http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0709

无对齐实体场景的多语言知识图谱补全

唐榕氚[1,2]，徐秋程[2]，汤闻易[2]，翟飞飞[1,3]，周玉[1,3,*]

(1. 中国科学院 自动化研究所，北京 100190；2. 南京航空航天大学 民航学院 空中交通管理系统全国重点实验室，南京 210007；
3. 北京中科凡语科技有限公司，北京 100190)

摘 要：多语言知识图谱补全(MKGC)旨在利用其他语言知识图谱的信息增强目标语言知识图谱上的链接预测性能。现有方法通常利用不同语言知识图谱之间预先对齐的实体对作为知识迁移的媒介，然而在实际场景中，不同语言知识图谱之间通常没有预先对齐的实体，导致难以实现知识迁移。针对上述无对齐实体场景，提出一种融合预训练语言模型信息的伪对齐实体生成模块，不断迭代生成新的对齐实体进行知识迁移。为区分不同语言知识图谱中信息对目标语言知识图谱的贡献度，提出一种基于多图注意力的图神经网络(MGA-GNN)用于对三元组进行编码，通过该网络输出的嵌入表征计算得到三元组的合理性得分，完成链接预测任务。为验证所提方法的有效性，在2个公开数据集DBP-5L和E-PKG上进行了实验验证，结果表明：所提方法在多个语言知识图谱上链接预测的性能超过了有对齐实体的MKGC方法，证明了该方法在更加实际场景下的优越性能。

关 键 词：多语言知识图谱补全；实体对齐；多图注意力；图神经网络；链接预测
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2026)01-0252-08

具有丰富结构化知识的知识图谱(knowledge graph, KG)被广泛应用于各种人工智能领域，如问答系统[1-2]、推荐系统[3-4]和信息抽取[5]等。在知识图谱中，知识以三元组的形式进行存储，包含头实体、关系和尾实体，知识图谱补全(knowledge graph completion, KGC)任务也称链接预测，旨在根据已有知识预测未知三元组中缺失的部分以进行知识的扩充，缓解知识图谱普遍存在的不完备问题。通常，KGC只在一个知识图谱上进行，而考虑到现实世界中存在许多不同语言的知识图谱，这些知识图谱之间存在大量的知识重叠和特有知识。为了利用不同语言的知识图谱信息来增强目标知识图谱的链接预测性能，近年来，多语言知识图谱补全(multilingual KGC, MKGC)引起了许多研究者的关注。在MKGC任务中，对于待补全的目标知识图谱，如何将其他语言的知识迁移到该知识图谱上成为了主要难点。现有工作，如KEnS[6]、AlignKGC[7]、SS-AGA[8]都将不同知识图谱之间存在的预先对齐的实体作为知识迁移的桥梁，通过这些实体将其他语言的知识聚集到目标知识图谱上。然而，在更加符合实际的场景下，通常只能获取不同语言的知识图谱，而无法直接获取这些知识图谱之间实体的对齐情况。但是，很少有研究关注无对齐实体场景下的MKGC任务。本文力图设计一种简单而有效的方法来实现无对齐实体场景下的MKGC，使其能够更好地在实际场景下进行应用。

KGC主要采用基于知识图谱嵌入的方法，如TransE[9]、DistMult[10]、RotatE[11]。上述方法先将实体和关系嵌入到低维向量空间中，为三元组建立不同的几何关系，然后设计不同的评分函数来计算三

收稿日期：2023-10-31；**录用日期**：2023-12-29；**网络出版时间**：2024-03-12 11:05
网络出版地址：link.cnki.net/urlid/11.2625.V.20240311.1607.006
基金项目：空中交通管理系统全国重点实验室开放课题(SKLATM202203)
*通信作者. E-mail：yzhou@nlpr.ia.ac.cn

引用格式：唐榕氚，徐秋程，汤闻易，等. 无对齐实体场景的多语言知识图谱补全[J]. 北京航空航天大学学报，2026，52(1)：252-259.
TANG R C，XU Q C，TANG W Y，et al. Multilingual knowledge graph completion without aligned entity pairs[J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52(1)：252-259 (in Chinese).

元组的合理性分数。由于知识图谱特有的图结构, 基于图神经网络(graph neural network, GNN)的方法在 KGC 任务上受到了广泛关注[12-13], 具有强大图结构建模能力的 GNN 在此任务上展示出了巨大潜力。

现有 MKGC 相关的方法常常与知识图谱的实体对齐任务紧密关联, 而跨语言知识图谱的实体对齐方法常常利用给定的对齐实体作为监督信号指导模型的训练[14-16], 因此, 当前的 MKGC 任务均在有对齐实体场景下才能实现。KEnS[6] 首次通过集成的方法利用多种不同语言知识图谱的互补知识提升了单语言 KGC 的性能。AlignKGC[7] 在多个语言的知识图谱上联合进行补全、实体对齐和关系对齐, 并提高了 KGC 的准确性和实体对齐性能。SS-AGA[8] 通过使用 GNN 动态添加对齐实体来改进 MKGC 任务。然而, 上述方法均需要预先对齐的实体对, 限制了进一步的应用。

