通过三元组感知注意力实现类型增强的集成三元组表示，用于跨语言实体对齐

Zhishuo Zhang 和**Chengxiang Tan***和**Haihang Wang** 和**Xueyan Zhao** 和**Min Yang**中国上海同济大学计算机科学与技术系{2110140,jerrytan,wanghh,1710839,2110136}@tongji.edu.cn

arXiv:2305.01556v1 [cs.CL] 2 May 2023

摘要

实体对齐（EA）是整合跨语言、跨领域知识图谱（KG）的关键任务，旨在从不同知识图谱中发现指向同一现实世界对象的实体。现有的大多数方法通过基于嵌入的方法挖掘三元组元素之间的相关性，生成对齐实体表示，却很少关注三元组的不可分性和实体角色的多样性。本文提出了一种名为 TTEA 的新框架——通过三元组感知注意力实现类型增强的集成三元组表示，用于跨语言实体对齐，以解决上述问题，同时考虑集成三元组特异性和实体角色特征。具体而言，该框架将关系视为语义空间与类型空间之间的信息载体，从而通过特异性感知三元组注意力平滑地控制空间变换和信息传播过程中的噪声影响。此外，我们的框架采用三元组感知的实体增强来建模三元组元素的角色多样性。在三个真实世界跨语言数据集上进行的大量实验表明，我们的框架优于当前最先进的方法。

1 引言

跨语言知识图谱（KG），例如 DBpedia（Bizer 等人，2009年）、YAGO（Suchanek 等人，2008年）和 ConceptNet（Speer 等人，2017年），已广泛应用于金融（Li 等人，2019b）、医疗（Song 等人，2019年；Sun 等人，2020a）和人工智能（Han 等人，2020年；Yang 等人，2021年；Pellegrino 等人，2021年）等诸多现实场景。由于大多数知识图谱是在不同语言或领域中独立构建的，因此，格式为（$head\ entity$，$relation$，$tail\ entity$）的数据会因异质性和规则特异性而难以有效整合。实体对齐（EA）是一项信息融合的关键任务，旨在从不同知识图谱中发现等价实体。

近年来，基于嵌入的方法受到了广泛关注。这类方法将实体和关系编码到潜在向量空间中，并通过测量嵌入距离进行实体对齐（Wang 等人，2018年；Mao 等人，2020b；Peng 等人，2020年；Zhang etal.，2021年）。为了获得适用于实体对齐的优质实体和关系表示，人们做出了许多努力。根据研究动机，这些方法大致可分为两类：基于 Trans 的方法和基于图神经网络的方法。

基于 Trans 的方法将三元组（$\boldsymbol{h}$、$\boldsymbol{r}$、$\boldsymbol{t}$）中的元素交互视为一种平移过程 $\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$。这些方法（Wang 等人，2014；Lin 等人，2015；Sun 等人，2019）有效且简单，但由于内部关联复杂且难以描述，无法形成三元组元素的完整表示。基于图神经网络的方法分为两类：基于图卷积网络的方法和基于图注意力网络的方法。前者通常通过邻居对齐和拓扑结构来体现实体对齐（Gao 等人，2022；Xie 等人，2021；Zhu 等人，2021b），后者则整合周围信息以增强嵌入（Wu 等人，2019a；Zhu 等人，2021a）。尽管这些方法可以通过融合邻居信息有效提升性能，但很少考虑集成三元组的特殊性和角色多样性。如图 1 所示，给定对齐实体对（e^1、e_1），实体 e^1 在知识图谱1中扮演头实体的角色，而 e_1 在 KG2 中扮演尾实体的角色；显然，e^1 对三元组（e^1、$r1$、e^2）的影响与 e_1 对三元组（e_2、$R1$、e_1）的影响并不一致。此外，头实体和尾实体之间可能存在多个不同类型的关系，如知识图谱1中的实体对（e^3、e^5）和 KG2 中的（e_3、e_5）所示。

为解决上述不足，本文提出 TTEA——**面向跨语言实体对齐的类型增强集成三元组表示与三元组感知注意力**，并基于以下

*通讯作者。

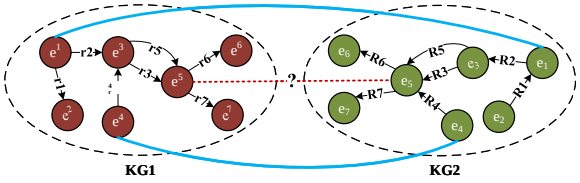


图 1：实体角色多样性以及实体之间多种关系的示例。知识图谱1和KG2中的实体之间的蓝色实线表示对齐关系。

直观假设：对于特定三元组，关系在语义空间和类型空间中具有通用性，能够充分利用三元组特异性和角色多样性。鉴于三元组中的元素不可分割，TTEA 引入类型增强集成三元组表示模块，在保留三元组特异性的同时捕获语义信息和类型信息。针对三元组中的多种关系和角色，我们设计了一种三元组感知的实体增强机制，以获得循环协同增强的头实体感知嵌入和尾实体感知嵌入。据我们所知，TTEA 是首个在实体对齐中利用头实体和尾实体的集成三元组特异性及角色多样性的工作。在三个跨语言知识图谱上的实验结果证明，TTEA 优于最先进的基线方法。源代码可在 GitHub 获取[1]。

总而言之，我们的主要贡献如下：

•我们提出一种新颖的视角，将关系视为空间变换过程中的信息载体，从而能够有效缓解映射过程中引入的噪声。

• 我们提出一种新颖的实体对齐框架，通过集成三元组表示和三元组感知实体增强，充分利用三元组特异性和角色多样性。

• 在公开数据集上进行的大量实验表明，TTEA 的表现显著且稳定地优于最先进的实体对齐基线方法。

2 问题定义

知识图谱可形式化为 $KG=(E,\ R,\ T)$，其中 E 和 R 分别是实体集和关系集，$T\subset E\times R\times E$ 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，发现指向 $KG=$ 和 E 中同一现实世界对象的实体对，记为 $\{(e_1,\ e_2)\ |e_1\in E1,\ e_2\in E2\}$，其中 e_1 和 e_2 等价。

[1]https://GitHub.com/CodesForNlp/TTEA

3 TTEA框架

我们提出了一种框架 TTEA，该框架基于类型增强的集成三元组表示，并通过三元组感知注意力机制实现三元组感知实体增强。TTEA 的整体架构如图 2 所示，主要由四个部分组成：拓扑结构聚合、类型增强的集成三元组表示、三元组感知实体增强和实体对齐策略。第一部分利用结构信息增强基于实体名称的嵌入，以进行初始化；随后，在类型增强的集成三元组表示部分生成具有特异性的集成三元组表示。接着，在三元组感知实体增强模块中，分别获得头实体和尾实体的三元组感知表示，并使二者相互循环增强。最后，在实体对齐策略部分，采用双向迭代策略扩充种子对，同时通过反向传播更新实体嵌入和参数。

