

基于联合教学的半监督实体对齐

卢凯^{1,2,3}, 赵晶^{1,2,3}

(1. 齐鲁工业大学(山东省科学院) 山东省计算中心(国家超级计算济南中心) 算力互联网与信息安全教育部重点实验室; 2. 齐鲁工业大学(山东省科学院) 计算机科学与技术学部 大数据应用技术山东省工程研究中心; 3. 山东省工业网络和信息网络安全重点实验室 山东省基础科学研究中心(计算机科学), 山东 济南 250353)

摘要: 实体对齐旨在识别并关联不同数据源中的等效实体, 然而训练数据不足仍是该领域面临的主要挑战。传统半监督方法虽尝试生成潜在映射, 但易受噪声干扰且未能充分考虑对齐过程中的不确定性。针对上述问题, 提出一种基于学生-教师架构的联合训练实体对齐方法(JTEA)。该方法通过联合训练种子实体对与潜在映射, 结合双向匹配策略及联合匹配置信度评分, 不仅提高了实体映射的准确性, 还能够有效表征对齐过程中的不确定性。此外, JTEA方法创新性地引入匹配多样性对齐选择, 进一步提升了模型性能与稳定性。实验证明, JTEA方法在基准数据集上表现优异。

关键词: 知识图谱; 实体对齐; 表示学习; 联合教学; 图注意力网络

DOI: 10.11907/rjtk.241613

开放科学(资源服务)标识码(OSID):

中图分类号: TP391.1

文献标识码: A

文章编号: 1672-7800(2025)009-0099-07



Semi-Supervised Entity Alignment Based on Joint Teaching

LU Kai^{1,2,3}, ZHAO Jing^{1,2,3}

(1. Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Ji'nan), Qilu University of Technology (Shandong Academy of Sciences); 2. Shandong Engineering Research Center of Big Data Applied Technology, Faculty of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences); 3. Shandong Provincial Key Laboratory of Industrial Network and Information System Security, Shandong Fundamental Research Center for Computer Science, Ji'nan 250353, China)

Abstract: Entity alignment aims to identify and correlate equivalent entities across different data sources, but the lack of sufficient training data remains a major challenge. Traditional semi-supervised methods attempt to generate potential mappings, but they are susceptible to noise and ignore uncertainties. The newly proposed JTEA method adopts a student-teacher architecture, which combines joint training of seed entity pairs with potential mappings, along with a bidirectional matching strategy and joint matching confidence scores, to improve mapping accuracy and represent uncertainties. Additionally, it introduces a matching diversity alignment selection method to further enhance model performance and stability. Experiments have demonstrated the outstanding performance of the JTEA method on benchmark datasets.

Key Words: knowledge graph; entity alignment; representation learning; joint-teaching; graph attention network

0 引言

实体对齐是知识图谱构建中的一项关键技术,其目标

是在不同来源的知识图谱中识别出指向现实世界中相同实体的不同表示。这些实体可能具有不同的名称、标识符和属性描述等,但它们在语义上指向同一个对象。实体对齐的目标是识别和连接这些数据中的相同实体,从而构建

收稿日期: 2024-07-19

扫描二维码阅读全文:



基金项目: 山东省自然科学基金-联合基金(ZR2022LZH008); 齐鲁工业大学(山东省科学院)研究生教改项目(YJG23YB002); 山东省本科教学改革研究重点项目(Z2023314); 济南市市校融合发展战略工程项目(JNSX2023045)

作者简介: 卢凯(2000-), 男, 齐鲁工业大学(山东省科学院)计算机科学与技术学部硕士研究生, 研究方向为实体对齐; 赵晶(1979-), 女, 博士, 齐鲁工业大学(山东省科学院)计算机科学与技术学部副教授、硕士生导师, 研究方向为智能优化算法、机器学习、图像处理。本文通讯作者: 赵晶。

一个更加全面且实用的知识图谱。

近年来,基于嵌入的实体对齐取得了重大进展^[1-5]。这类方法利用知识图谱嵌入技术,将知识图谱的符号表示转换为低维向量,从而通过嵌入空间的几何结构来表征实体的语义相关性。然而,这些方法严重依赖于足够数量的手动标注的对齐实体对,标注数据不仅成本高昂且耗时,用于链接不同的嵌入空间以支持对齐学习。特别是对于某些领域或任务,获取标记数据可能非常困难,对基于监督嵌入的实体对齐方法提出了挑战。

半监督方法为解决上述问题提供了一种有效的途径。他们通过迭代提出新的“可靠”实体对齐来扩展训练数据。例如,Zhu等^[6]提出了IPTransE,它通过迭代过程预测未标记的数据,将超过复杂阈值的预测(可靠预测)作为重新训练的潜在映射。BootEA可以通过启发式比对编辑方法来缓解错误累积的问题,但对齐模型的学习能力和比对选择偏差的限制了其性能^[7]。CycTEA设计了一个有效的集成框架,使多个对齐器能够交换其可靠的实体映射,但它们之间需要高度互补性^[8]。尽管半监督方法取得了积极的成果,但仍面临以下问题:①潜在映射的不确定性。先前的工作在很大程度上忽略了半监督训练期间潜在映射的不确定性。回顾自训练过程,潜在映射的生成呈现出非黑即白的特性,即实体对被明确判定为潜在映射或非潜在映射。而事实上,不同的潜在映射具有不同的不确定性,对模型学习的贡献也不同。②噪声潜在映射学习。半监督实体对齐方法的性能在很大程度上依赖于潜在映射的质量,然而这些潜在映射不可避免地包含大量噪声。更为严重的是,将这些含有噪声的映射直接加入训练数据会误导后续的训练过程,进而导致错误累积,最终损害实体对齐的整体性能。

