

《大词林》中实体类型获取及层次化 构建方法

哈尔滨工业大学计算机学院

秦兵

2015.11.9 湖北宜昌

《大词林》构建缘起

背景

■ 命名

- 本质是区分事物
- 具有不同的粒度
 - 姚明
 - 阿司匹林

■ 命名实体识别

- 识别文本中事物的名称并分类

乔布斯 创立 了 苹果 公司 。

背景

- MUC-6(1995)首次提出命名实体识别任务
 - 实体名：人名、地名、机构名。
 - 时间表达式
 - 数字表达式
- 缺陷：实体类型太局限，不能满足实际需求
 - 如开放域问答
 - 小孩感冒吃什么药比较好？



传统
命名实体

背景

- 关于更广泛类别命名实体识别的研究
 - ACE-2007: 7大类, 45小类
 - Sekine和Nobata (2004) : 200类
 - Sekine (2008) : 150类
 - Ling和Weld (2012) : 112类
 - Yosef等 (2013) : 505类
 -
- 类别均为事先定好的, 仍为分类问题

背景

■ 开放域实体

● 类别更多，且不限定

- 药品名：阿司匹林、双黄连口服液；
- 会议名：20国集团伦敦峰会、中央经济工作会议；
- 电影名：《泰囧》、《苹果》；
-

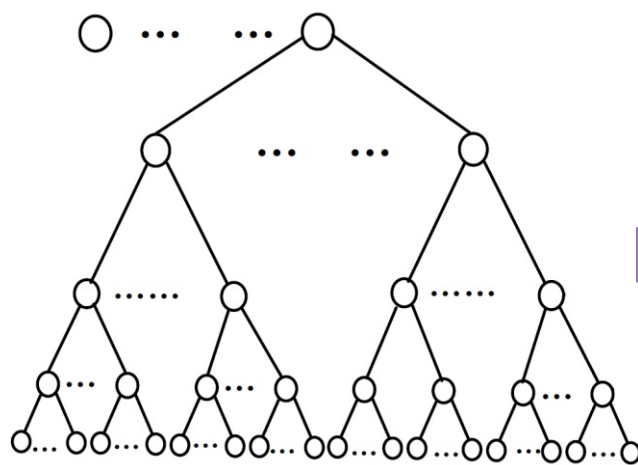
● 类别更细，且有层次

- 机构名 {
- 球队名：达拉斯小牛队、国际米兰；
 - 企业名：微软、中石油；
 - 高校名：哈尔滨工业大学、麻省理工学院；
 -

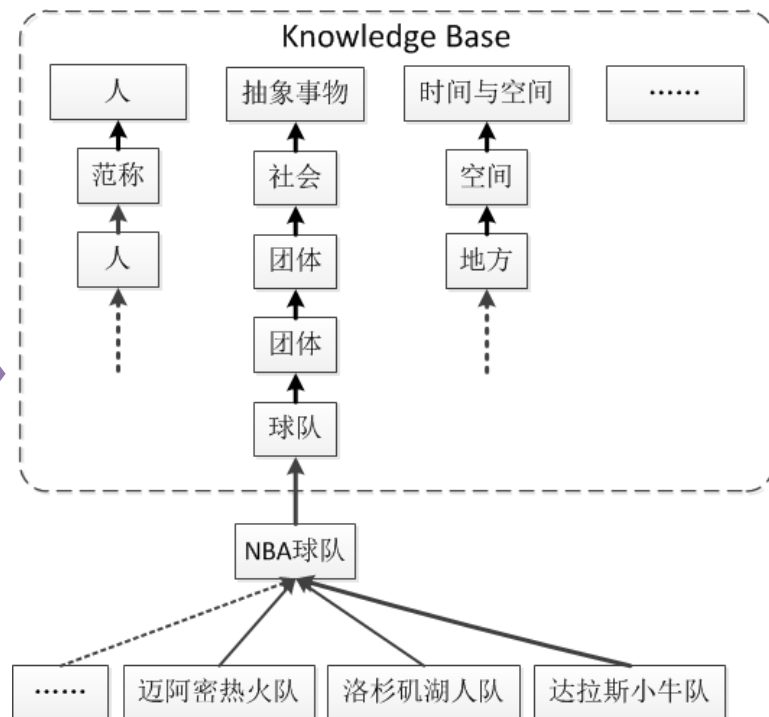
动机

NE₁
NE₂
·
·
·
NE_i
·
·
·
NE_n

+



同义词词林（扩展版）



- 上位词抽取
- 上位词层次化

开放域实体类别获取

问题

■ 开放域实体的类别

● 上位词

- 东北虎：动物、猫科动物、肉食动物
- 马尔克斯：作家、小说家、拉美作家

● 如果 x 是一个（种） y ，则

- y 是 x 的上位词
- x 是 y 的下位词

记为 $x \xrightarrow{H} y$

■ 任务

- 给定实体，挖掘其一个或多个上位词

获取上位词的方法

- 基于百科知识库对人工构建的词典进行细化和扩充
 - 如YAGO: Wikipedia + WordNet
 - Suchanek et al. (2008)
- 基于模式匹配的方法

咖啡 是一种 饮料。

苹果、梨 等 水果。

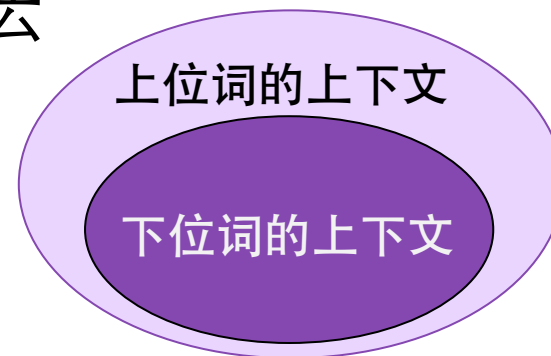
Pattern	Translation
w 是[一个 一种] h	w is a [a kind of] h
w [、] 等 h	w[,] and other h
h [,] 叫[做] w	h[,] called w
h [,] [像]如 w	h[,] such as w
h [,] 特别是 w	h[,] especially w

- Hearst (1992) ; Snow et al. (2005)

获取上位词的方法

■ 基于词语分布相似度的方法

- 一个假设
 - 上位词的上下文更宽泛
 - 真包含其下位词的上下文
- 有向的相似度计算方法



$$WeedsPrec(u, v) = \frac{\sum_{f \in F_u \cap F_v} w_u(f)}{\sum_{f \in F_u} w_u(f)}$$

$\sum_{f \in F_u \cap F_v} w_u(f)$	u和v上下文的交集
$\sum_{f \in F_u} w_u(f)$	u的上下文

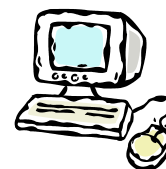
- Weeds et al. (2003; 2004)
- Kotlerman et al. (2010); Lenci and Benotto (2012)

