知识图谱中的实体对齐方法

郭晨亮

2020.6.22

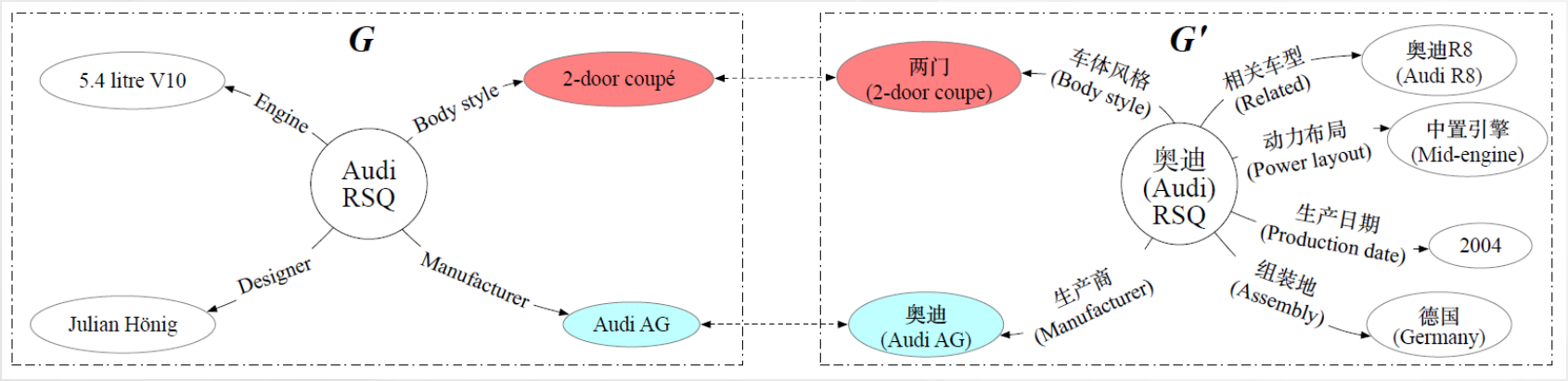
知识图谱的实体对齐任务

- 知识图谱：实体构成点，实体间的关系构成边，即(实体,边,实体)的关系三元组，实体具有属性，即(实体,属性,值)的属性三元组。
- 知识图谱的实体对齐任务：合并两个知识图谱，找到实体在不同图中的对应关系，已知一组预先匹配的实体对应。
- 分析：

(1) 实体表述的不同形式、不同语言应具有相似的语义。

(2) 现有知识图谱中，大多数实体关系很少，只有少量实体关系很多，存在长尾实体。

(3) 在不同知识图谱之间，通常存在较大的结构异质性，同一实体的属性、连接的其它实体只有部分相同。



常用方法

- 1. 两个不同的知识图谱KG
- 2. 嵌入模型：构造KG中点或边的嵌入向量表示
  - (1) 对实体点的嵌入表示通常用到关系(结构)信息或属性信息

    结构信息：三元组、相邻关系、路径关系

    属性信息：属性和值的类型、值的文本
  - (2) 负样本产生：随机选择、选择最接近正样本的K组数据
  - (3) 嵌入表示获取方式：三元组学习、GNN、对抗学习、多视图学习

常用方法

- 3. 对齐模型：组合两个KG中的向量表示，计算实体间相似度
  - (1) 组合方式

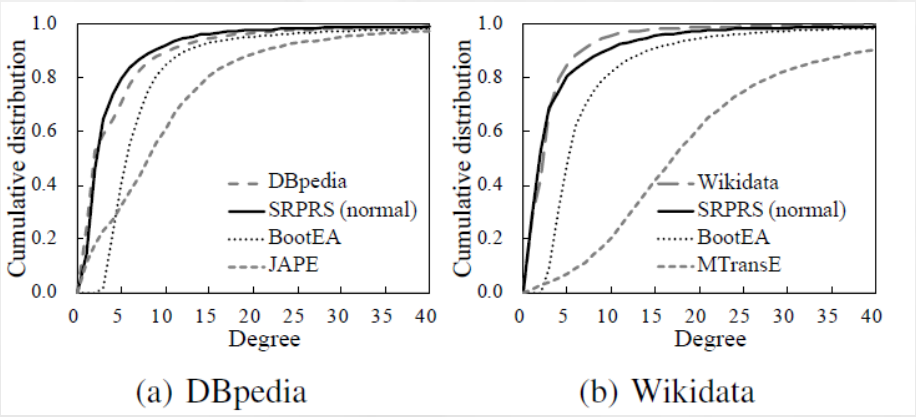
    两个KG嵌入到不同空间：嵌入空间变换

    两个KG嵌入到同一空间：嵌入空间校准、参数共享、参数交换
  - (2) 迭代更新
  - (3) 向量距离度量：余弦、欧几里得距离、曼哈顿距离
- 4. 根据相似度对齐实体
  - 对齐方式：直接寻找最相似的实体、综合考虑双向关系、二分图匹配

