第三届全国中文知识图谱研讨会

大规模知识图谱的表示学习

趋势与挑战

清华大学自然语言处理实验室

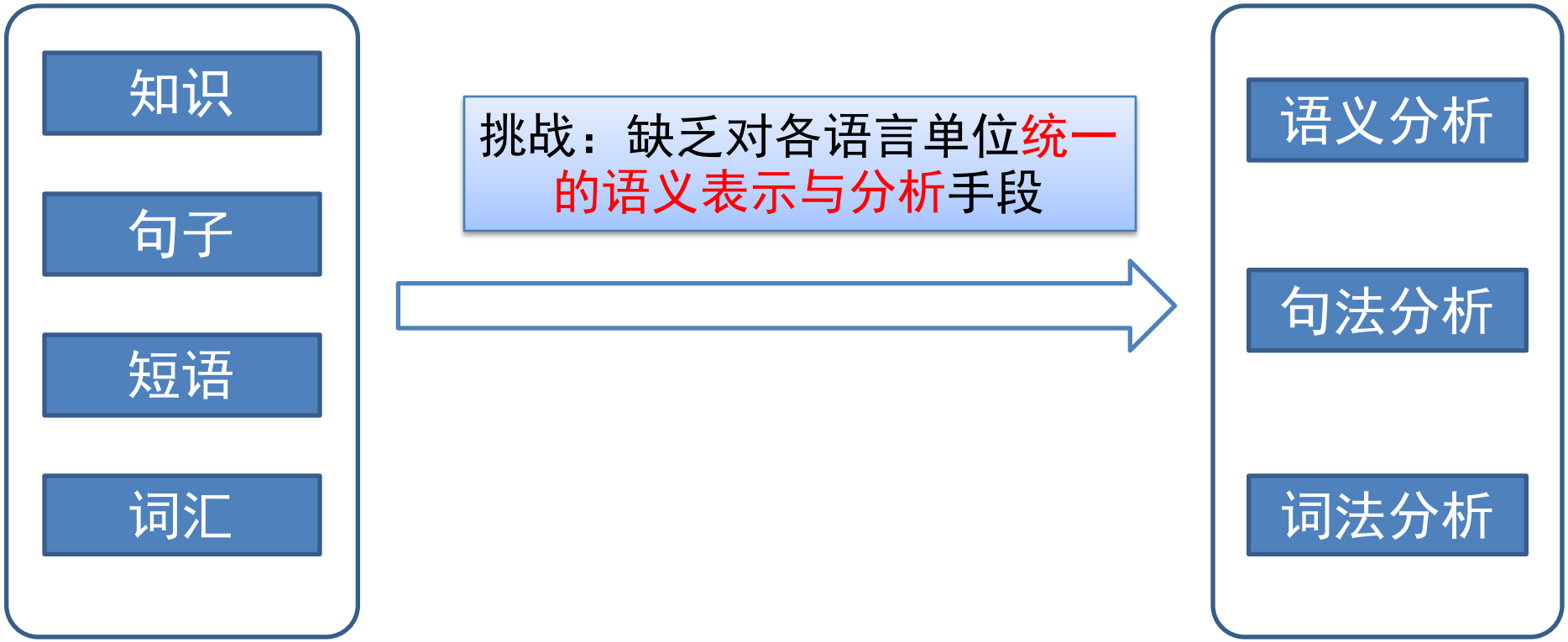
刘知远

liuzy@tsinghua.edu.cn

机器学习 = 数据表示 + 学习目标 + 优化方法

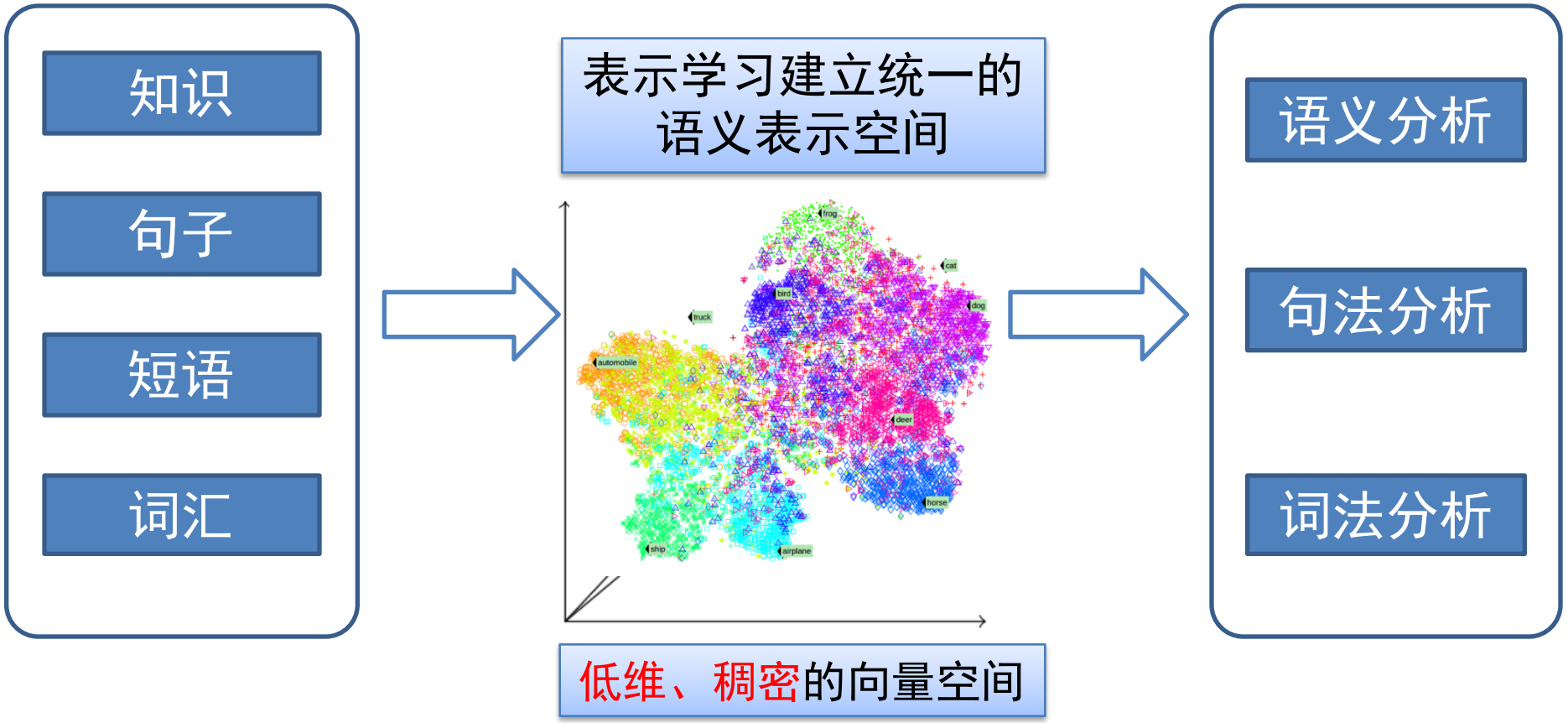
Yoshua Bengio. Deep Learning of Representations. AAAI 2013 Tutorial.

