·人工智能与模式识别·　　文章编号：1000-3428(2022)03-0074-07　　文献标志码：A　　中图分类号：TP391.1

基于双向对齐与属性信息的跨语言实体对齐

车　超，刘　迪

（大连大学 先进设计与智能计算省部共建教育部重点实验室，辽宁 大连 116622）

摘　要：实体对齐表示在不同的知识图谱中查找引用相同现实身份的实体。目前主流的基于图嵌入的实体对齐方法中的对齐实体通常具有相似的属性，有效利用属性信息可提升实体对齐效果，同时由于不同知识图谱之间的知识分布差异，仅考虑单个方向的对齐预测会导致预测结果出现偏差。针对上述问题，提出一种改进的跨语言实体对齐方法。利用融合属性信息的双向对齐图卷积网络模型，将前馈神经网络编码实体对应的属性信息与初始的实体嵌入相结合，得到联合属性信息的实体表示，并使用双向对齐机制实现跨语言的实体对齐预测。在3个跨语言数据集上的实验结果表明，该方法通过融合更多的知识图谱信息增强了实体表示能力，并且利用双向对齐机制缓解了数据分布差异问题，相比基于图嵌入的实体对齐方法整体性能更优。

关键词：实体对齐；知识图谱；属性信息；双向对齐；图卷积网络

开放科学（资源服务）标志码（OSID）：

中文引用格式：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.

英文引用格式：CHE C，LIU D. Cross-language entity alignment based on bidirectional alignment and attribute information[J].Computer Engineering，2022，48(3)：74-80.

Cross-language Entity Alignment Based on Bidirectional Alignment and Attribute Information

CHE Chao，LIU Di

（Key Laboratory of Advanced Design and Intelligent Computing，Ministry of Education，Dalian University，Dalian，Liaoning 116622，China）

【Abstract】Entity alignment is to find the entities that refer to the same real identity in different knowledge graphs.The alignment entities in most of the existing graph embedding-based entity alignment methods share similar attributes，which means utilizing attribute information can improve entity alignment performance. At the same time，due to the knowledge distribution differences between different knowledge graphs，alignment prediction that considers only a single direction will lead to deviation in alignment results. In response to the above problems，this paper proposes an improved cross-language entity alignment method. The method uses a Bidirectional alignment Graph Convolutional Network model with Attribution information（BiGCN-A），and combines the attribute information corresponding to the coded entity with the initial entity embedding through a feed-forward neural network to obtain an entity representation of joint attribute information.A bidirectional alignment mechanism is also used to realize cross-language entity alignment prediction. Experimental results on three cross-language datasets show that the proposed method enhances the entity representation ability by fusing more knowledge graph information.It also uses a bidirectional alignment mechanism to alleviate the problem of data distribution differences. Compared with the entity alignment method based on graph embedding，the proposed method displays better overall performance.

【Key words】entity alignment；knowledge graph；attribute information；bidirectional alignment；Graph Convolutional Network（GCN）

DOI：10. 19678/j. issn. 1000-3428. 0060540

0　概述

知识图谱将非结构化知识转化为结构化的三元组知识，广泛应用于机器阅读[1]、机器翻译[2]、推荐系统[3]、问答系统[4]等自然语言处理（Natural Language Processing，NLP）任务。随着知识图谱基础工程技术的完善和进步，人们已经建立了越来越多的单语言知识图谱，例如DBpedia[5]、YAGO[6-7]和BabelNet[8]，它

基金项目：国家自然科学基金面上项目（62076045）；辽宁省自然科学基金（2019-ZD-0569）。

作者简介：车　超（1981—），男，副教授、博士，主研方向为自然语言处理、数据挖掘；刘　迪，硕士研究生。

收稿日期：2021-01-11　　**修回日期：**2021-03-03　　**E-mail：**chechao101@163.com

们通常将现实世界的知识表示为一种特定结构的知识图谱。不同的知识图谱的创建目的不同，侧重点不同，通常会包含许多互补信息。将这些知识图谱整合在一起会大幅提高知识的利用效率，但是同一实体在不同的知识图谱中有不同的表现形式。如何在不同的知识图谱之间集成异构知识成为一个迫切需要解决的问题，实体对齐就是解决该问题的有效方法。

