面向知识图谱构建的信息抽取技术

韩先培
中国科学院软件研究所

2015-6-27

传统信息抽取

- Grishman（1997）：从自然语言文本中抽取指定类型的实体、关系、事件等事实信息，并形成结构化数据输出的文本处理技术
  - 实体
  - 实体和实体之间的关系
  - 实体参与的事件

传统信息抽取示例

据美联社消息，**当地时间7月7日清晨**，**英国伦敦金融中心的地铁**发生6次**爆炸**，其中还包括一辆满载乘客的双层公共汽车。由于事发当时处于上班的高峰时期，造成了大量人员伤亡。据初步统计的数字，多起爆炸已至少造成**４５人**死亡、**约１０００人**受伤。

信息抽取

类型	地点	时间	死亡人数	受伤人数
爆炸	英国伦敦金融中心的地铁	当地时间7月7日清晨	45人	约1000人

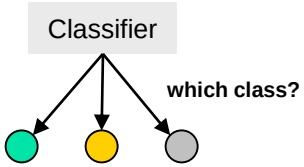
传统信息抽取的核心技术

字典匹配

Abraham Lincoln was born in Kentucky.

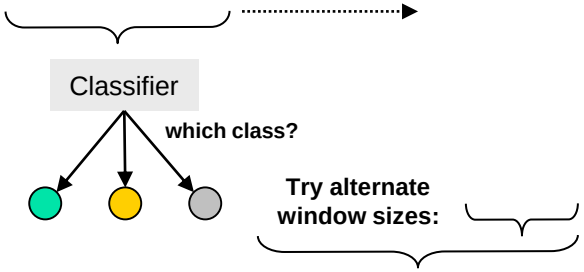
对每个分割候选进行分类

Abraham Lincoln was born in Kentucky.



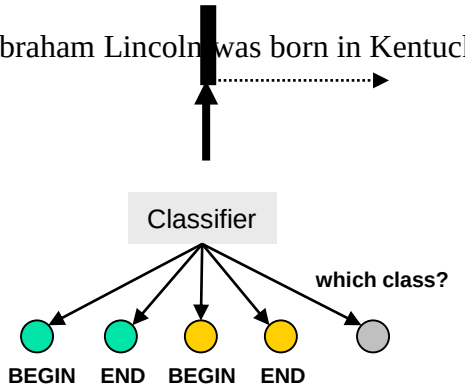
滑动窗口分类器

Abraham Lincoln was born in Kentucky.



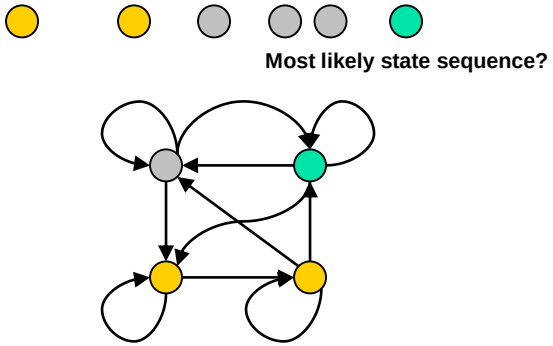
边界模型

Abraham Lincoln was born in Kentucky.



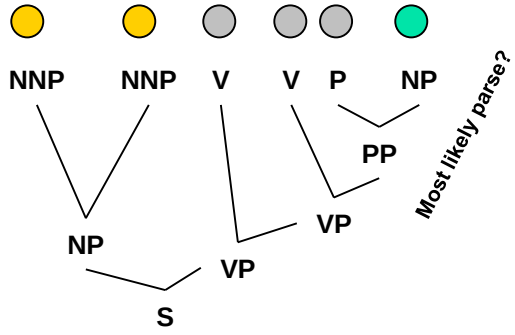
有限状态自动机

Abraham Lincoln was born in Kentucky.



上下文无关文法

Abraham Lincoln was born in Kentucky.



分水岭

Wikipedia

DBpedia

yago
select knowledge

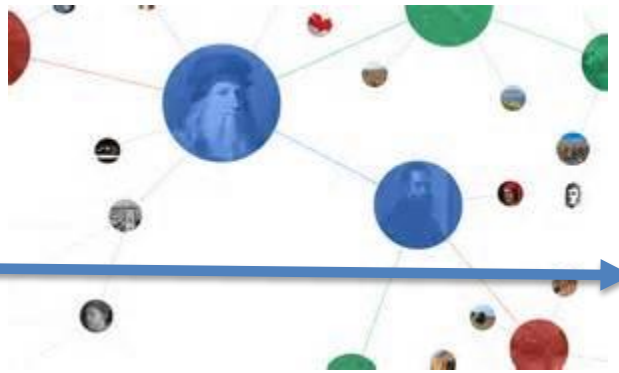
Cyc

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



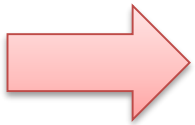
1985

1990

2005-2010

信息抽取目标的转变

从文本中抽取指定类型的实体、关系、事件等事实信息



从海量数据中发现实体相关的信息，并将其与现有知识库进行集成

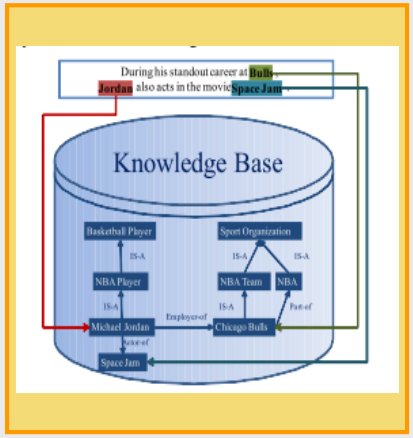
ACE

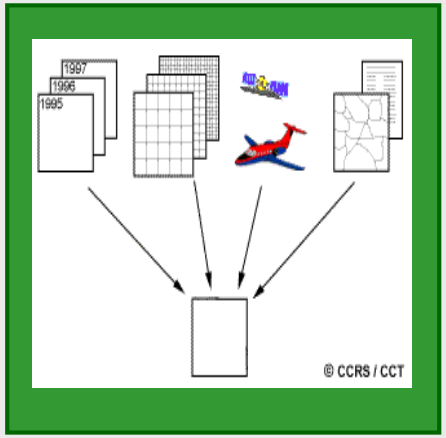
KBP

- 文本分析为核心 -->知识发现为核心
- 主要任务：抽取 --> 发现
- 数据源：文本 --> 海量数据
- 抽取对象：预先指定类型 -->Open Domain
- 与现有知识库的集成成为了新的核心任务

