

中文知识图谱：体系、获取与服务

中国科学院自动化研究所
模式识别国家重点实验室
赵军 刘康

什么是知识图谱

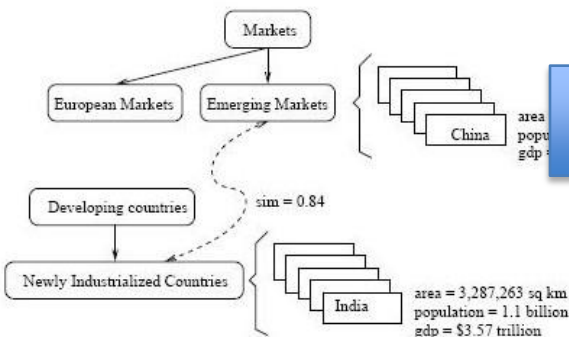
- The Knowledge Graph is a system that understands facts about people, places and things and how these entities are all connected.
- 知识图谱本质上是一种语义网络。其结点代表实体(entity)或者概念(concept)，边代表实体/概念之间的各种语义关系。



smartelearningmethods.com

ProBase

2,653,873概念



搜狗知立方

百度知心



出生：1873-02-23 广东新会
逝世：1929-01-19
妻子：李惠仙（正室）/王桂章（续聘）
人物关系：梁启超（父亲）/梁思达（儿子）/梁思忠（儿子）/
梁思懿（女儿）/梁思和（儿子）
个人标签：享受工作的同时享受生活

著作

梁启超

中国近三百年
思潮

中国历史研究

新大陆游记

李鸿章传

梁代文学研究



林依晨 [百度百科](#)

依晨，中国台湾女演员、歌手。2000年参加台北捷运举办的“第一届捷运超美少女比赛”夺冠。2001年考入国立政治大学新闻系，其后陆续参与MV与广告拍摄。2002年正式出道，荧屏处女作是《[薰衣草](#)》。>>> [查看完整资料](#)>>



实 (天家)



工作上如有任何問題，請洽：csa0312@yahoo.com.tw 或 090-2-2720811 周子誠 謝



7.3 概念



歌曲(40) 专辑(3)	热度	偏好	下载
这首歌可能不会让你听见那声音*		▷	⬇



5 普通通知		▶	主
5 通知100%保障 價格附保證...		▶	主

同系列詳情: **2,714,074**人 眾和友站: **4,973,721**



你，<是你俩的快偶就来推一下>  

欢历年“天天马来拜年”~青年快乐！ 

你，看看你在亲戚的地位。

点击: 6万	回复: 2912
点击: 430	回复: 21
点击: 2342	回复: 173

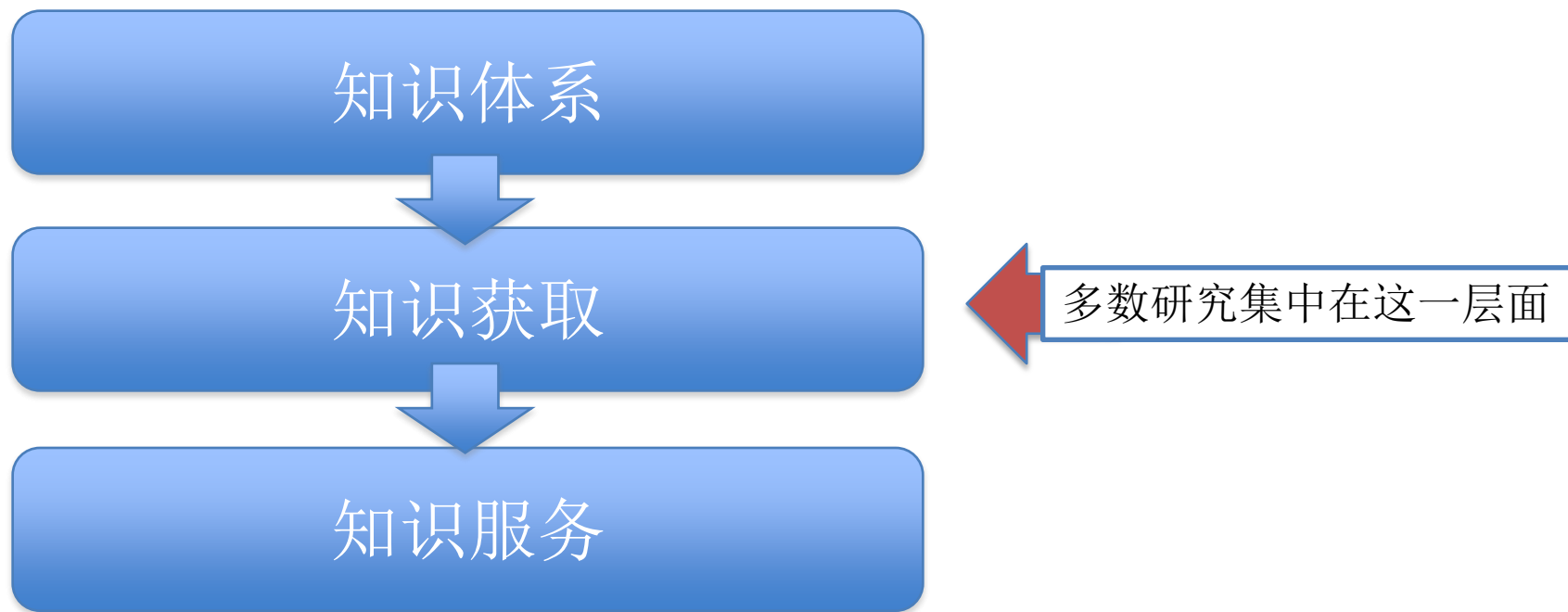


		
天外飞仙 全36集	东方圣朝叶 全17集	我的宝贝在囡 共2季

其他已有的知识库

Name	language	Year	Construction	Types
Hownet	Chinese	2000	Manual	Common Sense Knowledge
Wordnet	English	1985		
CYC		1984		
DBpedia		2007	Automatic	Common Sense Knowledge + Factual Knowledge
Yago		2007		
Freebase		2007	Crowding Sourcing	

知识工程：三个层面问题



知识体系

知识体系：几个术语

- **Ontology vs. Knowledge Base**

- **Ontology**: 共享概念化的规范，涉及**概念**、**关系**和**公理**三个要素
- **Knowledge Base**: 服从于ontology 控制的知识单元的载体
- **Ontology**是蛋糕的模具，**Knowledge Base**是蛋糕

- **Formal Ontology vs. Lightweight Ontology**

- **Formal Ontology**: 大量使用公理
- **Lightweight Ontology**: 不用或很少使用公理

知识体系：几个术语(cont)

- 关系

- 层级关系Hypernym-Hyponym

- Is-a (Kind-of)
 - Part-Whole

- 非层级关系

- Thematic roles 论旨角色
 - Possession 领属
 - Attribute 属性
 - Casuality 因果
 -

知识体系：三种组织形式

- 层级分类法

- **Ontology (狭义)**

- 树状结构，不同层节点之间具有严格的IsA关系
 - Human activities -> leisure activities -> sports -> golf
 - 优点：因为概念关系单一，方便于知识推理
 - 缺点：无法表示概念关系的多样性

- **Taxonomy**

- 树状结构，上下位节点之间非严格的IsA关系，而是Hypernym-Hyponym关系
 - Places -> Milky Way Galaxy -> Solar Systems -> Sol -> Inner Planets -> Earth -> North America -> United States -> California -> Cupertino.
 - 优点：可以表示比较丰富的概念关系
 - 缺点：给推理带来困难，无法避免概念冗余

知识体系：三种组织形式(Cont.)

