# 面向异构信息网络对齐的神经网络模型研究

吕青松

指导教师：王志春

北京师范大学信息科学与技术学院

2019 年 5 月

# 目录


1. 研究动机
   - 异构信息网络
   - 异构信息网络对齐
   - 相关工作
2. 模型介绍
   - 图卷积神经网络
   - 基于图卷积神经网络的对齐模型
3. 实验
   - 数据集和评测方法
   - 实验结果
4. 总结

# 异构信息网络

异构信息网络 $H = (\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T}_R, \mathcal{T}_A)$

- $\mathcal{E}$：实体集合
- $\mathcal{R}$：关系集合
- $\mathcal{A}$：属性集合
- $\mathcal{V}$：属性值集合
- $\mathcal{T}_R \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$：关系三元组集合
- $\mathcal{T}_A \subseteq \mathcal{E} \times \mathcal{A} \times \mathcal{V}$：属性三元组集合

以知识图谱为例，

- 关系三元组：（爱因斯坦，毕业于，苏黎世联邦理工学院）
- 属性三元组：（爱因斯坦，出生年份，1879）

# 异构信息网络对齐

给定:

- 两个信息网络 $H_1$ 和 $H_2$
- 匹配实体集合 $\mathcal{S} = \{(e_1, e_2)|e_1 \in \mathcal{E}_1 \wedge e_2 \in \mathcal{E}_2\}$

输出:

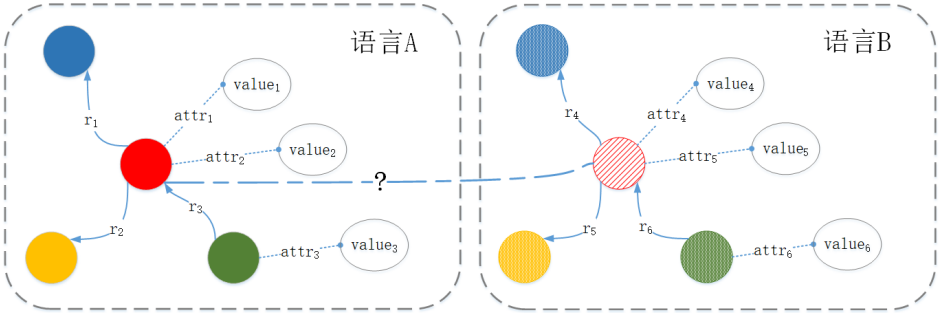
- 新的匹配实体集合 $\mathcal{S}' = \{(e_1, e_2)|e_1 \in \mathcal{E}_1 \wedge e_2 \in \mathcal{E}_2 \wedge (e_1, e_2) \notin S\}$

本质:

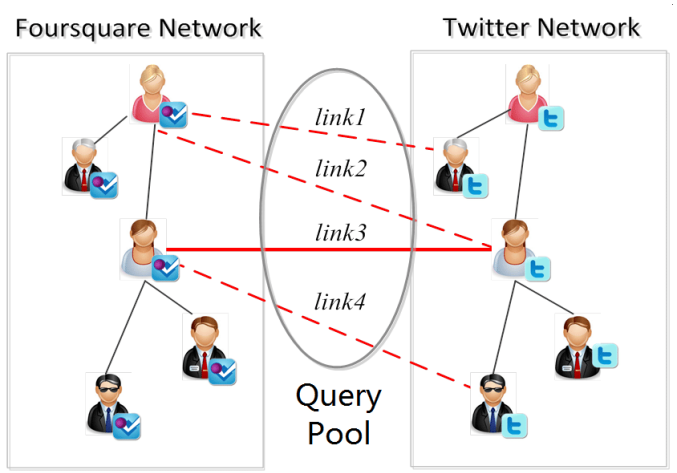
- 发现潜在的匹配实体

# 相关工作

## 跨语言知识图谱匹配




## 跨平台社交网络匹配



# 传统方法

| | 跨语言知识图谱匹配 | 跨平台社交网络匹配 |
|---|---|---|
| 基于实体名称 | 机器翻译 [2, 8] | 用户昵称 [5] |
| 基于人工特征 | 实体类别、邻居相似度等 [10] | 时间、地点、公共邻居等 [4] |