为解决上述问题,本文设计了一个简单而有效的半监督实体对齐框架JTEA。首先,利用一组有限的种子实体对作为基准来训练学生模型,学生模型逐步更新其参数,而教师模型则通过学生模型权重的指数移动平均进行更新^[9];其次,学生模型从教师模型生成的潜在映射中开展学习,提出一种双向匹配策略,集成来自不同方向的对齐决策,并将它们与匹配置信度得分相结合,以有效地评估对齐结果的不确定性;再次,提出了一种利用匹配多样性的对齐选择方法,为对齐模型选择可靠的实体。本文主要贡献:①提出了一种简单而有效的半监督实体对齐教学框架JTEA。它通过结合种子实体对和生成的潜在映射,共同指导模型进行对齐选择。联合迭代示教过程有助于克服单个模型学习能力的局限性,减少新训练数据中的噪声,从而提高性能。②引入双向匹配策略。通过在不同方向上联合调整决策生成更详细的潜在映射,并使用联合匹配置信度来表征其不确定性。此外,利用延迟接受算法为对齐模型选择可靠的实体。③在基准数据集上进行的的大量实验验证了该框架的有效性。

1 相关工作

1.1 基于嵌入的实体对齐

近年来,基于嵌入的实体对齐方法取得了重要进展。早期的研究,如MTransE利用TransE嵌入方法将不同的知识图谱嵌入到单独的向量空间中,并通过对齐模块将其进行转换^[10]。然而,在优化期间需要平衡编码器和对准模块。图神经网络,如GCN^[11]和GAT^[12],由于其鲁棒的结构学习能力,已被用作编码器。GCN-Align是第一个使用GCN来捕获邻域信息的方法,但它受到了不同知识图谱的结构异构性的影响^[13]。为缓解这个问题并改进结构学习,AliNet通过多跳聚合和门控机制扩展邻域,以实现更好的结构建模^[14]。KE-GCN将GCN与知识嵌入相结合,既能捕获实体结构,又能捕获关系语义^[15]。

1.2 半监督实体对齐

由于训练过程中可用的种子实体对通常数量有限,研究者提出了多种半监督实体对齐方法,这些方法通过同时利用有限的标注映射和大量未标注数据来实现实体对齐,从而为实际应用场景提供了更具实用性的解决方案。主流解决方案侧重于迭代生成潜在映射以弥补训练数据的缺乏。IPTransE应用基于阈值过滤的自训练来产生潜在映射,但由于引入了大量噪声数据,并对后续训练产生了误导,因此其性能未能达到预期^[6]。KDCoE对知识图谱嵌入模型和文字描述嵌入模型进行协同训练,以逐步提出新的潜在映射,从而增强对彼此对齐学习的监督^[16]。此外,CycTEA设计了一个有效的集成框架,使多个对齐模型能够交换其可靠的实体映射,以进行更鲁棒的半监督训练,但它需要不同对齐器之间的高度互补性^[8]。

实体对齐是一个双向的过程,部分模型如MRAEA已经针对这一点进行了设计^[17]。以实体对(ei, ej)为例,MRAEA会同时考虑ei与ej以及ej与ei之间的双向对齐关系。然而,尽管MRAEA采用了双向最近邻迭代策略,但它存在一个明显的不足:即使两个实体互为对方的最近邻,但如果它们的相似度较低,这种对齐关系也不应成立。因此,在进行半监督实体对齐时,需要一个更加精细的模型,它不仅要考虑实体的双向关系,还要设置一个合理的相似度阈值,以确保只有相似度足够高的实体才会被对齐。

1.3 一对一对齐选择

近年来,研究人员愈发重视实体对齐中的一对一问题。为此,CEA提出了一种创新的解决方案,将全局实体对齐问题转化为稳定匹配问题^[18]。这一转化确保了在匹配过程完成后,不存在源实体与目标实体均未被成功匹配的尴尬情况,从而提升了对齐结果的稳定性和可靠性。与此同时,GM-EHD-JEA^[19]与LatsEA^[20]等方法则另辟蹊径,将全局实体对齐视为一项任务分配问题,并巧妙地运用穷举枚举技术来优化计算过程,旨在降低复杂度,提升处理

效率。这些努力共同推动了实体对齐领域在一致性保证及效率提升方面的显著进步。

2 JTEA 方法

本文提出的半监督实体对齐方法,称为JTEA,如图1所示。JTEA的总体框架由2个编码器组成,遵循师生模型,通过利用种子实体对和生成的潜在映射来训练模型。同时,采用双向匹配策略和延迟接受算法生成潜在映射,并将生成的潜在映射合并到种子集中。

