南京大学万维网软件研究组

The Websoft Research Group, Nanjing University, China

知识图谱前沿动态综述

胡伟

Home: http://ws.nju.edu.cn/~whu E-mail: whu@nju.edu.cn

南京大学 计算机软件新技术国家重点实验室

CCL / NLP-NABD 2018, 长沙, 2018/10/21

知识图谱

- 知识图谱于2012年5月由Google正式提出，其初衷是为了提高搜索引擎的能力，改善用户的搜索质量以及搜索体验
  - Node: entity / concept
  - Edge: attribute / relationship
- Other famous knowledge bases
  - DBpedia, Freebase, Wikidata, YAGO, WordNet, Probase …
  - Linked Open Data (LOD) cloud

Nanjing University

Website | Directions | Save

Public university in Nanjing, China

Nanjing University, or Nanking University, is a prestigious public university, and is the oldest institution of higher learning, located in Nanjing, China. Wikipedia

Address: 22 Hankou Rd, Gulou Qu, Nanjing Shi, Jiangsu Sheng, China, 210008

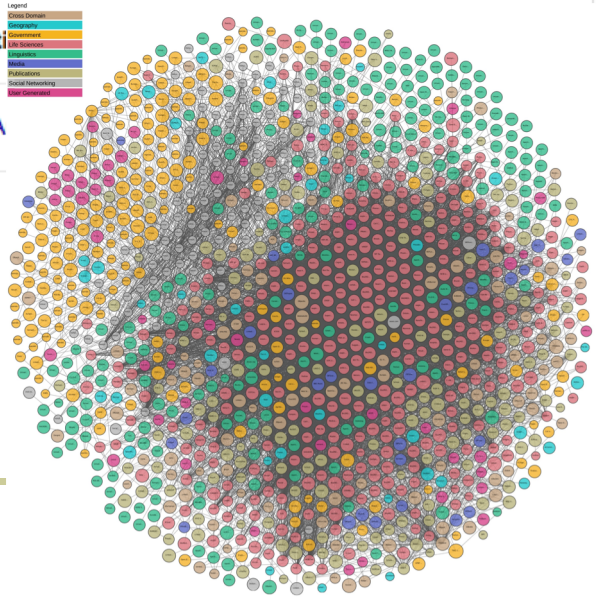
Total enrollment: 35,434 (2007)

President: Lv Jian (吕建)