## 缺点

- 基于机器翻译，依赖于翻译的效果
- 基于用户昵称，存在重名、匿名和多用户名的问题
- 基于人工特征，需要人工仔细设计特征，且无法迁移使用

# 基于嵌入表示的方法

## 跨语言知识图谱匹配

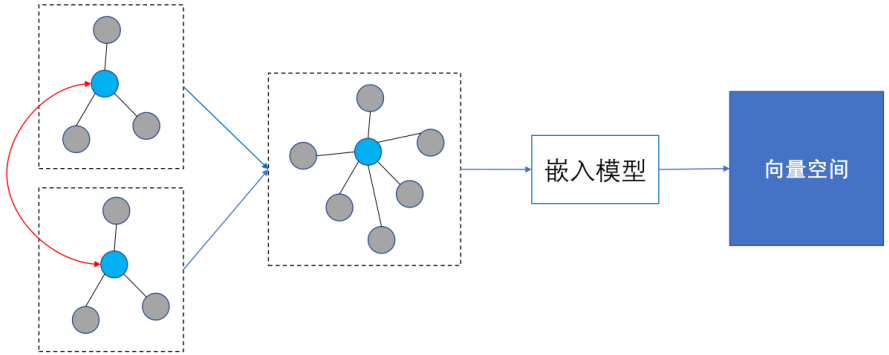
- JE [3]
- MTansE [1]
- JAPE [9]
- ITransE [11]

## 跨平台社交网络匹配

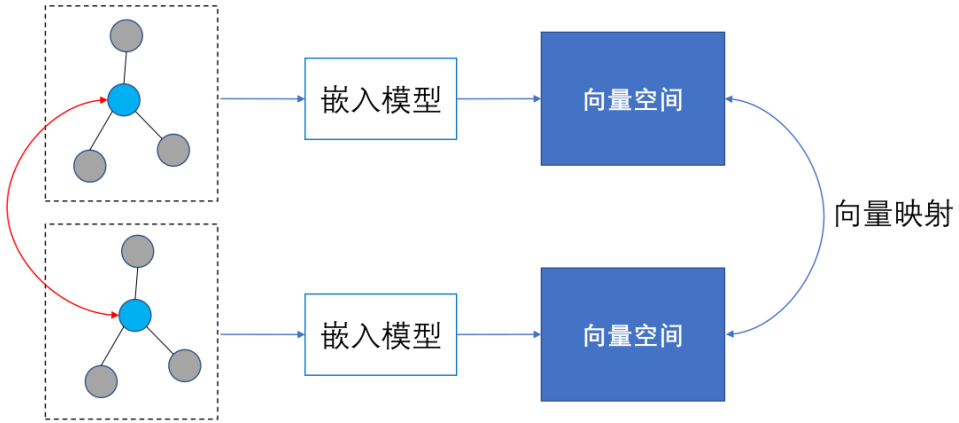
- IONE [6]
- PALE [7]

