数据集数据统计
Table 1 Statistic of the MMKGs datasets

| 数据集 | 实体 | 关系 | 三元组 | 图片 | SameAs |
|---------|--------|-------|---------|--------|--------|
| FB15K | 14 915 | 1 345 | 592 213 | 13 444 | |
| DB15K | 14 777 | 279 | 99 028 | 12 841 | 12 846 |
| Yago15K | 15 404 | 32 | 122 886 | 11 194 | 11 199 |

由于数据集不提供图片, 为获取实体相关图片, 本文基于数据集 MMKG 创建 URI (Uniform resource identifier) 数据, 并设计网络爬虫, 解析来自图像搜索引擎 (即 Google Images、Bing Images 和 Yahoo Image Search) 的查询结果. 然后, 将不同搜索引擎获取的图片分配给不同的 MMKG. 为模拟真实世界多模态知识图谱的构建过程, 去除等效实体图像集中相似度过高的图片, 并引入一定数量的噪声图片.

本文实验使用 Hits@ k ($k = 1, 10$) 和平均倒数排名 (Mean reciprocal rank, MRR) 作为评价指标.

对于测试集中每个实体, 另一个图谱中的实体根据它们与该实体的相似度得分以降序排列. Hits@ k 表示前 k 个实体中包含正确的实体的数量占总数量的百分比; 另一方面, MRR 表示正确对齐实体的倒数排序的平均值. MRR 是信息检索领域常用的评价指标之一, 表示目标实体在模型预测的实体相关性排序中排名的倒数的平均值. 注意, Hits@ k 和 MRR 数值越高表示性能越好, Hits@ k 的结果以百分比表示. 表 2 和表 3 中以粗体标注最好的效果. Hits@1 代表对齐的准确率, 通常视为最重要的评价指标.

4.2 参数设置和对比方法

实体结构特征由图卷积神经网络生成, 负例数量设定为 15, 边缘超参数 $\gamma = 3$, 训练 400 轮, 维度 $d_s = 300$. 视觉特征由第 3.1 节中提出的视觉特征处理模块生成, 维度 $d_v = 2 048$; 相似度阈值 α 的值是基于比例确定的, 对于每个实体的图片集, 保留相似度前 50% 的图片, 过滤其余 50% 的噪声图片. 基于文献 [5, 13] 的实验设置, 将种子实体的比例设置为 20% 和 50%, 并且选取 10% 的实体作为验证集, 用于调整式 (10) 和式 (12) 中超参数, 其中, $b = 1.5$, $a = 1$. 参数 K 的取值与种子实体的比例相关, 实验中设定的种子实体比例 $seed$ 不同, 则 K 取值也不同, 当 $seed = 20\%$ 时, K 取值为 0.6; 当 $seed = 50\%$ 时, K 取值为 0.8. 式 (12) 中超参数 w_1 和 w_2 分别取 0.8 和 0.1. 三元组筛选模块中的阈值 β 也是基于验证集调整得来, 取值为 0.3, 将 FB15K 的三元组量筛选至约 30 万.

此外, 将本文提出的模型 (AF²MEA) 与以下 4 种方法进行对比.

1) IKRL 方法^[32]. 通过基于注意力的方法, 将实体的图像表示与三元组知识聚合到实体的集成表示中以对齐实体.

2) GCN-align 方法^[14]. 利用 GCN 生成实体结构和视觉特征矩阵, 以固定权重结合两种特征以对齐实体.

3) PoE 方法^[5]. 基于提取的结构、属性和视觉特征, 综合各个特征的相似度得分以找到潜在对齐的实体.

4) HMEA (Hyperbolic multi-modal entity alignment) 方法^[3]. 利用双曲图卷积神经网络 HGCN 生成实体的结构和视觉特征矩阵, 并在双曲空间中以权重结合结构特征和视觉特征, 进行实体对齐.

4.3 主实验

通过表 2 可以明显看出, 与 IKRL、GCN-align、PoE 以及 HMEA 方法相比, 本文提出的方法取得

最好的实验结果. 在数据集 FB15K-DB15K 上, 本文提出的方法 AF²MEA 的 Hits@1 值显著高于当前最优方法 HMEA, 尤其在种子实体比例为 20% 条件下, Hits@1 指标的提升超过 5%, MRR 也取得大幅提升. 此外, 在各项指标上, 本文所提 AF²MEA 均大幅领先 IKRL、GCN-align 以及 PoE.

