·人工智能与模式识别·　　文章编号：1000-3428(2022)06-0065-08　　文献标志码：A　　中图分类号：TP391

基于邻域聚合的实体对齐方法

谭元珍，李晓楠，李冠宇

（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘　要：实体对齐旨在判断来自不同知识图谱的实体是否为指向真实世界的同一个对象。然而，知识图谱间的结构异质性往往会影响实体对齐的准确性。提出一种基于邻域聚合匹配网络（NAMN）模型的实体对齐方法。根据每跳邻居对中心实体重要性不同的特点，采用分层的思想区别处理每跳邻域信息，通过门控机制进行聚合以学习图结构的表征。在此基础上，为每个实体构建邻域局部子图进行跨图邻域匹配，并将匹配阶段的输出与通过门控机制所学习到的图结构表征进行联合编码，生成最终面向匹配的表征。采用DBP15K数据集进行实验，结果显示，Hits@1的所有值均在75%以上，Hits@10的所有值均在85%以上，最高可达到97%，平均倒数排名均高于80%，表明NAMN模型能够有效提高实体的匹配准确度。

关键词：实体对齐；知识图谱；门控邻域聚合；邻域匹配；对齐预测

开放科学（资源服务）标志码（OSID）：

中文引用格式：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.

英文引用格式：TAN Y Z，LI X N，LI G Y.Entity alignment method based on neighborhood aggregation[J].Computer Engineering，2022，48(6)：65-72.

Entity Alignment Method Based on Neighborhood Aggregation

TAN Yuanzhen，LI Xiaonan，LI Guanyu

（College of Information Science and Technology，Dalian Maritime University，Dalian，Liaoning 116026，China）

【Abstract】Entity Alignment（EA）aims to judge whether entities from different Knowledge Graph（KG）are the same object pointing to the real world.However，the structural heterogeneity between KG often affects the accuracy of EA.Hence，an EA method based on a Neighborhood Aggregation Matching Network（NAMN）model is proposed.Based on the different importance of each hop neighbor to the central entity，a hierarchical idea is applied to process the neighborhood information of each hop differently，and the gating mechanism is used to perform aggregation to learn the representation of a graph structure.Subsequently，a neighborhood local subgraph is constructed for each entity for cross graph neighborhood matching，and the output of the matching stage is jointly encoded with the graph structure representation learned through the gating mechanism to generate the final matching oriented representation.The experiment is performed using the DBP15K dataset.The experimental results show that all values of Hits@1 exceed 75%，all values of Hits@10 are between 85% and 97%，and the Mean Reciprocal Rank（MMR）exceeds 80%，indicating that the NAMN model can effectively improve the matching accuracy of entities.

【Key words】Entity Alignment（EA）；Knowledge Graph（KG）；gated neighborhood aggregation；neighborhood matching；alignment prediction

DOI：10.19678/j.issn.1000-3428.0061589

0　概述

随着智能信息服务应用的不断发展，知识图谱（Knowledge Graph，KG）已被广泛应用于智能问答[1]、智能信息处理[2-3]、个性化推荐[4]等领域。近年来，越来越多的知识图谱被构造以提供针对不同领域的知识，如DBpedia[5]、YAGO[6]、ConceptNet[7]、NELL[8]。研究人员发现，这些知识图谱通常不完整，相互之间包含着互补的事实，需要将不同的知识图谱整合到统一的知识图谱中，为不同的应用提供结构知识。然而，将来自不同知识图谱的实体链接到相同的真实世界知识并非易事，因为不同的知识

基金项目：国家自然科学基金（61976032，62002039）。

作者简介：谭元珍（1996—），女，硕士研究生，主研方向为智能信息处理；李晓楠，博士研究生；李冠宇（通信作者），教授、博士。

收稿日期：2021-05-10　　**修回日期：**2021-07-13　　**E-mail：**rabitlee@163.com

图谱基于不同的数据源所构造，所以同一实体在不同的知识库中也有着不同的表述。

在多语言知识图谱中查找等效实体，对于集成多源知识图谱起到重要的作用。实体对齐（Entity Alignment，EA）旨在从来自多个来源构成的知识图谱中找到表示真实世界知识的同一实体。目前，比较流行的实体对齐方法是基于知识图谱嵌入的方法，此方法主要是利用知识图谱的表示学习，克服了依靠人工创建规则或特征[9]的问题。该类方法假定基于不同数据源构造的知识图谱具有相似的结构，在向量空间中具有相对相似位置的实体为对齐实体，使用TransE等一系列模型[10-15]表示每个知识图谱中的实体和关系，然后将预对齐的实体投影至统一的向量空间。然而，基于知识图谱嵌入的实体对齐需要足够数量的种子序列，并且受不同知识图谱间不完整性和异质性的影响，对齐精确度往往不高。

图神经网络（Graph Neural Network，GNN）是学习图结构化数据的矢量表示和解决图上各种监督预测问题的强大模型[16-18]。GNN遵循递归邻域聚合方法，每个节点聚合其邻居的特征向量以计算新的特征向量[16,18]。在聚合k次迭代之后，节点由其变换后的特征向量表示，该特征向量可以捕获节点多跳邻居附近的结构信息，然后通过合并来获得整个图的表示[19]。文献[20]证明，GNN在识别同构子图方面具有与Weisfeiler-Leman（WL）检测[21]相同的表达能力。相似实体通常具有相似的邻域，这是GNN实现不同知识图谱之间实体对齐的理论基础。

然而，现有基于GNN的实体对齐模型依然面临着一个关键问题：由于不同的知识图谱具有结构异质性[22]，因此对应实体通常具有不同的邻域结构。解决此问题的关键在于要减小不同知识图谱实体邻域结构的异质性。本文提出一种邻域聚合匹配网络（Neighborhood Aggregation Matching Network，NAMN）模型，旨在从实体邻域角度对图结构信息进行编码以实现实体对齐，缓解结构异质性带来的影响。

