中国中文信息学会前沿技术讲习班
CIPS-ATT
第四期：深度学习与自然语言处理

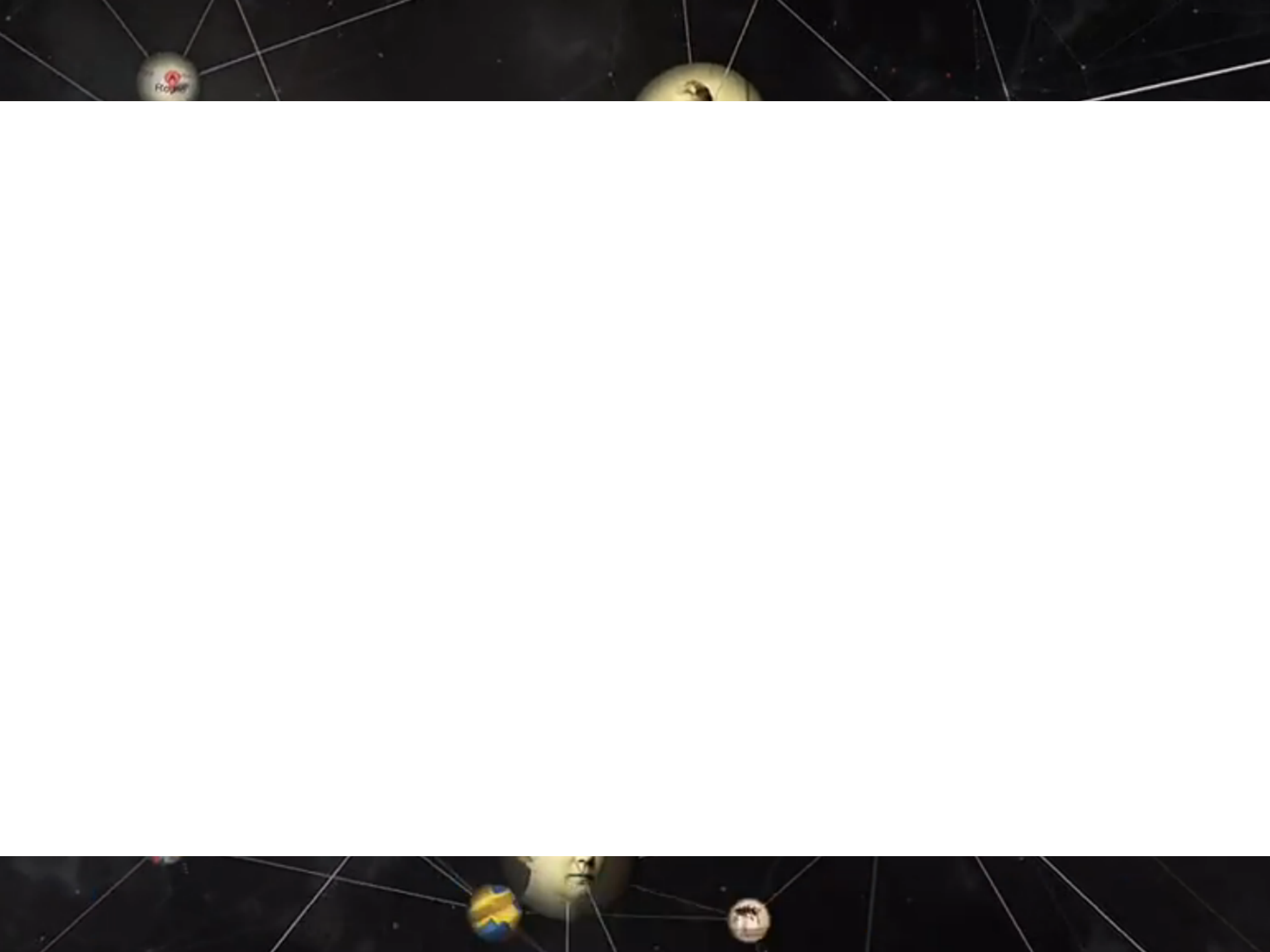
深度学习与知识图谱

刘知远
清华大学
liuzy@tsinghua.edu.cn

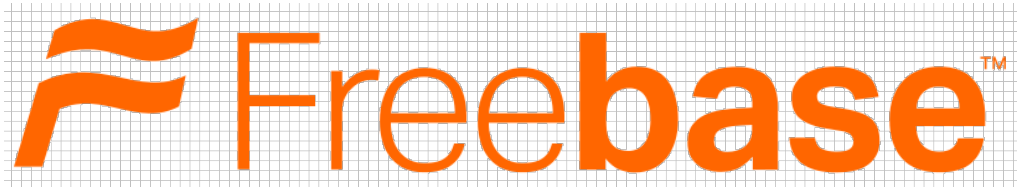
韩先培
中科院软件所
xianpei@nfs.iscas.ac.cn

什么是知识图谱？


NOBEL PRIZE



典型知识图谱


Freebase

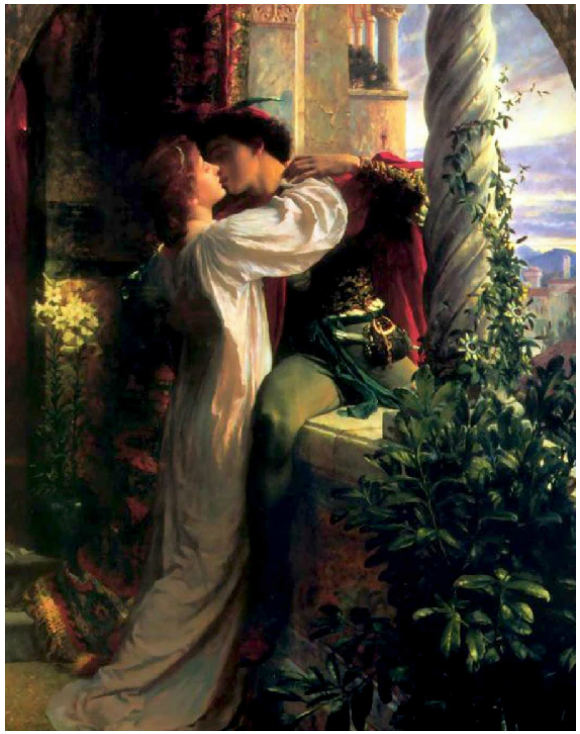

DBpedia


yago
select knowledge


WordNet
A lexical database for English

结构化知识

write

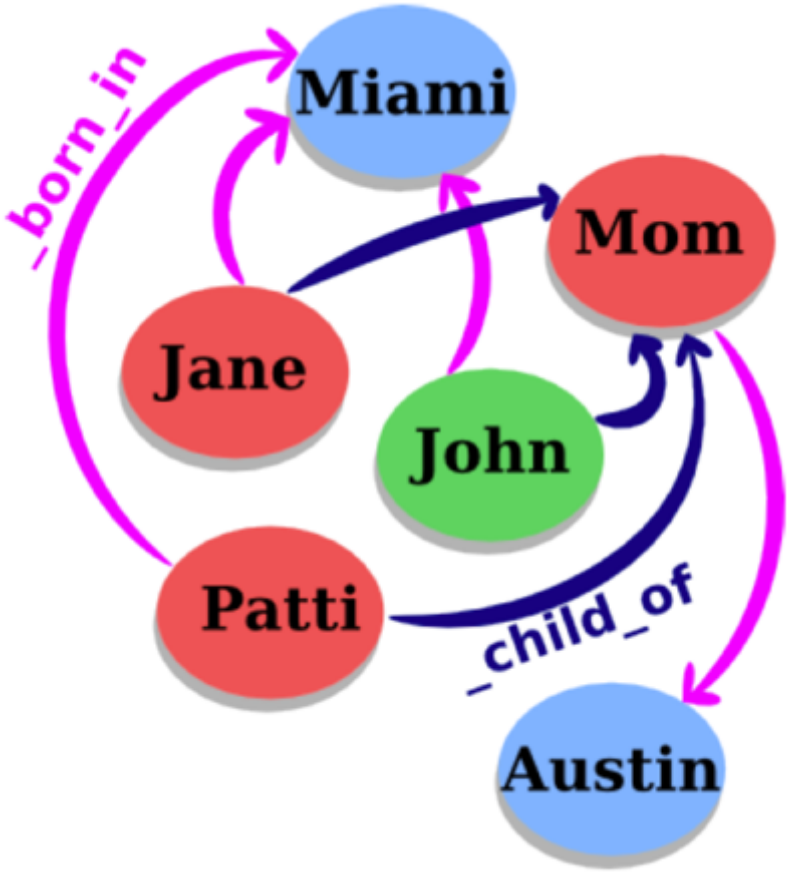


(***William Shakespeare***, book/author/works_written, ***Romeo and Juliet***)

head entity **relation** **tail entity**

知识图谱实体与关系

- 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):
- 代表知识图谱
  - WordNet: 语言知识
  - Freebase: 世界知识



代表知识图谱

- 语言知识图谱
  - WordNet：155, 327个单词，同义词集117,597个，同义词集之间由22种关系连接
- 事实性知识图谱
  - OpenCyc： 23.9万个实体，1.5万个关系属性，209.3万个事实三元组
  - Freebase：4000多万实体，上万个属性关系，24多亿个事实三元组
  - DBpedia：400多万实体，48,293种属性关系，10亿个事实三元组
  - YAGO2 ：980万实体，超过100个属性关系， 1亿多个事实三元组
  - 百度百科：词条数1000万个
  - 互动百科：800万词条，5万个分类，68亿文字

代表知识图谱

- 领域知识图谱
  - Kinships：描述人物之间的亲属关系，104个实体，26种关系，10,800个三元组
  - UMLS：医学领域，描述医学概念之间的联系，135个实体，49种关系，6,800个三元组。
  - Cora：2,497个实体，7种关系，39,255个三元组
- 机器自动构建的知识图谱
  - NELL： 519万实体，306种关系， 5亿候选三元组
  - Knowledge Vault: 4500万实体，4469种关系，2.7亿三元组

分水岭

Wikipedia

DBpedia

yago
select knowledge

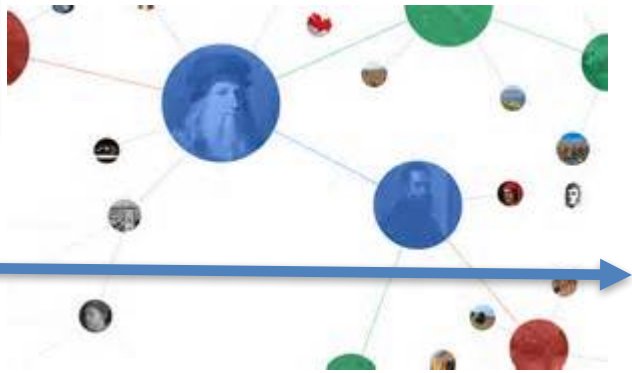
Cyc

WordNet

知网

高质量数据源
500万概念
多语言
富含丰富语义结构的文档：
Infobox，table，list，category…

freebase

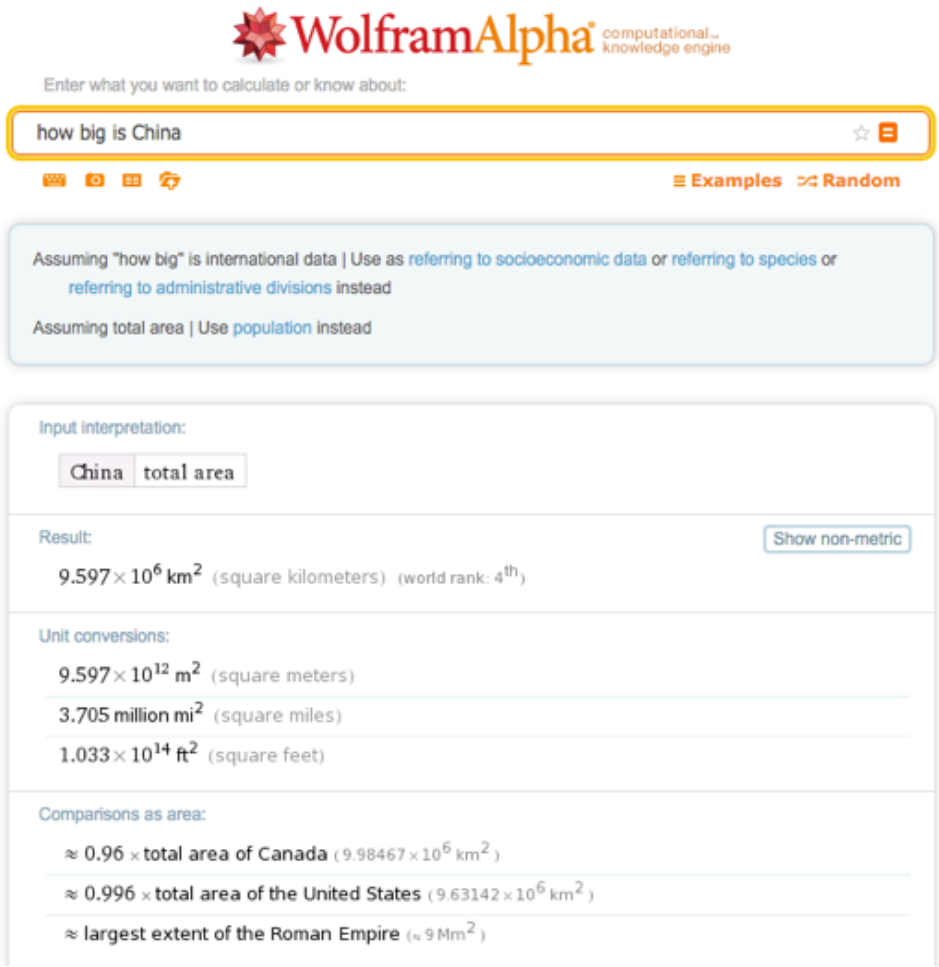


1985

1990

2005-2010

知识图谱应用：问答系统


WolframAlpha computational knowledge engine
Enter what you want to calculate or know about:
how big is China
Examples
Random
Assuming "how big" is international data | Use as referring to socioeconomic data or referring to species or referring to administrative divisions instead
Assuming total area | Use population instead
Input interpretation:
China
total area
Result:
Show non-metric
9.597 × 10^6 km^2 (square kilometers) (world rank: 4th)
Unit conversions:
9.597 × 10^12 m^2 (square meters)
3.705 million mi^2 (square miles)
1.033 × 10^14 ft^2 (square feet)
Comparisons as area:
≈ 0.96 × total area of Canada (9.98467 × 10^6 km^2)
≈ 0.996 × total area of the United States (9.63142 × 10^6 km^2)
≈ largest extent of the Roman Empire (≈ 9 Mm^2)


姚明个子有多少
网页
新闻
贴吧
知道
音乐
图片
视频
地图
文库
更多»
百度为您找到相关结果约3,030,000个
搜索工具
姚明身高：
226cm
姚明，1980年生于上海市徐汇区，祖籍吴江震泽。中国篮球运动员。1998年4月，他入选王非执教的国家队，开始篮球生涯。2002年，他以状元秀身份被NBA的休斯敦火箭队选中。20... 详情>>
来自百度百科 | 报错

知识图谱应用：搜索引擎

奥巴马 百度一下

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约86,500,000个 搜索工具

奥巴马_百度百科

姓名：贝拉克·侯赛因·奥巴马
生日：1961年8月4日 职业：政治家、律师、总统
简介：贝拉克·侯赛因·奥巴马（Barack Hussein Obama），1961年8月4日出生，美国民主党籍政治家，第44任美国总统，为美...
人物经历 执政表现 主要作品 家庭生活 人物评价 更多>>
查看“奥巴马”全部4个含义>>
baike.baidu.com/

奥巴马的最新相关信息

奥巴马卸任后当NBA球队老板? 白宫发言人:在讨论

ESPN消息,美国总统奥巴马即将卸任,而他未来要做的事情似乎已经确定好了,那就是当一支NBA球队的老板。 奥巴马卸任后当NBA球队老板? 白宫发言人:在讨...
网易体育 49分钟前