3.1 拓扑结构聚合

我们首先将关系扩展为原始关系、反向关系和自关系的组合，以充分描述知识图谱中的拓扑结构。受 RAGA（Zhu 等人，2021a）的启发，我们还使用基于实体名称的嵌入作为初始化表示，随后采用带有高速公路网络（Srivastava 等人，2015年）的两层图卷积网络来聚合拓扑信息，同时保留实体主要语义。第 l 个高速公路-GCN 层计算如下：

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

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

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

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，$\boldsymbol{X}^{l}\in\mathbb{R}^{n\times d_e}$ 表示第 l 个隐藏层中的输入实体嵌入

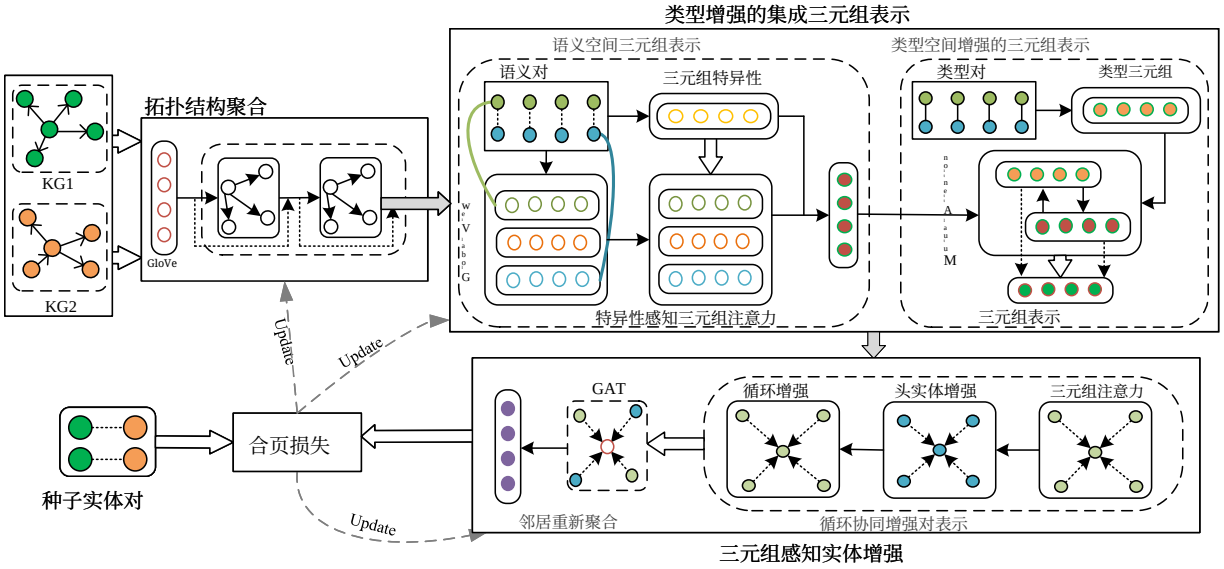


图2：TTEA框架的整体架构。

层， n 是知识图谱中的实体数量， d_e 是实体嵌入的维度， $\boldsymbol{X}^{l+1}$ 是第 l层的输出。 $\sigma(\cdot)$是激活函数， $\cdot$表示逐元素乘法， $\boldsymbol{W}^l$ 和$\boldsymbol{b}^l$ 分别是第 l个隐藏层中输入嵌入的权重矩阵和偏置向量。

3.2 类型增强的集成三元组表示

本模块通过挖掘三元组在语义空间和类型空间中的内在关联，生成集成三元组表示。

语义空间中的集成三元组表示

为描述关系整体性和多关系特征，将全局关系 $\bar{\boldsymbol{r}} \in \mathbb{R}^{2d_e}$ 计算为具有相同关系的头实体和尾实体拼接后的平均值。

$$\bar{\boldsymbol{r}}_r = \frac{1}{|T_r|} \sum_{(e_i,r,e_j)\in T_r} (\boldsymbol{x}_i \| \boldsymbol{x}_j) \tag{4}$$

其中， $\boldsymbol{x}_i$ 是 e_i的嵌入， $\|$ 表示拼接操作， T_r 是知识图谱中具有特定关系 r 的三元组集合。

此外，我们利用三元组特异性来缓解元素交互过程中的冗余和噪声。特定三元组（e_i, r, e_j）的局部关系首先定义为 $\tilde{\boldsymbol{r}}_{irj} = \boldsymbol{x}_i \| \boldsymbol{x}_j$, 基于此，三元组特异性表示为 $\boldsymbol{R}_{irj} = \boldsymbol{x}_i \| (\tilde{\boldsymbol{r}}_{irj} \boldsymbol{W}_e^l) \| \boldsymbol{x}_j$, 其中 $\boldsymbol{W}_e^l \in \mathbb{R}^{2d_e \times d_r}$ 是映射矩阵， d_r 是关系维度。

随后，我们设计了头实体感知注意力、尾实体感知注意力和关系感知注意力，以合理地整合全局三元组特征。值得注意的是，在关系感知三元组注意力机制中，我们使用整体关系 $\tilde{\boldsymbol{r}}_r = \boldsymbol{W}_e^g \bar{\boldsymbol{r}}_r + \boldsymbol{W}_e^l \tilde{\boldsymbol{r}}_{irj}$， 其中 $\boldsymbol{W}_e^g$, $\boldsymbol{W}_e^l \in \mathbb{R}^{2d_e \times d_r}$ 分别是全局关系和局部关系的映射矩阵。具体而言，头实体感知语义三元组表示 $\bar{\boldsymbol{x}}_{hr}$ 通过头实体感知三元组注意力获得：

$$c_{ir} = a^T \left(\boldsymbol{x}_i \boldsymbol{W}_h \| \tilde{\boldsymbol{R}}_{irj} \boldsymbol{W}_t^s \right) \tag{5}$$

$$\alpha_{ir} = \frac{\exp\left(\mathrm{LReLU}\left(c_{ir}\right)\right)}{\sum_{(e_{i'},r,e_{j'})\in T_r} \exp\left(\mathrm{LReLU}\left(c_{i'r}\right)\right)} \tag{6}$$

$$\bar{\boldsymbol{x}}_{hr} = \mathrm{LReLU}\left(\sum_{T_r} \left(\alpha_{ir} \boldsymbol{x}_i \boldsymbol{W}_h \right) \right) \tag{7}$$