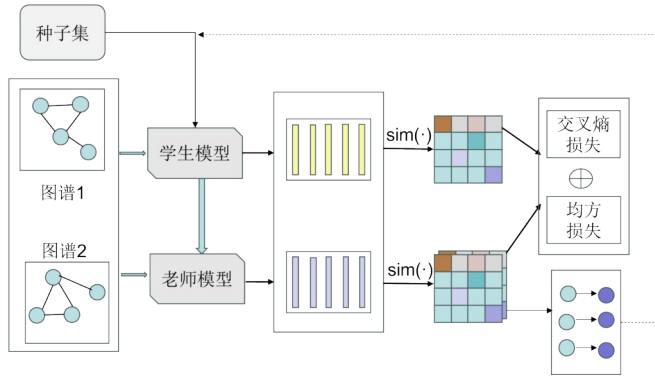


Fig. 1 The overall framework of JTEA

图1 JTEA 的总体框架

2.1 知识图谱嵌入

(1)引入编码器。利用邻域结构和关系语义将来自不同知识图谱的实体嵌入统一的向量空间。随机初始化可训练实体嵌入 $\mathbf{H}^{ent} \in \mathbb{R}^{(|\mathcal{E}_s| + |\mathcal{E}_t|) \times d_e}$ 和关系嵌入 $\mathbf{H}^{rel} \in \mathbb{R}^{|\mathcal{R}_s \cup \mathcal{R}_t| \times d_r}$, 其中, d_e 和 d_r 分别代表实体和关系的维度。

(2)结构建模。结构特征至关重要,因为等效实体往往具有相似的邻里环境。此外,利用多范围邻域结构能够提供更多的对齐证据,并减轻结构异质性问题。在这项工作中,我们应用图注意力网络来允许实体通过注意力机制选择性地聚合其周围信息,并通过堆叠L层递归地捕获多范围结构特征。如式(1)、式(2)所示。

$$\mathbf{h}_{e_i}^{(l)} = \sigma \left(\sum_{e_j \in \mathcal{N}_{e_i}} \alpha_{ij} \mathbf{W}_g \mathbf{h}_{e_j}^{(l-1)} \right) \quad (1)$$

$$\alpha_{ij} = \frac{\exp(\sigma(\mathbf{a}^\top [\mathbf{W}_g \mathbf{h}_{e_i} \oplus \mathbf{W}_g \mathbf{h}_{e_j}]))}{\sum_{e_j \in \mathcal{N}_{e_i}} \exp(\sigma(\mathbf{a}^\top [\mathbf{W}_g \mathbf{h}_{e_i} \oplus \mathbf{W}_g \mathbf{h}_{e_j}]))} \quad (2)$$

其中, $\top$ 表示转置, $\oplus$ 表示串联, $\mathbf{W}_g$ 和 $\mathbf{a}$ 分别是特定于层的变换参数和注意力变换向量。 $\mathcal{N}_{e_i}$ 表示 e_i 的邻居集合, α_{ij} 表示实体 e_j 对实体 e_i 的学习重要性。 $\mathbf{H}^{(l)}$ 表示第 l 层的实体嵌入矩阵, $\mathbf{H}^{(0)} = \mathbf{H}^{ent}$ 。 $\sigma(\cdot)$ 是非线性函数,使用ELU。

(3)关系建模。携带丰富语义的关系级信息对于对齐知识图谱中的实体至关重要,因为两个等效实体可能共享重叠关系。考虑到关系方向,即向外($e_i \rightarrow e_j$)和向内($e_i \leftarrow e_j$)对表征给定目标实体 e_i 有着微妙的影响,使用两个均值聚合器分别收集向外和向内关系语义,为异构知识图谱提

供补充特征。如式(3)、式(4)所示。

$$\mathbf{h}_{e_i}^{r+} = \frac{1}{|\mathcal{N}_{e_i}^{r+}|} \sum_{r \in \mathcal{N}_{e_i}^{r+}} \mathbf{h}_r^{rel} \quad (3)$$

$$\mathbf{h}_{e_i}^{r-} = \frac{1}{|\mathcal{N}_{e_i}^{r-}|} \sum_{r \in \mathcal{N}_{e_i}^{r-}} \mathbf{h}_r^{rel} \quad (4)$$

其中, $\mathcal{N}_{e_i}^{r+}$ 和 $\mathcal{N}_{e_i}^{r-}$ 分别是实体 e_i 的向外和向内关系的集合。

(4)加权串联。根据邻域结构和关系语义捕获实体的上下文信息,随后连接实体 e_i 的中间特征以获得最终的实体表示。如式(5)所示。

$$\mathbf{h}_{e_i} = \bigoplus_{k \in K} \left[\frac{\exp(\mathbf{w}_k)}{\sum \exp(\mathbf{w})} \cdot \mathbf{h}_{e_i}^k \right] \quad (5)$$

其中, $K = \{(1), \dots, (L), r+, r-\}$ 且 $\mathbf{w} \in \mathbb{R}^{|K|}$ 是可训练的注意力向量,用于自适应地控制每个特征的流。在相乘之前将 $\mathbf{w}$ 输入到 softmax, 以确保归一化权重之和为1。