奥巴马欲成NBA球队老板 总统助奇才抢杜兰... 腾讯体育 1小时前
奥巴马的算盘: 英国留欧利于美国外交 搜狐财经 2小时前
奥巴马称朝鲜仍是特别威胁 决定对朝制裁... 网易军事 18小时前
作为最亲密的盟友,奥巴马对英国“脱欧”... 新浪财经 18小时前

奥巴马_百度图片

im2.baidu.com - 查看全部767,024张图片

历年时代周刊年度风云人物 展开


普京
俄罗斯铁腕总统


克林顿
曾任美国总统


小布什
美国第43任总统


比尔·盖茨
微软公司创始人前首富


肯尼迪
美国第35任总统


罗斯福
美国蝉联4届的总统


斯大林
苏联共产党中央总书记


里根
演员出身的美国总统

美国民主党员 展开


希拉里
美国第67任国务卿


拜登
美国现任副总统


卡罗琳·肯尼迪
被美国人称全国的宝贝


克里
美国第68任国务卿

知识图谱应用：自动推理

梁启超的儿子的老婆的情人的父亲 搜狗搜索

网页 论坛 知识 新闻 博客 百科 更多 什么是分类搜索 找到约 173,657 条结果

梁启超的儿子的老婆的情人的父亲：

徐申如

推理说明： 梁启超的儿子是梁思成。梁思成的妻子是林徽因。林徽因的情人是徐志摩。徐志摩的父亲是徐申如

梁启超的儿子的老婆的情人的老婆-读书-DoNews.COM
2004年6月15日... 作者 帖子主题： 知道是谁不？ 作者 帖子主题： RE：梁启超的儿子的老婆的情人的老婆 【（shengfang）回复（cool）的大作】 陆小慢 作者 帖子主题： RE:...
donewsIT门户 - home.donews.com/.../477746.html - 2004-6-15 - 快照 - 预览

梁启超的儿子的老婆的情人的父亲 最佳答案 搜狗知识搜索
梁启超的儿子的老婆的情人的老婆是谁 - 已解决 搜搜问问 2007-11-25
答：梁启超的儿子呢是中国的著名建筑师梁思成 梁思成的老婆呢叫林薇茵看过《人间四月天》的人应该知道啊 那么林薇茵的情人呢就是大名鼎鼎的徐志摩啦 那么徐志摩的老婆...
梁启超的儿子的老婆的情人的老婆是谁？？？ - 已解决 搜搜问问 2011-3-4
梁启超的儿子的太太的情人的太太分别是谁 - 已解决 搜搜问问 2007-12-9
梁启超的儿子的妻子的情人的老婆是谁？ 百度知道 2006-10-4

梁启超

梁启超(1873.2.23—1929.1.19)，生于广东新会。1894年，梁启超提倡变法，并于上海主撰《时务报》，著《变法通议》，刊布报端，启发国人之革新思想。与谭嗣同..
相关阅读

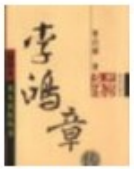
出生： 1873-02-23 / 广东新会
逝世： 1929-01-19
妻子： 李蕙仙 (正室) / 王桂荃 (老婆)
人物关系： 梁宝瑛 (父亲) / 梁思达 (儿子) / 梁思忠 (儿子) / 梁思懿 (女儿) / 梁思成 (儿子)
个人名言： 享受工作的同时享受生活

著作 更多>>


中国近三百年学..


中国历史研究法


新大陆游记



李鸿章传


清代学术概论

知识表示学习

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.

语言表示代表方案

- 1-hot representation: basis of Bag-of-Word Model

star [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ⋯]

sun [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ⋯]

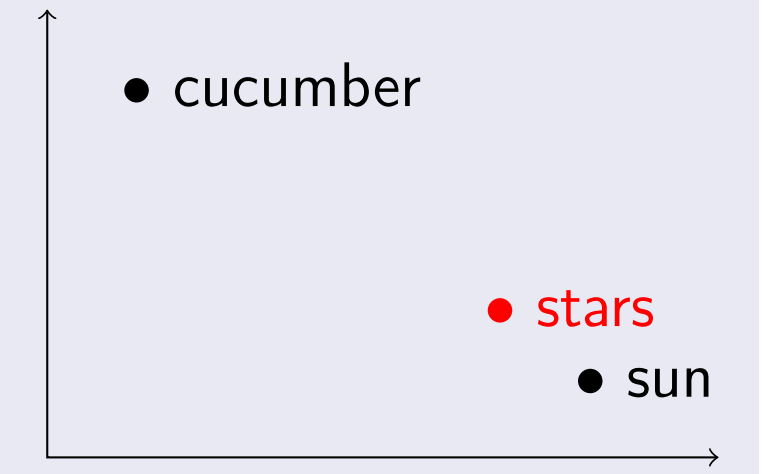
sim(star, sun) = 0

语言表示代表方案

- Count-based distributional representation

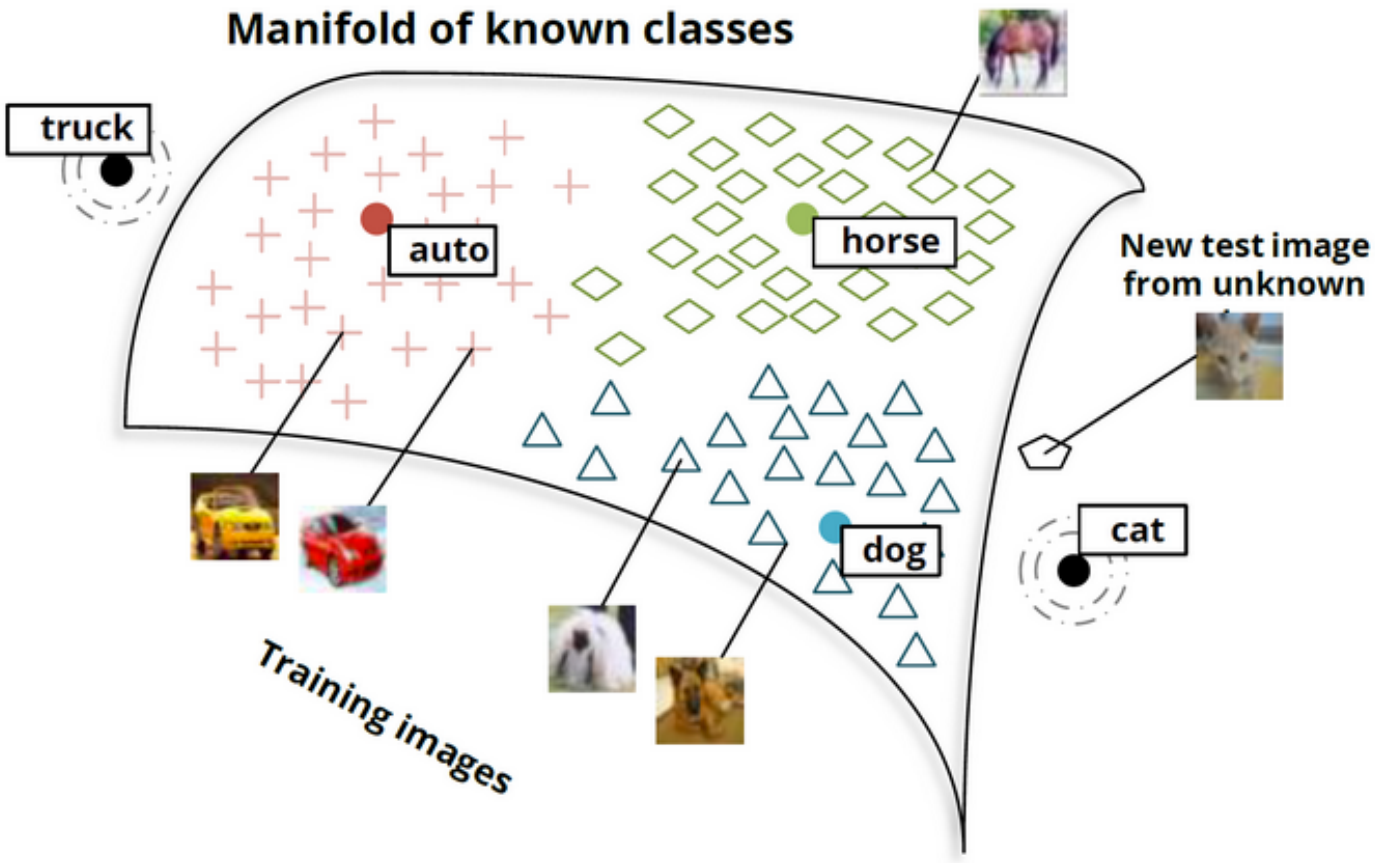
he curtains open and the stars shining in on the barely
ars and the cold , close stars " . And neither of the w
rough the night with the stars shining so brightly , it
made in the light of the stars . It all boils down , wr
surely under the bright stars , thrilled by ice-white
sun , the seasons of the stars ? Home , alone , Jay pla
m is dazzling snow , the stars have risen full and cold

