**文章编号**：1004-4353(2021)04-0356-05

# 融合 GAT 和 TransH 的中韩跨语言实体对齐方法研究

金城， 崔荣一， 赵亚慧， 张振国
（延边大学 工学院，吉林 延吉 133002）

**摘要**：为研究中韩双语实体自动对齐方法，提出了一种融合图注意力网络(GAT)和基于超平面平移的知识图谱嵌入模型(TransH)的跨语言实体对齐模型.使用中韩实体数据集对模型进行验证表明，该模型的 Hits@1、Hits@5 和 Hits@10 在韩文对齐中文时分别达到了 49.62%、80.89%和 91.76%，在中文对齐韩文时分别达到 49.79%、80.74%和 91.67%，且优于传统的基于知识嵌入或图嵌入的对齐方法.因此该模型可为构建中韩对齐知识图谱以及其他语言的对齐知识图谱提供参考.

**关键词**：中韩跨语言实体；实体对齐；知识图谱；图注意力网络
**中图分类号**：TP391 **文献标识码**：A

# Cross-lingual entity alignment in Chinese and Korean based on GAT and TransH

JIN Cheng， CUI Rongyi， ZHAO Yahui， ZHANG Zhenguo
（*College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*）

**Abstract**: To study the automatic alignment method of Chinese and Korean bilingual entities, a cross-language entity alignment model combining graph attention network (GAT) and knowledge graph embedding model based on hyperplane translation (TransH) is proposed. Validation of the model using the Chinese and Korean entity data sets shows that the Hits@1, Hits@5, and Hits@10 of the model reached 49.62%, 80.89% and 91.76% respectively, when aligning Korean to Chinese; when Chinese is aligned with Korean, it reaches 49.79%, 80.74% and 91.67% respectively. It is better than traditional alignment methods based on knowledge embedding or graph embedding. Therefore, the model can provide a reference for constructing a Chinese-Korean alignment knowledge graph and alignment knowledge graph of other languages.

**Keywords**: Chinese-Korean cross language entity; entity alignment; knowledge graph; graph attention network

## 0 引言

多语言知识图谱能够将客观世界中用不同语言描述的大量实体、属性和关系构建成一个庞大的知识网络.在面向多语言场景的人工智能应用中，建立丰富的多语言知识图谱库可为人工智能应用提供先验知识，提高其认知和理解能力.但构建多语言知识图谱往往需要耗费大量的人力及物力来处理多种语言的海量数据，因此研究如何低成本、高效率地建立多语言知识图谱具有重要的意义和价值.

在实现跨语言实体对齐时，目前通常采用基于知识嵌入或基于图嵌入的对齐方法.基于知识嵌入的

**收稿日期**：2021-05-13
**基金项目**：延边大学校企合作项目(延大科合字[2020]15 号)
**第一作者**：金城(1997—)，男，硕士，研究方向为知识图谱与跨语言对齐.
**通信作者**：崔荣一(1962—)，男，博士，教授，研究方向为自然语言文本处理与模式识别.

对齐方法是采用类似于词向量分布式的表示方法对知识图谱中的实体和关系进行表示的一种方法(TransH)，该方法虽具有准确率较高、模型复杂度低、训练相对简单等优点，但存在对数据量要求较高的缺点. 基于图嵌入的实体对齐方法是利用图神经网络学习知识图谱中的图结构信息和节点之间的相互依赖关系在向量空间中表示实体和关系的一种方法(GAT)，该方法虽然在数据量较低的情况下也具有较高的准确率，但存在占用内存高、训练速度慢和易受图结构异构影响等缺点. 2019 年，Li 等提出了 KECG 模型，该模型采用相似的方法进行跨语言实体对齐实验，并取得了较好的实验结果[1]. 目前为止还未发现有学者对中韩两种语言的实体的自动对齐和中韩对齐数据集进行研究，因此本文采用将 TransH[2] 和 GAT[3] 相融合的方法，研究如何在数据量较低以及图结构异构情况下实现中韩实体的自动对齐.