- 标签分类法

- **Folksonomy**

- 网络用户自发性定义的平面的、非层级的标签分类
 - 优点：灵活，可以表达更为丰富的概念关系
 - 缺点
 - 缺乏层次性，难以揭示复杂的关系
 - 自定义的标签缺乏语义精确性，标签缺乏组织与关联
 - 给推理带来很大的困难

目前网络知识资源的组织形式

- 目前网络知识资源（Wikipedia、百度百科、互动百科等）多是采用Taxonomy与Folksonomy相结合的组织形式，以Taxonomy为主。

人物 动漫人物 歌手 运动员 古代传记 演员 >> 文化 考古 网络用语 世界名著 诗词 神话 >>

技术 土木工程 移动通信 CPU MP3 电脑病毒 >> 历史 侏罗纪 清朝 明朝 洋务运动 先秦 >>

艺术 纪念碑 戏剧 音乐 绘画 雕塑 建筑 >> 生活 影视 动漫 游戏 服饰 美容 烹饪 >>

地理 大陆架 国家 地质 岛屿 山脉 河流 >> 社会 外交 军事 民俗 交通 法律 企业 >>

体育 冰雪运动 极限运动 电子竞技 篮球 足球 >> 自然 地震 气象 天文 花卉 恐龙 细菌 >>

科学 生物 遗传学 医学 化学 物理 数学 >> 经济 投资 保险 银行 期货 基金 股票 >>

航空母舰

编辑词条

开放分类：世界军事 军事 技术 武器 水面舰艇部队

图片 (4+) 讨论 知识模块



美国尼米兹号航空母舰

航空母舰(Aircraft Carrier)，简称“航母”、“空母”，前苏联称之为“载机巡洋舰”，是一种可以提供军用飞机起飞和降落的军舰。中文“航空母舰”一词来自日文汉字。航空母舰一般总是一支航空母舰舰队中的核心舰船，有时还作为航母舰队的旗舰。舰队中的其它船只为它提供保护和供给。依靠航空母舰，一个国家可以在远离其国土的地方、不依靠当地的机场情况施加军事压力和进行作战。

编辑摘要

相关百科观察

更多百科观察

关注新闻热点，解读背景知识

印度首艘国产航母下水不等于海试：印度媒体高调报道称，完全在印度国内制造的第一艘航母**维克兰特号航空母舰**将于8月12日在科钦船厂下水，这将是历史性的一天。首艘国产航母下水后，印度将成为继美、俄、英、法国之后少数能自行建造航母的国家。不过，美国“防务新闻网”报道说“维克兰特”号航母的建造工作仅完成30%，实际部署时间很可能推迟到2020年。更新时间：2013-08-14 08:45:16



目前的Folksonomy存在的问题

- Folksonomy的标签不能覆盖的所有的关系
 - 无论是开放分类标签
 - 还是Infobox属性标签
- 这些开放式类别标签存在冗余、不规范的问题，标签之间也缺乏关联
 - 1980年、购房、房产、房地产.....

全部	含有开放分类 (Folksonomy)的页 面数比例
互动百科	70.19%
百度百科	64.38%

目前的Taxonomy存在的问题

- 不同的知识资源采用不同的Taxonomy

		人物 政治人物 经济人物 文化人物 娱乐人物 体育人物 科学家 话题人物 热点人物	科学 自然科学 社会科学 应用科学 人文社科 科学奖项 交叉学科 科技新品	技术 通信技术 电信技术 计算机技术 互联网 航空航天 能源 科技产品 高新技术
		自然 植物 动物 微生物 宇宙天文 自然资源 自然现象 自然遗产 环境保护	历史 各国历史 各年代历史 区域历史 历史学 专门史 历史	文化 文学 哲学 神话传说 文化场馆 流行文化 文化遗产
人物	动漫人物 歌手			 世界各地 亚洲 - 非洲 - 大洋洲 - 北美洲 - 南美洲 - 欧洲 - 南极洲
技术	土木工程 移动	社会 政治 军事 教育 职业 法律 民族 宗教 社会问题 荣誉 公益	 生活、艺术与文化 收藏 - 饮食 - 服装 - 交通 - 体育 - 娱乐 - 旅游 - 游戏 - 嗜好 - 工具 - 音乐 - 舞蹈 - 电影 - 戏剧 - 电视 - 摄影 - 绘画 - 雕塑 - 手工艺 - 家庭 - 文明 - 文物 - 节日 - 虚构 - 符号 - 次文化 - 动画 - 漫画	
艺术	纪念碑 戏剧			 人文与社会科学 哲学 - 文学 - 艺术 - 语言学 - 历史学 - 地理学 - 心理学 - 社会学 - 政治学 - 法学 - 军事学 - 传播学 - 新闻学 - 考古学 - 人类学 - 民族学 - 教育学 - 图书资讯科学 - 经济学 - 人口学 - 家政学 - 管理学 - 性学
地理	大陆架 国家	艺术 雕塑 绘画 音乐 舞蹈 戏剧 曲艺 影视 摄影 建筑 艺术家	 中华文化 中国历史 - 中国神话 - 中国音乐 - 戏曲曲艺 - 中华民俗 - 中国文学 - 中文古典典籍 - 武术 - 中医 - 国画 - 书法 - 佛教 - 道教 - 生肖	 自然与自然科学 生物 - 动物 - 植物 - 气象 - 季节 - 化学元素 - 矿物 - 地理 - 数学 - 物理学 - 力学 - 化学 - 天文学 - 星座 - 地球科学 - 地质学 - 生物学 - 医学 - 药学 - 农学 - 资讯科学 - 系统科学 - 密码学
体育	冰雪运动 极限			 工程、技术与应用科学 交通 - 建筑学 - 土木工程 - 电气工程 - 计算机科学 - 机械工程 - 能源科学 - 测绘学 - 航空航天 - 矿业 - 冶金学 - 印刷 - 化学工程 - 水利工程 - 通信技术 - 生物工程 - 材料科学 - 环境科学
科学	生物 遗传学	生活 健康 时尚 礼仪	 社会 文化 - 历史 - 语言 - 宗教 - 教育 - 家庭 - 组织 - 族群 - 经济 - 政治 - 政府 - 国家 - 传统 - 产业 - 媒体 - 体育 - 安全 - 法律 - 犯罪 - 奖励 - 城市	
				 宗教及信仰 各国宗教 - 宗教人物 - 宗教史 - 宗教建筑 - 宗教节日 - 宗教哲学 - 宗教场所 - 宗教学 - 宗教组织 - 神祇 - 神话 - 神学

