arXiv:2305.01561v1 [cs.CL] 2 May 2023

OTIEA：面向跨语言实体对齐的本体增强型三元组内在关联

Zhishuo Zhang、Chengxiang Tan、Xueyan Zhao、Min Yang 和 Chaoqun Jiang同济大学计算机科学与技术系，中国上海。

*通讯作者。电子邮箱：jerrytan@tongji.edu.cn；贡献作者：2110140@tongji.edu.cn；1710839@tongji.edu.cn；2110136@tongji.edu.cn；tommmm@foxmail.com；

摘要

缺乏充足外部资源的跨语言和跨领域知识对齐，是融合非规则数据的一项基础且关键的任务。实体对齐（EA）旨在通过逐元素融合，从不同知识图谱（KG）中发现等价对象，近年来已受到业界和学术界的广泛关注。现有的大多数实体对齐方法通常借助邻居节点、结构信息和外部资源，探索实体与关系之间的关联。然而，这些方法很少对三元组元素之间复杂的内在交互及角色信息进行建模，可能导致对三元组的表征不充分。此外，在一些场景中，尤其是跨语言和跨领域应用中，外部资源通常不可用，这反映出这些方法的可扩展性较差。为解决上述不足，本文提出一种新颖的通用实体对齐框架（OTIEA），该框架基于本体对，并通过三元组感知注意力实现角色增强机制，无需引入外部资源。具体而言，本文通过挖掘内在相关性和本体对信息，而非将元素视为彼此独立的对象，设计了一种本体增强型三元组编码器。此外，通过融合角色多样性，三元组感知实体解码器能够生成面向实体对齐的表示。最后，采用双向迭代对齐策略来扩展种子实体对。在三个真实世界数据集上的实验结果表明，与基线方法相比，我们的框架取得了具有竞争力的性能。

关键词 **数据集：**三元组感知注意力、内在关联、本体对增强、角色多样性

1 引言

在过去十年中，知识图谱（KGs）已成功应用于许多现实场景。一些商业公司构建了跨语言知识图谱，例如 DBpedia[1], 、YAGO [2]和 ConceptNet [3]，用于推荐系统 [4]、信息检索 [5] 以及其他基于知识的下游任务 [6]。建立不同知识图谱之间的连通性，以实现融合。

近年来，有用信息一直是研究热点。然而，特定领域和特定语言的知识图谱之间存在异质性和冗余，极大地限制了知识整合。实体对齐（EA）是一项基础且关键的任务，它通过发现代表同一现实世界对象的等价实体，在连接不同知识图谱的同时，过滤来自不同领域和语言的冗余信息。

基于嵌入的方法一直是实体对齐的主流 [7–10],它通过不同技术将知识图谱中的实体和关系映射到不同的潜在低维向量空间。这些方法根据其动机大致可分为两类：**Trans基方法和图神经网络**基方法。前者将三元组元素之间的内部交互建模为显式的平移过程，例如 TransE [11]。后者则尝试通过图卷积网络和图注意力网络聚合邻居节点的信息传递，以丰富语义表示。这两种方法都显著提升了实体对齐效果，但仍有一些重要问题尚未得到充分考虑，这可能是影响以往方法性能的原因之一。

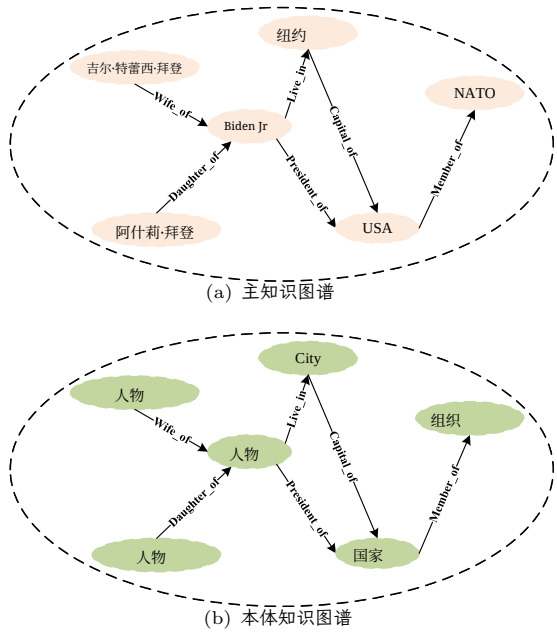


图 1：主知识图谱及其对应本体知识图谱示例

（1）一些研究者提出，仅依赖内部知识来表示知识图谱中的关系三元组并不足够，实体的外部资源可以有效补充并丰富实体表示。以往的工作 [12–21] 取得了相当不错的成果，尝试利用属性三元组、实体描述以及外部资源中的其他非文本信息。然而，尽管这些方法性能出色，但其通用性不足，尤其是在一些低资源场景中，由于知识具有领域特异性，外部资源可能无法获取或不可信。

（2）三元组不可分割，其元素彼此相互依赖。对于一个特定三元组（h，r，t）而言，头实体 h、关系 r 和尾实体t 在外部形式上具有天然的顺序，这不仅反映了浅层语义组合，也体现了集成三元组表示。然而，现有的大多数基于图神经网络的方法[17, 22–25] 分别对三元组元素进行嵌入，丢失了潜在语义，并忽略了知识图谱中实体的角色差异。此外，三元组元素之间的内在关联复杂且难以描述，无法通过试图对三元组元素之间简单映射过程进行建模的基于 Trans 的方法 [26–28] 得到充分表达。

(3) 成对本体是增强语义三元组表示的重要辅助资源。从直观角度来看，本体信息是最浅层、最简单的。实体的表达，其中包含了现有大多数方法所忽略的主要语义和类型信息。此外，作为一个整体，本体对能够提供更全面的语义表示，而单个实体的本体对三元组毫无意义。如图1所示，主知识图谱及其对应的本体知识图谱中，三元组 ***(***阿什莉·拜登***,*** 女儿***, Biden Jr)***对应的本体三元组***(***人物***,*** 女儿***,*** 人物***)*** 表示两个女儿 之间存在一种特定关系类型人物 实体应作为特定三元组的整体来处理，而不是作为两个毫无意义的独立元素，这一点显而易见。然而，以往相关研究很少包含成对特征。

鉴于上述洞见，本文提出 OTIEA——跨语言实体对齐的本体增强三元组内在关联，其中直观地认为，无论在语义空间还是本体空间中，三元组都是不可分割的，从而能够充分建模三元组内在关联和本体对增强。OTIEA 设计了一种三阶段注意力机制，用于挖掘交互过程中的潜在语义

在元素之间构建语义三元组表示。此外，还引入了一种本体增强方法，以融合本体对信息。进一步地，构建了一个三元组感知实体解码器，将集成三元组解码为角色感知实体，用于实体对齐。DBP15K 中三个跨语言知识图谱上的实验结果表明，与基线方法相比，OTIEA 取得了具有竞争力的性能。源代码可在 GitHub 上获取[1]。