面向知识图谱的IE—核心模块

高价值信息检测 → **知识链接** → **开放抽取** → **验证集成**

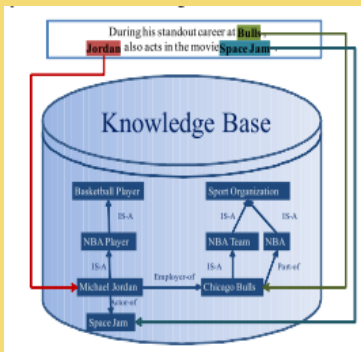




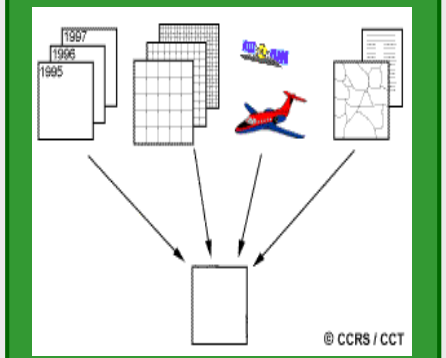
面向知识图谱的IE—核心模块

高价值信息检测

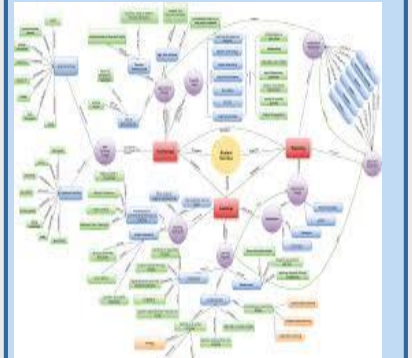
知识链接



开放抽取



验证集成



高价值信息检测

- 目的：对目标知识，找到容易抽取的数据块(Nugget)
  - 大大降低信息抽取的难度
  - 面向知识图谱的IE以知识为核心，目标是覆盖要抽取的知识，不需要覆盖所有文档
  - 数据规模导致无法覆盖数据的每一部分
- 高价值结构：Wikipedia Infobox，Web Table，List，...
- 高价值文本：匹配特定模板的文本，概念定义句...

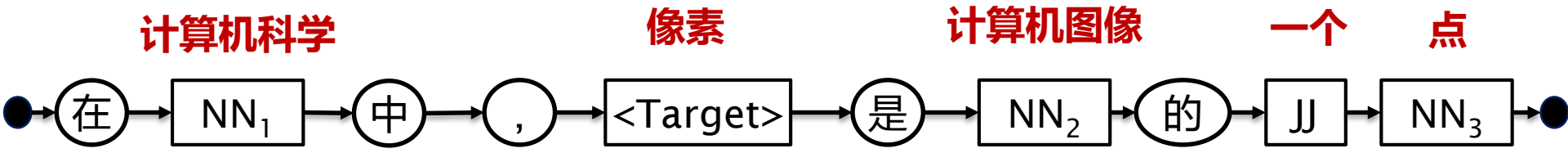
个人资料	
出生	1980年9月12日（34岁） 中国上海
国籍	中华人民共和国
登录身高	7英尺6英寸（2.29米）
登录体重	310英磅（141千克）

姚明身高2.29米

姚明父亲姚志源身高2.08米，姚明比他还高了21厘米

OntoLearn(Velardi et al., CL 2012)

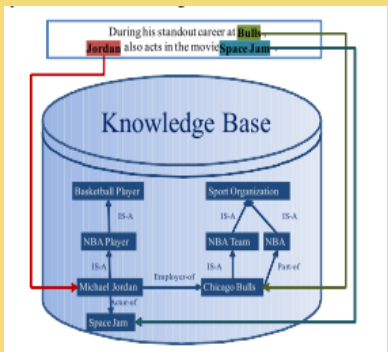
- 寻找概念的定义句来抽取IS-A关系
  - (像素，点)，（红色，颜色）,…
- 一个概念定义句必须包含四个域：
  - **DEFINIENDUM**: 被定义概念的声明，如*在计算机科学中，像素*
  - **DEFINITOR**: 用来引入定义的动词短语，如*指的是，被认为是*
  - **DEFINIENT**：概念的从属项，通常包含其上位词，如*一个点*
  - **REST**: 额外的从句，用于进一步明确或区分开类似概念的不同之处，如*计算机图像的*
- 构建了有限状态自动机来识别给定概念的所有定义句



