



大规模知识图谱表示学习

趋势与挑战

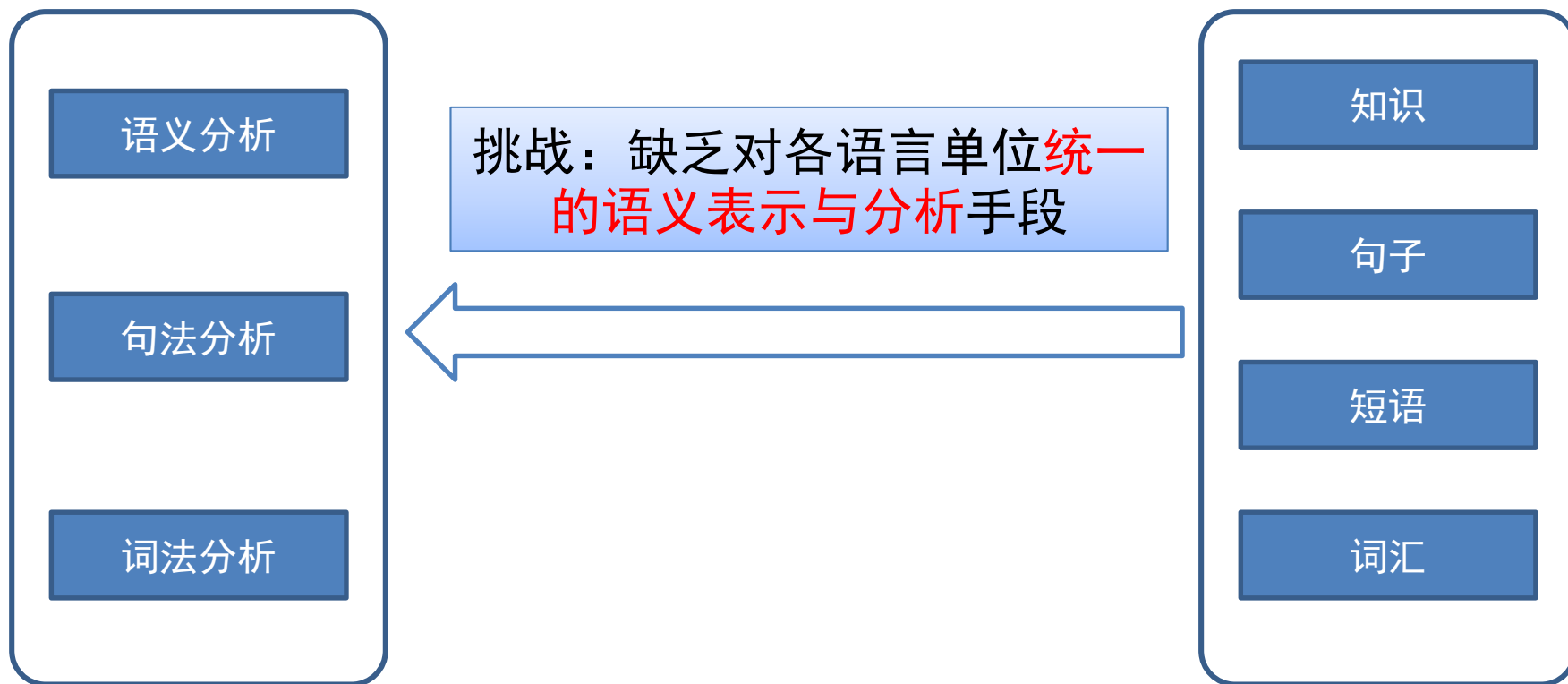
清华大学自然语言处理实验室

刘知远

liuzy@tsinghua.edu.cn

机器学习 = 数据表示 + 学习目标 + 优化方法

表示学习的意义



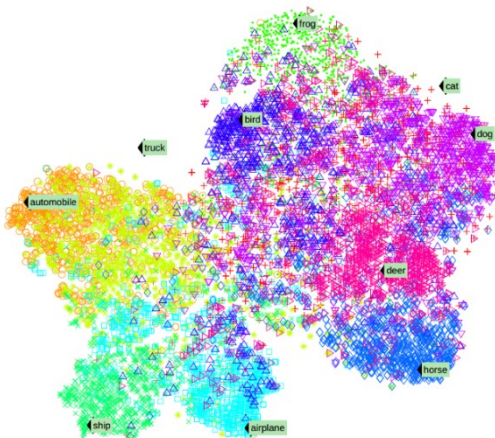
表示学习的意义

表示学习建立统一的
语义表示空间

语义分析

句法分析

词法分析



知识

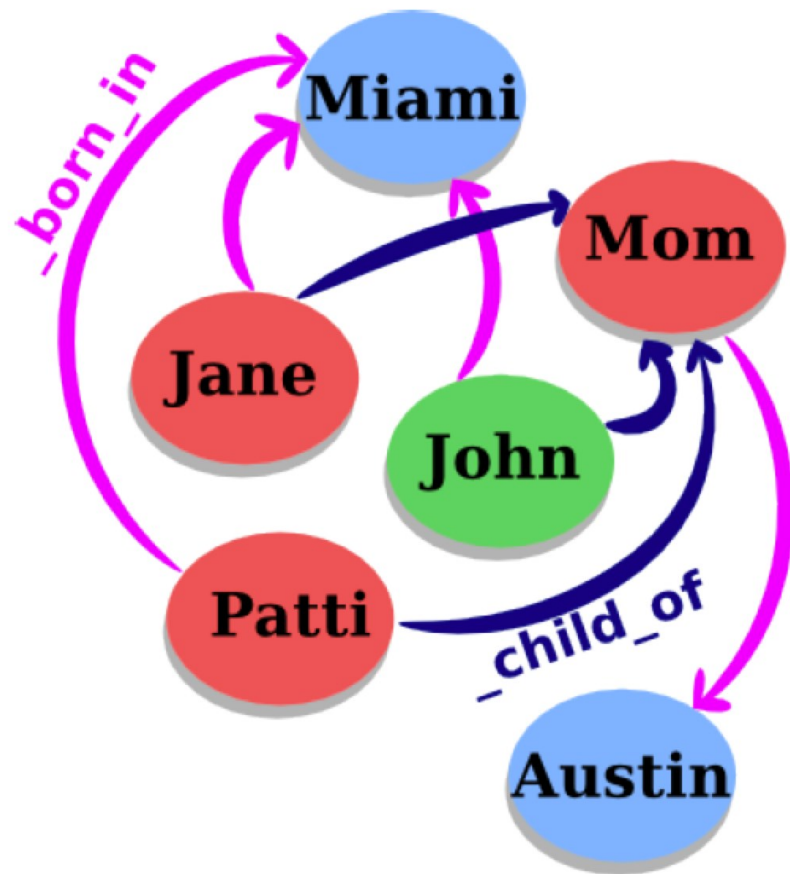
句子

短语

词汇

知识图谱实体与关系的表示

- 知识图谱包括实体与关系
 - 节点代表实体
 - 连边代表关系
- 事实可以用三元组表示
 - (head, relation, tail)
- 代表知识库
 - WordNet: 语言知识
 - Freebase: 世界知识