在数据集 FB15K-Yago15K 上, 与其他 4 种模型相比, AF²MEA 在全部指标上均有大幅提升, 进一步验证了本文提出的模型的有效性. 其中, 在种子实体比例为 20% 和 50% 的条件下, AF²MEA 的 Hits@1 指标较 HMEA 分别提升约 11% 和 8%.

4.4 消融实验

本文创新性地设计了模型的 3 个模块, 分别是视觉特征处理模块、三元组筛选模块和自适应特征融合模块. 为验证各模块对于多模态实体对齐任

务的有效性, 本节进一步设计了消融实验. 其中, AF²MEA_{Adaptive}、AF²MEA_{Visual} 和 AF²MEA_{Filter} 分别表示去除特征融合模块的模型、去除视觉特征处理模块的模型和去除三元组筛选模块的模型, 通过与本文提出的完整模型 AF²MEA 进行对比来检测各模块的有效性.

本文消融实验分别在数据集 FB15K-DB15K 和 FB15K-Yago15K 上进行, 并分别基于 20% 和 50% 种子实体比例进行实验对比. 表 3 展示了消融实验的结果, 完整模型在所有情况下均取得最好的实体对齐效果, 去除各个子模块都使得对齐准确率出现一定程度的下降.

对表 3 进行具体分析可知, 三元组筛选模块对实体对齐影响最大: 在种子实体占比 20% 的条件下, 去除该模块导致 Hits@1 指标在数据集 FB15K-DB15K 和 FB15K-Yago15K 上分别下降 3.6% 和

表 2 多模态实体对齐结果
Table 2 Results of multi-modal entity alignment

| 数据集 | 方法 | seed = 20% | | | seed = 50% | | |
|---------------|---------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | Hits@1 | Hits@10 | MRR | Hits@1 | Hits@10 | MRR |
| FB15K-DB15K | IKRL | 2.96 | 11.45 | 0.059 | 5.53 | 24.41 | 0.121 |
| | GCN-align | 6.26 | 18.81 | 0.105 | 13.79 | 34.60 | 0.210 |
| | PoE | 11.10 | 17.80 | — | 23.50 | 33.00 | — |
| | HMEA | 12.16 | 34.86 | 0.191 | 27.24 | 51.77 | 0.354 |
| | AF ² MEA | 17.75 | 34.14 | 0.233 | 29.45 | 50.25 | 0.365 |
| FB15K-Yago15K | IKRL | 3.84 | 12.50 | 0.075 | 6.16 | 20.45 | 0.111 |
| | GCN-align | 6.44 | 18.72 | 0.106 | 14.09 | 34.80 | 0.209 |
| | PoE | 8.70 | 13.30 | — | 18.50 | 24.70 | — |
| | HMEA | 10.03 | 29.38 | 0.168 | 27.91 | 55.31 | 0.371 |
| | AF ² MEA | 21.65 | 40.22 | 0.282 | 35.72 | 56.03 | 0.423 |

表 3 消融实验实体对齐结果
Table 3 Entity alignment results of ablation study

| 数据集 | 方法 | seed = 20% | | | seed = 50% | | |
|---------------|---|--------------|--------------|--------------|--------------|--------------|--------------|
| | | Hits@1 | Hits@10 | MRR | Hits@1 | Hits@10 | MRR |
| FB15K-DB15K | AF ² MEA | 17.75 | 34.14 | 0.233 | 29.45 | 50.25 | 0.365 |
| | AF ² MEA _{Adaptive} | 16.03 | 31.01 | 0.212 | 26.29 | 45.35 | 0.331 |
| | AF ² MEA _{Visual} | 16.19 | 30.71 | 0.212 | 26.14 | 45.38 | 0.323 |
| | AF ² MEA _{Filter} | 14.13 | 28.77 | 0.191 | 22.91 | 43.08 | 0.297 |
| FB15K-Yago15K | AF ² MEA | 21.65 | 40.22 | 0.282 | 35.72 | 56.25 | 0.423 |
| | AF ² MEA _{Adaptive} | 19.32 | 37.38 | 0.255 | 31.77 | 53.24 | 0.393 |
| | AF ² MEA _{Visual} | 19.75 | 36.38 | 0.254 | 32.08 | 51.53 | 0.388 |
| | AF ² MEA _{Filter} | 15.84 | 32.36 | 0.216 | 27.38 | 48.14 | 0.345 |

6.8%; 在种子实体占比 50% 的条件下, 去除三元组筛选模块导致的性能下降更多, 约为 7% 和 8%. 此外, 去除视觉特征处理模块和自适应特征融合模块也对实体对齐效果产生了一定程度的影响. 在数据集 FB15K-DB15K 上, 去除视觉特征处理模块和去除自适应特征融合模块导致近似相同程度的 Hits@1 指标的下降, 在种子实体占比为 20% 时下降 1.5% 以上, 在种子实体占比为 50% 时下降超过 3%.