1 相关工作

1.1 基于知识图谱嵌入的实体对齐

近年来，知识图谱嵌入学习已成功应用于实体对齐领域。当前的处理方法是将不同的知识图谱表示为嵌入，投影至同一向量空间，然后通过测量嵌入之间的相似性来进行实体对齐。MTransE[10]是基于嵌入的多语言实体对齐模型，其使用TransE模型学习两个知识图谱中实体的嵌入，然后学习连接两个嵌入空间之间的映射函数，以实现实体对齐。IPTransE[11]和BootEA[12]是通过联合嵌入进行迭代的自我训练方法，其使用预对齐的实体种子对来进行计算，并将迭代过程中新发现的实体对添加到训练数据集中，优化对齐效果。JAPE[13]使用Skip-Gram方法，利用种子对齐方式将两个知识图谱的实体嵌入到统一的向量空间中，将结构嵌入和属性嵌入结合在一起找到相似实体。文献[14]提出了一种RSN方法，结合递归神经网络（Recurrent Neural Network，RNN）和残差学习，以有效地捕获知识图谱内部和知识图谱之间的长期关系依赖性，优化实体对齐效果。MultiKE[15]分别从名称、属性和结构视图中学习实体的表示形式，集成3个特定的视图嵌入组合策略以提高实体对齐性能，并使用预先训练好的词嵌入来完善属性值的学习。但是，以上方法需要足够数量的种子对，成本较高，并且不同知识图谱的结构异质性对知识图谱的嵌入质量也产生了很大的负面影响，导致对齐效果变差。

1.2 基于GNN的实体对齐

与上述基于知识图谱嵌入的方法不同，图神经网络（GNN）使用图结构和节点特征来学习节点或整个图的表示向量，遵循邻域聚合策略，在图学习方面取得了显著进步。因此，一些工作试图将GNN应用在实体对齐方面以取得更好的对齐性能。GCN-Align[16]是一种基于GCN的实体对齐模型，其利用GCN将每个知识图谱的实体嵌入统一的向量空间，传播来自邻居的信息，通过结构知识进行实体对齐。然而，GCN-Align在训练过程中仅考虑实体之间的等效关系，没有在知识图谱中使用丰富的关系来区分共享邻居的实体。R-GCN[17]模型考虑到节点之间的关系，解决了GCN处理图结构中关系对节点的影响，其通过为每个关系设置转换矩阵来进一步合并关系类型信息，提高对齐效果。RDGCN[18]是一种新的关系感知双图卷积网络，其通过构建用于嵌入学习的对偶关系图，使用门控机制捕获邻域结构，缓解知识图之间的异构性，以学习更好的实体表示。R-GCN模型和RDGCN模型将预先对齐的实体和关系作为训练数据，这可能会导致昂贵的开销。AliNet[23]模型将门控机制和注意力机制结合在一起，以聚合多跳邻域来整合GCN，从而减少图异构性对实体对齐的影响，达到更好的对齐效果。然而，AliNet在汇总信息时将实体的所有一跳邻居同等对待，在没有仔细选择的情况下引入了噪声，影响了实体对齐性能。

2 NAMN模型设计

2.1 基础知识

2.1.1 知识图谱的实体对齐

本文将知识图谱形式表示为$G_i=(E_i,R_i,T_i)$，其中，E_i、R_i、T_i分别表示为G_i中实体、关系和三元组的

集合。$N_e=\{e'|(e,r,e')\in T\}\cup\{e'|(e',r,e)\in T\}$是$G_i$中实体$e$的邻居集。对齐的实体对形式化表示为$A=\{(e_1,e_2)\in E_1\times E_2|e_1\leftrightarrow e_2\}$，其中，$\leftrightarrow$表示等价关系，即$e_1$和$e_2$所表示的为真实世界中相同的实体。实体对齐的任务就是找到G_1和G_2之间的等效实体对。为方便起见，本文将G_1和G_2放到一个大图中，即$G=G_1+G_2$，$R=R_1\cup R_2$，$T=T_1\cup T_2$，实体的总个数$n=|E_1|+|E_2|$。

2.1.2　图神经网络

GNN通过递归聚合其邻居的特征向量来学习节点表示，不同的聚合策略产生了GNN的不同变体，其中的一种变体vanilla GCN[24]在第$l(l\geqslant1)$层处的节点i的隐藏表示为$h_i^{(l)}$，如式(1)所示：

$$h_i^{(l)}=\sigma\left(\sum_{a\in N(i)\cup\{i\}}\frac{1}{\varepsilon_i}W_1^{(l)}h_a^{(l-1)}\right)\tag{1}$$

其中：$\{h_1^{(l)},h_2^{(l)},\cdots,h_n^{(l)}|h_i^{(l)}\in\mathbb{R}^{d^{(l)}}\}$代表第$l$个GCN层的输出节点特征；$N(i)$代表实体$i$的邻居集合；$\varepsilon_i$代表归一化常数；$W_1^{(l)}\in\mathbb{R}^{d^{(l)}\times d^{(l-1)}}$代表第1层可训练权重矩阵元素；$\sigma(\cdot)$代表激活函数，通常采用ReLU(·)作为激活函数。Vanilla GCN是将节点编码为其邻居和最后一层自身表示的平均池(mean pooling)操作，第一层的输入表示为$h_i^{(0)}$。