本文的主要贡献如下：

•本文首次设计并将编码器-解码器架构应用于实体对齐任务，为编码集成三元组提供了全新视角。

• 在编码过程中，我们设计了一种三阶段交互注意力机制和一种本体对增强方法，以捕捉三元组中的复杂内在关联，同时利用本体信息补充语义三元组。在解码器中，我们引入了一种三元组感知实体解码策略，以考虑实体角色特征。

• 在三个公开数据集上进行的大量实验表明，与当前最先进的基线方法相比，OTIEA取得了具有竞争力的结果。

2 相关工作

近年来，基于嵌入的表示学习已成为工业应用和学术研究的热点。本节将详细介绍现有的基于嵌入的实体对齐方法，这些方法根据表示层级大致可分为两类：

基于 ***Trans*** 的方法：TransE [11] 是基于 Trans 的表示学习的首个范式，它将三元组语义建模为头实体 $\boldsymbol{h}$ 经由关系 $\boldsymbol{r}$ 转换为尾实体 $\boldsymbol{t}$ 的过程，即 $\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$。然而，由于表示能力不足，重叠关系无法得到有效解决。因此，一些研究者基于 TransE 及其他 Trans 模型（如 TransH[26]、TransR [27]和 TransEdge [28]），通过不同的实体和关系映射机制，探索可行方案来解决重叠问题。

TransR 将实体映射到与其相关的特定关系潜在空间，在一定程度上克服了 TransE 的局限性。TransEdge 基于 TransE，根据特定的头实体—尾实体对对关系表示进行上下文化。MTransE [29] 分别获取不同知识图谱中的实体和关系表示，并在两个知识图谱之间构建映射变换，以实现实体对齐。BootEA [30]采用自举策略扩展对齐种子。RpAlign[31]将实体对齐任务转换为知识图谱补全任务，并为已对齐实体提出一种新的“锚点”关系，以取代实体向量距离作为衡量方式。JAPE [14] 结合关系三元组、结构信息和属性，对实体和关系的表示进行优化。IPTransE [32] 将两个知识图谱的嵌入映射到统一的潜在向量空间，并采用迭代对齐策略增强实体表示。这些基于 Trans 的模型尝试以不同方式建立不同知识图谱中实体与关系之间的语义关联，但忽略了固有的结构信息；而且，三元组组成部分之间的内在关联未必能通过从头实体到尾实体的转换过程简单刻画。在我们的工作中，通过三阶段内部交互注意力机制对其进行建模，并验证了该机制的有效性。

基于图神经网络的方法： 图神经网络能够捕获结构信息并传播邻居消息，这些对于知识图谱中的知识表示至关重要。因此，近年来许多研究致力于通过图神经网络探索知识图谱内部的丰富信息。GCN-Align [7]是首项利用结构信息，通过图卷积网络增强实体表示的工作。MRAEA [33] 通过关注节点的入邻居和出邻居及其关联关系来建模实体嵌入，并提出了一种用于增强表示的双向迭代策略。RREA [8]利用关系反射变换，为每个实体获得特定于关系的嵌入。MCEA [34] 仅利用结构信息进行实体对齐，并扩展长尾实体的卷积区域，以提升图网络的能力。NAEA [35] 融合了实体的邻域子图级信息，并

[1]https://github.com/CodesForNlp/OTIEA

采用邻域感知注意力表示。RDGCN [24]通过关系知识图谱与双关系对应项之间的注意力交互来利用关系。NMN [36]设计邻居匹配机制，通过图采样方法构建面向匹配的实体表示。KAGNN [37]将图神经网络扩展为多视图版本，仔细融合知识事实和邻接结构，并提出知识感知注意力机制，以保留原始语义并确定知识事实的重要性。SHEA[38] 同时考虑关系语义和边方向，通过软注意力和硬注意力机制，在不同情况下聚合来自实体邻居的信息。RAGA[25] 采用自注意力机制将实体信息传播至关系，然后将关系信息聚合回实体，以处理知识图谱中的多重关系。现有基于图神经网络的表示方法有效提升了实体对齐的性能。然而，三元组内部的具体特征仍无法得到可靠建模，对实体对的本体和角色信息的利用也不充分。为解决这些不足，OTIEA 通过本体增强三元组编码器和三元组感知实体解码器，将本体对与角色多样性融合在一起。

3 问题定义

知识图谱可形式化为 $KG = (E, R, T)$，其中 $E = \{e_1, e_2, ..., e_n\}$ 和 $R = \{r_1, r_2, ..., r_k\}$分别是实体集和关系集，$n$和 k 分别是知识图谱中实体和关系的数量，$T \subset E \times R \times E$ 是关系三元组集。给定两个跨语言知识图谱 $KG1 = (E1, R1, T1)$ 和 $KG2 = (E2, R2, T2)$，实体对齐任务定义为：基于一组种子实体对，找出 $KG1$ 和 $KG2$ 中指向同一现实世界对象的实体对，该种子实体对集记为 $S = \{(e_1, e_1') | e_1 \in E1, e_1' \in E2\}$，其中 e_1 和 e_1' 等价。

4 OTIEA框架

我们提出了一种新颖的框架 OTIEA，该框架基于对三元组元素之间内在关联的建模，并利用本体对和角色特征的重要性来实现跨语言实体对齐。能够自然地解决重叠关系问题。OTIEA的整体架构如图2所示，可分为三个部分：本体增强三元组编码器、三元组感知实体解码器和对齐策略。首先，分别将两个知识图谱输入本体增强三元组编码器，其中，基于图卷积网络的拓扑聚合层聚合结构信息，随后通过三元组内在关联建模层构建集成语义三元组，再设计本体增强三元组生成层，融合本体信息以编码融合三元组表示。接着，利用三元组感知实体解码器解码融合三元组表示，得到兼顾元素多样性的实体表示。最后，采用双向迭代对齐策略，自动扩充种子实体对，以进行半监督训练。

4.1 本体增强三元组编码器

基于GCN的拓扑聚合层

为了充分考虑通过知识图谱的拓扑结构传递的丰富语义和关联信息，我们首先将原始关系、反向关系和自关系相结合，随后将扩展后的知识图谱分别输入双层GCN高速公路网络，如先前的工作所示 [25]，以嵌入实体节点。基于图卷积网络的拓扑聚合生成实体嵌入如下：

$$\boldsymbol{X}^{l+1} = \mathrm{ReLU}\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}^{l}\right) \quad (1)$$

其中，$\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$、$\boldsymbol{A}$ 是关系扩展图的邻接矩阵，I 是单位矩阵，$\tilde{\boldsymbol{D}}$ 是 $\tilde{\boldsymbol{A}}$ 的度矩阵，$\boldsymbol{X}^{l} \in \mathbb{R}^{n \times d_e}$ 表示 l隐藏层中的输入实体嵌入，n是知识图谱中的实体数量，d_e 是实体嵌入的维度，$\boldsymbol{X}^{l+1}$ 是l层的输出。

随后，采用逐层高速公路网络 [39] ，以保留原始实体信息并避免梯度爆炸。高速公路网络的工作方式如下：

$$T\left(\boldsymbol{X}^{l}\right) = \sigma\left(\boldsymbol{X}^{l}\boldsymbol{W}^{l} + \boldsymbol{b}^{l}\right) \quad (2)$$