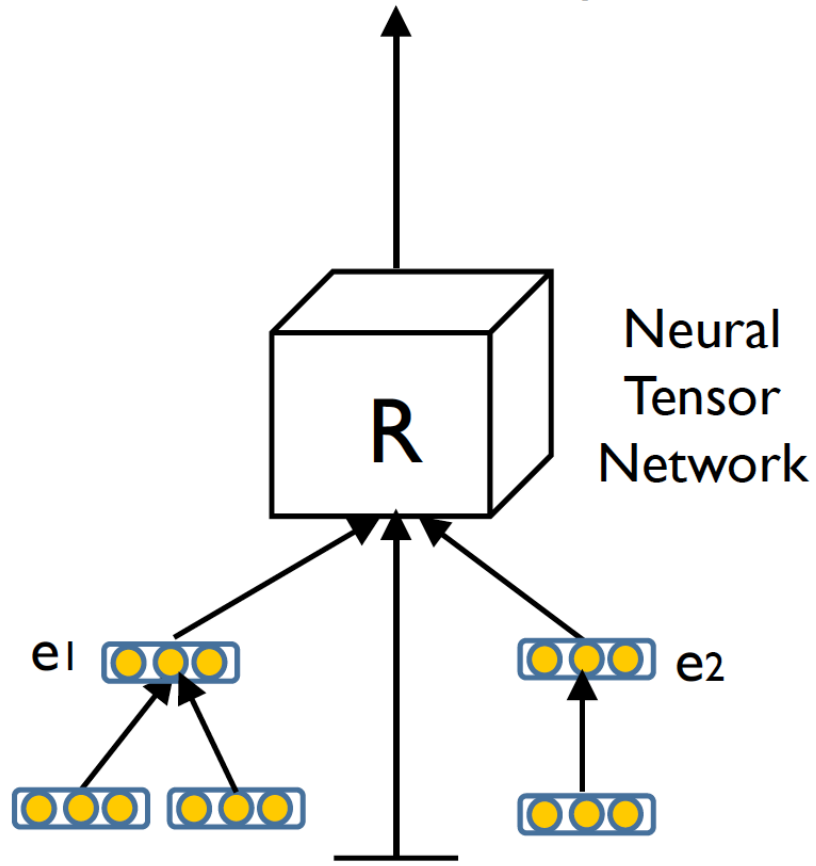


知识图谱表示学习

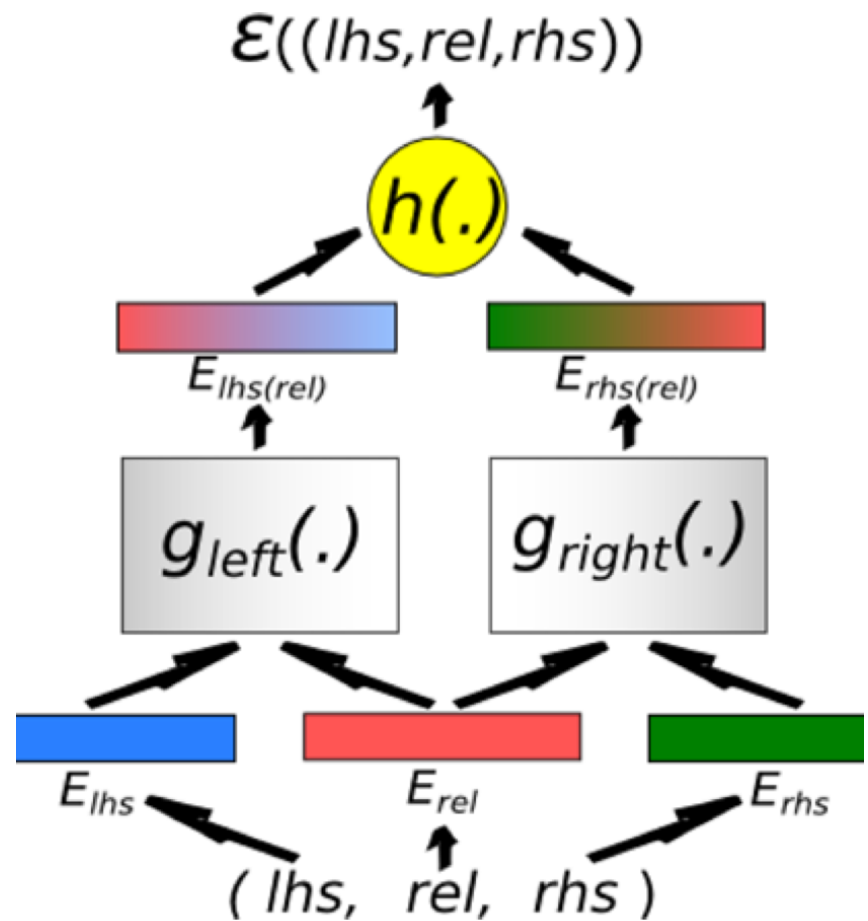
- 大规模知识获取
 - 从文本数据抽取关系：信息抽取任务
 - 从知识图谱抽取关系：知识图谱补全
- 研究挑战：如何表示与利用知识图谱信息
 - 高维： $10^5 \sim 10^8$ 个实体, $10^7 \sim 10^9$ 种关系
 - 稀疏
 - 高噪音、不完整
- 研究思路：将知识图谱嵌入到低维向量空间
- 应用场景：知识获取，知识推理，知识融合

知识表示代表模型

Confidence for Triplet



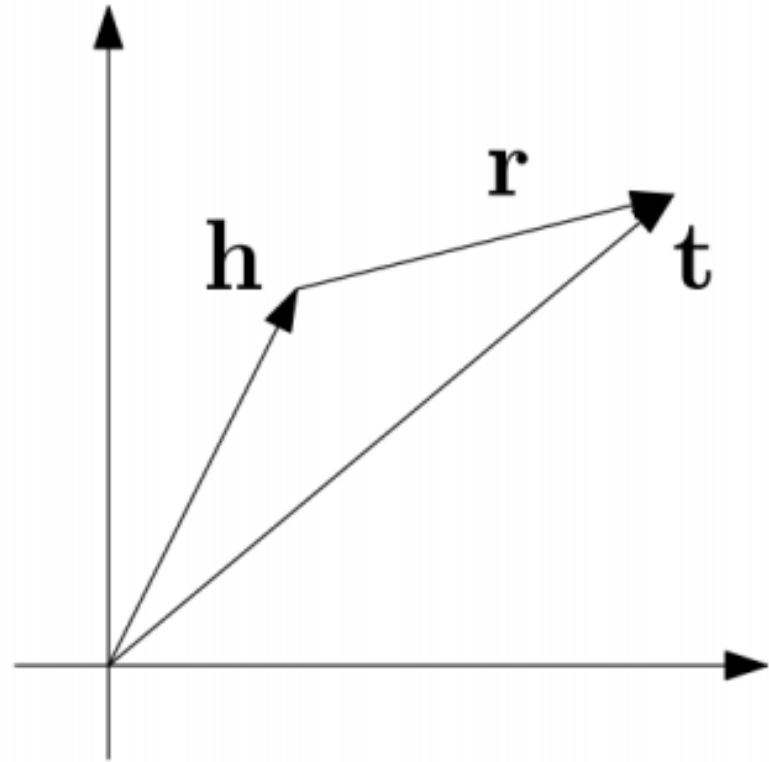
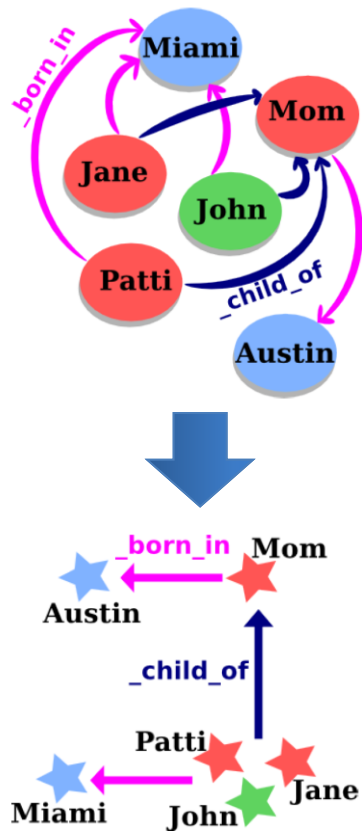
Neural Tensor Network (NTN)



Energy Model

知识表示代表模型：TransE

- 对每个事实 (head, relation, tail), 将其中的 relation 作为从 head 到 tail 的翻译操作
- 优化目标: $h + r = t$



评测任务：链接预测

电影WALL-E的风格是什么

WALL-E

电影风格

?