早期的实体对齐方法主要依赖于定义各种独立于语言的特征或者机器翻译技术来发现跨语言的连接。近年来，基于嵌入的实体对齐方法将知识图谱嵌入到低维向量空间中进行运算，显著提升了实体对齐效果。基于嵌入的实体对齐方法主要分为基于翻译模型[9]和基于图神经网络[10]两类。翻译模型利用头尾实体和关系在空间中的平移不变性计算知识图谱实体和关系的嵌入表示。在应用于实体对齐时，首先通过翻译模型学习实体和关系在不同知识图谱中的嵌入，然后利用已有的实体对齐种子作为纽带将它们对齐到统一的向量空间。该方法不仅可以保留知识图谱的结构，而且可以隐式地利用现有知识中缺失的连接补全知识图谱。基于图神经网络的实体对齐方法[11]利用图卷积网络（Graph Convolutional Network，GCN）[12]增强实体与其邻居信息的嵌入，可以更好地利用实体对齐种子来传播相似信息到整个图，仅需少量对齐种子便能达到较好的效果。为了实现对关系的编码，研究人员进行大量研究并取得了一系列重要成果。SCHLICHTKRULL等[13]提出关系图卷积网络（Relational Graph Convolutional Network，R-GCN）模型，该模型通过为每种关系分配一个权重矩阵来建模多关系图。WU等[14]提出高速门图卷积网络模型（HGCN-JE），该模型利用少量的对齐实体种子学习的实体嵌入来近似关系表示。通过近似关系表示和初步实体嵌入相结合得到实体联合表示，进一步训练模型，取得了不错的实体对齐效果。

除了关系信息以外，属性信息同样重要，知识图谱中存在大量属性信息，对实体对齐效果产生重要影响。同时，现有的大部分基于图神经网络的实体对齐方法并不重视预测阶段的处理，通常仅计算单个方向的实体相似度排行矩阵，利用单一矩阵进行对齐预测，导致预测结果出现偏差。针对以上问题，借鉴在HGCN-JE模型中联合生成实体和关系向量的思想，并加入属性信息与双向对齐机制，本文提出一种融合属性信息的双向对齐图卷积网络模型（Bidirectional alignment Graph Convolutional Network with Attribution information，BiGCN-A）进行实体对齐，将属性信息融入到实体对齐中，并且在对齐预测阶段进行双向实体对齐以获得更高的对齐准确率。

1　相关工作

1.1　图神经网络

近年来，由于图结构的强大表现力，利用机器学习方法分析图的研究越来越受到重视。图神经网络（Graph Neural Network，GNN）[10]是一类基于深度学习的图域信息处理方法，因较好的性能和可解释性，已成为一种被广泛应用的图分析方法。GCN是GNN的变体，是一种基于图操作的神经网络，它能高效地结合实体邻接节点信息，学习知识图谱的结构信息，对知识图谱进行编码。GCN对节点分类、关系抽取、语义角色标注等NLP问题均具有较好的应用效果。针对GCN无法编码关系信息的问题，研究人员进一步提出R-GCN，通过赋予每类关系一个权重矩阵编码关系信息，生成关系向量。图注意力（Graph Attention，GAT）网络[15]通过使用注意力机制对邻近节点特征加权求和，邻近节点特征的权重完全取决于节点特征，独立于图结构，在节点分类问题上取得了较好的效果。

1.2　跨语言实体对齐

一些研究人员使用字符串相似性作为主要对齐方法，例如NGOMO等[16]使用三角不等式来计算实体相似性的近似值，通过计算相似度高的实体对的实际相似度，返回实际字符串相似度最高的实体对。随着知识表示学习技术的发展，众多翻译模型被应用于实体对齐。由于TransE的简单性和有效性，因此大量的实体对齐工作使用TransE模型完成。联合嵌入方法（JE）[17]将翻译模型应用在实体对齐中，通过学习不同知识图谱在统一向量空间中的嵌入，在该空间中执行实体对齐。多语言知识图谱嵌入方法（MTransE）[18]将两个知识图谱嵌入到独立的低维向量空间，通过对齐实体种子产生映射矩阵实现实体对齐。联合属性保持嵌入方法（JAPE）[19]将结构嵌入和属性嵌入相结合，匹配不同知识图谱中的实体，结构嵌入使用TransE模型，属性嵌入使用Skip-gram[20]模型。自举法（BootEA）[21]通过迭代增加实体对齐种子方法学习知识图谱的嵌入。多视图嵌入法（MultiKE）[22]将单个知识图谱分成名称、属性、关系3个视图，分别训练实体向量并将3个视图的实体向量相结合进行对齐。多映射关系法（MMR）[23]提出一种新的知识表示方法，通过重新定义能量函数弥补了TransE在编码复杂关系问题上的劣势，提高了实体对齐性能。虽然基于TransE的实体对齐方法在三元组层面的表示上具有不错的效果，但是全局结构表示不理想。随着图神经网络的发展，研究的主要方向转到利用图卷积网络进行实体对齐。图卷积法（GCN-Align）[11]通过图卷积网络编码实体和属性进行实体对齐。对偶关系图卷积法（RDGCN）[24]通过构建一个对偶关系图，与原始知识图谱之间相互交互，使编码关系信息进入实体。门控多阶邻居信息法（Alinet）[25]使用图卷积网络结合实体的一阶邻域，利用图注意力网络结合二阶邻域使实体的嵌入更有表达力，从而提升对齐效果。混合多角度信息法（HMAN）[26]通过多语言BERT模型