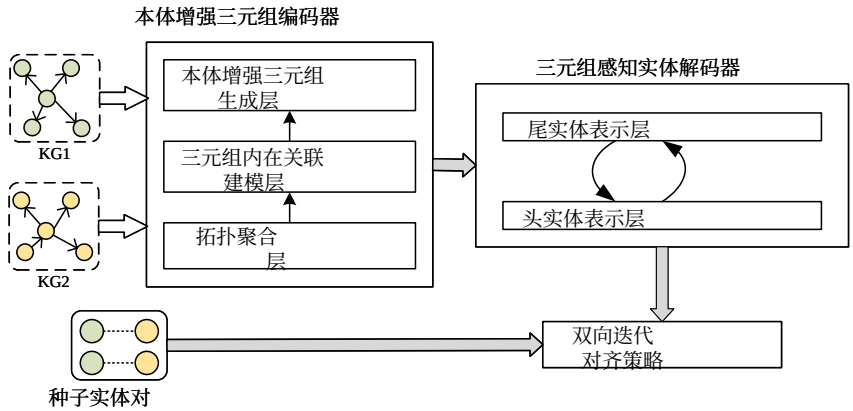


图 2：OTIEA框架的整体架构，采用编码器-解码器结构。两个知识图谱中的初始化实体分别输入本体增强三元组编码器，通过三阶段交互注意力机制和本体对增强生成集成三元组表示。随后，三元组表示在三元组感知实体解码器中融合角色特征，解码为面向实体对齐的实体表示。最后，采用双向迭代对齐策略来扩充训练种子对。

$$\boldsymbol{X}^{l+1}=T\left(\boldsymbol{X}^{l}\right)\cdot\boldsymbol{X}^{l+1}+\left(1-T\left(\boldsymbol{X}^{l}\right)\right)\cdot\boldsymbol{X}^{l} \tag{3}$$

其中，$T(\cdot)$ 为传输因子。$\sigma(\cdot)$ 为激活函数，$\cdot$表示逐元素乘法，$\boldsymbol{W}^{l}$ 和 $\boldsymbol{b}^{l}$ 分别为第 l个隐藏层中输入嵌入的权重矩阵和偏置向量。

三元组内在关联建模层

在这一层中，我们将三元组中任意两个元素的有序拼接视为另一个元素的潜在表示，并通过注意力机制对元素之间的内在关联进行建模，从而构建语义三元组的整体表示。在此过程中，对于特定三元组（e_i，r，e_j），其对应的头实体嵌入 $\boldsymbol{X}_i\in\mathbb{R}^{d_e}$和尾实体嵌入 $\boldsymbol{X}_j\in\mathbb{R}^{d_e}$可用于首先获得关系的潜在表示 $\boldsymbol{X}^{r}\in\mathbb{R}^{d_r}$。

$$\boldsymbol{X}^{r}=\sigma\left(\left(\boldsymbol{X}_i\|\boldsymbol{X}_j\right)\boldsymbol{W}_{sr}+\boldsymbol{b}_{sr}\right) \tag{4}$$

其中，d_e 和 d_r 分别表示实体和关系的维度，σ 表示激活函数，$\boldsymbol{W}_{sr}\in\mathbb{R}^{2d_e\times d_r}$和 $\boldsymbol{b}_{sr}\in\mathbb{R}^{d_r}$为可训练参数，$\|$ 表示拼接操作。

随后，设计了一种三阶段内部交互注意力机制，用于反映三元组中另外两个元素对某一元素的影响。由于计算过程相同，下面以头实体的内部交互注意力为例：首先，将关系 $\boldsymbol{X}^{r}$ 与尾实体 $\boldsymbol{X}_j$ 按顺序拼接，并将其视为头实体 $\boldsymbol{X}_i$的潜在表示，然后计算内部交互注意力：

$$c_{irj}=a^{T}\left(\boldsymbol{X}_i\boldsymbol{W}_h\|\left(\boldsymbol{X}^{r}\|\boldsymbol{X}_j\right)\boldsymbol{W}_{hT}\right)\right) \tag{5}$$

$$\begin{aligned}\alpha_{it}&=\operatorname{Softmax}\left(c_{irj}\right)\\&=\frac{\exp\left(\operatorname{LeakyRelu}\left(c_{irj}\right)\right)}{\sum\limits_{\left(e_{i'},r,e_{j'}\right)\in T_r}\exp\left(\operatorname{LeakyRelu}\left(c_{i'rj'}\right)\right)}\end{aligned} \tag{6}$$

$$\boldsymbol{X}_{hi}=\operatorname{Relu}\left(\sum_{e_i\in h_r}\left(\alpha_{it}\boldsymbol{X}_i\boldsymbol{W}_h\right)\right) \tag{7}$$

其中，$\boldsymbol{W}_h\in\mathbb{R}^{d_e\times d_r}$ 和 $\boldsymbol{W}_{hT}\in\mathbb{R}^{2d_e\times d_r}$ 分别是头实体和三元组内在关联的可训练参数。T_r 表示具有特定关系 r的三元组集合 T 的子集。h_r 是来自 T_r的头实体集合。$\boldsymbol{X}_{hi}$ 是头实体 e_i 的潜在语义表示。

同理，尾实体 e_j 和关系 r 的潜在语义表示可分别表示为 $\boldsymbol{X}_{tj}$ 和 $\boldsymbol{X}^{tr}$。因此，三元组 $(e_i,\ r,\ e_j)$ 的潜在语义表示为 $\boldsymbol{S}^{c}_{irj}=\boldsymbol{X}_{hi}+\boldsymbol{X}^{tr}+\boldsymbol{X}_{tj}$。

最后，我们考虑重叠的关系三元组，通过对同一关系对应的头实体与尾实体的拼接结果取平均值，生成三元组关系的全局特征。这样即可获得保留主要特征的全局三元组表示。

$$\boldsymbol{X}_g^r = \left(\frac{1}{|T_r|} \sum_{(e_i,r,e_j)\in T_r} (\boldsymbol{X}_i \| \boldsymbol{X}_j) \right) \boldsymbol{W}_{rg} + \boldsymbol{b}_{rg} \quad (8)$$

$$\boldsymbol{S}_{irj}^g = \sigma\left(\left(\boldsymbol{X}_i \left\| \boldsymbol{X}_g^r \right\| \boldsymbol{X}_j \right) \boldsymbol{W}_{sp} + \boldsymbol{b}_{sp} \right) \quad (9)$$

其中，$\boldsymbol{W}_{rg} \in \mathbb{R}^{2d_e \times d_r}$、$\boldsymbol{W}_{sp} \in \mathbb{R}^{(2d_e+d_r)\times d_r}$、$\boldsymbol{b}_{rg} \in \mathbb{R}^{d_r}$和 $\boldsymbol{b}_{sp} \in \mathbb{R}^{d_r}$为可训练参数。

之后，集成语义三元组表示为 $\boldsymbol{S}_{irj} = \boldsymbol{S}_{irj}^g + \boldsymbol{S}_{irj}^c$。