2.2 老师-学生模型

受 Mean Teacher^[9] 的启发,本文设计了2个具有相同结构的编码器,将其命名为教师模型和学生模型。首先,通过利用一组有限的种子实体对作为基准来训练学生模型,学生模型逐步更新其参数,而教师模型则通过学生模型权重的指数移动平均进行更新;然后,学生模型从教师模型生成的潜在映射中开展学习,并试图模仿教师模型;最后,教师模型将其关于理解实体关系和空间布局的知识以某种形式传递给学生模型。具体而言,更新教师模型如下:

$$\theta_{tea} \leftarrow m\theta_{tea} + (1 - m)\theta_{stu}, m \in [0, 1] \quad (6)$$

其中, θ 表示模型权重, m 是预设的动量超参数,控制教师模型平滑更新和演化。

为训练学生模型,本文定义了一个由交叉熵损失和均方误差损失组成的组合损失函数。交叉熵损失度量代表学生模型对已知实体对的预测准确度,而均方误差损失度量代表学生模型与教师模型之间预测的差异。对于每个已知实体对 (u, v) , u 是源知识图谱中的实体, v 是目标知识图谱中的实体,交叉熵函数可以表示为:

$$\mathcal{L}_{align} = -\frac{1}{N} \sum_{(u, v) \in \text{aligned pairs}} (\log P(v|u) + \log P(u|v)) \quad (7)$$

其中, $P(v|u)$ 表示学生模型在给定 u 的情况下预测 v 的概率, $P(u|v)$ 表示学生模型在给定 v 的情况下预测 u 的概率, N 表示已知实体对的数量。代表师生模型输出之间差异的损失如下:

$$\mathcal{L}_{distill} = \frac{1}{N} \sum_{i=1}^N (f_{stu}(u) - f_{tea}(u))^2 \quad (8)$$

其中, $f_{stu}(u)$ 是学生模型对实体 u 的表示, $f_{tea}(u)$ 是教师模型对实体 u 的表示, N 是实体的数量。整体损失函数定义为2个分量损失函数的加权和,具体如下:

$$\mathcal{L}_{total} = \lambda_{align} \cdot \mathcal{L}_{align} + \lambda_{distill} \cdot \mathcal{L}_{distill} \quad (9)$$

其中, λ_{align} 表示平衡交叉熵对齐损失, $\lambda_{distill}$ 表示师生模型输出差损失函数的权重参数。

本文通过迭代训练,使用对齐样本和定义的损失函数更新学生模型的参数。在每个训练步骤中,学生模型致力于通过最小化其与教师模型之间的差异来进行自我训练,直至学生模型的性能达到令人满意的水平,以此提升对齐模型的准确度与泛化能力。

2.3 潜在映射的生成和学习

受双向迭代策略的启发,本文提出了双向匹配策略。如图2所示,旨在通过联合调整不同方向的决策来生成更详细的潜在映射,并使用联合匹配置信度得分来表征其不确定性。首先,在完成编码之后,对2个知识图谱中的实体实施双向匹配,以此筛选出一组潜在的映射候选,并对每类潜在映射,采用相似性度量方法计算其对应的匹配置信度;其次,通过组合对准决策和从两个方向获得的匹配置信度得分(即,源实体到目标实体和目标实体到源实体),进一步丰富潜在映射的内容;最后,在联合匹配置信度得分的帮助下,识别哪些映射是可靠的,哪些映射可能需要进一步验证或消除。

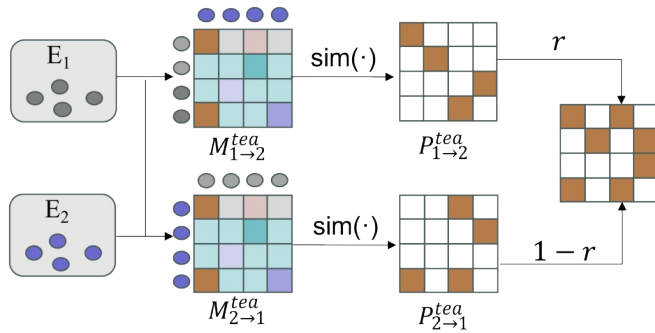


Fig. 2 Legend of bidirectional matching strategy

图2 双向匹配策略图例

具体而言,使用编码的实体表示,可以计算源实体和目标实体之间的成对相似度,以获得相似度矩阵。该相似度矩阵可以直观地显示源实体与目标实体之间的相似关系,为后续实体对齐任务提供重要依据。如式(10)所示:

$$\mathbf{M}_{s \rightarrow t}^{tea} = \text{sim}(\hat{\mathbf{e}}_s, \hat{\mathbf{e}}_t, \theta_{tea}) \in \mathbb{R}^{|\hat{\mathcal{E}}_s| \times |\hat{\mathcal{E}}_t|} \quad (10)$$