奥巴马 vs. 美国人

基于多信息源的上位词挖掘



对一个陌生实体的认知过程



猜

&

问

按字面信息推理

&

Web 挖掘

可口可乐公司

Bai du 百度

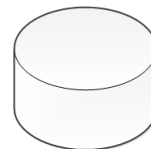
Google



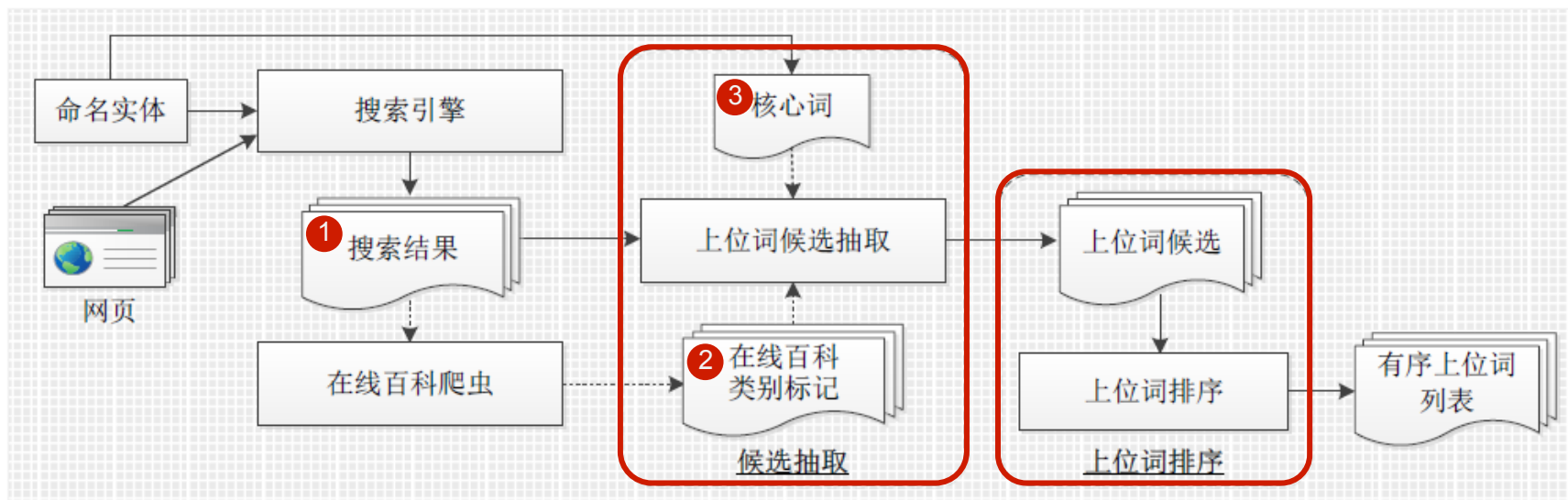
认知类别，融合到知识体系中



获取类别，融入语义词典中



方法框架



- 候选上位词抽取(多信息源):

- ① 搜索结果
- ② 在线百科类别标记
- ③ 核心词

- 上位词排序

上位词候选抽取

■ 搜索结果

- 选取与当前命名实体共现频率最高的N个名词或名词短语

假设：命名实体与其上位词高频共现



理由：命名实体通常有介绍性的网页

N = 10 时，覆盖率为 87%

Baidu 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

"奥扎格雷钠氯化钠"

[奥扎格雷钠氯化钠注射液_百度百科](#)

说明 【药品名称】：奥扎格雷钠氯化钠注射液 【英文或拉丁名】：Sodium Ozagrel and Sodium Chloride Injection 【汉语拼音】：Aozhageleina Lühuana Zhushen...
[说明 - 适应症 - 用法与用量 - 不良反应 - 禁忌症 - 更多>>](#)
[baike.baidu.com/view/17510...htm 2012-12-26 - 百度快照](#)

[奥扎格雷钠氯化钠注射液说明书_百度文库](#)

★★★★★ 评分:4/5 2页
 奥扎格雷钠氯化钠注射液说明书 隐藏>> 注射用奥扎格雷钠 资料编号 5 注射用奥扎格雷钠说明书 【药品名称】 通用名:注射用奥扎格雷钠 【适应症】 本品为血栓...
[wenku.baidu.com/view/b4003aebaeaad1f... 2012-2-9 - 百度快照](#)

[奥扎格雷钠氯化钠注射液_药品说明书_价格_适应证_作用_山东华鲁...](#)
 奥扎格雷钠氯化钠注射液 生产厂商:山东华鲁制药有限公司 功效主治: 用于治疗急性血栓性脑梗死和脑梗死所伴随的运动障碍。 用法用量:成人每天2次,每次100ml(80mg)...
[ypk.xywy.com/html/352/1758...html 2013-3-12 - 百度快照](#)

上位词候选抽取

■ 在线百科知识库中的类别标签

奥扎格雷钠氯化钠注射液

[编辑](#)


★ 收藏

👍 20

🔗 12

奥扎格雷钠氯化钠注射液用于治疗急性血栓性脑梗死和脑梗死所伴随的运动障碍。

英文名称	Ozagrel and Sodium Chloride Injection	药品类型	化学药品
	Sodium Ozagrel and Sodium Chloride Injection		脑代谢功能促进剂
是否处方药	处方药	规格	100ml: 奥扎格雷钠80mg与氯化钠0.9g
剂型	注射剂		250ml: 奥扎格雷钠80mg与氯化钠2.25g



奥扎格雷钠氯化钠注射液图

开放分类：

医学

应用科学

科学

胃药

药品

药物中文名称列表

西药

注射液

药物

上位词候选获取

■ 核心词

- 奥扎格雷钠氯化钠注射液； 大众公司
- 复仇者联盟； 水煮鱼
- 汉语命名实体末尾词即当做其核心词
- 实验表明，在我们的数据集中有**40.6%**的命名实体的核心词是一个正确的上位词

小丑鱼 (Clownfish)

Search Results

Baidu 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

小丑鱼

百度一下

Encyclopedia Category Tags

Baidubaike

新闻 网页 贴吧 知道 音乐 图片 视频 地图 百科 文库

小丑鱼

进入词条 搜索词条 帮助

首页 分类频道 特色百科 玩转百科 百科用户 百科校园 百科合作

小丑鱼

编辑 1451 1457

小丑鱼是对雀鲷科海葵鱼亚科鱼类的俗称，是一种热带咸水鱼。已知有28种，一种来自棘鲷属（*Premnas*），其余来自双锯鱼属（*Amphiprion*）。小丑鱼与海葵有着密不可分的共生关系，因此又称海葵鱼。带毒刺的海葵保护小丑鱼，小丑鱼则吃海葵消化后的残渣，形成一种互利共生的关系。