常用方法

5. 常用数据集

- DBP15K、DBP100K、DWY100K、SRPRS、WK3L等
- 没有统一规范的数据集，模型效果只能进行相对比较
- 有些数据集的数据分布与现实数据分布相差较大，过于密集，更容易完成对齐任务



(a) DBpedia (b) Wikidata

	#Rel. triples	#Entities	Deg.
DBpedia (EN)	17,788,314	5,136,559	6.93
DBP15K	101,058	14,984	13.49
WK3L	95,096	8,351	22.77

	#Rel. triples	#Entities	Deg.
DBpedia (FR)	4,873,571	1,794,708	5.43
DBP15K	93,016	14,991	12.41
WK3L	68,787	7,180	19.16

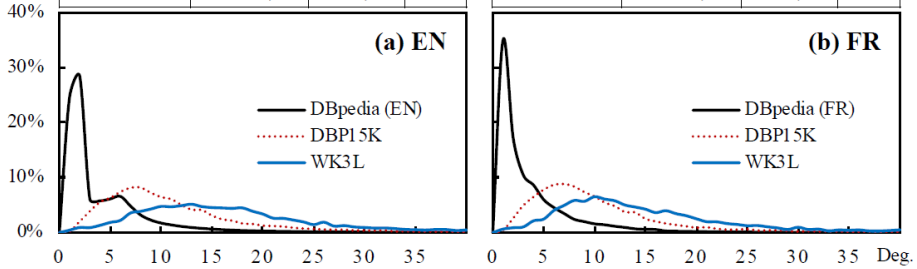


Figure 2: Degree distributions and average degrees of two popular datasets DBP15K [80] and WK3L [10] used in current approaches. Both are extracted from DBpedia [43], but their degree distributions are quite different from DBpedia and their average degrees are also much larger.

JE(Joint Embedding Method)

- TransE：对于h到t的关系l，h+l≈t，用d(h+l,t)和d(h,t-l)测试

$$\mathcal{L} = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} \left[\gamma + d(\boldsymbol{h}+\boldsymbol{\ell},\boldsymbol{t}) - d(\boldsymbol{h'}+\boldsymbol{\ell},\boldsymbol{t'})\right]_+$$

$$S'_{(h,\ell,t)} = \left\{(h',\ell,t)|h' \in E\right\} \cup \left\{(h,\ell,t')|t' \in E\right\}.$$

$$L = \sum_{(h,r,t)\in S} \sum_{(h',r,t')\in S'_{(h,r,t)}} \{[\gamma + d(h+r,t) - d(h'+r,t')]_+ +$$

$$\lambda_1 \sum_{y\in\{h,h',r,t,t'\}} |\|y\|_2 - 1|\} + \lambda_2 \sum_{(e_i,e'_i)\in A} \|M_d e_i - e'_i\|_2$$

A joint embedding method for entity alignment of knowledge bases. CCKS2016.

Translating embeddings for modeling multirelational data. In Advances in neural information processing systems, 2013.

MTransE(Multilingual Embedding)

- 知识模型(Knowledge Model)

对于语言为(Li,Lj)的两个知识图谱，共同训练关于ij的模型

$$S_K = \sum_{L \in \{L_i, L_j\}} \sum_{(h,r,t) \in G_L} \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$$

$$J = S_K + \alpha S_A$$

- 对齐模型(Alignment Model)

$$S_A = \sum_{(T,T') \in \delta(L_i, L_j)} S_a(T, T')$$

- Distance-based Axis Calibration

$$S_{a_1} = \|\mathbf{h} - \mathbf{h}'\| + \|\mathbf{t} - \mathbf{t}'\| \qquad S_{a_2} = \|\mathbf{h} - \mathbf{h}'\| + \|\mathbf{r} - \mathbf{r}'\| + \|\mathbf{t} - \mathbf{t}'\|$$