Province: Jiangsu

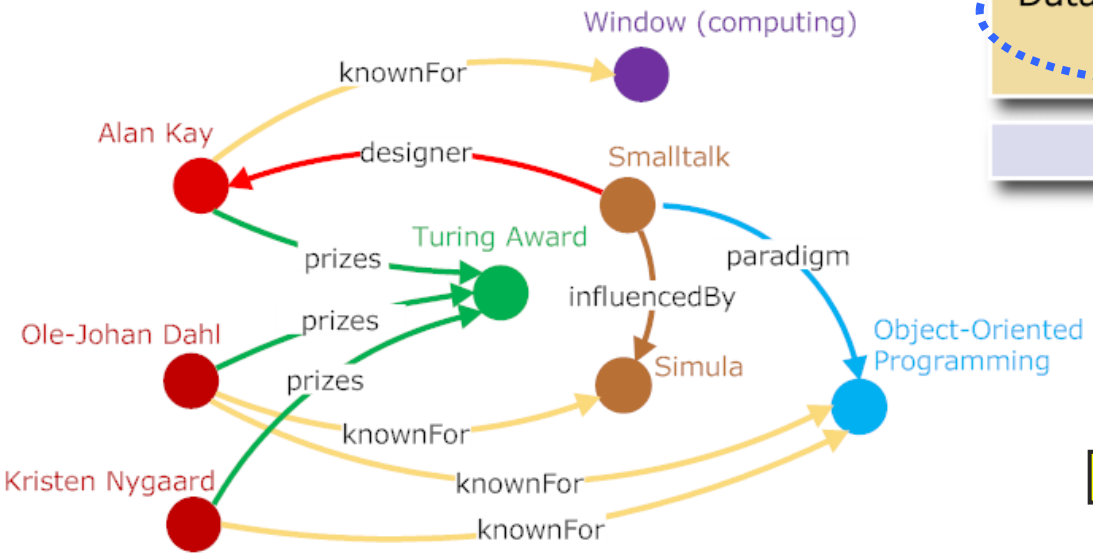
Undergraduate tuiti

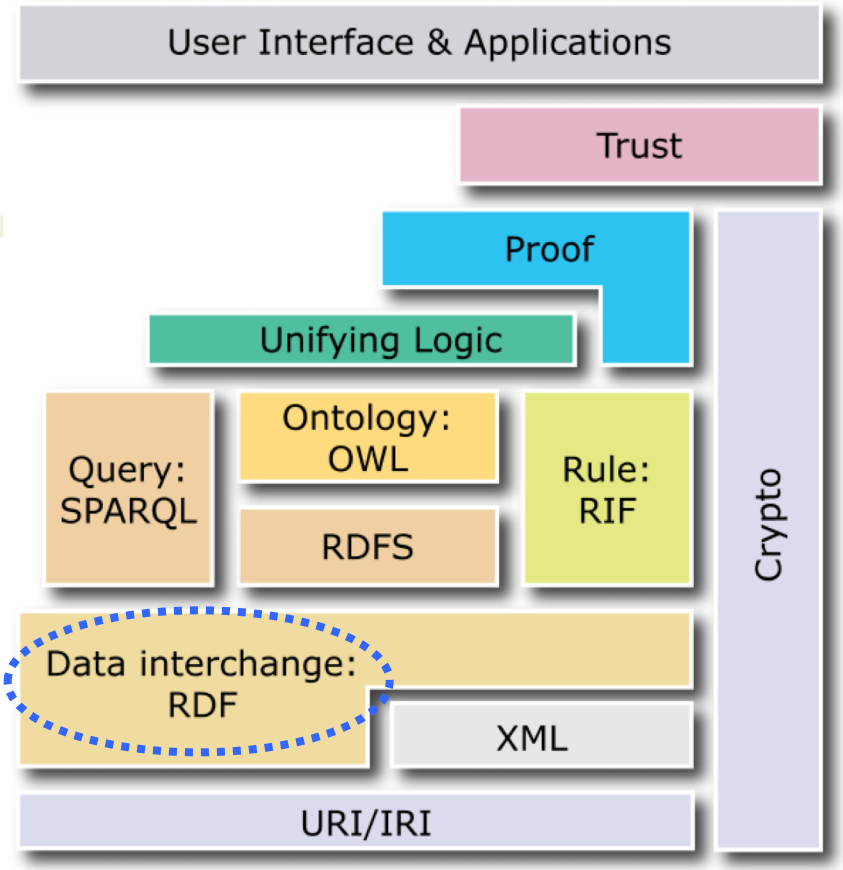
Know this place? A



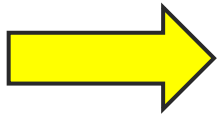
RDF (Resource Description Framework)

- RDF三元组





Semantic Web layers



*" The world is **not** made of strings, **but** is made of **things** "*

知识图谱研究框架

书本、字典、百科

知识获取

知识图谱

已有知识图谱

知识融合

知识图谱

知识推理

知识补充

调研

- 知识补充

1. T. Dettmers, P. Minervini, P. Stenetorp, S. Riedel. *Convolutional 2D Knowledge Graph Embeddings.* In: **AAAI 2018**
2. J. Lajus, F.M. Suchanek. *Are All People Married? Determining Obligatory Attributes in Knowledge Bases.* In: **WWW 2018**
3. P. Mirza, S. Razniewski, F. Darari, G. Weikum. *Enriching Knowledge Bases with Counting Quantifiers.* In: **ISWC 2018**

- 知识融合

1. Z. Sun, W. Hu, Q. Zhang, Y. Qu. *Bootstrapping Entity Alignment with Knowledge Graph Embedding.* In: **IJCAI 2018**
2. P. Kolyvakis, A. Kalousis, D. Kiritisis. *DeepAlignment: Unsupervised Ontology Matching with Refined Word Vectors.* In: **NAACL-HLT 2018**