其中， $a \in \mathbb{R}^{d_r \times 1}$ 是一个一维向量，用于将多维输入映射为标量。 $\boldsymbol{W}_h \in \mathbb{R}^{d_e \times d_r}$ 和 $\boldsymbol{W}_t^s \in \mathbb{R}^{(2d_e+d_r) \times d_r}$ 是头实体和语义空间中的集成三元组的线性变换矩阵。随后，可用相同的方法获得尾实体感知语义三元组表示 $\bar{\boldsymbol{x}}_{tr}$ 和关系感知语义三元组表示¯三元组表示 $\boldsymbol{R}_r$ 。

然后，将 $\bar{\boldsymbol{x}}_{hr}$、 $\bar{\boldsymbol{x}}_{tr}$¯和 $\boldsymbol{R}_r$组合，计算出语义空间中的融合集成三元组表示，并将其与主要特异性相加，得到语义

空间中的最终三元组表示，用于（e_i，r，e_j）$\in T$，以充分保留三元组的语义特异性：

$$\boldsymbol{S}_{irj} = \bar{\boldsymbol{x}}_{hr} + \bar{\boldsymbol{x}}_{tr} + \bar{\boldsymbol{R}}_r + (\tilde{\boldsymbol{R}}_{irj}\boldsymbol{W}_t^s) \quad (8)$$

类型空间增强的三元组表示

在此模块中，我们采用非线性映射，从语义嵌入 $\boldsymbol{X} \in \mathbb{R}^{n\times d_e}$生成类型嵌入 $\boldsymbol{X}^t \in \mathbb{R}^{n\times d_t}$，其中 d_t 为类型维度：

$$\boldsymbol{X}^t = \tanh(\boldsymbol{X}\boldsymbol{W} + b) \quad (9)$$

为了有效表征三元组，考虑到类型的不可分性，我们将类型三元组的各元素视为一个整体，以纳入类型信息。对于三元组 (e_i，r，e_j)，全局关系表示 $\bar{\boldsymbol{r}}_r^t \in \mathbb{R}^{2d_t}$ 通过对具有相同关系的实体对进行拼接并取平均值来计算r，然后将其与局部关系 $\tilde{\boldsymbol{r}}_{irj}^t = \boldsymbol{x_i} \| \boldsymbol{x_j}$相加，以生成类型三元组 $\tilde{\boldsymbol{R}}_{irj}^t$，如公式（10）-（12）所示，其中 $\boldsymbol{W}_t^r \in \mathbb{R}^{2d_t\times d_r}$ 为变换矩阵：

$$\bar{\boldsymbol{r}}_r^t = \frac{1}{|T_r|} \sum_{(e_i,r,e_j)\in T_r} (\boldsymbol{x}_i^t \| \boldsymbol{x}_j^t) \quad (10)$$

$$\boldsymbol{r}_{irj}^t = \bar{\boldsymbol{r}}_r^t + \tilde{\boldsymbol{r}}_{irj}^t \quad (11)$$

$$\tilde{\boldsymbol{R}}_{irj}^t = \boldsymbol{x}_i^t \| (\boldsymbol{r}_{irj}^t \boldsymbol{W}_t^r) \| \boldsymbol{x}_j^t \quad (12)$$

接着，我们设计了语义-类型互注意力，以获得增强的类型空间三元组表示$\bar{}\boldsymbol{T}_r$：

$$\alpha_{tr} = \frac{\exp\left(\mathrm{LReLU}\left(a^T\left(\boldsymbol{S}_{irj} \| \tilde{\boldsymbol{R}}_{irj}^t \boldsymbol{W}_t\right)\right)\right)}{\sum_{T_r} \exp\left(\mathrm{LReLU}\left(a^T\left(\boldsymbol{S}_{i'r j'} \| \tilde{\boldsymbol{R}}_{i'r j'}^t \boldsymbol{W}_t\right)\right)\right)} \quad (13)$$

$$\bar{\boldsymbol{T}}_r = \mathrm{ReLU}\left(\sum_{(e_i,r,e_j)\in T_r} (\alpha_{tr}\boldsymbol{S}_{irj})\right) \quad (14)$$

其中 $\boldsymbol{W}_t \in \mathbb{R}^{(2d_t+d_r)\times d_r}$是类型三元组的可训练参数。增强的全局表示$\bar{}\boldsymbol{S}_r$可以通过相同的方式生成。

最终，在保留主要类型特征的同时，得到类型增强的集成三元组表示 $\boldsymbol{T}_{ijr} \in \mathbb{R}^{(d_r+d_t)}$。

$$\boldsymbol{T}'_{irj} = \boldsymbol{S}_{irj} + \bar{\boldsymbol{S}}_r + (\tilde{\boldsymbol{R}}_{irj}^t \boldsymbol{W}_t) + \bar{\boldsymbol{T}}_r \quad (15)$$

$$\boldsymbol{T}_{irj} = \boldsymbol{T}'_{irj} \| \bar{\boldsymbol{r}}_r^t \quad (16)$$

3.3 三元组感知的实体增强

循环协同增强实体对表示

一个实体在不同三元组中可能分别扮演头实体或尾实体的不同角色，而头实体e_i和尾实体 e_j对三元组 (e_i，r，e_j) 的影响截然不同。在此模块中，头实体和尾实体分别通过三元组注意力生成，并以循环方式相互增强。

$$\alpha_{th} = \frac{\exp\left(\mathrm{LReLU}\left(a^T\left(\boldsymbol{T}_{ijr}\boldsymbol{W}_h^c \| \boldsymbol{x}_i\right)\right)\right)}{\sum_{(e_i,r',e'_j)\in T} \exp\left(\mathrm{LReLU}\left(a^T\left(\boldsymbol{T}_{ij'r'}\boldsymbol{W}_h^c \| \boldsymbol{x}_i\right)\right)\right)} \quad (17)$$

$$\boldsymbol{x}_i = \boldsymbol{x}_i + \mathrm{ReLU}\left(\sum_{(e_i,r,e_j)\in T} (\alpha_{th}\boldsymbol{T}_{ijr}\boldsymbol{W}_h^c)\right) \quad (18)$$

$$\alpha_{tt} = \frac{\exp\left(\mathrm{LReLU}\left(a^T\left(\boldsymbol{T}_{ijr}\boldsymbol{W}_t^c \| \boldsymbol{x}_j\right)\right)\right)}{\sum_{(e'_i,r',e_j)\in T} \exp\left(\mathrm{LReLU}\left(a^T\left(\boldsymbol{T}_{i'jr'}\boldsymbol{W}_t^c \| \boldsymbol{x}_j\right)\right)\right)} \quad (19)$$

$$\boldsymbol{x}_j = \boldsymbol{x}_j + \mathrm{ReLU}\left(\sum_{(e_i,r,e_j)\in T} (\alpha_{tt}\boldsymbol{T}_{ijr}\boldsymbol{W}_t^c)\right) \quad (20)$$