计算实体的描述信息相似度并将其与实体的结构嵌入相结合，在对齐阶段取得了不错的效果。上述方法均采用了图卷积网络对知识图谱进行编码，为本文方法提供了可参考的思路，因此本文方法在HGCN-JE的基础上融入了属性信息。

2 融合属性信息的双向对齐图卷积网络模型

在多语言知识图谱G中，使用L表示G所包含的语言的集合，使用$G_L=\{E_i,R_i,A_i,V_i\}$表示特定语言的知识图谱，其中，E_i、R_i、A_i、V_i分别表示实体、关系、属性、属性值。该知识图谱由关系三元组(h_i,r_i,t_i)和属性三元组(h_i,a_i,v_i)组成，其中，h_i，$t_i\in E_i$，$r_i\in R_i$，$a_i\in A_i$和$v_i\in V_i$。给定用源语言L_1和目标语言L_2表示的两个知识图谱G_1和G_2，存在一组预先对齐的实体集合$\mathbb{L}=\{(e,u)|e\in E_1,u\in E_2\}$，将其作为训练数据训练模型。跨语言实体对齐的任务是利用现有的对齐实体种子对模型进行训练，自动发现剩余的对齐实体对。

BiGCN-A模型整体框架如图1所示。给定知识图谱G_1和G_2，实体对齐种子$\mathbb{L}$，BiGCN-A模型通过具有高速网络机制[27]的GCN对知识图谱进行编码得到实体嵌入，利用高速网络和全连接网络得到属性嵌入，将实体和属性嵌入融合实现知识图谱的预对齐，然后采用预对齐稳定后的模型训练出的实体表示近似表示关系，通过将关系表示和实体表示结合生成联合实体表示，进一步使用多层GCN迭代集成邻居信息，以获得更好的实体和关系表示，最终通过双向对齐方法进行实体对齐预测。

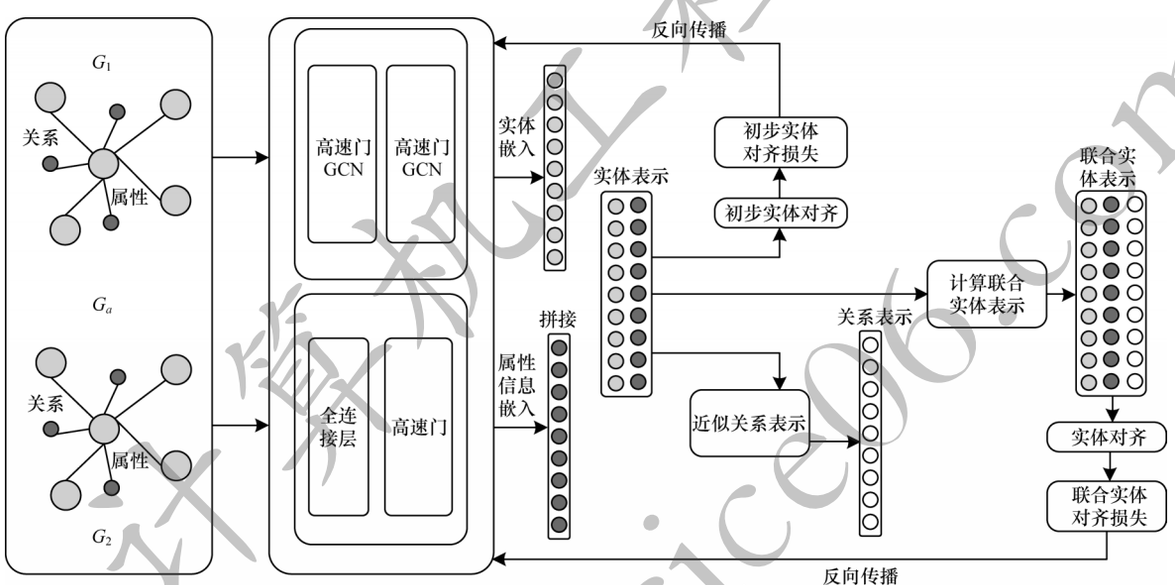

图1 BiGCN-A模型的整体框架

Fig.1 Overall framework of the BiGCN-A model

2.1 融入属性信息的初步实体对齐

如图1所示，将G_1和G_2放在图$G_a=(E_a,R_a,A_a,V_a)$中构成模型的输入。利用现有的对齐实体种子训练模型，采用训练出的稳定模型发现更多潜在的对齐实体，完成初步的实体对齐工作。使用两层GCN获取实体的嵌入表示，使其能够更好地结合邻居实体信息。根据Alinet[25]得出的结论，实体的直接邻居与远距离邻居相比异构性更小，因此不需要基于注意力的邻域聚集来选择相关的邻居实体。GCN层的输入是实体特征矩阵$\boldsymbol{H}$，l层的GCN将特征表示$\boldsymbol{H}^{(l)}$作为输入，输出$\boldsymbol{H}^{(l+1)}$可表示如下：

$$\boldsymbol{H}^{(l+1)}=\phi\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right) \tag{1}$$