## 1　融合 TransH 和 GAT 的跨语言实体对齐模型

本文提出的融合 TransH 和 GAT 的跨语言实体对齐模型如图 1 所示. 该模型主要包括图嵌入层和知识嵌入层两部分，分别用于提取知识图谱的图结构特征信息和实体间的关系特征信息. 模型的输入为中韩两种语言的知识图谱和预先对齐的实体对. 模型经过迭代训练后，将实体对映射到向量空间中. 在向量空间中，具有相同语义的等价实体相互靠近，其距离通过 $L_2$ 范数计算. 由图 1 可以看出，模型通过多轮迭代更新后，可为不同语言中具有相同语义的实体赋予没有冲突且一一对应的关系，同时也可为实体和关系赋予新的、合适的向量表示，以此计算出所有实体间可能存在的对齐关系.

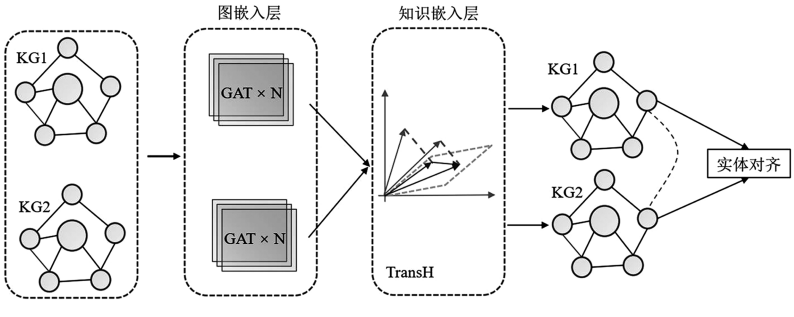


**图 1　融合 TransH 和 GAT 的跨语言实体对齐模型**

### 1.1　图嵌入层

图嵌入模型的目标是利用知识图谱的结构特征将双语知识图谱中的对齐实体嵌入到一个统一的向量空间中. GAT 与其他图神经网络相比具有如下优点：①可以并行计算相邻节点对，提高模型的计算速率；②可以为具有多个与节点相连的边赋予任意大小的权重；③可以适配和训练结构不同的数据集；④在引入注意力机制后，模型只关注邻居节点，因此模型的计算速率得到提高. 基于 GAT 的上述优点，本文采用 GAT 作为编码器(encoder)获取知识图谱的图结构信息，通过对不同邻居节点赋予不同的关注度来忽略一些重要度相对较低的节点，从而降低不同知识图谱异构带来的影响.

图嵌入层的输入为知识图谱的网络结构和实体嵌入矩阵 $\boldsymbol{X}\in\mathbf{R}^{x\times d}$. 知识图谱的网络结构用每个实体邻居节点的集合表示，其中实体的维度用 $d$ 表示. 编码器通过叠加多个图注意力层(graph attention layer) 来实现，其表达式为：

$$\boldsymbol{H}^{(l+1)}=\sigma(\boldsymbol{A}^{(l)}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}), \tag{1}$$

其中 $\boldsymbol{H}^{l}(\boldsymbol{H}^{l}\in\mathbf{R}^{n\times d^{(l)}})$ 和 $\boldsymbol{W}^{(l)}(\boldsymbol{W}^{(l)}\in\mathbf{R}^{d^{l}\times d^{(l+1)}})$ 分别是第 $l$ 层的隐藏状态和权重，$\boldsymbol{H}^{(0)}=\boldsymbol{X}$，$\sigma(\cdot)$ 为非线性激活函数(activation function) $\mathrm{ReLU}(\cdot)=\max(0,\cdot)$，$\boldsymbol{A}^{(l)}\in\mathbf{R}^{n\times n}$ 是一个利用自注意力机制对模型输入的图经过计算后得到的连通矩阵. 在连通矩阵 $\boldsymbol{A}^{(l)}$ 中，实体 $\boldsymbol{e}_i$ 到 $\boldsymbol{e}_j$ 的权重用元素 $a_{ij}^{(l)}$ 表示. $a_{ij}^{(l)}$ 由如下自注意力机制计算公式得到：

$$a_{ij}^{(l)}=\mathrm{softmax}(c_{ij}^{(l)})=\frac{\exp(c_{ij}^{l})}{\sum\limits_{\boldsymbol{e}_k\in N_{\boldsymbol{e}_i}\cup\{\boldsymbol{e}_i\}}\exp(c_{ij}^{(l)})},\tag{2}$$