词义标签：海洋动物 鱼类 鲈形目 雀鲷科 鱼纲 动物

Hudongbaike

互动百科 行业百科 移动版

小丑鱼

编辑词条



小丑鱼（学名：Amphiprioninae）是对雀鲷科海葵鱼亚科鱼类的俗称，是一种热带咸水鱼。因为脸上都有一条或两条白色条纹，好似京剧中的丑角，所以俗称“小丑鱼”。已知有28种，一种来自棘鲷属（*Premnas*），其余来自双锯鱼属（*Amphiprion*）。小丑鱼与海葵有着密不可分的关系，因此又称海葵鱼。

开放分类：宠物 海水观赏鱼 海洋生物 生物 自然 节肢动物 鱼类 鲈形目

我来补充

Head Word

小丑 鱼

HEAD WORD

鱼 (fish)
鱼类 (fish)
音乐 (music)
手机 (cellphone)
歌曲 (song)
图片 (image)
歌词 (lyric)
视频 (video)
海底 (seafloor)
网 (net)

海洋动物 (marine animal)
鱼类 (fish)
鲈形目 (Perciformes)
雀鲷科 (Pomacentridae)
鱼纲 (Pisces)
动物 (animal)

宠物 (pet)
海水观赏鱼 (seawater fancy fish)
海洋生物 (marine organism)
生物 (organism)
自然 (nature)
节肢动物 (Arthropod)
鱼类 (fish)
鲈形目 (Perciformes)

鱼 (fish)

Ranked Hypernyms

鱼类 (fish) : 1.00
鱼 (fish) : 1.00
动物 (animal) : 1.00
生物 (organism) : 1.00
海洋动物 (marine animal) : 0.99
鲈形目 (Perciformes) : 0.99
音乐 (music) : 0.94
海洋生物 (marine organism) : 0.84
手机 (cellphone) : 0.27
鱼纲 (Pisces) : 0.24
雀鲷科 (Pomacentridae) : 0.24
宠物 (pet) : 0.15
海水观赏鱼 (seawater fancy fish) : 0.14
节肢动物 (Arthropod) : 0.13
歌曲 (song) : 0.13
图片 (image) : 0.13
歌词 (lyric) : 0.13
视频 (video) : 0.13
海底 (seafloor) : 0.12
网 (net) : 0.12

上位词排序

■ 排序的特征

特征	含义	取值范围
Prior	上位词先验概率	[0, 1]
Is_Tag	是否为百科类别标签	0 or 1
Is_Head	是否为核心词	0 or 1
In_Titles	候选在搜索结果标题中出现比例是否符合一定要求	0 or 1
Synonyms	当前候选的同义词所占全部候选的比例	[0, 1]
Radicals	候选中汉字和实体最末一个字偏旁匹配的比例	[0, 1]
Source_Num	候选来源数	1, 2, 3, or 4
Lexicon	候选上位词本身及其核心词	0 or 1

■ 模型

- SVM; Logistic Regression

上位词排序

- 训练语料自动构建
 - 从百科知识库中抽取实体名
 - 从多信息源抽取上位词候选
 - 启发式选取正反例
 - 原则1:上位词的来源越多越可能是正确的
 - 原则2:上位词的先验概率越高,越可能是正确的
 - Top1 选为正例
 - Bottom1 选为反例

数据准备

■ 1,879 开放域实体

领域	命名实体数	
	开发集	测试集
生物	72	351
医疗	61	291
食品	75	303
电影	51	204
工业	56	224
其他	35	136
Total	350	1529

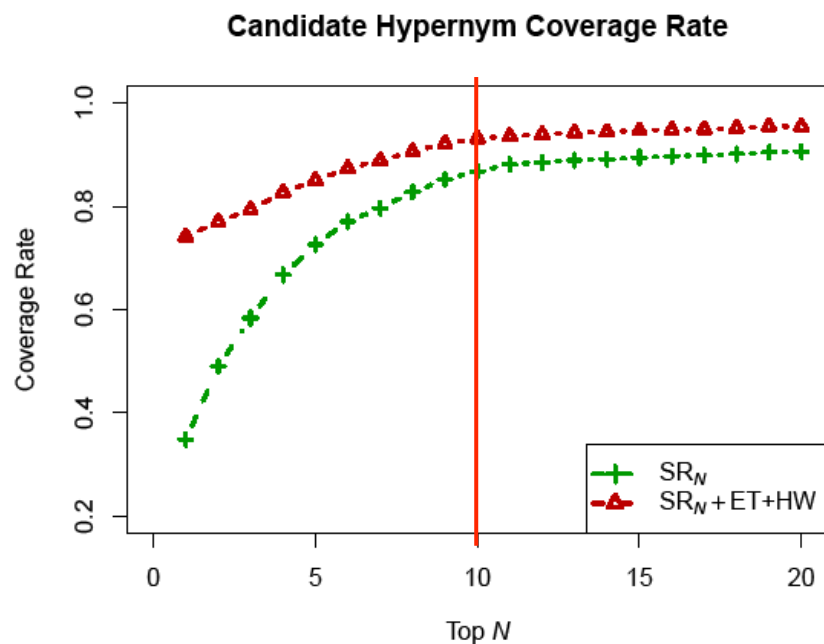
■ 评测语料标注

- 自动抽取上位词候选
- 人工标注是否正确

■ 评测语料统计分析

- 每个实体平均有12.53个候选
- 其中正确的上位词平均为2.09个
- Kappa值为0.79

候选上位词覆盖率

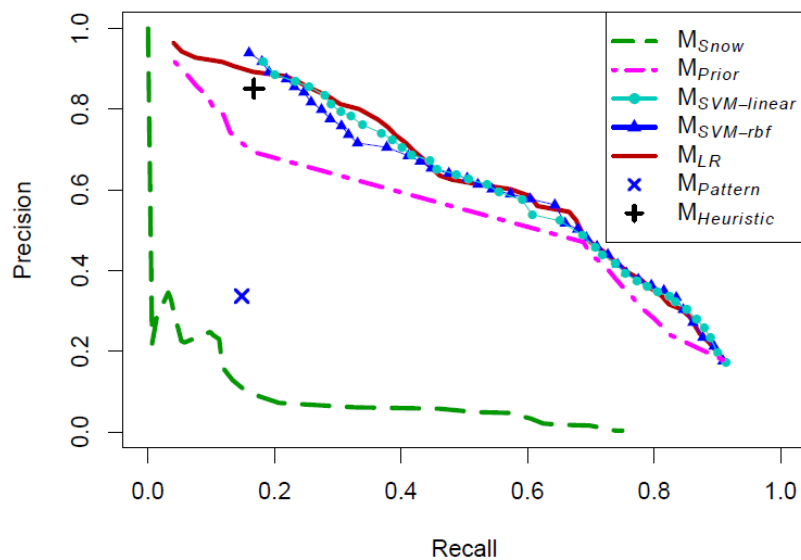


来源	覆盖率	平均数量
SR_{10}	0.8691	9.44*
ET	0.3938	3.07
HW	0.4135	0.87 [†]
$SR_{10} + ET$	0.8909	12.02
$SR_{10} + HW$	0.9117	9.75
ET + HW	0.7073	3.92
$SR_{10} + ET + HW$	0.9324	12.53