本体增强三元组生成层

在这一层中，将本体信息融合到三元组语义中，以形成增强的三元组表示。接下来，我们将详细介绍增强过程。

首先，我们将实体表示从语义空间转换到潜在本体空间，并将语义和本体视为不同的模态。

$$\boldsymbol{X}^o = \tanh\left(\boldsymbol{W}_{s2o} \boldsymbol{X} + \boldsymbol{b}_{s2o} \right) \quad (10)$$

其中，$\boldsymbol{X}^o$ 表示实体的本体表示，$\boldsymbol{W}_{s2o} \in \mathbb{R}^{d_e \times d_o}$和 $\boldsymbol{b}_{s2o} \in \mathbb{R}^{d_o}$，$d_o$ 是本体表示的维度。

然后，由于针对特定三元组处理单个实体本体毫无意义，因此在本体空间中将实体对作为一个整体处理。此外，具有相同本体表示的实体对可能关联多个关系，因此，全局本体关系表示的获取方式如下：

$$\boldsymbol{X}_g^{or} = \frac{1}{|T_r|} \sum_{(e_i,r,e_j)\in T_r} \left(\boldsymbol{X}_i^o \| \boldsymbol{X}_j^o \right) \quad (11)$$

$$\boldsymbol{X}^{or} = (\boldsymbol{X}_g^{or}) \boldsymbol{W}_{rg}^o + \boldsymbol{b}_{rg}^o \quad (12)$$

其中 $\boldsymbol{W}_{rg}^o \in \mathbb{R}^{d_o \times d_r}$ 和 $\boldsymbol{b}_{rg}^o \in \mathbb{R}^{d_r}$ 是可训练参数。然后，本体三元组可以构造如下：

$$\boldsymbol{O}_{irj} = \sigma\left(\left(\boldsymbol{X}_{ei}^o \| \boldsymbol{X}^{or} \| \boldsymbol{X}_{ej}^o \right) \boldsymbol{W}_t^o + \boldsymbol{b}_t^o \right) \quad (13)$$

其中，$\boldsymbol{W}_t^o \in \mathbb{R}^{(2d_o+d_r)\times d_r}$ 和 $\boldsymbol{b}_t^o \in \mathbb{R}^{d_r}$ 均为可训练参数。

我们认为，语义表示和本体表示这两种模态相互促进，因此设计了一种模态间的协同注意力机制，以获得融合语义和本体的集成三元组表示。协同注意力机制描述如下：

$$\alpha_{tr} = \frac{\exp\left(\mathrm{LReLU}\left(a^T (\boldsymbol{S}_{irj} \| \boldsymbol{O}_{irj}) \right) \right)}{\sum_{T_r} \exp\left(\mathrm{LReLU}\left(a^T (\boldsymbol{S}_{i'rj'} \| \boldsymbol{O}_{i'rj'}) \right) \right)} \quad (14)$$

$$\bar{\boldsymbol{O}}_r = \mathrm{ReLU}\left(\sum_{(e_i,r,e_j)\in T_r} (\alpha_{tr} \boldsymbol{S}_{irj}) \right) \quad (15)$$

通过上述计算过程即可获得增强本体三元组表示 $\boldsymbol{O}_r$。同样，也可以生成增强语义三元组表示 $\boldsymbol{S}_r$。

最后，在保留本体主要特征的同时，我们构建了本体增强型集成三元组表示 $\boldsymbol{T}_{irj} \in \mathbb{R}^{2d_o+d_r}$。

$$\boldsymbol{T}'_{irj} = \boldsymbol{S}_{irj} + \bar{\boldsymbol{S}}_r + \bar{\boldsymbol{O}}_r + \boldsymbol{O}_{irj} \quad (16)$$

$$\boldsymbol{T}_{irj} = \boldsymbol{T}'_{irj} \| \boldsymbol{X}_g^{or} \quad (17)$$

4.2 三元组感知实体解码器

循环协同增强机制

三元组感知实体解码器旨在区分角色多样性，并采用实体对循环协同增强机制来提升实体表示能力；其有效性已在第6节中得到验证。

实体对循环协同增强机制由三个分别对应不同实体角色的三元组感知注意力层组成：两个头实体感知表示层和一个尾实体感知表示层。头实体感知层、尾实体感知层以及另一个头实体感知层的组织形式，确保充分利用实体角色信息。考虑到

三个三元组感知注意力层的计算相似性，我们仅在此 formulate 头实体感知表示过程如下：

$$\alpha_{rj}=\frac{\exp\left(\mathrm{LReLU}\left(a^T\left(\boldsymbol{T}_{irj}\boldsymbol{W}^h\|\boldsymbol{X}_i\right)\right)\right)}{\sum_{\left(e_i,r',e'_j\right)\in T}\exp\left(\mathrm{LReLU}\left(a^T\left(\boldsymbol{T}_{ir'j'}\boldsymbol{W}^h\|\boldsymbol{X}_i\right)\right)\right)} \tag{18}$$

$$\boldsymbol{X}_i=\boldsymbol{X}_i+\mathrm{ReLU}\left(\sum_{(e_i,r,e_j)\in T}\left(\alpha_{rj}\boldsymbol{T}_{irj}\boldsymbol{W}^h\right)\right) \tag{19}$$

其中，$\boldsymbol{W}^h\in\mathbb{R}^{(d_r+2d_o)\times d_e}$ 是可训练的权重参数。随后，将头实体表示输入尾实体表示的计算中。由此可获得本体增强的实体表示，并将其输入初始图注意力网络层，重新聚合邻居信息，以生成最终实体表示 $\boldsymbol{X}^f=\{\boldsymbol{x}_1^f,\boldsymbol{x}_2^f,...,\boldsymbol{x}_n^f\}$。

4.3 迭代对齐策略

获得最终实体表示后，我们采用迭代对齐策略 [33] 执行半监督训练过程。基于间隔的损失函数使用 L1 范数距离定义如下：

$$dis\left(e_i,e_j\right)=\left\|\boldsymbol{x}_i^f-\boldsymbol{x}_j^f\right\|_1 \tag{20}$$

$$L=\sum_{(e_i,e_j)\in S}\max\left(dis(e_i,e_j)-dis\left(e'_i,e'_j\right)+\lambda,0\right) \tag{21}$$

其中，(e_i,e_j) 是训练集 S 中的预对齐实体对，(e'_i,e'_j) 是通过随机将 e_i 或 e_j 替换为其 k 个最近邻而生成的负样本，λ 是间隔超参数。在双向半监督训练过程中，分别计算从源知识图谱到目标知识图谱以及从目标知识图谱到源知识图谱的最小实体距离。只有在两个方向上均具有最小距离的实体对才会被添加到训练集中，从而有效缓解半监督训练过程中引入的错误对齐实体对。