表示学习的意义



表示学习的意义

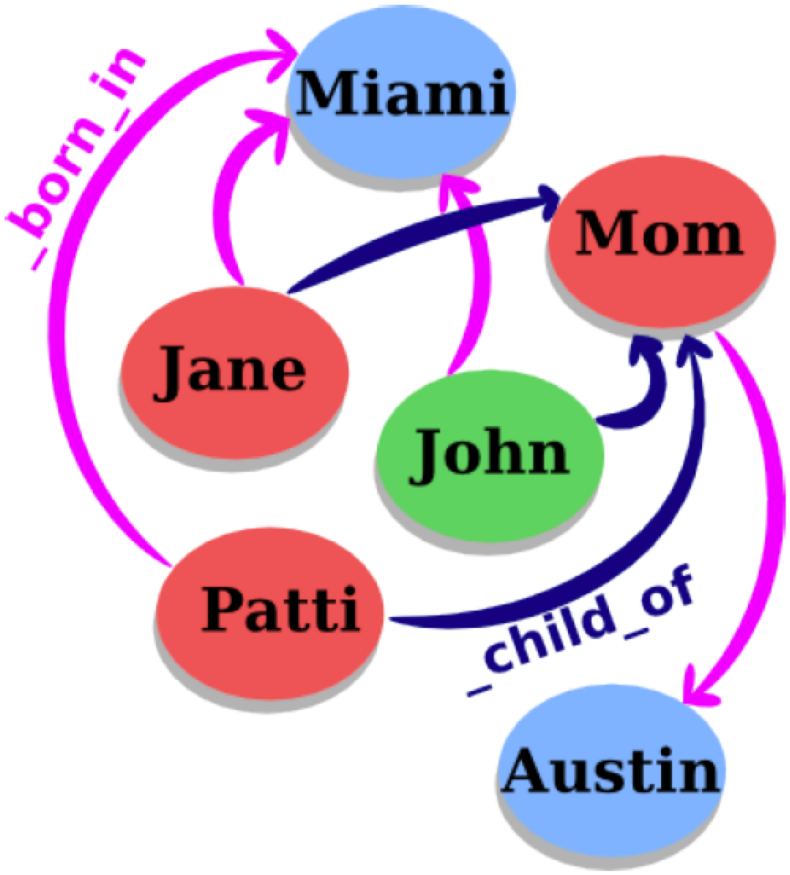
- 缓解数据稀疏，建立统一空间，实现知识迁移



分布式表示（distributed representation）/嵌入（embeddings）

知识图谱中的实体与关系

- 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

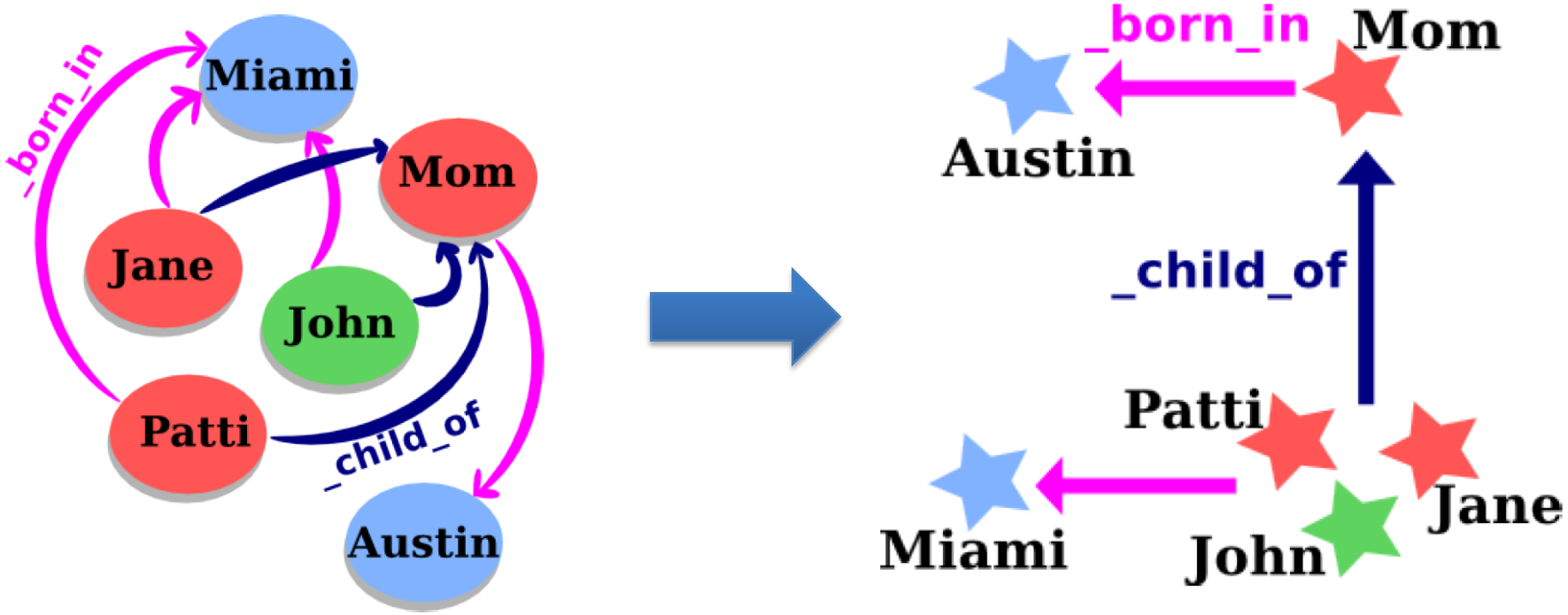


大规模知识图谱的表示学习问题

- 大规模知识图谱特点
  - 高维、稀疏、噪音、不完整
- 大规模知识图谱传统表示方法
  - 稀疏矩阵三元组，RDF，符号化表示每个实体/关系
  - 无法有效度量和利用实体间语义关联
- 研究思路：将知识图谱嵌入到低维向量空间
  - 实体和关系都表示为低维向量
  - 有效表示和度量实体、关系间的语义关联
- 应用场景：知识获取，知识推理，知识融合