- Translation Vectors

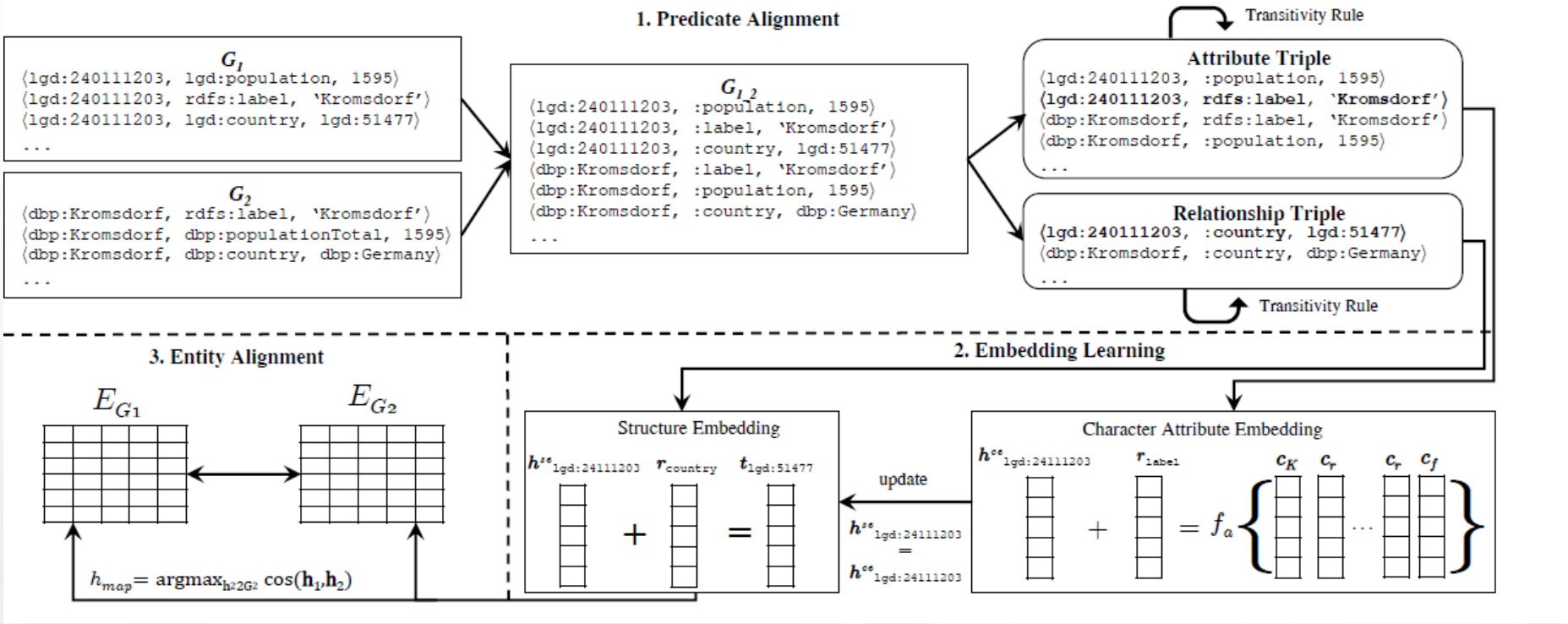
$$S_{a_3} = \left\|\mathbf{h} + \mathbf{v}_{ij}^e - \mathbf{h}'\right\| + \left\|\mathbf{r} + \mathbf{v}_{ij}^r - \mathbf{r}'\right\| + \left\|\mathbf{t} + \mathbf{v}_{ij}^e - \mathbf{t}'\right\|$$

- Linear Transformations

$$S_{a_4} = \left\|\mathbf{M}_{ij}^e \mathbf{h} - \mathbf{h}'\right\| + \left\|\mathbf{M}_{ij}^e \mathbf{t} - \mathbf{t}'\right\| \qquad S_{a_5} = \left\|\mathbf{M}_{ij}^e \mathbf{h} - \mathbf{h}'\right\| + \left\|\mathbf{M}_{ij}^r \mathbf{r} - \mathbf{r}'\right\| + \left\|\mathbf{M}_{ij}^e \mathbf{t} - \mathbf{t}'\right\|$$

Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. IJCAI2017.

AttrE



- 三元组，属性文本，余弦，参数共享
- 重点是属性，不针对跨语言3
- 首先用谓词对齐寻找相同的实体，比较编辑距离阈值95%
- 学习结构嵌入，不同的是按关系比例加权
- 学习属性嵌入，对属性值编码，字嵌入和/LSTM/n-gram合成函数，类似结构学习
- 联合学习，使对齐实体的属性嵌入与结构嵌入余弦相似度最大