# 基于嵌入表示的方法

## JE、JAPE、IONE




## MTransE、ITransE、PALE



# 基于嵌入表示的方法

## 优点

- 不依赖于实体名称
- 不依赖于人工特征

## 缺点

- 联合训练单一网络和跨网络的信息，在模型训练过程中需要平衡两种信息的损失函数
- 网络中的属性信息没有被充分利用

# 图卷积神经网络（GCN）

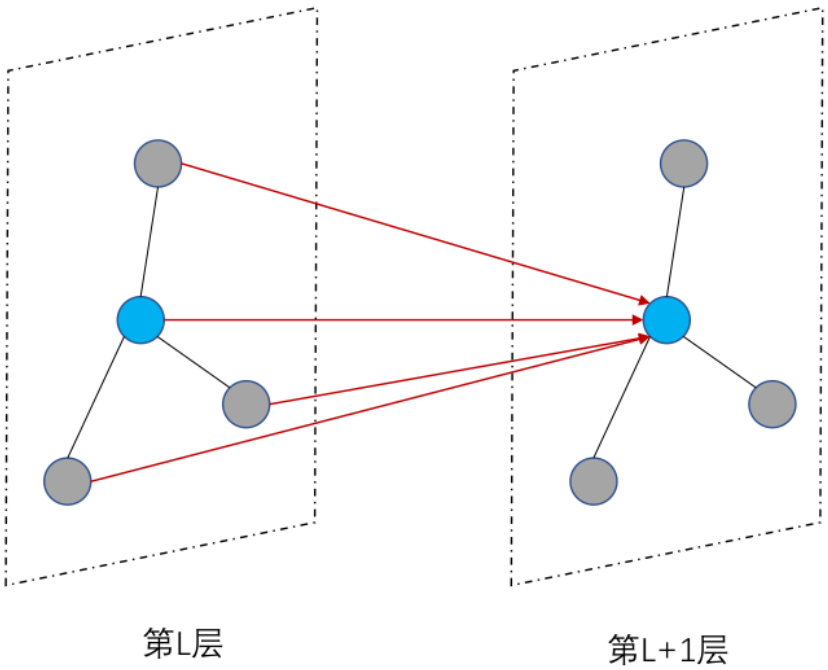



图: GCN 模型

- **输入**: 实体的特征矩阵 $X \in \mathbb{R}^{n \times d}$ 和实体的连接关系矩阵 $A \in \mathbb{R}^{n \times n}$
- **输出**: 实体的嵌入表示矩阵 $Z \in \mathbb{R}^{n \times k}$.