	shining	bright	trees	dark	look
stars	38	45	2	27	12



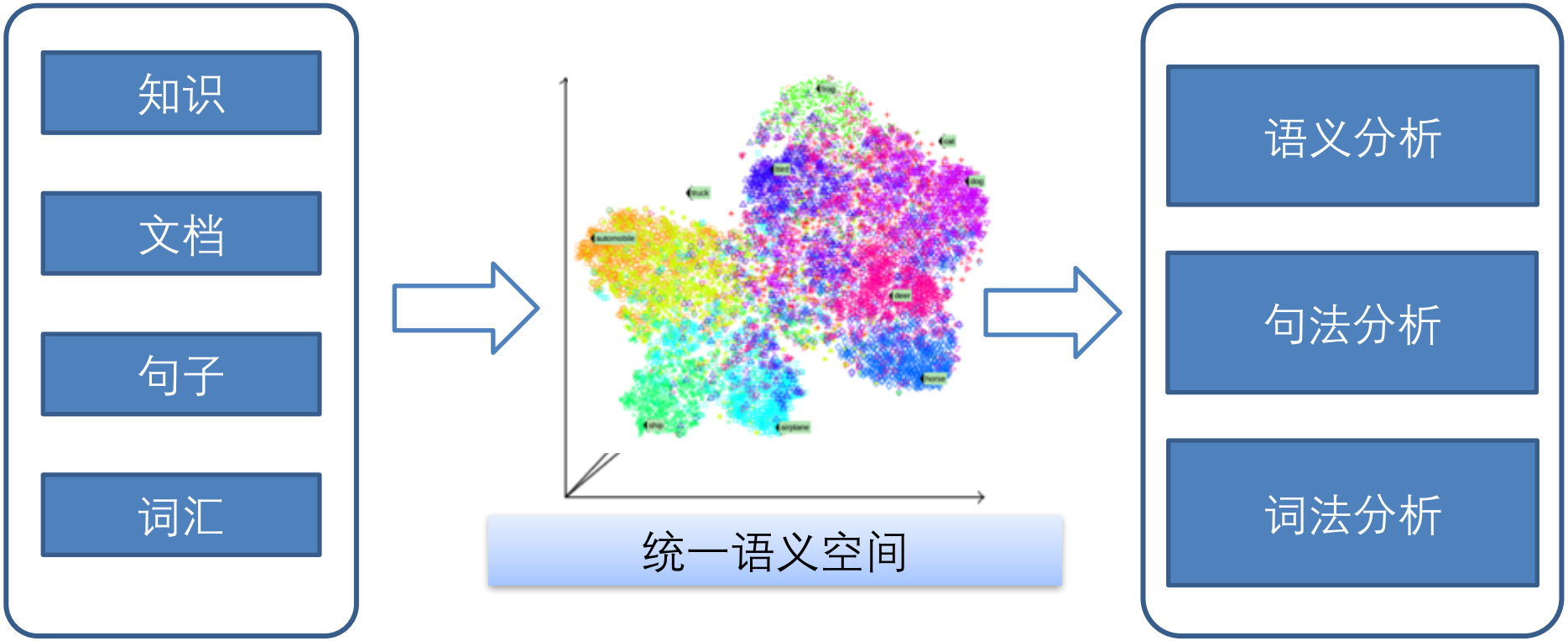
分布式表示学习

- Distributed Representation (Word Embeddings)
- 每个词被表示成稠密、实值、低维向量



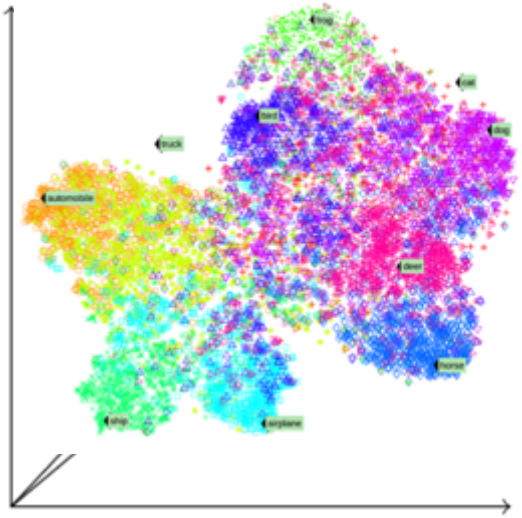
分布式表示对自然语言处理的意义

- 解决大数据NLP的数据稀疏问题
- 实现跨领域、跨对象的知识迁移
- 提供多任务学习的统一底层表示



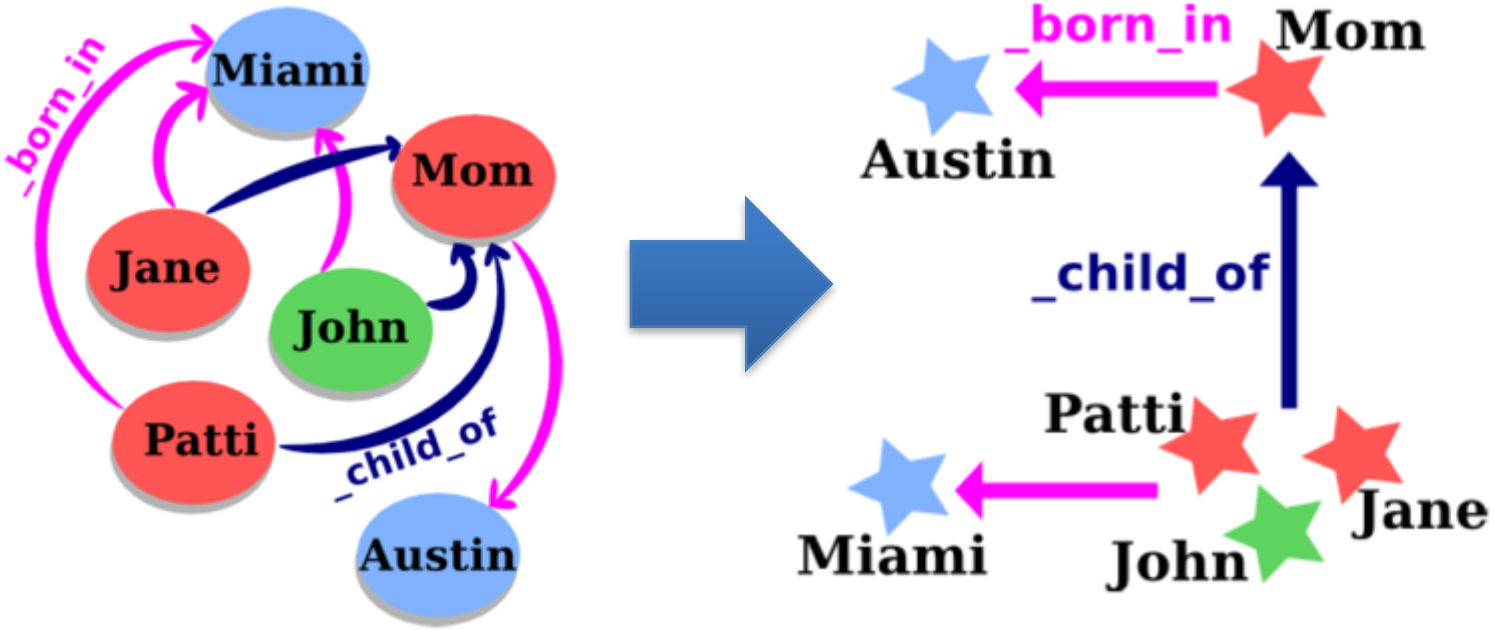
知识表示的挑战

- 知识图谱的典型表示方案
  - 基于符号表示的三元组（RDF）
  - 无法有效计算实体间的语义关系
- 解决方案：将知识映射到低维向量空间



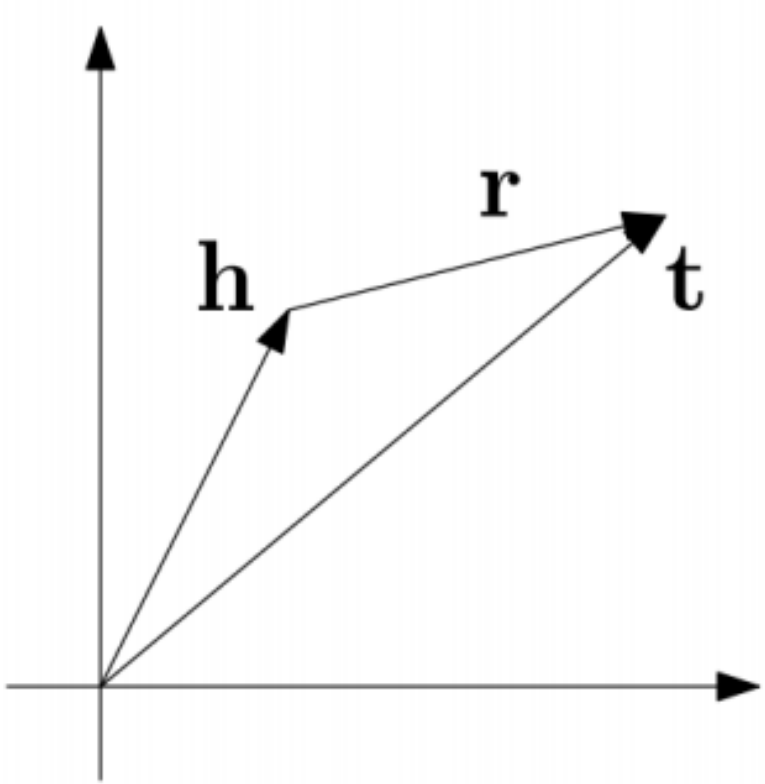
TransE: 将关系表示为翻译

- 对每个事实 (head, relation, tail)，将其中的relation作为从head到tail的翻译操作



TransE: 将关系表示为翻译

- 对每个事实 (head, relation, tail)， 将relation作为从head到tail的翻译操作



优化目标: **h + r = t**

翻译模型的学习

- 势能函数
  - 对于真实事实的三元组(h, r, t)，要求**h**+**r**=**t**
  - 对于错误的三元组则不满足该条件
  - 定义势能函数

$$f\left(h,r,t\right)=\left\|\mathbf{h}+\mathbf{r}-\mathbf{t}\right\|_2$$

f(姚明,出生于,北京) > f(姚明,出生于,上海)

翻译模型的学习

势能函数 $f(h,r,t)=\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_2$

目标函数 $\sum_{(h,r,t)\in\Delta}\sum_{(h',r,t')\in\Delta'}[\gamma+f(h,r,t)-f(h',r,t')]_+$

其中 $[x]_+=\max(0,x)$

约束条件 $\|\mathbf{h}\|\le 1,\|\mathbf{r}\|\le 1,\|\mathbf{t}\|\le 1$

Δ 表示知识库中三元组的集合

Δ' 表示三元组(h, r, t)负例样本集合

评价任务：链接预测

WALL-E _has_genre ?

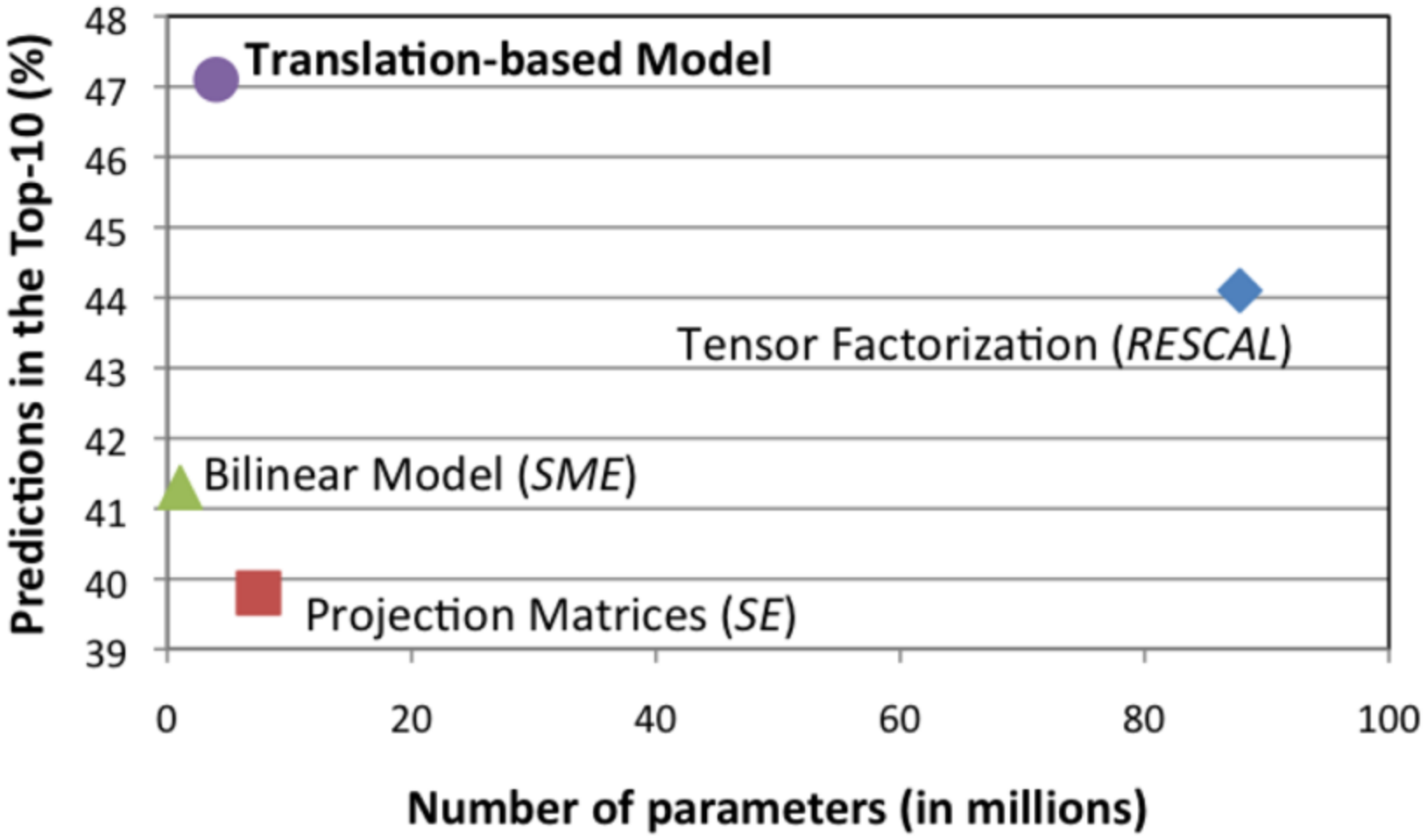
评价任务：链接预测

WALL-E _has_genre

Animation
Computer animation
Comedy film
Adventure film
Science Fiction
Fantasy
Stop motion
Satire
Drama
Connecting

链接预测性能比较

Freebase15K



TransE样例

Entity	**Tsinghua_University**	**A.C._Milan**
1	University_of_Victoria	Inter_Milan
2	St._Stephen's_College,_Delhi	Celtic_F.C.
3	University_of_Ottawa	FC_Barcelona
4	University_of_British_Columbia	Genoa_C.F.C.
5	Peking_University	Udinese_Calcio
6	Utrecht_University	Real_Madrid_C.F.
7	Dalhousie_University	FC_Bayern_Munich
8	Brasenose_College,_Oxford	Bolton_Wanderers_F.C.
9	Cardiff_University	Borussia_Dortmund
10	Memorial_University_of_Newfoundland	Hertha_BSC_Berlin

TransE样例

Entity	China	Barack_Obama	Apple
1	Japan	George_W._Bush	Onion
2	Taiwan	Nancy_Pelosi	Strawberries
3	South_Korea	John_Kerry	Avocado
4	Argentina	Hillary_Rodham_Clinton	Pear
5	North_Korea	Al_Gore	Cabbage
6	Hungary	George_H._W._Bush	Broccoli
7	Israel	John_McCain	Egg
8	Australia	Colin_Powell	Cheese
9	Iceland	Bill_Clinton	Bread
10	Hong_Kong	Charles_B._Rangel	Tomato

TransE样例

Relation	/people/person/nationality	/location/location/contains
1	/people/person/places_lived	/base/aareas/schema/administrative_area/administrative_children
2	/people/person/place_of_birth	/location/country/administrative_divisions
3	/people/person/spouse_s	/location/country/first_level_divisions
4	/base/popstra/celebrity/vacations_in	/location/country/capital
5	/government/politician/government_positions_held	/award/award_nominee/award_nominations
6	/people/deceased_person/place_of_death	/location/administrative_division/capital
7	/olympics/olympic_athlete/country	/location/us_county/county_seat
8	/olympics/olympic_athlete/medals_won	/base/aareas/schema/administrative_area/capital
9	/music/artist/origin	/location/us_county/hud_county_place
10	/people/person/employment_history	/award/award_winner/awards_won

TransE样例

Head	China	Barack_Obama
Relation	/location/location/adjoin	/education/education/institution
1	Japan	Harvard_College
2	Taiwan	Massachusetts_Institute_of_Technology
3	Israel	American_University
4	South_Korea	University_of_Michigan
5	Argentina	Columbia_University
6	France	Princeton_University
7	Philippines	Emory_University
8	Hungary	Vanderbilt_University
9	North_Korea	University_of_Notre_Dame
10	Hong_Kong	Texas_A&M_University

TransE样例

Head	Stanford_University	Apple	Titanic
Relation	/education/educational_institution/students_graduates	/food/food/nutrients	/film/film/genre
1	Steven_Spielberg	Lipid	War_film
2	Ron_Howard	Protein	Period_piece
3	Stan_Lee	Valine	Drama
4	Barack_Obama	Tyrosine	History
5	Milton_Friedman	Serine	Biography
6	Walter_F._Parkes	Iron	Film_adaptation
7	Michael_Cimino	Cystine	Adventure_Film
8	Gale_Anne_Hurd	Pantothenic_acid	Action_Film
9	Bryan_Singer	Vitamin_A	Political_drama
10	Aaron_Sorkin	Sugar	Costume_drama

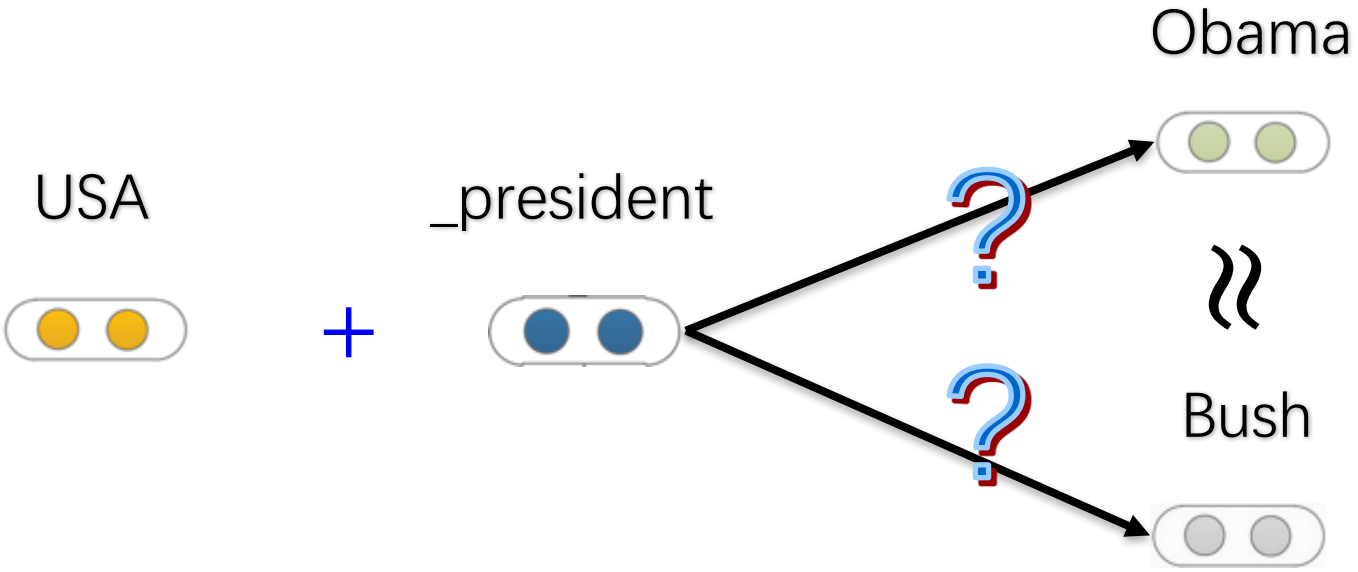
知识表示学习的主要挑战

- 复杂关系建模
- 考虑外部信息
- 关系路径建模

复杂关系建模

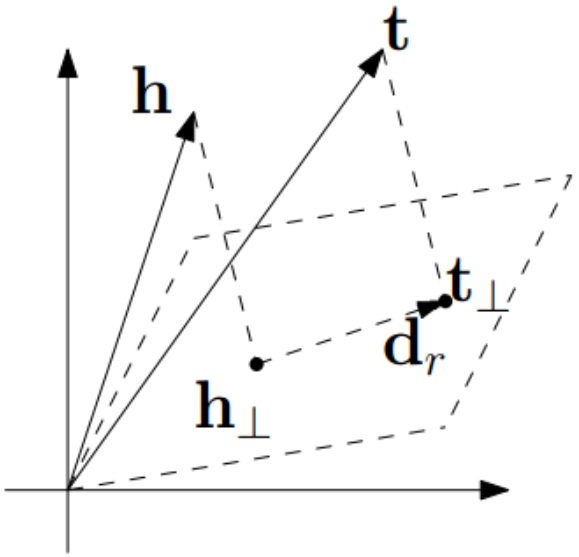
复杂关系的建模

- 1-to-N, N-to-1, N-to-N关系
  - (USA, _president, Obama)
  - (USA, _president, Bush)

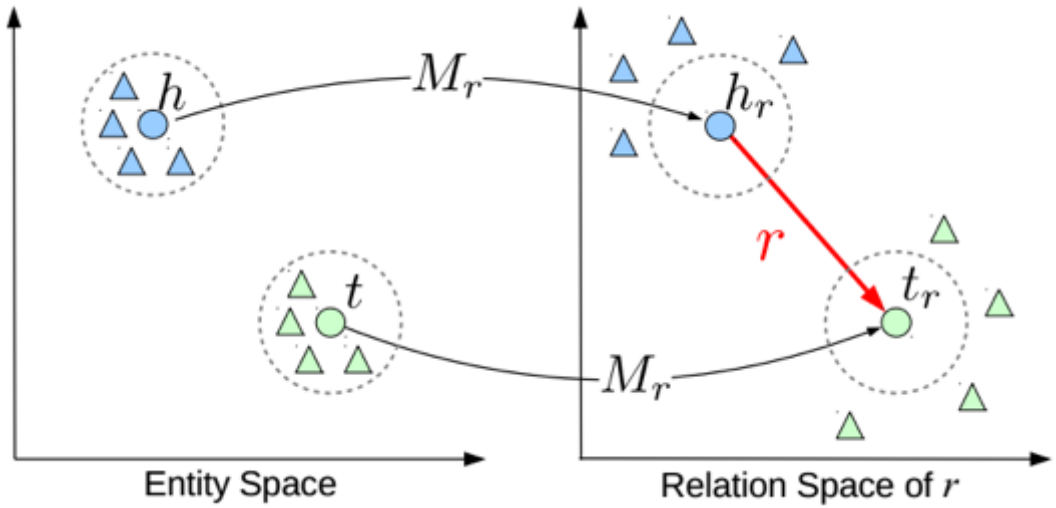


复杂关系的建模

- 建立与特定关系有关的实体表示



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.