AAAI 2019 Entity Alignment between Knowledge Graphs Using Attribute Embeddings

RSN4EA

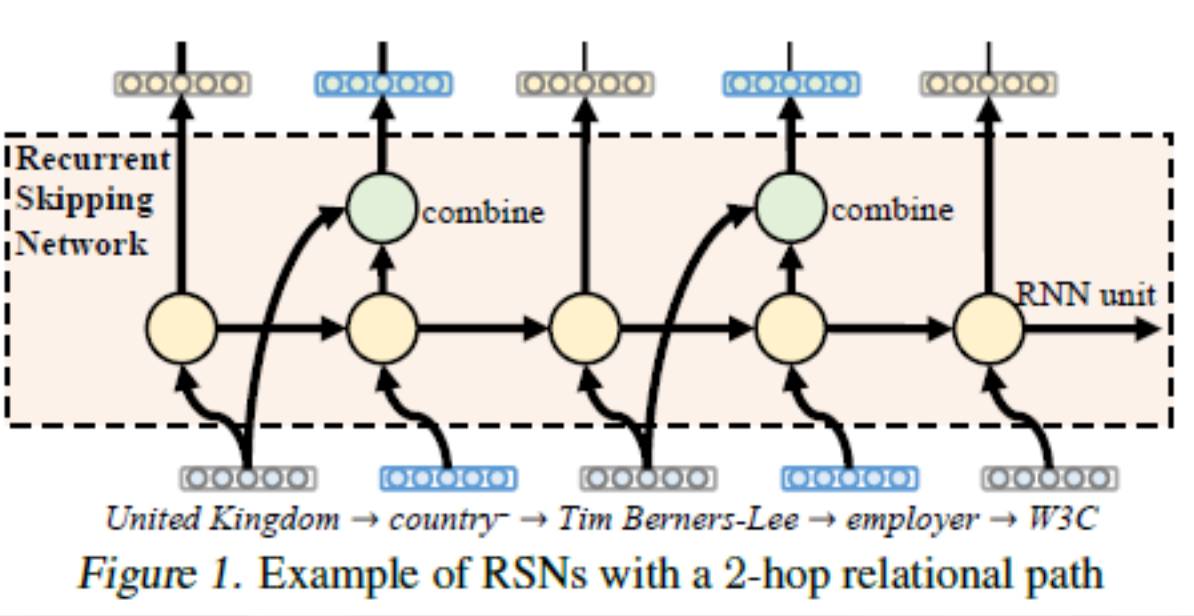
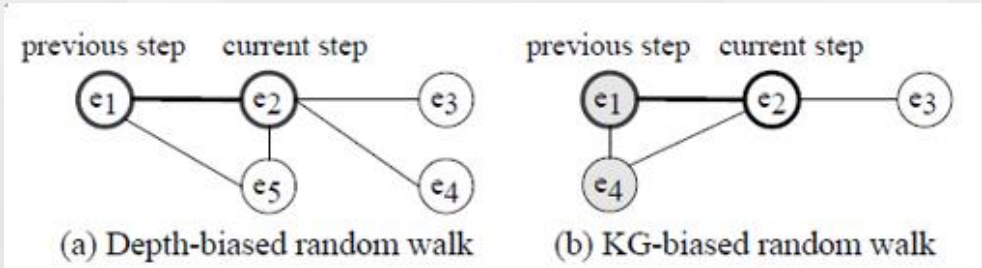
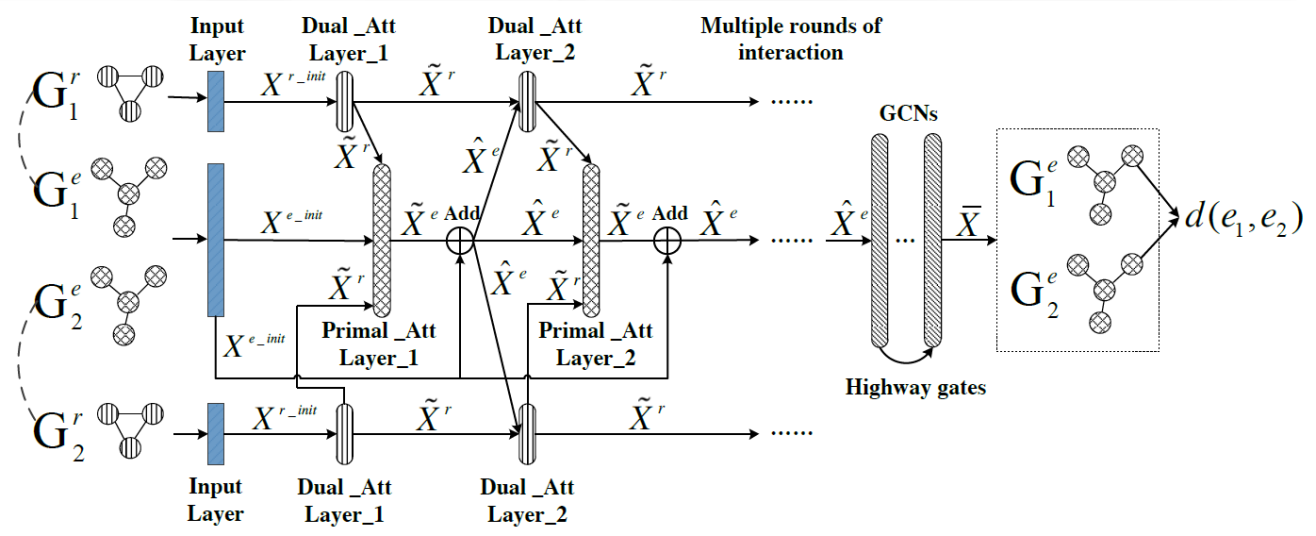


Figure 1. Example of RSNs with a 2-hop relational path



- 偏向随机游走抽样会产生深层和跨KG的关系路径
- 递归跳过网络模型关系路径以学习KG嵌入
- 基于类型的噪声对比NCE估计以优化的方式评估RSN的损失

RDGCN



首先对知识图谱建立对偶图，点改为边，边改为点

对偶图中边权值用两种关系的实体头尾交并比之和表示

$$w_{ij} = \frac{|H_i \cap H_j|}{|H_i \cup H_j|} + \frac{|T_i \cap T_j|}{|T_i \cup T_j|}$$