其中：$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{i}$是邻接矩阵，$\boldsymbol{i}$是单位矩阵；$\tilde{\boldsymbol{D}}$是$\tilde{\boldsymbol{A}}$的对角节点度矩阵；$\phi(\cdot)$是ReLU函数；$\boldsymbol{W}^{(l)}$表示第$l$层中的可学习参数，将$\boldsymbol{H}^{(0)}$作为初始输入。

为控制跨层积累的噪声并保存从交互中学习到的有用的关系信息，按照RAHIMI等[28]提出的方法，在GCN层之间引入高速网络机制，具体如下：

$$\boldsymbol{T}\left(\boldsymbol{H}^{(l)}\right)=\sigma\left(\boldsymbol{H}^{(l)}\boldsymbol{W}_T^{(l)}+\boldsymbol{b}_T^{(l)}\right) \tag{2}$$

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

式(1)为基本的GCN网络层结构。为了融入属性信息，将实体属性作为词袋特征进行显式建模。类似于One-Hot向量，构造基于计数的N-Hot向量$\boldsymbol{X}_a$，(i,j)项表示实体E_i的第j个属性的计数。值得注意的是，仅考虑最频繁的前F个属性，以避免数据稀疏问题。因此，对于每个实体，其属性特征都是F维向量，受到HMAN[26]的启发，如果通过图结构传播属性信息，邻居实体属性的传播会引入噪声，仅关注当前实体的属性效果更好。因此，通过一个前向神经网络获取相关属性信息的嵌入，同时在前向神经网络中加入高速网络机制，最终前馈神经网络定义如下：

$$\boldsymbol{S}_a=\phi\left(\boldsymbol{W}_a^{(1)}\boldsymbol{X}_a+\boldsymbol{b}_a^{(1)}\right) \tag{4}$$

$$\boldsymbol{T}_a=\sigma\left(\boldsymbol{W}_a^{t}\boldsymbol{S}_a+\boldsymbol{b}_a^{t}\right) \tag{5}$$

$$\boldsymbol{H}_{\mathrm{f}}=\phi\left(\boldsymbol{W}_a^{(2)}\boldsymbol{S}_a+\boldsymbol{b}_a^{(2)}\right)\cdot\boldsymbol{T}_a+\boldsymbol{S}_a\cdot\left(1-\boldsymbol{T}_a\right) \tag{6}$$

其中：$\boldsymbol{X}_a$对应于原始属性特征；$\boldsymbol{W}_a^{(1)}$、$\boldsymbol{W}_a^{(2)}$、$\boldsymbol{W}_a^{t}$、$\boldsymbol{b}_a^{(1)}$、$\boldsymbol{b}_a^{(2)}$、$\boldsymbol{b}_a^{t}$分别表示训练属性信息的模型参数；$\phi(\cdot)$是ReLU函数；$\sigma(\cdot)$是Sigmoid函数。至此，便获得初步的实体表示$\boldsymbol{H}=\boldsymbol{H}^{(l)}\oplus\boldsymbol{H}_a$，$\oplus$表示拼接操作。

训练阶段的目标是将跨语言实体嵌入到相同的低维向量空间中，在该空间中等价实体嵌入距离要尽量相近，非等价实体的嵌入距离要尽量远。给定两个知识图谱G_1和G_2，以及一组预先对齐的实体对$\mathbb{L}(G_1,G_2)$作为训练数据，模型使用基于边缘排名损失函数进行训练，定义如下：

$$L=\sum_{(\boldsymbol{p},\boldsymbol{q})\in\mathbb{L}}\sum_{(\boldsymbol{p}',\boldsymbol{q}')\in\mathbb{L}'}[d(\boldsymbol{p},\boldsymbol{q})+\gamma-d(\boldsymbol{p}',\boldsymbol{q}')]_+ \tag{7}$$

其中：$\mathbb{L}$表示种子对齐对；$\mathbb{L}'$表示由最近邻采样产生的$\mathbb{L}$的负例集合；$d(\boldsymbol{p},\boldsymbol{q})$表示$p,q$之间的曼哈顿距离。因为一个实体在另一个知识图谱中只能有一个对应的实体，最佳的负例实体是与目标实体最接近的实体。通过$d(\cdot)$计算两实体之间的距离得到距离最小的实体。给定预先对准的实体对$(p,q)\in\mathbb{L}$，其中，$p\in E_1$，$q\in E_2$，K是负样本的数目，选择E_2中最接近q的K个实体作为负例，反之亦然。至此，待模型稳定，便完成了融入属性的初步实体对齐工作。