$$H^{(l+1)} = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

其中，$\hat{A} = A + I$，$\hat{D} = D + I$，$D$ 是每个节点的度。

# 基于图卷积神经网络的对齐模型

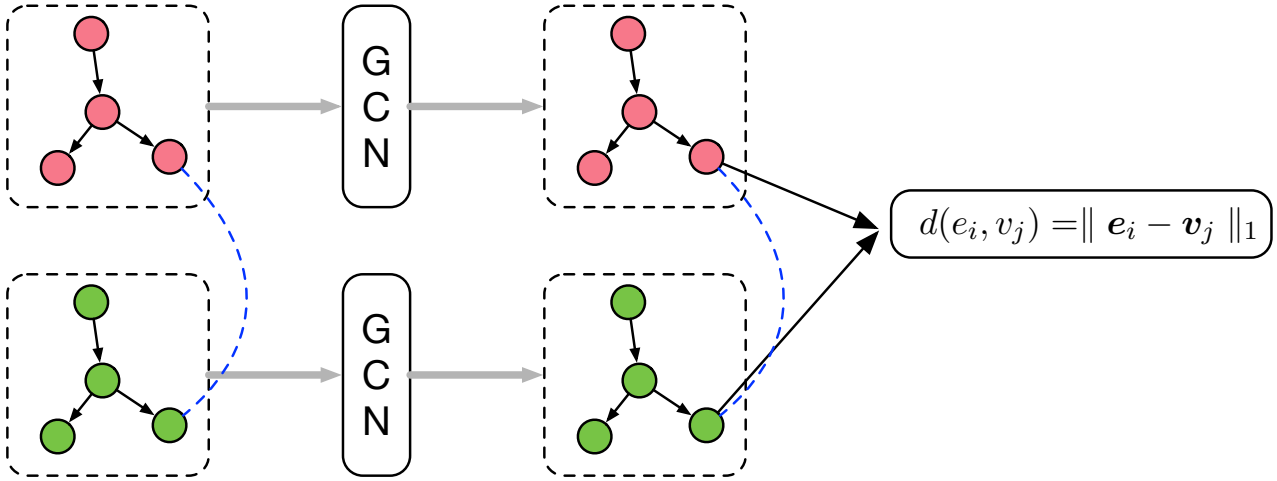


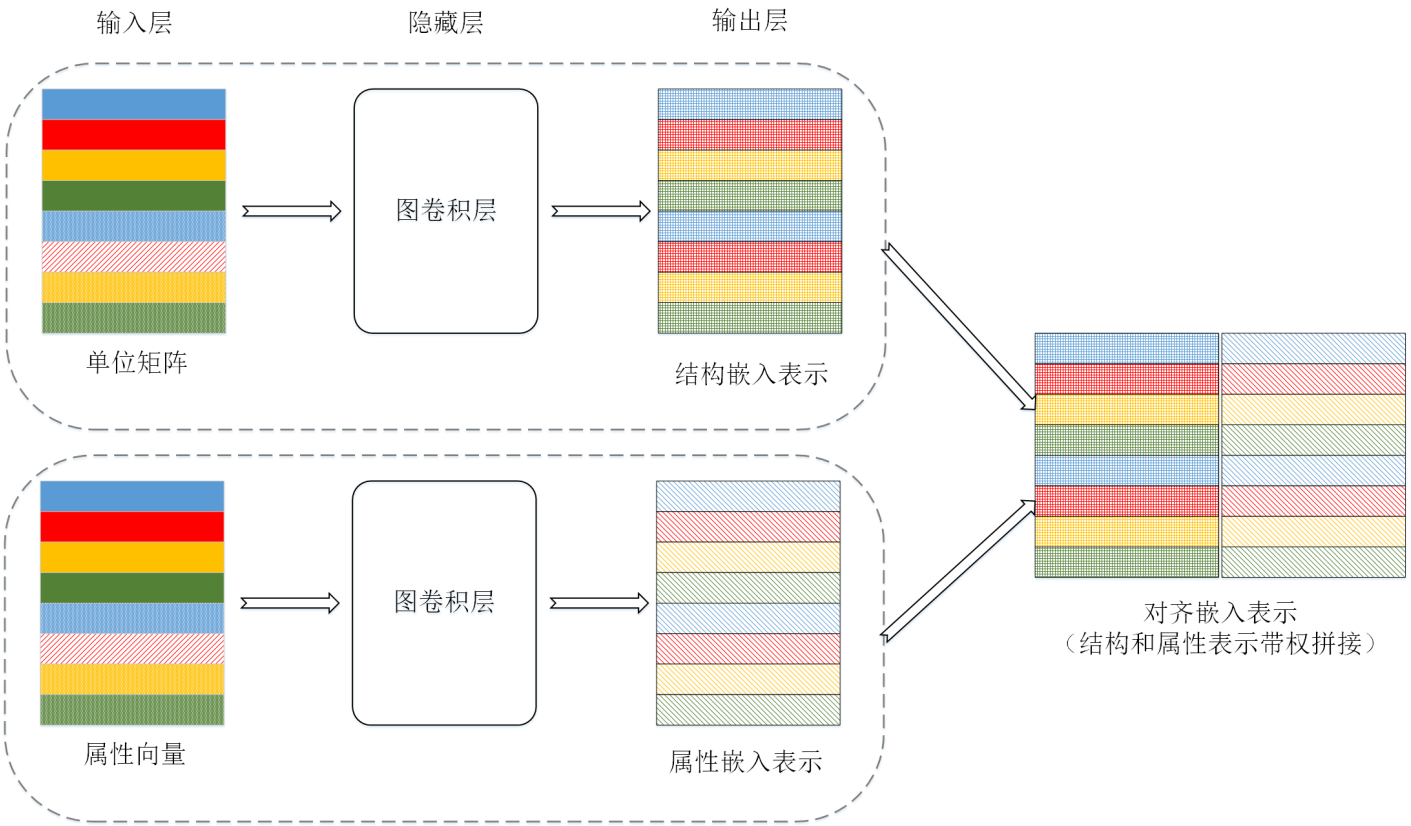

图: GCN 对齐模型

## 损失函数

$$\mathcal{L} = \sum_{(e,v)\in S} \sum_{(e',v')\in S'_{(e,v)}} [d(e,v) + \gamma - d(e',v')]_+$$