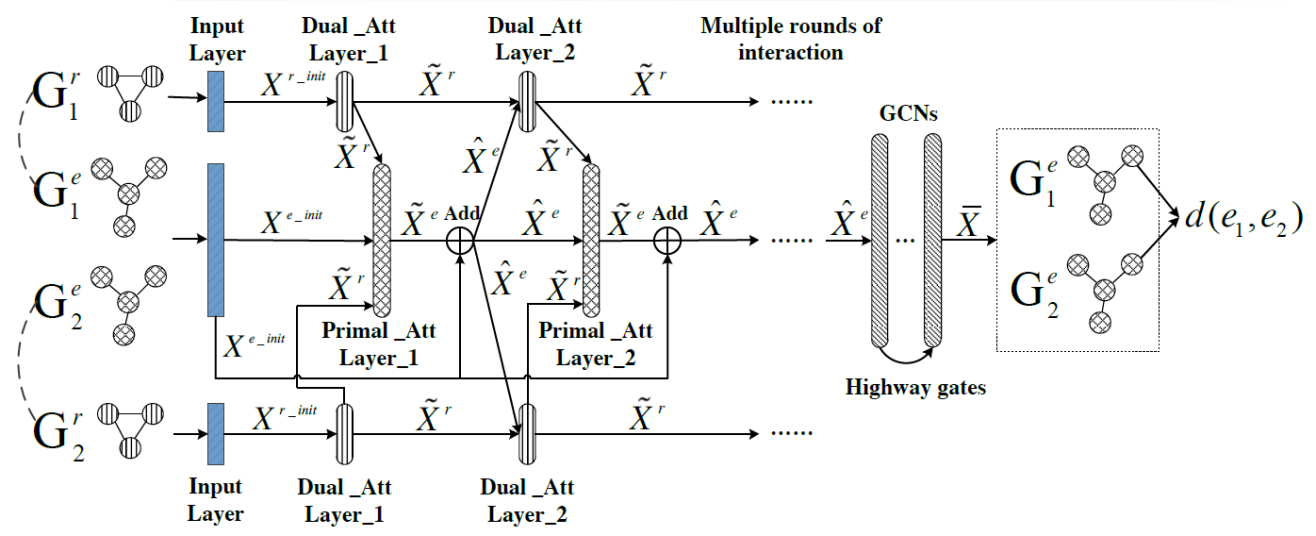
原图边/对偶图点表示为关系头实体向量均值和尾的拼接

$$c_i = [\frac{\sum_{k \in H_i} x_k}{|H_i|} || \frac{\sum_{k \in T_i} x_k}{|T_i|}]$$

原图点表示用点名称和属性的文本，结合翻译，用预训练词向量初始化

IJCAI 2019 Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs

RDGCN



对偶图GAT更新

$$\tilde{X}_i = \sigma\left(\sum_{j \in N_i^r} \alpha_{ij}^r X_j\right), \quad \alpha_{ij}^r = \frac{exp(\eta(w_{ij} a[c_i || c_j]))}{\sum_{k \in N_i^r} exp(\eta(w_{ik} a[c_i || c_k]))}$$

原图中GAT更新

$$\tilde{x}_i = \sigma^e\left(\sum_{j \in N_i^e} \alpha_{ij}^e x_j\right), \quad \alpha_{ij}^e = \frac{exp(\eta(a^e(X_{ij})))}{\sum_{k \in N_i^r} exp(\eta(a^e(X_{ik})))}$$

将GAT前后表示加权求和后通过两层HGCN层得到最终的实体向量表示

IJCAI 2019 Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs

HMAN

1.多视图特征表示

(1)拓扑特征：初始为单位阵，GCN学习

(2)关系特征：统计每种关系数量

(3)属性特征：类似关系

2.BERT文本特征

pointwise:[CLS]实体1[SEP]实体2[SEP]