个人概况

中文名：李娜
国籍：中国
民族：汉
出生地：湖北武汉
出生日期：1982年2月26日
身高：172cm
体重：65kg

个人背景

毕业院校：[华中科技大学](#)
运动项目：网球
所属运动队：[中国网球队](#)
专业特点：正手凶狠，灵活，底线好，力量大

个人贡献

重要事件：亚洲首位大满贯单打冠军得主

其他信息

主要奖项：[WTA](#)单打冠军头衔：7、[WTA](#)双打冠军头衔：2、[ITF](#)单打冠军头衔：19、[ITF](#)双打冠军头衔：16、2011法国网球公开赛女单冠军
启蒙教练：[余丽桥](#)
训练地：[马拉喀什](#)，[什切青](#)，[内乌姆](#)
教练：[卡洛斯·罗德里格斯](#)
丈夫：[姜山](#)

类别属性定义不统一

互动百科

中文名：	李娜	籍贯：	武汉市
性别：	女	民族：	汉族
国籍：	中国	出生年月：	1982年2月26日
星座：	双鱼座	职业：	运动员 女子网球选手
毕业院校：	华中科技大学	身高：	172厘米

Solution: Ontology Matching

- 建立体系间的Alignment

- 挖掘概念之间SameAs关系

- 评测：Ontology Alignment Evaluation Initiative

- 2004-2013

- Benchmarks (bibliographic references), Web directories, Anatomy (biomedical)

- 关键：概念之间的相似度计算

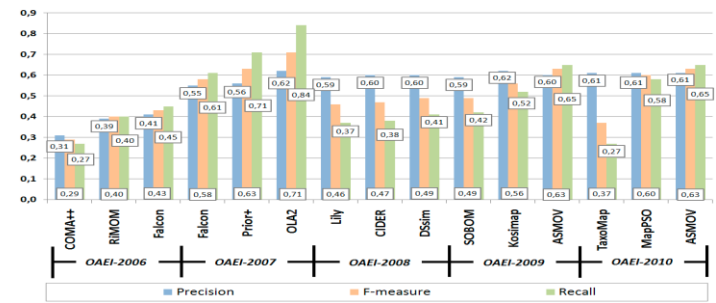
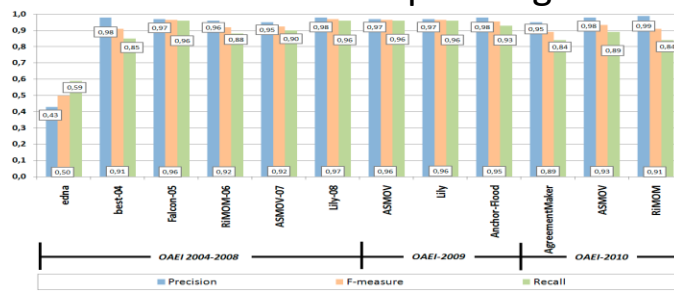
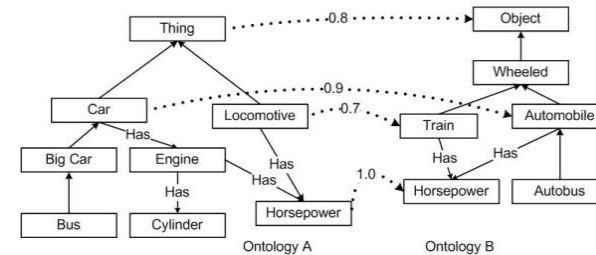
- 挑战

- Large-scale ontology matching and evaluation

- Matching with background knowledge (Increase recall but hurt precision)

- Multiple matchers and selection (Global Alignment)

- Incorporating social information





Solution: 建立框



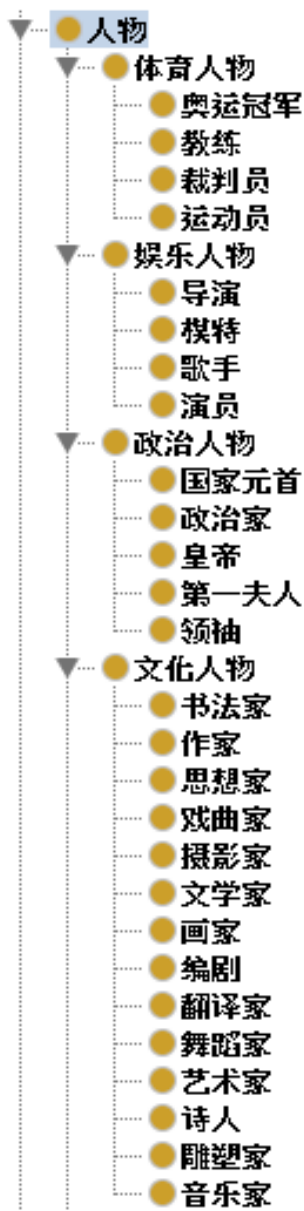
- Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe
- Embedded non-text objects: AudioObject, ImageObject, VideoObject
- Event
- Health and medical types: notes on the health and medical types
- Organization
- Person
- Place, LocalBusiness, Restaurant ...
- Product, Offer, AggregateOffer
- Review, AggregateRating

— Schema.org的翻译和扩展

- 体系覆盖度不足，局限于英文
- 细致化不足

— 百科知识描述体系的制订

- 中国大百科全书出版社



ogy

知识获取

文本信息结构

- 结构化数据 (Infobox)
 - 置信度高
 - 规模小
 - 缺乏个性化的属性信息
- 半结构化数据
 - 置信度较高
 - 规模较大
 - 个性化的信息
 - 形式多样
 - 含有噪声
- 纯文本
 - 置信度低
 - 复杂多样
 - 规模大

中文名:	姚明	专业特点:	20英尺外精确跳投
外文名:	Yao Ming	主要奖项:	NBA全明星赛(7次)
别名:	小巨人 移动长城		ESPN全球最有潜力运动员奖(2000)
国籍:	中国		劳伦斯世界最佳新秀奖(2003)
民族:	汉族		中国篮球杰出贡献奖
出生地:	上海	重要事件:	专题影片《姚明年》发行
出生日期:	1980年9月12日	祖籍:	江苏苏州吴江震泽
身高:	2.23米 (7.32英尺)	位置:	中锋
体重:	140.6kg	鞋码:	18码
运动项目:	篮球	自传:	《我的世界我的梦》
所属运动队:	NBA火箭队	生涯最高分:	41分