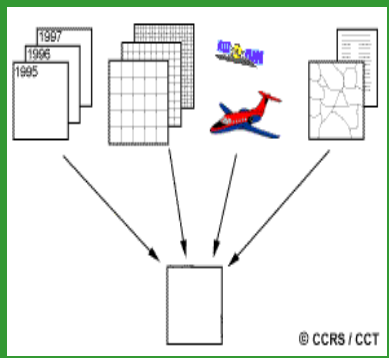
面向知识图谱的IE—核心模块

高价值信息检测

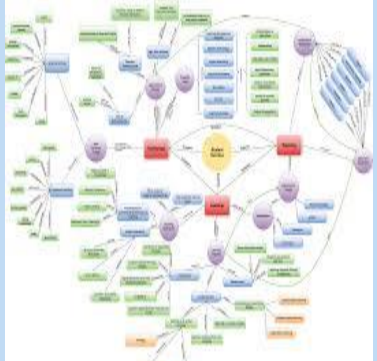
知识链接



开放抽取

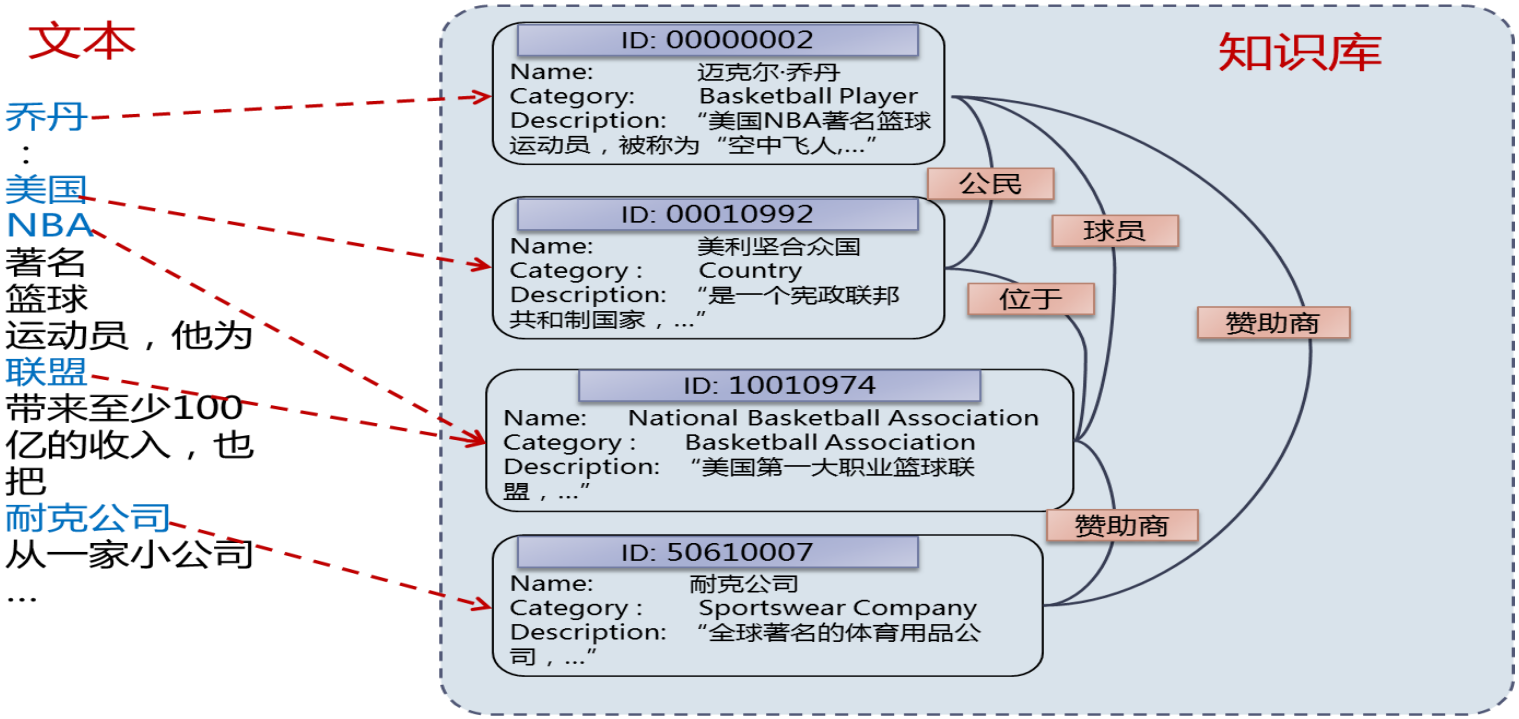


验证集成

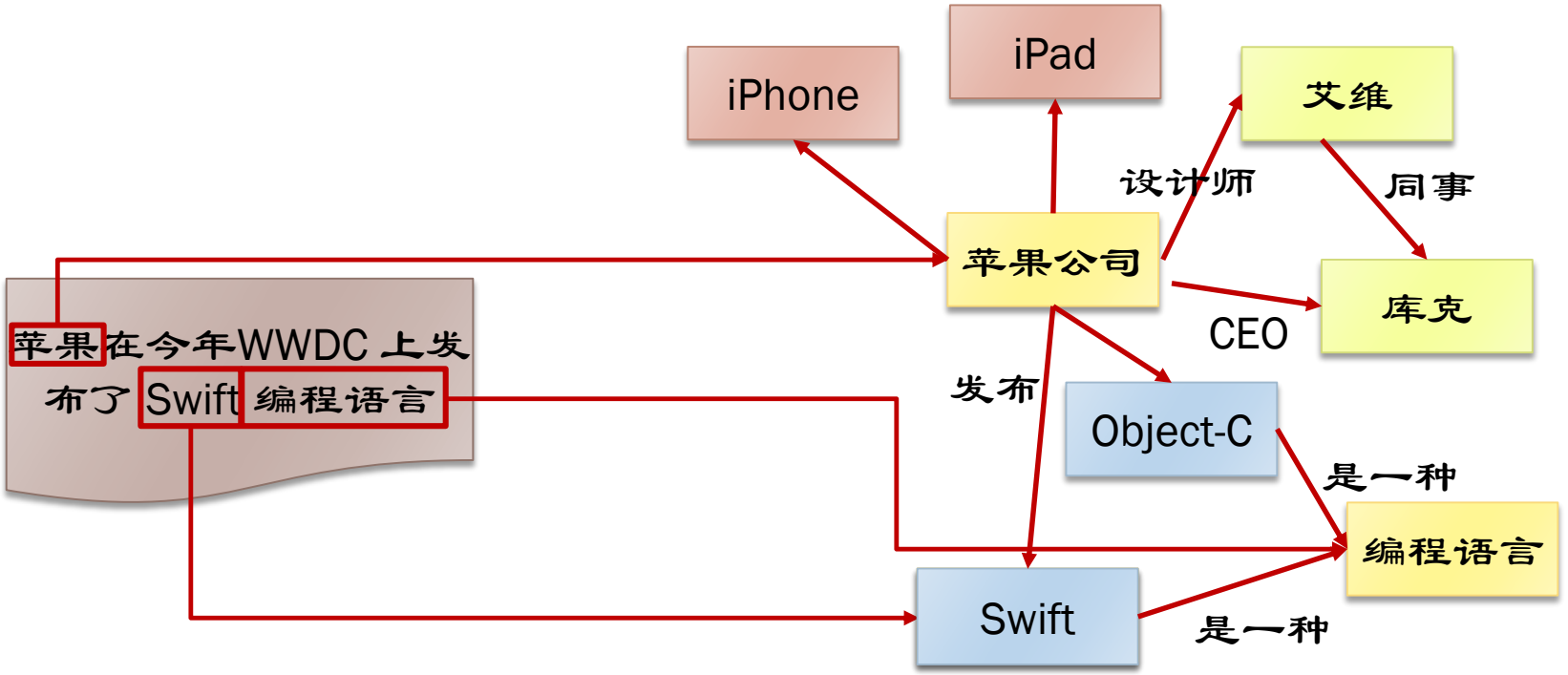


知识链接

- 将自然语言文本中的信息与知识库中的条目进行链接
- 作用：
  - 信息抽取的结果需要与现有知识图谱集成
  - 识别不同数据源中同一知识的冗余表示，处理表示的歧义性，提升信息抽取性能