计算量太大

pairwise:通过比较[CLS]实体[SEP]的结果计算相似度

3.组合

(1)由GCN生成top候选对，由BERT进行评价，受MAN或HMAN的效果影响

(2)加权组合

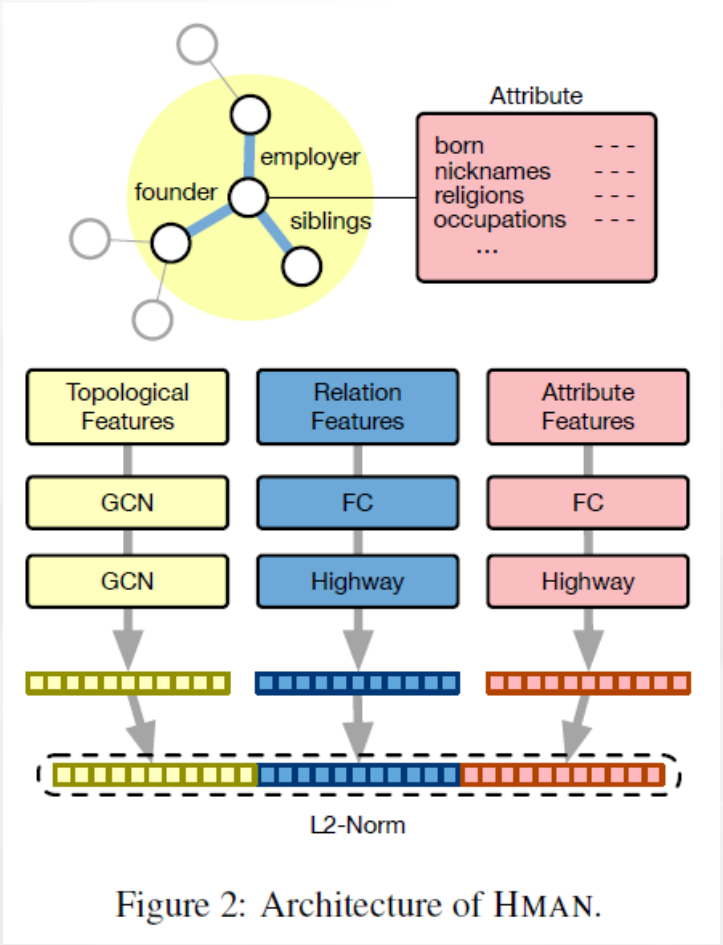


Figure 2: Architecture of HMAN.

HMAN

结论：

综合利用不同特征有效

拓扑结构具有重要作用

HMAN替换部分GCN有效

BERT的加权比候选效果明显更好

Model	ZH → EN			EN→ ZH			JA → EN			EN→ JA			FR → EN			EN→ FR		
	@1	@10	@50	@1	@10	@50	@1	@10	@50	@1	@10	@50	@1	@10	@50	@1	@10	@50
	DBP15K																	
Hao et al. (2016)	21.2	42.7	56.7	19.5	39.3	53.2	18.9	39.9	54.2	17.8	38.4	52.4	15.3	38.8	56.5	14.6	37.2	54.0
Chen et al. (2017a)	30.8	61.4	79.1	24.7	52.4	70.4	27.8	57.4	75.9	23.7	49.9	67.9	24.4	55.5	74.4	21.2	50.6	69.9
Sun et al. (2017)	41.1	74.4	88.9	40.1	71.0	86.1	36.2	68.5	85.3	38.3	67.2	82.6	32.3	66.6	83.1	32.9	65.9	82.3
Wang et al. (2018)	41.2	74.3	86.2	36.4	69.9	82.4	39.9	74.4	86.1	38.4	71.8	83.7	37.2	74.4	86.7	36.7	73.0	86.3
MAN	46.0	79.4	90.0	41.5	75.6	88.3	44.6	78.8	90.0	43.0	77.1	88.7	43.1	79.7	91.7	42.1	79.1	90.9
MAN w/o TE	21.5	55.0	79.4	20.2	53.6	78.8	15.0	44.0	69.9	14.3	44.0	70.6	10.2	34.5	59.5	10.8	35.2	60.3
MAN w/o RE	45.6	79.1	89.5	41.1	75.0	87.3	44.2	78.7	89.8	43.0	76.9	88.1	42.8	79.7	91.4	42.1	78.9	90.6
MAN w/o AE	43.7	77.1	87.8	39.2	72.9	85.5	43.2	77.6	88.4	41.2	74.9	86.6	42.9	79.6	91.0	41.5	78.9	90.5
HMAN	**56.2**	**85.1**	**93.4**	**53.7**	**83.4**	**92.5**	**56.7**	**86.9**	**94.5**	**56.5**	**86.6**	**94.6**	**54.0**	**87.1**	**95.0**	**54.3**	**86.7**	**95.1**
HMAN w/o TE	3.2	16.7	38.3	3.5	17.2	38.5	5.4	22.3	45.5	5.2	22.0	45.5	2.4	13.9	35.3	2.2	13.7	35.3
HMAN w/o RE	50.2	78.4	86.5	49.3	78.6	87.0	52.6	81.6	89.1	52.4	81.1	89.8	52.7	84.2	91.4	52.0	83.9	91.1
HMAN w/o AE	49.2	81.0	89.8	48.8	80.9	90.0	52.2	83.3	91.6	51.5	83.1	91.6	52.3	85.6	93.7	52.3	85.1	93.2
HMAN w/o HW	46.8	76.1	84.1	46.0	76.2	84.6	50.5	79.5	87.5	49.9	79.1	87.5	51.9	82.7	90.9	51.6	82.5	90.6