- 知识推理

1. P.G. Omran, K. Wang, Z. Wang. *Scalable Rule Learning via Learning Representation.* In: **IJCAI 2018**

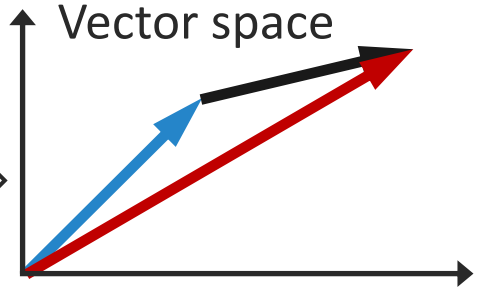
提纲

- 知识补充
- 知识融合
- 知识推理

知识补充

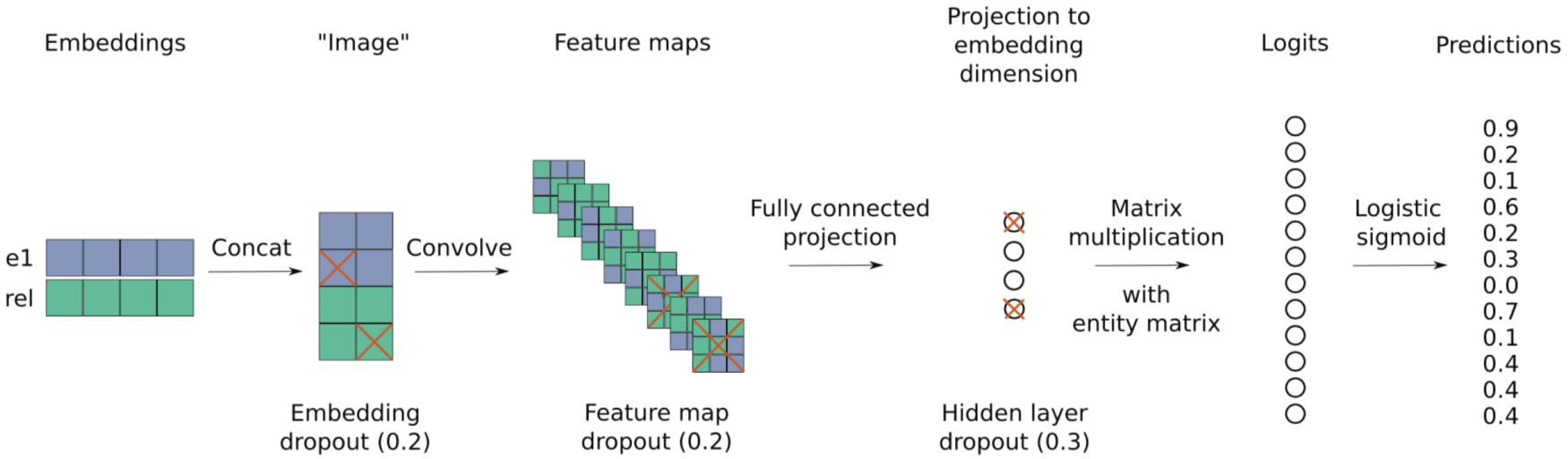
- 问题：现有知识图谱依然不完备
- 任务：**entity / link prediction**
  - given $(s, r, ?)$, complete a triple by predicting o // or s given $(?, r, o)$
- 方法：表示学习
  - 翻译模型（translational models）
    - TransE: model (s, r, o) as $\boldsymbol{s} + \boldsymbol{r} \approx \boldsymbol{o}$
    - 后续改进: TransH, TransR, PTransE, …
  - 语义模型
    - DistMult, ComplEx, RESCAL, NeuralLP, …

(*Washington*, *capitalOf*, *America*)



Convolutional 2D Knowledge Graph Embeddings

- 方法：a multi-layer **convolutional** network model
  - 步骤 1 & 2：变形并合并实体和关系embeddings
  - 步骤 3, 4 & 5：输入卷积层，将结果投影到一个k维空间，匹配候选