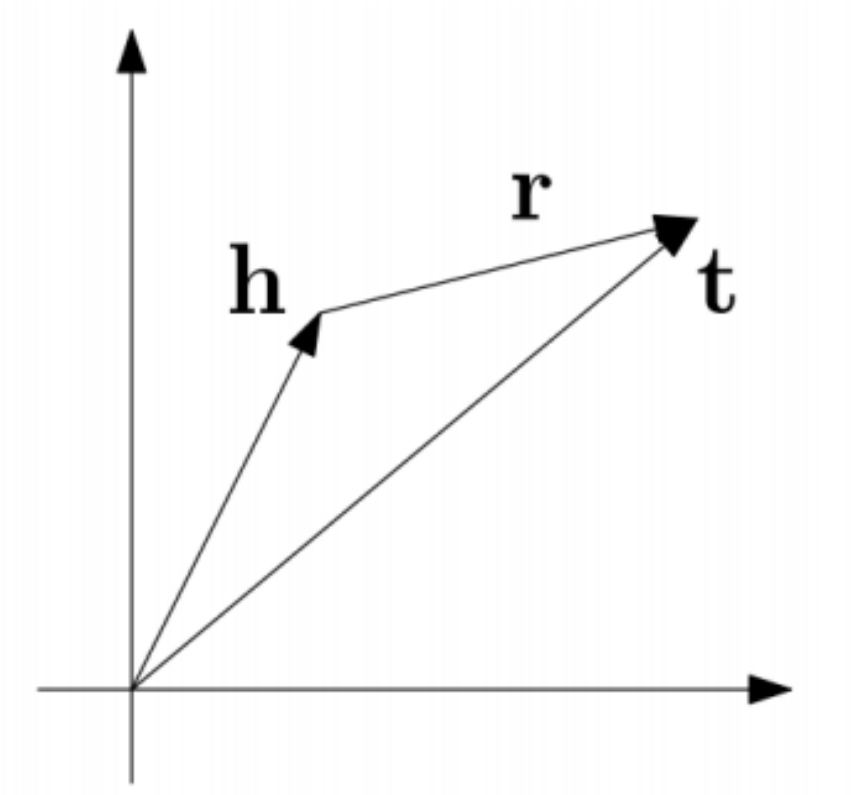
知识表示代表模型：TransE

- 对每个事实 (head, relation, tail)，将relation看做从head到tail的翻译操作
- 优化目标：h + r = t

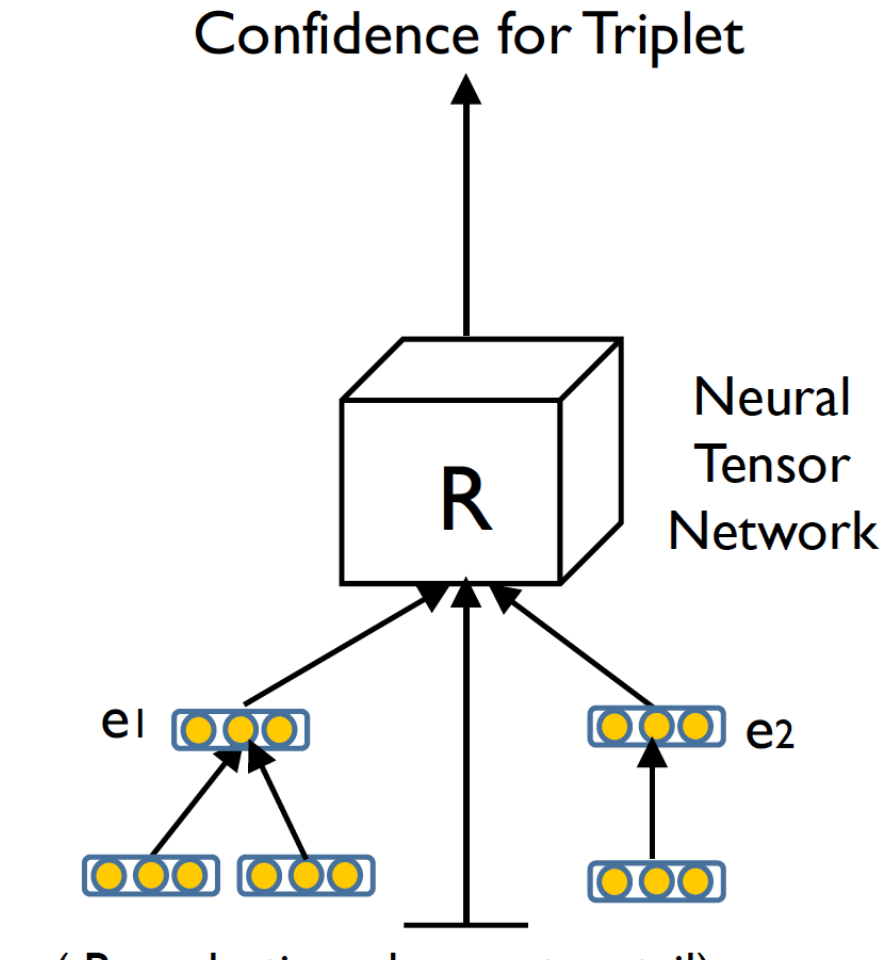


知识表示代表模型：TransE

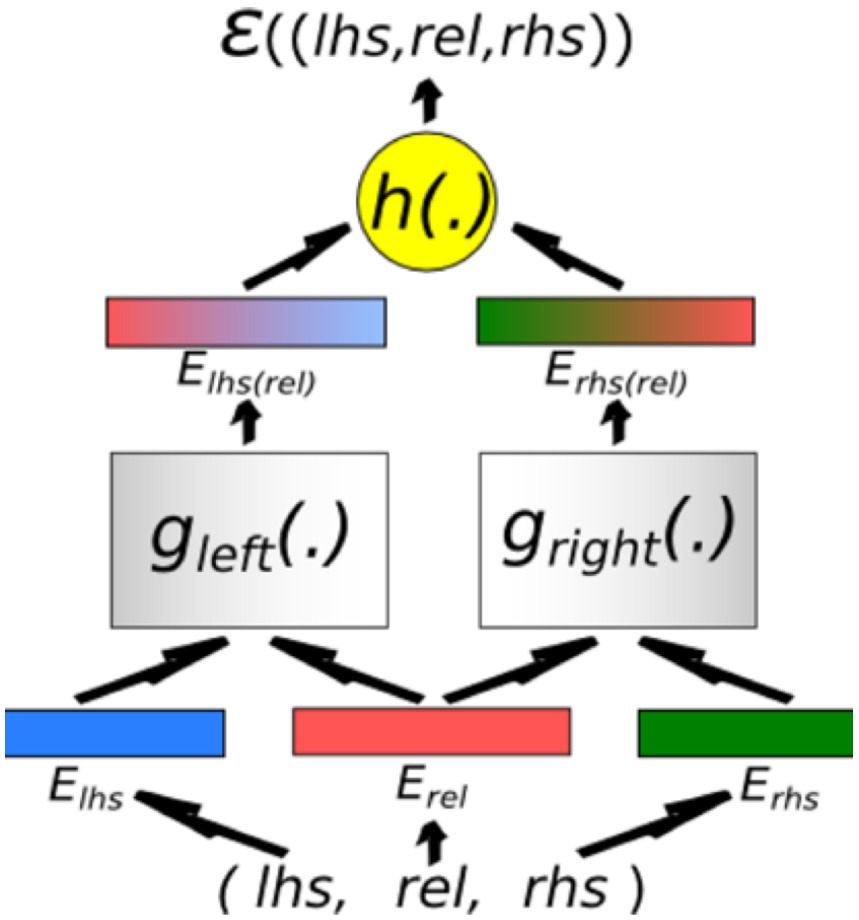
- 对每个事实 (head, relation, tail)，将relation看做从head到tail的翻译操作
- 优化目标：h + r = t



其他代表模型



Neural Tensor Network (NTN)



Energy Model

评测任务：链接预测

WALL-E　　电影风格　　？

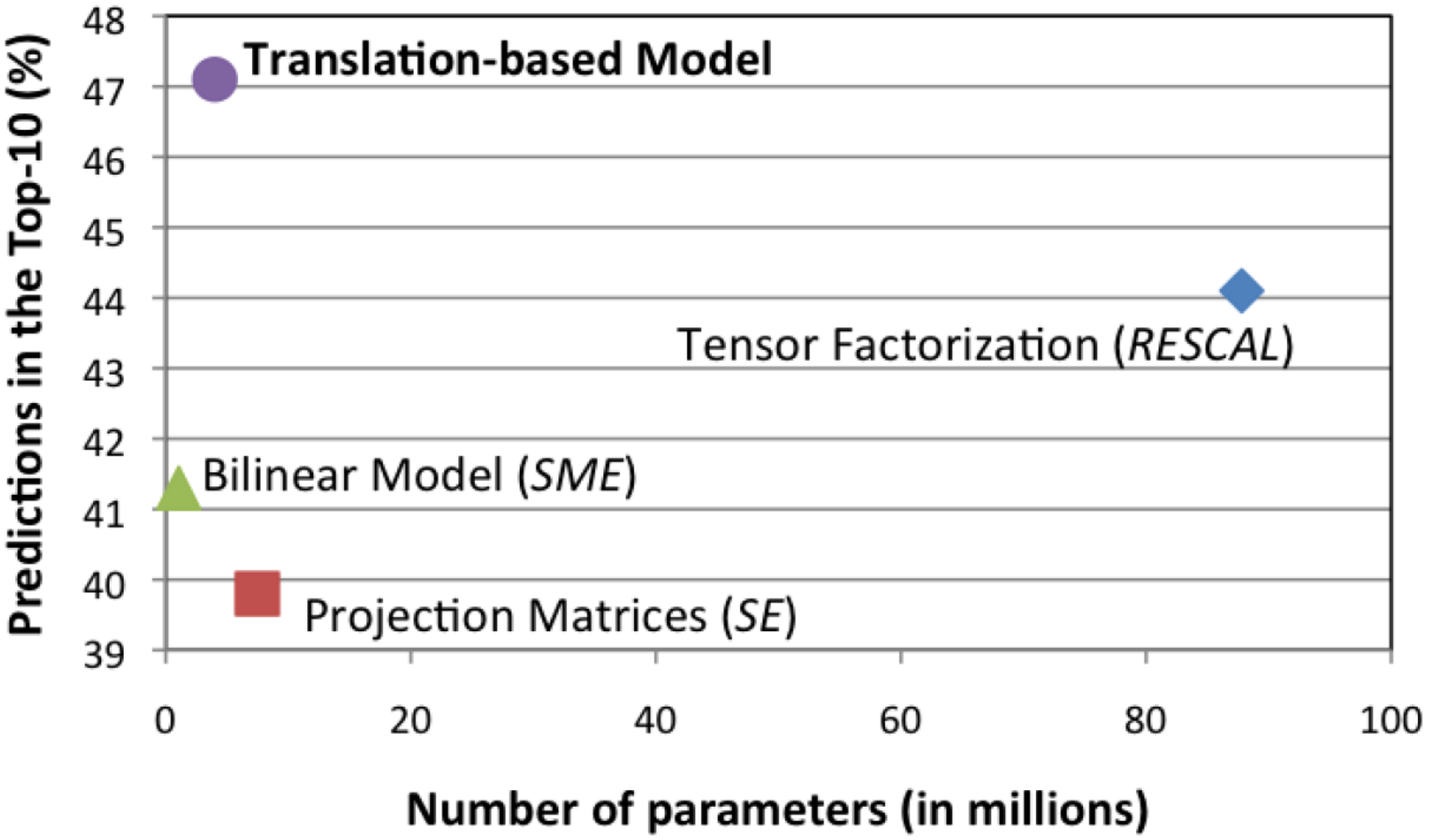
评测任务：链接预测

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


Translation-based Model
Tensor Factorization (*RESCAL*)
Bilinear Model (*SME*)
Projection Matrices (*SE*)
Predictions in the Top-10 (%)
48
47
46
45
44
43
42
41
40
39
0
20
40
60
80
100
Number of parameters (in millions)