其中，$[x]_+ = \max\{0, x\}$，$S'_{(e,v)}$ 是通过 $(e, v)$ 随机替换其中一个实体得到的反例集合，$\gamma > 0$ 是一个超参数，$d$ 在这里使用 $L_1$ 距离。

# 结构嵌入和属性嵌入




$$[H_s^{(l+1)}; H_a^{(l+1)}] = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}[H_s^{(l)}W_s^{(l)}; H_a^{(l)}W_a^{(l)}]\right)$$

**最终，实体的嵌入表示矩阵为 $[\beta H_s; (1-\beta)H_a]$，其中 $\beta > 0$ 是一个超参数。**

# 数据集

## 跨语言知识图谱匹配

DBpedia15k [9]，包含三个子集：汉语-英语、法语-英语、日语-英语。

| 数据集 | | 实体数 | 关系数 | 属性数 | 关系三元组数 | 属性三元组数 |
|---|---|---|---|---|---|---|
| 汉语-英语 | 汉语 | 66,469 | 2,830 | 8,113 | 153,929 | 379,684 |
| | 英语 | 98,125 | 2,317 | 7,173 | 237,674 | 567,755 |
| 日语-英语 | 日语 | 65,744 | 2,043 | 5,882 | 164,373 | 354,619 |
| | 英语 | 95,680 | 2,096 | 6,066 | 233,319 | 497,230 |
| 法语-英语 | 法语 | 66,858 | 1,379 | 4,547 | 192,191 | 528,665 |
| | 英语 | 105,889 | 2,209 | 6,422 | 278,590 | 576,543 |

# 数据集

## 跨平台社交网络匹配

FT5k [6]，包括 Foursquare 和 Twitter 各 5000 余个用户。

| 平台 | 用户数 | 关系三元组数 | 匹配用户数 |
|---|---|---|---|
| Twitter | 5,220 | 164,916 | 1,609 |
| Foursquare | 5,315 | 76,972 | |

# 评测方法

## Hits@k

- 把匹配实体集合 $\mathcal{S}$ 按一定比例划分为训练集 $\mathcal{S}_{train}$ 和测试集 $\mathcal{S}_{test}$
- 对于测试集 $\mathcal{S}_{test}$ 中的样本 $(e_1, e_2) \in \mathcal{E}_1' \times \mathcal{E}_2'$，计算 $e_1$ 与 $\mathcal{E}_2'$ 中所有实体的距离，如果 $e_2$ 在距离中排名前 k，则记一分
- 所有测试样本中得分的比例即为 Hits@k

## 距离计算方法

- L1 距离

# 训练集: 测试集 =3:7

| 法语-英语 | | 法语 → 英语 | | | 英语 → 法语 | | |
|---|---|---|---|---|---|---|---|
| | | Hits@1 | Hits@10 | Hits@50 | Hits@1 | Hits@10 | Hits@50 |
| JE | | 15.38 | 38.84 | 56.50 | 14.61 | 37.25 | 54.01 |
| MTransE | | 24.41 | 55.55 | 74.41 | 21.26 | 50.60 | 69.93 |
| JAPE | SE w/o neg. | 29.55 | 62.18 | 79.36 | 25.40 | 56.55 | 74.96 |
| | SE | 29.63 | 64.55 | 81.90 | 26.55 | 60.30 | 78.71 |
| | SE + AE | 32.39 | 66.68 | 83.19 | 32.97 | 65.91 | 82.38 |
| JAPE′ | SE w/o neg. | 28.23 | 60.99 | 78.47 | 24.68 | 55.25 | 74.19 |
| | SE | 27.58 | 62.03 | 79.98 | 24.93 | 58.95 | 77.79 |
| | SE + AE | 30.21 | 65.81 | 82.57 | 31.42 | 63.86 | 80.95 |
| GCN | SE | 42.26 | 78.48 | 88.86 | 40.86 | 76.68 | 88.04 |
| | SE + AE | **43.76** | **80.12** | **90.81** | **42.28** | **79.03** | **90.58** |