实体链接系统



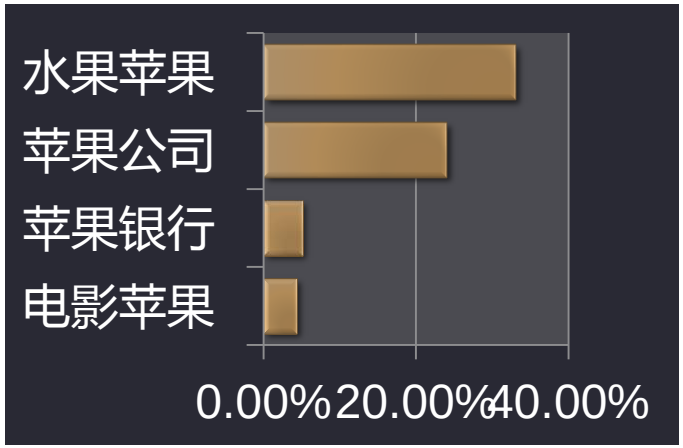
知识链接计算文本提及到实体之间的匹配程度，使用多方位的信息：

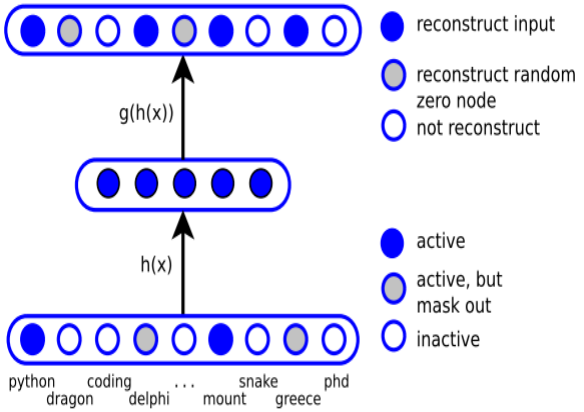
- 先验可能性（Popularity）
- 上下文相似度(Context Similarity)
- 文本的主题一致性(Coherence)