4.5 各子模块分析

1) 视觉特征处理模块. 视觉特征处理模块包含基于相似度注意力的图片特征融合机制和基于相似度的图片过滤机制. 为验证上述两种机制的有效性, 本节设计了对比实验, 其中 Att、Filter 分别表示基于相似度注意力的图片特征融合机制和基于相似度的图片过滤机制. Att+Filter 表示结合两种机制, 即本文提出的视觉特征处理模块. HMEA-v 表示文献 [13] 提出的视觉特征处理方法.

由表 4 可知, 本文提出的基于相似度注意力的图片特征融合机制与 HMEA-v 相比, 在所有指标

上均有较大提升, 在种子实体占比 20% 的情况下, Hits@1 提升超过 6%, MRR 也取得很大提升. 此外, 两个模块 Att、Filter 结合取得了最好的对齐效果, 相比单纯使用注意力模块有了小幅的提升.

2) 三元组筛选模块. 为验证本文提出的三元组筛选模块的有效性, 本文对比了 $F_{PageRank}$ 、 F_{random} 和 F_{our} 三种筛选机制, 分别代表基于 PageRank 评分筛选机制、随机筛选机制以及本文设计的筛选机制. 为控制实验变量, 本实验使用上述三种筛选机制筛选了相同数量的三元组, 约 30 万, 并基于图卷积神经网络学习结构特征, 保持各参数一致.

实验结果表明, 随机筛选 F_{random} 相较于保留所有三元组的基线, 其 Hits@1 在 $seed = 20\%$ 和 $seed = 50\%$ 的情况下分别提升约 1.5% 和 2.5%, 表明图谱结构差异性对于实体对齐存在一定的影响. 基于 PageRank 评分的筛选机制相比于随机筛选, 在种子实体比例为 50% 的情况下, 提升 3% 左右. 由表 5 可知, 本文提出的三元组筛选机制取得了最优对齐结果, 在 FB15K-DB15K 上与基线对比, Hits@1 指标在不同种子实体比例下分别提升约 3% 和 8%;

表 4 实体视觉特征的对齐结果
Table 4 Entity alignment results of visual feature

| 数据集 | 方法 | $seed = 20\%$ | | | $seed = 50\%$ | | |
|---------------|------------|---------------|--------------|--------------|---------------|--------------|--------------|
| | | Hits@1 | Hits@10 | MRR | Hits@1 | Hits@10 | MRR |
| FB15K-DB15K | HMEA-v | 2.07 | 9.82 | 0.058 | 3.91 | 14.41 | 0.086 |
| | Att | 8.81 | 20.16 | 0.128 | 9.57 | 21.13 | 0.139 |
| | Att+Filter | 8.98 | 20.52 | 0.131 | 9.96 | 22.58 | 0.144 |
| FB15K-Yago15K | HMEA-v | 2.77 | 11.49 | 0.072 | 4.28 | 15.38 | 0.095 |
| | Att | 9.25 | 21.38 | 0.137 | 10.56 | 23.55 | 0.157 |
| | Att+Filter | 9.43 | 21.91 | 0.138 | 11.07 | 24.51 | 0.158 |

表 5 不同三元组筛选机制下实体结构特征对齐结果
Table 5 Entity alignment results of structure feature in different filtering mechanism