评测数据集合

Dataset	#Rel	#Ent	#Train	#Valid	# Test
WN18	18	40,943	141,442	5,000	5,000
FB15K	1,345	14,951	483,142	50,000	59,071
WN11	11	38,696	112,581	2,609	10,544
FB13	13	75,043	316,232	5,908	23,733
FB40K	1,336	39528	370,648	67,946	96,678

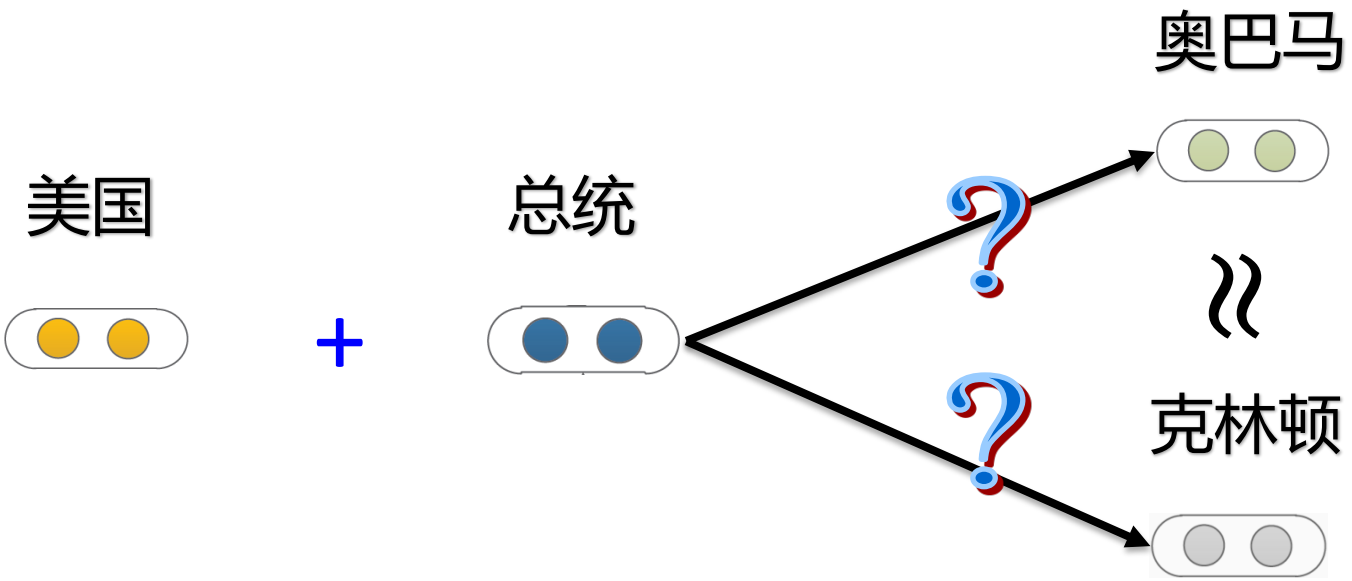
TransE给我们的启示

- 创新性地设计学习目标
- 模型复杂度与知识图谱稀疏性存在辩证关系

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**

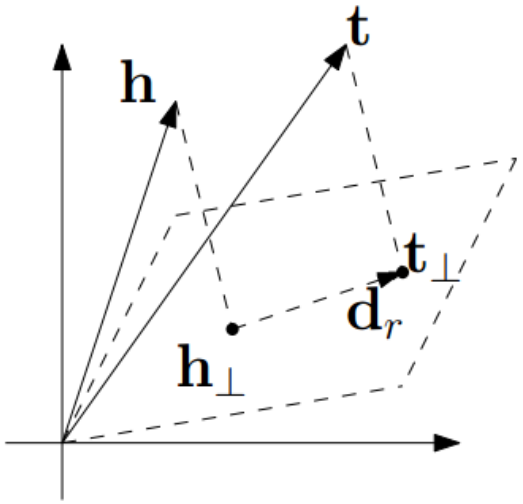
知识表示研究趋势：一对多关系处理

- TransE的假设无法较好处理一对多、多对一、多对多关系

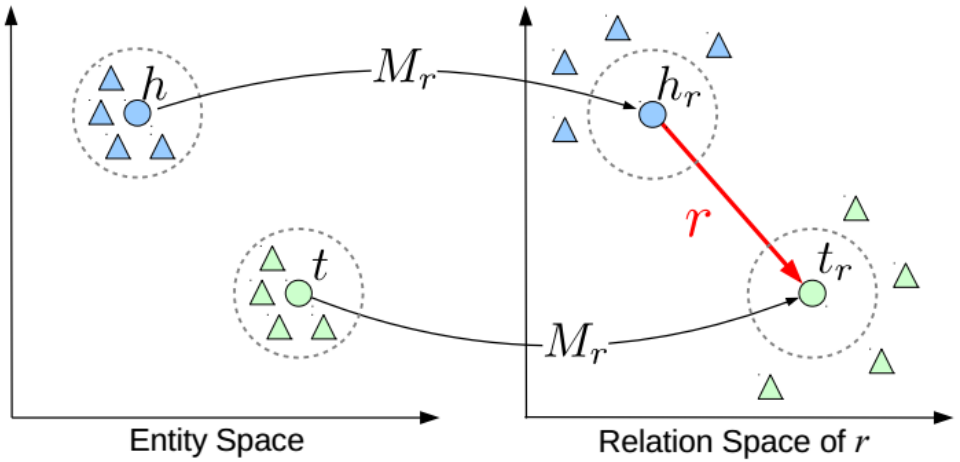


知识表示研究趋势：一对多关系处理

- 在TransE基础上考虑关系对实体的影响



TransH



TransR

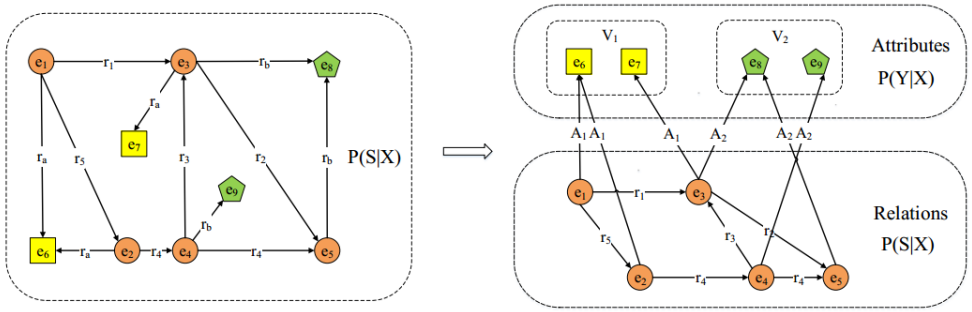
Wang, et al. (2014). Knowledge graph embedding by translating on hyperplanes. AAAI.
Lin, et al. (2015). Learning entity and relation embeddings for knowledge graph completion. AAAI.

实验结果：链接预测

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**

复杂关系建模的挑战

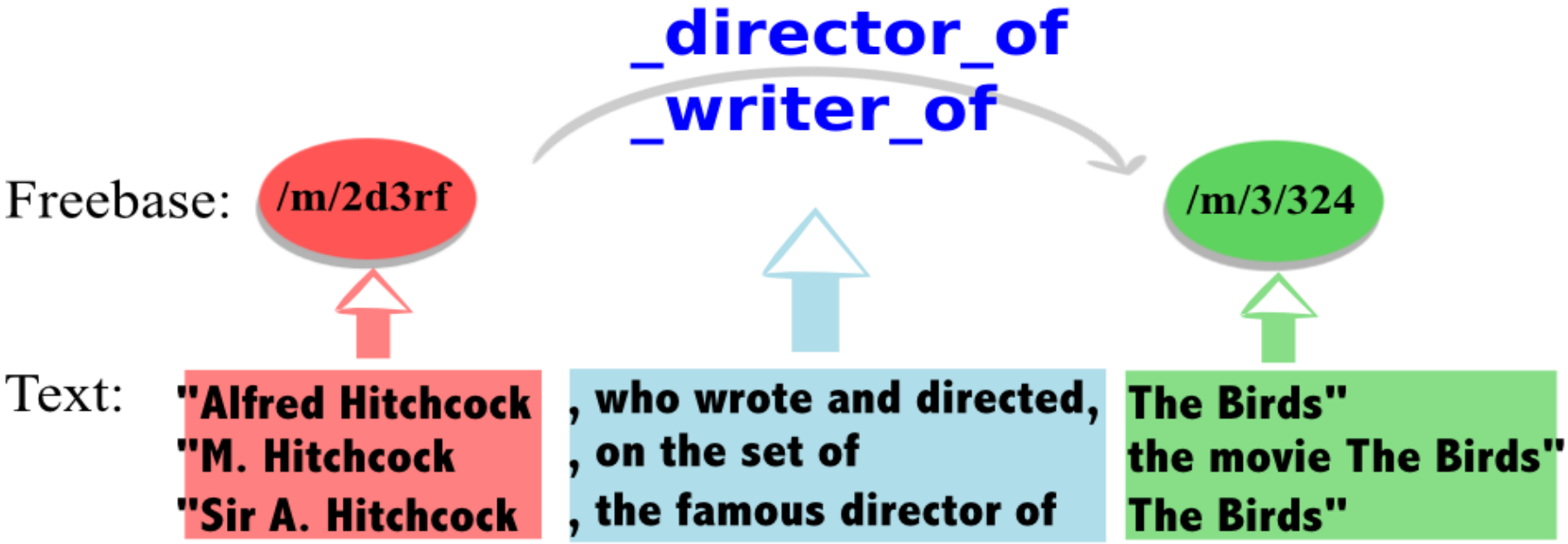
- 研究热点
  - TransA，TransD，TransE，TransG，TransH，TransR
- 挑战问题
  - TransE简单目标无法应对复杂关系（1-N、N-1、N-N）
  - 建议方案：将关系划分为不同类型，分而治之
- FB关系➔属性（性别、国籍）、关系