其中, $\text{sim}(\cdot)$ 表示余弦相似度函数。 $\mathbf{M}_{s \rightarrow t}^{tea}$ 和 $\mathbf{M}_{t \rightarrow s}^{tea}$ 表示源实体和目标实体之间不同方向的相似度矩阵, $\mathbf{M}_{t \rightarrow s}^{tea}$ 是 $\mathbf{M}_{s \rightarrow t}^{tea}$ 的转置。随后,针对每个矩阵,选取每行中预测相似度最高的实体对作为潜在映射,进而将不同方向上的潜在映射结果进行整合,并根据验证数据上的最后 Hits@1 分数进行加权,得到最终的潜在映射矩阵。如式(11)~式(13)所示:

$$\mathbf{P}^{tea} = \beta \cdot \mathbf{g}(\mathbf{M}_{s \rightarrow t}^{tea}) + (1 - \beta) \cdot \mathbf{g}(\mathbf{M}_{t \rightarrow s}^{tea})^T \quad (11)$$

$$\beta = \frac{\text{valid}(\mathbf{M}_{s \rightarrow t}^{tea})}{\text{valid}(\mathbf{M}_{s \rightarrow t}^{tea}) + \text{valid}(\mathbf{M}_{t \rightarrow s}^{tea})} \quad (12)$$

$$\mathbf{g}(\mathbf{M}) = [m_{i,j}] = \begin{cases} 1, & \text{if } j = \text{argmax}_j \mathbf{M}_{i,j} \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

其中, $\mathbf{g}(\cdot)$ 是将相似度矩阵转换为 one-hot 矩阵的函数,即矩阵每行唯一值为 1 的位置表示伪映射,从而生成教师模型生成的最终潜在映射矩阵 $\mathbf{P}_{tea}$ 。每个潜在映射都

与一个联合匹配置信度得分相关联,且联合匹配置信度越高,不确定性越小。与双向迭代策略不同,本研究采用不同方向对齐决策的投票一致性和匹配置信度来促进不确定性估计。具体而言,给定一个实体对 $(\hat{e}_i, \hat{e}_j)$,当且仅当两个方向一致投票该实体对为潜在映射时,其置信度 $\mathbf{P}_{i,j}^{tea}$ 才为 1;否则,若仅有一个方向投票支持该实体对,或者没有任何方向投票支持该实体对,则其置信度为 0(即该实体对不会被视作潜在映射)。

2.4 匹配多样性对齐选择

研究者通常期望源实体的嵌入和其相应的目标实体之间具有高度的相似性。与现有方法仅依赖最近邻搜索进行实体比对不同,本研究引入了匹配多样性的概念,用于衡量预测对齐与其他竞争对齐之间的差异。正确的对齐对应表现出高匹配多样性,表明高嵌入相似性,而竞争对齐则表现出较低相似性。

为保证所选比对的质量并满足一对一匹配的要求,本文将对齐选择问题归结为一个稳定匹配问题。研究将学生模型和教师模型计算的相似度矩阵以相等的权重进行组合,从而生成融合矩阵 $\mathbf{M}$ 。对于每个源实体 u ,在 $\mathbf{M}$ 中检索相应的行条目,并按降序对元素进行排序。具有最高排名的目标实体被选择为匹配伙伴。

在本研究中,我们采用延迟接受算法以确保每个实体均能实现成功匹配。该算法的原理如下:首先,每个源实体向具有最高相似性得分的目标实体发送配对请求,并形成临时匹配;其次,未成功匹配的源实体将继续向尚未请求的、优先级最高的目标实体发出配对请求;再次,如果目标实体尚未匹配或更喜欢新的请求者,它将拒绝当前的匹配并接受新的请求。重复此过程,直到所有源实体完成相应匹配。

图3展示了DAA规则的对齐过程。初始状态下,存在3个源实体(u_1, u_2, u_3)和3个目标实体(v_1, v_2, v_3)。根据初始偏好矩阵,每个实体依据其偏好(偏好值较大者优先)向其他实体发出配对请求。在第一轮中, u_1 和 u_2 同时向 v_1 发出匹配请求,而 u_3 则向 v_2 发出请求。由于 v_1 更倾向于 u_1 ,因此选择 u_1 并拒绝了 u_2 ; v_2 接受了 u_3 的请求。在第二轮中,仅 u_2 仍未匹配,它向其次选目标 v_2 发出请求。尽管 v_2 已与 u_3 暂时匹配,但由于 v_2 更偏好 u_2 ,因此选择与 u_2 配对。在第三轮中,先前配对失败的 u_3 向 v_3 发出请求,双方最终成功匹配。

3 实验

3.1 数据集和评价指标

本实验使用 OpenEA 的基准数据集进行评估,其与现实世界的知识图谱中的实体度分布相一致,包括2个跨语言设置(EN-FR-15K 和 EN-DE-15K)和2个单语言设置(DW-15K 和 D-Y-15K)^[21]。对于数据分割,遵循 OpenEA

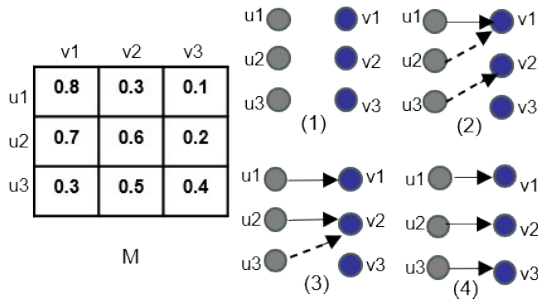


Fig. 3 Alignment process of DAA rules
图 3 DAA 规则的对齐过程

的标准集,分配 20% 用于训练,10% 用于验证,70% 用于测试。表 1 展示了数据集的细节。

本实验采用 Hits@k(k=1,5) 和平均倒数秩(MRR)作为评价指标,以评估实体对齐的性能。Hits@k 反映了对齐的精度,而 MRR 表示正确对齐实体的平均倒数秩。这些指标的值越高,表明实体对齐的性能越好。