使用知识库提供构建模块：

- 名字-实体词典
- 实体关系，类别
- 实体的文本描述和关键特征
- 用来构建权重的参数

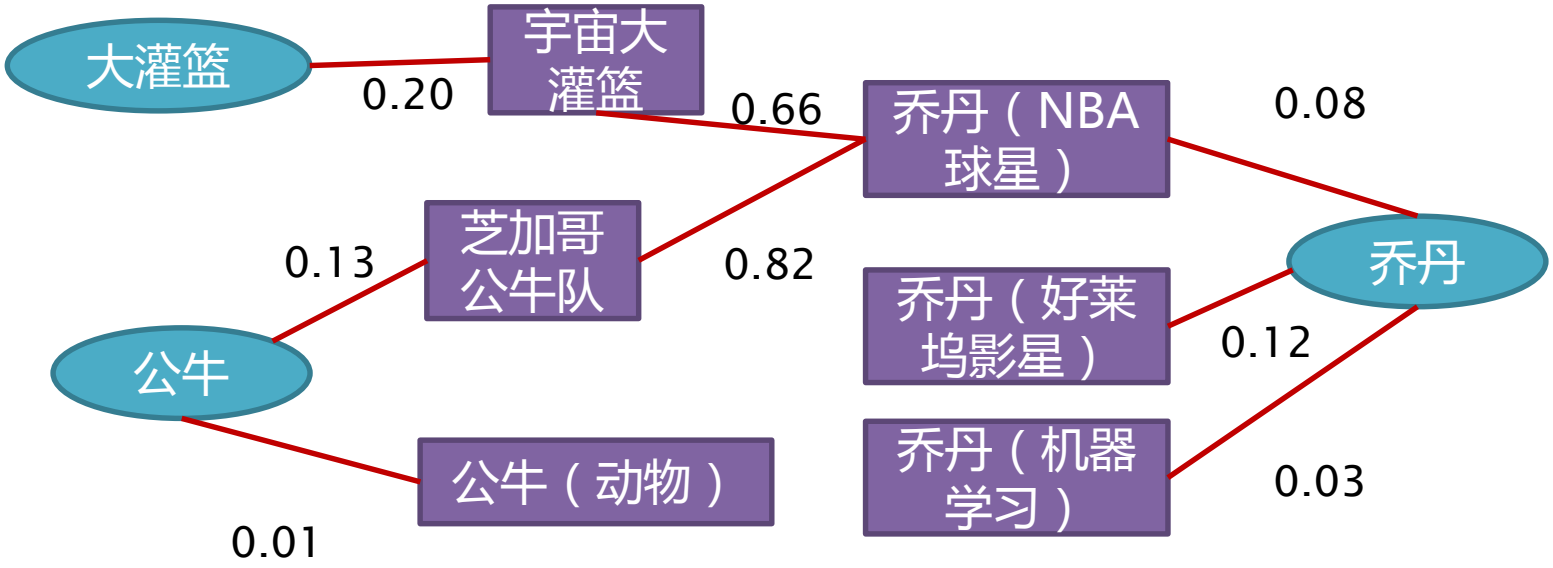
实体链接代表性方法—统计方法





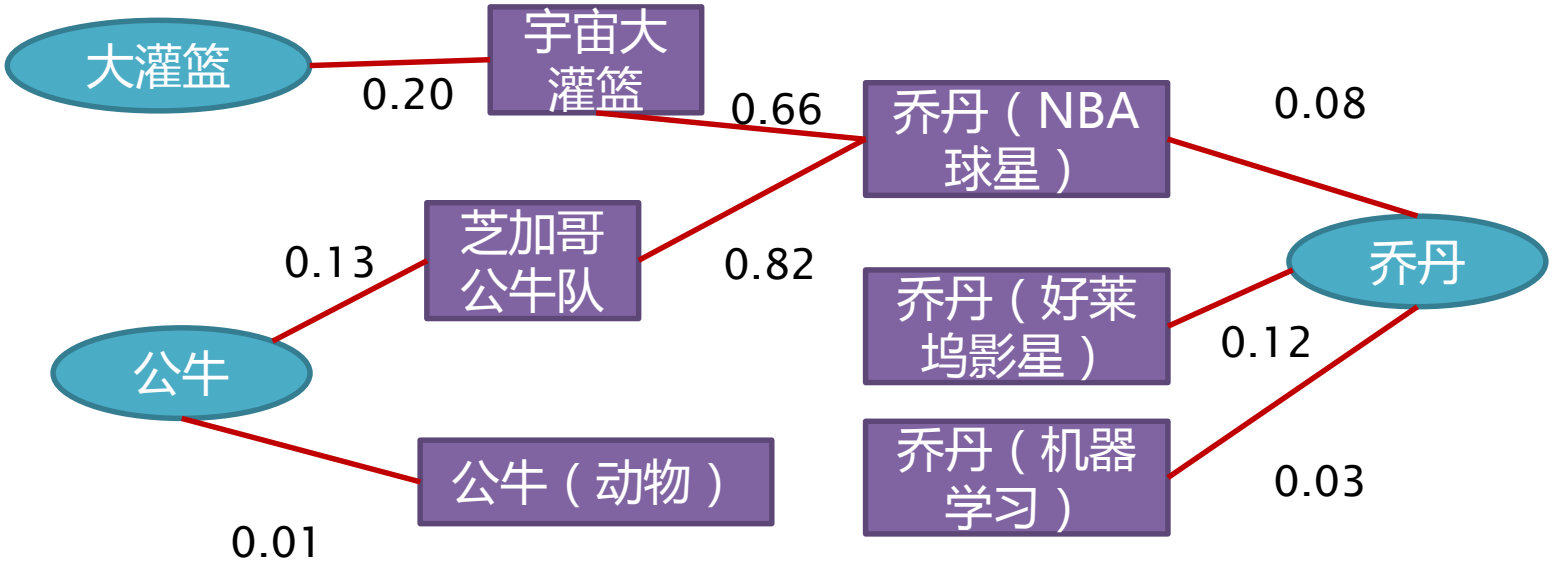
- 使用统计量来表示实体链接所需的知识
- 使用知识库和大规模语料库来估计上述统计量
- 设计统计模型综合多个不同的统计量来进行决策
  - 生成式模型（实体-提及模型 ACL 11, 实体-主题模型 EMNLP 12, …）
  - 深度学习模型（He et al., ACL 13, Sun et al., IJCAI’15,…）

实体链接代表方法—图方法



- 使用知识库中的知识来构建mention-entity graph
- 构建算法来计算最大似然链接结构
  - 同时考虑mention-entity的一致性和entity-entity之间的语义关联
  - 保证每一个mention指向且只指向一个目标实体

实体链接代表方法—图方法

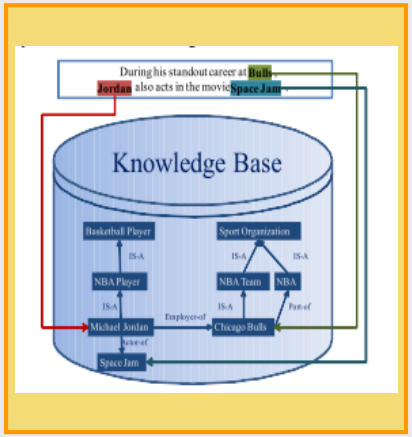


- 计算最大似然链接结构的算法
  - **寻找具有最大似然值的子图/最稠密子图**（Chakrabarti et al.: KDD’09，Hoffart et al., EMNLP’11,…）
  - **基于Graph Ranking寻找最大可能节点**（Han et al., SIGIR’11, Alhelbawy and Gaizauskas, ACL’14…）

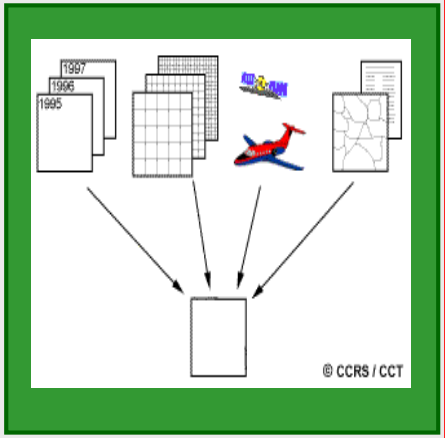
面向知识图谱的IE—核心模块

高价值信息检测

知识链接



开放抽取



验证集成

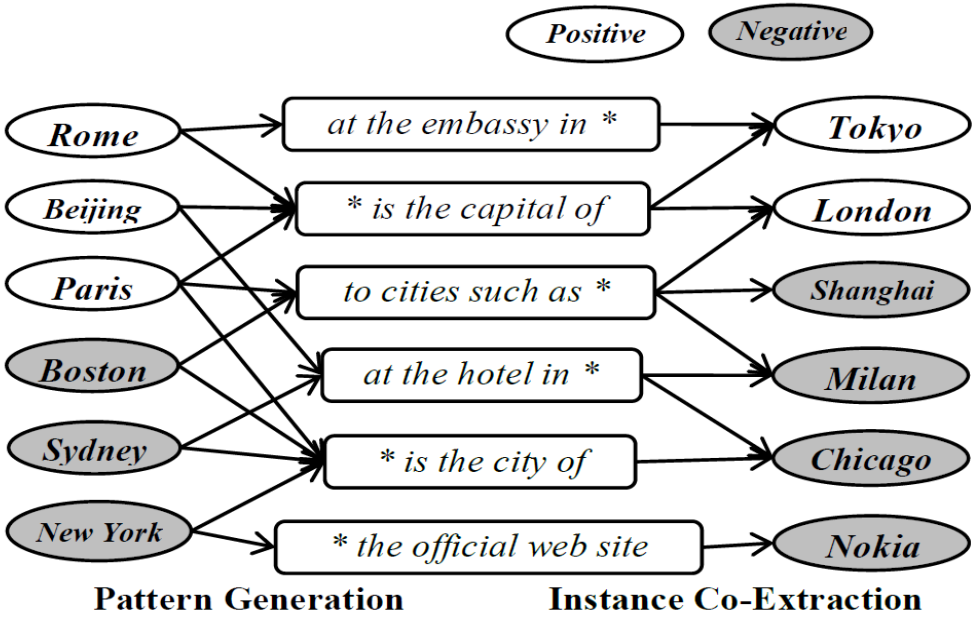
开放抽取

- 传统的**人工标注语料+机器学习算法**模式无法满足开放域开放语料下的信息抽取
  - 语料构建成本过高
  - 跨领域跨文本类别时抽取性能严重下降
  - 需要抽取的信息类别通常未预先指定

- 需要研究新的抽取方法
  - 按需抽取—Bootstrapping
  - 开放抽取—Open IE
  - 知识监督抽取—Distant Supervision
  - …(如知识库挖掘算法Path Ranking算法)

按需抽取：Bootstrapping

- Bootstrapping：模板生成->实例抽取->迭代直至收敛
- 语义漂移问题：迭代会引入噪音实例和噪音模板
  - 首都：*Rome* → 城市模板 "** is the city of*"
- (McIntosh et al. ACL 09)：同时扩展多个互斥类别
  - 同时扩展人物、地点、机构，一个实体只能属于一个类别
- COLING 14：引入负实例来限制语义漂移