Metric	Mean Rank		Hits@1 (%)	
	Raw	Filter	Raw	Filter
TransE	3.2	3.0	68.9	84.3
TransH	3.4	3.2	67.7	83.6
TrasnR	3.3	3.1	67.3	84.5
KR-EAR(TransE)	2.8	2.5	69.2	87.6
+ CRA	2.3	2.0	**73.7**	90.3
KR-EAR(TransR)	2.6	2.3	66.0	87.7
+ CRA	**2.0**	**1.7**	72.0	**90.7**

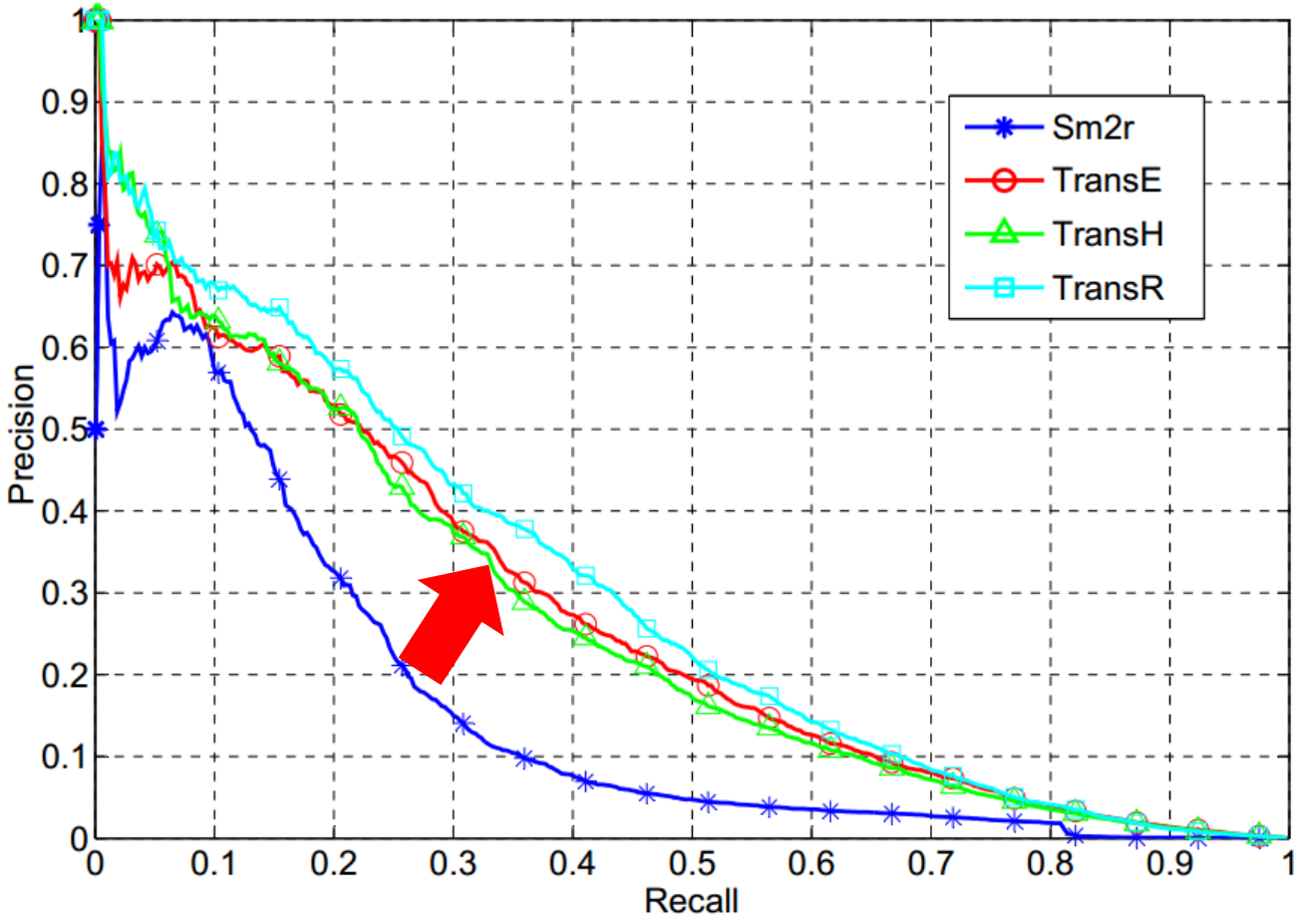
知识表示研究趋势：文本+KG融合

- 基于文本序列的关系抽取
- 基于知识图谱的链接预测



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

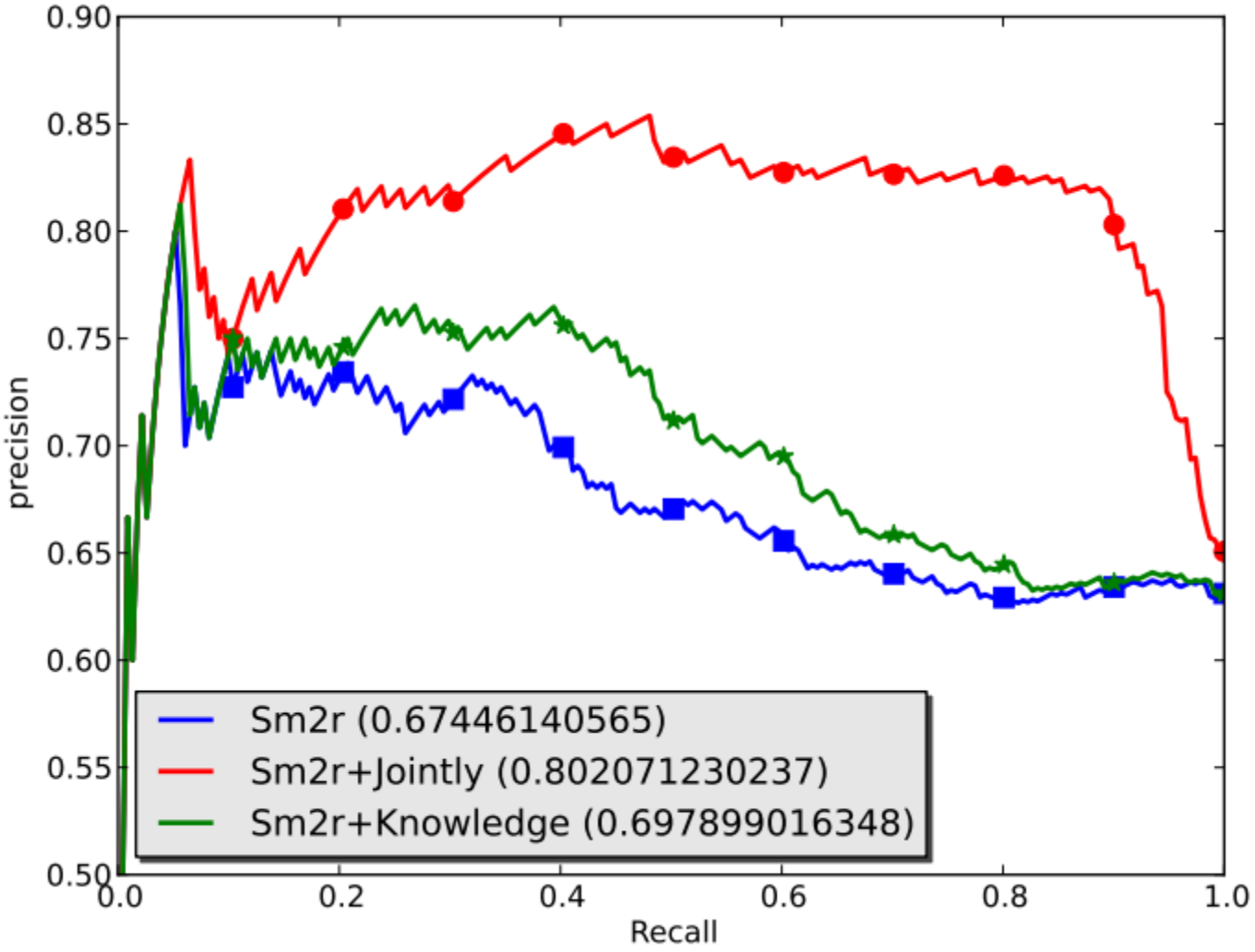
$$\mathcal{L}_K = \sum_{(h,r,t)\in\Delta} \mathcal{L}_f(h,r,t)$$