2.1.3　远距离邻居选择

为了减少邻域信息所带来的非同构影响，本文方法引入远距离邻居。如图1所示，两个中心实体对(a,A)的一跳邻居不同，只包含对等实体对(b,B)和(c,C)，而a的一跳邻居d和A的远距离邻居D对应，A的一跳邻居E和F与a的远距离邻居e和f对应。如果可以将远距离邻居e和f包含在a的邻域聚合中，并且也将A的远距离邻居D考虑在内，那么GNN将会学习到更多关于a和A的相似表示。但是，并非所有的远距离邻居都有积极贡献，因此，本文引入注意力机制，旨在找到对中心实体有积极贡献的远距离邻域。

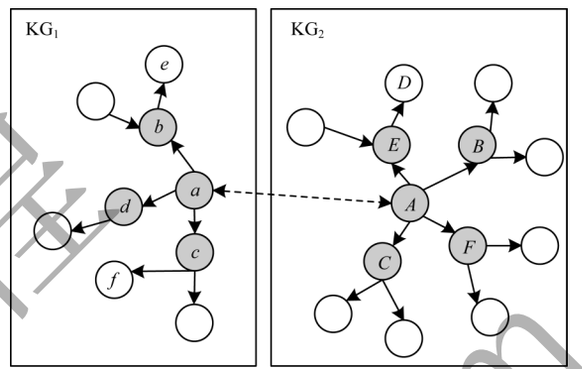

图1　远距离邻居选择示例

Fig.1　Example of selecting long-distance neighbors

2.1.4　图匹配

通过图的结构信息来度量两个图的相似性，进而估计G_1和G_2描述的为同一实体的可能性。在近期研究中，图匹配网络(GMN)[25]引入跨图关注机制对图进行联合推理，以区分跨图的节点并识别差异，计算两个图之间的相似度得分。受GMN模型的启发，本文也采用一跨图邻域匹配模块来识别两个实体邻域节点之间的差异。

2.1.5　距离函数

对于不同实体之间的相似性，通常采用计算实体之间的距离来度量，而计算距离的方法会直接关系到对齐的效果。表1中列举了一些常见的距离函数。

表1　常见的距离函数

Table 1　Common distance functions

距离函数	函数表达式	优点	缺点
欧式距离	$d=\|\|a-b\|\|_1$	适用于处理低维数据，“开箱即用”	需要对数据进行归一化处理
余弦相似度	$d=\cos(\theta)=\frac{a\cdot b}{\|\|a\|\|\,\|\|b\|\|}$	适用于文本分析	只考虑向量方向
曼哈顿函数	$d=\sum_{i=1}^{k}\|a_i-b_i\|$	适用于离散型数据	直观性差
切比雪夫距离	$d=\max(\|x-y\|)$	适用于仓库物流领域	只适用于特定的用例，通用性差

2.2　NAMN模型框架

为了缓解邻域异质性对实体对齐产生的影响，NAMN模型利用GNN对图结构信息进行建模，采用分层的思想对邻域信息进行区别处理。首先，对于一跳邻居进行全部采样，对于二跳及以上邻居，采用注意力机制进行局部采样；然后，引入门控机制对实体的k-hop邻居信息进行聚合，以挖掘图结构的隐藏信息；在此基础上，考虑到实体一跳邻居结构异质性的影响，为每个实体提取一个可区分的邻域，构建邻域局部子图进行跨图邻域匹配，将匹配阶段的输出与通过门控机制所学习的图结构表示进行联合编码，以生成面向匹配的实体表示；最后，对于面向匹配的实体表示，使用距离函数进行实体的对齐预测。NAMN模型框架如图2所示，在不失一般性的情况下，该图展示了一跳和二跳邻居信息的情况。NAMN遵循3个阶段的处理流程，即门控邻域聚合、邻域匹配和对齐预测。

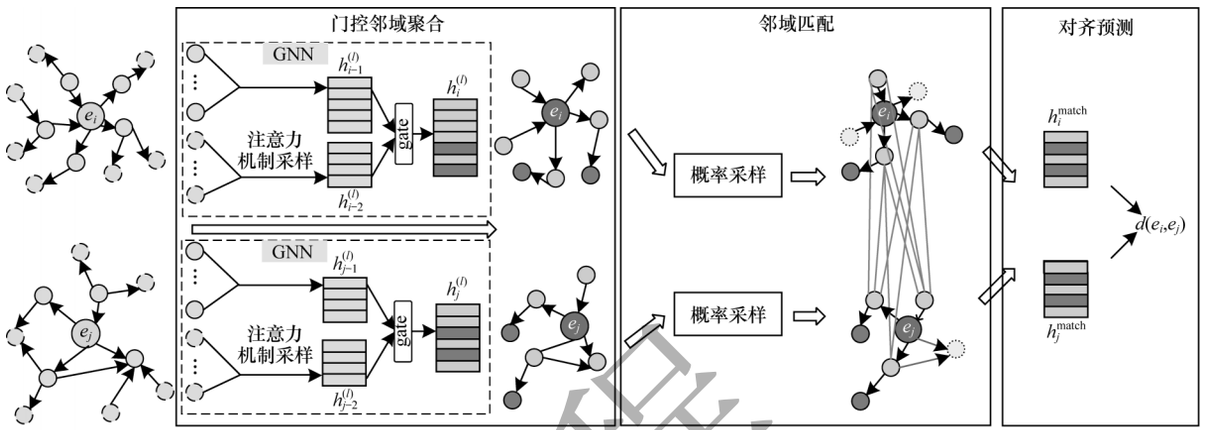

图2 NAMN模型框架

Fig.2 NAMN model framework

2.3 门控邻域聚合

按照每跳邻居对中心实体的重要性可知，实体的一跳邻居是最重要的邻域。本文使用vanilla GCN聚合实体的邻居信息，学习知识图谱结构嵌入。首先，使用预训练的词嵌入[26]来初始化GCN的方法。将两个知识图谱 $G_1=(E_1,R_1,T_1)$ 和 $G_2=(E_2,R_2,T_2)$ 作为NAMN模型的输入，使用式(1)来更新节点表示。为控制噪声的影响，还引入highway networks[27]方法，以避免噪声在GNN层之间传播。