评测任务：链接预测

电影WALL-E的风格是什么

WALL-E

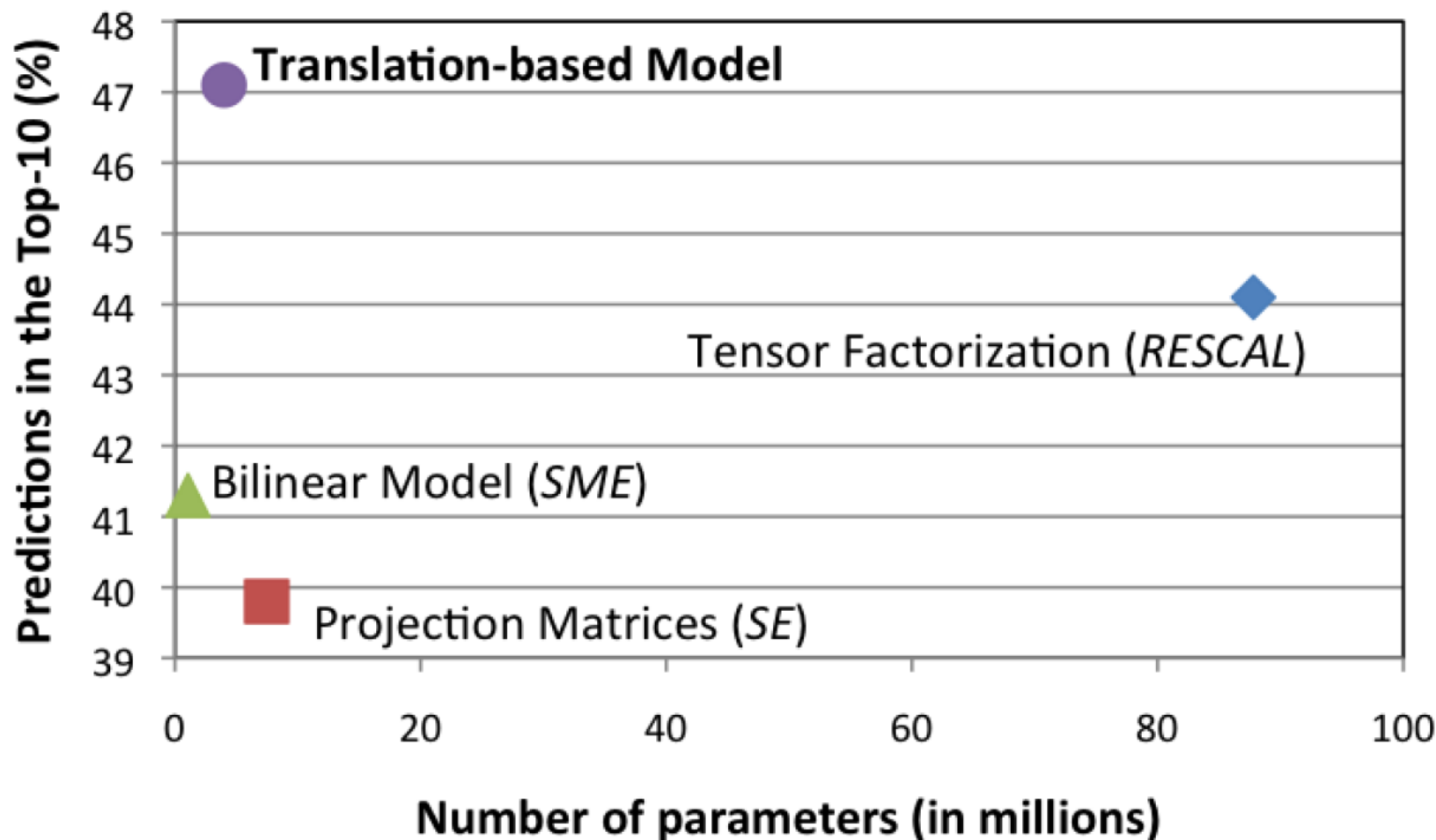
电影风格



1. Animation
2. Computer animation
3. Comedy film
4. Adventure film
5. Science Fiction
6. Fantasy
7. Stop motion
8. Satire
9. Drama
10. Connecting

链接预测性能比较

FB15K



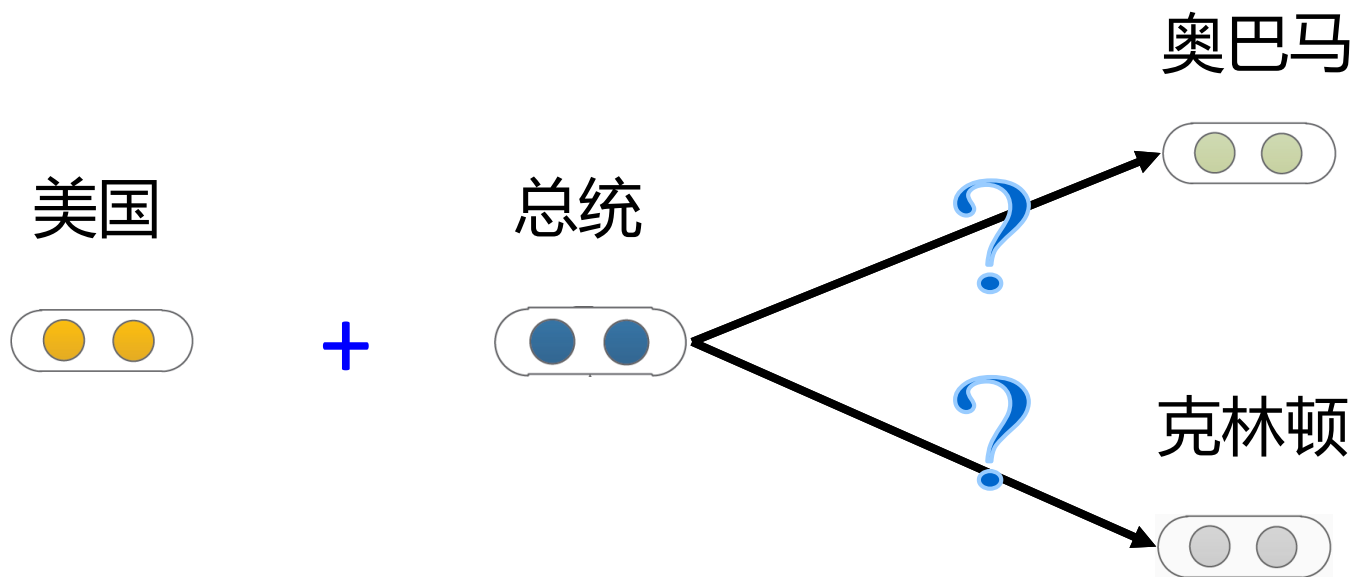
TransE的启示

- 合理设计学习目标的重要性：研究创新所在
- 模型复杂度与知识图谱稀疏性的辩证关系

Data Sets	WN11	FB13	FB15K
SE	53.0	75.2	-
SME (bilinear)	70.0	63.7	-
SLM	69.9	85.3	-
LFM	73.8	84.3	-
NTN	70.4	87.1	68.5
TransE (unif)	75.9	70.9	79.6
TransE (bern)	75.9	81.5	79.2
TransH (unif)	77.7	76.5	79.0
TransH (bern)	78.8	83.3	80.2
TransR (unif)	85.5	74.7	81.7
TransR (bern)	85.9	82.5	83.9
CTransR (bern)	85.7	-	84.5