2.2　联合实体关系对齐

因为无法通过GCN直接得到关系向量，所以通过上节得到的实体嵌入近似来表示关系嵌入，用于构建联合实体表示向量。通过观察发现，一个关系连接的头实体和尾实体的统计信息能够在一定程度上反映关系的浅层语义信息，因此可以通过聚合实体表示近似得到关系表示。给定一个关系$r\in R_a$，存在关系均为r的三元组集合，$T_r=(h_r,r,t_r)$，首先将关系r对应的头实体集向量和尾实体集向量分别求平均值，将得到的平均头尾实体向量进行拼接，之后引入一个矩阵$\boldsymbol{W}$对拼接后的向量进行一次线性变换得到关系的表示。因为实体和关系在知识图谱中密不可分，实体中会包含关系的语义信息，同时关系中也会包含实体的语义信息，并且具有对齐性质的实体通常具有相似的关系，具有相似关系的实体对齐可能性更大，所以利用先前得到的实体和关系的嵌入表示生成新的联合嵌入，用于进一步训练模型。具体而言，首先训练融入属性的GCN-A模型，对于每个实体，通过一个实体关系的邻接矩阵提供的信息，计算出一个实体的关系上下文向量，然后将这个关系上下文向量与预训练得到的实体向量进行拼接，形成实体的联合表示向量。

2.3　双向对齐

将实体对齐预测问题看作排序问题，当G_1的实体与G_2中的实体对齐时，需要计算$e_1\in G_1$与所有的$e_j\in G_2$之间的相似度，得到相似度矩阵$\boldsymbol{D}_s$。通过相似度大小进行排序，得到排行矩阵$\boldsymbol{D}_r$，$\boldsymbol{D}_r$中的每一行代表了实体对齐的排序结果。不同知识图谱之间知识分布的差异会影响对齐预测的准确性。具体而言，对于$e_1\in G_1$能够得到一个排名第一的$e_j\in G_2$作为预测结果，虽然表面上看似合理，因为e_j在e_1对G_2所有实体的相似度矩阵中的相似度分数最高且与e_1的距离最近，但在e_j对G_1的所有实体的相似度矩阵中e_1的相似度得分不一定最高，在其相似度排名中甚至会排在几十位，在这种情况下e_j并不是e_1的最优选择，反而产生了误差。因此，对齐预测过程需要考虑两个方向，而以往多数研究仅考虑一个方向。本文通过计算得到两个方向的排行矩阵来解决该问题，从G_1和G_2两个方面出发，分别得到G_1对应于G_2和G_2对应于G_1的两个方向的排行矩阵$\boldsymbol{D}_{r1}$、$\boldsymbol{D}_{r2}$，将排序矩阵重定义为$\boldsymbol{D}_r=\boldsymbol{D}_{r1}+\boldsymbol{D}_{r2}^{\mathrm{T}}$并将其作为最终的相似度排行矩阵，其中$\boldsymbol{D}_{r2}^{\mathrm{T}}$表示$G_2$排序矩阵的转置。

相似度排行矩阵实例如图2所示，其中，图2(a)为法语-英语方向相似度排行rank1，图2(a)为英语-法语方向相似度排行rank2，数字表示排名，数字越小排名越靠前，实体名称均来自DBP FR-EN数据集。法语实体Pi.l.du.Br的对齐实体对应于英语实体Pe.I.Br。从法语-英语方向看，与法语实体Pi.l.du.Br相似度最高的是英语的Pe.I.Br，但是对于英语-法语方向的相似度排行而言，Pe.I.Br对应的英语Em.du.Br，Pi.l.du.Br排名为3，在对齐预测时如果只考虑一个方向的相似度排行矩阵错误可能性就会增大，不能够正确预测出对齐实体对Pi.l.du.Br和Pe.I.Br。综合两个相似度排行矩阵，将rank2的相似度矩阵进行转置，与rank1的相似度矩阵相加得到最终的排名，Pi.l.du.Br和Pe.I.Br的最终排名为2，Pe.I.Br在所有排名中最靠后(Pi.l.du.Br对于英语实体的相似度排名分别为2、8、4、3、4)，从而正确预测出对齐实体。

法语-英语	Pe.I.Br	Brazil	Pe.II.Br	Em.of.Br	Pr.Im.l.Br
P.l.du.Br	0	4	3	1	2
Brésil	4	0	1	2	3
Pi.II.du.Br	2	0	1	3	4
Em.du.Br	1	2	4	0	3
Is.du.Br	4	3	1	2	0

(a)法语-英语

英语-法语	Pi.l.du.Br	Brésil	Pi.II.du.Br	Em.du.Br	Is.du.Br
Pe.I.Br	2	3	1	0	4
Brazil	4	0	1	2	3
Pe.II.Br	1	2	0	4	3
Em.of.Br	2	4	1	0	3
Pr.Im.l.Br	2	3	1	3	0