对于二跳邻居，若再直接采用GNN层来聚合，会导致更多的噪声信息。为找到对中心实体有积极贡献的远距离邻域，本文引入注意力机制[23]来计算实体 e_i 的二跳邻居信息(表示为 $h_{i-2}^{(l)}$)，如式(2)所示：

$$h_{i-2}^{(l)}=\mathrm{ReLU}\left(\sum_{a\in N_2(i)\cup\{i\}}\beta_{ia}\boldsymbol{W}_2h_a^{(l-1)}\right) \tag{2}$$

其中：$N_2(\cdot)$ 代表的是实体的二跳邻居集合；$\boldsymbol{W}_2$ 是可训练的权重矩阵；β_{ia} 是中心实体 i 与邻居 a 的一个可学习的注意力权重。

为进一步聚合邻域信息，使用门控机制将邻域信息结合在一起，以挖掘实体 e_i 的隐藏表示：

$$g(h_i^{(l)})=\mathrm{ReLU}(\boldsymbol{M}h_i^{(l)}+\boldsymbol{b}) \tag{3}$$

$$h_i^{(l)}=g\left(h_{i-2}^{(l)}\right)\cdot h_{i-1}^{(l)}+\left(1-g(h_{i-2}^{(l)})\right)\cdot h_{i-2}^{(l)} \tag{4}$$

其中：$g(h_i^{(l)})$ 为控制一跳和二跳邻居组合的门；“·”表示逐元素乘法；$\boldsymbol{M}$ 和 $\boldsymbol{b}$ 分别为权重矩阵和偏差向量。

对于two-hop邻域聚合，本文引入注意力权重 β_{ia} 以突出重要邻居。图注意力网络(Graph Attention Network，GAT)[28]是在实体中采用共享的线性变换，但是却忽略了中心实体和邻居之间可能完全不同，这种共享的转换会导致无法正确区分。为此，本文分别使用两个矩阵 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$ 对中心实体和邻居进行线性变换[23]。形式上，中心实体 i 和邻居 a 之间的注意力权重计算方法如式(5)所示：

$$c_{ia}=\mathrm{attn}(\boldsymbol{M}_1h_i^{(l)},\boldsymbol{M}_2h_a^{(l)})=\mathrm{LeakyReLU}(p[\boldsymbol{M}_1h_i^{(l)}\|\boldsymbol{M}_2h_a^{(l)}]) \tag{5}$$

其中：p、$\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$ 为可训练的参数；$\|$ 表示级联；c_{ia} 是衡量 e_i 和 e_a 重要性的注意力权重；attn(·)是注意力函数。在此基础上，使用softmax(·)函数进行标准化处理，以使其在不同实体之间具有可比性，从而有效地编码实体名称的语义信息：

$$\beta_{ia}=\mathrm{softmax}\left(c_{ia}\right)=\frac{\exp(c_{ia})}{\sum_{k\in N_2(i)\cup\{i\}}\exp(c_{ik})} \tag{6}$$

将两个知识图谱 $G_1=(E_1R_1T_1)$ 和 $G_2=(E_2R_2T_2)$ 作为NAMN模型的输入，使用式(1)来更新节点表示。为控制噪声的影响，还引入highway networks[27]方法，以避免噪声GNN层之间传播。

2.4 邻域匹配

2.4.1 邻域局部子图构建

实体的一跳邻居是决定该实体与其他实体是否对齐的关键，但是并非所有的一跳邻居都对实体对齐有着积极的影响。为此，本文引入局部子图，应用向下采样过程(down-sampling process)，旨在选择对中心实体信息量最大的一跳邻居。对于每个实体对 (e_i,e_j)，如果 e_i 和 e_j 有关系(例如 r)直接连接，在局部子图中为其对应节点添加一有向边，但只保留 r 的方向。

为了选择合适的邻居，采用邻里采样策略[29]。给定实体 e_i，对其一跳邻居 e_{i_1} 进行采样的概率如式(7)所示：

$$p(h_{i\_1}|h_i)=\mathrm{softmax}\left(h_i\boldsymbol{W}_3h_{i\_1}^{\mathrm{T}}\right)=\frac{\exp(h_i\boldsymbol{W}_3h_{i\_1}^{\mathrm{T}})}{\sum_{n\in N_i}\exp(h_i\boldsymbol{W}_3h_{i\_n}^{\mathrm{T}})} \tag{7}$$

其中：$\boldsymbol{W}_3$ 是共享的权重矩阵；N_i 是中心实体 e_i 的一跳邻居集；h_i 和 h_{i_1} 分别是中心实体 e_i 和一跳邻居 e_{i_1} 通过式(1)计算的学习嵌入表示。

2.4.2 跨图邻域匹配

确定对中心实体应考虑的邻居之后，也即产生了邻域局部子图。在跨图邻域匹配过程中，为减少匹配开销，首先进行的为候选人的选择。计算 G_1 中的实体 e_i 与 G_2 中的所有实体 $\{e_2\}$ 在其表示空间中的相似性，

找到 G_2 在嵌入空间中最接近 e_i 的实体，作为候选者 $D_i=\{D_{i_1},D_{i_2},\cdots,D_{i_l}|D_{i_k}\in E_2\}$，计算公式如式(8)~式(9)所示：

$$\alpha_{ij}=s_h(h_i,h_j),j\in\{1,2,\cdots,|E_2|\} \tag{8}$$

$$p(h_i|h_j)=\frac{\sum_{j=1}^{|E_2|}\alpha_{ij}\cdot h_j}{\sum_{j=1}^{|E_2|}\alpha_{ij}} \tag{9}$$