上位词排序

	方法	在线百科中的命名实体		在线百科外的命名实体		全部命名实体	
		P@1	R-Prec	P@1	R-Prec	P@1	R-Prec
模式匹配	$M_{Pattern}$	0.5542	0.4937	0.4306	0.3638	0.5229	0.4608
	M_{Snow}	0.3199	0.2592	0.2827	0.2610	0.3092	0.2597
词语分布	$M_{balAPin}$	0.2564	0.2450	0.1492	0.1693	0.1981	0.2038
	M_{invCL}	0.3266	0.2840	0.1540	0.1769	0.2328	0.2258
先验概率	M_{Prior}	0.7339	0.5483	0.3940	0.3531	0.5494	0.4423
本文方法	$M_{SVM-linear}$	0.8569	0.6899	0.6157	0.5837	0.7260	0.6322
	$M_{SVM-rbf}$	0.8484	0.6940	0.6241	0.5901	0.7266	0.6376
	M_{LR}	0.8612	0.7052	0.6807	0.6258	0.7632	0.6621

上位词排序



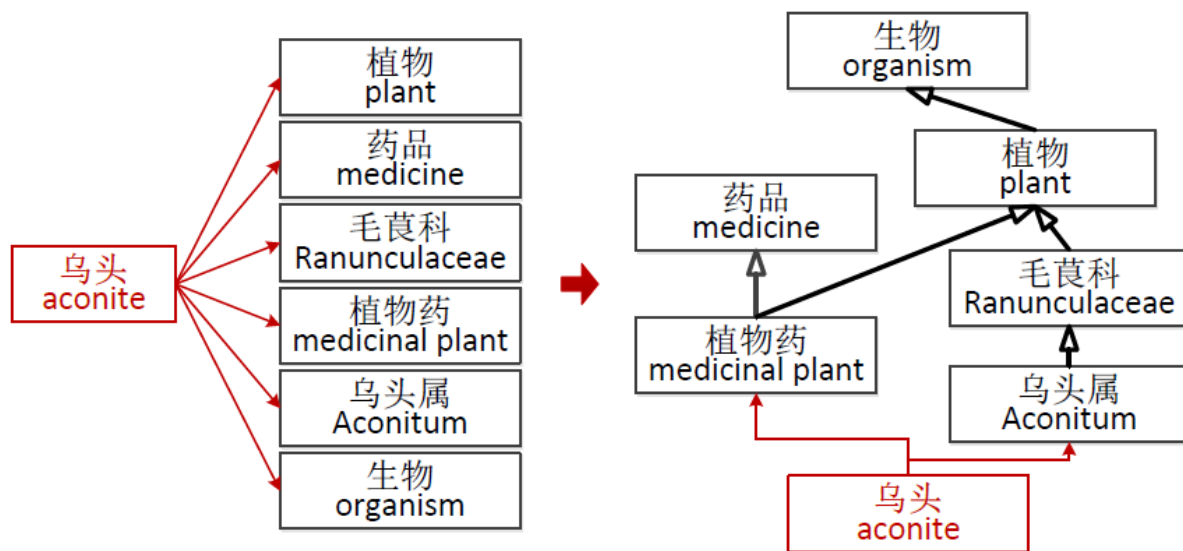
方法	准确率	召回率	F1值
$M_{Pattern}$	0.3366	0.1485	0.2061
M_{Snow}	0.2300	0.1128	0.1514
$M_{balAPin}$	0.1772	0.3324	0.2312
M_{invCL}	0.2261	0.2617	0.2426
$M_{Heuristic}$	0.8516	0.1678	0.2803
M_{Prior}	0.4706	0.6885	0.5591
$M_{SVM-linear}$	0.5857	0.5880	0.5868
$M_{SVM-rbf}$	0.5721	0.6338	0.6014
M_{LR}	0.5458	0.6657	0.5998

各种方法最好的F1值

开放式实体类别层次化

问题

■ 上位词之间的层次化



相关工作

- 基于百科知识库扩展人工词典的方法
 - 直接构成层次化结构

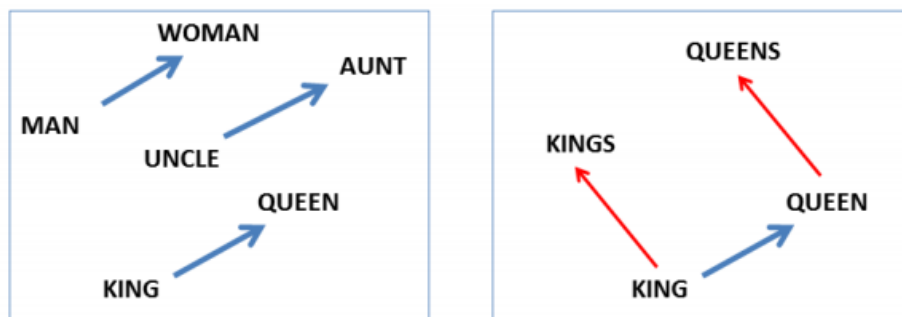
 - 基于模式匹配的方法
 - 基于词语分布相似度的方法
 - 基于前述的方法
- } 挖掘上下位词对，
组合成层次化结构

词汇分布表示简介

- 词汇分布表示 (Word Embedding)
 - 稠密的、低位的、实数值的词向量

词汇 $\Rightarrow$ $[0.28, 0.05, \dots, 0.56]$

$$v(\text{king}) - v(\text{queen}) \approx v(\text{man}) - v(\text{woman})$$



Mikolov et al. (2013a)