本文提出一种不需要预先对齐实体的 MKGC 方法。首先, 设计了翻译模块和预训练语言模型相结合的对齐实体获取模块, 能够在初始阶段获取高质量的对齐实体。其次, 设计了基于多图注意力的图神经网络(graph neural network based on multi-graph attention, MGA-GNN), 并以此分别构建对齐模型和知识模型, MGA-GNN 能将对齐模型和预训练语言模型的表征结合, 迭代生成新的对齐实体, 使用知识模型将不同来源知识图谱的信息聚合到目标知识图谱上, 提升 MKGC 的性能。

1　本文方法

1.1　预备知识和相关符号

知识图谱可以用$G=(E,R,T)$表示, 其中, E表示实体的集合, R表示实体与实体之间的关系集合, T表示实体与关系组成的三元组集合。每条知识用一个三元组$(h,r,t)\in T$表示, 其中, $h\in E$表示头实体, $r\in R$表示关系, $t\in E$表示尾实体。

KGC 任务的目标是基于当前知识图谱中存在的事实来预测出新的事实三元组。具体而言, 给定$(h,r,?)$或$(?,r,t)$这样的问题, 问号表示缺失的实体, 模型需要推理出缺失的尾实体或头实体。

对于 MKGC 任务, 可以利用多个使用不同语言构建的知识图谱。假设有N个知识图谱$G_1,G_2,\cdots,G_N$, 每个由不同的语言进行构建。在有已知对齐实体的场景下, 任意 2 个知识图谱$G_i=(E_i,R_i,T_i)$和$G_j=(E_j,R_j,T_j)$之间存在一些已知的对齐实体对$\{(\boldsymbol{e}_i,\boldsymbol{e}_j):\boldsymbol{e}_i\in E_i,\boldsymbol{e}_j\in E_j\}$, 其中, $\boldsymbol{e}$表示实体。对于本文所关注的无对齐实体场景, 无法获取真实的对齐实体对。另外, 对于这 2 类场景有一个关于关系集合的约束, 即所有知识图谱中的关系集合都属于一个公共的关系集合R。该约束表示每个知识图谱中的关系集合与语言无关, 即对于任意的$i\in 1,2,\cdots,N$, $R_i\in R$。

1.2　整体框架

目前已有的 MKGC 框架, 如 KEnS[6] 和 SS-AGA[8], 只能用于有对齐实体场景下的 MKGC, 为了实现无对齐实体场景下的 MKGC 任务, 本文构建了无对齐实体的 MKGC 框架, 整体结构如图 1 所示。该框架主要包含 2 个部分: 对齐实体获取模块和 GNN 编码模块。

1.3　对齐实体获取模块

对于 MKGC, 对齐实体起着连接不同知识图谱之间知识的桥梁作用, 但在实际场景中通常无法直接获取真实的对齐实体, 因此, 本文首先构建一个对齐实体获取模块, 目标是寻找出不同知识图谱之间的对齐实体, 进而为后续的知识迁移服务。知识图谱之间实体对齐的任务已经得到广泛研究[17-18], 但是常需要一定数量的真实对齐实体对作为监督信号来训练模型, 这不符合本文研究的场景。为此, 本文提出一个简单而有效的初始对齐实体获取方法。该方法包括 2 个阶段: 对齐实体初始获取和对齐实体迭代获取。

1.3.1　对齐实体初始获取

在任务初始阶段, 无预先给定的对齐实体, 而对于给定的多个不同语言的知识图谱, 每个实体的文字名称和文字描述是容易获取的, 将每个实体的这些文字信息经过预训练语言模型 mBERT[19] 进行编码, 得到实体的文本特征。mBERT[19] 是一个多语言预训练语言模型, 能够捕捉实体文本中丰富的语义信息。随后, 对于任意 2 个知识图谱中的实体$\boldsymbol{e}_i\in E_i$和$\boldsymbol{e}_j\in E_j$, 按照式(1)和式(2)计算它们之间的跨域相似性局部缩放(cross-domain similarity local scaling, CSLS)[20] 距离$\mathrm{CSLS}(\boldsymbol{e}_i,\boldsymbol{e}_j)$, 如果两者在 CSLS 距离上互为最近邻, 说明其语义信息相近, 即很可能是对齐实体, 故将这 2 个实体添加进伪对齐实体对集合中(后文方法获取的对齐实体均指伪对齐实体)。

$$\mathrm{CSLS}(\boldsymbol{e}_i,\boldsymbol{e}_j)=2\cos(\boldsymbol{e}_i^{\mathrm{t}},\boldsymbol{e}_j^{\mathrm{t}})-r_j(\boldsymbol{e}_i)-r_i(\boldsymbol{e}_j)\tag{1}$$

$$r_j(\boldsymbol{e}_i)=\frac{1}{K}\sum_{\boldsymbol{e}_j\in N_j(\boldsymbol{e}_i)}\cos(\boldsymbol{e}_i^{\mathrm{t}},\boldsymbol{e}_j^{\mathrm{t}})\tag{2}$$

式中: $\boldsymbol{e}_i^{\mathrm{t}}$和$\boldsymbol{e}_j^{\mathrm{t}}$分别为$\boldsymbol{e}_i$和$\boldsymbol{e}_j$对应的文本特征表示; $N_j(\boldsymbol{e}_i)$表示在向量空间中使用实体$\boldsymbol{e}_i$的文本特征向量$\boldsymbol{e}_i^{\mathrm{t}}$与知识图谱$G_j$中所有实体的文本特征向量计算距离, 选取$K$个最近邻的实体组成的集合; K为