其中：s_h 是向量空间相似性度量，如 Euclidean 或 cosine；α_{ij} 是注意力权重；$p(h_j|h_i)$ 为 G_2 中的实体 e_j 被采样为 e_i 候选者的概率。

在邻域匹配模块中，G_1 和 G_2 叠加在一起作为一个大的输入图，引入一匹配向量来计算 G_1 中的实体邻域和 G_2 中所有实体的匹配程度[25]。形式上，令 (e_i,D_{i_k}) 为要测量的实体对，其中 $e_i\in E_1$，而 $D_{i_k}\in E_2$ 为候选者之一，设定 x 和 y 分别是 e_i 和 D_{i_k} 的两个邻居，得到邻居 x 的匹配向量 $\boldsymbol{m}_x$：

$$\alpha_{xy}=\frac{\exp(h_x\cdot h_y)}{\sum_{y'\in N_{i_k}^s}\exp(h_x\cdot h_{y'})} \tag{10}$$

$$\boldsymbol{m}_x=\sum_{y\in N_{i_k}^s}\alpha_{xy}(h_x-h_y)=h_x-\sum_{y\in N_{i_k}^s}h_y \tag{11}$$

其中：$N_{i_k}^s$ 是 D_{i_k} 的采样邻居集；h_x 和 h_y 是邻居 x 和 y 通过式(1)计算的GCN输出嵌入表示。

然后，将邻居 x 的输出嵌入与匹配向量相结合：

$$\hat{h}_x=[h_x\|m_x\cdot\tau] \tag{12}$$

其中：τ 是超参。此处，匹配向量 $\boldsymbol{m}_x$ 可以区分两个邻居之间的差异。当两个邻居表示相似时，$\boldsymbol{m}_x$ 趋向于零向量；当邻居表示不同时，匹配向量 $\boldsymbol{m}_x$ 将会变大。

最后，汇总其采样的邻居表示集为 $\{\hat{h}_x\}$，并使用式(13)的聚合方法[30]计算中心实体的 e_i 的 one-hop 的邻域表示：

$$g_{i\_1}=\left(\sum_{x\in N_i^s}\mathrm{ReLU}(\hat{h}_x\boldsymbol{W}_{\mathrm{gate}})\cdot\hat{h}_x\right)\boldsymbol{W}_N \tag{13}$$

其中：$\boldsymbol{W}_{\mathrm{gate}}$ 和 $\boldsymbol{W}_N$ 分别是可学习的门控制矩阵和共享矩阵；N_i^s 是采样的邻居集。

在此基础上，将式(4)所计算的实体的隐藏表示 $h_i^{(l)}$ 及其一跳邻居表示连接起来，生成最终的面向匹配的表示：

$$h_i^{\mathrm{match}}=[g_{i\_1}\|h_i^{(l)}] \tag{14}$$

2.5　对齐预测

对于最终生成的面向匹配的实体表示 h^{match}，可以简单地通过测量两个实体之间的距离来判定两个实体是否应该对齐：

$$d(e_i,e_j)=\pi(e_i,e_j)=\left\|h_i^{\mathrm{match}}-h_j^{\mathrm{match}}\right\|_1 \tag{15}$$

其中：$\|\cdot\|_1$ 表示 L_1 范数。将基于边际的排名损失函数作为NAMN模型的目标：

$$O=\sum_{(c,v)\in Z}d(c,v)+\sum_{(c',v')\in R}\alpha[\lambda-d(c',v')] \tag{16}$$

其中：c 和 v、D_c 和 D_v 组成负样本的候选集 $R=\{(c',v')|(c'=c\cap v'\in D_c)\cup(v'=v\cap c'\in D_v)\}$；$Z$ 是候选的种子集；α 是平衡的超参数。本文目标是使对齐的实体具有很小的距离，未对齐实体表示具有较大的距离，即负样本的距离应该大于 λ，也即 $d(c',v')>\lambda$。

此外，使用Adam优化器[31]对目标进行优化，通过Xavier初始化[11]对所有可学习的参数(包括实体的输入特征向量)进行初始化。

3　实验

3.1　数据集

为了评估NAMN模型性能，参考最近的研究[12,32]，本文使用大型数据集DBpedia[33]下的子集DBP15K作为实验数据进行验证。这些数据集包括3个跨语言数据集，分别是英语、中文、日语和法语的不同语言版本，即 $\mathrm{DBP15K_{ZH\text{-}EN}}$(中文-英语)、$\mathrm{DBP15K_{JA\text{-}EN}}$(日语-英语)、$\mathrm{DBP15K_{FR\text{-}EN}}$(法语-英语)，每个数据集由15 000个对齐的实体对和约40万个三元组组成。3个数据集的详细信息如表2所示。

表2　数据集统计

Table 2　Data set statistics

数据集	Ent.	Rel.	Tri.
$\mathrm{DBP15K_{ZH\text{-}EN}}$	66 469	2 830	153 929
	98 125	2 317	237 674
$\mathrm{DBP15K_{JA\text{-}EN}}$	65 744	2 043	164 373
	95 680	2 096	233 319
$\mathrm{DBP15K_{FR\text{-}EN}}$	66 858	1 379	192 191
	95 680	2 209	278 590

3.2　评估指标与参数设置

按照惯例，将数据集的30%作为训练数据，剩下的70%用作测试数据。在以下超参数中进行搜索：学习率 $R_{\mathrm{learning\_rate}}=\{0.001,0.005,0.01\}$，$\tau=\{0.1,0.2,\cdots,0.5\}$，$\alpha=\{0.01,0.05,\cdots,0.1,0.2\}$，$\lambda=\{1.5,1.4,\cdots,1.0\}$，每层的隐藏层层数 $L=\{1,2,3,4\}$，维度为 {100, 200, 300, 400, 500}，最终实验设置如表3所示。此外，本文设置候选人的大小为20个，并为每个预先对齐的实体对抽取10个负样本，以简化训练。