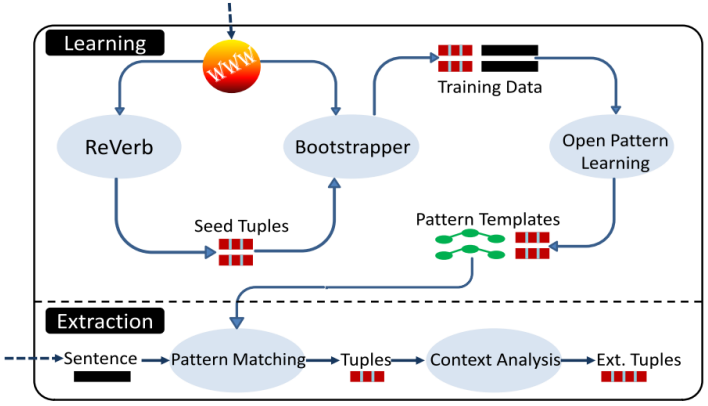


开放抽取：ReVerb

- 通过识别表达语义关系的短语来抽取实体之间的关系
  - *(华为，总部位于，深圳)，(华为，总部设置于，深圳)，(华为，将其总部建于，深圳)*
- 同时使用句法和统计数据来过滤抽取出来的三元组
  - 关系短语应当是一个以动词为核心的短语
  - 关系短语应当匹配多个不同实体对
- 优点：无需预先定义关系类别
- 缺点：语义没有归一化，同一关系有不同表示

$$V \mid VP \mid VW^*P$$
$$V = \text{verb particle? adv?}$$
$$W = (\text{noun} \mid \text{adj} \mid \text{adv} \mid \text{pron} \mid \text{det})$$
$$P = (\text{prep} \mid \text{particle} \mid \text{inf. marker})$$

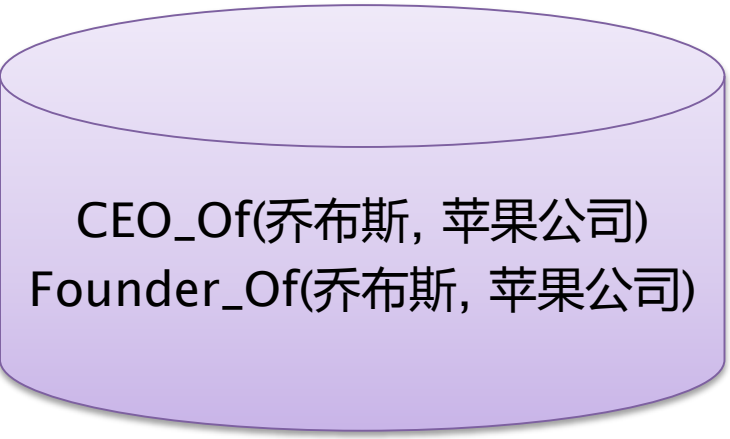
关系短语的句法结构约束



知识监督开放抽取-Distant Supervision

- 开放域信息抽取的一个主要问题是缺乏标注语料
- **Distant Supervision**: 使用知识库中的关系启发式的标注训练语料

知识库



标注训练语料

Relation Instance	Label
S1: 乔布斯是苹果公司的创始人之一	Founder-of, CEO-of
S2: 乔布斯回到了苹果公司	Founder-of, CEO-of

简单远距离监督方法(Mintz et al.，ACL09)

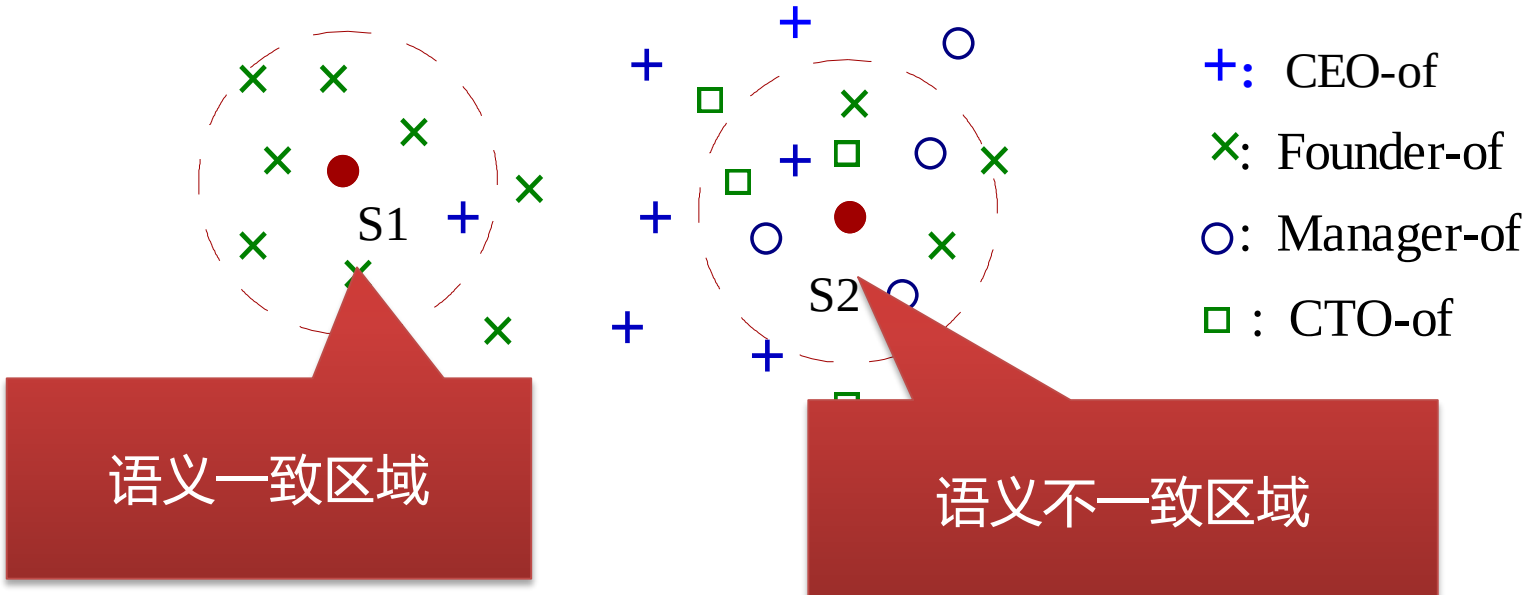
- **DS假设**: 每一个同时包含两个实体的句子都会表述这两个实体在知识库中的对应关系
- 基于上述假设标注所有句子作为训练语料
- 使用最大熵分类器来构建IE系统
- 最大的问题：噪音训练实例