1984:《一个和八个》摄影	个人简介
1984:《黄土地》摄影	姓名: 姚明
1986:《大阅兵》摄影	祖籍: 江苏苏州吴江震泽
1987:《老井》主演	出生地: 上海市徐汇区
1987:《红高粱》导演	出生医院: 上海交通大学附属第六人民医院
1989:《古今大战秦俑情》	小学: 上海市高安路第一小学
1989:《代号美洲豹》	初中: 上海市第二中学
1990:《菊豆》导演	大学: 上海交通大学
1991:《大红灯笼高高挂》	星座: 处女座
	曾效力球队: CBA上海东方大鲨鱼 、 NBA休斯敦火箭
	前任主教练: 里克·阿德尔曼
	休斯敦火箭队球衣号码: 11
	国家队球衣号码: 13
	2008奥运会号码: 13
	手掌: 19厘米
	净身高: 2.26米
	臂展: 7英尺4.75英寸 (2.25米)
	站立摸高: 9英尺7英寸 (2.91米)
	原地弹跳: 19英寸 (约48厘米)

张艺谋, [陕西省西安人](#), [中国电影导演](#), [北京奥运会开幕式总导演](#)。在[电影学院](#)学的是[摄影专业](#)。张艺谋是中国在国际影坛最具影响力的导演, 曾多次荣获国际电影节大奖并成功执导北京奥运会开幕式。其代表作《[红高粱](#)》被认为是中国电影走向世界的新开始, 《[英雄](#)》则是开启了中国电影的大片时代。

抽取方法

- 结构化与半结构化文本信息（利用网页结构）
 - 信息块的识别（Record Identification）
 - 模板的学习（Pattern Learning）
 - 属性值的抽取（Attribute Value Extraction）

基本信息

出版社：作家出版社；第1版（2007年1月1日）

平装: 284页

语种： 简体中文

开本: 32

ISBN: 9787309054111 基本信息

条形码: 9 787309 781000 >
出版社: 作家出版社; 第1版 (2007年7月1日)

商品尺寸: 平装: 346页

商品重印: 开本: 16开

品牌·作家 ISBN: 9787506339391

ASIN: B001 条形码: 9787506339391

用白评分: 商品尺寸: 22.8 x 15 x 2.4 cm

商品重量: 340 g

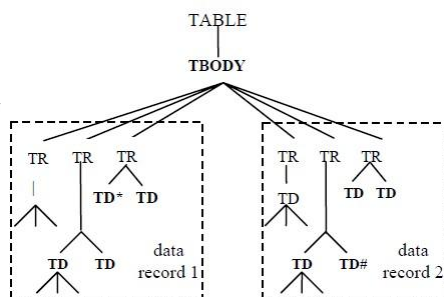
第21卷

第28位

用户评分: (19 条商品评论)

亚马逊热销商品排名: 图书商品排名第123, 470名 (亚马逊图书商品排名排行第...)

您想告诉我们您发现了更低的价格?



The diagram illustrates the mapping of attributes from an Infobox to a structured data table. On the left, a blue box labeled 'Infobox中的属性名' (Attributes in Infobox) lists: 身高 (Height), 语言 (Language), 国籍 (Nationality), and 体重 (Weight). On the right, a light purple box represents the data table with rows for: 身高: 168厘米 (Height: 168cm), 地区: 大陆 (Region: Mainland), 国籍: 中国 (Nationality: China), 语言: 普通话 (Language: Mandarin), 特长: 表演, 舞蹈 (Specialty: Performance, Dance), 曾就职于: 吉林市歌舞团 (Formerly employed at: Jilin City Dance Troupe), 担任: 舞蹈演员 (Position: Dance Performer), 职业: 影视演员 (Occupation: Film and Television Actor), and 毕业院校: 北京电影学院 (Alma Mater: Beijing Film Academy). Arrows show the mapping: '身高' maps to '身高: 168厘米'; '语言' maps to '语言: 普通话'; '国籍' maps to '国籍: 中国'; and '体重' maps to '地区: 大陆'.

身高:	168厘米
地区:	大陆
国籍:	中国
语言:	普通话
特长:	表演, 舞蹈
曾就职于:	吉林市歌舞团
担任:	舞蹈演员
职业:	影视演员
毕业院校:	北京电影学院

第一次的演戏: 《奋斗》

电视剧作品

电视剧《**奋斗**》导演: 赵宝刚 饰演: 露露

电视剧《**别无选择**》导演: 曹东 饰演: 刘英

电视剧《**生死桥**》导演: 田沁鑫 饰演: 龙小姐

Bootstrapping

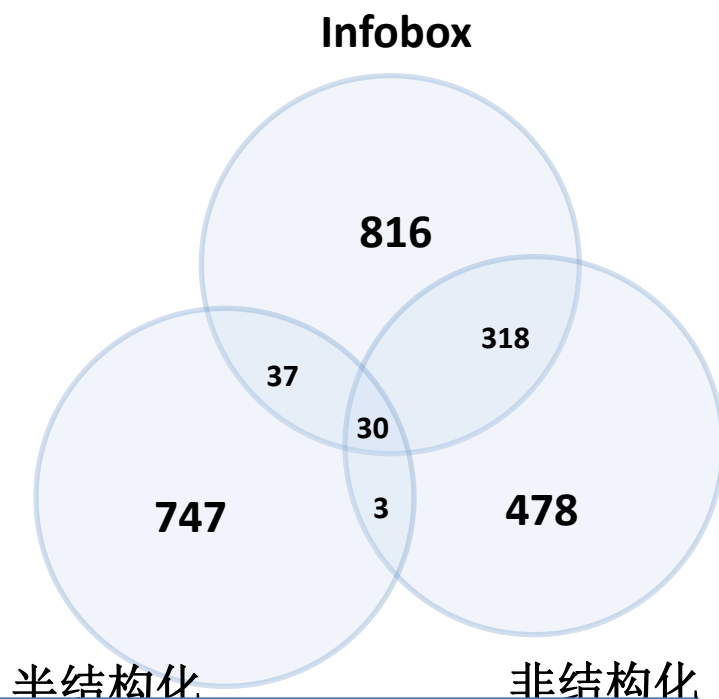
抽取方法（续）

- 相对于工业界，学术界更加侧重于从纯文本中抽取实体知识
 - 传统关系抽取
 - 给定关系类别和训练语料
 - 开放式关系抽取
 - 已有关系类别，缺乏训练语料
 - Distant Supervision
 - 完全开放式
 - 从句法到语义

结构化 vs. 半结构化 vs. 非结构化

- 随机抽取100篇百科文档（共5类）
 - 对于其中三部分都包含的网页进行了统计

	人物	地理	电影	动物	图书
InfoBox	87	182	260	183	104
非结构化	79	147	109	107	36
半结构化	119	96	327	129	76
OverLap Info vs. 非	62	101	87	72	26
OverLap Info vs. 半	11	11	21	7	17
OverLap 非 vs. 半	8	7	4	4	10
三方OverLap	7	7	3	4	9