链接预测结果

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**

Examples

Head Entity	**Titanic**		
Relation	**/film/film/genre**		
Model	TransE	TransH	TransR
1	War_film	Drama	Costume_drama
2	Period_piece	Romance_Film	Drama
3	Drama	Costume_drama	Romance_Film
4	History	Film_adaptation	Period_piece
5	Biography	Period_piece	Epic_film
6	Film_adaptation	Adventure_Film	Adventure_Film
7	Adventure_Film	LGBT	LGBT
8	Action_Film	Existentialism	Film_adaptation
9	Political_drama	Epic_film	Existentialism
10	Costume_drama	War_film	War_film

Examples

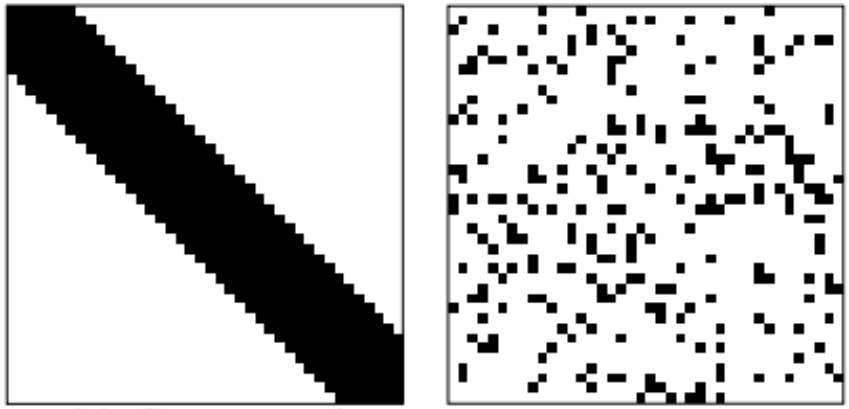
Head	**University_of_Cambridge**		
Relation	**/education/education/student**		
Model	TransE	TransH	TransR
1	John_Cleese	Stephen_Fry	David_Attenborough
2	Samuel_Beckett	David_Attenborough	Stephen_Fry
3	Harold_Pinter	Ralph_Vaughan_Williams	Stephen_Hawking
4	Virginia_Woolf	Alan_Bennett	Ralph_Vaughan_Williams
5	Graham_Chapman	Francis_Bacon	Alan_Bennett
6	Philip_Pullman	Julian_Fellowes	Julian_Fellowes
7	Ian_McEwan	Hugh_Bonneville	Ernest_Rutherford
8	Douglas_Adams	Graham_Chapman	Jonathan_Lynn
9	Terry_Gilliam	Miriam_Margolyes	Tom_Hollander
10	Richard_Dawkins	Stephen_Hawking	Chris_Weitz

TransSparse

- 实体的语义向量与其语义关系密切相关

$$\mathbf{h}_p = \mathbf{M}_r^h(\theta_r^h)\mathbf{h}, \quad \mathbf{t}_p = \mathbf{M}_r^t(\theta_r^t)\mathbf{t}$$

- 基于动态稀疏矩阵的语义关系建模

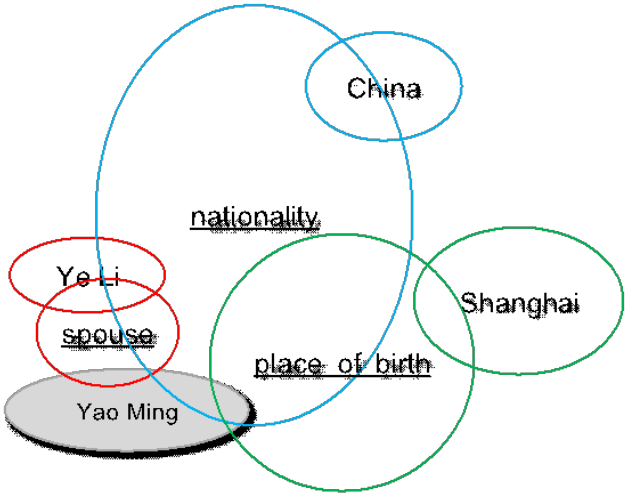


(a) Structured (b) Unstructured

Ji, G., Liu, K., He, S., & Zhao, J. (2016). Knowledge graph completion with adaptive sparse transfer matrix. In *Proceedings of AAAI*.

KG2E

- 多维高斯分布表示符号
- 均值向量表示该符号的位置(含义)
- 协方差矩阵表示该符号的多样性(不确定性)
  - 包含事实越多，该实体语义越明确
  - 关系越复杂，该关系确定性越弱



(b) density-based embedding

He, S., Liu, K., Ji, G., & Zhao, J. (2015). Learning to represent knowledge graphs with gaussian embedding. In *Proceedings of CIKM*.

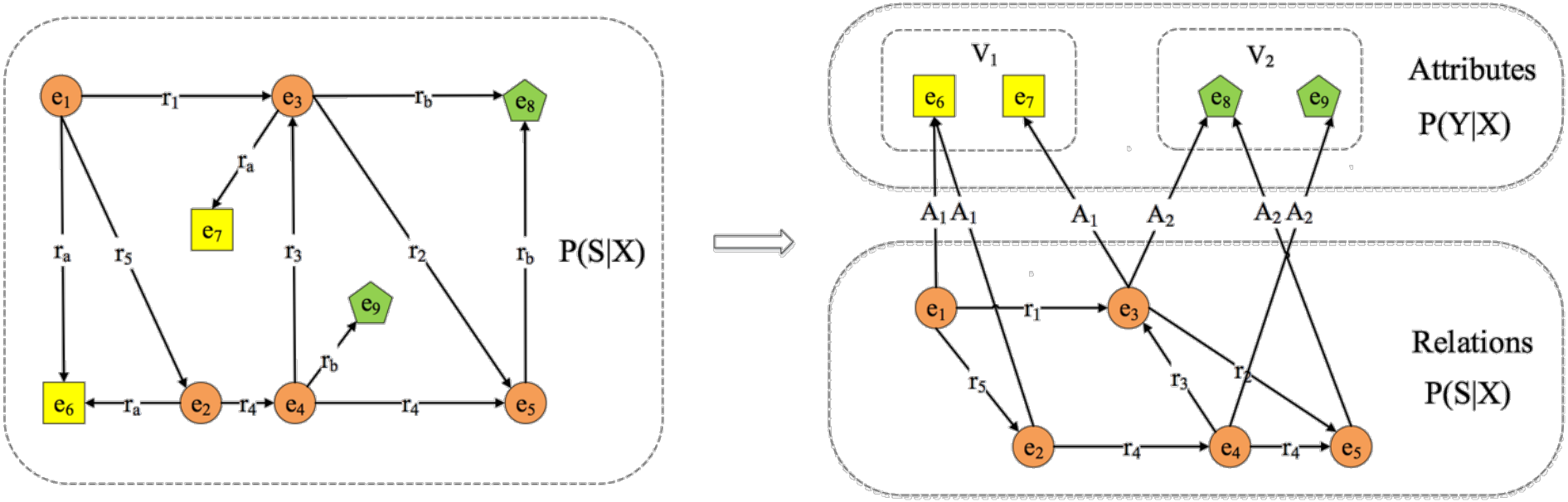
实体的属性与关系

- 知识图谱中大量“关系”实际是实体属性
  - 国籍、性别、宗教信仰等
- 用TransE对实体属性建模并不合理
  - 人物 + 性别 = 男/女 (?)

类型	名称	E_t	E_h
属性	国籍	1.05	1,551.90
	性别	1.00	637,333.33
	种族	1.12	41.52
	宗教信仰	1.09	107.40
关系	父母	1.58	1.67
	首都	1.29	1.42
	作者	1.02	2.17
	创始人	1.37	1.31

实体、属性与关系的表示学习

- 实体与关系均用向量表示，仍用TransE学习
- 实体与属性改用线性预测模型学习



评测结果：实体预测

Entity	Head				Tail				Total			
Metric	Mean Rank		Hits@10 (%)		Mean Rank		Hits@10 (%)		Mean Rank		Hits@10 (%)	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
TransE	385	277	20.2	39.2	134	124	51.4	66.7	259	200	35.8	53.0
TransH	416	309	17.7	35.4	147	138	50.0	65.0	282	224	33.9	50.2
TransR	394	285	20.5	41.2	125	116	53.4	71.0	260	200	37.0	56.1
KR-EAR(TransE)	295	198	22.7	39.6	77	69	54.2	69.5	186	133	38.5	54.5
KR-EAR(TransR)	**268**	**170**	**23.4**	**43.0**	**75**	**66**	**55.7**	**71.5**	**172**	**118**	**39.5**	**57.3**

评测结果：关系预测

- 新模型能够更好的实现关系预测
- 先验知识：实体属性也可以用来预测关系(CRA)

Metric	Mean Rank		Hits@1 (%)	
	Raw	Filter	Raw	Filter
TransE	3.1	2.8	65.9	83.8
TransH	3.4	3.1	64.9	84.1
TrasnR	3.4	3.1	65.2	84.5
KR-EAR(TransE)	2.4	2.1	67.9	86.2
+ CRA	**1.8**	**1.6**	70.9	88.7
KR-EAR(TransR)	2.6	2.2	66.8	89.0
+ CRA	1.9	**1.6**	**71.5**	**90.4**

属性间关联样例

- 先验知识：统计表明实体部分属性间具有高度关联

属性	关联属性
职业	婚姻状况，国籍，性别，语言，种族
电影发布地区	电影国家，电影语言，电影发布日期，电影题材
地点时区	国家位置，地点货币
音乐类型	使用乐器，所属专辑，职业，乐器发声部位
电视剧题材	电视剧国家，电视剧语言，电视剧发行网络

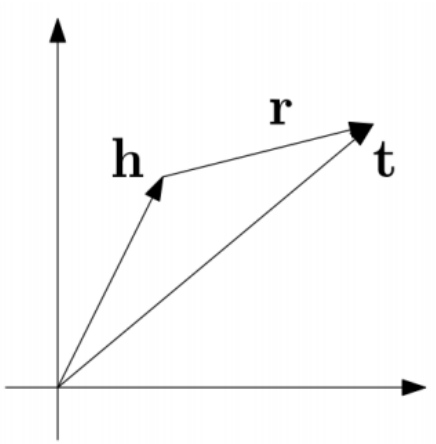
评测结果：属性预测

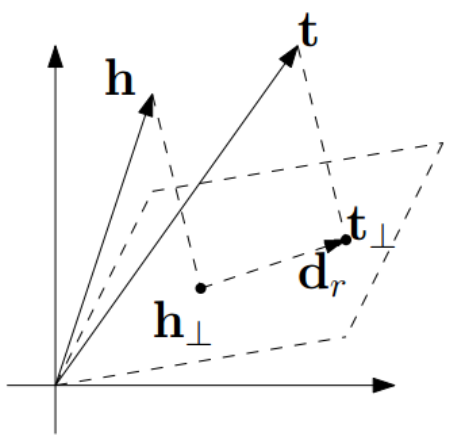
- 新模型能够更好地实现实体的属性预测
- 先验知识：属性关联关系也可用来预测属性

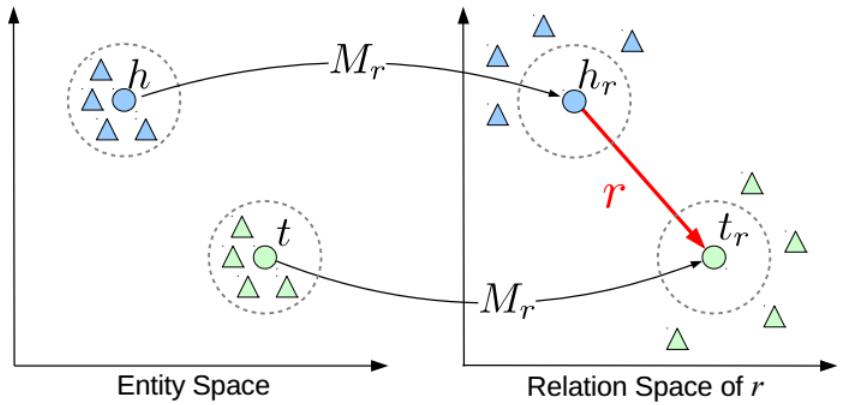
Metric	Mean Rank		Hits@1 (%)	
	Raw	Filter	Raw	Filter
TransE	10.7	5.6	36.5	55.9
TransH	10.7	5.6	38.5	57.9
TrasnR	9.0	3.9	42.7	65.6
KR-EAR(TransE)	8.3	3.2	47.2	69.0
+AC	**7.5**	**3.0**	49.4	70.4
KR-EAR(TransR)	8.3	3.2	47.6	69.8
+AC	**7.5**	**3.0**	**49.8**	**70.8**

小结

- TransE无法较好处理1-N, N-1, N-N等复杂关系
- 面向该问题已经产生大量工作
  - TransA，TransD，TransE，TransG，TransH，TransR, KG2E, TranSparse, Hole