表3　参数设置

Table 3　Parameters setting

参数	参数值
λ	1.5
τ	0.1
α	0.1
$R_{\mathrm{learning\_rate}}$	0.001
AliNet维度	400,300
g_{i_1} 维度	50

对于评估指标，使用Hits@K和平均倒数排名(Mean Reciprocal Rank，MRR)评估对齐性能。Hits@K通过排名在前K个的正确对齐实体的比例来进行计算，MRR是指所有正确实体的平均倒数排名。这两个指标值越高，表明实体对齐模型效果越好。

3.3 对比方法与实验结果

本文将NAMN模型与最近提出的基于嵌入的实体对齐模型进行比较，并将其分为2类：1）基于嵌入的模型，如MTransE、IPTransE、JAPE、BootEA和RSN；2）基于图的模型，如GCN-Align和RDGCN。同时，引入近期考虑到知识图谱邻域异质性的两个最新成果进行比较，即MuGNN和AliNet模型。

表4列出了在DBP15K数据集上所有方法的实体对齐性能。实验结果表明，NAMN明显优于3个数据集上的所有基线模型。NAMN模型可以实现Hits@1的所有值均高于75%，Hits@10的所有值均高于85%，MRR的所有值均不低于80%，这进一步证实了本文方法的有效性。具体来说，在基于嵌入模型中，BootEA模型表现最佳，通过引导过程可以从更多训练实例中受益。对于仅考虑结构信息的基于GNN的模型，RDGCN明显优于其他模型，这是因为RDGCN模型考虑从邻域结构入手，缓解了结构异质性所带来的影响，体现出解决结构异质性的重要性。

表4 不同实体对齐方法性能比较

Table 4 Performance comparison of different entity alignment methods %

模型	$DBP15K_{ZH-EN}$			$DBP15K_{JA-EN}$			$DBP15K_{FR-EN}$		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
MTransE	30.8	61.4	36.4	27.9	57.5	34.9	24.4	55.6	33.5
IPTransE	40.6	73.5	51.6	36.7	69.3	47.4	33.3	68.5	45.1
JAPE	41.2	74.5	49.0	36.3	68.5	47.6	32.4	66.7	43.0
BootEA	62.9	84.8	70.3	62.2	85.4	70.1	65.3	87.4	73.1
RSN	50.8	74.5	59.1	50.7	73.7	59.0	51.6	76.8	60.5
GCN-Align	41.3	74.4	54.9	39.9	74.5	54.6	37.3	74.5	53.2
RDGCN	70.8	84.6	74.6	76.7	89.5	81.2	88.6	95.7	91.1
MuGNN	49.4	84.4	61.1	50.1	85.7	62.1	49.5	87.0	62.1
AliNet	53.9	82.6	62.8	54.9	83.1	64.5	55.2	85.2	65.7
NAMN	76.8	89.4	82.1	79.2	93.6	89.4	92.9	97.4	96.4

为进一步证明NAMN模型的有效性，在DBP15K的另外3个数据集上进行对比实验。这3个数据集分别是$DBP15K_{EN-ZH}$（英语-中文），$DBP15K_{EN-JA}$（英语-日语），$DBP15K_{EN-FR}$（英语-法语），实体对齐的结果比对如表5所示。可以看出，所有模型的性能都有所下降，但NAMN模型明显优于另外3个模型，NAMN模型的Hits@1的值要高于另外3个模型约30%以上，其中，在$DBP15K_{EN-FR}$数据集中Hits@1达到了最高，充分证明了NAMN模型的有效性和鲁棒性。

表5 实体对齐结果比较

Table 5 Comparison of entity alignment results %

模型	$DBP15K_{EN-ZH}$		$DBP15K_{EN-JA}$		$DBP15K_{EN-FR}$	
	Hits@1	Hits@10	Hits@1	Hits@10	Hits@1	Hits@10
MTransE	24.78	52.42	23.72	49.92	21.26	50.60
JAPE	40.15	71.05	38.37	67.27	32.97	65.91
GCN-Align	36.49	69.94	38.42	71.81	36.77	73.06
NAMN	70.23	86.26	77.46	91.52	89.26	97.42

为更直观地表现NAMN模型的性能，在DBP15K数据集上，采用Hits@1到Hits@50以10为步长的多个基准进行比较，选择JAPE、GCN-Align和AliNet作为对比模型，具体如图3所示，其中横坐标为Hits@K。可以看出：NAMN模型的Hits@K值均高于其他模型，在$DBP15K_{JA-EN}$和$DBP15K_{FR-EN}$上都取得了最高的得分；AliNet模型的Hits@K值在K取20之后，得分接近NAMN模型，说明缓解实体邻域结构的异质性有利于实体对齐，但AliNet模型的Hits@1明显低于NAMN模型，说明NAMN模型具有更好的对齐性能。