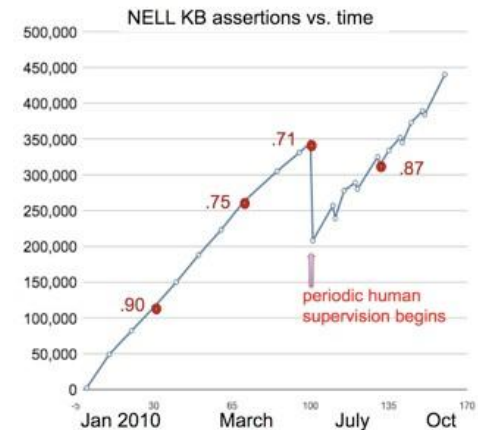
半结构化和非结构化文本的实体关系抽取非常重要
非结构化文本的实体关系抽取：对于文本进行结构化
半结构化文本实体关系抽取：抽取个性化的实体属性

CMU: NELL(Never-Ending Language Learning)

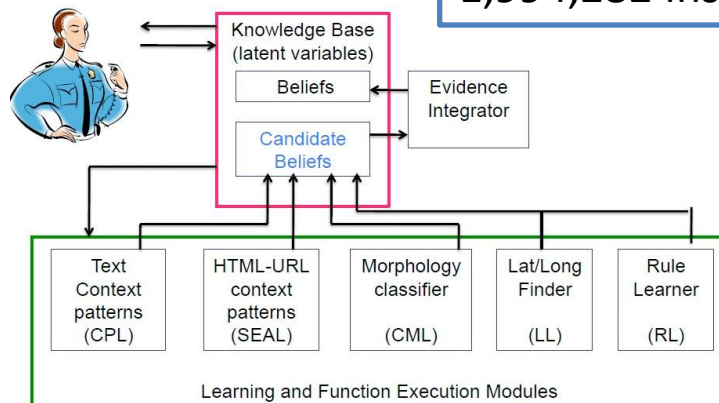
Read the Web

Research Project at Carnegie Mellon University

- Input
 - Initial ontology
 - 500 million web pages
- Aim
 - Extract new instances of categories and relations
 - Learn to read better than yesterday



1,994,282 Instances of 874 different categories and relations



Recently-Learned Facts [twitter](#)

Refresh

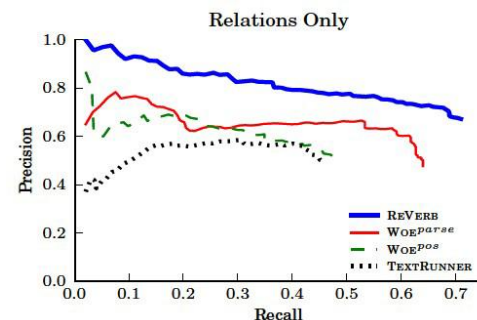
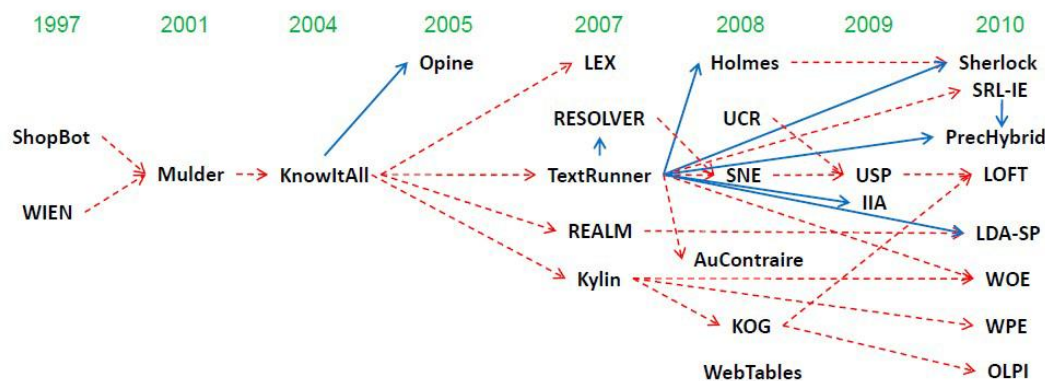
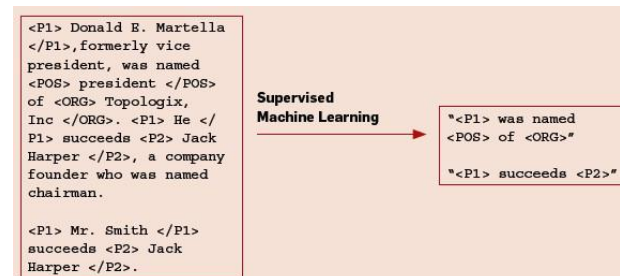
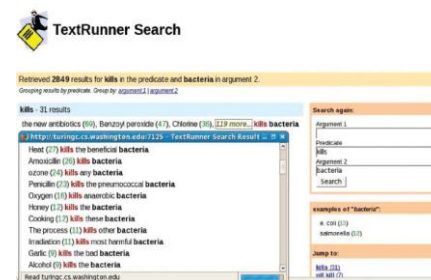
instance	iteration	date learned	confidence		
michael_cerveris is an actor	766	04-sep-2013	100.0	👍	👎
michael_w_hicks is a professor	766	04-sep-2013	95.7	👍	👎
lawn_green is a color	767	06-sep-2013	91.7	👍	👎
caloramator_proteoclasticus is a bacterium	766	04-sep-2013	94.9	👍	👎
grotduiiken_met_video is a cave	766	04-sep-2013	92.5	👍	👎
alarm has color red	771	26-sep-2013	100.0	👍	👎
beneficial_insects feeds on bugs	769	13-sep-2013	98.4	👍	👎
flowers is an agricultural product that attracts bugs	769	13-sep-2013	100.0	👍	👎
wsjx_tv is a TV affiliate of the network fox	771	26-sep-2013	100.0	👍	👎
mt is the capital city of the state or province iowa	771	26-sep-2013	99.6	👍	👎



UW: Machine Reading

- TexRunner、ReVerb、WOE、OLLIE

- 从Wikipedia Infobox获得关系名
- 通过在句法树上回标获得句法关系模板



思考

- NELL:
 - 给定了Ontology，约束了关系的类别，很难发现未知的实体关系
- University of Washington :
 - 从句法结构判别实体关系，可以发现未知的实体关系，但是所抽取的都是关系的mention，缺乏对于关系语义的确定
 - 需要对于关系的语义进行挖掘
- 已有方法都是集中于英文，在中文方面表现如何