(b)英语-法语

图2　相似度排行矩阵实例

Fig.2　Examples of similarity ranking matrixes

3　实验

3.1　数据集与实验设置

采用DBP15K数据集进行测试，DBP15K数据集包含DBP ZH-EN(汉语-英语)、DBP FR-EN(法语-英

语)和DBP JA-EN(日语-英语)3个跨语言的真实世界数据集,3个数据集的统计信息如表1所示,其中每一个数据集都是通过抽取DBpedia多语言版本的15 000个对齐实体链接构建的。为了方便和之前的工作[14,19]进行对比,使用30%的预对齐实体对作为训练数据,70%用于测试,使用Hits@k作为评价指标,即通过计算排名在相似度排名列表前k个中正确对齐的实体的比例来得到Hits@k分值。

表1 DBP15K数据集

Table 1 DBP15K dataset

数据集	语言	实体数	关系数	属性数	关系元组数	属性元组数
DBP ZH-EN	汉语	66 469	2 830	8 113	153 929	379 684
	英语	98 125	2 317	7 173	237 674	567 755
DBP JA-EN	日语	65 744	2 043	5 882	164 373	354 619
	英语	95 680	2 096	6 066	233 319	497 230
DBP FR-EN	法语	66 858	1 379	4 547	192 191	528 665
	英语	105 889	2 209	6 422	278 590	576 543

实验设置阈值γ为1、学习率为0.005。每经过50代对负例实体进行采样,其中k=125。实验基于两层GCN,并使用前1 000个最频繁的属性(即F=1 000)来构建N-Hot特征向量。为更好地进行模型初始化,在不同的知识图谱中使用实体名称。通过谷歌翻译器将非英文实体翻译为英文实体,并利用预训练好的词向量得到初始化表示。需要注意的是,因为人名、地名等词汇对应不同语言,然而不同语言的表示风格不同,所以可能会导致谷歌翻译的结果出现部分错误。

3.2 实验结果与分析

3.2.1 与图嵌入方法的性能对比

实验选取GCN-Align[11]、HGCN-JE[14]、JE[17]、MTransE[18]、JAPE[19]、HMAN[26]等6种主流的图嵌入方法与本文BiGCN-A模型进行比较,实验结果如表2所示,其中:JE、MTransE和JAPE是基于翻译模型进行实体对齐;GCN-Align、HGCN-JE和HMAN是基于GCN进行实体对齐,均属于跨语言实体对齐的SOTA方法,GCN-Align方法使用属性信息,通过GCN将属性信息与结构信息进行聚合,由于BERT模型对于知识图谱嵌入方面效果不好,因此实验未涉及与基于BERT的实体对齐方法的比较。

表2 与其他图嵌入方法的实体对齐结果对比

Table 2 Comparison of entity alignment results with other graph embedding methods %

方法与模型	DBP ZH-EN数据集		DBP JA-EN数据集		DBP FR-EN数据集	
	Hits@1	Hits@10	Hits@1	Hits@10	Hits@1	Hits@10
JE	21.27	42.77	39.97	39.97	15.38	38.84
MTransE	30.83	61.41	27.86	57.45	24.41	55.55
JAPE	41.18	74.46	36.25	68.50	32.39	66.68
GCN-Align	41.25	74.38	39.91	74.46	37.29	74.49
HMAN	56.20	85.10	56.70	86.90	54.00	87.10
HGCN-JE	71.86	85.37	75.99	88.92	89.21	96.18
BiGCN-A	76.10	88.97	78.61	91.72	90.25	96.47

在表2中,HMAN方法中的结果保留小数点后1位,为了更好对比,用0补全到小数点后2位。由表2可以看出:1)通过捕获丰富的相邻结构信息,基于GCN的实体对齐方法在Hits@1上的性能优于基于翻译的实体对齐方法,在Hits@10上的性能优于MTransE和JE方法;2)HMAN使用知识图谱中实体的描述信息,相较其他未使用实体名称嵌入的向量作为实体初始化嵌入的方法,在所有数据集上都取得了最优结果;3)HGCN-JE方法因为使用高速网络的GCN并融入关系信息,同时利用实体名称嵌入的向量作为实体初始化嵌入,在所有数据集上的效果明显优于HMAN方法;4)BiGCN-A模型因为使用了实体初始化嵌入以及属性信息并在对齐预测阶段使用双向对齐机制,所以在所有数据集上的效果均达到最优,特别是在DBP ZH-EN数据集上Hit@1比HGCN-JE提升了4.24个百分点;5)BiGCN-A模型在DBP ZH-EN数据集和DBP JA-EN数据集上有大幅的性能提升,即使在效果已经非常好的DBP FR-EN数据集上仍有小幅的性能提升,这充分验证了其有效性。