# 训练集: 测试集 =3:7

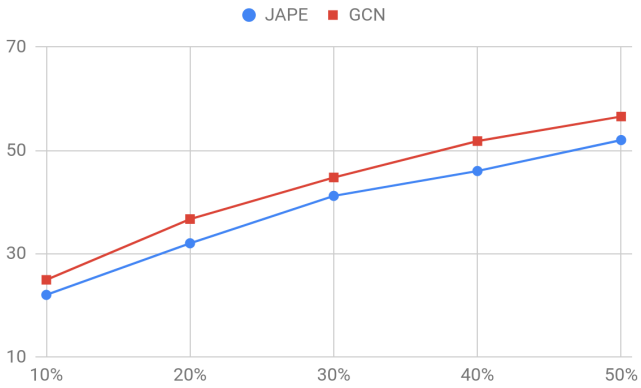
| 日语-英语 | | 日语 → 英语 | | | 英语 → 日语 | | |
|---|---|---|---|---|---|---|---|
| | | Hits@1 | Hits@10 | Hits@50 | Hits@1 | Hits@10 | Hits@50 |
| JE | | 18.92 | 39.97 | 54.24 | 17.80 | 38.44 | 52.48 |
| MTransE | | 27.86 | 57.45 | 75.94 | 23.72 | 49.92 | 67.93 |
| JAPE | SE w/o neg. | 33.10 | 63.90 | 80.80 | 29.71 | 56.28 | 73.84 |
| | SE | 34.27 | 66.39 | 83.61 | 31.40 | 60.80 | 78.51 |
| | SE + AE | 36.25 | 68.50 | 85.35 | 38.37 | 67.27 | 82.65 |
| JAPE′ | SE w/o neg. | 28.90 | 60.61 | 80.03 | 25.34 | 53.36 | 71.94 |
| | SE | 29.35 | 63.31 | 82.76 | 26.37 | 57.35 | 76.87 |
| | SE + AE | 31.06 | 64.11 | 81.57 | 32.45 | 62.21 | 79.08 |
| GCN | SE | 42.73 | 75.63 | 84.61 | 40.59 | 72.51 | 82.15 |
| | SE + AE | **45.10** | **78.34** | **88.57** | **42.62** | **75.41** | **86.62** |