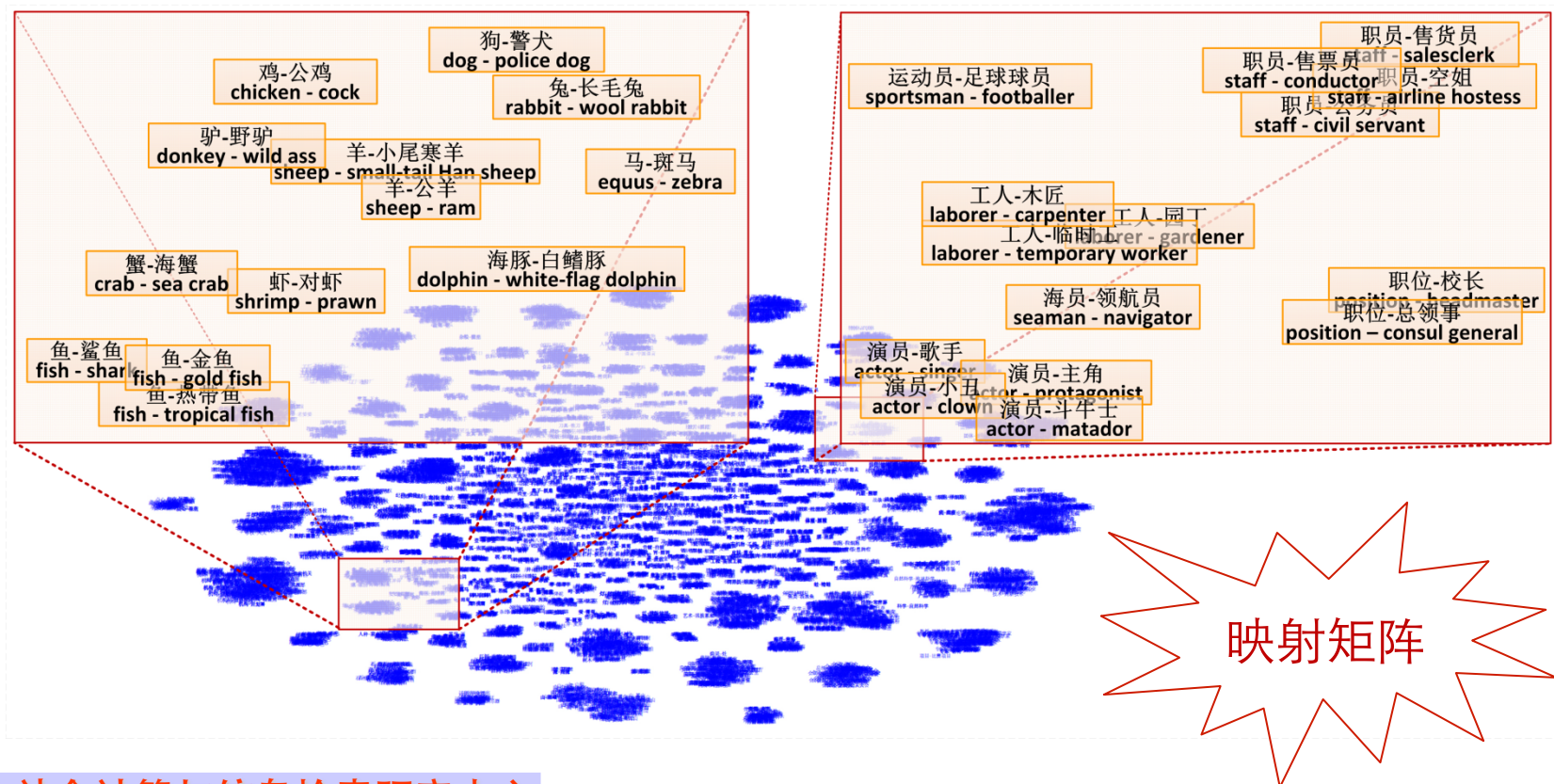
动机

- 用词向量之间的差值（偏移）能否刻画上下位关系呢？

No.	Examples
1	$v(\text{虾}) - v(\text{对虾}) \approx v(\text{鱼}) - v(\text{金鱼})$ $v(\text{shrimp}) - v(\text{prawn}) \approx v(\text{fish}) - v(\text{gold fish})$
2	$v(\text{工人}) - v(\text{木匠}) \approx v(\text{演员}) - v(\text{小丑})$ $v(\text{laborer}) - v(\text{carpenter}) \approx v(\text{actor}) - v(\text{clown})$
3	$v(\text{工人}) - v(\text{木匠}) \not\approx v(\text{鱼}) - v(\text{金鱼})$ $v(\text{laborer}) - v(\text{carpenter}) \not\approx v(\text{fish}) - v(\text{gold fish})$

动机

- “上位词 - 下位词”的分布
 - 数据来源于同义词词林（扩展版）

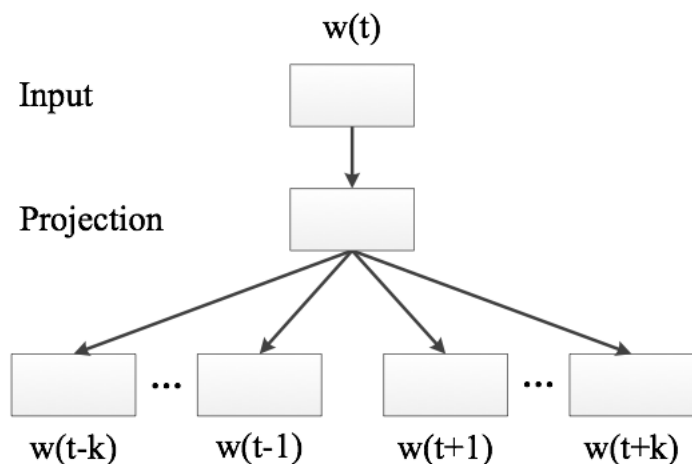


方法

- 词汇分布表示的训练
- 映射矩阵的学习
- 上下位关系的判别

词汇分布表示的训练

■ 跨词元模型 (Skip-gram)



$$\mathcal{L} = \sum_{t=1}^T \sum_{-k \leq j \leq k, j \neq 0} \log p(w_{t+j} | w_t)$$

$$p(w_i | w_j) = \frac{\exp(v_{w_i}^\top v_{w_j})}{\sum_{t=1}^V \exp(v_{w_t}^\top v_{w_j})}$$

Mikolov et al. (2013)

映射学习

■ 单一线性映射

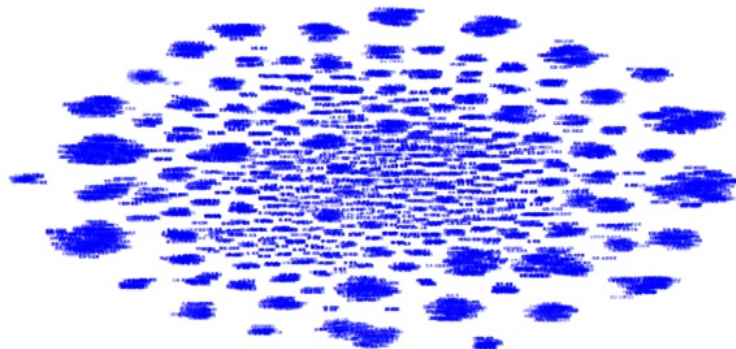
- 给定一个上下位词对($\mathbf{x}$, $\mathbf{y}$), 存在一个矩阵 Φ , 使得 $\mathbf{y} = \Phi \mathbf{x}$
- 训练方法