| 数据集 | 方法 | $seed = 20\%$ | | | $seed = 50\%$ | | |
|---------------|----------------|---------------|--------------|--------------|---------------|--------------|--------------|
| | | Hits@1 | Hits@10 | MRR | Hits@1 | Hits@10 | MRR |
| FB15K-DB15K | Baseline | 6.26 | 18.81 | 0.105 | 13.79 | 34.60 | 0.210 |
| | $F_{PageRank}$ | 8.03 | 21.37 | 0.125 | 18.90 | 39.25 | 0.259 |
| | F_{random} | 7.57 | 20.76 | 0.120 | 16.32 | 36.48 | 0.231 |
| | F_{our} | 9.74 | 25.28 | 0.150 | 22.09 | 44.85 | 0.297 |
| FB15K-Yago15K | Baseline | 6.44 | 18.72 | 0.106 | 15.88 | 36.70 | 0.229 |
| | $F_{PageRank}$ | 9.54 | 23.45 | 0.144 | 21.67 | 42.30 | 0.290 |
| | F_{random} | 8.17 | 20.86 | 0.126 | 18.22 | 38.55 | 0.254 |
| | F_{our} | 11.59 | 28.44 | 0.175 | 24.88 | 47.85 | 0.327 |

在 FB15K-Yago15K 上, Hits@1 指标分别提升约 5% 和 9%。

3) 自适应特征融合模块. 本文提出的自适应特征融合, 结合实体度以及实体与种子实体的关联程度, 赋予不同模态信息动态的贡献率权重. 第 4.4 节中消融实验结果已证明自适应特征融合机制的有效性, 为进一步验证该机制对结构信息匮乏的实体的对齐效果, 本节对比自适应特征融合机制和固定权重特征融合两种方法.

由于结构信息的丰富程度与实体的度相关, 我们按照实体度的数量将实体划分为 3 类, 在这 3 类实体上分别测试本文提出的自适应融合机制和固定权重机制下多模态实体对齐的准确率. 本实验种子实体比例设置为 20%, 分别在数据集 FB15K-DB15K 与 FB15K-Yago15K 上进行, 相关参数与第 4.4 节中消融实验保持一致.

表 6 展示了自适应特征融合与固定权重融合的多模态实体对齐结果. 其中 Fixed 和 Adaptive 分别代表固定权重融合机制和自适应特征融合机制; Group 1、Group 2 和 Group 3 分别表示前 1/3、中间 1/3 和后 1/3 部分实体, 基于实体度从小到大划分. 由表 6 可知, 自适应特征融合机制相比固定权重融合, 在各类实体上均取得更好的实体对齐效果. 图 5 表示自适应特征融合与固定权重融合的实体对齐 Hits@1 对比, 可以清晰地看出, 在 Group 1 上提升显著高于 Group 2 和 Group 3, 证明本文提出的自适应特征融合机制可显著提升结构信息匮乏的实体即长尾实体的对齐准确率.

表 6 自适应特征融合与固定权重融合多模态实体对齐结果

Table 6 Multi-modal entity alignment results of fixed feature fusion and adaptive feature fusion

| 方法 | Group 1 | | Group 2 | | Group 3 | |
|---------------|---------|---------|---------|---------|---------|---------|
| | Hits@1 | Hits@10 | Hits@1 | Hits@10 | Hits@1 | Hits@10 |
| FB15K-DB15K | | | | | | |
| Adaptive | 16.44 | 32.97 | 17.43 | 33.47 | 19.29 | 35.40 |
| Fixed | 13.87 | 28.91 | 15.82 | 31.08 | 18.12 | 34.33 |
| FB15K-Yago15K | | | | | | |
| Adaptive | 16.44 | 32.97 | 17.43 | 33.47 | 19.29 | 35.40 |
| Fixed | 16.21 | 33.23 | 19.55 | 37.11 | 22.27 | 45.52 |

4.6 补充实验

本工作旨在结合知识图谱中普遍存在的结构信息和不同模态的视觉信息, 并提升视觉信息的有效

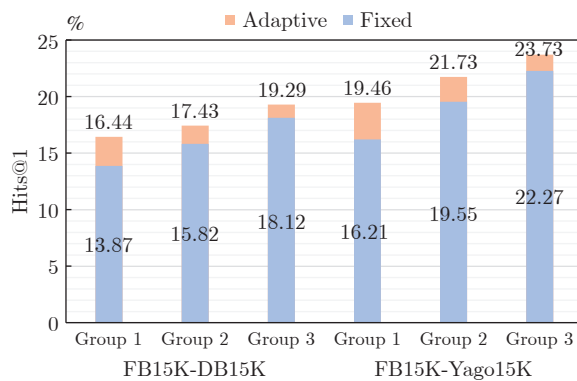


图 5 自适应特征融合与固定权重融合的实体对齐 Hits@1 对比

Fig. 5 Entity alignment Hits@1's comparison of adaptive feature fusion and fixed feature fusion