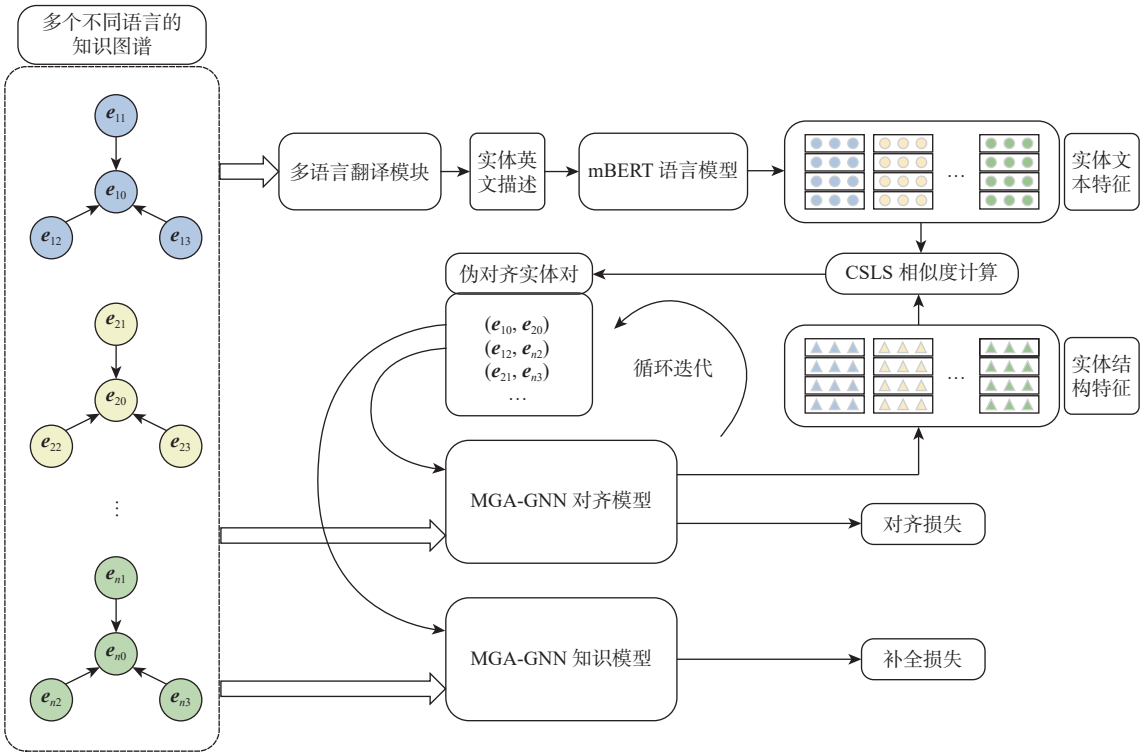

图 1 无对齐实体的多语言知识图谱补全框架

Fig. 1 The framework of multilingual knowledge graph completion without alignment pairs

$N_j(\boldsymbol{e}_i)$中实体的数量，作为超参数进行设置。

虽然经过 mBERT 编码得到的实体向量在不同语言之间已经有一定的对齐，但不足以生成足够的高质量对齐实体，因此，本文使用一个高质量的多语言翻译模块将不同语言知识图谱内实体的文本信息翻译为英文的表述，再将该英文表述经过 mBERT 编码。这里翻译到英文的原因有 2 点：①相比于翻译到其他语言，目前常见的多语言翻译模块翻译到英文的性能更高；②由于训练语料的原因，mBERT 对英文的表征能力更强。通过上述流程，可以大大提高对齐实体初始获取的质量和数量。

1.3.2 对齐实体迭代获取

初始阶段获取的对齐实体数量较少，因此，在模型训练阶段使用一个对齐模型不断迭代生成新的对齐实体。对齐模型的具体结构将在 1.4 节进行详细说明，实体$\boldsymbol{e}_i$和$\boldsymbol{e}_j$经过对齐模型编码后得到对应的结构特征表示记为$\boldsymbol{e}_i^{\mathrm{s}}$和$\boldsymbol{e}_j^{\mathrm{s}}$。随后将文本特征与结构特征相结合，按照式(3)和式(4)计算 2 个知识图谱之间所有实体之间的 CSLS 距离，如果两者在 CSLS 距离上互为最近邻，则将这 2 个实体添加进对齐实体对集合中。

$$\mathrm{CSLS}(\boldsymbol{e}_i,\boldsymbol{e}_j)=2\max(\cos(\boldsymbol{e}_i^{\mathrm{s}},\boldsymbol{e}_j^{\mathrm{s}}),\cos(\boldsymbol{e}_i^{\mathrm{t}},\boldsymbol{e}_j^{\mathrm{t}}))-r_j(\boldsymbol{e}_i)-r_i(\boldsymbol{e}_j) \tag{3}$$

$$r_j(\boldsymbol{e}_i)=\frac{1}{K}\sum_{\boldsymbol{e}_j\in N_j(\boldsymbol{e}_i)}\max(\cos(\boldsymbol{e}_i^{\mathrm{s}},\boldsymbol{e}_j^{\mathrm{s}}),\cos(\boldsymbol{e}_i^{\mathrm{t}},\boldsymbol{e}_j^{\mathrm{t}})) \tag{4}$$