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

Wang, et al. (2014). Knowledge graph and text jointly embedding. In EMNLP.

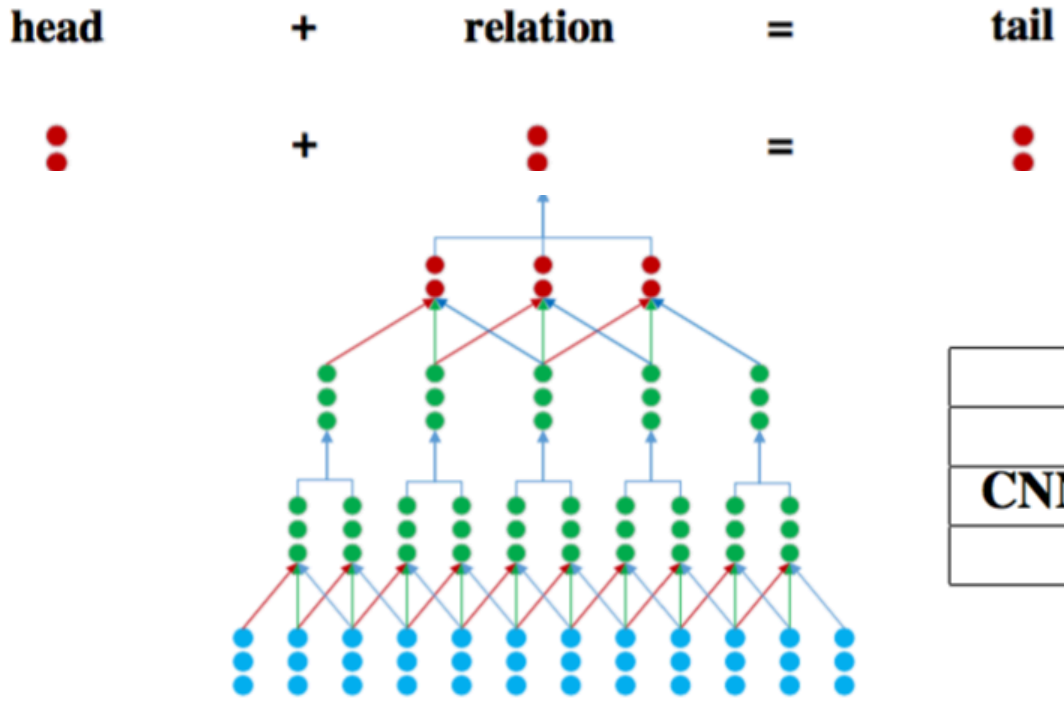
TransE+Word2Vec的效果

- 数据 NYT+FB



文本+KG融合的挑战问题

- 结合文本关系抽取的最新方法：CNN等
- 建立对词汇、实体和关系的统一表示空间

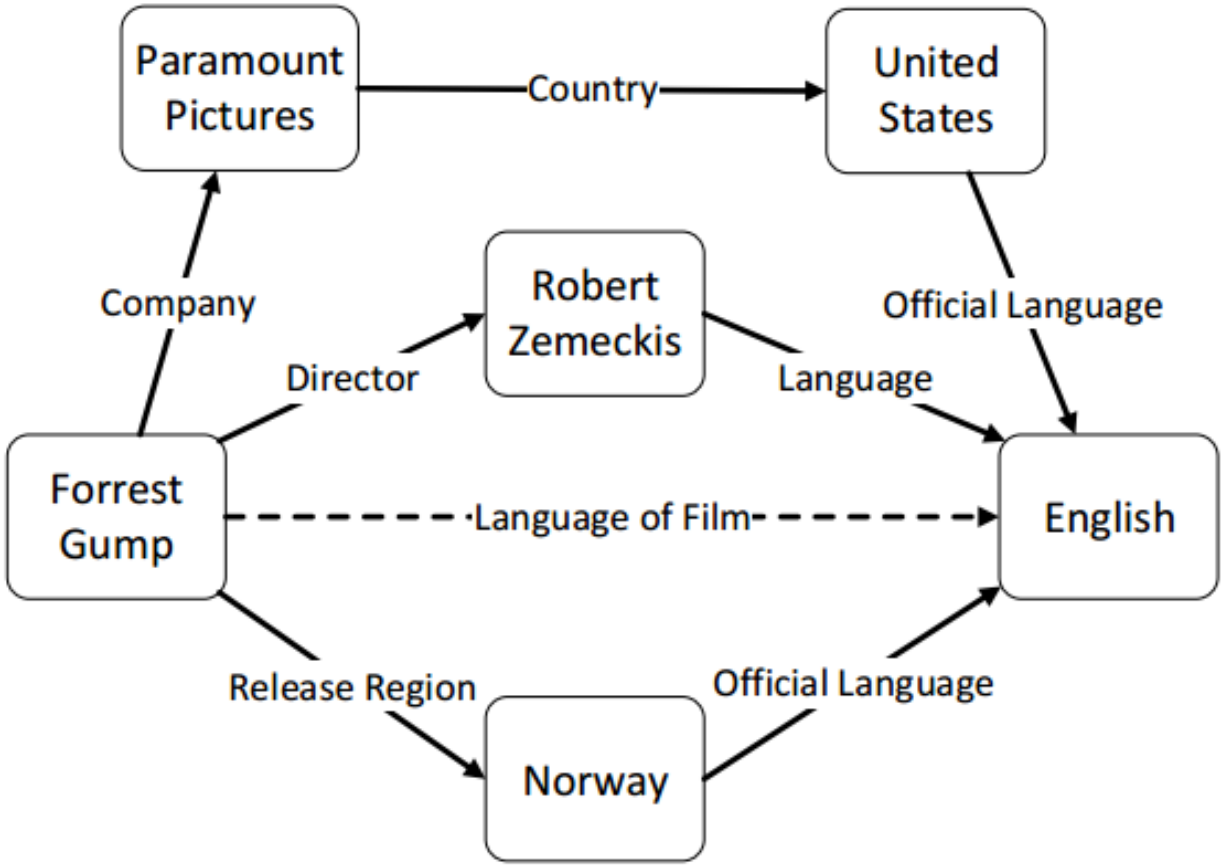


Type	Accuracy
TransE	0.38
CNN(We implemented)	0.68
TransE+CNN	0.74

Zeng, et al. (2014). Relation classification via convolutional deep neural network. COLING.
Santos, et al. (2015). Classifying Relations by Ranking with Convolutional Neural Networks. In ACL.