OTIEA框架的整体流程如算法所示 1。

算法 1 OTIEA的整体流程框架

要求： 源 KG_s，目标 KG_t

初始化 E_s 和 E_t 通过 Glove
对于 $epoch\ i=1\rightarrow E$，**执行**
  聚合 E_s 的拓扑结构并 E_t 根据公式（1）–（3）
  生成语义三元组 S_s 和 S_t
  通过 T_s 和 T_t 获得集成三元组公式（17）
  生成面向实体对齐的实体表示 E_s 并 E_t 根据公式（18）–（19）
  使用公式（20）–（21）进行训练
  如果 $n\%5==0$ **则**
    扩充训练集 S
  结束如果
结束

5 实验设置

5.1 数据集

为了与基线方法进行公平比较，我们在简化版DBP15K上评估OTIEA框架。该数据集是一个公开数据集，几乎所有用于跨语言实体对齐任务的相关方法都广泛使用该数据集，其描述见表1。需要注意的是，简化版与原始版本的区别在于移除了孤立实体，因为这些实体对于大多数依赖知识图谱结构进行消息传播的实体对齐方法几乎没有帮助。

如表1 所示，简化版DBP15K包含三个中文、英语和法语的跨语言子数据集：ZH-EN、日英和FR-EN。三个子数据集中的链接三元组数量均为15000，实体数量均为19k，关系数量处于同一水平；FR-EN中的三元组数量略多于ZH-EN和日英。

5.2 基线

为了从多个方面评估 OTIEA，我们选择不引入外部资源的方法作为基线方法，这些方法可分为两类。此外，我们还比较了由我们实现或由原论文提供半监督训练技术的一些相关方法，以从不同层面证明模型的优越性。

表 1： Statistical data of the SimplifiedDBP15K.

DBP15K		实体	关系	关系三元组	链接
中文-英文	ZH	19388	1700	70414	15000
	EN	19572	1322	95142	
日文-英文	JA	19814	1298	77214	15000
	EN	19780	1152	93484	
法文-英文	FR	19661	902	105998	15000
	EN	19993	1207	115722	

基于 Trans 的方法（基础）： MTransE[29], BootEA [30], TransEdge [28], RpAlign [31]。

基于图神经网络的方法（基础）： GCN-Align[7], MCEA [34]。NAEA [35], RDGCN [24], NMN[36], KAGNN [37], SHEA [38], RAGA [25]。

半监督方法： MRAEA-semi[33], RREA-semi [8],RAGA-semi [25], MCEA- semi [34]。

据我们所知，RAGA 在不使用外部资源的情况下，通过探索实体与关系之间的交互，取得了当前最先进水平的结果。

5.3 模型变体

我们构建了 OTIEA-base 的三个变体，以验证不同组件或模块的有效性。

(1) wo-E：简化版 OTIEA，在三元组内在关联建模层中移除了三元组的全局特征。

(2) wo-O：简化版 OTIEA，移除了本体增强三元组生成层。

（3）wo-C：三元组感知实体解码器中不包含循环协同增强机制的简化版 OTIEA。

5.4 实现细节

我们使用来自 RAGA 的初始化实体嵌入。RAGA 首先将法语和日语的实体名称翻译成英语，然后利用 Glove 模型获取实体嵌入。对于三个子数据集，我们采用相同的超参数设置：本体维度 d_o 和关系维度 d_r 均为 100，图卷积网络的深度 l 为 2，损失函数中的间隔超参数 λ 设为 3.0。我们还采样最近的k =5 个邻居来构造负样本。按照惯例，30% 的种子实体对用作训练集，其余用作测试集。我们采用 Hits@k 和平均倒数排名来衡量所有方法的实体对齐性能，所有评估指标的分数越高，表示性能越好。

6 结果与分析

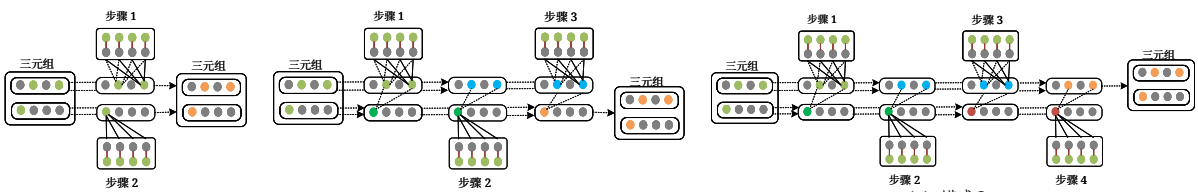
表2 展示了不同实体对齐方法在简化版 DBP15K的三个子数据集上的总体结果，这些结果要么通过源代码实现，要么由原论文提供。虚线将基于 Trans 的方法与基于图神经网络的方法分隔开来，实线则分别分隔了基线方法、不同的 OTIEA 变体和半监督方法。

总体实验结果的比较表明，我们的框架在三个子数据集和各项评估指标上均取得了具有竞争力的性能。与几乎所有不使用外部资源的基线方法相比，OTIEA 在基础版本和半监督版本中都表现更好。可能的原因是，OTIEA 通过三元组级注意力对三元组集成信息进行建模，这进一步表明，在实际应用中，应将三元组作为一个整体进行处理，而不是将其视为松散的组成部分。具体而言，基于 Trans 的方法由于表征能力不足，通常表现较差。相比之下，近年来基于图神经网络的方法取得了很大进展，其中 RAGA [25] 是与我们工作最相关的解决方案之一，它在实体和关系之间引入了注意力机制，但忽略了三元组集成特征以及实体类型的影响。具体来说，OTIEA-base 在 Hits@1、Hits@10 和平均倒数排名上均超过了当前最先进水平的 RAGA，而在日英数据集上，OTIEA-semi 的 Hits@1 略低于 RAGA-semi，这可能是由于数据集中的数据分布存在差异。对于不同变体，与 OTIEA-base(wo-E)、OTIEA-base(wo-O) 和 OTIEA-base(wo-C) 相比，OTIEA-base 在几乎所有评估指标上都取得了更好的性能，证明全局三元组表征、本体信息和循环协同增强机制对整体框架具有一定的积极影响。

为了进一步验证我们框架的鲁棒性，我们在三个子数据集上使用不同比例的训练种子实体对，将 OTIEA-base 和 OTIEA-semi 与 RAGA-base 和 RAGA-semi 进行了比较。图 5、图 6和图 7展示了

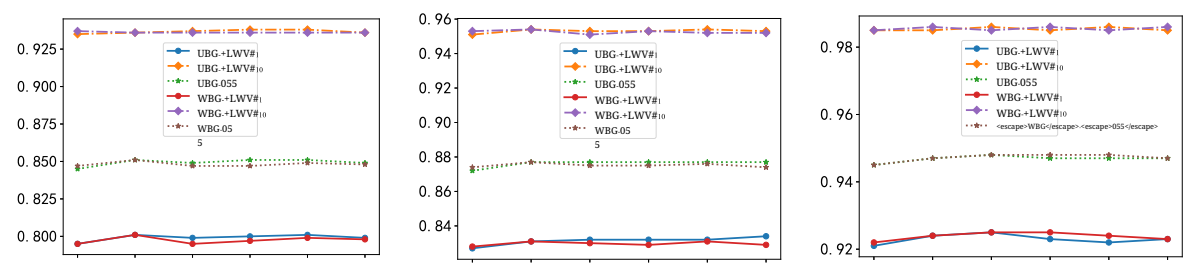
表2： Overall performance of entity alignment.