# 训练集: 测试集 =3:7

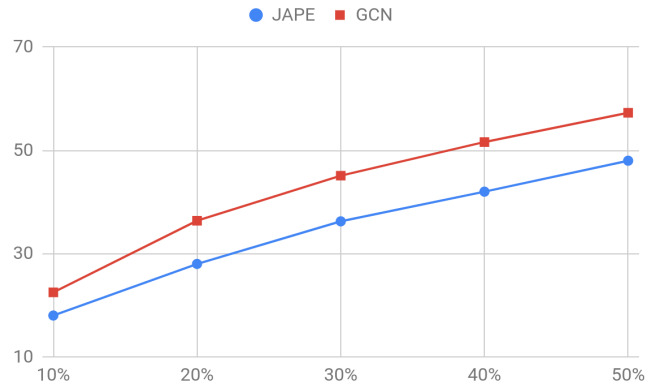
| 汉语-英语 | | 汉语 → 英语 | | | 英语 → 汉语 | | |
|---|---|---|---|---|---|---|---|
| | | Hits@1 | Hits@10 | Hits@50 | Hits@1 | Hits@10 | Hits@50 |
| JE | | 21.27 | 42.77 | 56.74 | 19.52 | 39.36 | 53.25 |
| MTransE | | 30.83 | 61.41 | 79.12 | 24.78 | 52.42 | 70.45 |
| JAPE | SE w/o neg. | 38.34 | 68.86 | 84.07 | 31.66 | 59.37 | 76.33 |
| | SE | 39.78 | 72.35 | 87.12 | 32.29 | 62.79 | 80.55 |
| | SE + AE | 41.18 | 74.46 | **88.90** | 40.15 | **71.05** | **86.18** |
| JAPE′ | SE w/o neg. | 30.10 | 62.58 | 80.28 | 23.04 | 52.91 | 72.17 |
| | SE | 30.54 | 66.41 | 83.94 | 23.91 | 57.02 | 77.31 |
| | SE + AE | 33.32 | 69.28 | 86.40 | 33.02 | 66.92 | 85.15 |
| GCN | SE | 41.96 | 74.10 | 83.78 | 38.62 | 69.58 | 80.70 |
| | SE + AE | **44.75** | **77.31** | 87.76 | **40.61** | 72.61 | 84.10 |