其中，$\boldsymbol{W}_h^c$，$\boldsymbol{W}_t^c \in \mathbb{R}^{(d_r+d_t)\times d_e}$是三元组的可训练权重参数。

邻居重新聚合

在 TTEA 的最后一个部分，我们应用一个 GAT 层，将邻居信息与建模后的表示重新聚合，并生成最终的实体表示 $\boldsymbol{x}_i^f$，用于实体对齐。

$$\alpha_{ij} = \frac{\exp\left(\mathrm{LReLU}\left(a^T\left(\boldsymbol{x}_i \| \boldsymbol{x}_j\right)\right)\right)}{\sum_{e_k\in N_i} \exp\left(\mathrm{LReLU}\left(a^T\left(\boldsymbol{x}_i \| \boldsymbol{x}_k\right)\right)\right)} \quad (21)$$

$$\boldsymbol{x}_i^f = \boldsymbol{x}_i \| \left(\mathrm{ReLU}\left(\sum_{e_j\in N_i} (\alpha_{ij}\boldsymbol{x}_j)\right)\right) \quad (22)$$

其中，N_i 是 e_i的邻居实体集合。

3.4 实体对齐策略

采用曼哈顿距离衡量实体之间的相似度，并据此将基于间隔的损失函数 L 定义为式（24）。此外，我们参照 MRAGA（Mao 等人，2020a），采用双向迭代策略，基于负样本方法扩充训练种子对。

$$dis\left(e_i, e_j\right) = \left\| \boldsymbol{x}_i^f - \boldsymbol{x}_j^f \right\|_1 \tag{23}$$

$$L = \sum_{(e_i, e_j) \in S} \max\left(dis(e_i, e_j) - dis\left(e_i', e_j'\right) + \lambda, 0\right) \tag{24}$$

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。

4 实验设置

4.1 数据集

为了与以往方法进行可靠且公平的比较，我们在三个真实世界的多语言数据集上评估 TTEA。这些数据集来自简化的 DBP15K，如表 1 所述。该数据集通过从初始 DBP15K 中移除大量无关实体和关系构建而成，几乎所有相关研究都采用了该数据集。

表 1：简化 DBP15K 的统计数据。

DBP15K		实体	关系	关系三元组	链接
中英	ZH	19388	1700	70414	15000
	EN	19572	1322	95142	
日英	JA	19814	1298	77214	15000
	EN	19780	1152	93484	
法英	FR	19661	902	105998	15000
	EN	19993	1207	115722	

4.2 基线方法

为全面评估我们的方法，我们将 TTEA 与基于 Trans、基于图神经网络和半监督实体对齐方法进行比较：

– **基于 Trans 的方法：** MTransE(Chen 等人，2016年)、JAPE(Sun 等人，2017年)、BootEA(Sun 等人，2018年)、TransEdge(Sun 等人，2019年)、RpAlign(Huang 等人，2022a)。

–**基于图神经网络的方法：** ***(1)*** 基于 ***GCN*** 的方法：GCN-Align(Wang 等人，2018年)，HMAN(Yang 等人，2019年)，HGCN(Wu 等人，2019b)，MCEA(Qi 等人，2022年)。 ***(2)*** 基于 ***GAT*** 的方法：NAEA(Zhu 等人，2019年),RDGCN(Wu 等人，2019a), NMN(Wu 等人，2020年),KAGNN(Huang 等人，2022b), MRGA(Ding 等人，2021年), SHEA(Yan 等人，2021年), RAGA-l(Zhu 等人，2021a).

– **半监督方法：** MRAEA(Mao 等人，2020a), RREA-semi(Mao 等人，2020b),RAGA-semi(Zhu 等人，2021a), MCEA-semi(Qi 等人，2022)。

需要指出的是，为保证公平性，像 RAEA（Zhu 等人，2022年）和 RNM（Zhu 等人，2021b）这类需要额外信息的方法不作为基线方法。

4.3 模型变体

为对框架中的不同组件进行有效评估，我们实现了 TTEA 的三个变体，以验证其有效性：

(1) 不含E：不使用集成三元组注意力的简化版 TTEA。(2) 不含T：不使用类型空间增强模块的简化版 TTEA。

(3) 不含C：不使用 Cycleco 增强模块的简化版 TTEA。

4.4 实现细节

我们使用 Glove(Pennington 等人，2014)生成初始实体嵌入。为与基线方法进行公平比较，我们使用 30% 的对齐种子进行训练，其余用于测试。高速公路图卷积网络的深度 l 为 2，关系维度 d_r 和类型维度 d_t 均为 100。用于更新负样本的训练轮次 p 为 5，最近邻负样本的数量 k 为 5。间隔超参数 λ 为 3.0。

4.5 评估指标

按照惯例，我们报告 Hits@1、Hits@10 和平均倒数排名，以评估实体对齐性能。Hits@k 衡量排名位于前 k 位的正确对齐所占百分比，平均倒数排名则是结果排名倒数的平均值。Hits@k 和平均倒数排名的得分越高，表示性能越好。