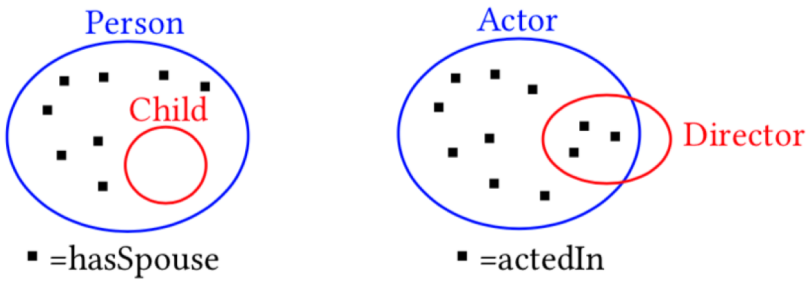
- 效果：FB15K-237测试集上实验结果优异

知识补充

- 问题：现有知识图谱有待进一步挖掘
- 任务：**obligatory attributes (必要属性)**
  - Determine whether all instances of a given class have a given attribute in the real world – while all we have is an incomplete KB
    - *hasBirthDate* is an obligatory attribute for class *Person*, while *hasSpouse* is not
- 任务：**counting quantifiers (计数量词)**
  - Text often contains only counting information: the number of objects that stand in a specific relation with a certain entity, without mentioning the objects themselves
    - Given the sentence "*Trump has three sons and two daughters*", the output for predicate *numberOfChildren* should be **5**.
    - Children CQ: Wikipedia 12%; name mentioned: 7%; Wikidata: 2.5%

Determining Obligatory Attributes in Knowledge Bases

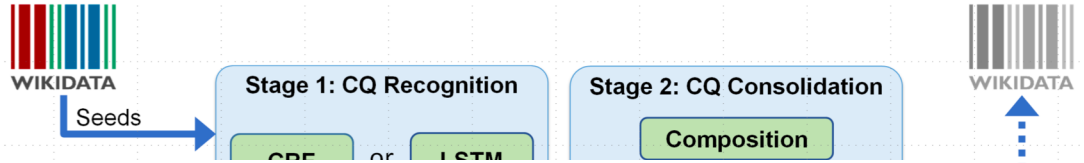
- 问题：必要属性
- 挑战
  1. 数十万概念，难以人工判断；自动化处理很难，比如没有反例数据
  2. **开放世界假设：**无法判断一条没有包含在知识库中的三元组的真伪
- 方法
  - 基于概念层次结构来推断概念的必要属性
  - 基本假设
    - 假设知识库的不完全性在知识库中所有概念上都是**均匀分布**的
    - 如果一个属性在某个概念上分布较稀疏，而在其他概念上分布较稠密，则可以推断它一定不是分布较稀疏的概念的必要属性



$$s_p^K(c, c') = \frac{conf(c \setminus c' \subseteq p_K)}{conf(c \cap c' \subseteq p_K)}$$

Enriching Knowledge Bases with Counting Quantifiers

- 问题：计数量词
- 挑战：Non-maximal seeds; Sparse, skewed observations; Linguistic diversity
- 方法
  - **步骤 1：识别CQ**



sentence	*Jolie*	*brought*	*her*	*twins*	,	*one*	*daughter*	*and*	*three*	*adopted*	*children*	*to*	*the*	*gala*	.
pre-processed	*Jolie*	*brought*	*her*	NUMTERM	,	CARDINAL	*daughter*	*and*	CARDINAL	*adopted*	*children*	*to*	*the*	*gala*	.
output tags	O	O	O	COUNT	COMP	COUNT	O	COMP	COUNT	O	O	O	O	O	O

  - **步骤 2：合并CQ**

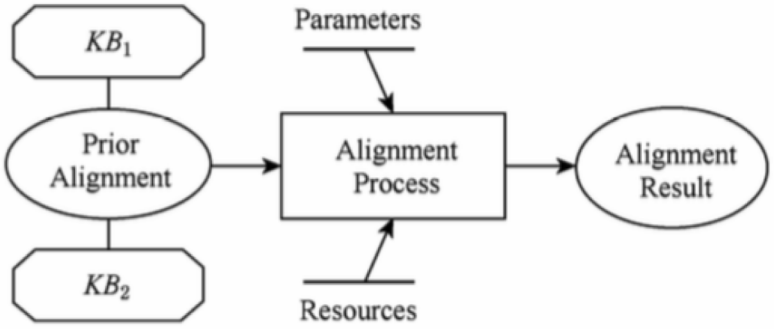
l_1: *Angelina has a grand total of* $\underline{six}_{0.4}$ *children together:* $\underline{three}_{0.3}$ *biological* $[and]_{0.6}$ $\underline{three}_{0.5}$ *adopted.*

l_2: *The arrival of the* $\underline{first}_{0.5}$ *biological child of Jolie and Pitt caused an excited flurry with fans.*