方法	中文-英文			日文-英文			法文-英文		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
MTransE(2017)	30.8	61.4	0.364	27.8	57.4	0.349	24.4	55.5	0.335
BootEA(2018)	62.9	84.7	0.703	62.2	85.4	0.701	65.3	87.4	0.731
TransEdge(2019)	73.5	91.9	0.801	71.9	93.2	0.795	71.0	94.1	0.796
RpAlign(2022)	74.8	88.8	0.794	72.9	89.0	0.872	75.2	89.9	0.801
GCN-Align(2018)	41.2	74.4	0.549	39.9	74.4	0.546	37.3	74.5	0.532
MCEA(2022)	72.4	93.4	0.800	71.9	94.0	0.800	73.9	95.3	0.820
NAEA(2019)	65.0	86.7	0.720	64.1	87.3	0.718	67.3	89.4	0.752
RDGCN（2019）	70.8	84.6	0.751	76.7	89.5	0.812	88.6	95.7	0.908
NMN(2020)	73.3	86.9	0.781	78.5	91.2	0.827	90.2	96.7	0.924
KAGNN(2022)	73.6	87.3	0.786	79.4	91.1	0.837	92.0	97.6	0.941
SHEA（2021年）	76.3	91.4	0.835	82.1	93.8	0.860	90.5	97.0	0.902
RAGA(2021年)	79.8	93.3	0.847	82.9	95.0	0.875	91.4	98.2	0.940
OTIEA-base(wo-E)	79.6	93.4	0.846	82.9	95.1	0.874	**92.4**	98.5	**0.947**
OTIEA-base(wo-O)	78.8	**93.7**	0.842	81.3	94.9	0.864	91.3	98.3	0.940
OTIEA-base(wo-C)	79.4	93.4	0.845	81.7	95.0	0.866	91.9	98.5	0.944
OTIEA-base(ours)	**80.1**	93.6	**0.851**	**83.1**	**95.4**	**0.877**	**92.4**	**98.6**	**0.947**
MRAEA-semi(2020)	75.2	92.3	0.824	75.3	93.3	0.825	78.1	94.7	0.843
RREA-semi(2020)	80.1	94.8	0.857	80.2	95.2	0.858	82.7	96.6	0.881
MCEA-semi(2022)	81.4	95.6	0.867	80.7	95.7	0.864	84.1	97.0	0.891
RAGA-semi(2021)	85.7	96.0	0.896	**88.9**	97.1	**0.920**	94.0	98.8	0.958
OTIEA-semi(ours)	**86.3**	**96.2**	**0.900**	88.6	**97.3**	**0.920**	**94.8**	**99.0**	**0.965**



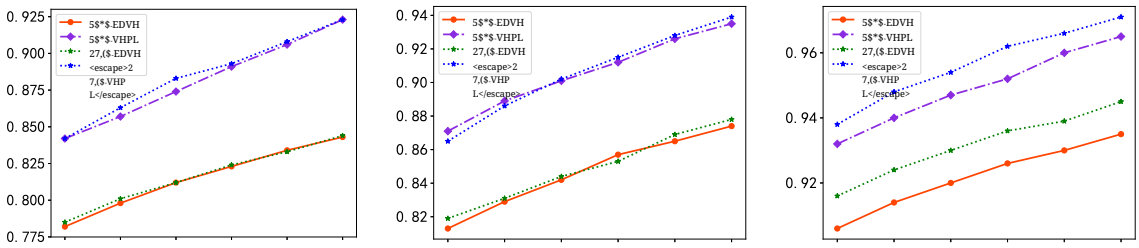
(a) 模式*1* (b) 模式*2* (c) 模式*3*

图3：循环协同增强机制的三种不同模式。



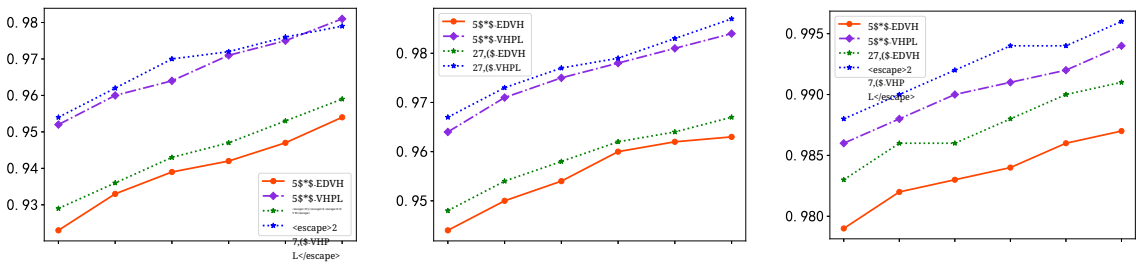
(a) 中文 英文 (b) 日文 英文 (c) FR-EN

图4：不同关系维度和本体维度下的实体对齐性能，rd 表示关系维度的数量，$t\ d$ 表示本体维度的数量。



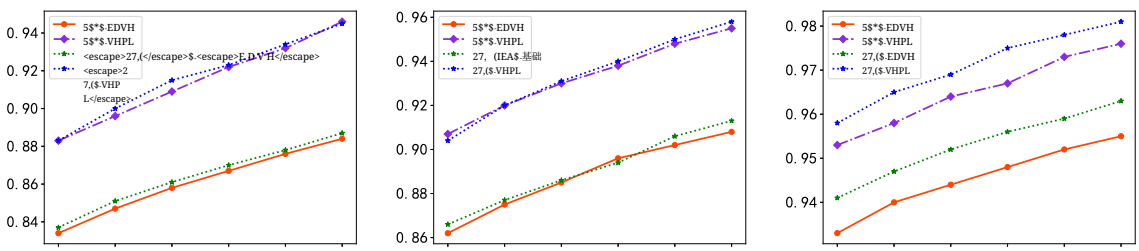
(a) 中文-英文　(b) 日文-英文　(c) 法文-英文

图5：不同训练种子实体对比例下的 Hits@1。



(a) 中文-英文　(b) 日文-英文　(c) 法文-英文

图6：不同训练种子实体对比例下的 Hits@10。



(a) 中文-英文　(b) 日文-英文　(c) FR-EN

图7：训练种子实体对比例不同时的平均倒数排名。

种子实体对比例分别从 25% 变化到 50% 时的 Hits@1、Hits@10 和平均倒数排名结果。结果表明，OTIEA 在 FR-EN 的所有评估指标上均取得了显著提升，在 ZH-EN 和日英上也表现出略微优势或相近的结果，尤其是在 Hits@10 上。

为探究不同图卷积网络深度的有效性，我们比较了三个具有不同图卷积网络深度的 OTIEA基础变体，如 $l = 1$, $l = 2$ 和 $l = 3$ 见表 4。结果表明，数据集对图卷积网络深度较为敏感。具体而言，两层图卷积网络在中文-英文和日文-英文上优势最大，而整体而言，一层图卷积网络在 FR-EN 上表现最佳。