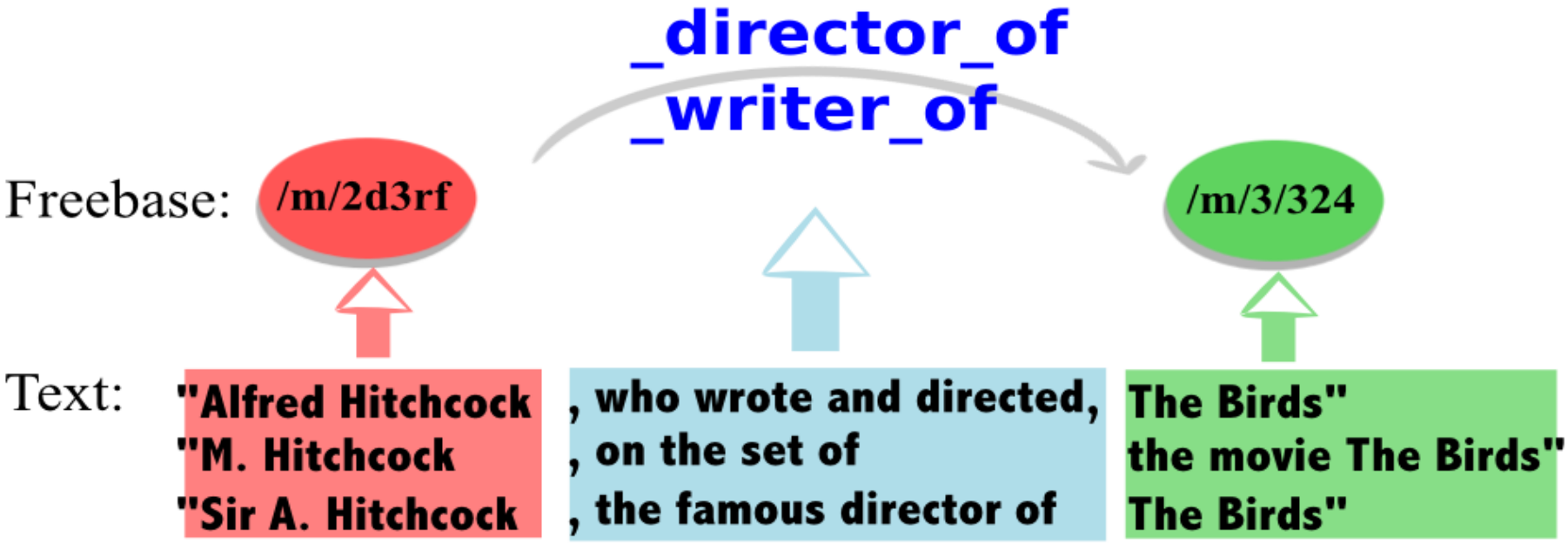
考虑外部信息

文本与知识的融合

- 基于知识图谱的关系预测

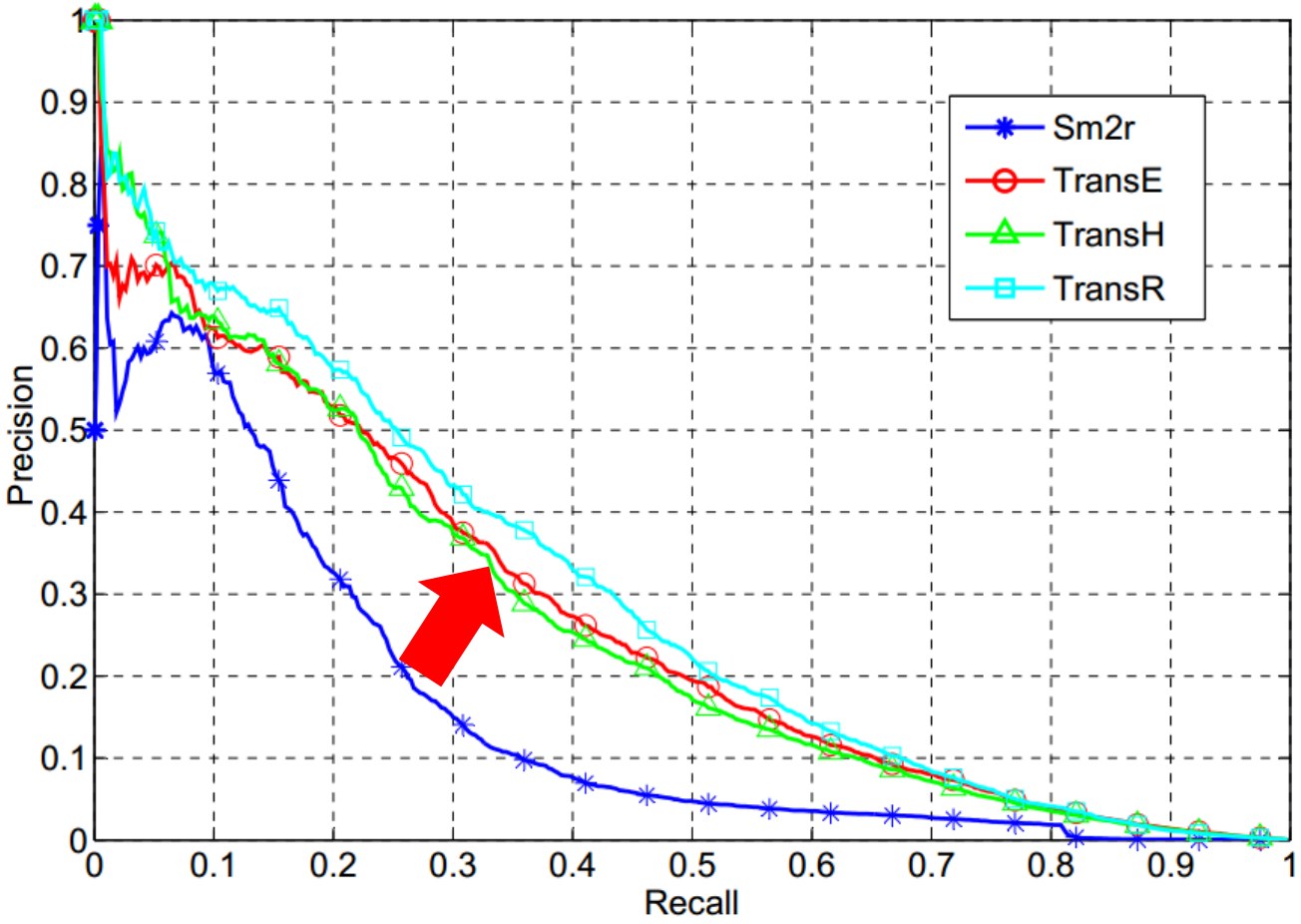
$$r \sim t\text{-}h$$

- 基于文本信息的关系预测



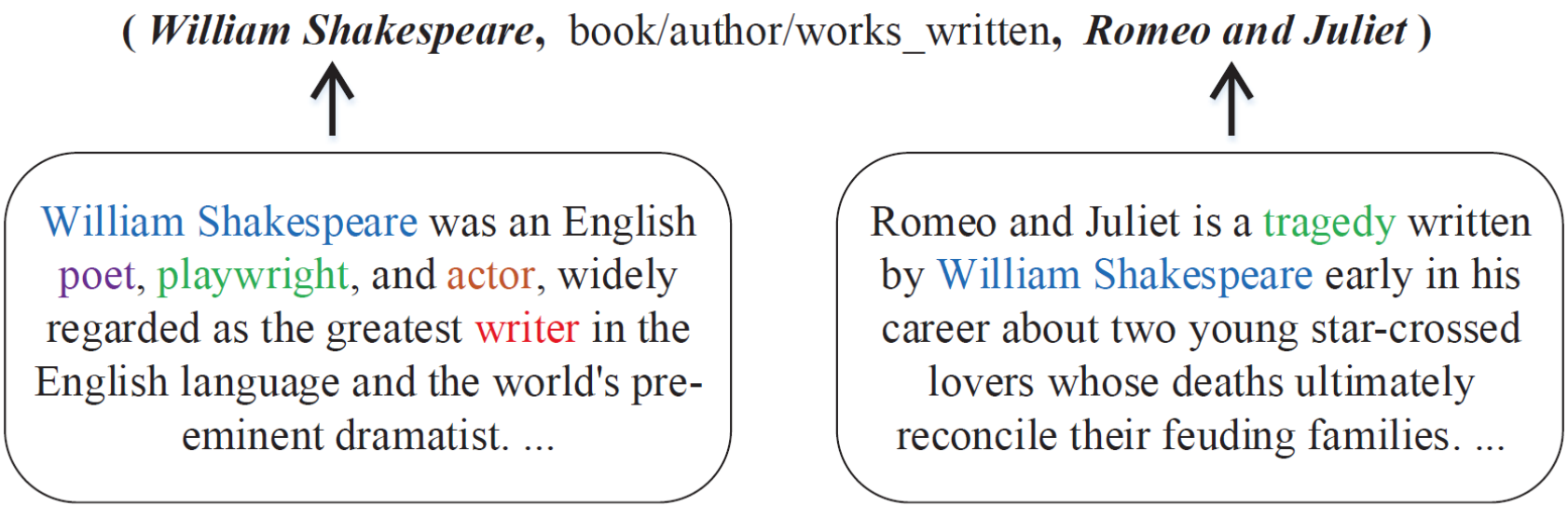
融合文本与知识的关系抽取

- NYT+FB (Weston et al.2013)

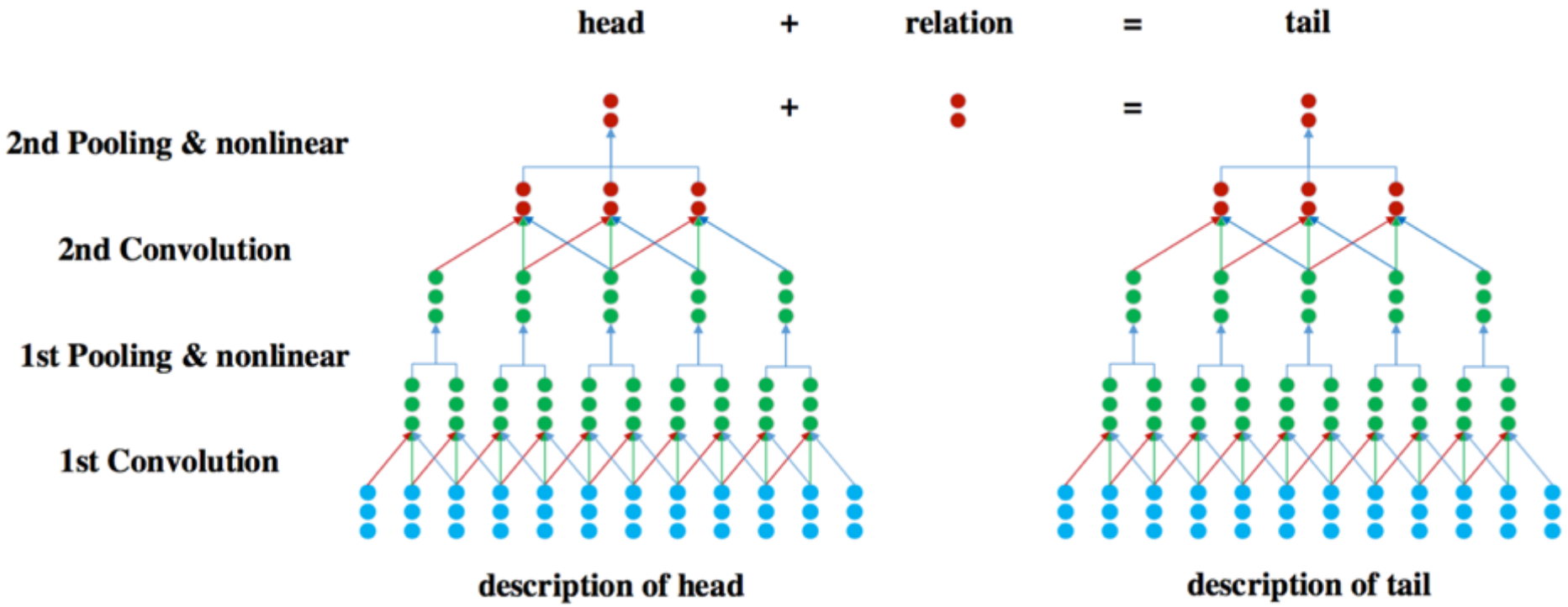


融合实体描述的知识表示

- 利用实体描述信息提供关于实体的语义信息



融合实体描述的知识表示



Xie, et al. (2016). Representation Learning of Knowledge Graphs with Entity Descriptions. AAAI.

Zero-shot场景下的链接预测

- 对于新实体，根据描述信息有效得到实体表示

Metric	$d-e$	$e-d$	$d-d$	Total
Partial-CBOW	26.5	20.9	67.2	24.6
CBOW	27.1	21.7	66.6	25.3
Partial-CNN	26.8	20.8	69.5	24.8
CNN	**31.2**	**26.1**	**72.5**	**29.5**

Entity Prediction

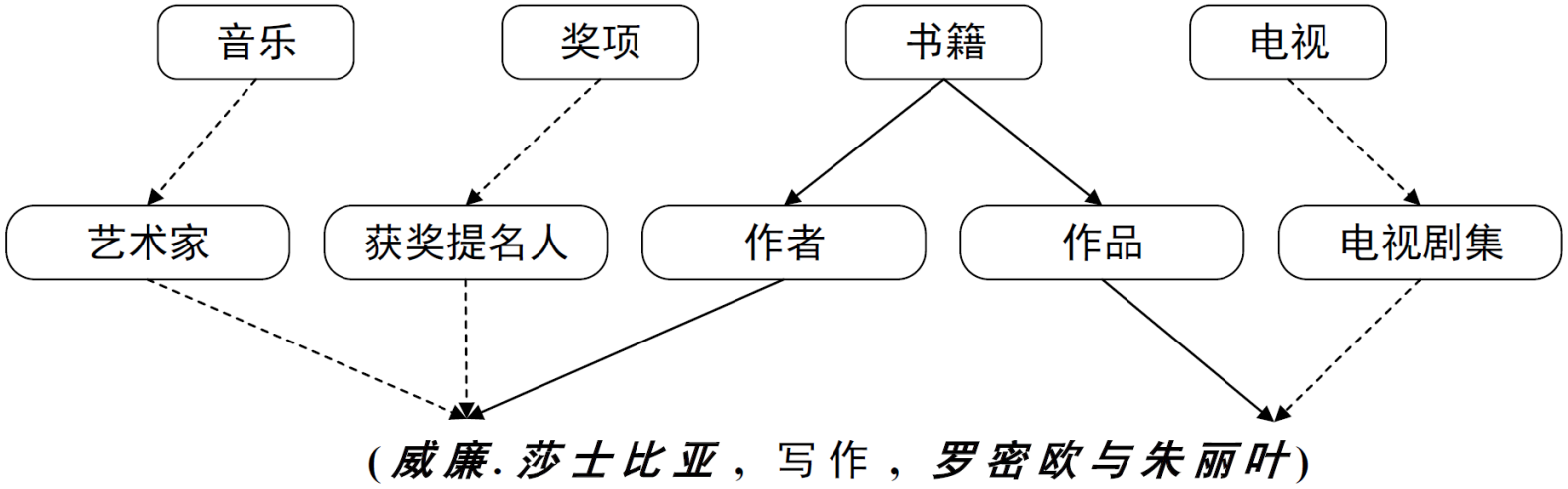
Metric	$d-e$	$e-d$	$d-d$	Total
Partial-CBOW	49.0	42.2	0.0	46.2
CBOW	52.2	47.9	0.0	50.3
Partial-CNN	56.6	52.4	4.0	54.8
CNN	**60.4**	**55.5**	**7.3**	**58.2**

Relation Prediction

* d-e 表示头实体的表示通过CNN从实体描述学习得到

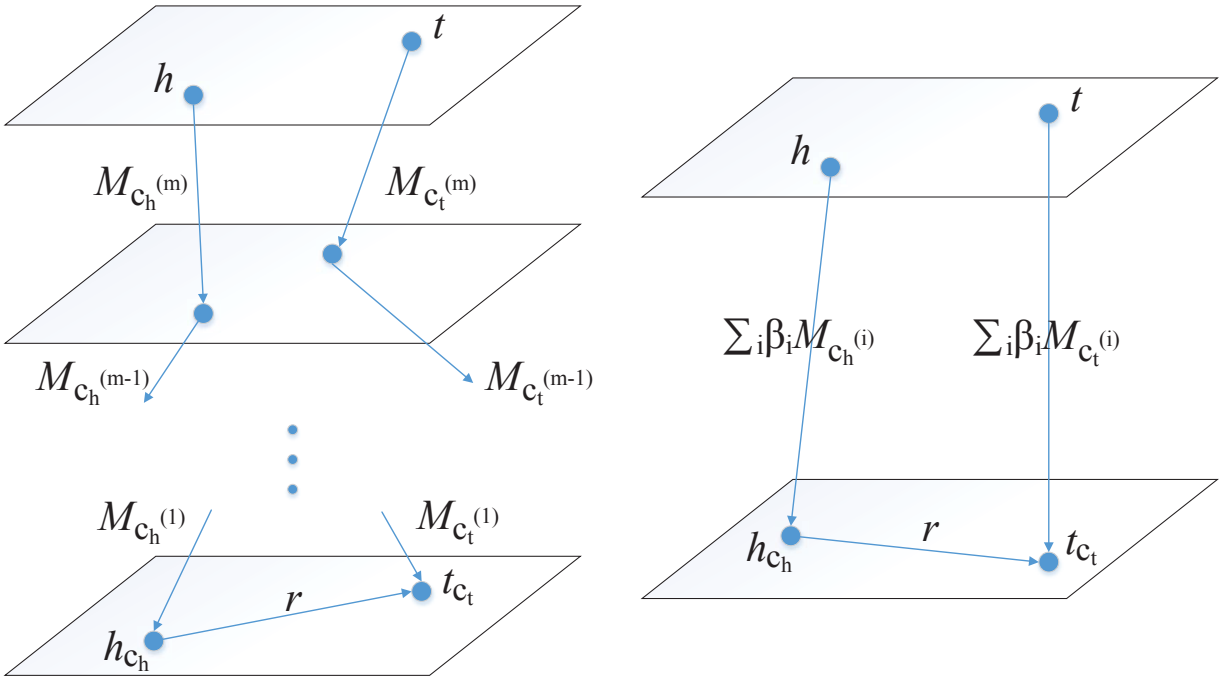
知识图谱包含丰富的外部信息

- 除三元组外，知识图谱包含丰富的外部信息
- 举例：实体的层次类别



融合实体层次类别的表示学习

- 利用矩阵映射技术融合实体层次类别信息
  - Recursive Hierarchy Encoder（RHE）
  - Weighted Hierarchy Encoder（WHE）