1.4 图神经网络编码模块

GNN 对图结构数据具有强大的编码表征能力，为了更好实现多个知识图谱之间知识互补的能力，本文设计了 MGA-GNN。基于该网络，构建了一个用于实体对齐的模型和一个用于 MKGC 的模型。网络构建分为 2 个步骤：联合建图和 GNN 编码。

1.4.1 联合建图

给定的N个知识图谱为独立不连通的图，无法直接通过 GNN 进行有效的信息传播。本文利用 1.3.1 节中初始获取的对齐实体将N个独立的图连接为一个统一的图。具体而言，在关系集合中增加一个对齐关系，将对齐实体对$(\boldsymbol{e}_i,\boldsymbol{e}_j)$转换为三元组$(\boldsymbol{e}_i,\text{对齐},\boldsymbol{e}_j)$，通过增加的这些三元组将所有知识图谱连接起来，得到大图G。在该统一的大图基础上，对所有知识图谱中的实体和关系在同一个向量空间中进行初始化，得到初始的实体向量表示$\boldsymbol{h}^0$和关系向量表示$\boldsymbol{r}'$，特征维度为d。

1.4.2 GNN 编码

建立图G和初始化实体关系向量后，采用多层 GNN 对实体和关系进行编码。从第l层网络到第$l+1$层网络的传递过程如下：

$$\boldsymbol{h}_i^{l+1}=\boldsymbol{h}_i^{l}+\sigma\left(\sum_{\boldsymbol{e}_j\in N(\boldsymbol{e}_i)}\mathrm{Att}(\boldsymbol{h}_i^{l},\boldsymbol{h}_{j(r)}^{l})\boldsymbol{h}_{j(r)}^{l}\right) \tag{5}$$

式中：$\boldsymbol{h}_i^{l}$和$\boldsymbol{h}_i^{l+1}$分别为中心实体$\boldsymbol{e}_i$在第l层和第$l+1$网络的隐层表示；$\sigma(\cdot)$为 GELU[21] 实现的非线性激

活函数; $N(\boldsymbol{e}_i)$为实体$\boldsymbol{e}_i$的邻域实体集合; $\boldsymbol{e}_j$为实体$\boldsymbol{e}_i$的邻居实体; $\boldsymbol{h}^l_{j(r)}$为实体$\boldsymbol{e}_j$经过与关系r关联之后的隐层表示; $\mathrm{Att}(\boldsymbol{h}^l_i,\boldsymbol{h}^l_{j(r)})$为注意力函数, 用于计算中心实体对邻居实体的注意力权重。

在传统的基于查询-键-值的图注意力网络中, 邻居节点对中心节点的贡献采用同一套查询-键-值的变换矩阵。然而对于 MKGC, 不同语言的知识图谱对于目标知识图谱的知识补充能力应该是不同的, 因此, 本文提出一种多图注意力机制, 将中心实体的邻居实体按照不同来源采取不同的注意力权重计算方式, 用于区分不同来源知识图谱的信息对中心实体的贡献程度。具体而言, 对于每种语言的知识图谱, 各自维护不同的查询-键-值变换矩阵 $\boldsymbol{W}_{\mathrm{q}}$、$\boldsymbol{W}_{\mathrm{k}}$、$\boldsymbol{W}_{\mathrm{v}}$, 随后, 注意力权重计算方式如下:

$$\mathrm{Att}(\boldsymbol{h}^l_i,\boldsymbol{h}^l_{j(r)})=\frac{\exp(\alpha^r_{ij})}{\sum\limits_{\boldsymbol{e}_j\in N(\boldsymbol{e}_i)}\exp(\alpha^r_{ij})} \tag{6}$$

$$\alpha^r_{ij}=\frac{\lambda\beta_r}{\sqrt{d}}(\boldsymbol{W}_{\mathrm{q}i}\boldsymbol{h}^l_i)^{\mathrm{T}}(\boldsymbol{W}_{\mathrm{k}j}\boldsymbol{h}^l_{j(r)}) \tag{7}$$

$$\boldsymbol{h}^l_{j(r)}=\boldsymbol{W}_{\mathrm{v}j}[\boldsymbol{h}^l_j,\boldsymbol{r}'] \tag{8}$$

式中: $\boldsymbol{W}_{\mathrm{q}i}\in\mathbf{R}^{d\times d}$、$\boldsymbol{W}_{\mathrm{k}j}\in\mathbf{R}^{d\times d}$、$\boldsymbol{W}_{\mathrm{v}j}\in\mathbf{R}^{d\times 2d}$分别为实体$\boldsymbol{e}_i$所属知识图谱对应的查询变换矩阵、实体$\boldsymbol{e}_j$所属知识图谱对应的键变换矩阵、值变换矩阵; β_r为可学习的、与关系$\boldsymbol{r}'$相对应的关系变量, 用于衡量关系$\boldsymbol{r}'$的权重; λ用于权衡通过前述方法获取的伪对齐实体的可靠性, 当关系$\boldsymbol{r}'$为对齐关系时, $\lambda=\max(\cos(\boldsymbol{e}^{\mathrm{s}}_i,\boldsymbol{e}^{\mathrm{s}}_j),\cos(\boldsymbol{e}^{\mathrm{t}}_i,\boldsymbol{e}^{\mathrm{t}}_j))$, 否则为 1; $[\cdot]$为向量拼接函数。