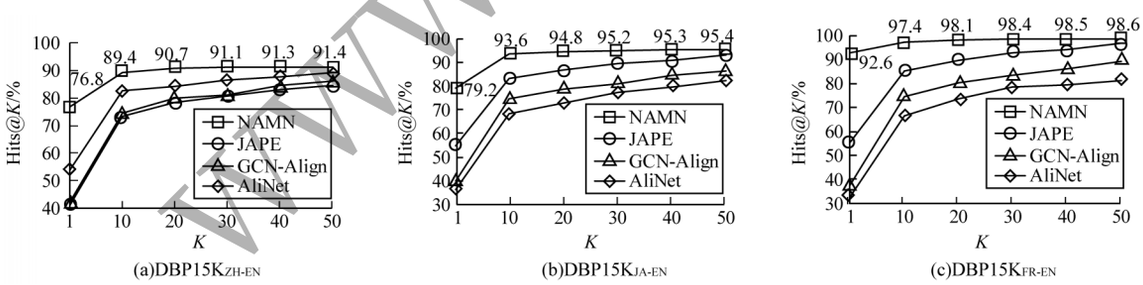

图3 DBP15K数据集上Hits@K得分结果比较

Fig.3 Comparison of Hits@K score results on DBP15K dataset

3.4 结果分析

NAMN模型使用门控机制和邻域采样策略来实现实体对齐，因此，分别对这两个策略进行分析。

将NAMN模型在DBP15K数据集上采用随机采样策略来进行比较，具体结果如图4所示。可以看出，NAMN模型可以提供更好的结果，本文的采样策略可以有效地选择信息量更大的邻居。对DBP15K数据集，两个模型均可达到性能平稳状态，当采样大小为3时，两个模型的性能更高。但随着采样大小的增大，性能会有所下降，说明较大的采样会引入更多的噪声。

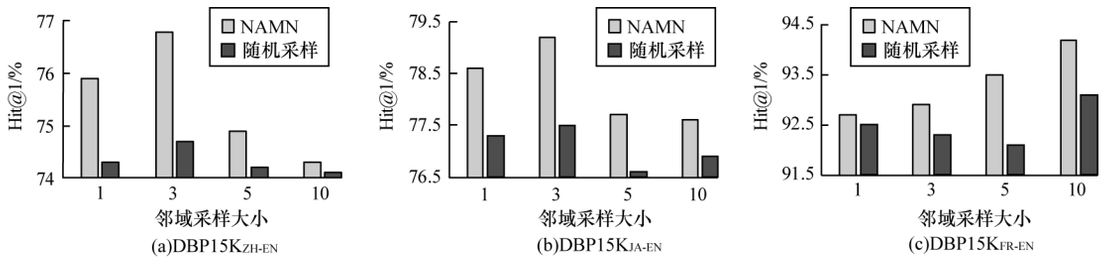

图4　DBP15K数据集上邻域抽样策略与随机采样策略的结果比较

Fig.4　Result comparison of neighborhood sampling strategy and random sampling strategy on DBP15K dataset

在聚合多跳邻居方面，本文使用不同的策略来设计NAMA的不同变体。变体1（NAMN-1）将实体的一跳和二跳邻居平等对待，使用GNN层直接聚合邻居信息；变体2（NAMN-2）用加法运算符替换门控机制；变体3（NAMN-3）用GAT来替换本文所用的注意力机制。由表6可以看出：NAMN-1的实验结果很差，这表明使用GNN层来直接聚合二跳邻居会引入很多的噪声信息，严重影响对齐性能；NAMN-2的实验结果较差，这表明加法机制只是简单的将邻居信息结合，并不会像门控机制那样选择性地组合各个维度上的邻居信息表示；NAMN-3的实验结果比NAMN略差，这表明本文所用注意力机制能够优化对齐效果。因此，对于远距离邻居选择，门控机制和注意力机制至关重要。

表6　NAMN不同变体对齐结果比较

Table 6　Comparison of alignment results of different variants of NAMN　%

模型	DBP15K$_{ZH-EN}$		DBP15K$_{JA-EN}$		DBP15K$_{FR-EN}$	
	Hits@1	Hits@10	Hits@1	Hits@10	Hits@1	Hits@10
NAMN-1	32.3	66.1	42.4	78.4	43.4	77.1
NAMN-2	70.9	86.7	74.3	91.6	84.3	95.9
NAMN-3	73.7	86.9	76.6	91.2	88.1	96.6
NAMN	76.8	89.4	79.2	93.6	92.9	97.4

在DBP15K数据集上1~4层的AliNet实验结果如图5所示，其中横坐标为AliNet的层数。可以看出：当AliNet的层数为2时，所有指标达到了最佳性能；当AliNet具有更多层时，其性能也会下降。

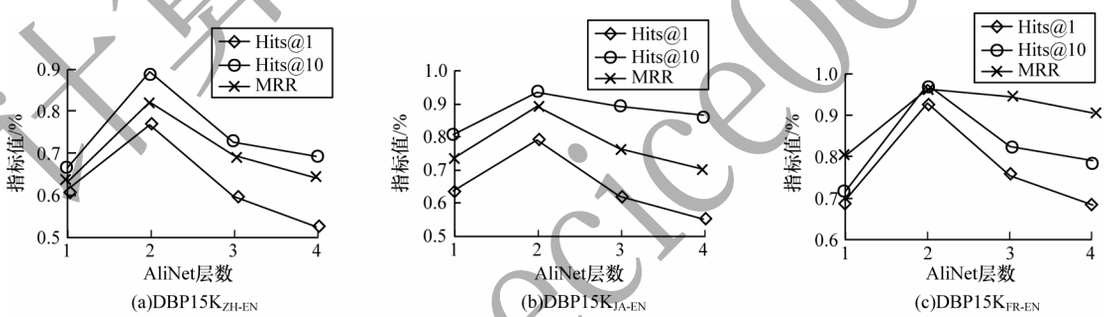

图5　DBP15K数据集上不同AliNet层数的实验结果比较

Fig.5　Experimental result comparison of different AliNet layers on DBP15K dataset

4　结束语

为提高实体对齐的准确性，本文提出邻域聚合匹配网络（NAMN）模型。从实体邻域角度出发，通过门控邻域聚合、邻域匹配和对齐预测3个阶段判定实体是否对齐，解决知识图谱间普遍存在的结构异质性问题。实验结果表明，在DBP15K数据集上，该模型的Hits@K指标达到75%以上。后续将利用实体的语义信息和关系的映射属性提高实体对齐的准确度，并进一步改进邻域的匹配策略，降低模型的复杂度，从而扩大模型的应用范围。

参考文献

[1] 孟明明，张坤，论兵，等. 一种面向知识图谱问答的语义查询扩展方法[J]. 计算机工程，2019，45(9)：276-283，290.
MENG M M，ZHANG K，LUN B，et al. A semantic query expansion method for question answering based on knowledge graph[J]. Computer Engineering，2019，45(9)：276-283，290. (in Chinese)

[2] 王辉，郁波，洪宇，等. 基于知识图谱的Web信息抽取系统[J]. 计算机工程，2017，43(6)：118-124.
WANG H，YU B，HONG Y，et al. Web information extraction system based on knowledge graph[J]. Computer Engineering，2017，43(6)：118-124. (in Chinese)

[3] CAO Y X，HOU L，LI J Z，et al. Joint representation learning of cross-lingual words and entities via attentive distant supervision[C]//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg，USA：Association for Computational Linguistics，2018：1-11.