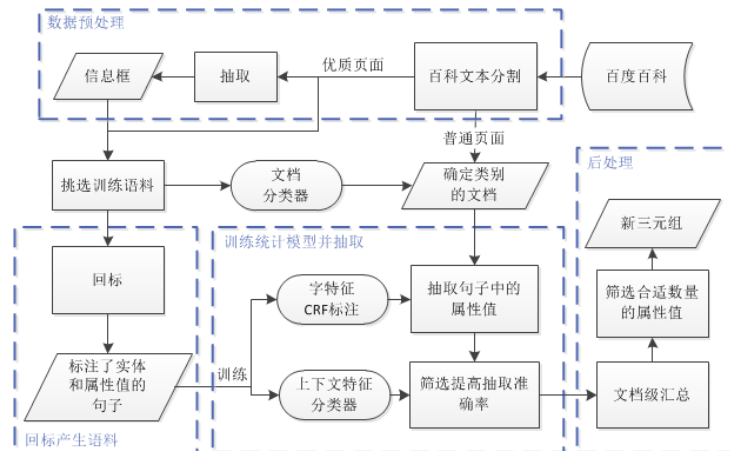
开放式中文实体关系抽取

- 已有百科知识进行回标产生训练语料并训练CRF抽取器

中文名:	姚明
外文名:	Yao Ming
别名:	小巨人 移动长城
国籍:	中国
民族:	汉族
出生地:	上海
出生日期:	1980年9月12日
身高:	2.23米 (7.32英尺)
体重:	140.6kg
运动项目:	篮球
所属运动队:	NBA火箭队

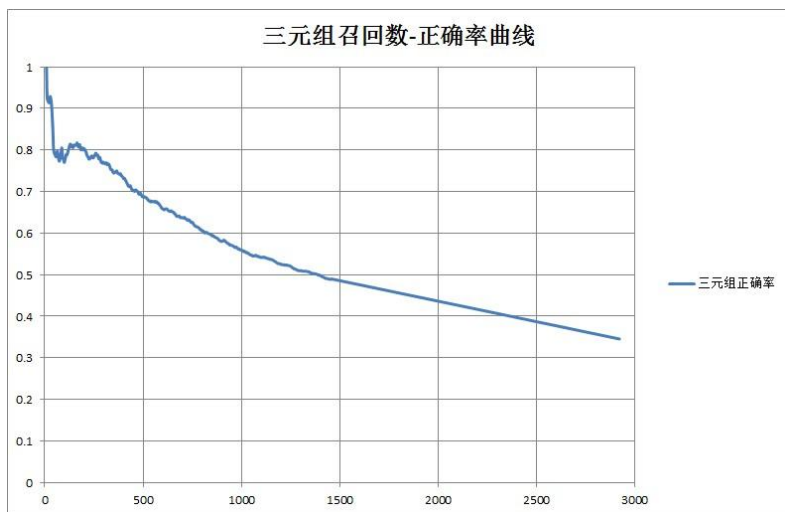
姚明 1980年生于上海。

- 对于新文档
 - 文档分类
 - 选择抽取器进行抽取
 - 句子级验证



开放式中文实体关系抽取

五个类别上测试：人物、植物、地理、电影、书籍



百度百科非结构化关系抽取

人物 动植物 图书 地理 电影

返回

下级地区
1个县 - 69.35%
行政区划类别
道 - 77.28%
所属地区
上海市 - 95.19%
著名景点
静安寺 - 98.39%
外滩 - 95.22%
大金山 - 90.04%
豫园 - 82.67%
枫泾古镇 - 79.63%
别名
沪 - 98.16%
政府驻地
黄浦区 - 53.53%
中文名称
华亭县 - 84.71%
面积
6340.5平方公里 - 95.99%
气候条件
北亚热带季风性气候 - 94.45%
火车站
上海站 - 93.90%
方言
吴语 - 70.93%
机场
浦东国际机场 - 85.32%
人口

上海 - 地理

上海，中国大陆第一大城市，四个中央直辖市之一，是中国大陆的经济、金融、贸易和航运中心。

上海创造和打破了中国世界纪录协会多项世界之最、中国之最。

上海位于我国大陆海岸线中部的长江口，拥有中国最大的外贸港口、最大的工业基地。

有超过2000万人居住和生活在上海地区，其中大部分属汉族江浙民系，通行吴语上海话。

上海又是一座新兴的旅游目的地，具有深厚的近代城市文化底蕴和众多的历史古迹。

如今上海已经发展成为一个闪耀全球的国际化大都市，并致力于在2020年建设成为国际金融中心和航运中心。

上海是2010年世界博览会举办城市。

上海，简称申或沪，地处中国漫长海岸线的最正中，亚洲第一大河长江的入海口以及亚太城市群的最核心。

常住人口2302万，其中户籍人口1412万，是中国第一大城市，也是世界人口最多的城市之一，城市规模已经超过了首都北京。

上海属亚热带湿润季风气候，四季分明。

一、二月最冷，最低气温为-5℃至-8℃，通常七月最热，最高气温达35℃—38℃。

每年六月中旬至七月上旬是梅雨季节。

上海历史悠久，已有两千多年历史。

上海春秋为吴国地，战国时为楚国春申君封邑，开始建城。

“申城”是上海地区最早的城市。

后来申城城址几经变迁，地名已经多次更改。

终于在三国时期于佘山附近固定了下来，并更名为“华亭”，唐朝设县，同时华亭县北部的上海镇也逐渐发展起来。

元朝至元二十八年七月，朝廷批准上海镇建独立县。

此日定为上海建城纪念日，距今已有700多年历史。

于是上海和华亭成为双子城。

思考

- Sentence Level vs. Set level
 - 构建知识图谱不需要正确识别每个句子中的实体关系
 - 充分利用网络数据的冗余特性
 - 根据数据源、文本信息结构的置信度进行投票

中文名: 姚明

外文名: Yao Ming

别名: 小巨人 移动长城

国籍: 中国

民族: 汉族

出生地: 上海市徐汇区

出生日期: 1980年9月12日

身高: 2.26米 (7.32英尺)

体重: 140.6kg

姚明，1980年生于上海市徐汇区，祖籍江苏省苏州市吴江区。美国NBA及世界篮球巨星，中国篮球史上里程碑式人物。原中国国家篮球队队员，曾效力于中国篮球职业联赛（CBA）上海大鲨鱼篮球俱乐部和美国国家篮球协会（NBA）休斯敦火箭队。姚明是中国最具影响力的人物之一，同时也是世界最著名的华人运动员之一，曾获7次NBA“全明星”，被美国《时代周刊》评为“2005年世界最具影响力100人”，被中国体育总局授予“体育运动荣誉奖章”“中国篮球杰出贡献奖”。2009年，姚明收购上海男篮，成为上海大鲨鱼篮球俱乐部老板。

姚明（1980年9月12日），著名篮球运动员，出生于中国上海的一个篮球世家，身高：2.26米。父亲姚志源，身高2.08米，曾效力于上海男篮；母亲方凤娣身高1.88米，70年代是中国女篮的主力队员。姚明17岁入选国家青年队；18岁穿上了中国队服。22岁以“状元身份”加入美国职业篮球队“火箭队”，开始了自己的NBA征程。2009年7月，收购上海男篮，成为国内球员兼老板第一人。2007年与同为篮球运动员的叶莉结为夫妻。2011年7月20日，姚明和他的团队

- 中文 vs. 英文
 - 中文文本缺乏严格的句法信息
 - Yao Ming was born in 1980.
 - 姚明，1980，上海人，篮球运动员.....