利用. MMEA (Multi-modal entity alignment)^[19] 模型取得了较好的实验结果, 但本文使用的数据集与其使用的数据集存在一定差异, 因此没有将 MMEA 作为主实验中的对比模型. 为证明本文提出方法的有效性, 我们在 AF²MMEA 原有的结构信息和视觉信息的基础上, 添加属性信息, 并在数据集 FB15K-Yago15K 上进行对比实验. 我们对属性信息进行简单处理: 首先基于种子实体找到对应属性, 利用对应属性的数值对实体对进行相似度打分. 由于实体属性值不受实体结构的影响, 我们再次使用自适应特征融合模块以融合属性信息, 寻找潜在的对齐实体.

如表 7 所示, 基于相同实验条件, 本文提出的模型 AF²MMEA 的效果显著优于 PoE 模型及 MMEA 模型. 在种子实体比例为 20% 的情况下, 与 MMEA 相比, 本文提出的方法在 Hits@1 指标上取得 5% 以上的提升. 在种子实体比例为 50% 的情况下, AF²MMEA 的 Hits@1 值达到 48.25%, 高出 MMEA 约 8%. 在指标 Hits@10 以及 MRR 上, AF²MMEA 也有较大的提升. 这进一步证明了本文提出框架的有效性和可扩展性.

5 结束语

为解决多模态知识图谱不完整的问题, 本文提出自适应特征融合的多模态实体对齐方法 AF²MMEA, 设计自适应特征融合机制实现多种模态信息有效融合, 充分利用多模态信息间的互补性. 并且, 当前多模态知识图谱中视觉信息利用率不高, 本文基于预训练的图像-文本匹配模型, 设计了视觉特征处理模块, 为实体生成更精确的视觉特征表示. 此外, 注

表 7 补充实验多模态实体对齐结果
Table 7 Multi-modal entity alignment results of additional experiment

| 方法 | seed = 20% | | | seed = 50% | | |
|---------------------|------------|---------|--------|------------|---------|-------|
| | Hits@1 | Hits@10 | MRR | Hits@1 | Hits@10 | MRR |
| PoE | 16.44 | 32.97 | 17.430 | 34.70 | 53.60 | 0.414 |
| MMEA | 13.87 | 28.91 | 15.820 | 40.26 | 64.51 | 0.486 |
| AF ³ MEA | 28.65 | 48.22 | 0.382 | 48.25 | 75.83 | 0.569 |

意到不同知识图谱之间存在较大的结构差异限制实体对齐的效果, 本文设计三元组筛选机制, 缓和结构差异. 该模型在多模态实体对齐数据集上取得最好的效果, 并显著提升实体对齐准确率.

后续工作将进一步研究多模态特征联合表示、预训练实体对齐模型等多模态实体对齐的相关问题, 构建高效可行的多模态知识图谱融合系统.