[4] CAO Y X，WANG X，HE X N，et al. Unifying knowledge graph learning and recommendation：towards a better understanding of user preferences[C]//Proceedings of World Wide Web Conference. Washington D. C.，USA：IEEE Press，2019：151-161.

[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] SPEER R, CHIN J, HAVASI C. ConceptNet 5.5: an open multilingual graph of general knowledge[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2017: 4444-4451.
[8] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an arcHitsecture for never-ending language learning[C]// Proceedings of AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2010: 1-10.
[9] MAHDISOLTANI F, BIEGA J, SUCHANEK F. YAGO3: a knowledge base from multilingual Wikipedias[C]// Proceedings of the 7th Biennial Conference on Innovative Data Systems Research. Asilomar, USA: [s.n.], 2014: 4-7.
[10] CHEN M H, TIAN Y T, YANG M H, et al. Multi-lingual knowledge graph embeddings for cross-lingual knowledge alignment[EB/OL]. [2021-05-05]. https://arxiv.org/abs/1611.03954.
[11] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. [S.l.]: JMLR, 2010: 249-256.
[12] SUN Z, HU W, ZHANG Q, et al. Bootstrapping entity alignment with knowledge graph embedding [C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. Washington D. C., USA: IEEE Press, 2018: 4396-4402.
[13] SUN Z, HU W, LI C. Cross-lingual entity alignment via joint attribute-preserving embedding[C]//Proceedings of International Semantic Web Conference. Berlin, Germany: Springer, 2017: 628-644.
[14] GUO L, SUN Z, HU W. Learning to exploit long-term relational dependencies in knowledge graphs [C]// Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2019: 2505-2514.
[15] ZHANG Q H, SUN Z Q, HU W, et al. Multi-view knowledge graph embedding for entity alignment[EB/OL]. [2021-05-05]. https://arxiv.org/abs/1906.02390.
[16] WANG Z C, LV Q S, 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.
[17] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks [C]//Proceedings of European Semantic Web Conference. Berlin, Germany: Springer, 2018: 593-607.
[18] WU Y, LIU X, FENG Y, et al. Relation-aware entity alignment for heterogeneous knowledge graphs[EB/OL]. [2021-04-10]. https://arxiv.org/abs/1908.08210.
[19] YING R, YOU J, MORRIS C, et al. Hierarchical graph representation learning with differentiable pooling[EB/OL]. [2021-05-05]. https://proceedings.neurips.cc/paper/2018/hash/e77dbaf6759253c7c6d0efc5690369c7-Abstract.html.
[20] MORRIS C, RITZERT M, FEY M, et al. Weisfeiler and leman go neural: higher-order graph neural networks[C]// Proceedings of 2019 AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2019: 4602-4609.
[21] LEMAN A A, WEISFEILER B. A reduction of a graph to a canonical form and an algebra arising during this reduction [J]. Nauchno-Technicheskaya Informatsiya, 1968, 2(9): 12-16.
[22] PUJARA J, MIAO H, GETOOR L, et al. Knowledge graph identification[C]//Proceedings of International Semantic Web Conference. Berlin, Germany: Springer, 2013: 542-557.
[23] SUN Z, WANG C, HU W, et al. Knowledge graph alignment network with gated multi-hop neighborhood aggregation[C]//Proceedings of 2020 AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2020: 222-229.
[24] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. [2021-05-05]. https://arxiv.53yu.com/abs/1609.02907.
[25] LI Y, GU C, DULLIEN T, et al. Graph matching networks for learning the similarity of graph structured objects[C]// Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2019: 3835-3845.
[26] XU K, WANG L, YU M, et al. Cross-lingual knowledge graph alignment via graph matching neural network[EB/OL]. [2021-05-05]. https://arxiv.53yu.com/abs/1905.11605.
[27] SRIVASTAVA R K, GREFF K, SCHMIDHUBER J. Highway networks [EB/OL]. [2021-05-05]. https://arxiv.53yu.com/abs/1505.00387.
[28] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL]. [2021-05-05]. https://arxiv.53yu.com/abs/1710.10903.
[29] WU Y, LIU X, FENG Y, et al. Neighborhood matching network for entity alignment[EB/OL]. [2021-05-05]. https://arxiv.53yu.com/abs/2005.05607.
[30] LI Y J, TARLOW D, BROCKSCHMIDT M, et al. Gated graph sequence neural networks[EB/OL]. [2021-05-05]. https://arxiv.org/abs/1511.05493.
[31] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. [2021-05-05]. https://arxiv.53yu.com/abs/1412.6980.
[32] CAO Y, LIU Z, LI C, et al. Multi-channel graph neural network for entity alignment[EB/OL]. [2021-05-05]. https://arxiv.org/abs/1905.11605.
[33] LEHMANN J, ISELE R, JAKOB M, et al. DBpedia-a large-scale, multilingual knowledge base extracted from Wikipedia[J]. Semantic Web, 2015, 6(2): 167-195.

编辑 金胡考