由于 1.3 节中得到的对齐实体存在噪声问题, 直接将伪对齐实体的信息传递给中心实体可能会造成误差传递, 为了缓解这类噪声, 在式(7)中引入对齐实体的可靠性参数λ。对于生成的伪对齐实体, 两者的相似度越高, 则认为该实体对对齐的可靠性更高, 进而通过该对齐关系传递给中心实体的信息应该更加可靠, 反之, 则应该抑制该信息传递。

基于上述 MGA-GNN, 本文采取同样的结构分别建立实体对齐模型和知识模型。实体对齐模型用于拉进对齐实体之间的距离, 进而通过训练迭代生成新的对齐实体, 并将新的对齐实体添加进大图G中, 不断提高多个知识图谱之间知识互补的能力; 知识模型用于对所有实体和关系进行编码, 以 KGC 为目标进行训练, 最终预测三元组中缺失的实体。

1.5　模型训练

对齐模型训练的目标是拉近对齐实体之间的距离, 故对于任意 2 个知识图谱G_i和G_j, 设已经生成的它们之间的对齐实体集合为P_{ij}, 则对齐损失函数采用如下对比损失:

$$L^{ij}_{\mathrm{A}}=\sum_{\substack{(\boldsymbol{e}_i,\boldsymbol{e}_j)\in P_{ij}\\(\boldsymbol{e}_{i'},\boldsymbol{e}_{j'})\notin P_{ij}}}\max\left(\gamma_{\mathrm{A}}+\left\|\boldsymbol{e}_i-\boldsymbol{e}_j\right\|_2-\left\|\boldsymbol{e}_{i'}-\boldsymbol{e}_{j'}\right\|_2,0\right) \tag{9}$$

式中: γ_{A}为对比损失的间隔超参数。

知识模型训练的目标是能够预测三元组中缺失的实体, 本文采用 TransE[9] 模型对三元组进行合理性打分, 在知识图谱G_i上的补全损失函数为

$$L^i_{\mathrm{K}}=\sum_{\substack{(h,r,t)\in T_i\\(h',r,t')\notin T_i}}\max\left(\gamma_{\mathrm{K}}-\phi(h,r,t)+\phi(h',r,t'),0\right) \tag{10}$$

$$\phi(h,r,t)=-\|\boldsymbol{e}_h+\boldsymbol{r}-\boldsymbol{e}_t\|_2 \tag{11}$$

式中: γ_{K}为补全损失的间隔超参数。

基于上述损失函数, 将对齐模型和知识模型进行联合训练, 总的损失函数为

$$L=\sum_{i=1}^{N}\sum_{j=i+1}^{N}L^{ij}_{\mathrm{A}}+\sum_{i=1}^{N}L^i_{\mathrm{K}} \tag{12}$$

2　实验准备

2.1　数 据 集

在 2 个数据集上进行实验以评估本文方法的性能, 数据集的统计数据数量如表 1 所示。DBP-5L[6] 数据集是从维基百科知识图谱 DBpedia[22] 抽取获得, 包括 5 个不同语言的知识图谱, 分别为希腊语(EL)、英语(EN)、西班牙语(ES)、法语(FR)和日语(JA)。E-PKG[8] 数据集是一个多语言的、与手机相关的电子产品知识图谱, 包括 6 种语言, 分别为德语(DE)、英语(EN)、西班牙语(ES)、法语(FR)、意大利语(IT)和日语(JA)。2 个数据集中的关系都用英语表示, 并且由不同语言的知识图谱共享。2 个数据集中具有最丰富知识的语言都是英语, 对

表 1　数据集统计数据数量

Table 1　The statistic numbers of datasets

数据集	语言	实体数	关系数	训练数据量	验证数据量	测试数据量
DBP-5L[6]	EL	5 231	111	8 670	4 152	1 017
	EN	13 996	831	48 652	24 051	7 464
	ES	12 382	144	33 036	16 220	4 810
	FR	13 176	178	30 139	14 705	4 171
	JA	11 805	128	17 979	8 633	2 162
E-PKG[8]	DE	17 223	21	45 515	22 753	7 602
	EN	16 544	21	60 310	39 150	10 071
	ES	9 595	21	18 090	9 039	3 034
	FR	17 068	21	47 999	23 994	8 022
	IT	15 670	21	42 767	21 377	7 148
	JA	2 642	21	10 013	5 002	1 688

应知识最少的语言分别为希腊语和日语。需要注意的是，DBP-5L[6]数据集给出了实体对应的名称，而E-PKG[8]数据集中的实体没有实体名称，只有实体相应的具体文字描述。

2.2 基线模型

为了验证本文方法的有效性，选取2类基线模型进行对比，分别为单语言KGC模型和有对齐实体的MKGC模型。对于单语言模型，选取经典模型TransE[9]和RotatE[11]，以及采用预训练语言模型的KG-BERT[23]。对于多语言模型，KEnS[6]首次提出了MKGC方法，并设计了一种集成方法来进行知识迁移；AlignKGC[7]将MKGC同实体和关系对齐进行联合学习；SS-AGA[8]利用已有对齐实体进行实体对齐的自监督学习，进而提升MKGC性能。