其中 $N_{\boldsymbol{e}_i}\cup\{\boldsymbol{e}_i\}$ 是 $\boldsymbol{e}_i$ 的自环边(self-loop)的邻居节点的集合,$c_{ij}^{(l)}$ 是实体 $\boldsymbol{e}_i$ 到 $\boldsymbol{e}_j$ 的注意力系数.注意力系数 $c_{ij}^{(l)}$ 的计算公式为:

$$c_{ij}^{(l)}=\mathrm{LeakyReLU}(\boldsymbol{q}^{\mathrm{T}}[\boldsymbol{W}^{(l)}\boldsymbol{h}_i^{(l)}\oplus\boldsymbol{W}^{(l)}\boldsymbol{h}_j^{(l)}]).\tag{3}$$

其中:$\boldsymbol{h}_i^{(l)}$ 和 $\boldsymbol{h}_j^{(l)}$($\boldsymbol{h}_i^{(l)},\boldsymbol{h}_j^{(l)}\in\boldsymbol{H}^{(l)}$)分别表示实体 $\boldsymbol{e}_i$ 和 $\boldsymbol{e}_j$ 的隐藏态;$\boldsymbol{q}$($\boldsymbol{q}\in\mathbf{R}^{2d^{(l)}}$)是一个可学习的参数,用于表示神经网络中连接层之间的权重;$\oplus$ 表示对两个向量进行拼接操作.图嵌入层经过 $t+1$ 次迭代更新后,将实体和其邻居节点的特征整合成特征向量.在非线性激活函数 LeakyReLU($x$)(LeakyReLU$(x)=\begin{cases}x, & x\geqslant 0\\ ax, & x<0\end{cases}$)中,$a$ 为$(0,+\infty)$区间内的一个固定参数,本文取 $a=0.2$.在对齐实体时,最小化损失函数 $O_{\mathrm{G}}$ 为:

$$O_{\mathrm{G}}=\sum_{(\boldsymbol{e}_i,\boldsymbol{e}_j)\in S}\sum_{(\boldsymbol{e}_i',\boldsymbol{e}_j')\in S'}[\mathrm{dist}(\boldsymbol{e}_i,\boldsymbol{e}_j)+\gamma_1-\mathrm{dist}(\boldsymbol{e}_i',\boldsymbol{e}_j')]_+,\tag{4}$$

其中 $\mathrm{dist}(\boldsymbol{e}_i,\boldsymbol{e}_j)$($\mathrm{dist}(\boldsymbol{e}_i,\boldsymbol{e}_j)=\boldsymbol{e}_i-\boldsymbol{e}_j$)是两个对齐实体对$(\boldsymbol{e}_i,\boldsymbol{e}_j)$之间的 $L_2$ 范数,$S'$ 表示从样本集合 $S$ 通过最近邻取样(nearest neighbor sampling)生成的负样本对的集合,$\gamma_1$($\gamma_1>0$)是一个边界超参数.

### 1.2 知识嵌入层

基于平移的 TransE 模型是目前知识图谱表示学习模型中的一种经典模型,它将实体和关系映射至同一个低维向量空间,并将实体与实体之间的关系表示为实体向量之间的平移操作.TransE 模型具有算法简单的优点,但该模型在处理自反关系、一对多、多对一和多对多关系时会将完全不同的实体在向量空间中赋予非常相似的向量,进而会降低实体的向量表示效果.为此,学者在 TransE 模型的基础上提出了一种改进模型——TransH 模型.TransH 模型将每个三元组中的关系定义为一个超平面 $W_r$ 和一个关系向量 $\boldsymbol{r}$,并将头实体 $\boldsymbol{e}_{\mathrm{h}}$ 和尾实体 $\boldsymbol{e}_{\mathrm{t}}$ 投影到超平面 $W_r$ 上(以此获得投影 $\boldsymbol{e}_{\mathrm{h}\perp}$ 和 $\boldsymbol{e}_{\mathrm{t}\perp}$),使得与实体对应的向量能够满足一定的线性关系.通过上述方法 TransH 模型即可使同一个实体在相同的关系中具有相同的语义,而在不同的关系中具有不同的语义.