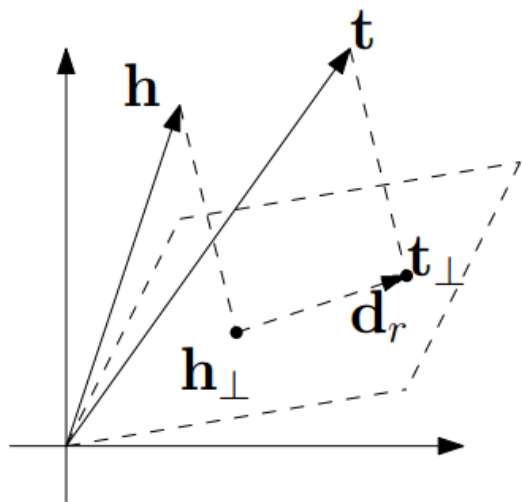
知识表示研究趋势：一对多关系处理

- TransE的假设无法较好处理一对多、多对一、多对多关系

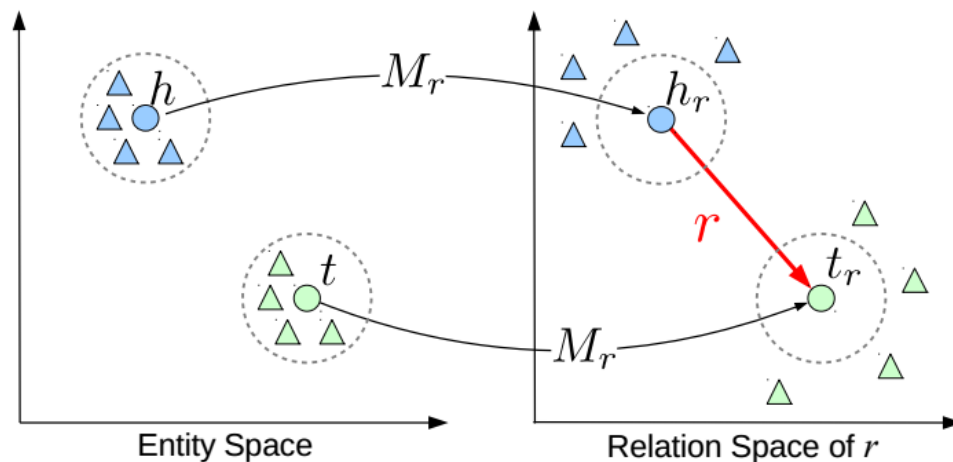


知识表示研究趋势：一对多关系处理

- 在TransE基础上考虑**关系对实体的影响**



TransH



TransR

实验结果：实体预测

Data Sets	WN18				FB15K			
Metric	Mean Rank		Hits@10 (%)		Mean Rank		Hits@10 (%)	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
Unstructured (Bordes et al. 2012)	315	304	35.3	38.2	1,074	979	4.5	6.3
RESCAL (Nickel, Tresp, and Kriegel 2011)	1,180	1,163	37.2	52.8	828	683	28.4	44.1
SE (Bordes et al. 2011)	1,011	985	68.5	80.5	273	162	28.8	39.8
SME (linear) (Bordes et al. 2012)	545	533	65.1	74.1	274	154	30.7	40.8
SME (bilinear) (Bordes et al. 2012)	526	509	54.7	61.3	284	158	31.3	41.3
LFM (Jenatton et al. 2012)	469	456	71.4	81.6	283	164	26.0	33.1
TransE (Bordes et al. 2013)	263	251	75.4	89.2	243	125	34.9	47.1
TransH (unif) (Wang et al. 2014)	318	303	75.4	86.7	211	84	42.5	58.5
TransH (bern) (Wang et al. 2014)	401	388	73.0	82.3	212	87	45.7	64.4
TransR (unif)	232	219	78.3	91.7	226	78	43.8	65.5
TransR (bern)	238	225	79.8	92.0	198	77	48.2	68.7
CTransR (unif)	243	230	78.9	92.3	233	82	44	66.3
CTransR (bern)	231	218	79.4	92.3	199	75	48.4	70.2


+20%




实验结果：关系预测

knowledge_completion.pdf	Data Sets	WN11	FB13	FB15K
	SE	53.0	75.2	-
	SME (bilinear)	70.0	63.7	-
	SLM	69.9	85.3	-
	LFM	73.8	84.3	-
	NTN	70.4	87.1	68.5
	TransE (unif)	75.9	70.9	79.6
	TransE (bern)	75.9	81.5	79.2
	TransH (unif)	77.7	76.5	79.0
	TransH (bern)	78.8	83.3	80.2
	TransR (unif)	85.5	74.7	81.7
	TransR (bern)	85.9	82.5	83.9
	CTransR (bern)	85.7	-	84.5





TransE、TransH、TransR开放源码



- https://github.com/mrlyk423/relation_extraction

 **Mrlyk423 / Relation_Extraction**



 Watch **8**  Star **38**  Fork **21**









AAAI2015 Learning Entity and Relation Embeddings for Knowledge Graph Completion

 **10** commits  **1** branch  **0** releases  Fetching contributors


 branch: **master** **Relation_Extraction** / + 


Add data download link.


 lyk423 authored on 2 Mar latest commit a72d562841 


 CTransR	Add source code of TransR and CTransR.	5 months ago
 TransE	Add data download link.	2 months ago
 TransH	Add training code of TransH	3 months ago
 TransR	Add data download link.	2 months ago
 cluster	Add source code of TransR and CTransR.	5 months ago
 .DS_Store	Add data download link.	2 months ago
 README.md	Add data download link.	2 months ago
 makefile	Add source code of TransR and CTransR.	5 months ago

<> Code


 **Issues**


 **Pull requests**


 **Pulse**


 **Graphs**

HTTPS clone URL

<https://github.com> 

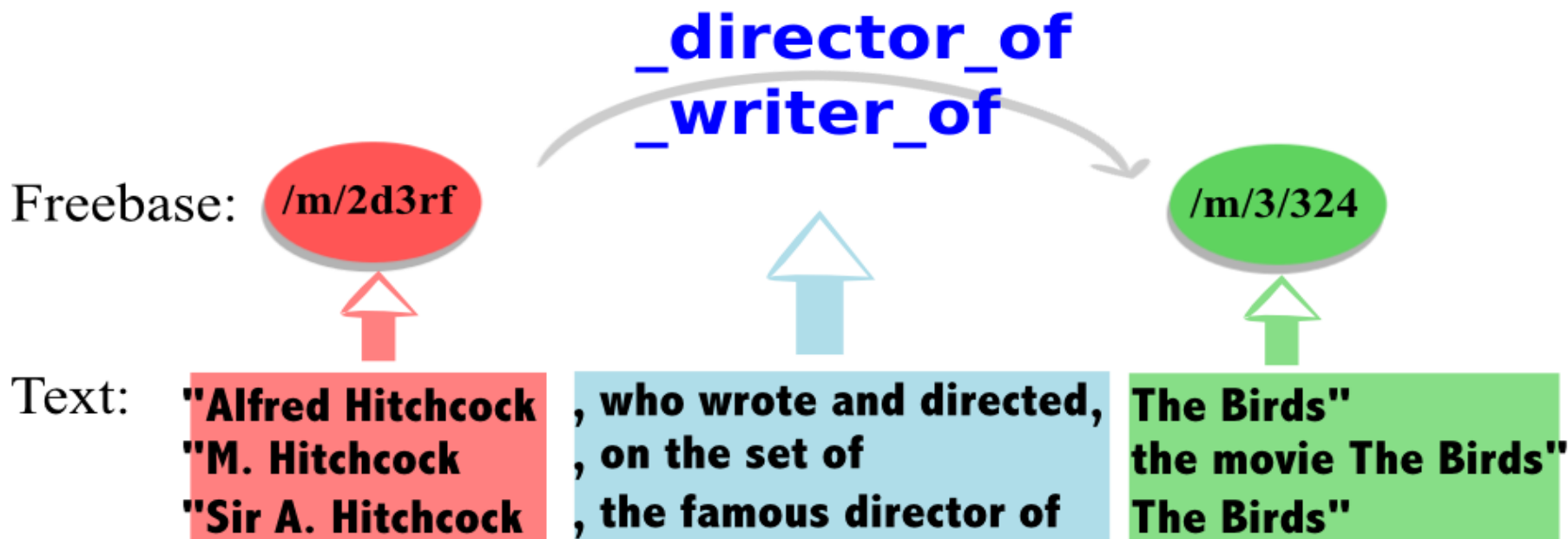
You can clone with **HTTPS** or **Subversion**. 