(a) RHE　　(b) WHE

评测结果：实体预测

- 层次类别能够显著提升表示学习区分能力

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	238	143	46.4	62.1
TransR	199	77	47.2	67.2
TKRL (RHE)	**184**	**68**	49.2	69.4
TKRL (WHE)	186	**68**	49.2	69.6
TKRL (RHE+STC)	202	89	**50.4**	73.1
TKRL (WHE+STC)	202	87	50.3	**73.4**

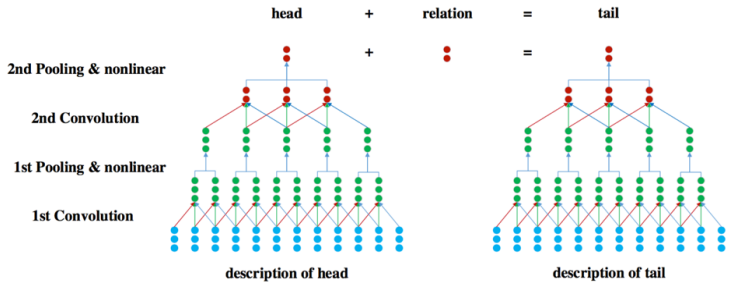
评测结果：长尾链接预测

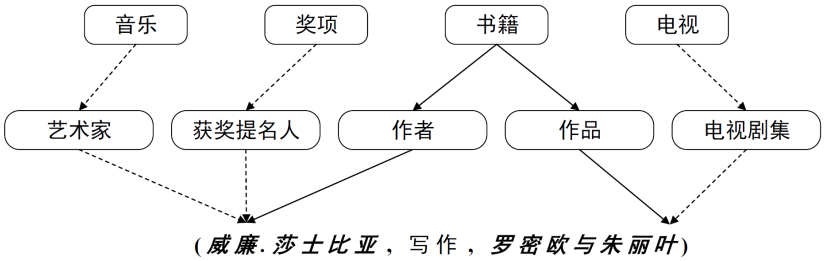
- 层次类别对长尾关系上的实体预测和关系预测效果更加显著
- 在知识表示中引入先验知识能够明显提升稀疏数据上的性能

Relation Frequency	Test Number	Hits@10 for Entity (%)			Hits@1 for Relation (%)		
		TransE	TransR	TKRL (WHE)	TransE	TransR	TKRL (WHE)
$f_r <= 10$	1,444	28.0	32.4 (+4.4)	38.1 (**+10.1**)	13.2	17.0 (+3.8)	21.5 (**+8.3**)
$f_r <= 100$	4,763	49.9	54.5 (+4.6)	57.9 (**+8.0**)	45.7	50.5 (+4.8)	54.3 (**+8.6**)
$f_r <= 1000$	18,296	66.1	69.1 (+3.0)	71.6 (**+5.5**)	70.9	75.4 (+4.5)	77.8 (**+6.9**)
$total$	62,374	61.9	67.2 (+5.3)	69.2 (**+7.3**)	80.4	88.8 (+8.4)	89.7 (**+9.3**)

小结

- 知识图谱包含文本、类别等丰富的外部信息，能够有效辅助知识表示学习



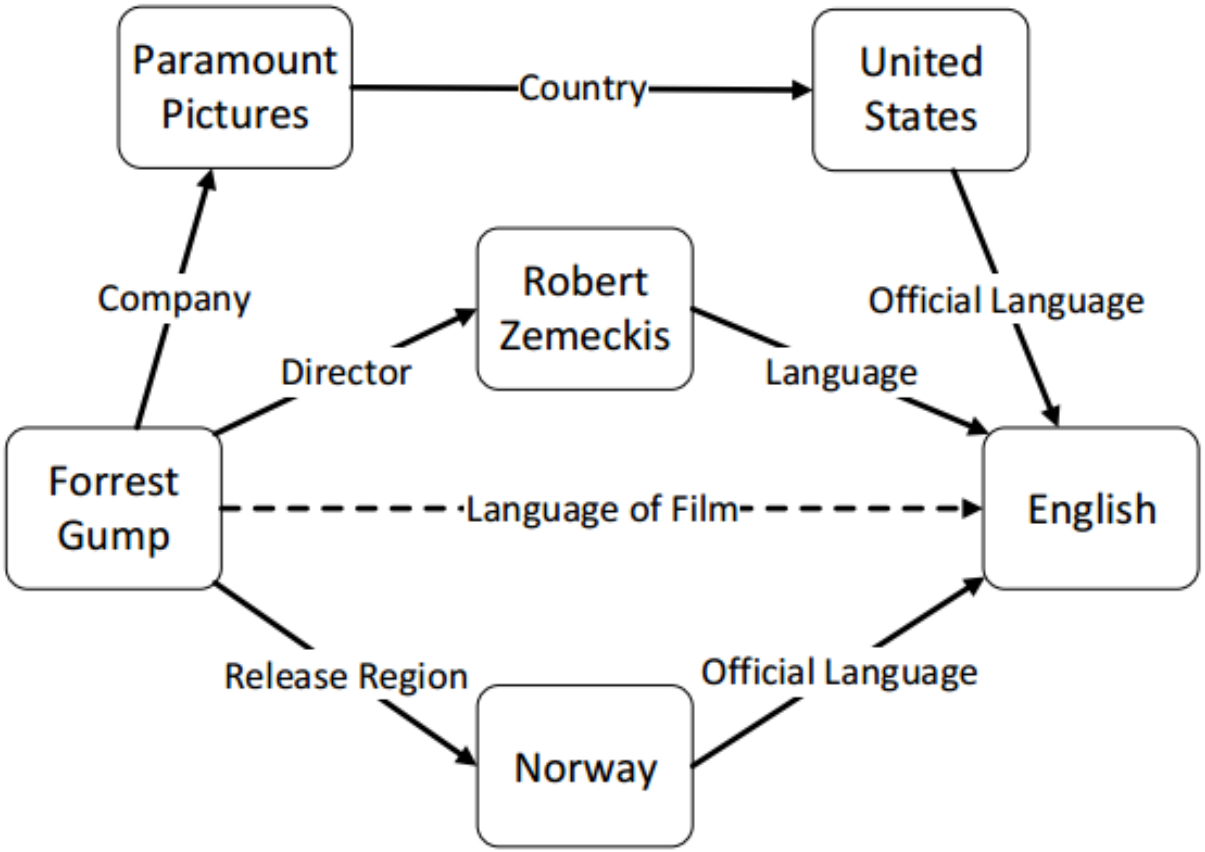


- 未来需要考虑更丰富的信息
  - 跨语言信息
  - 脑电信息
  - 社会网络结构
  - …

关系路径建模

关系路径

- 目前模型孤立地学习每个事实三元组
- 关系之间存在复杂的关系，涉及关系推理



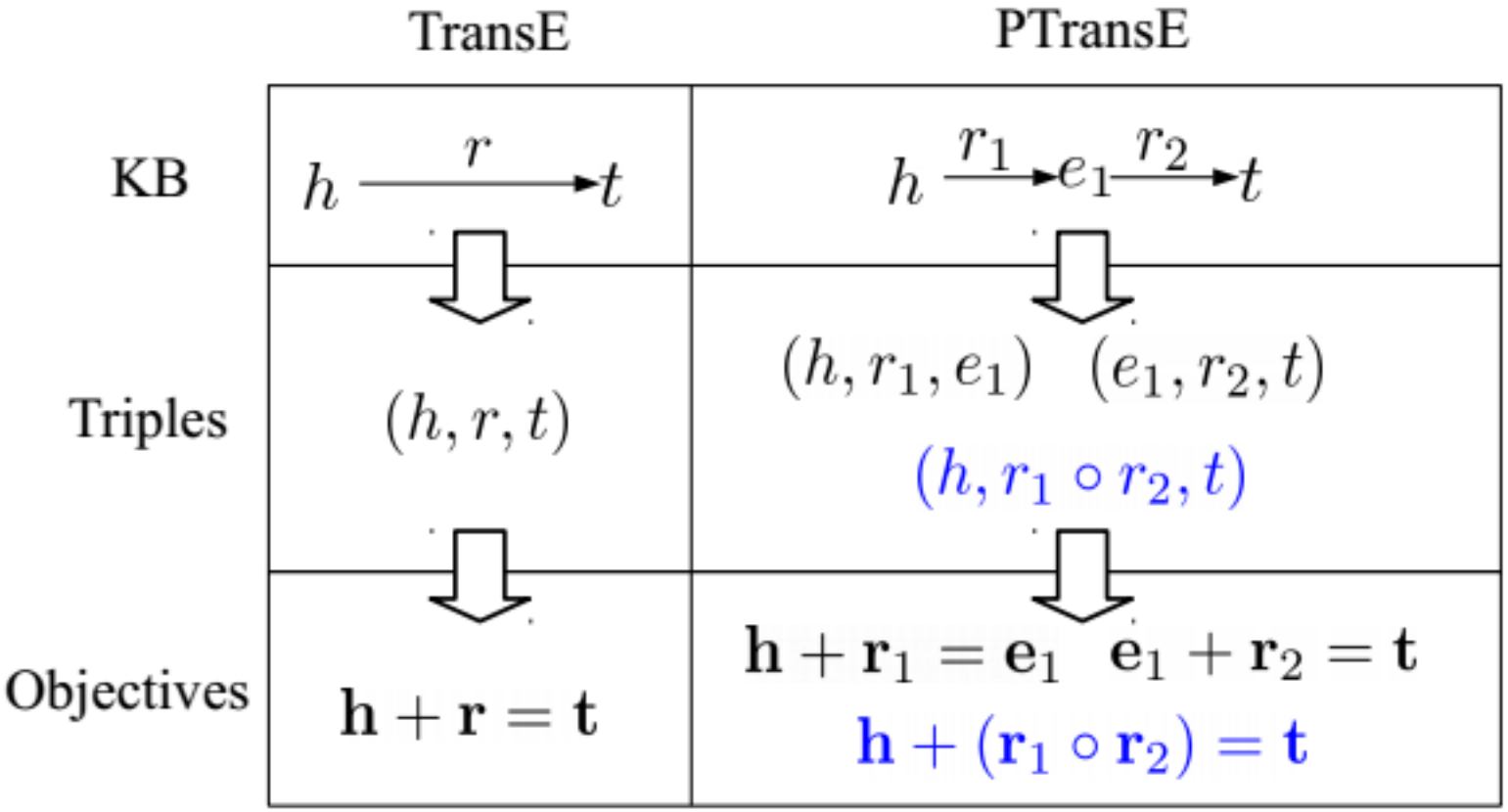
关系路径

- 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.

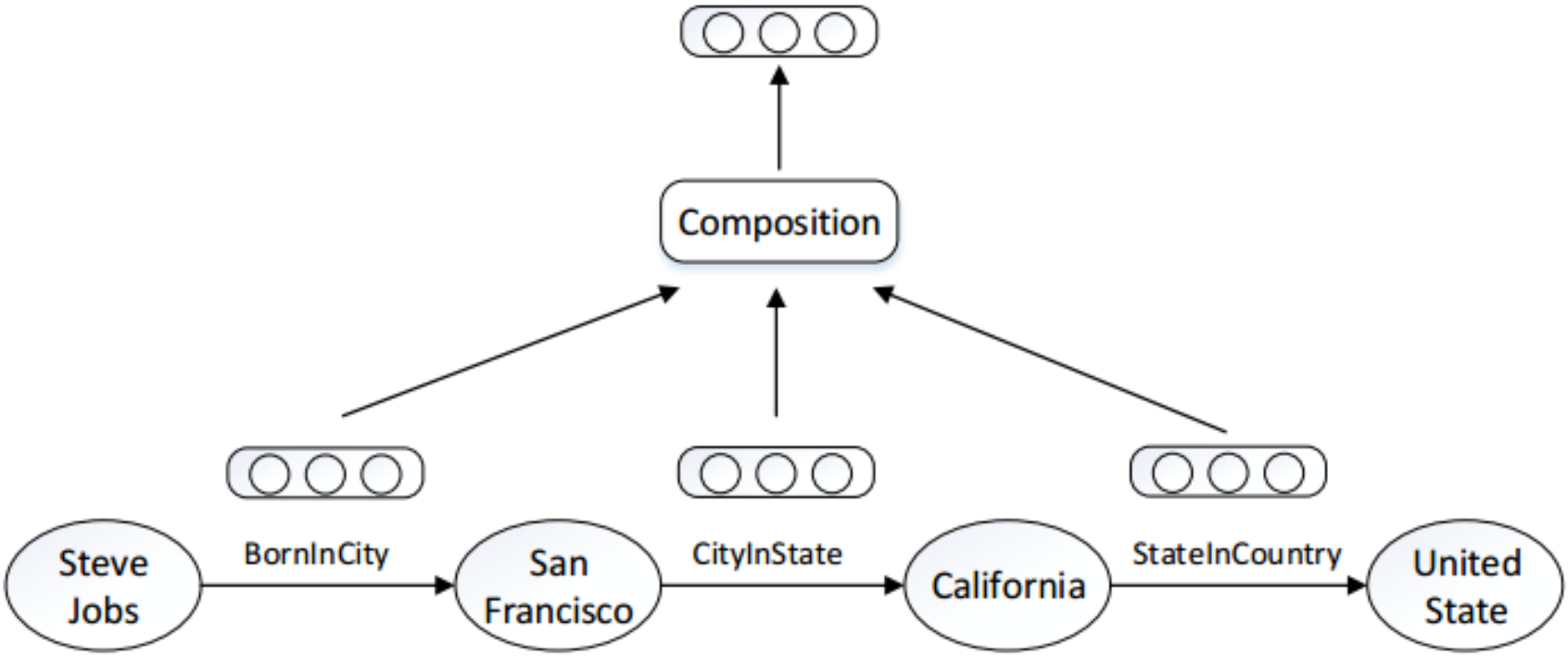
PTransE：考虑关系路径的TransE



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

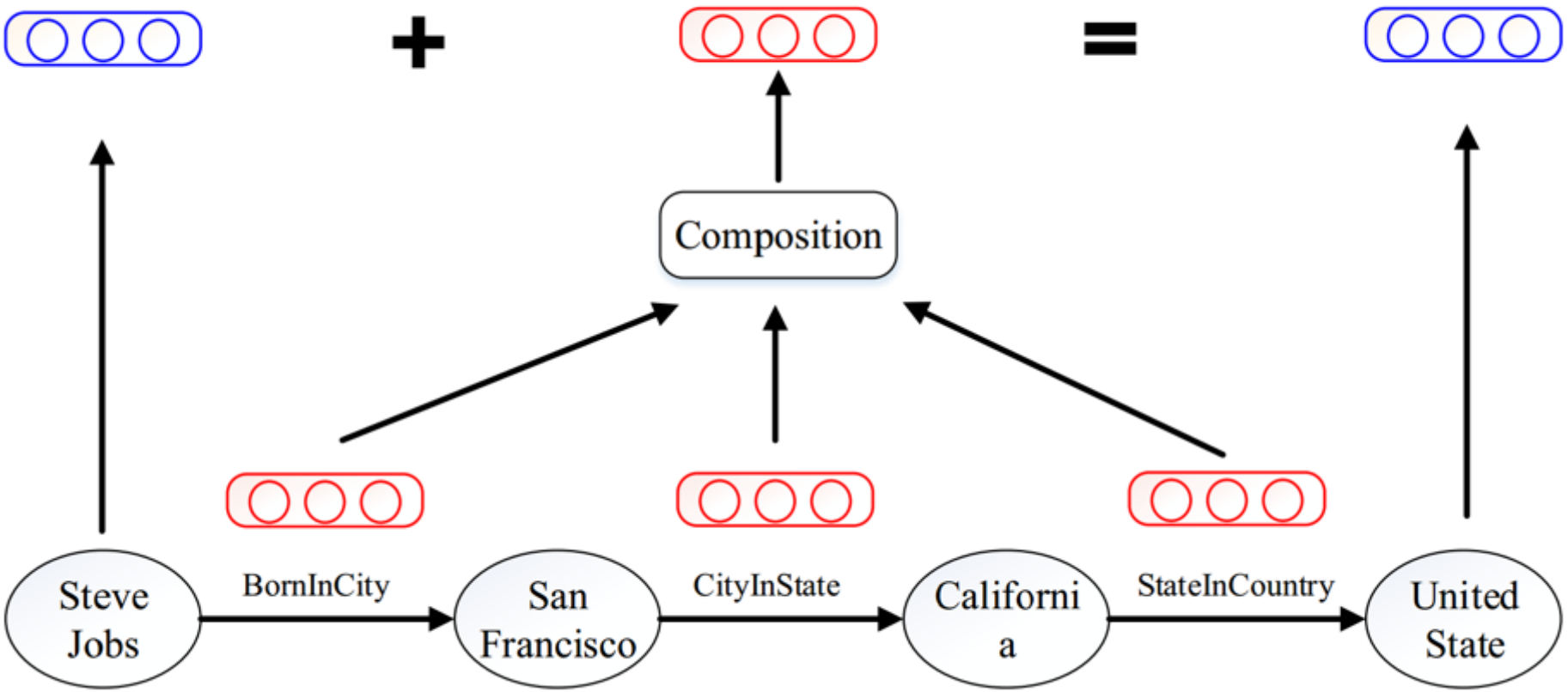
PTransE：考虑关系路径的TransE

- 关键问题：如何得到关系路径的表示
- 解决方案：语义组合（相加，相乘，RNN）



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

Path-based TransE



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

实体预测结果

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.