Table 1 Detailed information of datasets
表 1 数据集的详细信息

数据集	语言	实体	关系	三元组
EN-FR-15K	English	15 000	267	47334
	French	15 000	210	40864
EN-DE-15K	English	15 000	215	47676
	German	15 000	131	50419
D-W-15K	DBPedia	15 000	248	38265
	Wikidata	15 000	169	42746
D-Y-15K	DBPedia	15 000	165	30292
	YAGO	15 000	28	26638

3.2 基线方法

本实验从相关研究中选择如下模型作为实验基线: MTransE^[10]、GCN^[11]、GAT^[12]、GCN-Align^[13]、IPTransE^[6]、SEA^[22]、KDCoE^[16]、AliNet^[14]、KE-GCN^[15]、RHGN^[23]、MRAEA^[17]、BootEA^[7]。

3.3 实施细节

所有实验均基于 PyTorch 框架,并在 NVIDIA GeForce RTX 3090 GPU 上执行。遵循 OpenEA 的实验设置,本文报告了五重交叉验证的平均结果。实体 de 和关系 dr 的嵌入维数分别设为 256 和 128,GAT 层数 L 为 2,动量 m 为 0.99。在实体对齐推理阶段,使用余弦距离作为距离度量,并应用 Faiss 2 来有效地执行 NN 搜索,默认对齐方向是从左到右。例如,在 D-W-15K 中,将 DBpedia 视为源知识图谱,并试图在目标知识图谱 Wikidata 中找到源实体的对应物。

3.4 实验结果

对于所有基线,通过运行基线代码或参考相应的论文来获得报告的分。所用基线和实验结果如表 2 所示。粗体表示最佳结果,下划线次之。

表 2 展示了在 OpenEA 15K 数据集上的实验结果,表明本文提出的 JTEA 模型优于其他方法。MTransE、SEA 和

Table 2 Entity alignment results on 15K datasets
表 2 在 15K 数据集上的实体对齐结果

模型	EN-FR-15K			EN-DE-15K			D-W-15K			D-Y-15K		
	Hits	Hits	MRR	Hits	Hits	MRR	Hits	Hits	MRR	Hits	Hits	MRR
	@1	@5		@1	@5		@1	@5		@1	@5	
MTransE	.247	.467	.351	.307	.518	.407	.259	.461	.354	.463	.675	.559
GCN	.210	.414	.304	.304	.497	.394	.208	.367	.284	.343	.503	.416
GAT	.297	.585	.426	.542	.737	.630	.383	.622	.489	.468	.707	.573
GCN-Align	.338	.589	.451	.481	.679	.571	.364	.580	.461	.465	.626	.536
AliNet	.364	.597	.467	.604	.759	.673	.440	.628	.522	.559	.690	.617
KE-GCN	.408	.670	.524	.658	.822	.730	.519	.727	.608	.560	.750	.644
IPTransE	.169	.320	.243	.350	.515	.430	.232	.380	.303	.313	.456	.378
SEA	.280	.530	.397	.530	.718	.617	.360	.572	.458	.500	.706	.591
KDCoE	.581	.680	.628	.529	.629	.580	.247	.412	.325	.661	.764	.710
BootEA	.507	.718	.603	.675	.820	.740	.572	.744	.649	.739	.849	.788
MRAEA	.537	.779	.642	.681	.844	.754	.574	.772	.661	.725	.838	.769
RHGN	.500	.739	.603	.704	.859	.771	.560	.753	.644	.708	.831	.762
JTEA	.580	.808	.681	.727	.877	.799	.645	.828	.733	.736	.869	.800

IPTransE 等知识表示学习编码器的性能不如 AliNet 和 KE-GCN 等 GNN 编码器,突出了 GNN 在实体对齐方面的强大能力。普通的 GCN 不能超越其他基于 GCN 的方法,强调了设计有效的编码器的重要性。RHGN 利用关系门控卷积层来区分关系和实体,但由于该方法依赖于整个图结构来学习实体和关系的嵌入,所需的计算资源和处理时间随着 KG 的规模呈线性增加。这在一定程度上限制了 RHGN 在大规模、多实体知识图中的性能。BootEA 和 MRAEA 都是半监督实体对齐领域的竞争性基线方法。然而,BootEA 的缺点是需要仔细地微调置信度阈值来过滤潜在的映射,这可能会导致性能不稳定。同时,MRAEA 在处理潜在的映射时忽略了不确定性因素,这可能会导致误差的积累。相比之下,JTEA 通过双向匹配策略和稳定的匹配对齐选择方法,为对齐模型提供了可靠的解决方案,并在各项评估指标中均取得了不同程度的性能提升,充分展现了其作为一种轻量级且功能强大的实体对齐方法的优势。

