基于上下文的跨语言知识图谱实体对齐方法

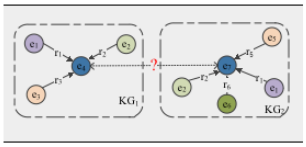
Northeastern University

马新月, 聂铁铮, 申德荣, 寇月

东北大学 计算机科学与工程学院, 沈阳, 中国

背景

- 知识图谱之间存在知识丢失、知识重复、知识关联不明确等问题；
- 传统的实体对齐方法存在训练缓慢、精确度低等问题，有局限性；
- 现有的实体对齐方法未充分考虑实体的**上下文信息**，当中心实体没有传出任何信息时，很难为其匹配到相似实体。



当知识图谱中的实体不存在传出三元组时，无法为其进行有效建模，可以考虑传入关系来提升实体嵌入效果。

引言

面向知识图谱的实体对齐技术可以实现知识的连接，将同类知识图谱融合为规模更大、质量更权威的领域知识图谱。**实体对齐的目标**是识别出不同知识图谱中指代对象为现实世界中同一事物的实体对

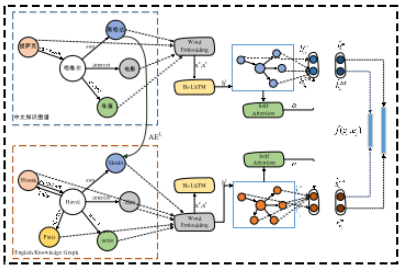
实体对齐方法有传统的实体对齐方法, 以基于知识表示学习的实体对齐方法。

- 基于**翻译**的表示学习方法能够在低维稠密空间中学习实体的向量表示。
- 基于**图卷积神经网络**的嵌入方法在三元组的维度上结合节点的属性信息进行复杂信息建模。

为解决多语言知识图谱实体对齐的精确度问题，在实体嵌入的过程中充分考虑实体的**上下文信息**以**关系的指向信息**，提出基于上下文的知识图谱实体对齐方法。

方法

本文的基于上下文的跨语言实体对齐方法，通过中心实体的上下文信息来解决前期工作的缺陷，总体框架如下图所示：



*上图为知识图谱实体对齐方法总体框架，首先根据**词嵌入**来分别对两个知识图谱中实体以关系信息进行语义编码，利用Bi-LSTM得到节点以关系的初始特征表示，接着利用**图注意力机制**进一步处理实体的嵌入表示，捕获每个三元组中的**内部依赖性**以邻居节点对中心实体嵌入表示的**重要程度**，最后利用基于距离的度量函数进行**实体对齐**。*

特征提取层

通过基于词的Bi-LSTM来学习三元组中的实体以关系的嵌入表示，将深层次的语义特征向量表示作为实体的初始特征向量表示。

实体嵌入层

结合**图注意力机制**对知识图谱中的中心实体的上下文建模，最终的嵌入表示结合了所有中心实体相关的邻居节点信息以关系信息，添加超参数β来平衡邻居类别(传入/传出)的重要程度。

$$\tilde{h}_{i_k} = \mathrm{ReLU}\Big(\sum_{j \in \{i_k \cup N_{i_k}\}} \alpha_{ij} W_1^{in} h_j^{in} + \sum_{j \in N_{i_k}} \alpha_{ij} W_1^{out} h_j^{out}\Big)$$

$$h_{i_k}^{in} = (1-\beta)\,\|_{p=1}^{|\mathcal{R}|} \sum_{j \in N_{i_k}^{in}(p)} \alpha_{ij} W_2^{in} h_j^{in} + \beta \sum_{j \in N_{i_k}^{in}} h_j^{in}$$

$$h_{i_k}^{out} = \|_{p=1}^{|\mathcal{R}|} \sum_{j \in N_{i_k}^{out}(p)} \alpha_{ij} W_2^{out} h_j^{out} + \beta \sum_{j \in N_{i_k}^{out}} h_j^{out}$$

$$e_{i_k}^{in} = \sigma\Big[\,\|_{p=1}^{|\mathcal{R}|} h_{i_k}^{in} \|\, \frac{1}{|N_{i_k}^{in}|} \sum_{j \in N_{i_k}^{in}} h_j \,\Big]$$

$$e_{i_k}^{out} = \sigma\Big[\,\|_{p=1}^{|\mathcal{R}|} h_{i_k}^{out} \|\, \frac{1}{|N_{i_k}^{out}|} \sum_{j \in N_{i_k}^{out}} h_j \,\Big]$$

$$\alpha_{ij} = \mathrm{softmax}_j(e_{ij}) = \frac{\exp(\mathrm{LeakyRelu}(a_{ij}))}{\sum_{j \in N_i} \exp(\mathrm{LeakyRelu}(a_{ij}))}$$

$$\alpha_{ij} = \mathrm{softmax}_j(e_{ij}) = \frac{\exp(\mathrm{LeakyRelu}(a_{ij}))}{\sum_{j \in N_i} \exp(\mathrm{LeakyRelu}(a_{ij}))}$$