为进一步探究关系维度和本体维度的影响，我们针对关系维度和本体维度分别取值为 50 至 300 的情况进行了消融实验。不同关系维度和本体维度下的 Hits@1、Hits@10 和平均倒数排名结果

如图所示4。从维度曲线可以看出，关系维度和本体维度不同的框架变体性能相近，表明维度对模型的影响较弱。可能的原因是，嵌入维度表示从真实空间到潜在低维向量空间的映射；只要满足表达需求，维度数量就不重要。

为证明循环协同增强机制的有效性，我们对不同模式的循环协同增强机制进行了详细比较，如图 3所示：

模式***1***：采用头尾顺序进行协同增强，且不包含循环过程的模式。

模式***2***：采用头尾头顺序进行循环协同增强的模式。

模式***3***：采用头尾头尾顺序进行循环协同增强的模式。

表 3 显示，模式***2***在我们的主要 OTIEA 实体对齐框架中使用时，在所有评估指标上均在三个子数据集上取得了最佳性能，这表明采用 模式***2*** 风格的循环协同增强机制是有效的。

表3: Comparison of different modes of Cycle Co-enhanced mechanism.

	中文-英文			日文-英文			法文-英文		
模式	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
mode1	79.4	93.4	0.846	81.7	95.0	0.866	91.9	98.5	0.944
mode2	**80.1**	**93.6**	**0.851**	**83.1**	**95.4**	**0.877**	**92.4**	**98.6**	**0.947**
mode3	79.8	**93.6**	0.849	83.0	95.3	0.876	92.3	98.5	**0.947**

表 4: Comparison of different depths ofGCNs.

	中文-英文			日文-英文			法文-英文		
深度	H@1	H@10	平均倒数排名	H@1	H@10	平均倒数排名	H@1	H@10	平均倒数排名
l=1	79.2	92.1	0.840	**83.6**	94.6	0.876	**93.2**	98.5	**0.952**
l=2	**80.1**	**93.6**	**0.851**	83.1	**95.4**	**0.877**	92.4	**98.6**	0.947
l=3	76.8	**93.5**	0.830	78.8	94.4	0.846	88.4	97.5	0.919

7 结论

本文提出了一种新颖的基于多层注意力机制的实体对齐模型 OTIEA。该模型采用端到端框架，学习用于实体对齐的代表性实体嵌入。本体增强三元组编码器通过三个阶段的内部交互注意力机制生成集成三元组表示，同时考虑三元组的内在关联性和本体对增强；三元组感知实体解码器则依据角色特征，通过三元组感知注意力循环传播三元组信息，以生成面向实体对齐的实体表示。此外，编码器-解码器架构通过考虑三元组不可分性，并在本体增强阶段将实体对视为整体，自然地解决重叠问题。在三个真实世界数据集上的实验以及进一步的消融研究表明，该框架具有有效性。

致谢。 本研究得到国家自然科学基金委员会资助，项目编号为 72274138。

声明与披露

- **利益冲突/竞争性利益：**作者没有需要披露的相关财务或非财务利益。
- **参与同意：** 所有作者声明同意参与。
- **发表同意：** 所有作者均声明同意发表。
- **作者贡献：** 所有作者共同参与了研究构思与设计。材料准备、数据收集和分析由Chengxiang Tan、Zhishuo Zhang完成。文章修改由Xueyan Zhao、Min Yang和Chaoqun Jiang完成。论文初稿由Zhishuo Zhang撰写，所有作者均对论文的先前版本提出了意见。所有作者均已阅读并批准最终稿件。
- **数据可用性和访问：** 本文使用的数据已在网络上开源，我们在 https://github.com/CodesForNlp/OTIEA 提供了可用的在线资源。
- **遵守伦理标准：**作者声明已取得知情同意，且不涉及任何伦理问题。

参考文献

[1]Bizer, C. 等. DBpedia——数据网络的一个结晶点。《网络语义学杂志》7（3），154–165（2009）。https://doi.org/https://doi.org/10.1016/j.websem.2009.07.002 .[2] Suchanek, F. M., Kasneci, G. & Weikum, G. YAGO：来自维基百科和 WordNet 的大型本体。《网络语义学杂志》6（3），203–217（2008）。https://doi.org/https: // doi.org/10.1016/j.websem.2008.06.001 .[3] Speer, R., Chin, J. & Havasi, C. ConceptNet 5.5：一个开放的多语言常识图谱。第三十一届 AAAI 人工智能大会论文集，4444–4451（2017）。[4] Zhang, R., Ma, H., Li, Q., Wang, Y. & Li, Z. Fire：通过特征交互和意图感知注意力网络增强知识的推荐方法。Applied Intelligence，1573–7497（2022）。https://doi.org/https:// doi.org/10.1007/s10489-022-04300-x .[5] Lingfei, Q., Wang, J., Lin, H. & Yang, L. 用于社区问答答案排序的多视角答题者表示。《信息科学》624，37–48（2023）。https://doi.org/https://doi.org/ 10.1016/j.ins.2022.12.040 .[6]Song, Y. 等. 中文儿科医学知识图谱的构建。语义技术，213–220（2020）。https://doi.org/https://doi.org/10.1007/ 978-981-15-3412-6 21 .[7]Wang, Z., Lv, Q., Lan, X. & Zhang, Y. 基于图卷积网络的跨语言知识图谱对齐。2018 年自然语言处理经验方法会议论文集，349–357（2018）。https://doi.org/10.18653/ v1/D18-1032 .[8] Mao, X., Wang, W., Xu, H., Wu, Y. & Lan, M. 关系反射实体对齐。第 29 届 ACM 国际会议论文集，关于信息与知识管理1095–1104（2020）。https://doi.org/https://doi.org/10.1145/3340531.3412001 。

[9]Peng, Y., Zhang, J., Zhou, C. & Xu, J. 利用关系结构相似性进行基于嵌入的实体对齐。*2020*年*IEEE*国际知识图谱会议（*ICKG*）123–130（2020年）。https://doi.org/10.1109/ICBK50248.2020.00027 。

[10] Zhang,Y., Gao,T., Lu,J., Cheng, Z.&Xiao, G. 面向跨语言知识图谱的自适应实体对齐。知识科学、工程与管理**12816**，474–487（2021年）。https://doi.org/https://doi.org/10.1007/978-3-030-82147-039。

[11]Bordes, A., Usunier, N., Garcia- Duran, A., Weston, J. & Yakhnenko, O. 用嵌入表示对多关系数据进行建模。*AdvancesinNeural Information Processing Systems***26** （2013年）。URLhttps://proceedings.neurips.cc/paper files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper. pdf 。

[12] Trisedya, B. D., Qi, J. & Zhang, R. 使用属性嵌入进行知识图谱间的实体对齐。*AAAI*人工智能会议论文集**33**，297–304（2019）。https://doi.org/https://doi.org/10.1609/aaai.v33i01.3301297 。