$$\Phi^* = \arg \min_{\Phi} \frac{1}{N} \sum_{(x,y)} \| \Phi x - y \|^2$$

- 但是, 上下位关系是多对多关系

映射学习

- 分段线性映射
 - 首先对 $y - x$ 聚类

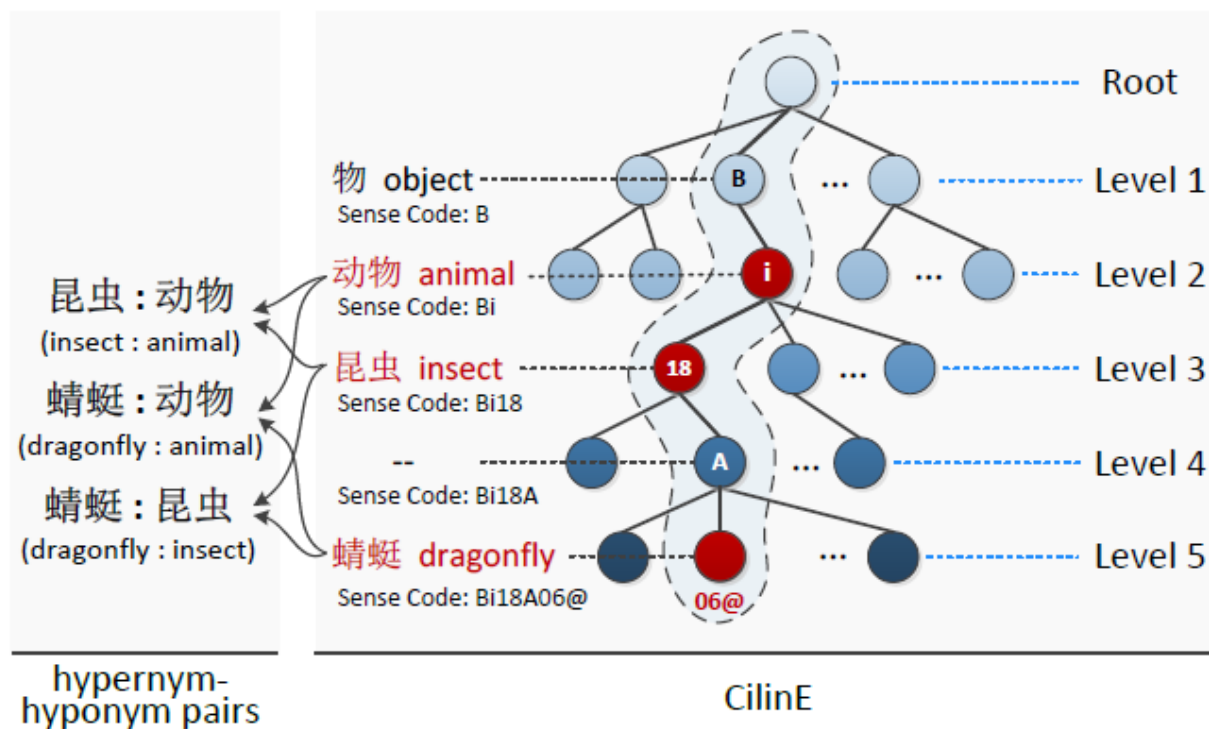


- 然后在每一个簇中分别学习映射矩阵

$$\Phi_k^* = \arg \min_{\Phi_k} \frac{1}{N_k} \sum_{(x,y) \in C_k} \| \Phi_k x - y \|^2$$

映射学习

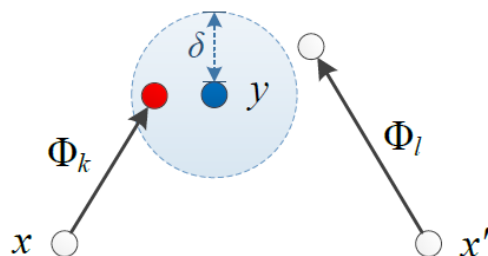
- 训练语料
 - 来自同义词词林扩展版



上下位关系判别

- 给定一个词对(x, y)
 - 如果满足一下两个条件之一，则 y 被判别为 x 的一个上位词

条件1: $d(\Phi_k x, y) = \| \Phi_k x - y \|^2 < \delta$



条件2: 存在另外一个词 z ，使得 $x \xrightarrow{H} z$ and $z \xrightarrow{H} y$

上下位关系判别

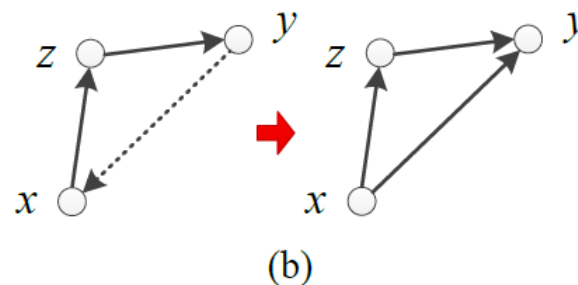
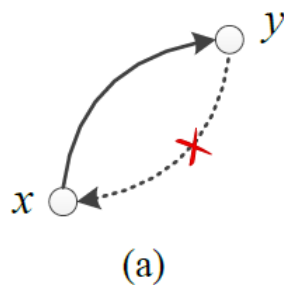
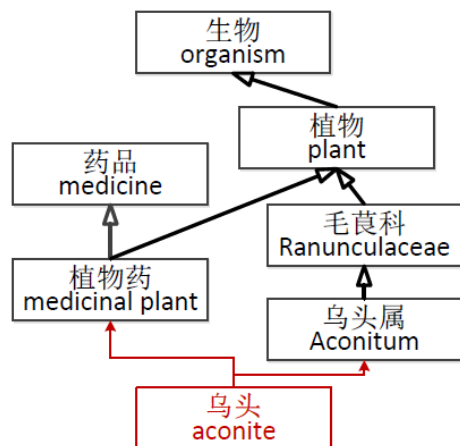
■ 层次整体约束

• 上下位关系的性质

非对称性: $\forall x, y \in L : x \xrightarrow{H} y \Rightarrow \neg(y \xrightarrow{H} x)$

传递性: $\forall x, y, z \in L : (x \xrightarrow{H} z \wedge z \xrightarrow{H} y) \Rightarrow x \xrightarrow{H} y$

• 层次化结构为有向无环图 (DAG)



数据

- 词语分布表示的训练语料
 - 百度百科正文
 - 约3000万个句子（7.8亿词）
- 映射学习训练语料
 - 同义词词林（扩展版）
 - 15,247 个上下位关系对
- 评测语料
 - 使用上述的方法为418个命名实体抽取上位词
 - 人工标注层次关系（按词对）

关系类型	词对数	
	开发集	测试集
上位-下位关系	312	1,079
下位-上位关系*	312	1,079
无关系	1,044	3,250
总计	1,668	5,408