3.5 消融实验

为了深入分析 JTEA 模型中各个组件对整体性能的影响,本研究设计了一系列消融实验。

消融实验结果如表 3 所示,当模型保持完整时,其实验结果是最突出的,证明了模型设计的每个组件都在提高性能方面发挥了积极的作用。当从实体表示中删除关系建模时,模型的性能会受到影响,显示出下降的趋势。这进一步强调了关系语义识别在丰富实体表示能力方面的重要性。在没有潜在映射学习(PML)的情况下,模型性能显著下降,尤其是在单语环境中,突出了潜在映射学习的关键重要性。此外,从 w/o BMS 和 w/o MDS 的结果中可以看出,双向匹配策略和稳定匹配对齐选择都有效地提高了潜在映射学习的质量。特别值得注意的是,w/o B&M 的实验结果进一步验证了 BMS 和 MDS 的结合够显著提升模型性能,从而优化实体对齐的效果。

Table 3 Ablation study results

表3 消融实验结果

模型	EN-FR-15K			EN-DE-15K			D-W-15K			D-Y-15K		
	Hits	Hits	MRR	Hits	Hits	MRR	Hits	Hits	MRR	Hits	Hits	MRR
	@1	@5		@1	@5		@1	@5		@1	@5	
w/o rel.	.472	.732	.585	.678	.840	.747	.553	.757	.641	.700	.842	.759
w/o PML	.485	.717	.590	.671	.836	.744	.520	.708	.605	.541	.684	.605
w/o BMS	.556	.781	.656	.707	.863	.777	.629	.812	.709	.725	.856	.783
w/o MDS	.560	.791	.663	.711	.863	.779	.635	.822	.718	.731	.862	.788
w/o B&M	.542	.771	.644	.694	.857	.766	.609	.802	.693	.712	.849	.774
JTEA	.580	.808	.681	.727	.877	.799	.645	.828	.733	.736	.869	.800

3.6 辅助实验

本文设计了辅助实验来研究超参数的变化。其中,包括但不限于图注意力网络的层数L和动量参数m。

本文在实验中调整了GAT层(L)的数量,从1层逐步增加到4层,以观察其对结果的影响,结果如图4所示。当L=1时,由于模型的结构建模能力较为有限,导致了较差的对齐性能。值得注意的是,在大多数情况下(除了D-Y-15K任务),当L=2时,模型达到了最佳的性能表现。但随着L值的进一步增加,模型性能并未呈现显著提升。研究推测,这一现象可能是由于邻域信息聚合过程中出现的过拟合或过平滑效应所导致的。

然而,在D-Y-15K这一特定任务中,实验结果呈现出不同的趋势。当L增加到4层时,系统性能反而达到了最佳。研究分析认为,这可能是由于D-Y-15K中的2个知识图谱的结构信息相对较稀疏,因此模型需要通过捕获更远距离的邻居信息来收集足够的对齐证据,从而提高了对齐的准确性。该研究结果为后续优化知识图谱对齐模型提供了重要的理论依据和实践启示。

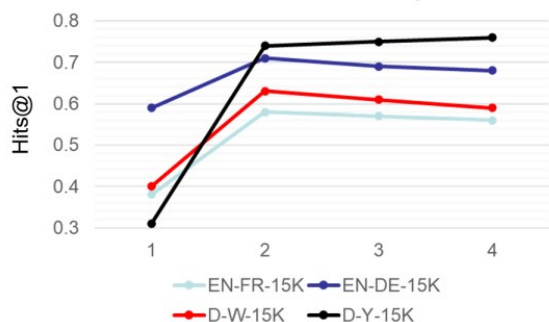


Fig. 4 The number of GAT layers

图4 GAT的层数

本文探究了动量参数m在[0, 0.9, 0.99]范围内的效果,针对EN-DE-15K和D-W-15K任务的结果如图5所示。实验结果表明,本方法在不同动量设置下均能保持稳定的性能,这证明了该方法对m的变动并不敏感。此外,当选择适当的m值(如0.99)时,本方法展现出了性能上的提升,这表明采用指数移动平均(EMA)更新方式有助于教师模型生成更为稳定和鲁棒的潜在映射。

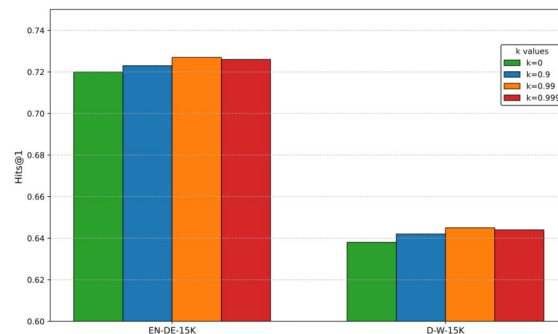


Fig. 5 Momentum parameter

图5 动量超参数

4 结语

本文提出了一种新的半监督实体对齐框架JTEA,旨在解决传统实体对齐模型中过度依赖种子数据的问题。通过集成师生模型,联合教学基于种子实体对的模型对齐决策,生成潜在映射来指导对齐过程。这种联合迭代教学的方法能够有效克服模型学习的局限性,降低新训练数据中噪声的干扰,从而显著提升模型性能。为更有效地学习潜在映射,本文设计了一个双向匹配策略。该策略集成了来自2个方向的对齐决策,并结合匹配置信度分数以评估对齐结果。在此基础上,进一步提出了一种基于匹配多样性的对齐选择方法,以确保对齐模型的实体对齐选择。本文通过大量实验验证了JTEA框架的有效性,实验结果充分证明了其在实体对齐任务中的优越性能。在未来的工作中,将进一步扩大JTEA模型的应用范围,将其应用于更多样化类型的数据集,以验证其泛化能力。