[13]He, F. 等。使用属性三元组和关系三元组进行无监督实体对齐。高级应用数据库系统**11446**，367–382（2019）。https://doi.org/https://doi.org/10.1007/978-3-030-18576-322。[14] Sun, Z., Hu, W. & Li, C. 通过联合属性保持嵌入实现跨语言实体对齐。*The SemanticWeb – ISWC2017* 628–644（2017）。https://doi.org/https://doi.org/10.1007/978-3-319-68288-437。[15]Yang, K., Liu, S., Zhao, J., Wang, Y. & Xie, B. COTSAE：用于实体对齐的结构与属性嵌入协同训练。*AAAI*会议论文集

关于人工智能34 (03), 3025–3032 (2020)。
https://doi.org/https://doi.org/10.1609/aaai.v34i03.5696 。

[16] Zhu, B. 等基于关系感知和属性参与的跨语言知识图谱实体对齐。*Applied Intelligence*(2022)。
https://doi.org/https://doi.org/10.1007/s10489-022-03797-6 。

[17]Gao, J., Liu, X., Chen, Y. 和 Xiong, F. Mhgcn：用于跨语言实体对齐的多视图高速公路图卷积网络。清华大学学报（自然科学版）**27** (4), 719–728 (2022)。
https://doi.org/10.26599/TST.2021.9010056 。

[18]Tang, X. 等. BERT-INT：一种基于 BERT 的知识图谱对齐交互模型. 第二十九届国际人工智能联合会议论文集 3174–3180（2020）. https://doi.org/https://doi.org/10.24963/ijcai.2020/439 .[19]Yang, H.-W. 等. 利用多方面信息对齐跨语言实体. arXiv 预印本 arXiv:1910.06575 4431– 4441（2019）. URL https://aclanthology. org/D19-1451. https://doi.org/10.18653/v1/ D19-1451 .[20] Zhang, F., Li, J. & Cheng, J. 通过属性和外部知识过滤改进实体对齐. Applied Intelligence（2022）. https://doi.org/https://doi.org/10.1007/s10489-022-03744-5 .[21]Tam, N. T. 等. 基于多阶卷积网络的知识图谱实体对齐. IEEE知识与数据工程汇刊 34 (9), 4201–4214（2022）. https://doi.org/10.1109/ TKDE.2020.3038654 .[22]Xie, Z. 等. 采用动态迭代训练的双门控图注意力网络，用于跨语言实体对齐. ACM信息系统汇刊 40 (3)（2021）. https://doi.org/10.1145/3471165 .

[23] Zhu, Y., Liu, H., Wu, Z. & Du, Y. 面向实体对齐的关系感知邻域匹配模型. *AAAI*人工智能会议（2020）.

[24]Wu, Y. 等面向异构知识图谱的关系感知实体对齐。第二十八届国际人工智能联合会议论文集5278–5284 (2019)。
https://doi.org/https://doi.org/10/gh7zth, https://arxiv.org/abs/1908.08210 。

[25] Zhu,R., Ma,M.& Wang, P. RAGA：用于全局实体对齐的关系感知图注意力网络。知识发现与数据挖掘进展。*PAKDD 2021*。501–513（2021年）。URLhttps://doi.org/10.1007/978-3-030-75762-540。https://doi.org/10.1007/978-3-030-75762-5 40。

[26]Wang, Z., Zhang, J., Feng, J. & Chen, Z. 通过在超平面上进行平移来嵌入知识图谱。*AAAI*会议论文集

I

人工智能会议**28**（1）（2014年）。
https://doi.org/https://doi.org/10.1609/aaai.v28i1.8870 。

[27] Lin, Y., Liu, Z., Sun, M., Liu, Y. & Zhu, X. 学习用于知识图谱补全的实体和关系嵌入。*AAAI*人工智能会议论文集**29** （1）（2015年）。
https://doi.org/https://doi.org/10.1609/aaai.v29i1.9491 。

[28]Sun, Z. 等人。TransEdge：为知识图谱翻译关系上下文化嵌入。语义网——*ISWC 2019* **11778**，612–629（2019）。https://doi.org/https://doi.org/10.1007/978-3-030-30793-635。

[29] Chen, M.、Tian, Y.、Yang, M. 和 Zaniolo, C.。用于跨语言知识对齐的多语言知识图谱嵌入。*IJCAI'17*1511–1517（2017）。[30]Sun, Z.、Hu, W.、Zhang, Q. 和 Qu, Y.。利用知识图谱嵌入自举实体对齐。第二十七届国际联合

*f*会议：人工智 *fi*能 *gence* 4396–4402

(2018)。
https://doi.org/https://doi.org/10.24963/ijcai.2018/611 。

[31]Huang, H. 等 通过关系预测进行跨知识图谱实体对齐。 *Knowledge-BasedSystems* **240**, 107813（2022）。
h
tt
ps://doi.org/https://doi.org/10.1016/j.knosys.2021.107813 。

[32] Zhu, H., Xie, R., Liu, Z. 和 Sun, M. 基于知识嵌入的迭代式实体对齐。 国际人工智能联合会议论文集（*IJCAI*）（2017）。
https://doi.org/https://doi.org/10.24963/ijcai.2017/595 。

[33]Mao, X., Wang, W., Xu, H., Lan, M. & Wu, Y. MRAEA：一种高效且稳健的跨语言知识图谱实体对齐方法。 第*13*届国际网络搜索与数据挖掘会议论文集 420–428（2020）。
h
ttps://doi.org/https://doi.org/10.1145/3336191.3371804 。
[34]Qi, D., Chen, S., Sun, X., Luan, R. & Tong, D. 一种仅使用结构信息进行实体对齐的多尺度卷积图网络。 *Applied Intelligence*（2022）。
h
t
t
p
s://doi.org/https://doi.org/10.1007/s10489-022-03916-3 。
[35] Zhu, Q., Zhou, X., Wu, J., Tan, J. & Guo, L. 面向多语言知识图谱的邻域感知注意力表征。第二十八届国际人工智能联合会议论文集 1943–1949（2019）。
https://doi.org/https://doi.org/10.24963/ijcai.2019/269 。
[36]Wu, Y., Liu, X., Feng, Y., Wang, Z. & Zhao, D. 用于实体对齐的邻域匹配网络。 第*58*届计算语言学协会年会论文集 6477– 6487（2020）。网址
https://aclanthology.org/2020.acl-main.578。
https://doi.org/10.18653/v1/2020.acl-main.578 。[37] Huang, Z. 等。 基于知识感知注意力图神经网络的多视图知识图谱融合。 *Applied Intelligence*（2022）. https://doi .org/https://doi .org/10.1007/s10489-022-03667-1 。[38]Yan, Z., Peng, R., Wang, Y. & Li, W. 面向知识图谱实体对齐的软自对齐与硬交叉对齐图注意力网络。 *Knowledge-Based Systems* **231**, 107415（2021年）。
h
ttps://doi.org/https://doi.org/10.1016/j.knosys.2021.107415 。
[39] Srivastava, R. K., Greff, K. & Schmidhuber, J. 高速公路网络。 *arXiv:1505.00387 [cs]*（2015年）。
https://doi.org/https://doi.org/10.48550/arXiv.1505.00387,
https://arxiv.org/abs/1505.00387 [cs]。