Kappa=0.96

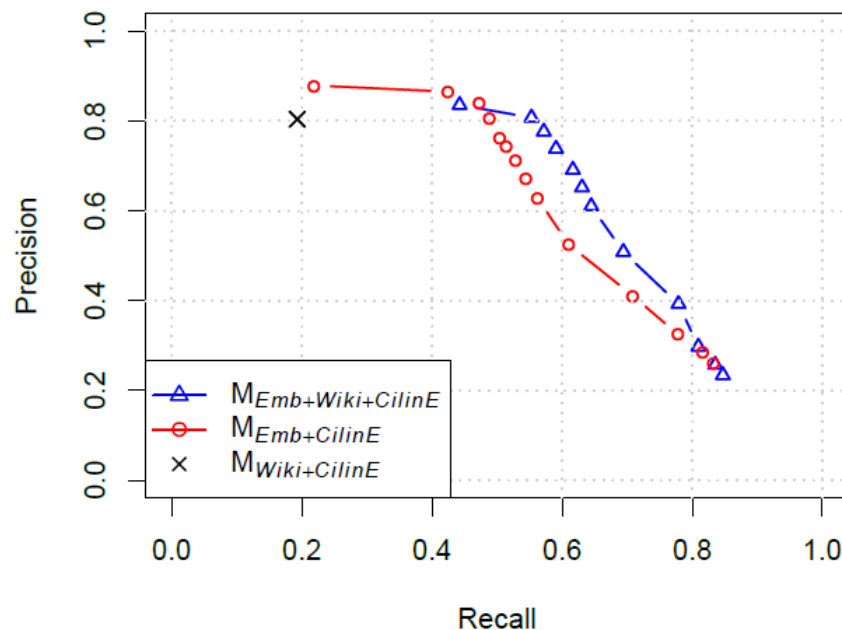
方法对比

	准确率	召回率	F1值	
$M_{Wiki+CilinE}$	0.9241	0.6061	0.7320	Suchanek et al. (2008)
$M_{Pattern}$	0.9747	0.2141	0.3511	Hearst (1992)
M_{Snow}	0.6088	0.2567	0.3611	Snow et al. (2005)
$M_{balApinc}$	0.5496	0.5338	0.5416	Kotlerman et al. (2010)
M_{invCL}	0.4963	0.6284	0.5546	Lenci and Benotto (2012)
M_{Fu}	0.8740	0.4819	0.6213	基于多信息源的方法
M_{Emb}	0.8054	0.6799	0.7374	我们的方法
$M_{Emb+CilinE}$	0.8059	0.7242	0.7629	与同义词词林结合
$M_{Emb+Wiki+CilinE}$	0.7978	0.8081	0.8029	与Suchanek et al. (2008)结合

在同义词词林之外的结果

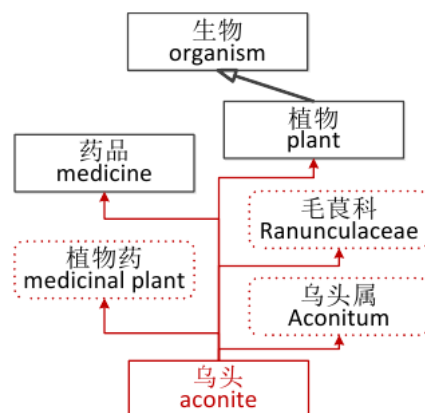
- 词对中至少有一个词不在词林中 (62%)

	准确率	召回率	F1值
$M_{Wiki+CilinE}$	0.8039	0.1929	0.3112
$M_{Pattern}$	0.7032	0.1125	0.1940
M_{Snow}	0.6547	0.1523	0.2471
$M_{balApinc}$	0.4858	0.4124	0.4461
M_{invCL}	0.3566	0.6101	0.4501
M_{Fu}	0.7804	0.3374	0.4711
M_{Emb}	0.7255	0.5180	0.6044
$M_{Emb+CilinE}$	0.7116	0.5280	0.6062
$M_{Emb+Wiki+CilinE}$	0.6913	0.6165	0.6517

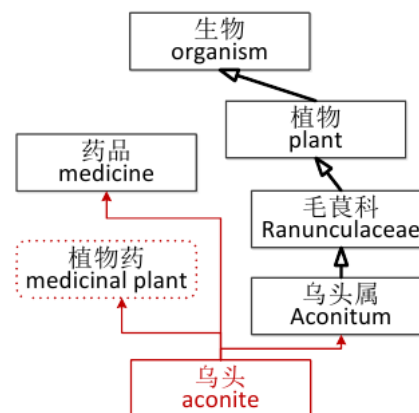


不同方法示例

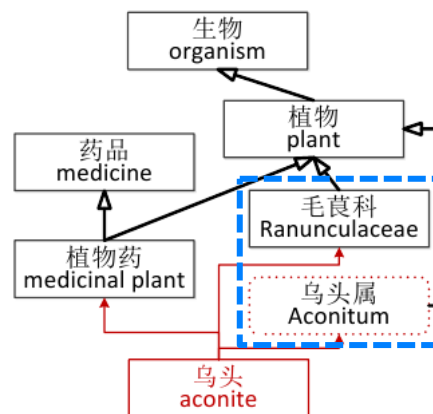
ACL2014



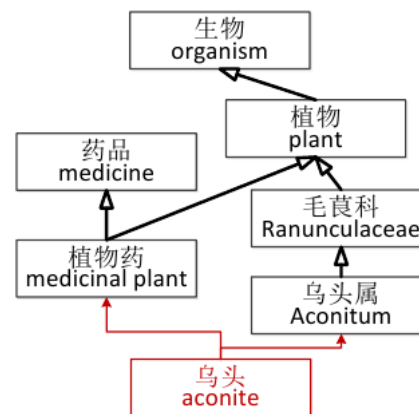
(a) CilinE



(b) Wikipedia+CilinE



(c) Embedding



(d) Embedding+Wikipedia+CilinE

演示系统

■ www.bigcilin.com



大词林树状结构

[illegible]

截止到目前的数据统计 (2015.11.8)