 **Clone in Desktop**

 **Download ZIP**

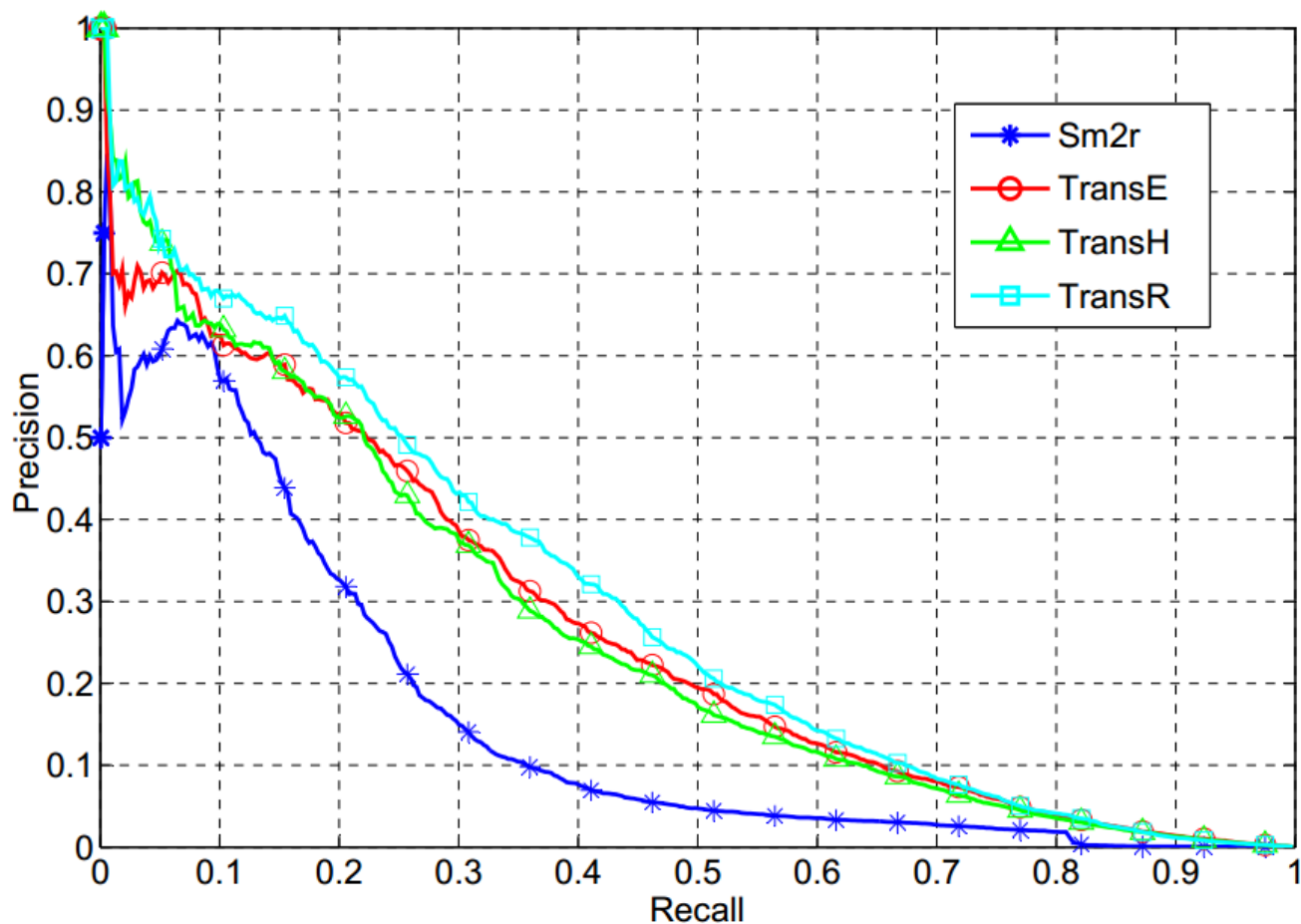
知识表示研究趋势：文本+KG融合

- 将基于文本的关系抽取与知识图谱链接预测相结合



文本+KG融合对关系抽取的帮助

- 数据 NYT+FB (Weston et al.2013)



TransE+Word2Vec

- KG=>TransE, Text=>Word2Vec
- 强制要求同时在KG和文本中出现的实体共享相似的向量

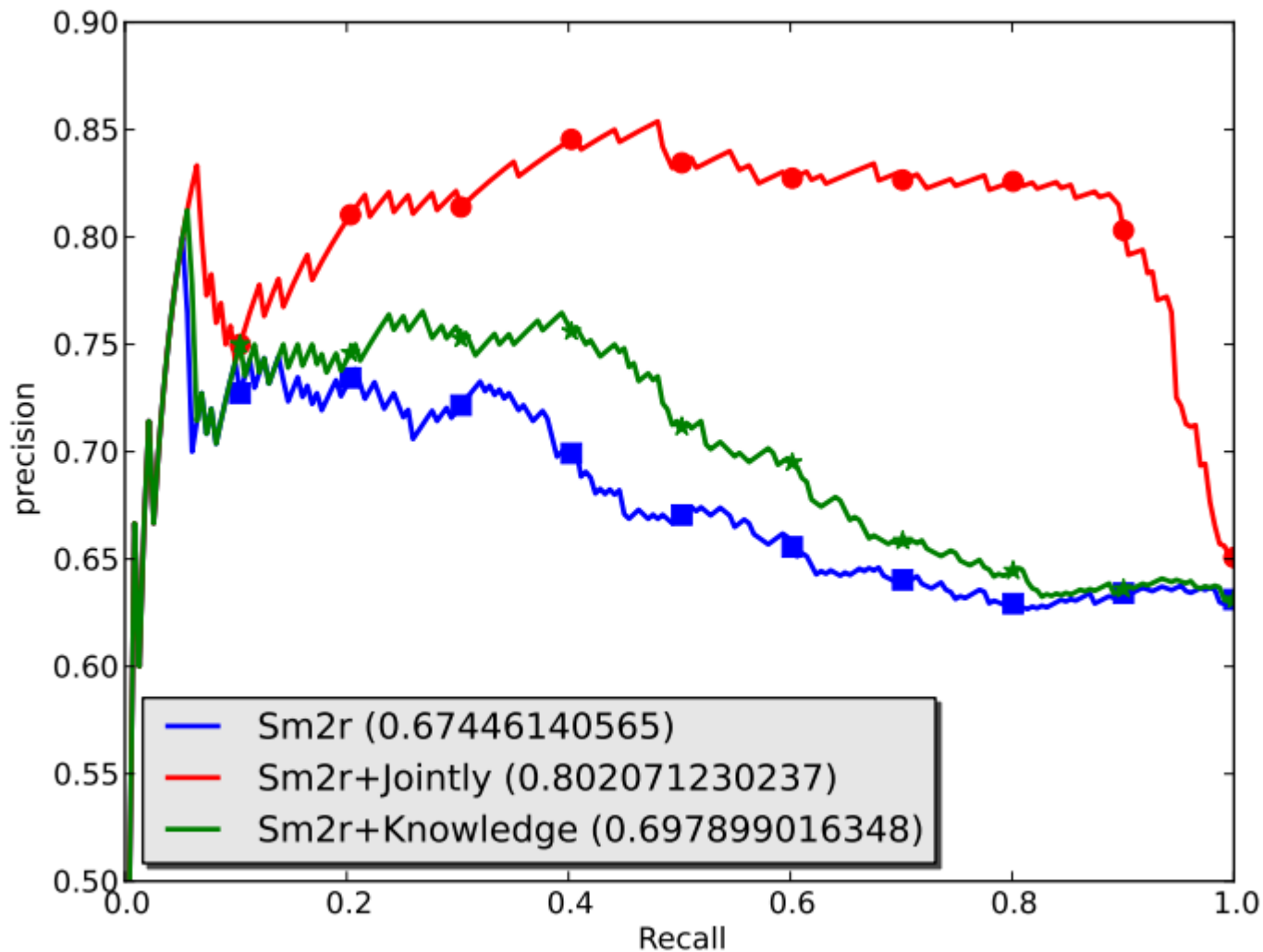
$$\mathcal{L}_K = \sum_{(h,r,t) \in \Delta} \mathcal{L}_f(h, r, t)$$