5 结果与分析

5.1 总体性能

三个数据集上的基线方法结果列于表 2中，结果来自使用源代码复现的方法或原始论文。The

表 2：实体对齐总体性能。

方法	中英			日英			法英		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
MTransE(2017年)	30.8	61.4	0.364	27.8	57.4	0.349	24.4	55.5	0.335
JAPE（2017年）	41.2	74.4	0.490	36.2	68.5	0.476	32.4	66.7	0.430
BootEA（2018年）	62.9	84.7	0.703	62.2	85.4	0.701	65.3	87.4	0.731
TransEdge（2019年）	73.5	91.9	0.801	71.9	93.2	0.795	71.0	94.1	0.796
RpAlign（2022年）	74.8	88.8	0.794	72.9	89.0	0.872	75.2	89.9	0.801
GCN-Align（2018年）	41.2	74.4	0.549	39.9	74.4	0.546	37.3	74.5	0.532
HMAN（2019年）	56.1	85.9	0.67	55.7	86.0	0.67	55.0	87.6	0.66
HGCN(2019年)	72.0	85.7	0.768	76.6	89.7	0.813	89.2	96.1	0.917
MCEA(2022年)	72.4	93.4	0.800	71.9	94.0	0.800	73.9	95.3	0.820
NAEA(2019年)	65.0	86.7	0.720	64.1	87.3	0.718	67.3	89.4	0.752
RDGCN(2019年)	70.8	84.6	0.751	76.7	89.5	0.812	88.6	95.7	0.908
NMN(2020年)	73.3	86.9	0.781	78.5	91.2	0.827	90.2	96.7	0.924
KAGNN(2022年)	73.6	87.3	0.786	79.4	91.1	0.837	92.0	97.6	0.941
MRGA(2021年)	75.5	90.5	0.783	73.4	90.3	0.771	75.7	91.7	0.791
SHEA(2021年)	76.3	91.4	0.835	82.1	93.8	0.860	90.5	97.0	0.902
RAGA-l(2021年)	79.8	93.3	0.847	82.9	95.0	0.875	91.4	98.2	0.940
TTEA-base(不含E)	78.9	93.4	0.842	82.0	95.0	0.868	91.9	98.5	0.944
TTEA-base(不含T)	78.7	93.4	0.841	81.4	95.1	0.864	91.4	98.4	0.940
TTEA-base(不含C)	79.9	93.5	0.849	82.9	95.1	0.875	92.2	98.3	0.946
TTEA-base（我们的方法）	**80.2**	**93.8**	**0.852**	**83.1**	**95.4**	**0.876**	**92.4**	**98.6**	**0.947**
MRAEA（2020年）	75.2	92.3	0.824	75.3	93.3	0.825	78.1	94.7	0.843
RREA-semi（2020年）	80.1	94.8	0.857	80.2	95.2	0.858	82.7	96.6	0.881
MCEA-semi（2022年）	81.4	95.6	0.867	80.7	95.7	0.864	84.1	97.0	0.891
RAGA-semi（2021年）	85.7	96.0	0.896	88.9	97.1	0.920	94.0	98.8	0.958
TTEA-semi（我们的方法）	**86.3**	**96.2**	**0.901**	**89.2**	**97.6**	**0.924**	**94.7**	**99.0**	**0.964**

实线用于区分基于 Trans 的方法、基于图神经网络的方法和基于半监督的方法；点线用于区分基于图卷积网络的方法和基于图注意力网络的方法；TTEA 的变体位于虚线下方。

对于基于 Trans 的方法，TransEdge 和 RpAlign 凭借对三元组元素的独特表示，优于 MTransE、JAPE、BootEA 和 NAEA。具体而言，RpAlign 凭借其关系预测和自训练机制，在 Hits@1 上表现更佳；而 TransEdge 通过针对特定头尾实体对对关系表示进行上下文化，在中文-英文和日文-英文数据集上取得了更优异的 Hits@10 和平均倒数排名。对于基于图卷积网络的方法，GCN-Align 对关系三元组的利用较为浅显，因此结果最差；MCEA 则通过扩展长尾实体的卷积区域，表现优于其他方法。此外，NAEA、RDGCN、NMN、KAGNN、MRGA、SHEA 和 RAGA-l 均采用图注意力网络来获得细粒度表示，因此无疑取得了出色的性能。其中，RAGA-l 的结果最佳，它通过……

从实体中生成关系，然后通过注意力机制将关系聚合到实体中。与基线方法相比，我们的 TTEA 在三个数据集的所有评估指标上均表现最佳，同时考虑了三元组特异性和角色多样性。

5.2 消融分析

TTEA 组件的影响

表明，表2 中 TTEA-base(wo-E)、TTEA-base(wo- T) 和 TTEA-base(wo-C) 的结果显示，尽管 TTEA 中的集成三元组注意力、类型空间增强模块和循环协同增强模块均有所改进，但类型空间增强模块的效果更为显著。此外，表 3 对循环协同增强模块的三种不同循环顺序模式进行了比较，以探究合适的循环形式。

模式***1***：采用头尾顺序的协同增强过程模式。

模式***2***：采用头尾头顺序的循环协同增强过程模式。

模式***3***：循环协同增强过程的模式

按头尾头尾顺序。

从表3 可以看出，TTEA 采用的模式2在中英和法英数据集上表现更好_ _，这得益于合理的循环过程；而模式3在日英数据集上取得了几乎相同的表现_。

表3：循环协同增强模块不同模式的比较。

	中英			日英			法英		
模式	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
模式1	79.8	93.5	0.849	**83.1**	95.0	0.875	92.2	98.4	0.946
模式2	**80.2**	**93.8**	**0.852**	**83.1**	**95.4**	0.876	**92.4**	**98.6**	**0.947**
模式3	79.9	93.6	0.849	**83.1**	**95.4**	**0.877**	92.3	98.5	**0.947**

表4：不同深度的高速公路图卷积网络比较。

	中英			日英			法英		
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	79.4	92.4	0.841	**83.4**	94.5	0.875	**92.9**	98.4	**0.950**
l=2	**80.2**	**93.8**	**0.852**	83.1	**95.4**	**0.876**	92.4	**98.6**	0.947
l=3	77.0	**93.8**	0.832	79.4	94.8	0.851	88.6	97.7	0.921

GCN深度的影响

为探究不同高速公路图卷积网络深度的影响 l，我们将具有不同深度的 TTEA 变体与 $l = 1$、$l = 2$ 和 $l = 3$ 进行了比较，结果见表4。结果表明，对于结构复杂的 ZH_EN 和 JA_EN，具有两层高速公路图卷积网络的 TTEA 变体表现出最大优势；而对于实体语义依赖较强的 FR_EN，一层高速公路图卷积网络的表现最佳。

关系与类型维度的影响

TTEA 中有两个维度超参数：关系维度 d_r 和类型维度 d_t。我们分别在六种不同的关系与类型维度（50、100、150、200、250 和 300）下评估 TTEA，以探究维度的影响。图3中的结果表明，不同的关系与类型维度下性能相近，这说明维度对 TTEA 的影响很小。特别是在 ZH_EN 上，当 $d_r = d_t = 100$时，可以取得最佳结果。