噪音训练实例

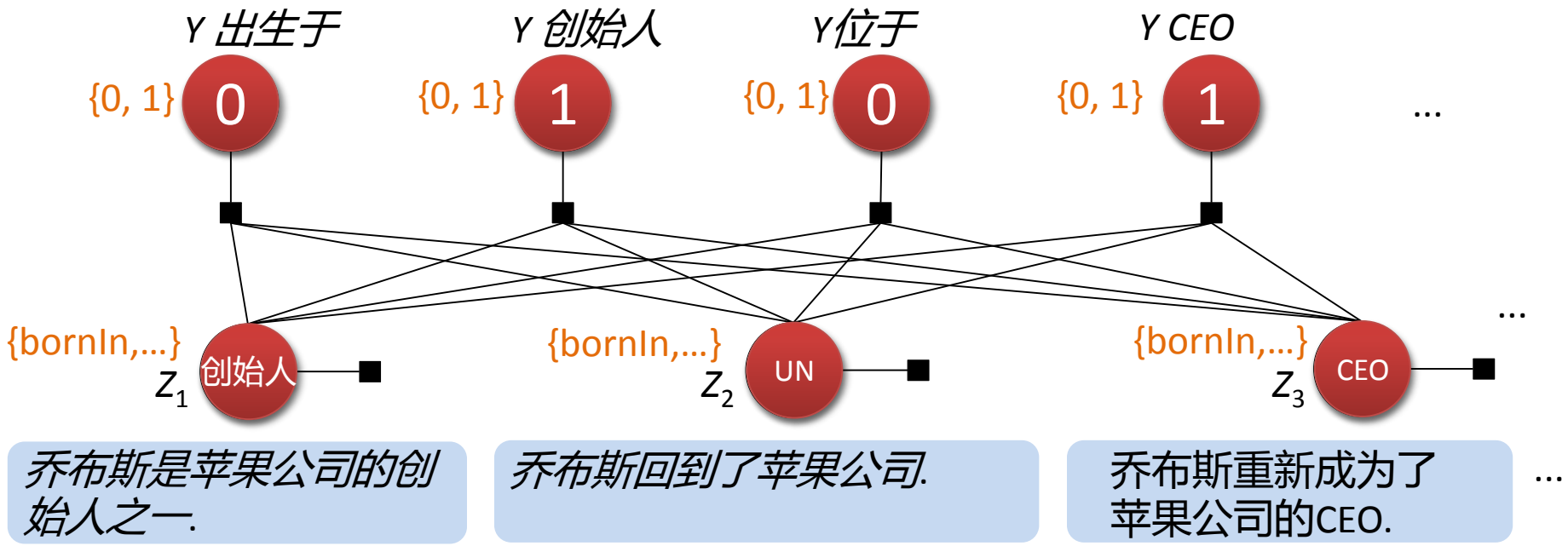
Relation Instance	Label	
S1:乔布斯是苹果公司的创始人之一	Founder-of	√
S1:乔布斯是苹果公司的创始人之一	CEO-of	X
S2:乔布斯回到了苹果公司	Founder-of	X
S2:乔布斯回到了苹果公司	CEO-of	X

基于噪音实例去除的DS方法

- 通过去除噪音实例来提升远距离监督方法的性能
- 假设：一个正确的训练实例会位于语义一致的区域，也就是其周边的实例应当都有相同一致的Label
  - 基于生成式模型的方法（Takamatsu et al. ACL 12）
  - 基于稀疏表示的方法（Han et al. ACL 14）



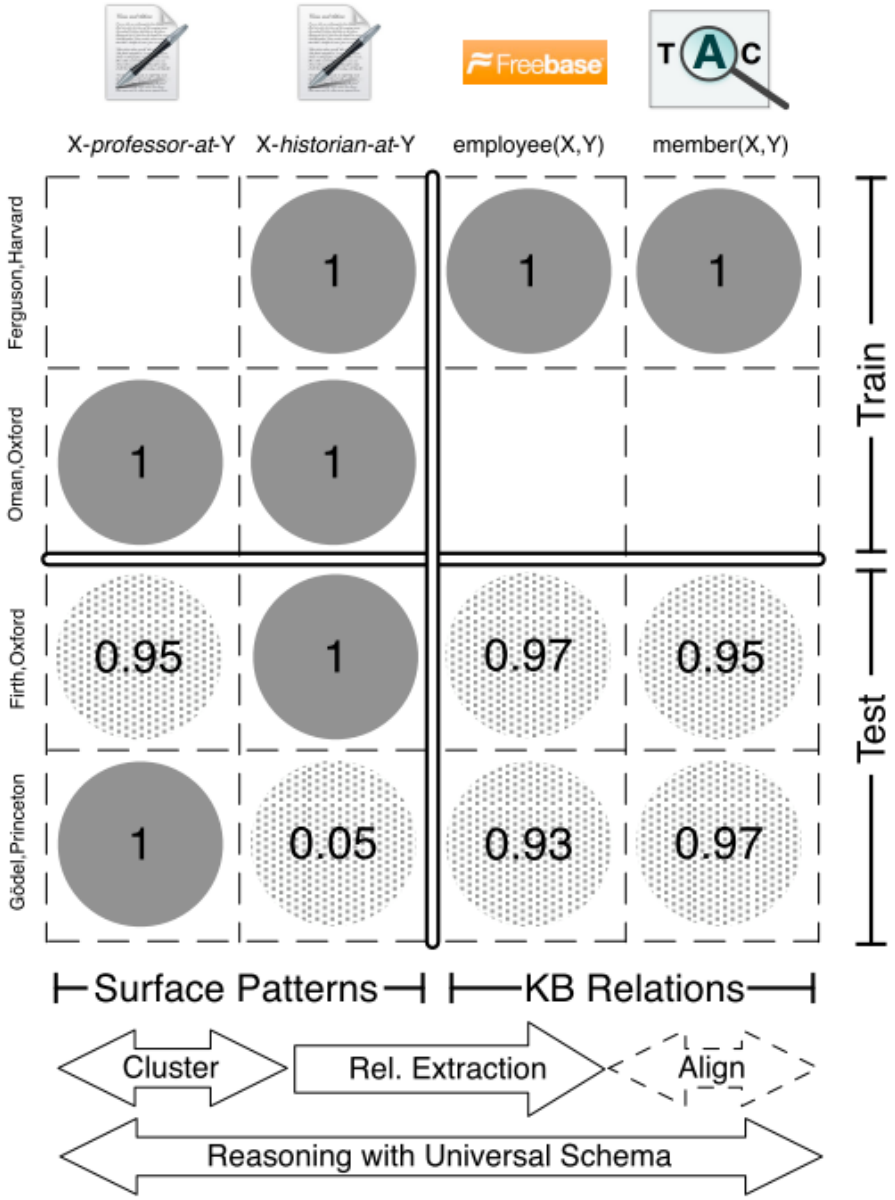
基于多实例学习的DS方法



- 一个实体对由一个句子集合表示
- AtLeastOne假设：只要实体对的一个句子具有特定关系，那么该实体对也就具有该关系
- 使用Factor Graph来表示多个变量之间的关系（Surdeanu et al. EMNLP 12, ... ）