本文采用 TransH 模型作为知识嵌入层的模型,模型训练采用边界排名损失函数(margin ranking loss)作为知识嵌入模型的目标函数.首先将其中一个知识图谱三元组中的头实体 $\boldsymbol{e}_{\mathrm{h}}$ 和尾实体 $\boldsymbol{e}_{\mathrm{t}}$ 映射到超平面上,由此得到映射向量 $\boldsymbol{e}_{\mathrm{h}\perp}$ 和 $\boldsymbol{e}_{\mathrm{t}\perp}$.$\boldsymbol{e}_{\mathrm{h}\perp}$ 和 $\boldsymbol{e}_{\mathrm{t}\perp}$ 的计算公式为:

$$\begin{cases}\boldsymbol{e}_{\mathrm{h}\perp}=\boldsymbol{e}_{\mathrm{h}}-\boldsymbol{e}_{\mathrm{h}w_r}=\boldsymbol{e}_{\mathrm{h}}-w^{\mathrm{T}}\boldsymbol{e}_{\mathrm{h}}w,\\ \boldsymbol{e}_{\mathrm{t}\perp}=\boldsymbol{e}_{\mathrm{t}}-\boldsymbol{e}_{\mathrm{t}w_r}=\boldsymbol{e}_{\mathrm{t}}-w_r^{\mathrm{T}}\boldsymbol{e}_{\mathrm{t}}w_r,\end{cases}\tag{5}$$

其中$\begin{cases}\boldsymbol{e}_{\mathrm{h}w_r}=w^{\mathrm{T}}\boldsymbol{e}_{\mathrm{h}}w\\ \boldsymbol{e}_{\mathrm{t}w_r}=w^{\mathrm{T}}\boldsymbol{e}_{\mathrm{t}}w\end{cases}$.$w^{\mathrm{T}}\boldsymbol{e}_{\mathrm{h}}w$ 中的 $w^{\mathrm{T}}\boldsymbol{e}_{\mathrm{h}}=|w||h|\cos\theta$,其表示 $\boldsymbol{e}_{\mathrm{h}}$ 在 $w_r$ 方向上的投影长度.将 $w^{\mathrm{T}}\boldsymbol{e}_{\mathrm{h}}$ 乘以 $\boldsymbol{e}_{\mathrm{h}}$ 并在 $w_r$ 上投影即可得到 $\boldsymbol{e}_{\mathrm{h}}$ 和 $\boldsymbol{e}_{\mathrm{t}}$ 在超平面上的投影.三元组的评分公式为:

$$f_r(\boldsymbol{e}_{\mathrm{h}},\boldsymbol{e}_{\mathrm{t}})=h-w_r^{\mathrm{T}}\boldsymbol{e}_{\mathrm{h}}w_r+\boldsymbol{r}-\boldsymbol{e}_{\mathrm{t}}+w_r^{\mathrm{T}}\boldsymbol{e}_{\mathrm{t}}w_r.\tag{6}$$

如果三元组关系是正确的,则 $f_r(\boldsymbol{e}_{\mathrm{h}},\boldsymbol{e}_{\mathrm{t}})$ 值较小;如果三元组关系是错误的,则 $f_r(\boldsymbol{e}_{\mathrm{h}},\boldsymbol{e}_{\mathrm{t}})$ 值较大.图嵌入模型的目标函数为:

$$O_{\mathrm{K}}=\sum_{(\boldsymbol{e}_{\mathrm{h}},r,\boldsymbol{e}_{\mathrm{t}})\in T}\sum_{(\boldsymbol{e}_{\mathrm{h}}',r',\boldsymbol{e}_{\mathrm{t}}')\in T'}[\gamma_2+f_r(\boldsymbol{e}_{\mathrm{h}},\boldsymbol{e}_{\mathrm{t}})-f_r(\boldsymbol{e}_{\mathrm{h}}',\boldsymbol{e}_{\mathrm{t}}')]_+,\tag{7}$$

其中 $[\cdot]_+=\max\{0,\cdot\}$,$\boldsymbol{e}_{\mathrm{h}}$ 和 $\boldsymbol{e}_r$ 是嵌入实体(来自更新后的图嵌入模型的嵌入矩阵 $\boldsymbol{H}^{(\mathrm{L})}$),$\boldsymbol{e}_{\mathrm{h}}'$ 和 $\boldsymbol{e}_{\mathrm{t}}'$ 为从负样本三元组集合取样的实体,$\boldsymbol{r}$ 是关系(来自需要被学习的关系矩阵 $\boldsymbol{R}\in\mathbf{R}^{|R|\times d}$),$T'$ 是负三元组样

本的集合，$\gamma_2(\gamma_2>0)$ 是一个边界超参数.