海量数据下的实体关系抽取

- 回标产生的训练语料越准确，训练得到的模型就越准确？
 - 增加两条规则以保证训练语料的正确性
 - TopN规则（回标后选取实体1与实体2最近的N个句子）
 - Top1 vs. Top5
 - 最近邻规则（当一个句子中出现多个实体1与多个实体2，则取最近邻的那个规则）
 - 最近邻 vs. 无最近邻
 - 用不同的训练语料训练抽取器

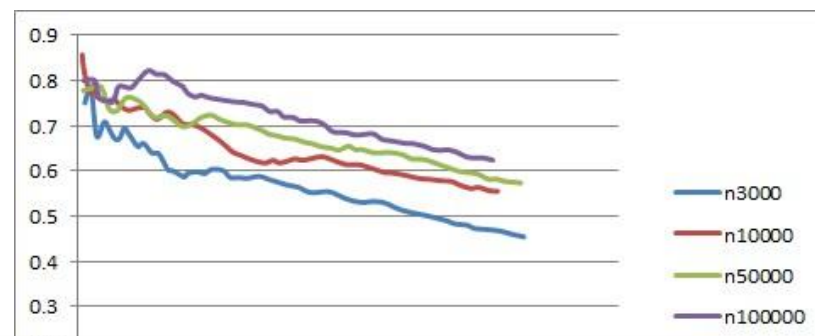
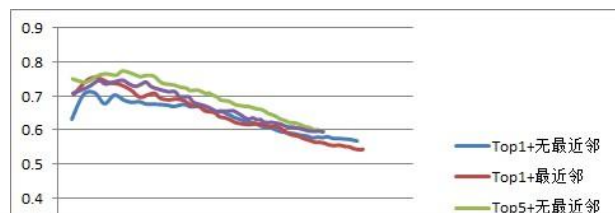
规则所产生的训练语料规模：

Top1+无最近邻 12.8 MB

Top1+最近邻 12.8 MB

Top5+无最近邻 25.4 MB

Top5+最近邻 25.4 MB



在大数据环境下，细致的处理不再重要
训练语料量的增加比训练语料质的提升更为重要

知识服务

已有的知识服务：检索与问答

李彦

Web Images Maps Shopping More Search tools

About 1,970,000 results (0.28 seconds)

李彦 - 北京大学化学与分子工程学院
www.chem.pku.edu.cn/liy/ Translate this page
Copper Catalyzing Growth of SWCNTs. Ultra-Low Feeding Gas Guided Non-Fast - Heating CVD Growth of Oriented Ultra-Long SWCNTs. SWCNTs Grown on ...

Li Yan
Soccer player

梁启超的儿子的老婆的情人的父亲

李彦(足球运动员) - 维基百科，自由的百科
zh.wikipedia.org/zh/李彦 (足球运动员) Translate this page
李彦 (1980年6月20日 -)，是中国退役足球运动员，曾效力于中国国家队，也是前中国国家足球队成员。

李彦 - 维基百科，自由的百科全书
zh.wikipedia.org/zh/李彦 Translate this page
这是一个消歧义页，罗列了有相同或相近的标题，但目的内部链接而转到本页，希望您能协助修正该处。

李彦(北宋) - 维基百科，自由的百科全书
zh.wikipedia.org/zh/李彦(北宋) Translate this page
李彦 (11世纪 - 1126年)，北宋太师，名列六贼之一。1121年，被杀。李彦为大内总管，将相之前。

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...
dnewsIT门户 - home.dnews.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徽因。看过《人间四月天》的人应该知道吧。那么林徽因的情人呢就是大名鼎鼎的徐志摩啦。那么徐志摩的老婆是谁呢？
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
梁启超的儿子的老婆的情人的老婆是谁？ 百度知道 2006-10-4

WolframAlpha computational knowledge engine

who is obama

Examples Random

Input interpretation:
Barack Obama (politician)

Basic information:

full name	Barack Hussein Obama II
date of birth	Friday, August 4, 1961 (age: 52 years)
place of birth	Honolulu, Hawaii, United States

Image:


Leadership position:

official position	President (44 th)
-------------------	-------------------------------



基于知识图谱的检索或问答的核心问题： Semantic Parsing

- 自然语言句子到知识库中概念和关系的映射



Semantic Parsing

- 传统semantic parsing
 - 在一个限定的领域中做semantic parsing
 - Ontology规模小
 - 基于关键词匹配或者人工书写模板
 - CCG(Combinatory Categorical Grammar)
 - PCCG(P_{Example Lexical Entries} Categorical Grammar)

New York City $\vdash NP : \text{new_york}$

neighborhoods in \vdash


$S \backslash NP / NP : \lambda x \lambda y. \text{neighborhoods}(x, y)$

Example CCG Grammar Rules

$X/Y : f \quad Y : g \Rightarrow X : f(g)$

$Y : g \quad X \backslash Y : f \Rightarrow X : f(g)$

评测: QALD



QALD-3 » Home

September 2013 · Co-located with: **CLEF 2013**

Multilingual Question Answering over Linked Data (QALD-3)

The **CLEF 2013 lab QALD-3** is the third in a series of evaluation campaigns on question answering over linked data, this time with a strong emphasis on multilinguality.