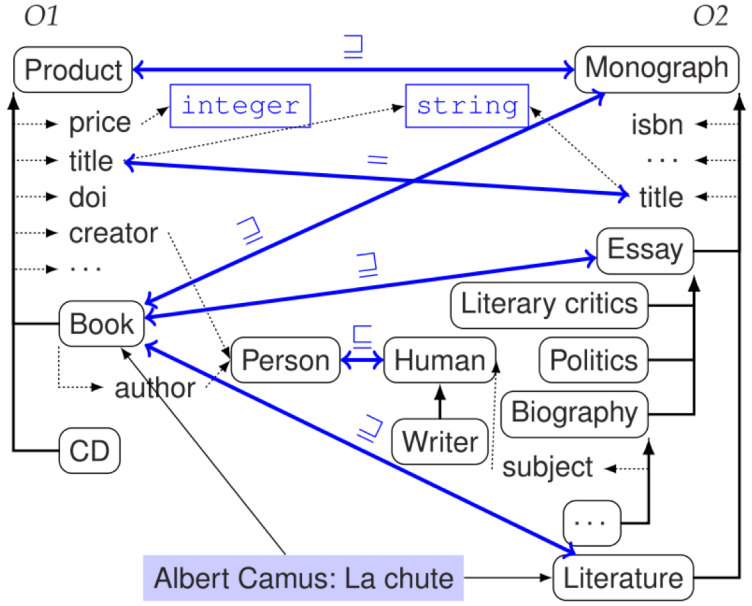
l_3: *On July 12, 2008, she gave birth to* $\underline{twins}_{0.8}$: $\underline{a}_{0.1}$ *son, Knox Léon,* $[and]_{0.5}$ $\underline{a}_{0.2}$ *daughter, Vivienne Marcheline.*

知识融合

- 问题：相似领域存在多个异构的知识图谱
- 任务：**实体对齐** & **本体匹配**
  - 实体对齐侧重发现指称真实世界相同对象的不同实例
  - 本体匹配侧重发现（模式层）等价或相似的类、属性或关系
- 流程