2.3 评价指标

通过尾实体预测任务来对模型性能进行评估。首先，对于测试集中的三元组(h, r, t)，给定h和r，将其他所有实体作为候选实体计算三元组合理性得分，根据该得分进行降序排序，以此预测尾实体t。然后，根据该排序计算得到评价指标Hits@1、Hits@10和平均倒数排名（mean reciprocal ranks, MRR）。其中，Hits@1和Hits@10分别表示正确实体在该排序中排名为1和排在前10的比例，MRR表示所有测试样例排名倒数的平均值。另外，在测试阶段，沿用文献[9]中的过滤设置，即将训练集、验证集和测试集中已知三元组的得分从排序中删掉。

2.4 参数设置

实验中使用的翻译模块为谷歌翻译，另外，采用网格搜索法对超参数进行调优，各个超参数的搜索范围为：实体和关系嵌入维度{128,256}，网络层数{2,3}，间隔参数{0.2,0.3,0.5}，学习率{0.000 1,0.001,0.005,0.01}。最终参数设置为：实体和关系嵌入维度为256，网络层数为2，学习率为0.01，在DBP-5L数据集上间隔参数为0.5，在E-PKG上间隔参数为0.3，训练的批量大小为200，使用Adam优化器对网络参数进行优化。

3 实验分析

3.1 实验结果

在数据集DBP-5L和E-PKG上的实验结果分别如表2和表3所示。可以看到，MKGC方法的结果普遍优于单语言KGC方法，这验证了不同语言知识图谱之间可以进行知识互补。可以发现，在DBP-5L数据集上，本文方法在4个语言的知识图谱上取得了最好性能，在E-PKG数据集上也获得了

表2 DBP-5L数据集实验结果

Table 2 Experimental results on the DBP-5L dataset

方法分类	方法	语言	H@1	H@10	MRR
单语言KGC方法	TransE[9]	EL	13.1	43.7	24.3
		EN	7.3	29.3	16.9
		ES	13.5	45.0	24.4
		FR	17.5	48.8	27.6
		JA	21.1	48.5	25.3
	RotatE[11]	EL	14.5	36.2	26.2
		EN	12.3	30.4	20.7
		ES	21.2	53.9	33.8
		FR	23.2	55.5	35.1
		JA	26.4	60.2	39.8
	KG-BERT[23]	EL	17.3	40.1	27.3
		EN	12.9	31.9	21.0
		ES	21.9	54.1	34.0
		FR	23.5	55.9	35.4
		JA	26.9	59.8	38.7
有对齐实体的MKGC方法	KEnS[6]	EL	28.1	56.9	
		EN	15.1	39.8	
		ES	23.6	60.1	
		FR	25.5	62.9	
		JA	32.1	65.3	
	AlignKGC[7]	EL	27.6	56.2	33.8
		EN	15.5	39.2	22.3
		ES	24.2	60.9	35.1
		FR	24.1	62.3	37.4
		JA	31.6	64.3	41.6
	SS-AGA[8]	EL	30.8	58.6	35.3
		EN	16.3	41.3	23.1
		ES	25.5	61.9	36.6
		FR	27.1	65.5	38.3
		JA	**34.6**	**66.9**	**42.9**
无对齐实体的MKGC方法	MGA-GNN	EL	**33.6**	**77.1**	**49.3**
		EN	**19.5**	**59.8**	**33.7**
		ES	**26.1**	**64.7**	**39.3**
		FR	**31.2**	**70.5**	**44.6**
		JA	28.5	60.6	39.8

注：加粗数值表示性能最优。

其中4个语言的知识图谱上最高的MRR和其中5个语言的知识图谱上最高的Hits@1和Hits@10。这主要得益于2点：①对齐实体生成模块提供了足够的对齐实体，使在目标知识图谱上进行补全推理时，可以通过这些对齐实体获取其他知识图谱中的知识；②提出的MGA-GNN充分考虑了邻域实体的来源多样性，而不是按照单一语言进行建模。

3.2 实体对齐分析

DBP-5L和E-PKG数据集本身是给有对齐实体场景下的MKGC建立的，因此，数据集提供了一部

表 3　E-PKG 数据集实验结果

Table 3　Experimental results on the E-PKG dataset

方法分类	方法	语言	H@1	H@10	MRR
单语言KGC方法	TransE[9]	DE	21.2	65.5	37.4
		EN	23.2	67.5	39.4
		ES	17.2	58.4	33.0
		FR	20.8	66.9	37.5
		IT	22.0	63.8	37.8
		JA	25.1	72.7	43.6
	RotatE[11]	DE	22.3	64.3	38.2
		EN	24.2	66.8	40.0
		ES	18.3	58.9	33.7
		FR	22.1	64.3	38.2
		IT	22.5	64.0	38.1
		JA	26.3	71.9	41.8
	KG-BERT[23]	DE	21.8	64.7	38.4
		EN	24.3	66.4	39.6
		ES	18.7	58.8	33.2
		FR	22.3	67.2	38.3
		IT	22.9	63.7	37.2
		JA	26.9	72.4	44.1
有对齐实体的MKGC方法	KEnS[6]	DE	24.3	65.8	
		EN	26.2	69.5	
		ES	21.3	59.5	
		FR	25.4	68.2	
		IT	25.1	64.6	
		JA	33.5	73.6	
	AlignKGC[7]	DE	22.1	65.1	38.5
		EN	25.6	68.3	40.5
		ES	19.4	59.1	34.2
		FR	22.8	67.2	38.8
		IT	24.2	63.4	37.3
		JA	31.2	72.3	46.2
	SS-AGA[8]	DE	24.6	66.3	39.4
		EN	26.7	69.8	41.5
		ES	21.0	60.1	**36.3**
		FR	**25.9**	68.7	**40.2**
		IT	24.9	63.8	38.4
		JA	33.9	**74.1**	48.3
无对齐实体的MKGC方法	MGA-GNN	DE	**24.9**	**68.4**	**40.3**
		EN	**29.3**	**70.4**	**44.4**
		ES	**22.2**	**60.3**	35.6
		FR	24.1	**68.9**	39.6
		IT	**26.1**	**65.9**	**40.9**
		JA	**43.0**	73.2	**54.5**