基于协同推荐的DS方法

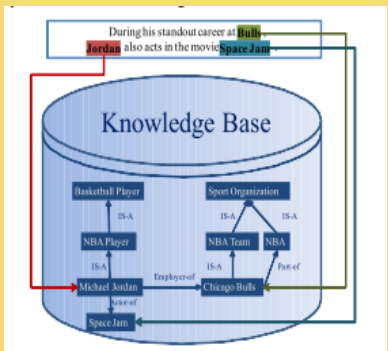
- 使用矩阵来表示实体对与Pattern，实体对与语义关系，Pattern与语义关系之间的关联
- 关系抽取任务被建模为矩阵填空问题
- 基于协同过滤推荐的方法(Riedel et al. NAACL 13)
- 基于Low-Rank矩阵分解的方法(Fan et al. ACL 14)



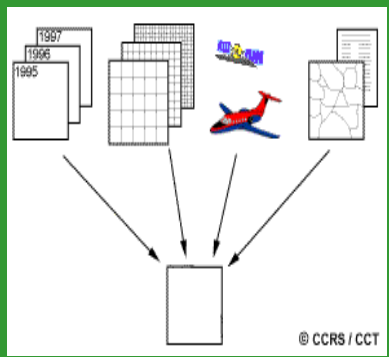
面向知识图谱的IE—核心模块

高价值信息检测

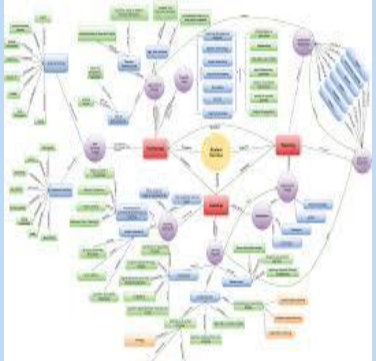
知识链接



开放抽取



验证集成



验证与集成

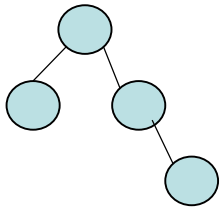
- 同一条知识可以从多个不同的数据源中抽取
  - 如何综合多个数据源中的证据来提升抽取的准确度和可靠性？
- 知识图谱构建不是一个静态的过程，需要及时更新动态知识并加入新知识
  - 如何判断新知识是否正确？
  - 如何判断新知识与已有知识是否一致？

Google‘s Knowledge Vault

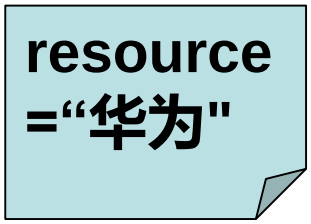
[L. Dong et al, SIGKDD 2014]

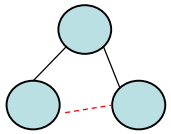
Sources:

Priors:



IT公司	总部
华为	深圳
中兴	深圳
...	...





文本

DOM Trees

HTML：表格

RDFa

Path Ranking Algorithm

使用最大熵模型来融合上述输出的证据

对每一个基本分类器，使用两个特征：

- 该抽取结果的数据源的数量（平方根）
- 抽取的置信度平均值

System	#	# > 0.7	# > 0.9	Frac. >0.9	AUC
Web 表格	9.4M	3.8M	0.59M	0.06	0.856
语义数据	140M	2.4M	0.25M	0.002	0.920
文本抽取	330M	20M	7.1M	0.02	0.867
DOM Tree	1200M	150M	94M	0.08	0.928
FUSED-EX.	1600M	160M	100M	0.06	0.927

开放抽取知识与知识库的集成

- Open IE可以抽取大量的知识，但是其关系使用自然语言短语表示，而不是知识库中的关系类别
  - 需要建立Relation Phrase到知识库关系类别之间的映射
  - *总部位于，总部设置于，将其总部建于* --> 总部位置
- 通常分为两个步骤：
  - 基于Relation Phrase之间的相似度进行聚类
  - 基于Relation Phrase聚类与知识库类别之间的相似度构建映射

开放抽取结果与知识库的集成

- Relation Phrase和关系类别都可以表示为与其共现的实体对集合
- 可以使用不同的集合相似度来进行计算
  - DIRT(Lin and Pantel, KDD01)， Soft Set inclusion (Nakashole, EMNLP 12)，Topic Model(Melamud et al., ACL 13), Jaccard (Dutta et al. WWW 15)

总部位于
(亚投行，北京)
(华为，深圳)
(IBM, 阿蒙克)
(联合国，纽约)
…

总部设置于
(亚投行，北京)
(联合国，纽约)
(红十字会，日内瓦)
(世界卫生组织, 日内瓦)
…

总结

面向知识图谱的IE

- **目标**：从海量数据中发现实体相关的信息，并将其与现有知识库集成
- **核心任务**：高价值信息检测、知识链接、开放抽取、集成与验证
- **有利**：可以使用多个大数据源，只使用最容易的抽取的信息源，已有大规模语义知识库
- **不利**：开放域, 无标注语料
- **底层技术:** 联合推理, 知识监督/无监督/自学习,多源信息集成，…

从传统IE到面向知识图谱的IE

- **文本分析为核心 → 知识获取为核心**
- **封闭信息类别 → 开放信息类别**
- **人工标注语料 → 自然标注语料/知识监督**
- 小规模文本→ 海量规模文本
- 规范文本（新闻） → 多源异质文本（UGC…）
- 规范文本的规律性 → 海量文本的冗余性和多样性
- 深度分析技术 → 基于冗余的浅层分析技术
- 准确率 → 召回率+时间效率
- **…**

请大家批评和指导