$$\mathcal{L}_{AN} = \sum_{(h,r,t) \in \Delta} \mathbf{I}_{[w_h \in \mathcal{V} \wedge w_t \in \mathcal{V}]} \cdot \mathcal{L}_f(w_h, r, w_t) +$$

$$\mathbf{I}_{[w_h \in \mathcal{V}]} \cdot \mathcal{L}_f(w_h, r, t) + \mathbf{I}_{[w_t \in \mathcal{V}]} \cdot \mathcal{L}_f(h, r, w_t)$$

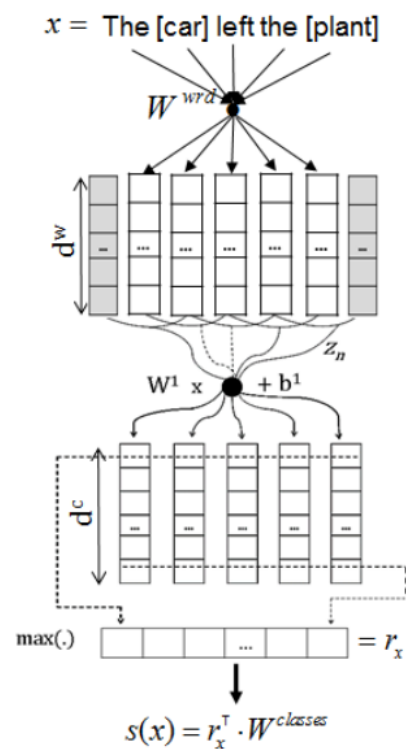
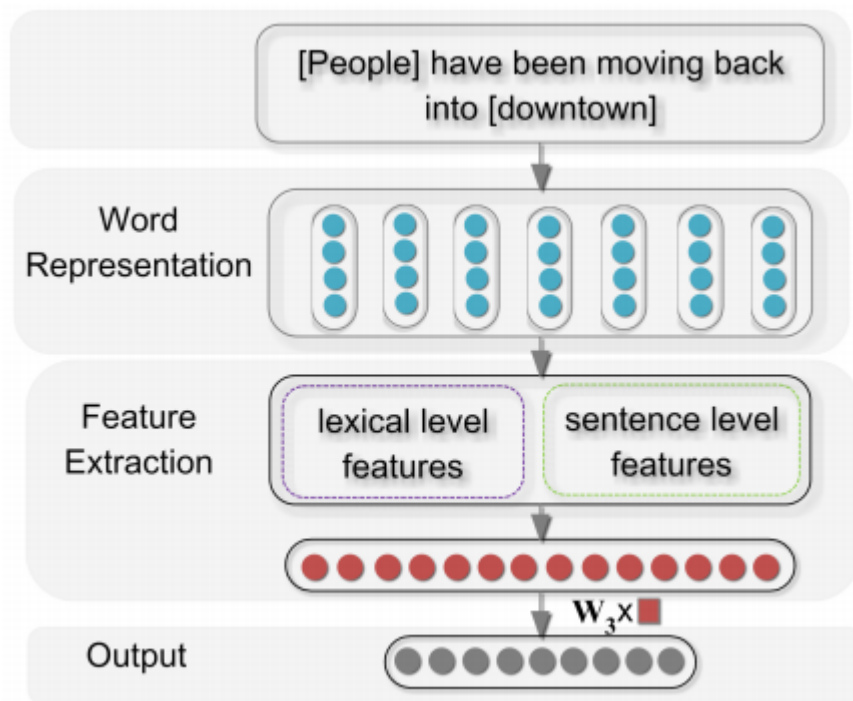
TransE+Word2Vec的效果