### 1.3　目标函数

本文提出的将图嵌入模型和知识嵌入模型进行融合的跨语言实体对齐模型的目标函数为：

$$O=O_G+O_K, \tag{8}$$

其中 $O_G$ 和 $O_K$ 由式(4)和式(7)给出. 在进行实体对齐推理时，本文模型首先计算每个实体在向量空间中的距离($L_2$)，然后根据计算结果找出新的具有相同语义的实体.

## 2　实验结果与分析

BabelNet[4]是一个可以查询多语言对齐的结构化数据网站. 本文利用 Python 编写了一个爬虫程序，通过结合维基百科的索引数据实现了结构化数据的自动查询与收集. 本文共爬取了 702 645 个三元组(含 13 种语言)作为实验数据来源. 由于本研究仅针对中文和韩文，因此仅对中文和韩文的数据进行了处理. 处理后所得的中韩对齐数据集如表 1 所示. 实验中将数据集中的 70%数据作为训练集，30%数据作为测试集.

**表 1　中韩跨语言对齐数据集的统计数据**

| 数据集 | 数量 |
|---|---|
| 三元组 | 86 934 |
| 实体 | 54 795 |
| 关系 | 1 196 |
| 训练集 | 14 289 |
| 测试集 | 33 344 |

实验条件：服务器的操作系统为 Ubuntu 20.04 LTS，内存为 128 GB，GPU 为 NVIDIA Quadro RTX 5000，CPU 为 Intel Xeon® Gold 6128. 程序采用 Python 3.7 编写，同时采用 Pytorch 1.6 实现数据加载和模型构建. 训练数据时，AdaGrade 的学习率($\lambda$)取 0.01，间隔排序损失函数的参数 $\gamma_1$ 和 $\gamma_2$ 均取 3，迭代次数为 500 次. 模型的对齐效果使用 Hits@$k$ 进行评估. 为筛选出效果最好的模型组合，本文进行了多组对比试验，实验结果如表 2 所示.

**表 2　不同模型的中韩跨语言实体对齐的实验结果**

| 模型 | 韩→中 | | | 中→韩 | | |
|---|---|---|---|---|---|---|
| | Hits@1 | Hits@5 | Hits@10 | Hits@1 | Hits@5 | Hits@10 |
| MTransE | 0.429 4 | 0.683 3 | 0.762 2 | 0.428 2 | 0.675 0 | 0.753 3 |
| GCN-Align | 0.435 9 | 0.708 9 | 0.802 9 | 0.438 7 | 0.707 7 | 0.798 1 |
| TransR+GAT | 0.459 4 | 0.764 3 | 0.892 5 | 0.455 1 | 0.758 7 | 0.891 0 |
| TransD+GAT | 0.463 2 | 0.770 9 | 0.894 2 | 0.457 0 | 0.761 8 | 0.894 0 |
| TransE+GAT | 0.444 3 | 0.742 4 | 0.871 9 | 0.441 1 | 0.739 3 | 0.867 8 |
| RotatE+GAT | 0.477 0 | 0.753 1 | 0.835 0 | 0.475 8 | 0.744 0 | 0.822 6 |
| TransH+GAT | 0.496 2 | 0.808 9 | 0.917 6 | 0.497 9 | 0.807 4 | 0.916 7 |

由表 2 中的结果可知：

1)融合图嵌入和知识嵌入的对齐模型(TransR+GAT、TransD+GAT、TransE+GAT、RotatE+GAT、TransH+GAT)的准确率显著高于基于知识嵌入和基于图嵌入的跨语言实体对齐模型 MTransE[5]和 GCN-Align[6]，其中在 Hits@1 指标上提高了 1.9%～15.6%，在 Hits@5 指标上提高了 3.4%～18.4%，在 Hits@10 指标上提高了 9.6%～20.4%.