References

- Zhu S G, Cheng X, Su S. Knowledge-based question answering by tree-to-sequence learning. *Neurocomputing*, 2020, **372**: 64–72
- Martinez-Rodriguez J L, Hogan A, Lopez-Arevalo I. Information extraction meets the semantic web: A survey. *Semantic Web*, 2020, **11**(2): 255–335
- Yao X C, Van Durme B. Information extraction over structured data: Question answering with freebase. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA: ACL, 2014. 956–966
- Sun Z, Yang J, Zhang J, Bozzon A, Huang L K, Xu C. Recurrent knowledge graph embedding for effective recommendation. In: Proceedings of the 12th ACM Conference on Recommender Systems. Vancouver, Canada: ACM, 2018. 297–305
- Wang M, Qi G L, Wang H F, Zheng Q S. Richpedia: A comprehensive multi-modal knowledge graph. In: Proceedings of the 9th Joint International Conference on Semantic Technology. Hangzhou, China: Springer, 2019. 130–145
- Liu Y, Li H, Garcia-Duran A, Niepert M, Onoro-Rubio D, Rosenblum D S. MMKG: Multi-modal knowledge graphs. In: Proceedings of the 16th International Conference on the Semantic Web. Portorož, Slovenia: Springer, 2019. 459–474
- Shen L, Hong R C, Hao Y B. Advance on large scale near-duplicate video retrieval. *Frontiers of Computer Science*, 2020, **14**(5): Article No. 145702
- Han Y H, Wu A M, Zhu L C, Yang Y. Visual commonsense reasoning with directional visual connections. *Frontiers of Information Technology & Electronic Engineering*, 2021, **22**(5): 625–637
- Zheng W F, Yin L R, Chen X B, Ma Z Y, Liu S, Yang B. Knowledge base graph embedding module design for visual question answering model. *Pattern Recognition*, 2021, **120**: Article No. 108153
- Zeng W X, Zhao X, Wang W, Tang J Y, Tan Z. Degree-aware alignment for entities in tail. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Event: ACM, 2020. 811–820
- Zhao X, Zeng W X, Tang J Y, Wang W, Suchanek F. An experimental study of state-of-the-art entity alignment approaches. *IEEE Transactions on Knowledge and Data Engineering*, 2022, **34**(6): 2610–2625
- Zeng W X, Zhao X, Tang J Y, Li X Y, Luo M N, Zheng Q H. Towards entity alignment in the open world: An unsupervised approach. In: Proceedings of the 26th International Conference Database Systems for Advanced Applications. Taipei, China: Springer, 2021. 272–289
- Guo H, Tang J Y, Zeng W X, Zhao X, Liu L. Multi-modal entity alignment in hyperbolic space. *Neurocomputing*, 2021, **461**: 598–607
- Wang Z C, Lv Q S, Lan X H, Zhang Y. Cross-lingual knowledge graph alignment via graph convolutional networks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL, 2018. 349–357
- Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR, 2015.
- He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770–778
- Chen M H, Tian Y T, Yang M H, Zaniolo C. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: IJCAI.org, 2017. 1511–1517
- Sun Z Q, Hu W, Zhang Q H, Qu Y Z. Bootstrapping entity alignment with knowledge graph embedding. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: IJCAI.org, 2018. 4396–4402
- Chen L Y, Li Z, Wang Y J, Xu T, Wang Z F, Chen E H. MMEA: Entity alignment for multi-modal knowledge graph. In: Proceedings of the 13th International Conference on Knowledge Science, Engineering and Management. Hangzhou, China: Springer, 2020. 134–147
- Guo L B, Sun Z Q, Hu W. Learning to exploit long-term relational dependencies in knowledge graphs. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 2505–2514
- Zhuang Yan, Li Guo-Liang, Feng Jian-Hua. A survey on entity alignment of knowledge base. *Journal of Computer Research and Development*, 2016, **53**(1): 165–192
(庄严, 李国良, 冯建华. 知识库实体对齐技术综述. 计算机研究与发展, 2016, **53**(1): 165–192)
- Qiao Jing-Jing, Duan Li-Guo, Li Ai-Ping. Entity alignment algorithm based on multi-features. *Computer Engineering and Design*, 2018, **39**(11): 3395–3400
(乔晶晶, 段利国, 李爱萍. 融合多种特征的实体对齐算法. 计算机工程与设计, 2018, **39**(11): 3395–3400)
- Trisedya B D, Qi J Z, Zhang R. Entity alignment between knowledge graphs using attribute embeddings. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu,