知识表示研究趋势：关系路径表示

- KG的实体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

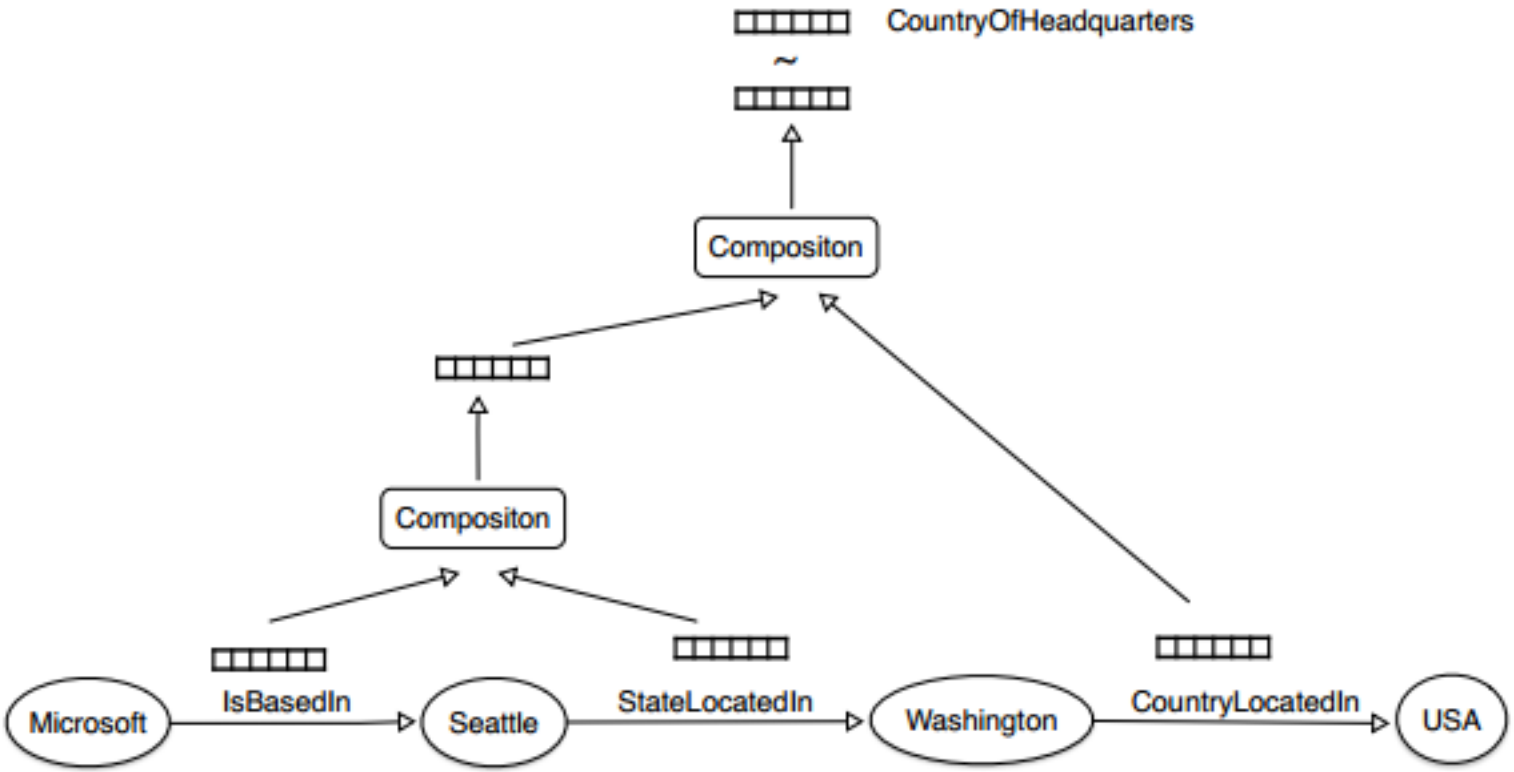
- Path Ranking Algorithm

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)

Lao, et al. (2011). Random walk inference and learning in a large scale knowledge base. EMNLP.

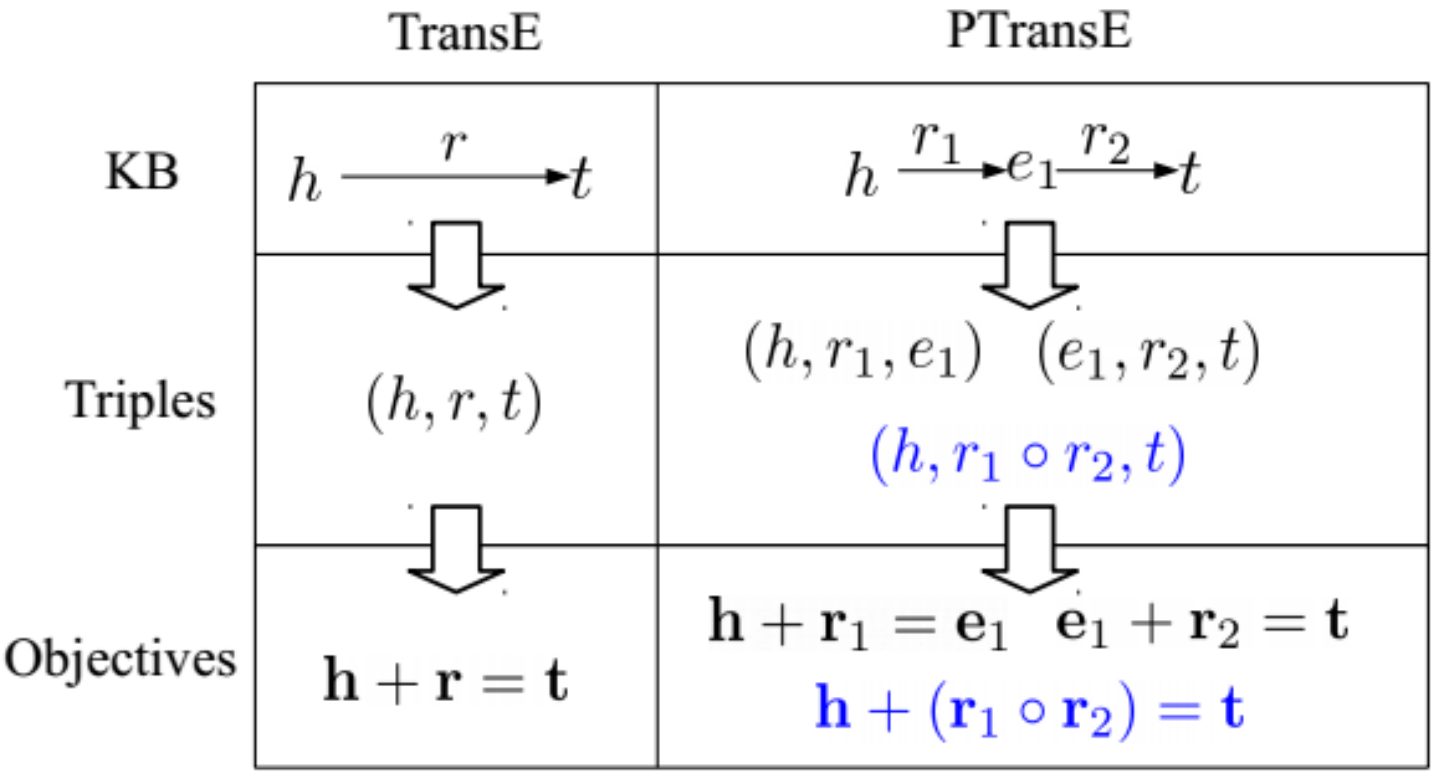
关系路径的表示学习

- Recursive Neural Network (RNN)



Neelakantan, et al. (2015). Compositional Vector Space Models for Knowledge Base Completion. ACL.