关系预测结果

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%

PTransE样例

Head	**Barack_Obama**	
Relation	**/education/education/institution**	
Model	TransE	PTransE
1	Harvard_College	Columbia_University
2	Massachusetts_Institute_of_Technology	Occidental_College
3	American_University	Punahou_School
4	University_of_Michigan	University_of_Chicago
5	Columbia_University	Stanford_University
6	Princeton_University	Princeton_University
7	Emory_University	University_of_Pennsylvania
8	Vanderbilt_University	University_of_Virginia
9	University_of_Notre_Dame	University_of_Michigan
10	Texas_A&M_University	Yale_University

PTransE样例

Head	**Stanford_University**	
Relation	**/education/educational_institution/students_graduates**	
Model	TransE	PTransE
1	Steven_Spielberg	Raymond_Burr
2	Ron_Howard	Ted_Danson
3	Stan_Lee	Delmer_Daves
4	Barack_Obama	D.W._Moffett
5	Milton_Friedman	Gale_Anne_Hurd
6	Walter_F._Parkes	Jack_Palance
7	Michael_Cimino	Kal_Penn
8	Gale_Anne_Hurd	Kurtwood_Smith
9	Bryan_Singer	Alexander_Payne
10	Aaron_Sorkin	Richard_D._Zanuck

PTransE样例

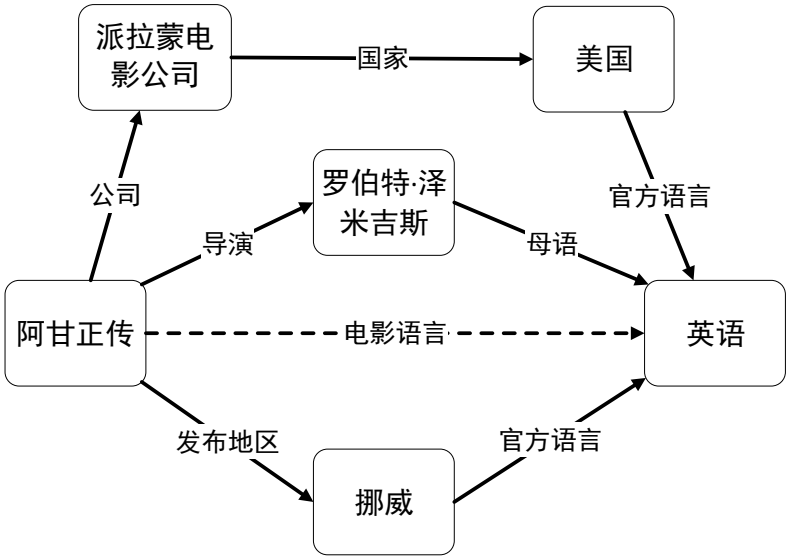
Relation1	**/people/person/place_of_birth**
Relation2	**/location/administrative_division/country**
1	/people/person/nationality
2	/people/person/places_lived./people/place_lived/location
3	/people/person/place_of_birth
4	/music/artist/origin
5	/olympics/olympic_athlete_affiliation/country
6	/government/politician/government_positions_held
7	/base/popstra/vacation_choice/location
8	/people/deceased_person/place_of_death
9	/government/political_appointer/appointees
10	/location/administrative_division/country

PTransE样例

Relation1	**/location/location/contains**
Relation2	**/location/location/contains**
1	/location/location/contains
2	/location/country/second_level_divisions
3	/location/country/administrative_divisions
4	/location/administrative_division/capital
5	/base/locations/continents/countries_within
6	/base/aareas/schema/administrative_area/administrative_children
7	/location/us_county/hud_county_place
8	/location/country/capital
9	/location/country/first_level_divisions
10	/travel/travel_destination/tourist_attractions

小结

- 考虑关系间更复杂的推理规则
  - 头、尾实体不完全一致的情况



(奥巴马，总统，美国)

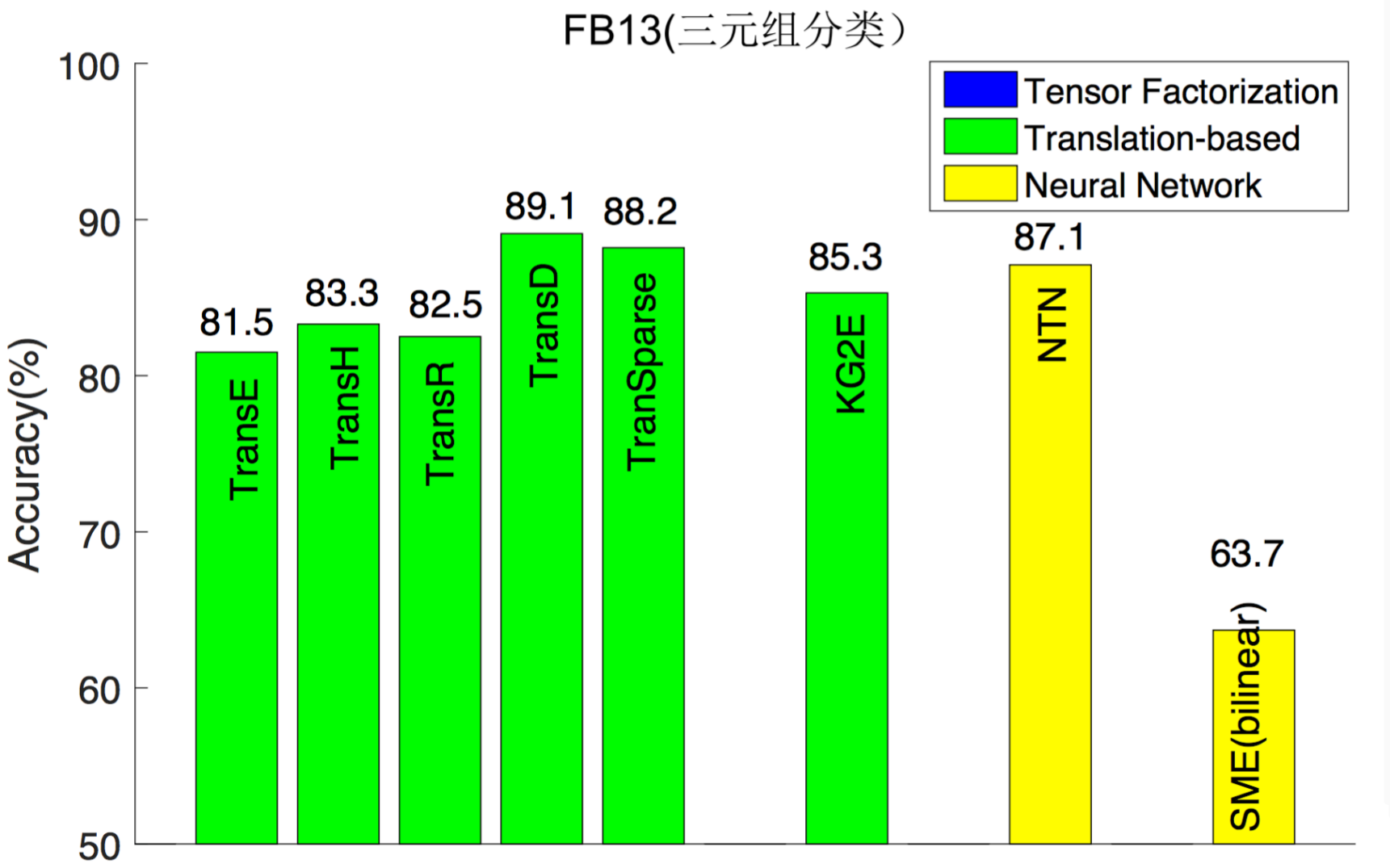
(奥巴马，是，美国人)

- 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.