种子实体对的影响

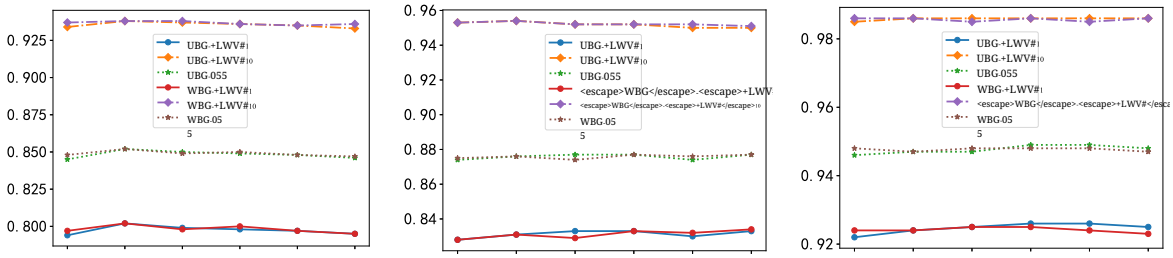
为探究不同训练种子对结果的影响，我们通过将训练种子的比例从25%逐步提高至50%，步长为5%，比较了RAGA-l和RAGA-semi与TTEA- base及TTEA-semi的表现。图4、图5和图6 分别展示了不同种子比例下的Hits@1、Hits@10和平均倒数排名。结果表明在三个数据集的所有评估指标上，无论是基础模式还是半监督局部对齐方法这两种模式，TTEA都优于RAGA。随着训练种子的增加，TTEA-base在三个数据集上的Hits@1、Hits@10和平均倒数排名曲线均比RAGA-l更加陡峭，表现出更好的性能和潜力。

6 相关工作

大多数基于 Trans 的方法采用 TransE（Bordes 等人，2013）及其变体（Wang 等人，2014；Lin 等人，2015）来嵌入实体和关系。一类方法将不同知识图谱中的实体和关系嵌入到不同的潜在空间中，然后构建用于实体对齐的映射变换（Chen 等人，2016；Zhu等，2017；Sun 等人，2019；Song 等人，2021；Xiang 等人，2021）。另一类方法则通过参数共享（Kang 等人，2020）和扩展对齐关系（Huang 等人，2022a），将不同知识图谱中的实体和关系嵌入到统一的潜在空间中。然而，这些方法忽略了特定三元组内部的相关性。

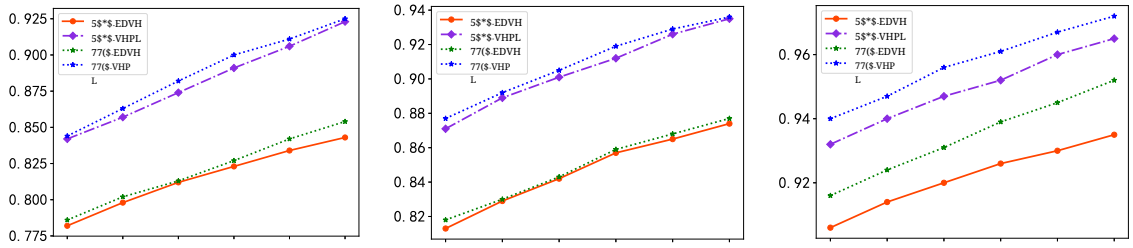
随着图神经网络应用于实体对齐，研究人员借助图神经网络取得了显著进展。GCN-Align（Wang 等人，2018年）是首个通过图卷积网络增强实体嵌入的工作，此后，基于图卷积网络的方法扩展为聚合邻居信息。一些工作利用邻居实体和关系对齐来解决实体对齐问题，其假设是等价实体拥有近似的邻居（Wu 等人，2020年）；另一些相关工作则利用拓扑结构，通过图卷积网络强化实体嵌入（Wu 等人，2019b；Zhu 等人，2021b；Peng 等人，2020年；Li 等人，2022年；Sun 等人，2020b）。

随后，基于图注意力网络的方法会考虑到不同邻居实体的重要性各不相同。有些工作采用邻居注意力来学习实体嵌入（Xie 等人，2021年；Huang 等人，2022b；Jiang 等人，2021年；Ding 等人，2021年；Zhu 等人，2019年）。据我们所知，RAGA（Zhu 等人，2021a）通过对实体与关系之间的相关性进行建模，取得了当前最先进的结果。另一些工作则利用跨知识图谱注意力，在对齐实体对之间传播邻居信息（Wu 等人，2019a；Yan 等人，2021年；Li 等人，2019a）。此外，一些基于图神经网络的方法还整合了外部资源，以增强嵌入（Yang 等人，2019年；Trisedya 等人，2019年；Yang 等人，2020年；Zhu 等人，2022年；Gao 等人，2022年）。



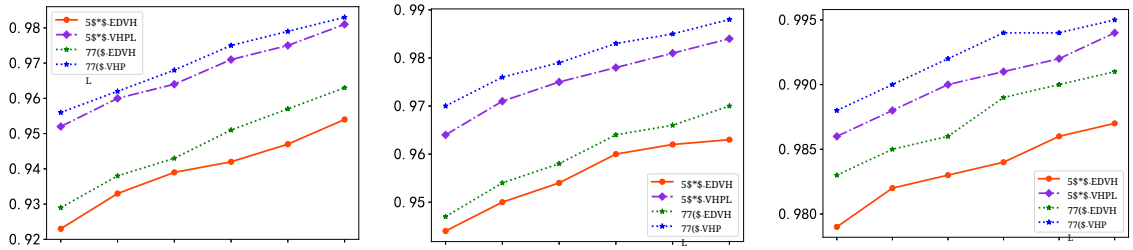
(a) ZH_EN (b) JA_EN (c) FR_EN

图 3：不同关系与类型维度下的实体对齐性能。



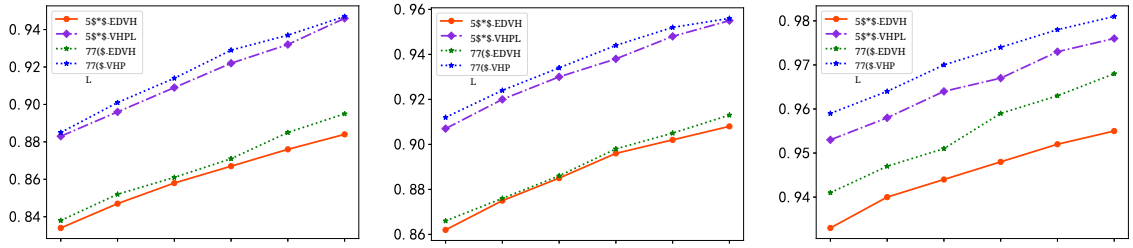
(a) ZH_EN (b) JA_EN (c) 法语_英语

图4：不同训练种子对下的Hits@1。



(a) 中文_英语 (b) 日语_英语 (c) 法语_英语

图5：不同训练种子对下的Hits@10。



(a) ZH_EN (b) JA_EN (c) FR_EN

图6：不同训练种子对下的平均倒数排名。

现有基于图神经网络的方法有效提升了实体对齐的性能，但未考虑实体角色的多样性以及集成三元组的多层级表示。

近年来，自举学习（Sun 等人，2018年）和迭代训练策略（Zhu 等人，2017年）被引入以应对种子实体对不足的问题。具体而言，双向迭

代训练策略（Mao 等人，2020a）近年来得到了广泛应用（Mao 等人，2020b；Trung 等人，2020年；Zhang 等人，2021年；Qi 等人，2022年），TTEA 也采用了该策略来提升性能。