参考文献:

- [1] TAN Y Z, LI X N, LI G Y. Entity alignment method based on neighborhood aggregation[J]. Computer Engineering, 2022, 48(6): 65-72.
谭元珍, 李晓楠, 李冠宇. 基于邻域聚合的实体对齐方法[J]. 计算机工程, 2022, 48(6): 65-72.
- [2] SU M, BU G G, FAN Q F, et al. Entity alignment algorithm based on attribute embedding and graph attention network[J]. Computer Systems and Applications, 2023, 32(3): 202-208.
苏谟, 步格格, 范秋枫, 等. 基于属性嵌入与图注意力网络的实体对齐算法[J]. 计算机系统应用, 2023, 32(3): 202-208.
- [3] ZHAO D, ZHANG J. Entity alignment method based on dual attention and relational semantic modeling[J]. Application Research of Computers, 2022, 39(1): 64-69, 79.
赵丹, 张俊. 基于双重注意力和关系语义建模的实体对齐方法[J]. 计算机应用研究, 2022, 39(1): 64-69, 79.
- [4] CHE C, LIU D. Cross-language entity alignment based on bidirectional alignment and attribute information[J]. Computer Engineering, 2022, 48(3): 74-80.
车超, 刘迪. 基于双向对齐与属性信息的跨语言实体对齐[J]. 计算机工程, 2022, 48(3): 74-80.
- [5] XIE F, ZENG X, ZHOU B, et al. Improving knowledge graph entity

- alignment with graph augmentation [C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2023: 3–14.
- [6] ZHU H, XIE R B, LIU Z Y, et al. Iterative entity alignment via joint knowledge embeddings[C]//International Joint Conference on Artificial Intelligence, 2017: 4258–4264.
- [7] SUN Z Q, HU W, ZHANG Q H, et al. Bootstrapping entity alignment with knowledge graph embedding [C]//International Joint Conference on Artificial Intelligence, 2018: 4396–4402.
- [8] XIN K X, SUN Z Q, HUA W, et al. Ensemble semi-supervised entity alignment via cycle-teaching[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 4281–4289.
- [9] TARVAINEN T, VALPOLA H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results [C]//Advances in Neural Information Processing Systems (NeurIPS), 2017: 1195–1204.
- [10] 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 (IJCAI), 2017: 1511–1517.
- [11] KIPF T N, WELLMING M. Semi-supervised classification with graph convolutional networks [C]//Toulon: International Conference on Learning Representations, 2017.
- [12] CASANOVA P, LIO A R P, BENGIO Y. Graph attention networks [C]//Vancouver: International Conference on Learning Representations, 2018.
- [13] WANG Z C, LV Q S, LAN X H, et al. Cross-lingual knowledge graph alignment via graph convolutional networks[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018: 349–357.
- [14] SUN Z Q, WANG C M, HU W, et al. Knowledge graph alignment network with gated multi-hop neighborhood aggregation [C]//Proceedings of the Aaai Conference on Artificial Intelligence, 2020: 222–229.
- [15] YU D H, YANG Y M, ZHANG R H, et al. Knowledge embedding based graph convolutional network [C]//Proceedings of the Web Conference, 2021: 1619–1628.
- [16] CHEN M H, TAIN Y T, CHANG K W, et al. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment [C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 3998–4004.
- [17] MAO X, WANG W T, XU H M, et al. Mraea: an efficient and robust entity alignment approach for cross-lingual knowledge graph [C]//Proceedings of the 13th International Conference on Web Search and Data Mining, 2020: 420–428.
- [18] ZHENG W X, ZHAO X, TANG J Y, et al. Collective entity alignment via adaptive features [C]//2020 IEEE 36th International Conference on Data Engineering (ICDE), 2020: 1870–1873.
- [19] XU K, SONG L F, FENG Y S, et al. Coordinated reasoning for cross-lingual knowledge graph alignment [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 9354–9361.
- [20] CHEN W, CHEN X Y, XIONG S W. Global entity alignment with gated latent space neighborhood aggregation [C]//China National Conference on Chinese Computational Linguistics, 2021: 371–384.
- [21] SUN Z Q, ZHANG Q H, HU W, et al. A benchmarking study of embedding-based entity alignment for knowledge graphs [C]//Proceedings of the VLDB Endowment, 2020: 2193–2206.
- [22] PEI S C, YU L, ZHANG X L, et al. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference [C]//The World Wide Web Conference, 2019: 3130–3136.
- [23] LIU X K, ZHANG K, LIU Y, et al. Rhgn: relation-gated heterogeneous graph network for entity alignment in knowledge graphs [C]//Findings of the Association for Computational Linguistics, 2023: 8683–8696.

(责任编辑:陈维捷)