注: 加粗数值表示性能最优。

分已对齐的实体对, 为了验证本文对齐实体获取模块的有效性, 表 4 和表 5 给出了最终通过该模块获取的对齐实体对数据集中给定的对齐实体的覆盖程度。精确率表示生成的对齐实体对在已给对齐实体对中的正确率, 召回率表示生成的对齐实体对对已给对齐实体对的召回比例。

表 4　DBP-5L 数据集中生成实体精确率和召回率

Table 4　The precision and recall of the generated aligned entities on the DBP-5L dataset　%

语言对	精确率	召回率
EL-JA	97.1	25.4
EL-ES	98.6	68.4
EL-EN	99.2	66.8
EL-FR	99.0	64.6
JA-ES	97.6	22.1
JA-EN	98.0	25.3
JA-FR	98.6	21.2
ES-EN	99.7	74.5
ES-FR	99.6	70.7
EN-FR	99.7	71.1

表 5　E-PKG 数据集中生成实体精确率和召回率

Table 5　The precision and recall of the generated aligned entities on the E-PKG dataset　%

语言对	精确率	召回率
DE-EN	58.5	24.5
DE-ES	51.9	21.0
DE-FR	56.1	21.4
DE-IT	55.6	22.2
DE-JA	60.9	16.3
ES-EN	51.7	23.5
ES-FR	50.1	22.8
ES-IT	55.0	25.2
ES-JA	50.6	16.0
FR-EN	56.1	23.9
FR-IT	58.7	25.2
FR-JA	58.6	16.6
IT-EN	60.3	24.3
IT-JA	61.6	15.8
JA-EN	42.8	17.7

可以发现, DBP-5L 数据集中的对齐精确率都较高, 这表明对齐实体模块获取的对齐实体较为精确, 为不同语言知识图谱之间的知识迁移起了很好的支撑作用。另外可以发现, DBP-5L 数据集中有关日语知识图谱的实体对齐的召回率都比较低, 均不超过 30%, 这是因为从日语翻译到英语的准确率不如其他 3 种语言(希腊语、西班牙语、法语), 导致获取的对齐实体数量不够, 这也解释了在 DBP-5L 数据集上本文方法在日语知识图谱上补全结果不如有对齐实体的方法。

对于 E-PKG 数据集, 从表 5 可以发现, 对齐实体的精确率均在 55% 左右, 远低于 DBP-5L 数据集

中的结果，这是因为E-PKG数据集中没有实体的文本名称，只有对于实体的文本描述，而不同语言中对相同实体的描述侧重不同，引入的语言差异也给翻译模块和文本表征模型带来了困难，导致精确率较低，同样召回率也不高。本文方法在多个知识图谱上的补全结果仍超越了其它模型，证明了本文方法的有效性。

4 结 论

1）本文提出了一种由伪对齐实体生成与多语言GNN编码组成的双模块框架，有效解决了多语言知识图谱补全中缺失实体对齐的关键瓶颈问题。

2）通过融合预训练语言模型与迭代式对齐扩展机制，实现了无需种子对齐实体的跨语言知识自动迁移，提升了方法在实际场景中的适用性。设计的多图注意力编码机制能够区分不同语言知识图谱的信息贡献差异，实现了跨语言知识的选择性融合与针对性表征。

3）在完全无对齐实体的实验设定下，本文方法不仅实现了有效的知识迁移，其性能甚至超越了依赖已知对齐实体的传统多语言知识图谱补全方法，为跨语言知识融合提供了不依赖预设对齐资源的新思路。

参考文献（References）

[1] REN H Y, DAI H J, DAI B, et al. LEGO: latent execution-guided reasoning for multi-hop question answering on knowledge graphs [C]//Proceedings of the 38th International Conference on Machine Learning. San Diego: PMLR, 2021, 139: 8959-8970.

[2] SAXENA A, KOCHSIEK A, GEMULLA R. Sequence-to-sequence knowledge graph completion and question answering[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 2814-2828.

[3] WANG X, HUANG T L, WANG D X, et al. Learning intents behind interactions with knowledge graph for recommendation[C]// Proceedings of the Web Conference 2021. New York: ACM, 2021: 878-887.

[4] WANG X T, LIU K P, WANG D J, et al. Multi-level recommendation reasoning over knowledge graphs with reinforcement learning [C]//Proceedings of the ACM Web Conference 2022. New York: ACM, 2022: 2098-2108.