7 结论

本文针对三元组特异性利用不足和实体角色多样性问题，提出了一种新颖的框架 TTEA——通过三元组感知注意力实现跨语言实体对齐的类型增强的集成三元组表示。TTEA 对语义空间与类型空间中的角色特征和关系交互进行建模，能够融合集成三元组特异性，并学习头实体与尾实体表示的循环协同增强。与最先进的基线方法相比，我们的模型在三个真实世界的跨语言数据集上取得了最佳性能。

局限性

我们的框架能够通过有效挖掘集成三元组特异性和实体角色多样性，摆脱对外部资源和结构异质性的限制，因此适用于大多数知识图谱的知识补全。然而，TTEA 所采用的基于实体名称的初始嵌入可能无法获得；这是一个至关重要的改进部分，我们将在未来加以解决。此外，我们的框架针对跨语言事件整合中的元素级任务，而事件级对齐任务则是我们未来工作的另一部分。

伦理声明

我们的论文提出了 TTEA，这是一种新颖的跨语言实体对齐框架，用于建模三元组特异性和角色多样性。TTEA 既不会在模型中引入任何社会/伦理偏见，也不会放大数据中的偏见。我们的模型基于 PyTorch 的公共库构建。此外，TTEA 使用公开数据集进行训练和测试。我们预计不会产生任何直接的社会后果或伦理问题。

参考文献

Christian Bizer、Jens Lehmann、Georgi Kobilarov、Sören Auer、Christian Becker、Richard Cyganiak 和 Sebastian Hellmann。2009年。DBpedia——数据网络的一个结晶点。*Journal of Web Semantics*，7(3):154–165。

Antoine Bordes、Nicolas Usunier、Alberto Garcia-Duran、Jason Weston 和 Oksana Yakhnenko。2013年。用于多关系数据建模的平移嵌入。载于神经信息处理系统进展，第 26 卷。Curran Associates, Inc. Muhao Chen、Yingtao Tian、Mohan Yang 和 Carlo Zaniolo。2016年。用于跨语言知识对齐的多语言知识图谱嵌入。Linyi Ding、Weijie Yuan、Kui Meng 和 Gongshen Liu。2021年。用于知识图谱对齐的多角色图注意力网络。载于*2021*年国际联合神经网络会议（*IJCNN*），第 1–8 页。

Jianliang Gao、Xiangyue Liu、Yibo Chen 和 Fan Xiong。2022年。Mhgcn：用于跨语言实体对齐的多视图高速公路图卷积网络。清华科学与技术，27(4):719–728。

Jiale Han、Bo Cheng 和 Xu Wang。2020年。基于文本增强知识图谱与超边注入的开放域问答。载于计算语言学协会论文集：*EMNLP 2020*年，第 1475–1481 页，线上。计算语言学协会。

Hongren Huang、Chen Li、Xutan Peng、Lifang He、Shu Guo、Hao Peng、Lihong Wang 和 Jianxin Li。2022a。通过关系预测实现跨知识图谱实体对齐。*Knowledge-BasedSystems*，240：107813。

Zhichao Huang、Xutao Li、Yunming Ye、Baoquan Zhang、Guangning Xu 和 Wensheng Gan。2022b。通过知识感知注意力图神经网络实现多视图知识图谱融合。*Applied Intelligence*，第 1–20 页。

Shixian Jiang、Tiezheng Nie、Derong Shen、Yue Kou 和 Ge Yu。2021年。通过联合图注意力和翻译表示实现知识图谱实体对齐。载于 Chunxiao Xing、Xiaoming Fu、Yong Zhang、Guigang Zhang 和 Chaolemen Borjigin 编，*Web InformationSystems and Applications*，第 12999 卷，第 347–358 页。施普林格国际出版，Cham。

Shize Kang、Lixin Ji、Zhenglian Li、Xindi Hao 和 Yuehang Ding。2020年。基于 TransC 的迭代式跨语言实体对齐。*IEICE TRANSACTIONSon Information and Systems*，103(5):1002–1005。Chengjiang Li、Yixin Cao、Lei Hou、Jiaxin Shi、Juanzi Li 和 Tat-Seng Chua。2019a。通过联合知识嵌入模型和跨图模型实现半监督实体对齐。第 2723–2732 页。Juncheng Li、Guanghui Yan、Shan Wang 和 Junbo Bao。2019b。面向时间序列投资行为的金融网络表示学习。见*2019*年 *IEEE* 第十届软件工程与服务科学国际会议（*ICSESS*），第 277–281 页。IEEE。

Xinhang Li、Yong Zhang 和 Chunxiao Xing。2022年。通过关系知识蒸馏联合学习知识嵌入与邻域共识，实现实体对齐。

Yankai Lin、Zhiyuan Liu、Maosong Sun、YangLiu 和 Xuan Zhu。2015年。学习用于知识图谱补全的实体和关系嵌入。

Xin Mao、Wenting Wang、Huimin Xu、Man Lan 和 YuanbinWu。2020a。MRAEA：一种高效且稳健的跨语言知识图谱实体对齐方法。美国纽约州纽约市。美国计算机协会。

Xin Mao、Wenting Wang、Huimin Xu、Yuanbin Wu 和 Man Lan。2020b。关系反射实体对齐。美国纽约州纽约市。美国计算机协会。

Maria Angela Pellegrino、Mario Santoro、Vittorio Scarano 和 Carmine Spagnuolo。2021年。用于知识图谱问答的自动技能生成。载于 欧洲语义网会议，第38–43页。施普林格。

Yanhui Peng、Jing Zhang、Cangqi Zhou 和 Jian Xu。2020年。基于嵌入的实体对齐：利用关系结构相似性。载于 *2020*年 *IEEE* 国际知识图谱会议（*ICKG*），第123–130页。