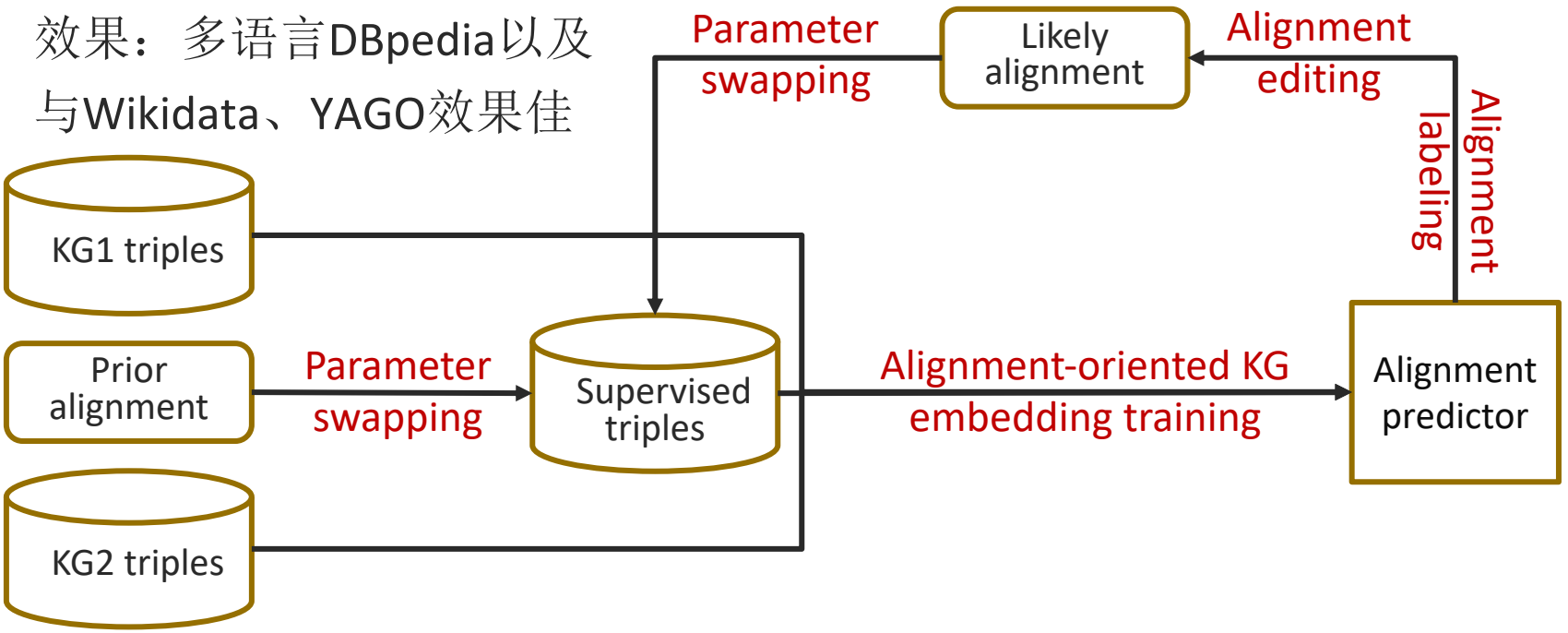


- 方法
  - 文本相似性：编辑距离、向量空间模型 ...
  - 结构相似性：子图匹配 ...



Bootstrapping Entity Alignment with KG Embedding

- 目标：通过embedding度量实体相似度
- 挑战：1) 面向对齐的知识图谱表示学习; 2) 已知的先验对齐数量少
- 方法：bootstrapping
- 效果：多语言DBpedia以及与Wikidata、YAGO效果佳



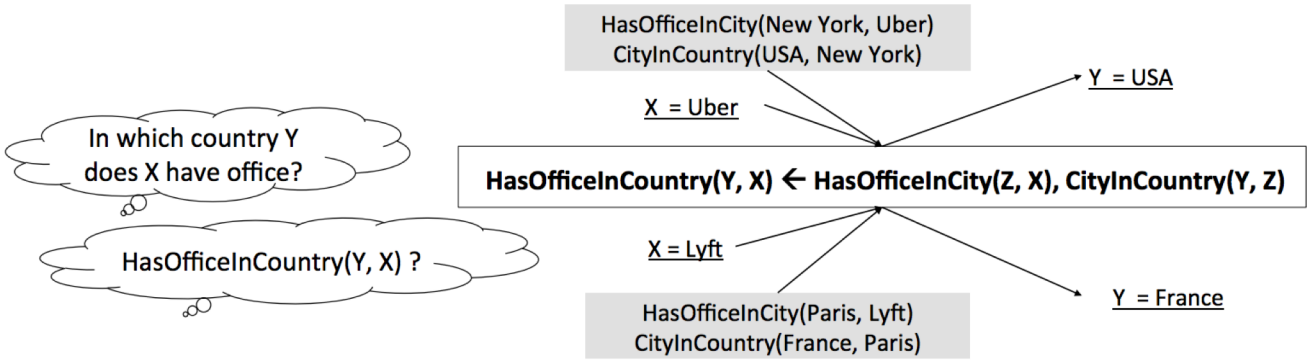
Unsupervised Ontology Matching with Refined Word Vectors

- 目标：无监督表示学习
- 挑战：tailor word embeddings to the domains and ontologies
- 方法
  - Learn domain-specific word vectors
    - antonym repel
    - synonym attract
    - original word-vector distance
    - WordNet, PPDP, WikiSynonyms
    - 相似性度量：Dual embedding space model
  - Extend stable marriage algorithm
    - ε-optimal mappings: 补充少量多对多映射
- 效果
  - OAEI Conference数据集
  - Schema.org – DBpedia alignment

System	Precision	Recall	Micro-F1
DeepAlignment	0.71	**0.80**	**0.75**
CroMatch	0.76	0.69	0.72
AML	0.79	0.65	0.71
DeepAlignment$_*$	0.68	0.68	0.68
XMap	0.81	0.58	0.67
LogMap	0.79	0.58	0.66
LogMapBio	0.75	0.58	0.65
StringEquiv	0.83	0.50	0.62