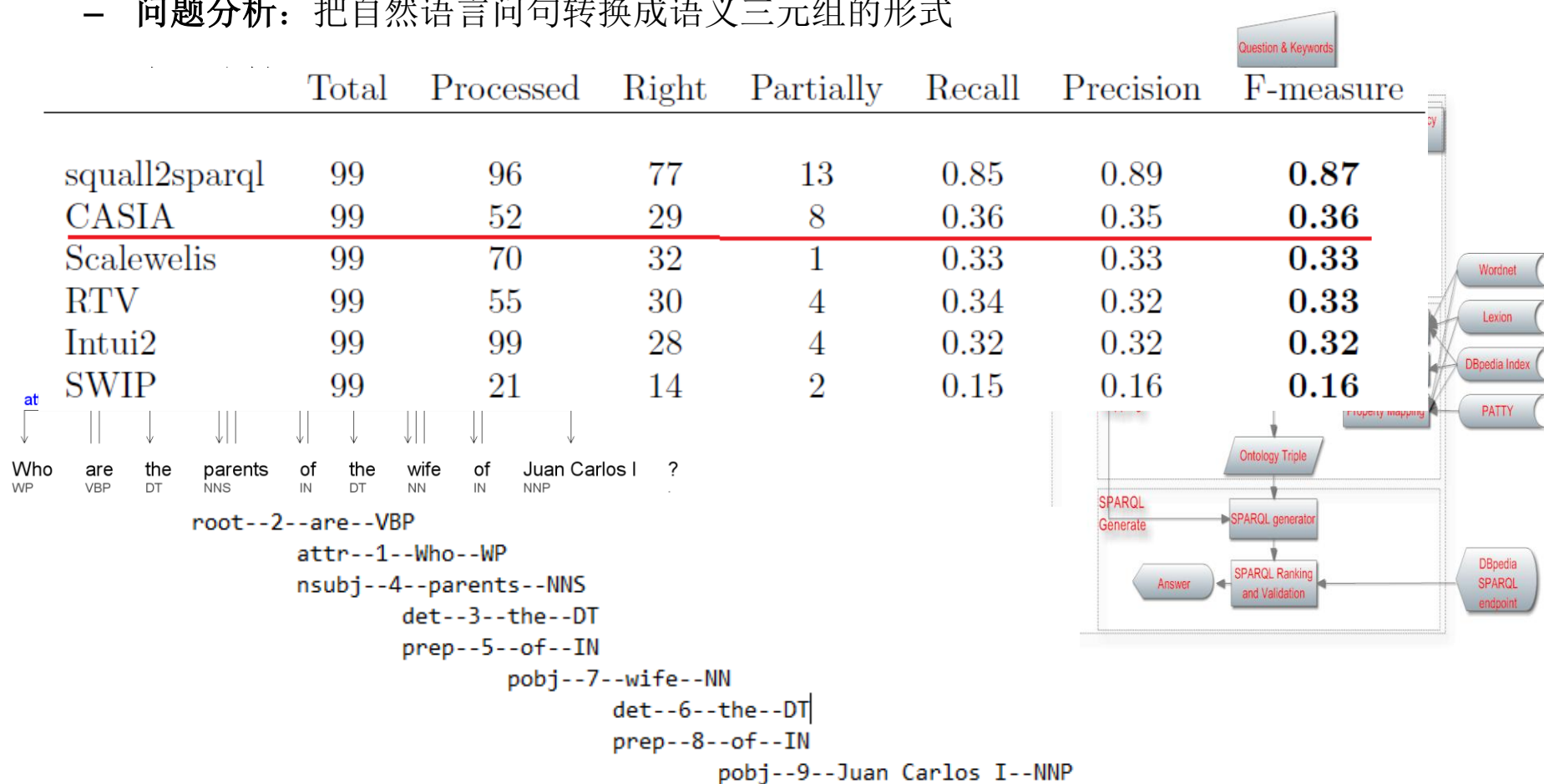
It offers two **open challenges**:

- Task 1: Multilingual question answering
- Task 2: Ontology lexicalization

面向复杂问句的知识问答

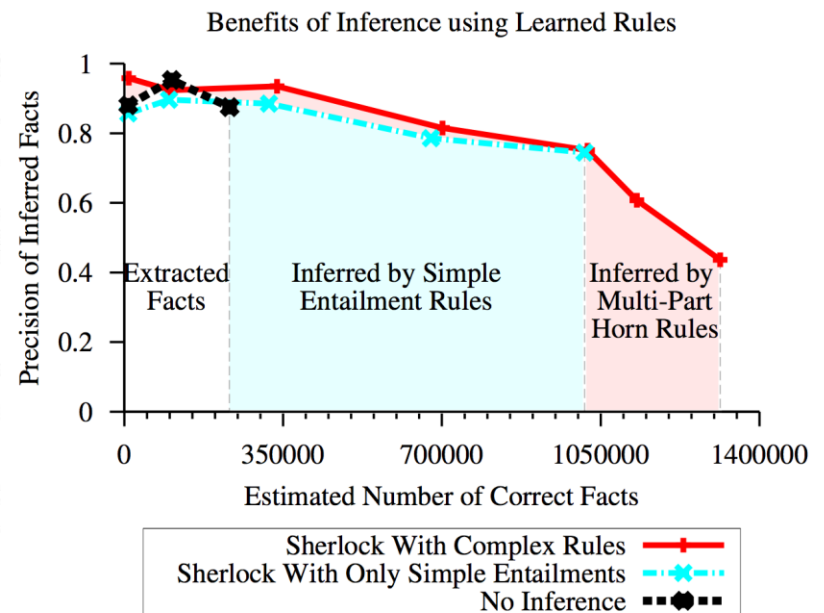
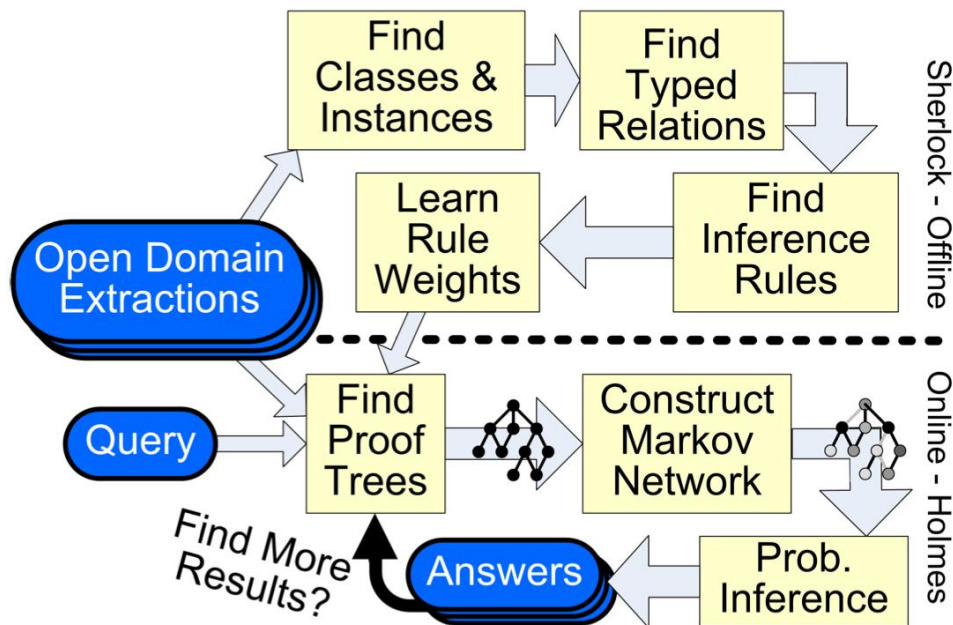
Who are the parents of the wife of Juan Carlos I?

- 问题分析：把自然语言问句转换成语义三元组的形式



Inference over the Web

- 关键难点
 - 如何学习鲁棒的推理规则
 - 如何推理、验证新的知识



小结

- 知识体系
 - 何种知识体系是有效的？
 - 是否需要建立知识体系的框架？或者建立进行ontology matching，或者Tag matching
- 知识获取
 - 非结构化文本的实体关系抽取是构建知识图谱的重要组成部分，目前的性能还未达到实用
 - 开放式关系抽取中，确定关系元组的语义重要
 - 中文知识抽取与英文有很大区别
- 知识服务
 - Semantic Parsing
 - 知识推理

谢谢