- 数据 NYT+FB



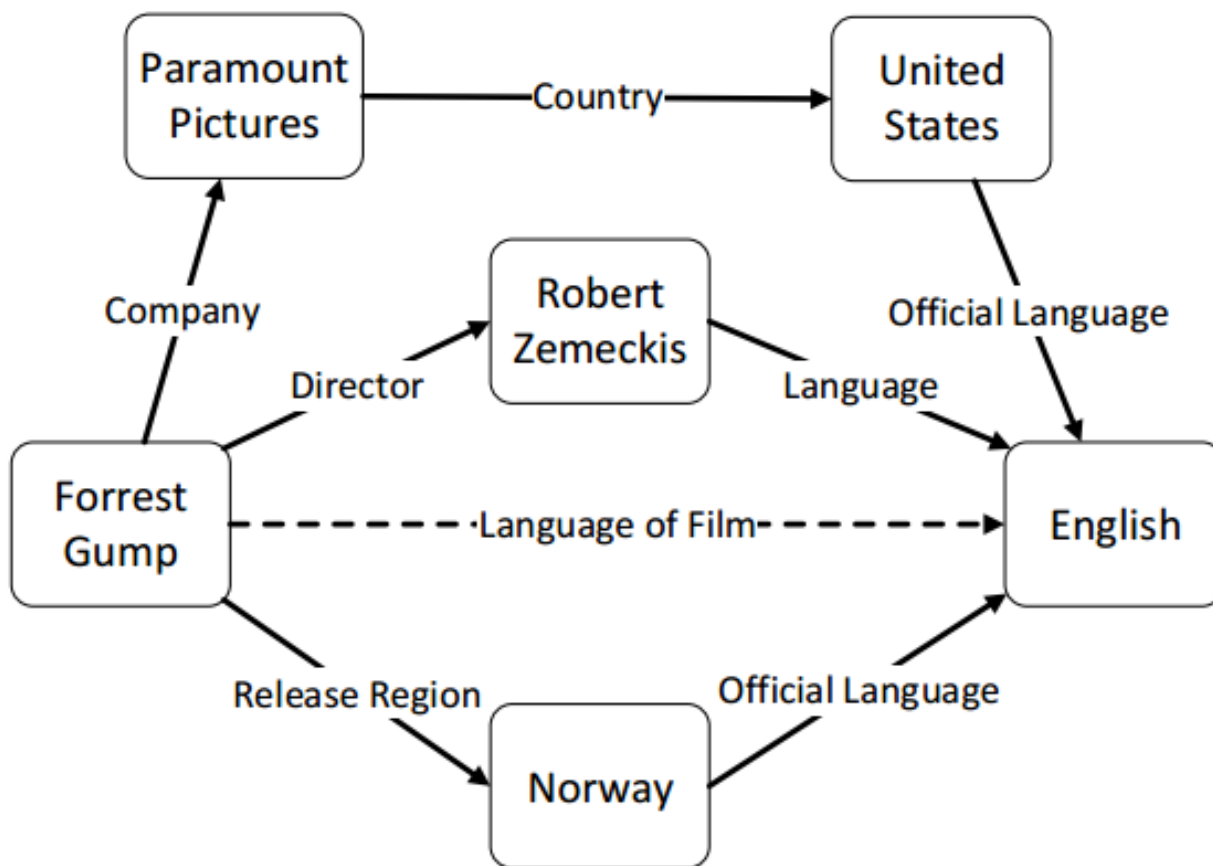
文本+KG融合的挑战问题

- 如何有效结合面向文本关系抽取的最新成果
 - 基于CNN的关系抽取模型
- 建立对词汇、实体和关系的统一表示空间



知识表示研究趋势：关系路径表示

- KG中的关系之间存在复杂的推理关系



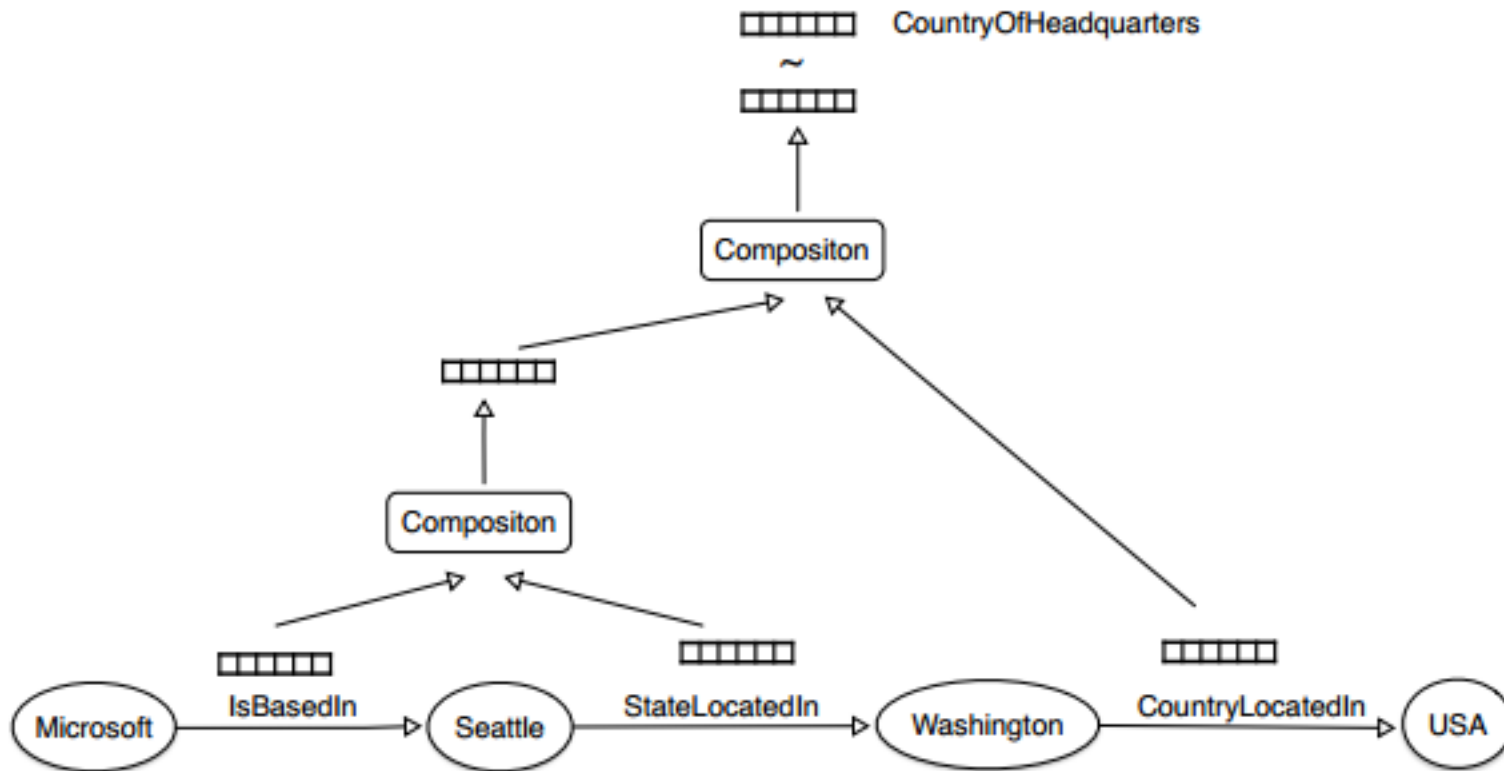
对关系路径建模的传统方法

- Path Ranking Algorithm

ID	PRA Path (Comment)
athletePlaysForTeam	
1	$c \xrightarrow{\text{athletePlaysInLeague}} c \xrightarrow{\text{leaguePlayers}} c \xrightarrow{\text{athletePlaysForTeam}} c$ (teams with many players in the athlete's league)
2	$c \xrightarrow{\text{athletePlaysInLeague}} c \xrightarrow{\text{leagueTeams}} c \xrightarrow{\text{teamAgainstTeam}} c$ (teams that play against many teams in the athlete's league)
athletePlaysInLeague	
3	$c \xrightarrow{\text{athletePlaysSport}} c \xrightarrow{\text{players}} c \xrightarrow{\text{athletePlaysInLeague}} c$ (the league that players of a certain sport belong to)
4	$c \xrightarrow{\text{isa}} c \xrightarrow{\text{isa}^{-1}} c \xrightarrow{\text{athletePlaysInLeague}} c$ (popular leagues with many players)
athletePlaysSport	
5	$c \xrightarrow{\text{isa}} c \xrightarrow{\text{isa}^{-1}} c \xrightarrow{\text{athletePlaysSport}} c$ (popular sports of all the athletes)
6	$c \xrightarrow{\text{athletePlaysInLeague}} c \xrightarrow{\text{superpartOfOrganization}} c \xrightarrow{\text{teamPlaysSport}} c$ (popular sports of a certain league)
stadiumLocatedInCity	
7	$c \xrightarrow{\text{stadiumHomeTeam}} c \xrightarrow{\text{teamHomeStadium}} c \xrightarrow{\text{stadiumLocatedInCity}} c$ (city of the stadium with the same team)
8	$c \xrightarrow{\text{latitudeLongitude}} c \xrightarrow{\text{latitudeLongitudeOf}} c \xrightarrow{\text{stadiumLocatedInCity}} c$ (city of the stadium with the same location)
teamHomeStadium	
9	$c \xrightarrow{\text{teamPlaysInCity}} c \xrightarrow{\text{cityStadiums}} c$ (stadiums located in the same city with the query team)
10	$c \xrightarrow{\text{teamMember}} c \xrightarrow{\text{athletePlaysForTeam}} c \xrightarrow{\text{teamHomeStadium}} c$ (home stadium of teams which share players with the query)
teamPlaysInCity	
11	$c \xrightarrow{\text{teamHomeStadium}} c \xrightarrow{\text{stadiumLocatedInCity}} c$ (city of the team's home stadium)
12	$c \xrightarrow{\text{teamHomeStadium}} c \xrightarrow{\text{stadiumHomeTeam}} c \xrightarrow{\text{teamPlaysInCity}} c$ (city of teams with the same home stadium as the query)
teamPlaysInLeague	
13	$c \xrightarrow{\text{teamPlaysSport}} c \xrightarrow{\text{players}} c \xrightarrow{\text{athletePlaysInLeague}} c$ (the league that the query team's members belong to)
14	$c \xrightarrow{\text{teamPlaysAgainstTeam}} c \xrightarrow{\text{teamPlaysInLeague}} c$ (the league that the query team's competing team belongs to)
teamPlaysSport	
15	$c \xrightarrow{\text{isa}} c \xrightarrow{\text{isa}^{-1}} c \xrightarrow{\text{teamPlaysSport}} c$ (sports played by many teams)
16	$c \xrightarrow{\text{teamPlaysInLeague}} c \xrightarrow{\text{leagueTeams}} c \xrightarrow{\text{teamPlaysSport}} c$ (the sport played by other teams in the league)

关系路径的表示学习方法

- Recursive Neural Network (RNN)