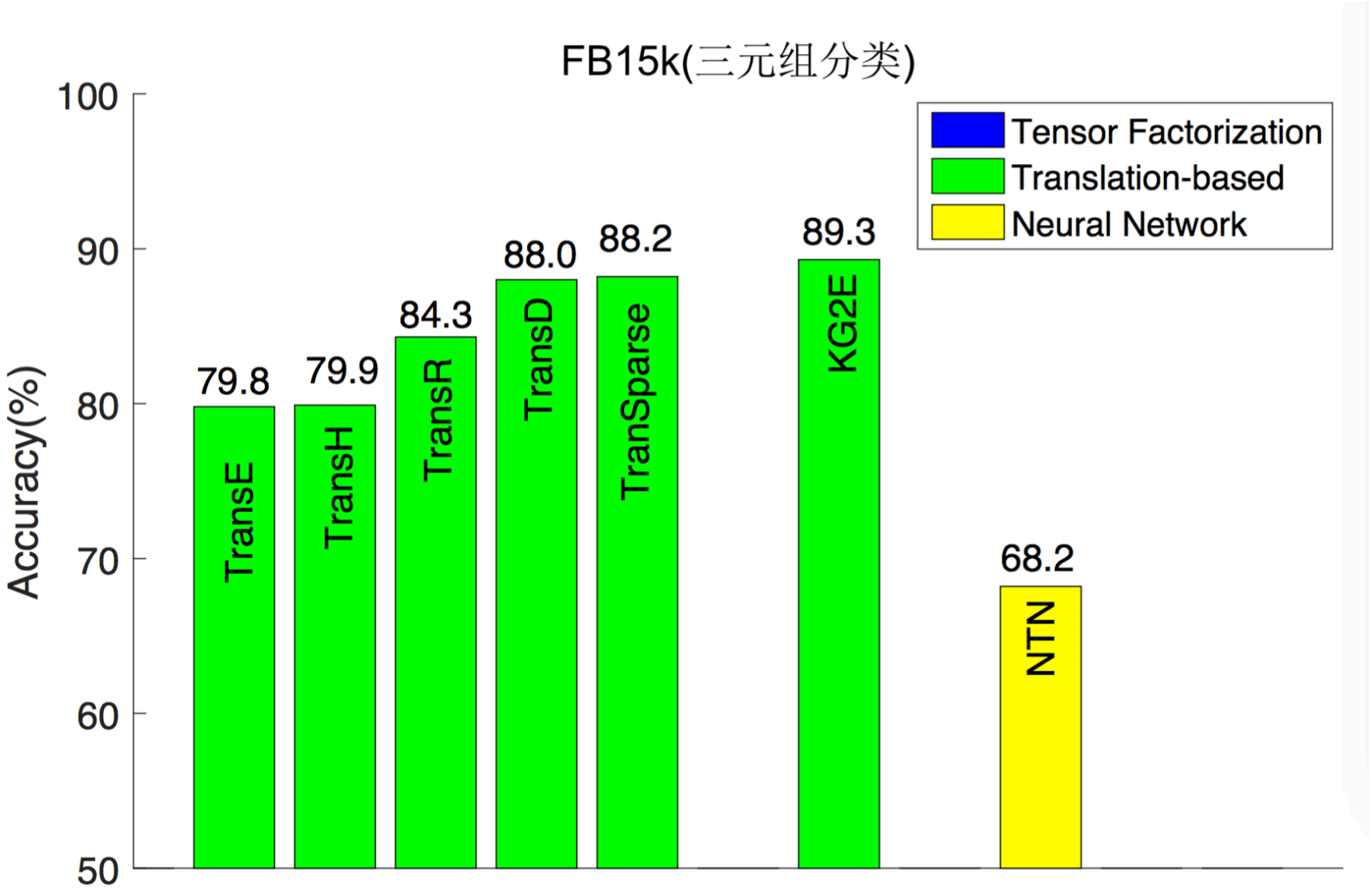
性能比较

三元组分类



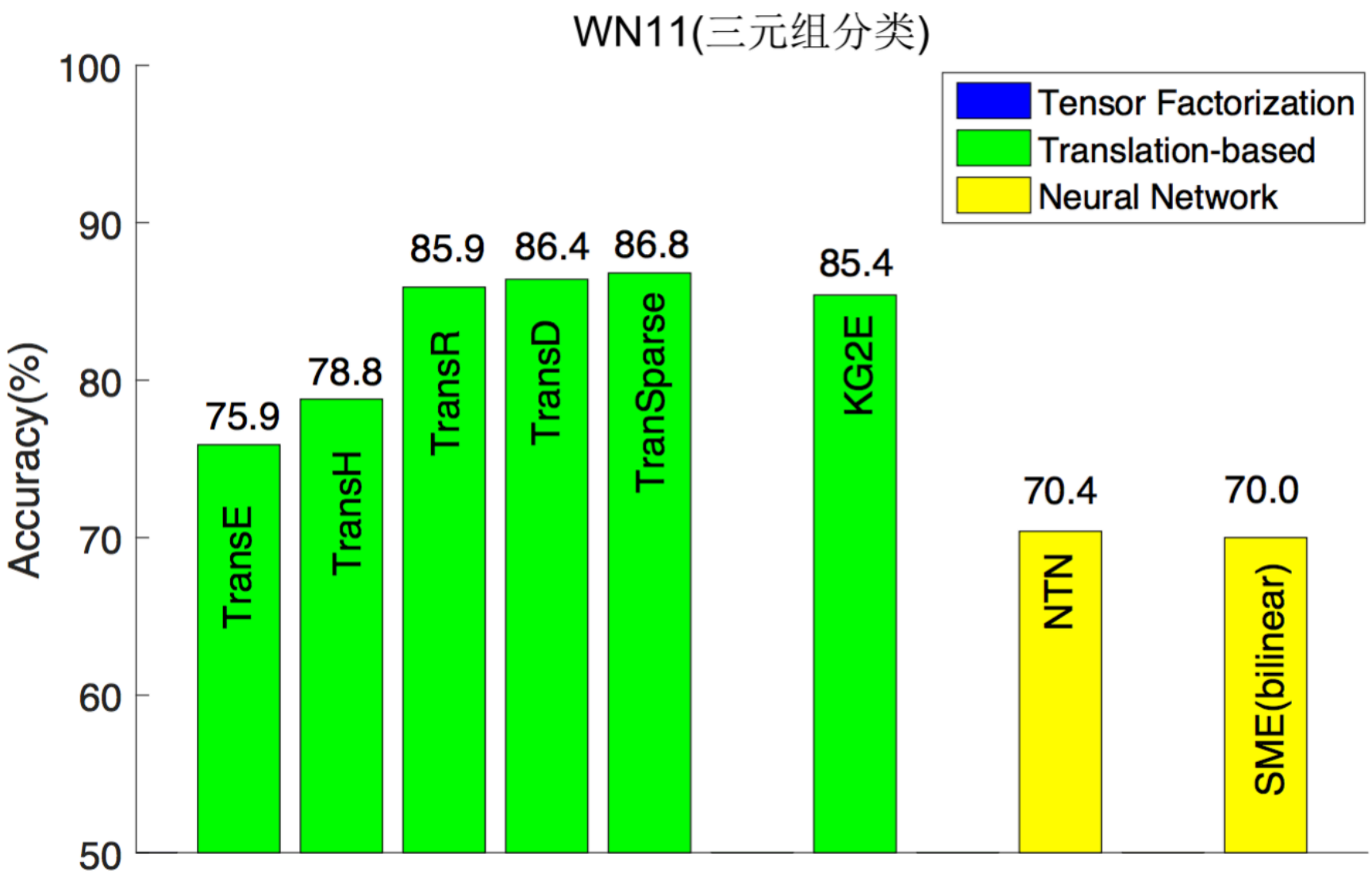
来自中科院自动化所赵军、刘康老师组

三元组分类



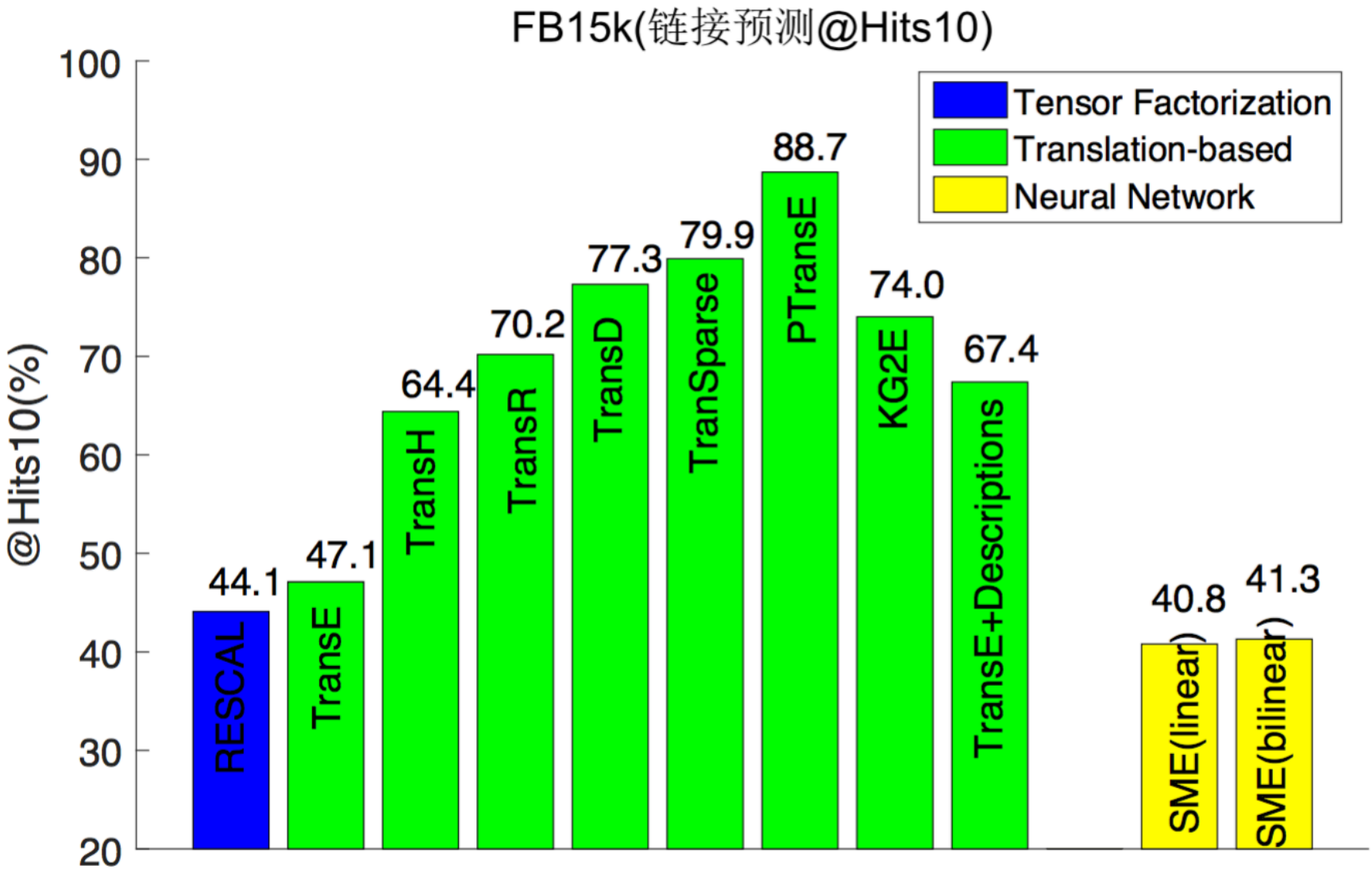
来自中科院自动化所赵军、刘康老师组

三元组分类



来自中科院自动化所赵军、刘康老师组

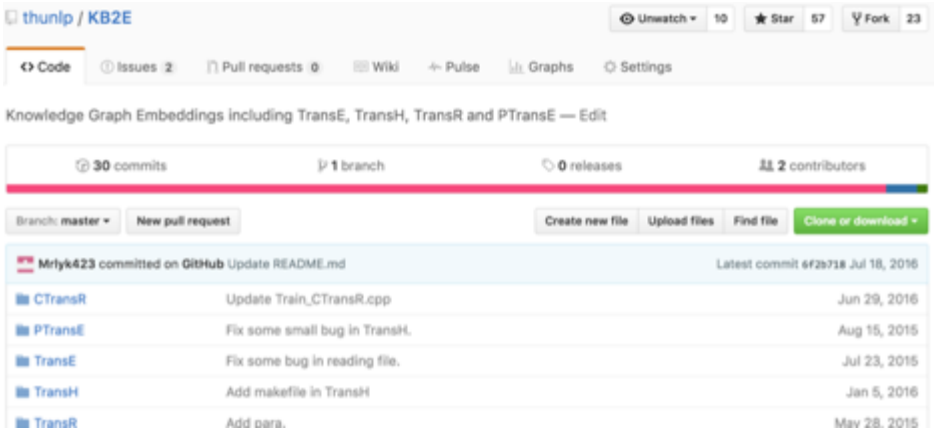
链接预测

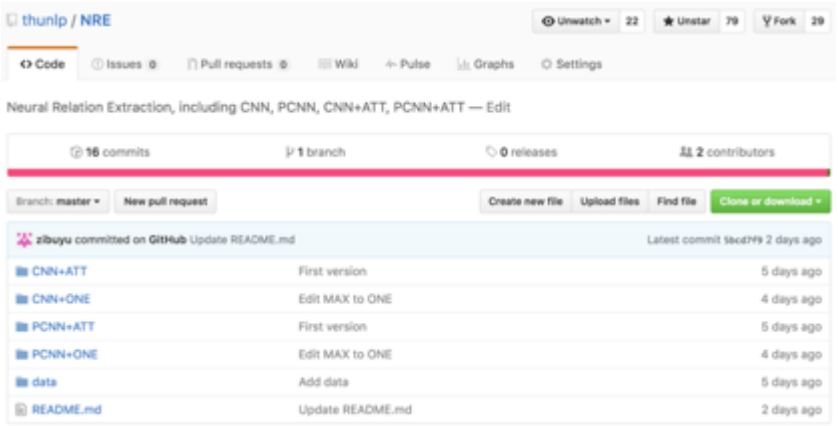


来自中科院自动化所赵军、刘康老师组

开源代码

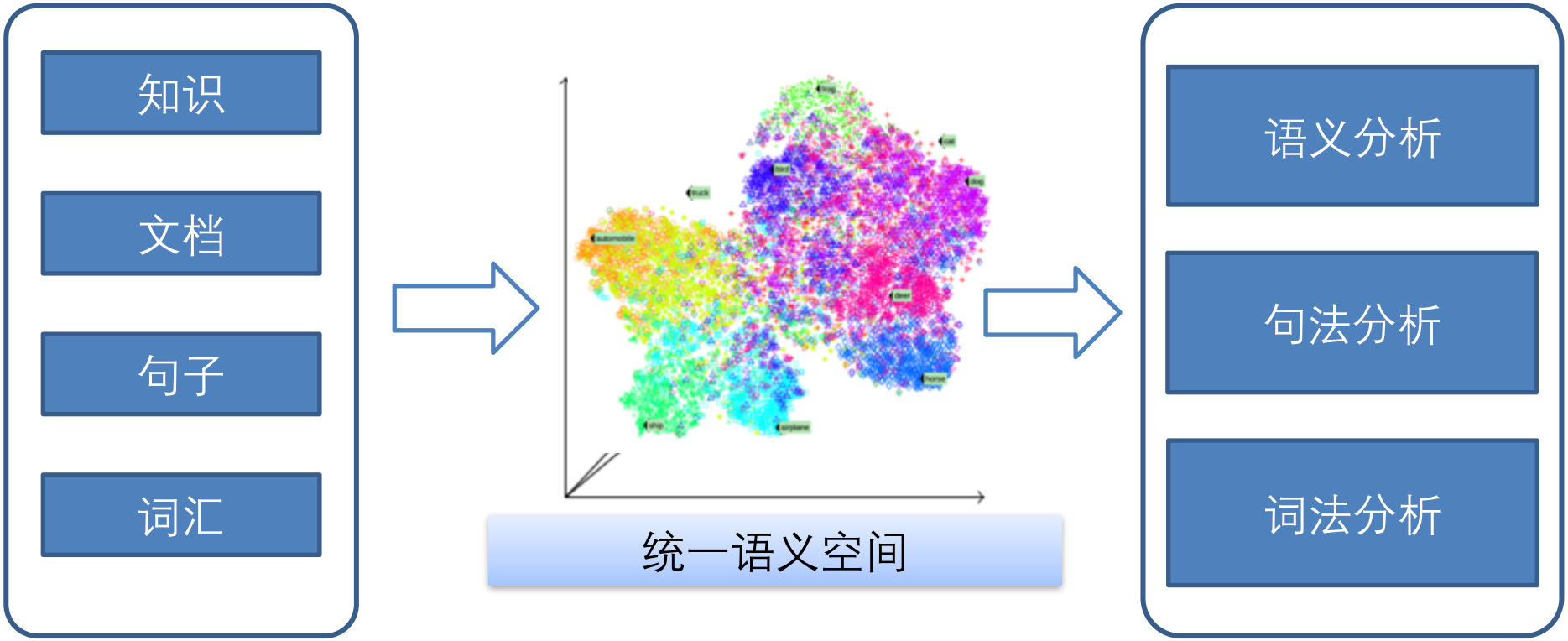
- KB2E: TransE、TransH、TransR、PTransE
  - https://github.com/thunlp/KB2E
- NRE: CNN、PCNN、x+ATT
  - https://github.com/thunlp/NRE

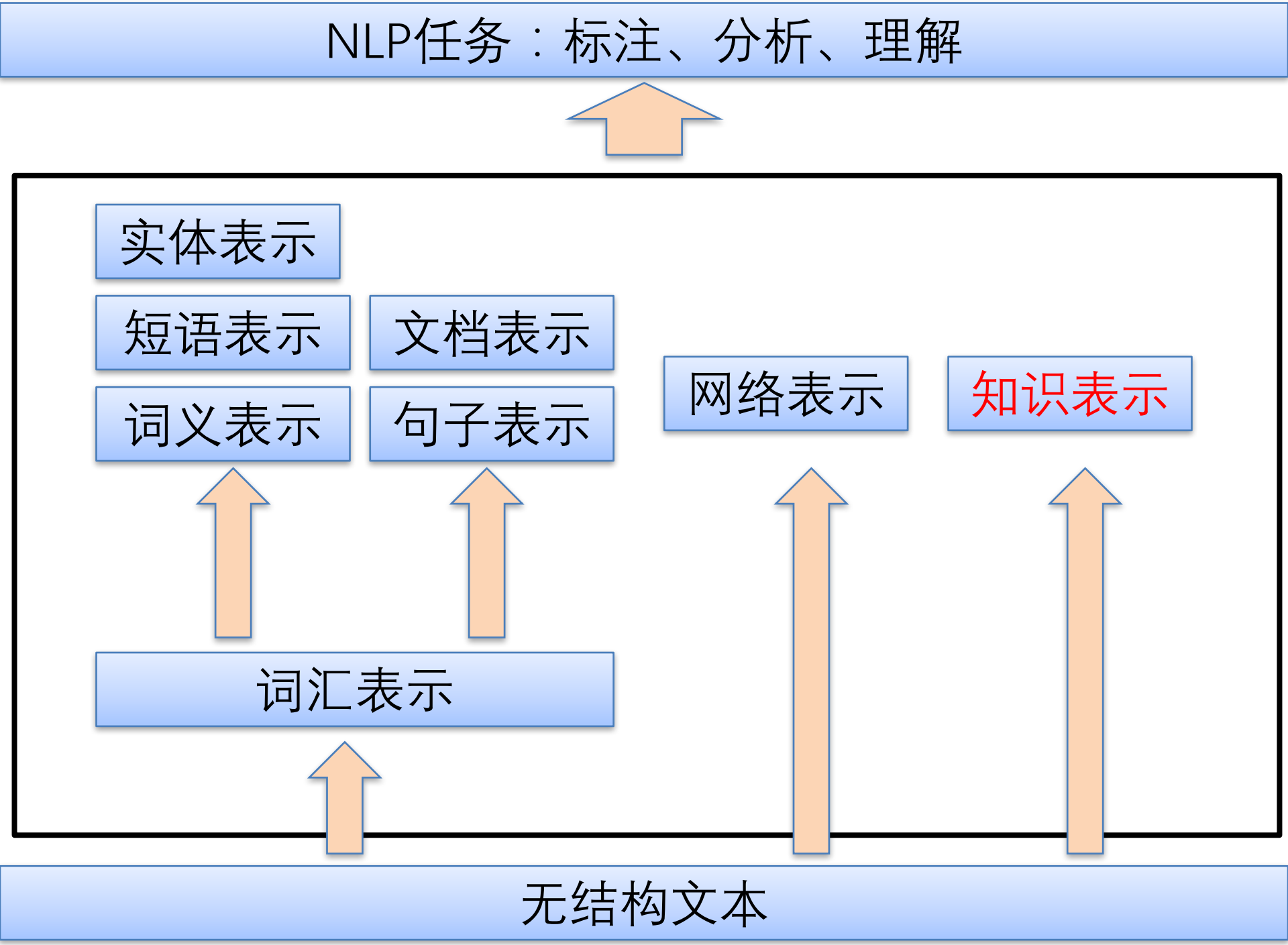




分布式表示对自然语言处理的意义

- 解决大数据NLP的数据稀疏问题
- 实现跨领域、跨对象的知识迁移
- 提供多任务学习的统一底层表示




NLP任务：标注、分析、理解
实体表示
短语表示
文档表示
词义表示
句子表示
网络表示
知识表示
词汇表示
无结构文本

自然语言处理发展趋势

- 深度学习和知识图谱为NLP发展带来无限可能
- 融合知识与文本，实现知识驱动的文本理解
  - 句子层面：问答系统、人机对话
  - 篇章层面：文档摘要、阅读理解
- 结合领域知识，实现知识驱动的文本生成
  - 法律：法律文书
  - 知识产权：专利
  - 金融投资：资讯
  - 科学研究：论文

总结

- 分布式表示将对象语义信息编码到低维向量空间中
  - 分布式表示已被广泛应用于汉字、词汇、词义、实体、短语、句子、文档、网络和知识的表示
  - 分布式表示可扩展性强，可有效解决数据稀疏问题，用于跨领域、跨对象的语义计算和知识迁移
- 知识图谱是对人类知识的结构化总结
  - 知识表示学习能够高效编码结构知识的语义信息
  - 知识图谱能够支持智能关联与推理
- 深度学习与知识图谱为自然语言处理带来无限可能
  - 知识驱动的文本理解
  - 知识驱动的文本生成

Thanks!

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

liuzy@tsinghua.edu.cn