3.2.2 消融实验

为验证属性信息和双向对齐机制的有效性,将BiGCN-A模型与只使用属性信息的GCN-A模型和只使用双向对齐的BiGCN模型进行对比,实验结果如表3所示。由表3可以看出,相较GCN-A模型和BiGCN模型,除了DBP FR-EN数据集之外,BiGCN-A模型均达到了最优的效果,这证明了属性信息结合双向对齐机制的有效性。与不使用属性信息的BiGCN模型相比,BiGCN-A模型在DBP ZH-EN、DBP JA-EN数据集上效果均有所提升,这表明添加属性信息是非常有效的。因为相似的实体倾向于拥有相似的属性,增加了属性信息,丰富了实体嵌入的要素,效果自然会有提升。但是观察到在DBP FR-EN数据集上增加属性信息效果会略微下降,这是因为输入实体嵌入的初始化是先通过谷歌翻译器得到英文实体,再使用训练好的词向量对实体特征初始化,而法英语言比较接近,翻译错误率小,得到的实体的初始化特征好,当拼接融入属性的实体特征时,反而使得相似度下降,导致结果略微下降。因此,只使用属性信息的GCN-A模型效果劣于只使用双向对齐机制的BiGCN模型,更劣于结合属性信息和双向对齐机制的BiGCN-A模型。实体对齐仅考虑一个方向会忽略实体分布的差异,对实体对齐结果造成误导,而两个方向的相似度排行相互叠加可以中和实体分布差异,减少对实体对齐的影响。

表3 基于属性信息和双向对齐的实体对齐结果对比

Table 3 Comparison of entity alignment results based on attribute information and bidirectional alignment %

模型	DBP ZH-EN数据集		DBP JA-EN数据集		DBP FR-EN数据集	
	Hits@1	Hits@10	Hits@1	Hits@10	Hits@1	Hits@10
GCN-A	72.70	86.91	76.76	89.75	88.11	95.37
BiGCN	75.23	88.23	79.30	91.28	91.10	97.09
BiGCN-A	76.10	88.97	78.61	91.72	90.25	96.47

3.2.3　对齐种子比率敏感度分析

为探究对齐种子比率对实体对齐效果的影响，分别按照10%、20%、30%、40%、50%的对齐种子比率划分训练集，并与不同对齐种子比率的JAPE、GCN-Align方法进行对比，结果如图3所示。由图3可以看出，BiGCN-A模型在不同的对齐种子比率和数据集下的表现均远优于JAPE与GCN-Align方法，在仅有10%的对齐种子比率作为训练集时Hits@1仍能达到67.99%（DBP ZH-EN数据集）、74.73%（DBP JA-EN数据集）、87.56%（DBP FR-EN数据集），远优于另外两种方法在有50%的对齐种子比率作为训练集时的结果。可见，BiGCN-A模型对于对齐种子比率的变化不敏感，具有较强的鲁棒性。

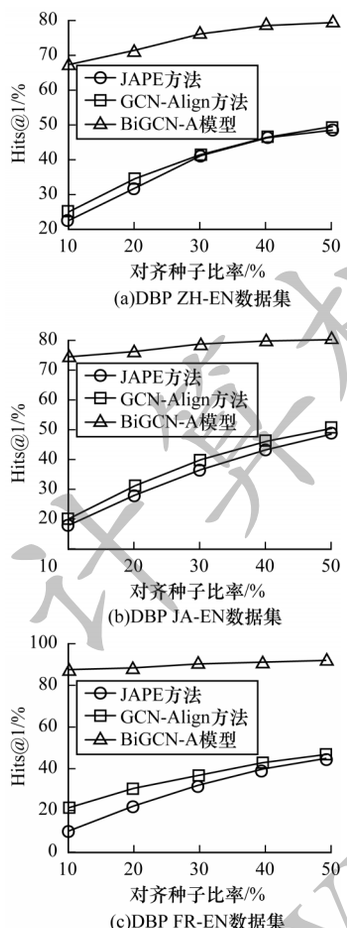

图3　不同对齐种子比率对实体对齐效果的影响

Fig.3　Effect of different alignment seed ratios on entity alignment effect

4　结束语

本文提出一种基于BiGCN-A模型的跨语言实体对齐方法，通过实体属性的相似性提高实体对齐的准确率，利用双向对齐机制求得两个方向的相似度排行矩阵并进行融合，得到最终的相似度排行矩阵，实现跨语言实体对齐的预测。在DBP15K数据集上的实验结果表明，基于BiGCN-A模型的实体对齐方法整体性能优于目前主流的基于图嵌入的实体对齐方法。后续将尝试引入知识图谱中的实体描述等信息来进一步提高实体对齐的准确率。另外，BiGCN-A模型在初始化向量时通过谷歌翻译器得到实体的英文表示，其中可能存在一些翻译错误，这也是下一步工作的重点方向。

参考文献

[1] YANG B S, MITCHELL T. Leveraging knowledge bases in LSTMs for improving machine reading[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2017: 1436-1446.