EMNLP 2019 Aligning Cross-Lingual Entities with Multi-Aspect Information

BERT-INT

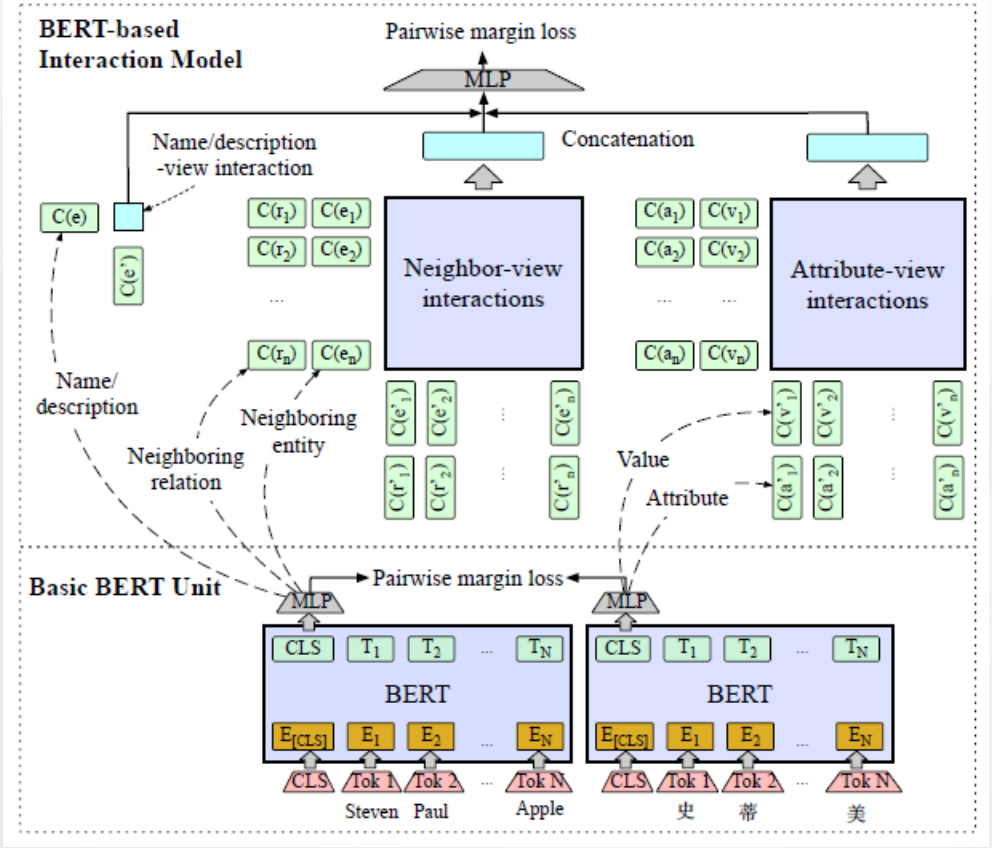
- 使用邻居关系、属性文本、余弦、嵌入空间校准、多视图学习、使用多语言的预训练BERT模型

1.基础BERT嵌入单元

选择最接近正样本的K组数据

$$C(e) = \text{MLP}(\text{CLS}(e)),$$

$$\mathcal{L} = \sum_{(e,e'^{+},e'^{-})\in\mathcal{D}} \max\{0, g(e,e'^{+}) - g(e,e'^{-}) + m)\},$$



IJCAI 2020 BERT-INT: A BERT-based Interaction Model For Knowledge Graph Alignment

BERT-INT

1.基础BERT嵌入单元

(1)名称视图：名称/描述的余弦相似度

(2)邻居视图：计算两个实体相邻结点名称/描述余弦相似度的矩阵

由于邻居无序，不适用CNN或RNN，使用RBF内核聚集函数提取有关相似性累积的特征，分别按行列最大池化后通过RBF层得到相似度并拼接

$$
\begin{aligned}
s_i^{max} &= \max_{j=0}^{n}\{s_{i0},\cdots,s_{ij},\cdots,s_{in}\} \qquad (3)\\
K_l(s_i^{max}) &= \exp\left[-\frac{(s_i^{max}-\mu_l)^2}{2\sigma_l^2}\right]\\
\mathbf{K}^r(\mathbf{S}_i) &= [K_1(s_i^{max}),\cdots,K_l(s_i^{max}),\cdots,K_L(s_i^{max})]\\
\phi^r(\mathcal{N}(e),\mathcal{N}(e')) &= \frac{1}{|\mathcal{N}(e)|}\sum_{i=1}^{|\mathcal{N}(e)|}\log\mathbf{K}^r(\mathbf{S}_i),
\end{aligned}
$$

(3)属性视图：用类似邻居的方式计算，忽略了邻居的属性

最终将所有相似度拼接，通过MLP计算综合相似度

IJCAI 2020 BERT-INT: A BERT-based Interaction Model For Knowledge Graph Alignment