[5] LIN Q K, LIU J, XU F Z, et al. Incorporating context graph with logical reasoning for inductive relation prediction[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2022: 893-903.

[6] CHEN X L, CHEN M H, FAN C J, et al. Multilingual knowledge graph completion via ensemble knowledge transfer[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: ACL, 2020: 3227-3238.

[7] SINGH H, JAIN P, MAUSAM, et al. Multilingual knowledge graph completion with joint relation and entity alignment[EB/OL]. (2021-04-18)[2023-10-29]. https://arxiv.org/abs/2104.08804.

[8] HUANG Z J, LI Z, JIANG H M, et al. Multilingual knowledge graph completion with self-supervised adaptive graph alignment[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 474-485.

[9] BORDES A, USUNIER N, GARCIA-DURÁN A, et al. Translating embeddings for modeling multi-relational data[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. New York: ACM, 2013: 2787-2795.

[10] YANG B S, YIH W T, HE X D, et al. Embedding entities and relations for learning and inference in knowledge bases[C]//Proceedings of the International Conference on Learning Representations. Phoenix: ICLR, 2014: 2768038.

[11] SUN Z Q, DENG Z H, NIE J Y, et al. RotatE: knowledge graph embedding by relational rotation in complex space[EB/OL]. (2019-02-26)[2023-10-29]. https://arxiv.org/abs/1902.10197.

[12] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]//Proceedings of the Semantic Web. Berlin: Springer, 2018: 593-607.

[13] ZHANG Y Q, YAO Q M. Knowledge graph reasoning with relational digraph[C]//Proceedings of the ACM Web Conference. New York: ACM, 2022: 912-924.

[14] CHEN M H, TIAN Y T, YANG M H, et al. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment[C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. New York: ACM, 2017: 1511-1517.

[15] GUO L B, ZHANG Q, SUN Z Q, et al. Understanding and improving knowledge graph embedding for entity alignment[C]//Proceedings of the 39th International Conference on Machine Learning. San Diego: PMLR, 2022, 162: 8145-8156.

[16] ZHU Y, LIU H Z, WU Z H, et al. Relation-aware neighborhood matching model for entity alignment[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(5): 4749-4756.

[17] SUN Z Q, WANG C M, HU W, et al. Knowledge graph alignment network with gated multi-hop neighborhood aggregation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 222-229.

[18] YAN Y C, LIU L H, BAN Y K, et al. Dynamic knowledge graph alignment[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(5): 4564-4572.

[19] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24)[2023-10-29]. https://arxiv.org/abs/1810.04805.

[20] CONNEAU A, LAMPLE G, RANZATO M, et al. Word translation without parallel data[EB/OL]. (2018-01-30)[2023-10-29]. https://arxiv.org/abs/1710.04087.

[21] HENDRYCKS D, GIMPEL K. Gaussian error linear units (GELUs) [EB/OL]. (2023-06-06)[2023-10-29]. https://arxiv.org/abs/1606.08415.

[22] LEHMANN J, ISELE R, JAKOB M, et al. DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia[J]. Semantic

Web, 2015, 6(2): 167-195.

[23] YAO L, MAO C S, LUO Y. KG-BERT: BERT for knowledge graph completion[EB/OL]. (2019-09-11)[2023-10-29]. https://arxiv.org/abs/1909.03193.

Multilingual knowledge graph completion without aligned entity pairs

TANG Rongchuan[1, 2], XU Qiucheng[2], TANG Wenyi[2], ZHAI Feifei[1, 3], ZHOU Yu[1, 3, *]

(1. Institute of Automation, Chinese Academy of Science, Beijing 100190, China;
2. State Key Laboratory of Air Traffic Management System, College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 210007, China;
3. Beijing Fanyu Technology Co., Ltd., Beijing 100190, China)

Abstract: The goal of multilingual knowledge graph completion (MKGC) is to improve link prediction performance on the target knowledge graph by leveraging data from other language-specific knowledge graphs. Existing methods usually use pre-aligned entities between different knowledge graphs to accomplish knowledge transfer. However, there are usually no pre-aligned entities between different knowledge graphs in practical scenarios, making knowledge transfer difficult to achieve. Considering the MKGC without aligned entity pairs, a pseudo-aligned entity generation module that integrates a pre-trained language model is proposed to iteratively generate new aligned entities for knowledge transfer. It is suggested to use a graph neural network based on multi-graph attention (MGA-GNN) to encode the triples in order to differentiate the information in various language-specific wisdom graphs. Finally, the plausibility of the triples is calculated via the embeddings output by the network to conduct the link prediction task. Experimental results on the DBP-5L and E-PKG datasets show the effectiveness of the proposed method and its superior performance in more practical scenarios.

Keywords: multilingual knowledge graph completion; entity alignment; multi-graph attention; graph neural network; link prediction

Received: 2023-10-31; **Accepted**: 2023-12-29; **Published Online**: 2024-03-12 11:05

URL: link.cnki.net/urlid/11.2625.V.20240311.1607.006

Foundation item: Open Research Projects of State Key Laboratory of Air Traffic Management System (SKLATM202203)

* **Corresponding author**. E-mail: yzhou@nlpr.ia.ac.cn