- USA: AAAI Press, 2019. 297–304
- 24 Zhu H, Xie R B, Liu Z Y, Sun M S. Iterative entity alignment via joint knowledge embeddings. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: IJCAI.org, 2017. 4258–4264
- 25 Chen M H, Tian Y T, Chang K W, Skiena S, Zaniolo C. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: IJCAI.org, 2018. 3998–4004
- 26 Cao Y X, Liu Z Y, Li C J, Liu Z Y, Li J Z, Chua T S. Multi-channel graph neural network for entity alignment. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019. 1452–1461
- 27 Li C J, Cao Y X, Hou L, Shi J X, Li J Z, Chua T S. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: ACL, 2019. 2723–2732
- 28 Mao X, Wang W T, Xu H M, Lan M, Wu Y B. MRAEA: An efficient and robust entity alignment approach for cross-lingual knowledge graph. In: Proceedings of the 13th International Conference on Web Search and Data Mining. Houston, USA: ACM, 2020. 420–428
- 29 Sun Z Q, Hu W, Li C K. Cross-lingual entity alignment via joint attribute-preserving embedding. In: Proceedings of the 16th International Semantic Web Conference on the Semantic Web (ISWC). Vienna, Austria: Springer, 2018. 628–644
- 30 Galárraga L, Razniewski S, Amarilli A, Suchanek F M. Predicting completeness in knowledge bases. In: Proceedings of the 10th ACM International Conference on Web Search and Data Mining. Cambridge, United Kingdom: ACM, 2017. 375–383
- 31 Ferrada S, Bustos B, Hogan A. IMGpedia: A linked dataset with content-based analysis of Wikimedia images. In: Proceedings of the 16th International Semantic Web Conference on the Semantic Web (ISWC). Vienna, Austria: Springer, 2017. 84–93
- 32 Xie R B, Liu Z Y, Luan H B, Sun M S. Image-embodied knowledge representation learning. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: IJCAI.org, 2017. 3140–3146
- 33 Mousselly-Sergieh H, Botschen T, Gurevych I, Roth S. A multimodal translation-based approach for knowledge graph representation learning. In: Proceedings of the 7th Joint Conference on Lexical and Computational Semantics. New Orleans, USA: ACL, 2018. 225–234
- 34 Tan H, Bansal M. LXMERT: Learning cross-modality encoder representations from transformers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: ACL, 2019. 5100–5111
- 35 Li L H, Yatskar M, Yin D, Hsieh C J, Chang K W. What does BERT with vision look at? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual Event: ACL, 2020. 5265–5275
- 36 Wang H R, Zhang Y, Ji Z, Pang Y W, Ma L. Consensus-aware visual-semantic embedding for image-text matching. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 18–34
- 37 Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 740–755
- 38 Plummer B A, Wang L W, Cervantes C M, Caicedo J C, Hockenmaier J, Lazebnik S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 2641–2649
- 39 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1137–1149
- 40 Schuster M, Paliwal K K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 1997, **45**(11): 2673–2681
- 41 Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 1025–1035
- 42 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net, 2017.
- 43 Wu Y T, Liu X, Feng Y S, Wang Z, Yan R, Zhao D Y. Relation-aware entity alignment for heterogeneous knowledge graphs. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2019. 5278–5284
- 44 Xing W P, Ghorbani A. Weighted pagerank algorithm. In: Proceedings of the 2nd Annual Conference on Communication Networks and Services Research. Fredericton, Canada: IEEE, 2004. 305–314
- 45 Zhang Q H, Sun Z Q, Hu W, Chen M H, Guo L B, Qu Y Z. Multi-view knowledge graph embedding for entity alignment. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: IJCAI.org, 2019. 5429–5435
- 46 Pang N, Zeng W X, Tang J Y, Tan Z, Zhao X. Iterative entity alignment with improved neural attribute embedding. In: Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG2019) Co-located With the 16th Extended Semantic Web Conference (ESWC). Portorož, Slovenia: CEUR-WS, 2019. 41–46
- 47 Huang B, Yang F, Yin M X, Mo X Y, Zhong C. A review of multimodal medical image fusion techniques. *Computational and Mathematical Methods in Medicine*, 2020, **2020**: Article No. 8279342
- 48 Atrey P K, Hossain M A, El Saddik A, Kankanhalli M S. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 2010, **16**(6): 345–379
- 49 Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 2017, **37**: 98–125



fusion.)

郭浩 国防科技大学博士研究生。主要研究方向为知识图谱构建与融合技术。E-mail: guo_hao@nudt.edu.cn (GUO Hao Ph.D. candidate at National University of Defense Technology. His research interest covers knowledge graph construction and



李欣奕 博士, 国防科技大学讲师. 主要研究方向为自然语言处理和信息检索. 本文通信作者.

E-mail: lixinyimichael@163.com

(**LI Xin-Yi** Ph.D., lecturer at National University of Defense Technology. His research interest covers

natural language processing and information retrieval. Corresponding author of this paper.)



郭延明 国防科技大学副教授. 主要研究方向为深度学习, 跨媒体信息处理与智能博弈对抗.

E-mail: guoyanming@nudt.edu.cn

(**GUO Yan-Ming** Associate professor at National University of Defense Technology. His research interest covers deep learning, cross-media information

processing, and adversarial intelligent game.)



唐九阳 国防科技大学教授. 主要研究方向为智能分析, 大数据和社会计算. E-mail: 13787319678@163.com

(**TANG Jiu-Yang** Professor at National University of Defense Technology. His research interest covers intelligence analysis, big data, and

social computing.)



赵翔 国防科技大学教授. 主要研究方向为图数据管理与挖掘和智能分析.

E-mail: xiangzhao@nudt.edu.cn

(**ZHAO Xiang** Professor at National University of Defense Technology. His research interest covers

graph data management and mining, and intelligence analysis.)