知识推理

- 问题：谓词逻辑存在高效证明系统的子集包括描述逻辑和**规则系统**
- 任务：规则学习
  - 规则：$P_t(x,y) \leftarrow P_1(x,z_1) \wedge P_2(z_1,z_2) \wedge \cdots \wedge P_n(z_{n-1},z_n)$
- 方法
  - 传统方法：归纳逻辑程序设计（inductive logic programming）
    - 例如，AMIE+
  - 统计预测模型：表示学习
- 应用
  - 问答
  - …



Scalable Rule Learning via Learning Representation

- 目标：基于表示学习的规则学习
- 挑战：可伸缩性
- 方法
  - 采样：只包含和目标谓词相关的实体和事实
  - 表示学习 & 规则搜索
    - 评分函数：相似性和共现性
  - 规则评估：确信度和覆盖率
- 效果

Algorithm 1 Learn rules for a KG and a target predicate

Input: a KG K, a predicate P_t, an integer $len \geq 2$, and two real numbers $MinSC, MinHC \in [0, 1]$

Output: a set $Rule$ of CP rules

1: $K' := \mathsf{Sampling}(K, P_t, len)$
2: $(\mathcal{P}, \mathcal{A}) := \mathsf{Embeddings}(K')$
3: $Candidates := \emptyset$
4: **for** $2 \leq l \leq len$ **do**
5: Add $\mathsf{RuleSearch}(K', P_t, \mathcal{P}, \mathcal{A}, l)$ to $Candidates$
6: **end for**
7: $Rules := \mathsf{Evaluate}(Candidates, K)$
8: $Rules := \mathsf{Filter}(Candidates, MinSC, MinHC)$
9: **return** $Rules$

KG	RLvLR			AMIE+		
	#R	#QR	Time	#R	#QR	Time
YAGO2s	**6.3**	**1.8**	**0.96**	5.65	0.5	10.00
DBpedia 3.8	**42.7**	**9.2**	**3.88**	9.05	0.5	4.59
Wikidata	**56.8**	**25.6**	**2.41**	0.95	0.3	10.00

Table 2: Rule mining comparison between RLvLR and AMIE+

KG	RLvLR		AMIE+	
	#Facts	#QFacts	#Facts	#QFacts
YAGO2s	**1.1M**	**7K**	0.27M	1
DBpedia 3.8	**16.6M**	**162K**	1.6M	1.8K
Wikidata	**2.1M**	**99K**	0.17M	4.6K

Table 3: The numbers of new facts predicted by RLvLR and AMIE+

总结

- 报告从知识补充、知识融合、知识推理三个方面介绍了知识图谱的前沿技术动态
- 报告还有很多未尽之处
  - 人机协作：众包、摘要 ……
  - 机器学习：强化学习、生成对抗网络 ……
  - 知识图谱应用：问答、推荐 ……
  - ……

其他进展

- L. Cai, W.Y. Wang. *KBGAN: Adversarial learning for knowledge graph embeddings.* In: **NAACL-HLT 2018**
- M. Chen, Y. Tian, M. Yang, C. Zaniolo. *MTransE: Multilingual knowledge graph embeddings for cross-lingual knowledge alignment.* In: **IJCAI 2017**
- S. Guo, Q. Wang, L. Wang, B. Wang, L. Guo. *Knowledge graph embedding with iterative guidance from soft rules.* In: **AAAI 2018**
- J. Huang, W. Hu, H. Li, Y. Qu. *Automated comparative table generation for facilitating human intervention in multi-entity resolution.* In: **SIGIR 2018**
- Z. Wang, Q. Lv, X. Lan, Y. Zhang. *Cross-lingual knowledge graph alignment via graph convolutional networks.* In: **EMNLP 2018**
- W. Xiong, T. Hoang, W.Y. Wang. *DeepPath: A reinforcement learning method for knowledge graph reasoning.* In: **EMNLP 2017**
- H. Zhu, R. Xie, Z. Liu, M. Sun. *IPTransE: Iterative entity alignment via joint knowledge embeddings.* In: **IJCAI 2017**
- Y. Zhuang, G. Li, Z. Zhong, J. Feng. *Hike: A hybrid human-machine method for entity alignment in large-scale knowledge bases.* In: **CIKM 2017**

谢谢！

致谢

- CCL / NLP-NABD 2018
- 国家重点研发计划 (2018YFB1004304), 国家自然科学基金 (61872172)