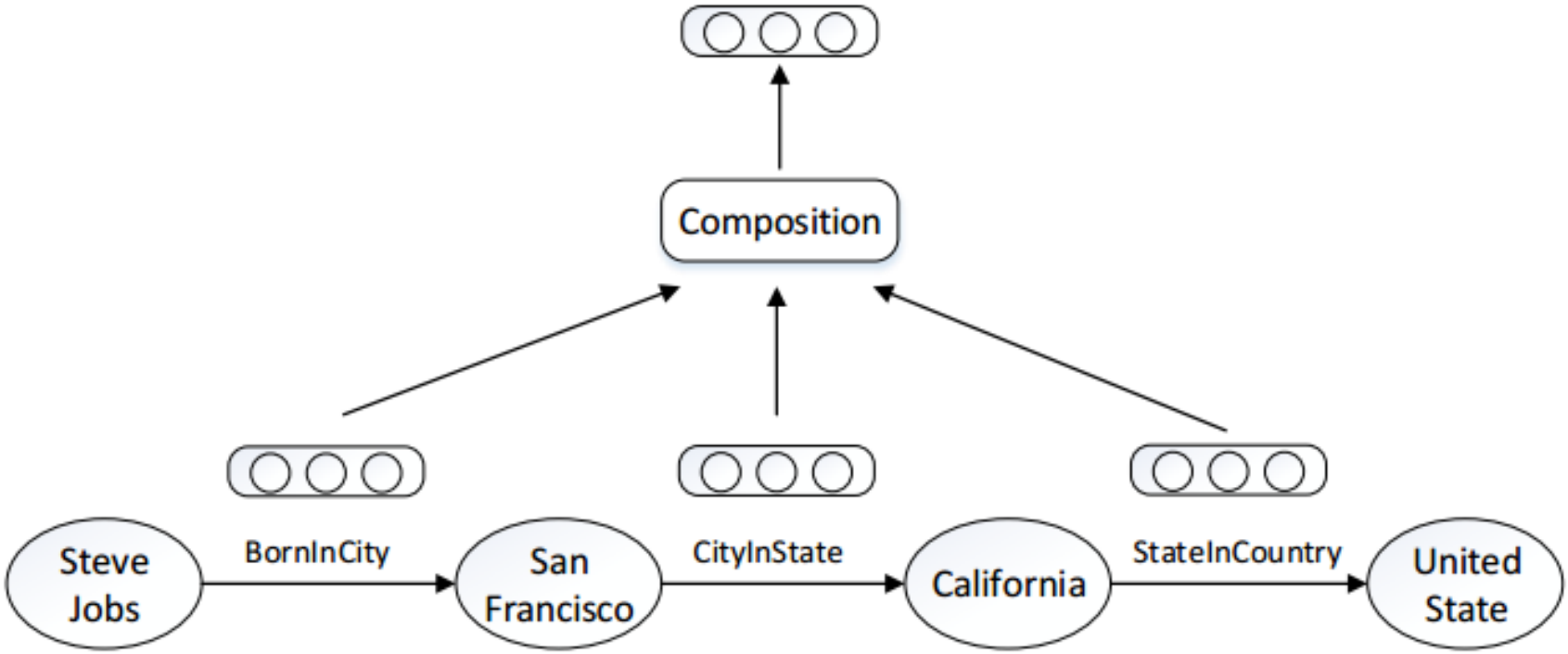
考虑关系路径的TransE: PTransE



Lin, et al. (2015). Modeling Relation Paths for Representation Learning of Knowledge Bases. EMNLP.

考虑关系路径的TransE: PTransE

- 关系路径的表示：组合语义
- ADD，MULTIPLY，RNN



Gardner, et al. (2013). Improving learning and inference in a large knowledge-base using latent syntactic cues. EMNLP.

评测结果：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, et al. (2015). Modeling Relation Paths for Representation Learning of Knowledge Bases. EMNLP.

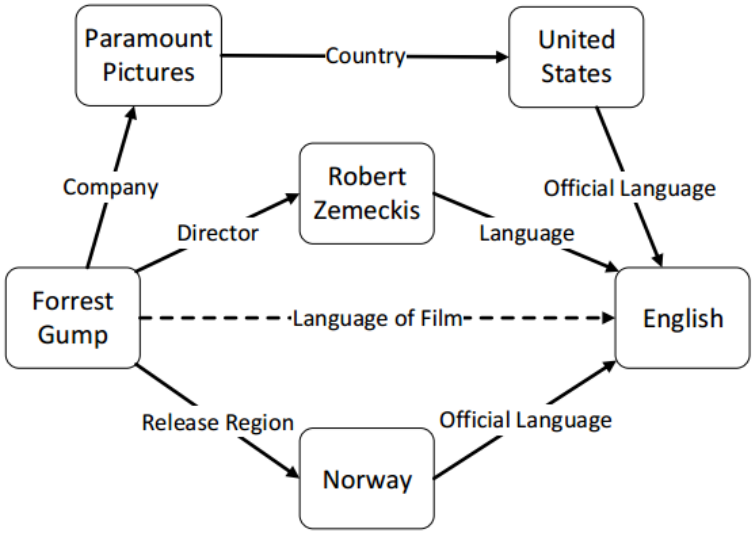
评测结果：关系预测

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%

关系路径表示的挑战问题

- 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)

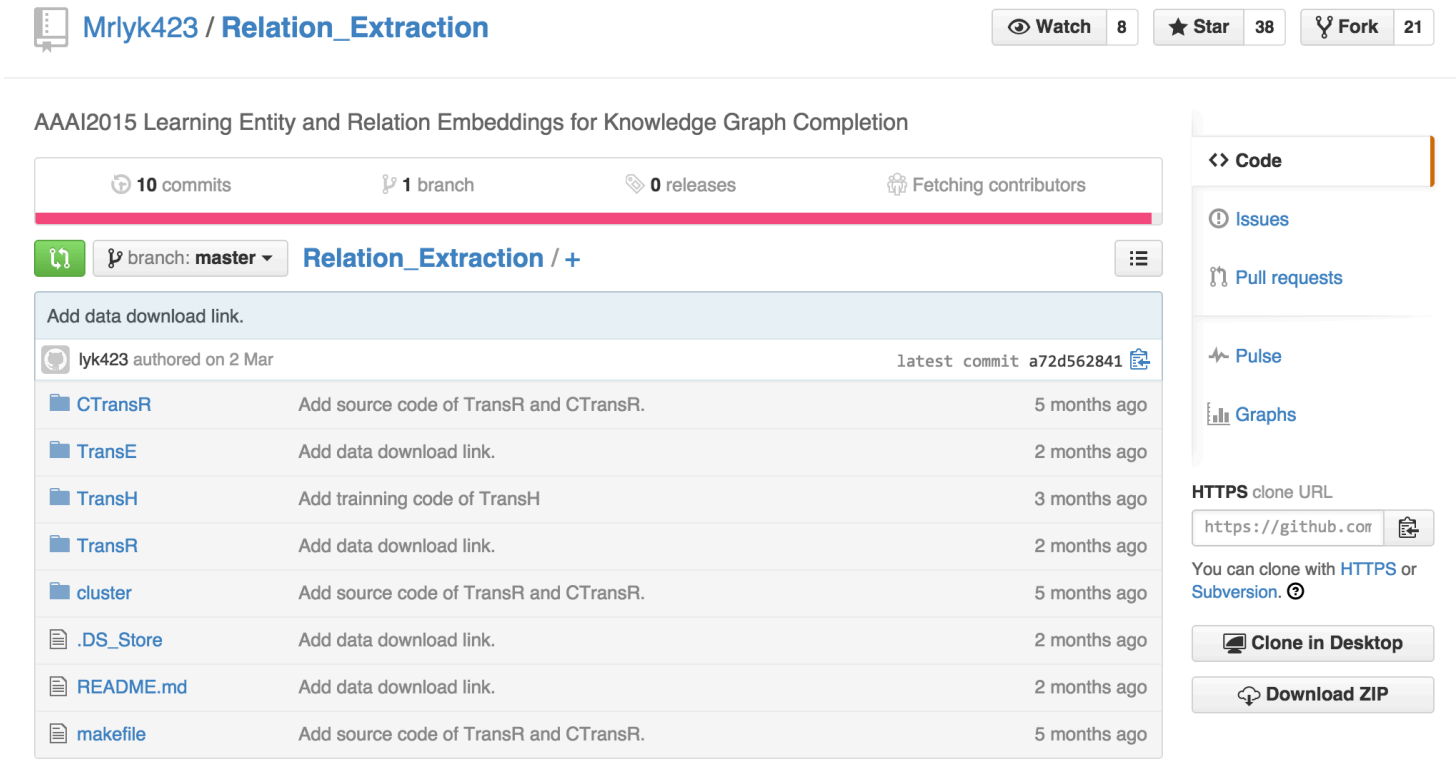
(奥巴马，是，美国人)

  - 推理关系可信性
- 如何更好地表示关系之间的复杂推理关系
  - 组合语义模型：RNN、NTN、…
- 应用：QA (Guu, et al. 2015)

Guu, et al. Traversing Knowledge Graphs in Vector Space. EMNLP 2015.

开放源码

- TransE、TransH、TransR、PTransE
  - https://github.com/mrlyk423/relation_extraction
- 将持续发布更多知识表示学习模型



知识表示学习的其他重要挑战

- 大规模知识图谱表示的快速学习
  - 长尾数据上的在线学习、分布式学习
- 融合知识图谱丰富信息的知识表示学习
  - 利用实体和关系的属性、描述、层次类型等信息
  - 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

广阔天地，大有可为

谢谢大家，敬请指正

http://nlp.csai.tsinghua.edu.cn/~lzy

liuzy@tsinghua.edu.cn