[2] MOUSSALLEM D, WAUER M, NGOMO A C N. Machine translation using semantic web technologies: a survey[J]. Journal of Web Semantics, 2018, 51: 1-19.

[3] ZHANG F Z, YUAN N J, LIAN D F, et al. Collaborative knowledge base embedding for recommender systems[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2016: 353-362.

[4] ZHANG Y Y, DAI H J, KOZAREVA Z, et al. Variational reasoning for question answering with knowledge graph [EB/OL]. [2020-12-07]. http://arxiv.org/abs/1709.04071v1.

[5] BIZER C, LEHMANN J, KOBILAROV G, et al. DBpedia—a crystallization point for the Web of data[J]. Journal of Web Semantics, 2009, 7(3): 154-165.

[6] SUCHANEK F M, KASNECI G, WEIKUM G. YAGO: a large ontology from Wikipedia and WordNet[J]. Journal of Web Semantics, 2008, 6(3): 203-217.

[7] REBELE T, SUCHANEK F, HOFFART J, et al. YAGO: a multilingual knowledge base from Wikipedia, WordNet, and GeoNames[C]//Proceedings of International Semantic Web Conference. Berlin, Germany: Springer, 2016: 177-185.

[8] NAVIGLI R, PONZETTO S P. BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network[J]. Artificial Intelligence, 2012, 193: 217-250.

[9] BORDES A, USUNIER N, GARCIADURAN A, et al. Translating embeddings for modeling multi-relational data[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2013: 2787-2795.

[10] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2009, 20(1): 61-80.

[11] WANG Z C, LÜ Q, LAN X H, et al. Cross-lingual knowledge graph alignment via graph convolutional networks[C]// Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2018: 349-357.

[12] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. [2020-12-07]. https://arxiv.org/abs/1609.02907.

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

[14] WU Y T, LIU X, FENG Y S, et al. Jointly learning entity

and relation representations for entity alignment[C]// Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2019: 240-249.
[15] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL]. [2020-12-07]. http://arxiv. org/pdf/1710. 10903.
[16] NGOMO A N, AUER S. LIMES: a time-efficient approach for large-scale link discovery on the Web of data[C]// Proceedings of International Joint Conference on Artificial Intelligence. New York, USA: ACM Press, 2011: 2312-2317.
[17] HAO Y C, ZHANG Y Z, HE S Z, et al. A joint embedding method for entity alignment of knowledge bases[M]. Berlin, Germany: Springer, 2016: 3-14.
[18] CHEN M H, TIAN Y T, YANG M H, et al. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment[C]//Proceedings of International Joint Conference on Artificial Intelligence. Berlin, Germany: Springer, 2017: 1511-1517.
[19] SUN Z Q, HU W, LI C K. Cross-lingual entity alignment via joint attribute-preserving embedding[C]//Proceedings of International Semantic Web Conference. Berlin, Germany: Springer, 2017: 628-644.
[20] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2020-12-07]. http://arxiv. org/abs/1301. 3781.
[21] SUN Z Q, HU W, ZHANG Q H, et al. Bootstrapping entity alignment with knowledge graph embedding [C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. Berlin, Germany: Springer, 2018: 4396-4402.
[22] ZHANG Q H, SUN Z Q, HU W, et al. Multi-view knowledge graph embedding for entity alignment[EB/OL]. [2020-12-07]. http://arxiv. org/abs/1906. 02390v1.
[23] SHI X F, XIAO Y H. Modeling multi-mapping relations for precise cross-lingual entity alignment[C]//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2019: 813-822.
[24] WU Y T, LIU X, FENG Y S, et al. Relation-aware entity alignment for heterogeneous knowledge graphs[EB/OL]. [2020-12-07]. https://arxiv. org/abs/1908. 08210.
[25] SUN Z Q, WANG C M, HU W, et al. Knowledge graph alignment network with gated multi-hop neighborhood aggregation[C]//Proceedings of 2020 AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2020: 222-229.
[26] YANG H W, ZOU Y Y, SHI P, et al. Aligning cross-lingual entities with multi-aspect information[C]//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2019: 4430-4440.
[27] SRIVASTAVA R K, GREFF K, SCHMIDHUBER J. Highway networks[EB/OL]. [2020-12-07]. https://zhuanlan. zhihu. com/p/38130339.
[28] RAHIMI A, COHN T, BALDWIN T. Semi-supervised user geolocation via graph convolutional networks[EB/OL]. [2020-12-07]. https://arxiv. org/abs/1804. 08049.

编辑 陆燕菲