2)所有模型中韩文对齐中文实体的准确率均高于中文对齐韩文实体的准确率(约为 1.0%)，其原因是韩文实体比中文实体在文字表示上更具有辨识度，即文字相同但表达含义不同的中文实体多于韩文中的韩文实体.

3)除 TransE+GAT 外，其他融合 GAT 和 Trans 系列模型的 Hits@5 和 Hits@10 均高于 GAT 与 RotatE 融合[5]的模型. 其原因是 GAT 模型与 Trans 系列模型相融合时，GAT 模型的泛化能力优于 GAT 模型与 RotatE 融合的模型，实体和关系得到了更合适的向量表示.

4)TransH+GAT 模型的对齐准确率高于 TransD+GAT 和 TransR+GAT 模型，虽然 TransH

模型的复杂度低于TransD模型[7]和TransR模型[8]的复杂度.其原因是复杂度高的模型在和GAT结合时,会导致不同实体的向量表示存在区分度小的问题,使得许多语义相近的实体被相对密集的表示,进而影响实体对齐的效果.

由于TransH+GAT组合模型的准确率高于其他组合,因此本文选用该组合模型作为最终的跨语言实体对齐方案,模型的损失值如图2所示.由图2可以看出:知识嵌入模型的Loss曲线整体较为平稳,损失值较低;图嵌入模型的Loss曲线随迭代次数的增加呈现平缓下降的趋势,且着随迭代次数的增加模型的Loss曲线与知识嵌入模型的Loss曲线逐渐接近.以上结果表明模型的训练效果较好,能够满足对齐任务的使用.

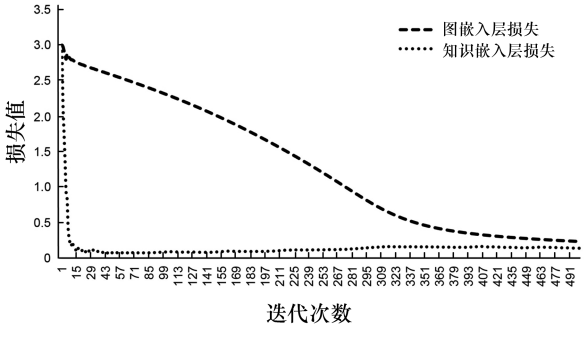


**图2 模型的损失值曲线**

## 3 结论

研究显示,本文提出的融合图嵌入和知识嵌入的中韩双语实体对齐模型的Hits@1、Hits@5和Hits@10,在韩文对齐中文时分别达到了49.62%、80.89%和91.76%,在中文对齐韩文时分别达到49.79%、80.74%和91.67%,且优于传统的基于知识嵌入和图嵌入的对齐方法,因此该模型可为构建中韩对齐知识图谱以及其他语言的对齐知识图谱提供参考.在今后的研究中,我们将对影响图神经网络和知识表示模型效果的因素(如知识表示模型的复杂程度)做进一步的分析,并研究其他图嵌入方法与知识嵌入方法相融合的效果,以探索更为有效的跨语言实体对齐策略.

## 参考文献:

[1] LI C, CAO Y, HOU L, et al. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019:2723-2732.

[2] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Québec: AAAI Press, 2014:1112-1119.

[3] SONG W, XIAO Z, WANG Y, et al. Session-based social recommendation via dynamic graph attention networks [C]//Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. New York: Association for Computing Machinery, 2019:555-563.

[4] NAVIGLI R, PONZETTO P, BABEINET S. Building a very large multilingual semantic network[C]//Proceedings of the 48th Annual Meeting of the Associationfor Computational Linguistics. Uppsala: Association for Computational Linguistics, 2010:216-225.

[5] JI S, PAN S, CAMBRIA E, et al. A survey on knowledge graphs: representation, acquisition, and applications [J]. IEEE T Neural Network, 2021,8(1):1-21.

[6] WANG Z, LV Q, LAN X, et al. Cross-lingual knowledge graph alignment via graph convolutional networks[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: EMNLP, 2018:349-357.

[7] JI G, HE S, XU L, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: Association for Computational Linguistics, 2015:687-696.

[8] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph completion[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Austin, Texas: AAAI Press, 2015:2181-2187.

[9] CHEN M, TIAN Y, YANG M, et al. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne: AAAI Press, 2017:1511-1517.