项目	数量
实体词	1,670,029
上位词	154,044
实体上位关系	19,114,186
上位词层次之间关系	50124

目前大词林已经被腾讯、讯飞、360等公司及研究机构付费使用

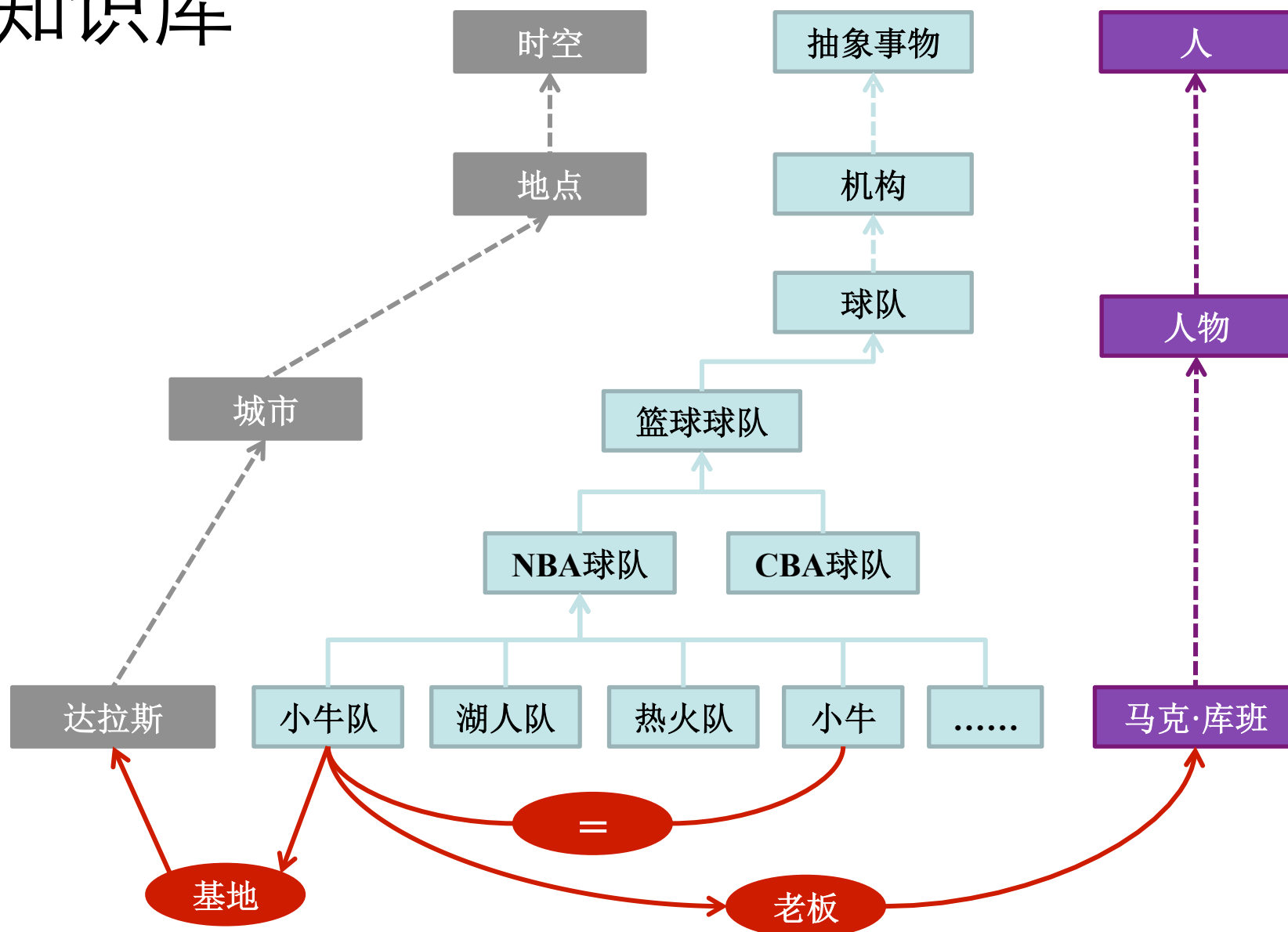
上述工作发表文章列表

- Exploiting Multiple Sources for Open-domain Hypernym Discovery. **EMNLP 2013**.
- Learning Semantic Hierarchies via Word Embeddings. ACL 2014

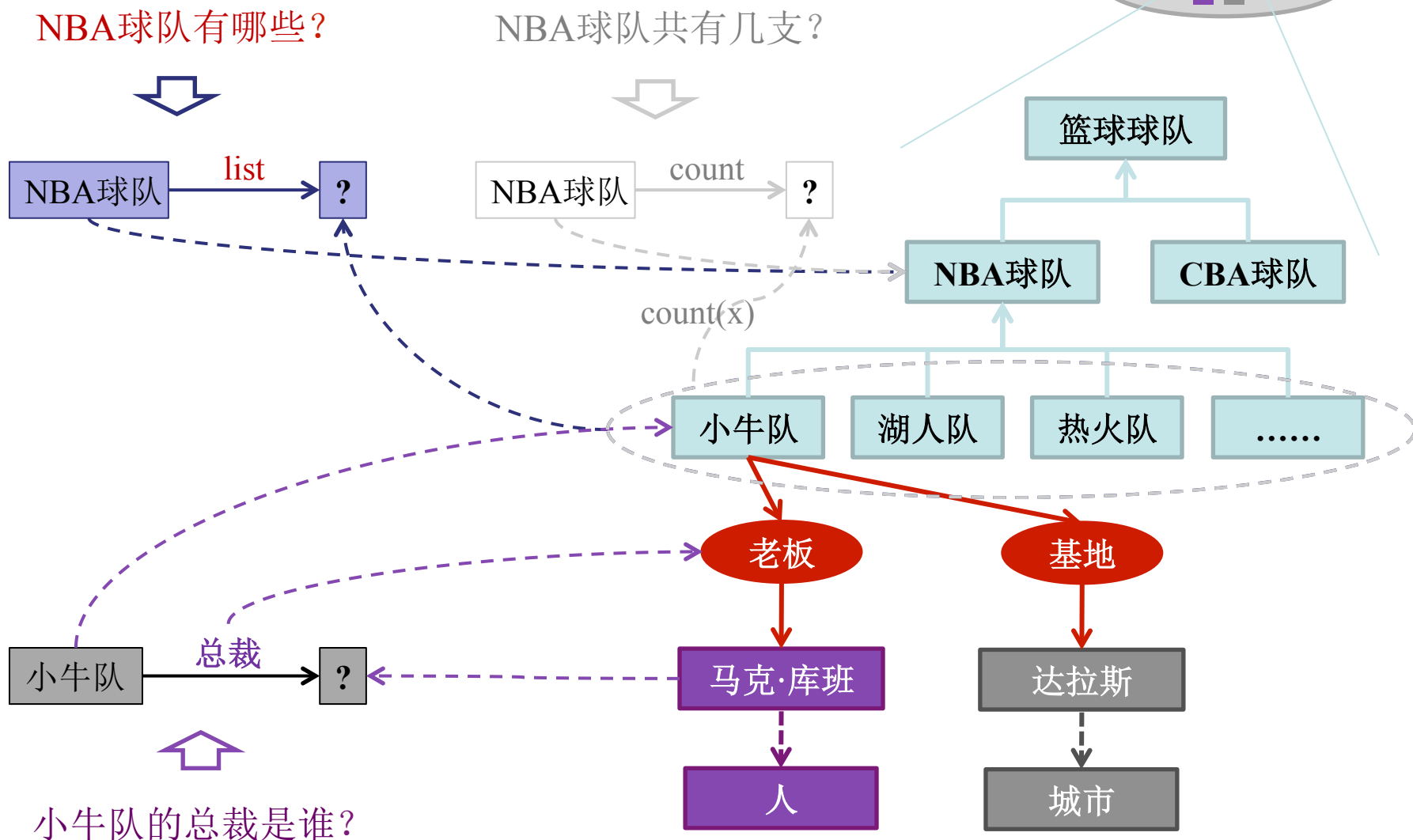
未来工作

- 大词林目前是基于上位关系的知识库
 - 无监督自动构建
 - 数据规模大
 - 准确率较高
 - 效果随着使用次数增加而提高
- 下一步工作将实体关系加入大词林，构建网状关系的知识库

知识库



应用：基于知识库的问答示例



谢谢！