# 不同比例的训练集测试集划分

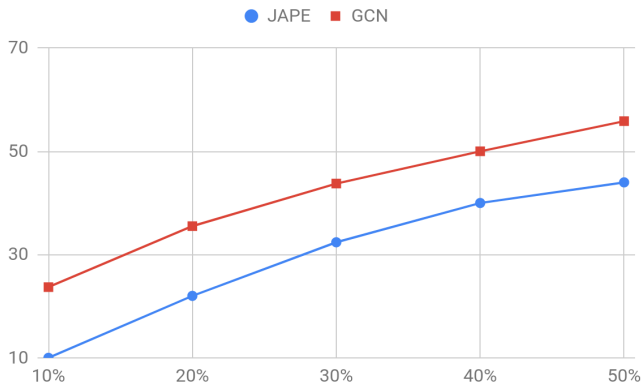
## JAPE 和 GCN Hits@1 结果对比




(a) 汉语-英语




(b) 日语-英语




(c) 法语-英语

# 不同比例的训练集测试集划分

IONE 和 GCN Hits@30 结果对比

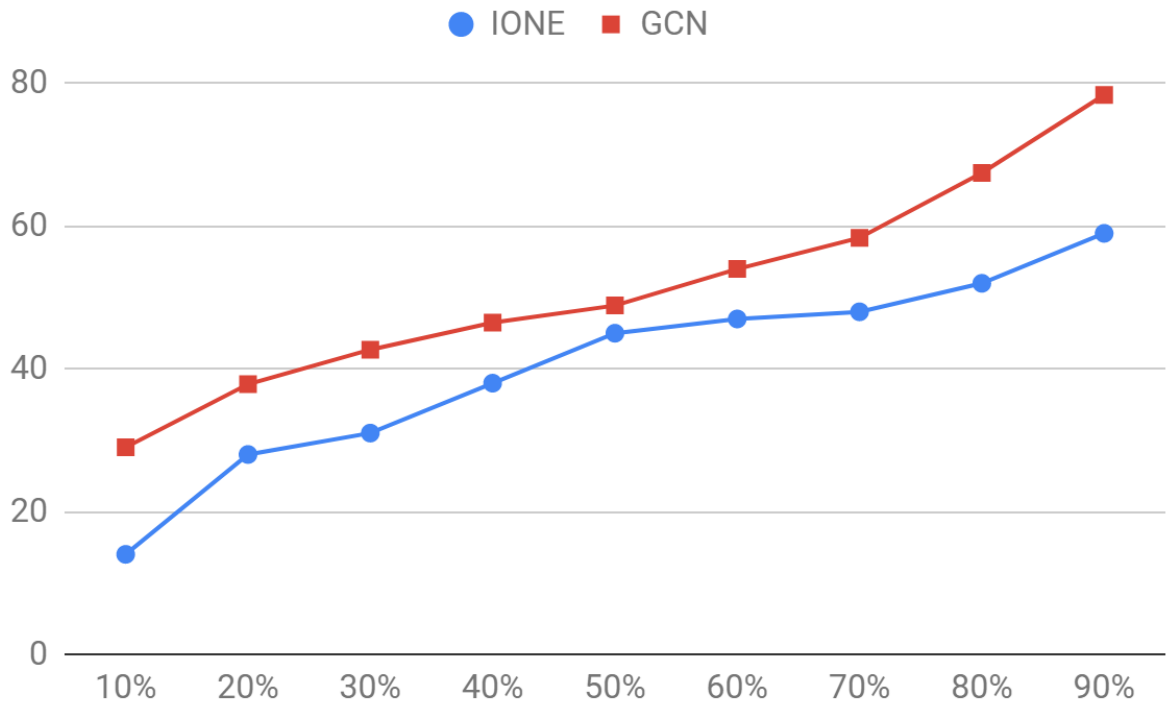



图: Foursquare-Twitter 数据集

# 总结

我们提出了一种基于图卷积神经网络的异构信息网络对齐模型，该模型：

- 直接面向对齐任务训练实体的嵌入表示，无需权衡多个损失函数
- 可以同时处理网络的结构信息和属性信息
- 在知识图谱领域和社交网络领域都能达到很好的效果

# 参考文献 I



Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo.
Multilingual knowledge graph embeddings for cross-lingual knowledge alignment.
In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1511–1517, 2017.


Bo Fu, Rob Brennan, and Declan O'Sullivan.
Cross-lingual ontology mapping – an investigation of the impact of machine translation.
In Asunción Gómez-Pérez, Yong Yu, and Ying Ding, editors, *The Semantic Web*, pages 1–15, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.


Yanchao Hao, Yuanzhe Zhang, Shizhu He, Kang Liu, and Jun Zhao.
A joint embedding method for entity alignment of knowledge bases.
In *China Conference on Knowledge Graph and Semantic Computing*, pages 3–14. Springer, 2016.


Xiangnan Kong, Jiawei Zhang, and Philip S Yu.
Inferring anchor links across multiple heterogeneous social networks.
In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 179–188. ACM, 2013.

# 参考文献 II



Jing Liu, Fan Zhang, Xinying Song, Young-In Song, Chin-Yew Lin, and Hsiao-Wuen Hon.
What's in a name?: an unsupervised approach to link users across communities.
In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 495–504. ACM, 2013.


Li Liu, William K Cheung, Xin Li, and Lejian Liao.
Aligning users across social networks using network embedding.
In *IJCAI*, pages 1774–1780, 2016.


Tong Man, Huawei Shen, Shenghua Liu, Xiaolong Jin, and Xueqi Cheng.
Predict anchor links across social networks via an embedding approach.
In *IJCAI*, volume 16, pages 1823–1829, 2016.


Dennis Spohr, Laura Hollink, and Philipp Cimiano.
A machine learning approach to multilingual and cross-lingual ontology matching.
In Lora Aroyo, Chris Welty, Harith Alani, Jamie Taylor, Abraham Bernstein, Lalana Kagal, Natasha Noy, and Eva Blomqvist, editors, *The Semantic Web – ISWC 2011*, pages 665–680, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.


Zequn Sun, Wei Hu, and Chengkai Li.
Cross-lingual entity alignment via joint attribute-preserving embedding.
In *International Semantic Web Conference*, pages 628–644. Springer, 2017.

# 参考文献 III


Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang.
Cross-lingual knowledge linking across wiki knowledge bases.
In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 459–468, New York, NY, USA, 2012. ACM.


Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun.
Iterative entity alignment via joint knowledge embeddings.
In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4258–4264. AAAI Press, 2017.

# 谢谢！