BERT-INT

实验效果

Model	$DBP15K_{ZH-EN}$			$DBP15K_{JA-EN}$			$DBP15K_{FR-EN}$		
	HR1	HR10	MRR	HR1	HR10	MRR	HR1	HR10	MRR
Only use graph structures by variant TransE									
MTransE [Chen *et al.*, 2017]	0.308	0.614	0.364	0.279	0.575	0.349	0.244	0.556	0.335
IPTransE [Zhu *et al.*, 2017]	0.406	0.735	0.516	0.367	0.693	0.474	0.333	0.685	0.451
BootEA [Sun *et al.*, 2018]	0.629	0.848	0.703	0.622	0.854	0.701	0.653	0.874	0.731
RSNs [Guo *et al.*, 2019]	0.508	0.745	0.591	0.507	0.737	0.590	0.516	0.768	0.605
TransEdge [Sun *et al.*, 2019]	0.735	0.919	0.801	0.719	0.932	0.795	0.710	0.941	0.796
MRPEA [Shi and Xiao, 2019]	0.681	0.867	0.748	0.655	0.859	0.727	0.677	0.890	0.755
Only use graph structures by variant TransE plus GCN									
MuGNN [Cao *et al.*, 2019]	0.494	0.844	0.611	0.501	0.857	0.621	0.495	0.870	0.621
NAEA [Zhu *et al.*, 2019]	0.650	0.867	0.720	0.641	0.873	0.718	0.673	0.894	0.752
KECG [Li *et al.*, 2019]	0.478	0.835	0.598	0.490	0.844	0.610	0.486	0.851	0.610
AliNet [Sun *et al.*, 2020]	0.539	0.826	0.628	0.549	0.831	0.645	0.552	0.852	0.657
Only use graph structures by variant TransE plus adversarial learning									
AKE [Lin *et al.*, 2019]	0.325	0.703	0.449	0.259	0.663	0.390	0.287	0.681	0.416
SEA [Pei *et al.*, 2019]	0.424	0.796	0.548	0.385	0.783	0.518	0.400	0.797	0.533
Combine graph structures and side information by variant GCN									
GCN-Align [Wang *et al.*, 2018]	0.413	0.744	0.549	0.399	0.745	0.546	0.373	0.745	0.532
GM-Align [Xu *et al.*, 2019]	0.679	0.785	-	0.740	0.872	-	0.894	0.952	-
RDGCN [Wu *et al.*, 2019a]	0.708	0.846	0.746	0.767	0.895	0.812	0.886	0.957	0.911
HGCN [Wu *et al.*, 2019b]	0.720	0.857	0.768	0.766	0.897	0.813	0.892	0.961	0.917
DGMC [Fey *et al.*, 2020]	0.772	0.897	-	0.774	0.907	-	0.891	0.967	-
Combine graph structures and side information by multi-view learning									
JAPE [Sun *et al.*, 2017]	0.412	0.745	0.490	0.363	0.685	0.476	0.324	0.667	0.430
MultiKE [Zhang *et al.*, 2019]	0.509	0.576	0.532	0.393	0.489	0.426	0.639	0.712	0.665
JarKA [Chen *et al.*, 2020]	0.706	0.878	0.766	0.646	0.855	0.708	0.704	0.888	0.768
HMAN [Yang *et al.*, 2019]	0.871	0.987	-	0.935	0.994	-	0.973	0.998	-
CEAFF [Zeng *et al.*, 2020]	0.795	-	-	0.860	-	-	0.964	-	-
BERT-INT	**0.968**	**0.990**	**0.977**	**0.964**	**0.991**	**0.975**	**0.992**	**0.998**	**0.995**

IJCAI 2020 BERT-INT: A BERT-based Interaction Model For Knowledge Graph Alignment

BERT-INT

实验效果

邻居关系掩码矩阵：与邻居点相似度矩阵相乘，用关系相关头尾实体均值拼接表示关系

多条邻居间的交互：对距离为1-1,1-m,m-1,m-m的情况建立上述的相似度矩阵，拼接所有最终结果。

名称/描述信息更有用

BERT对效果提升由帮助

BERT+GCN效果不好

可迁移

多视图的关系应分别计算

Model	$DBP15K_{ZH-EN}$			$DBP15K_{JA-EN}$			$DBP15K_{FR-EN}$		
	HR1	HR10	MRR	HR1	HR10	MRR	HR1	HR10	MRR
BERT-INT	**0.968**	**0.990**	**0.977**	**0.964**	**0.991**	**0.975**	**0.992**	**0.998**	**0.995**
Remove components									
-max pooling	0.962	0.989	0.973	0.959	0.991	0.973	0.992	0.998	0.995
-column aggregation	0.960	0.989	0.971	0.959	0.990	0.971	0.991	0.998	0.994
-neighbors	0.947	0.987	0.963	0.937	0.986	0.956	0.988	0.998	0.992
-attributes	0.919	0.984	0.945	0.938	0.987	0.957	0.983	0.998	0.990
-neighbors & attributes	0.830	0.970	0.883	0.848	0.974	0.897	0.965	0.995	0.978
Change the interaction component to variant GCN									
BERT-GCN	0.736	0.950	0.799	0.767	0.960	0.824	0.914	0.992	0.936
BERT-RDGCN	0.847	0.974	0.896	0.857	0.969	0.900	0.952	0.990	0.967
BERT-HMAN	0.911	0.993	0.943	0.937	0.994	0.960	0.982	0.999	0.989
Add components									
+relation mask	0.966	0.989	0.975	0.962	0.990	0.973	0.992	0.998	0.995
+attribute mask	0.942	0.986	0.959	0.950	0.990	0.966	0.989	0.998	0.993
+2-hop neighbors	0.965	0.990	0.975	0.964	0.991	0.975	0.992	0.998	0.995

IJCAI 2020 BERT-INT: A BERT-based Interaction Model For Knowledge Graph Alignment