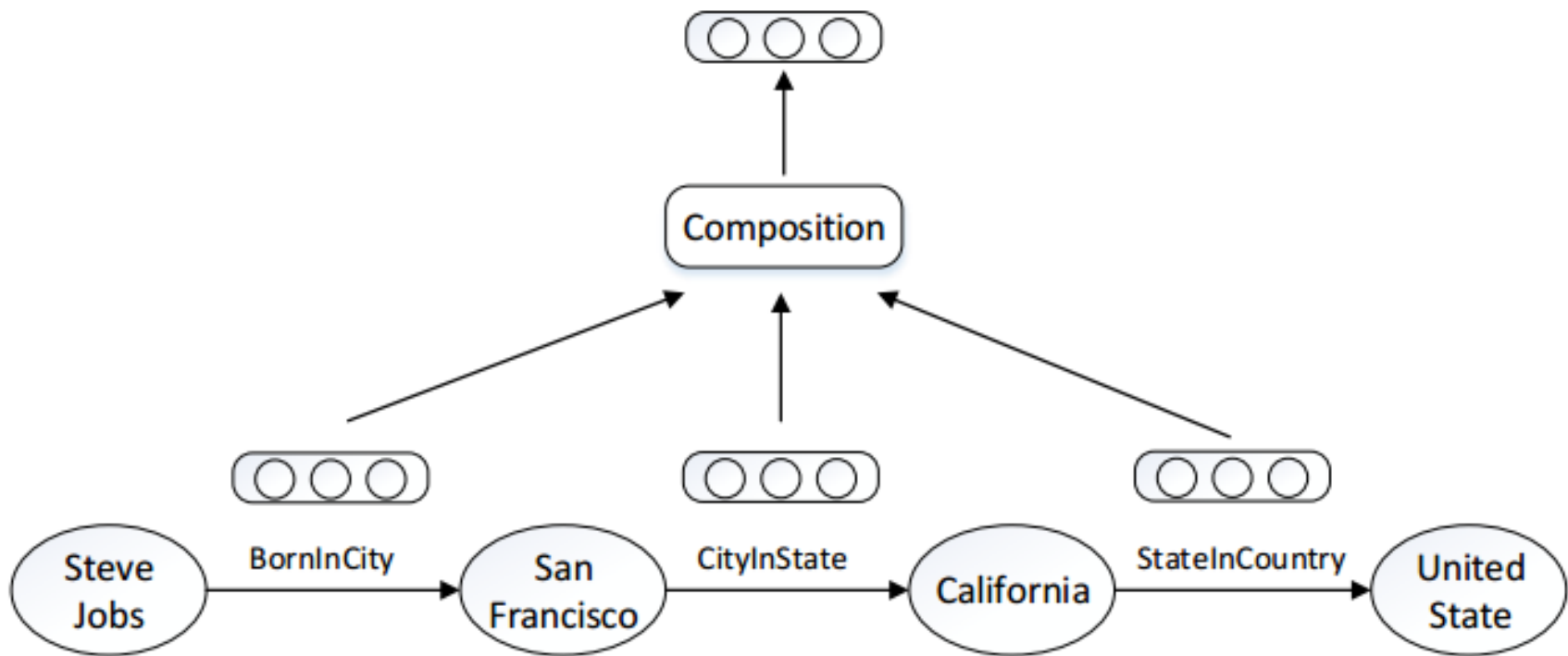


考虑关系路径的TransE

	TransE	PTransE
KB	$h \xrightarrow{r} t$	$h \xrightarrow{r_1} e_1 \xrightarrow{r_2} t$
Triples	(h, r, t)	$(h, r_1, e_1) \quad (e_1, r_2, t)$ $(h, r_1 \circ r_2, t)$
Objectives	$\mathbf{h} + \mathbf{r} = \mathbf{t}$	$\mathbf{h} + \mathbf{r}_1 = \mathbf{e}_1 \quad \mathbf{e}_1 + \mathbf{r}_2 = \mathbf{t}$ $\mathbf{h} + (\mathbf{r}_1 \circ \mathbf{r}_2) = \mathbf{t}$

考虑关系路径的TransE

- 关系路径的向量表示问题
- 组合语义：相加，相乘，RNN



Gardner, M., Talukdar, P. P., Kisiel, B., & Mitchell, T. (2013). Improving learning and inference in a large knowledge-base using latent syntactic cues. In EMNLP.

考虑关系路径的TransE：实体预测

Metric	Mean Rank		Hits@10 (%)	
	Raw	Filter	Raw	Filter
RESCAL	828	683	28.4	44.1
SE	273	162	28.8	39.8
SME (linear)	274	154	30.7	40.8
SME (bilinear)	284	158	31.3	41.3
LFM	283	164	26.0	33.1
TransE	243	125	34.9	47.1
TransH	212	87	45.7	64.4
TransR	198	77	48.2	68.7
TransE (Our)	205	63	47.9	70.2
PTransE (ADD, 2-step)	200	54	51.8	83.4
PTransE (MUL, 2-step)	216	67	47.4	77.7
PTransE (RNN, 2-step)	242	92	50.6	82.2
PTransE (ADD, 3-step)	207	58	51.4	84.6

+35%

Lin, Y., Liu, Z., & Sun, M. (2015). Modeling Relation Paths for Representation Learning of Knowledge Bases. arXiv preprint arXiv:1506.00379.

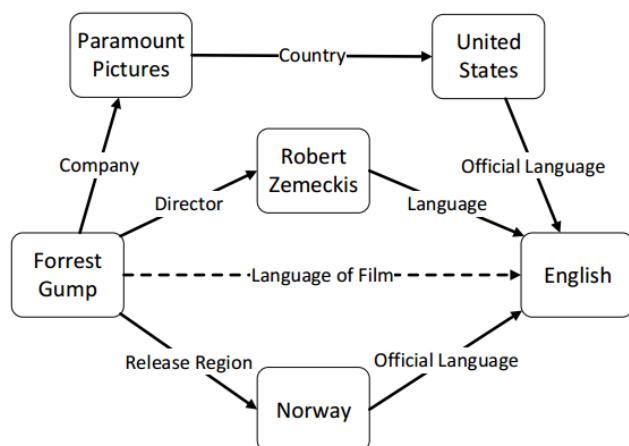
考虑关系路径的TransE：关系预测

Metric	Mean Rank		Hits@1 (%)	
	Raw	Filter	Raw	Filter
TransE	2.8	2.5	65.1	84.3
+Rev	2.6	2.3	67.1	86.7
+Rev+Path	2.4	1.9	65.2	89.0
PTransE (ADD, 2-step)	1.7	1.2	69.5	93.6
-TransE	135.8	135.3	51.4	78.0
-Path	2.0	1.6	69.7	89.0
PTransE (MUL, 2-step)	2.5	2.0	66.3	89.0
PTransE (RNN, 2-step)	1.9	1.4	68.3	93.2
PTransE (ADD, 3-step)	1.8	1.4	68.5	94.0

+10%

关系路径建模的挑战问题

- 如何寻找关系间的复杂推理关系
 - 更多类型推理关系



(奥巴马, 总统, 美国)



(奥巴马, 是, 美国人)

- 推理关系可信性
- 如何表示利用关系间的推理关系
 - 利用组合语义特性: RNN、NTN、...

知识表示学习的其他挑战问题

- 大规模知识图谱表示的**快速学习**
 - 长尾数据
 - 在线学习、分布式学习
- 融合**外部信息**的知识表示学习
 - 利用文本、实体和关系的属性等外部信息
 - 建立统一的知识表示空间
- 考虑**常识信息**的知识表示学习与信息抽取
 - 先验知识（如人的结婚年龄、毕业年龄等）
- 知识表示在信息融合、知识推理中的应用
 - 跨语言、跨知识库的知识融合
 - 在低维向量空间中的知识推理

谢谢老师同学！
欢迎批评指正！