实体对齐

度量实体对之间的距离，从距离排序列表中选择候选匹配实体，距离越近的实体越相似，在候选匹配列表中的排名越靠前。

$$D(e_1, e_2) = \delta\big(h_{e_1}^{in}, h_{e_2}^{in}\big) + \theta\big(h_{e_1}^{out}, h_{e_2}^{out}\big)$$

模型训练

利用基于距离排序损失函数，使两个知识图谱中相似的实体对在向量空间中的距离尽可能地接近。

$$L_1 = \sum_{(e_1,e_2) \in L_{train}} \sum_{(e_1',e_2') \in L_{e_1}^{-}} \Big[\gamma_1 + d\big(h_{e_1}^{in}, h_{e_2}^{in}\big) - \gamma_1^{-} d\big(h_{e_1'}^{in}, h_{e_2'}^{in}\big)\Big]_{+}$$

$$L_2 = \sum_{(e_1,e_2) \in L_{train}} \sum_{(e_1',e_2') \in L_{e_1}^{-}} \Big[\gamma_2 + d\big(h_{e_1}^{out}, h_{e_2}^{out}\big) - \gamma_2^{-} d\big(h_{e_1'}^{out}, h_{e_2'}^{out}\big)\Big]_{+}$$

结果

实验对比结果如下表所示，粗体数字代表最佳性能。

	DBP15K$_{FR-EN}$			DBP15K$_{ZH-EN}$			DBP15K$_{JA-EN}$		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
MTransE	30.8	61.4	0.214	27.9	57.5	0.349	24.9	49.4	0.335
JAPE	41.3	74.5	0.490	39.3	78.5	0.476	32.4	66.7	0.430
GCN-Align	41.3	72.1	0.549	39.91	74.5	0.546	37.3	69.5	0.532
KGEA	**56.4**	**72.4**	**0.613**	**64.6**	**75.5**	**0.682**	**79.6**	**86.8**	**0.826**

本文的方法KEGA与对比方法MTransE、JAPE以GCN-Align在DBP15K数据集上的实体对齐结果

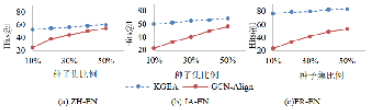
可以看出本文的实体对齐方法KEGA对比之前的方法有一定的提升，这表明将Bi-LSTM以注意力机制相结合可以挖掘出更多元组间的潜在关联信息。

	DBP-WD			DBP-YG		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
MTransE	28.1	52.6	0.363	25.2	49.3	0.334
JAPE	31.8	58.9	0.411	23.6	48.4	0.320
GCN-Align	53.6	75.7	0.600	59.7	77.8	0.683
KGEA	65.1	77.5	**0.695**	68.9	80.6	0.726

本文的方法KEGA与对比方法MTransE、JAPE以GCN-Align在DWY100K数据集上的实体对齐结果

使用不同比例的先验种子集作为训练数据来验证种子集数目对实体对齐结果的影响，对比实验结果如下图所示。

从图中可以看出，在所有的数据集中，KGEA的性能均为最优，KGEA利用了详细的实体关联信息进行嵌入表示，即使种子集比例较小，也能取得良好的效果。



不同的种子集比例对本文方法KGEA对齐效果的影响，其中x轴是用于对齐的种子集的比例，y轴是Hits@1分数

结论

- 本文提出了一种新的基于上下文的知识图谱实体对齐方法，通过探索中心实体周围的复杂关联信息来学习实体嵌入。
- 利用Bi-LSTM模型以图注意力机制构造包含邻居的关系信息以结构信息，更好得学习实体的向量表示来实现实体对齐。
- 实验结果表明本文的方法在三个多语言知识图谱数据集以两个大规模数据集中均取得了最好的实体对齐效果。

致谢

本文得到国家重点研发计划项目(No. 2018YFB1003404)、国家自然科学基金(No. U1811261, 61672142, 61472070)、国防基础科研计划(JCKY2018205C012)资助。