Jeffrey Pennington、Richard Socher 和 Christopher D Manning。2014年。GloVe：用于词表示的全局向量。载于 *2014*年自然语言处理实证方法会议论文集（*EMNLP*），第1532–1543页。Donglin Qi、Shudong Chen、Xiao Sun、Ruipeng Luan 和 Da Tong。2022年。一种仅使用结构信息进行实体对齐的多尺度卷积图网络。*AppliedIntelligence*，第1–11页。Xiuting Song、Han Zhang 和 Luyi Bai。2021年。使用实体类型匹配进行知识图谱之间的实体对齐。载于知识科学、工程与管理国际会议，第578–589页。Yu Song、Linkun Cai、Kunli Zhang、Hongying Zan、Tao Liu 和 Xiaohui Ren。2019年。中文儿科医学知识图谱的构建。载于联合国际语义技术会议，第213–220页。施普林格。Robyn Speer、Joshua Chin 和 Catherine Havasi。2017年。ConceptNet 5.5：一个开放的多语言通用知识图谱。Rupesh Kumar Srivastava、Klaus Greff 和 Jür- gen Schmidhuber。2015年。高速公路网络。*arXiv:1505.00387 [cs]*。Fabian M Suchanek、Gjergji Kasneci 和 Gerhard Weikum。2008年。YAGO：来自维基百科和 WordNet 的大型本体。*Journal of Web Semantics*，6(3):203–217。

Haixia Sun、Jin Xiao、Wei Zhu、Yilong He、Sheng Zhang、Xiaowei Xu、Li Hou、Jiao Li、Yuan Ni、Guo- tong Xie 等。2020a。用于增强理赔数据中欺诈、浪费和滥用检测的医学知识图谱：模型开发与性能评估。*JMIR* 医学信息学，8(7):e17653。

Zequn Sun、Wei Hu 和 Chengkai Li。2017年。通过联合属性保留嵌入实现跨语言实体对齐。载于语义网——*ISWC 2017*，第 628–644 页，Cham。施普林格国际出版公司。

Zequn Sun、Wei Hu、Qingheng Zhang 和 Yuzhong Qu。2018年。利用知识图谱嵌入自举实体对齐。载于*IJCAI*，第 18 卷，第 4396–4402 页。

Zequn Sun、Jiacheng Huang、Wei Hu、Muhao Chen、Lingbing Guo 和 Yuzhong Qu。2019年。Transedge：为知识图谱转换关系上下文化嵌入。载于国际语义网会议，第 612–629 页。

ZequnSun、Chengming Wang、WeiHu、Muhao Chen、Jian Dai、Wei Zhang 和 Yuzhong Qu。2020b。采用门控多跳邻域聚合的知识图谱对齐网络。34(01):222–229。

Bayu Distiawan Trisedya、Jianzhong Qi 和 Rui Zhang。2019年。利用属性嵌入实现知识图谱间的实体对齐。33(01):297–304。Huynh Thanh Trung、Tong Van Vinh、Nguyen Thanh Tam、Hongzhi Yin、Matthias Weidlich 和 Nguyen Quoc Viet Hung。2020年。基于无监督多阶卷积网络的自适应网络对齐。载于*2020*年 *IEEE*第*36*届国际数据工程会议（*ICDE*），第85–96页。Zhen Wang、Jianwen Zhang、Jianlin Feng 和 Zheng Chen。2014。通过在超平面上平移实现知识图谱嵌入。28(1)。Zhichun Wang、Qingsong Lv、Xiaohan Lan 和 Yu Zhang。2018年。通过图卷积网络实现跨语言知识图谱对齐。载于*2018*年自然语言处理经验方法会议论文集，比利时布鲁塞尔。计算语言学协会。Yuting Wu、Xiao Liu、Yansong Feng、Zheng Wang、Rui Yan 和 Dongyan Zhao。2019a。面向异构知识图谱的关系感知实体对齐。第二十八届人工智能国际联合会议论文集，第5278–5284页。Yuting Wu、Xiao Liu、Yansong Feng、Zheng Wang 和 Dongyan Zhao。2019b。联合学习实体对齐中的实体和关系表示。*arXiv*预印本 *arXiv:1909.09317*。

Yuting Wu、Xiao Liu、Yansong Feng、Zheng Wang 和 Dongyan Zhao。2020年。用于实体对齐的邻域匹配网络。载于第*58*届计算语言学协会年会论文集，第6477–6487页，线上。计算语言学协会。

Yuejia Xiang、Ziheng Zhang、Jiaoyan Chen、Xi Chen、Zhenxi Lin 和 Yefeng Zheng。2021年。Ontoea：通过联合知识图谱嵌入实现本体引导的实体对齐。载于计算语言学协会论文集：*ACL-IJCNLP 2021* 研究成果，第1117–1128页，线上。计算语言学协会。

Zhiwen Xie、Runjie Zhu、Kunsong Zhao、Jin Liu、Guangyou Zhou 和 Jimmy Xiangji Huang。2021年。用于跨语言实体对齐的双门控图注意力网络与动态迭代训练。40(3)。

Zhihuan Yan、Rong Peng、Yaqian Wang 和 Wei- dong Li。2021年。用于知识图谱实体对齐的软自注意力与硬交叉图注意力网络。*Knowledge-BasedSystems*，231:107415。

Hsiu-Wei Yang、Yanyan Zou、Peng Shi、Wei Lu、Jimmy Lin 和 Xu Sun。2019年。利用多方面信息进行跨语言实体对齐。*arXiv* 预印本 *arXiv:1910.06575*。

Kai Yang、Shaoqin Liu、Junfeng Zhao、Yasha Wang 和 Bing Xie。2020年。Cotsae：用于实体对齐的结构与属性嵌入协同训练。*AAAI* 人工智能会议论文集，34(03):3025–3032。

Xu Yang、Ziyi Huan、Yisong Zhai 和 Ting Lin。2021年。基于知识图谱的个性化推荐技术研究。*Applied Sciences*，11(15):7104。

Yuanming Zhang、Tianyu Gao、Jiawei Lu、Zhenbo Cheng 和 Gang Xiao。2021年。面向跨语言知识图谱的自适应实体对齐。载于国际知识科学、工程与管理会议，第474–487页。

Beibei Zhu、Tie Bao、Lu Liu、Jiayu Han、Junyi Wang 和 Tao Peng。2022年。基于关系感知和属性参与的跨语言知识图谱实体对齐。*AppliedIntelligence*。

Hao Zhu、Ruobing Xie、Zhiyuan Liu 和 Maosong Sun。2017年。基于知识嵌入的迭代实体对齐。发表于人工智能国际联合会议（*IJCAI*）论文集。

Qiannan Zhu、Xiaofei Zhou、Jia Wu、Jianlong Tan 和 Li Guo。2019年。面向多语言知识图谱的邻域感知注意力表示。载于第二十八届人工智能国际联合会议论文集，第 1943–1949 页，中国澳门。人工智能国际联合会议组织。

Renbo Zhu、Meng Ma 和 Ping Wang。2021a。Raga：面向全局实体对齐的关系感知图注意力网络。*arXiv:2103*。*00791 [cs]*。

Yao Zhu、Hongzhi Liu、Zhonghai Wu 和 Yingpeng Du。2021b。关系感知邻域匹配模型用于实体对齐。35(5):4749–4756。