

法律声明

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院和讲者共同拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



第三讲 知识抽取与挖掘I

大纲

- 知识抽取任务定义和相关比赛
- 面向结构化数据的知识抽取
- 面向半结构化数据的知识抽取
- 实践展示：基于百科数据的知识抽取

第一部分:知识抽取任务定义和相关比赛

背景

互联网信息呈爆炸式增长



来源: CNNIC



中国网民规模达7.31亿
网站总数达到482万个
网页总数超过2360亿个

——中国互联网络发展状况统计报告2017

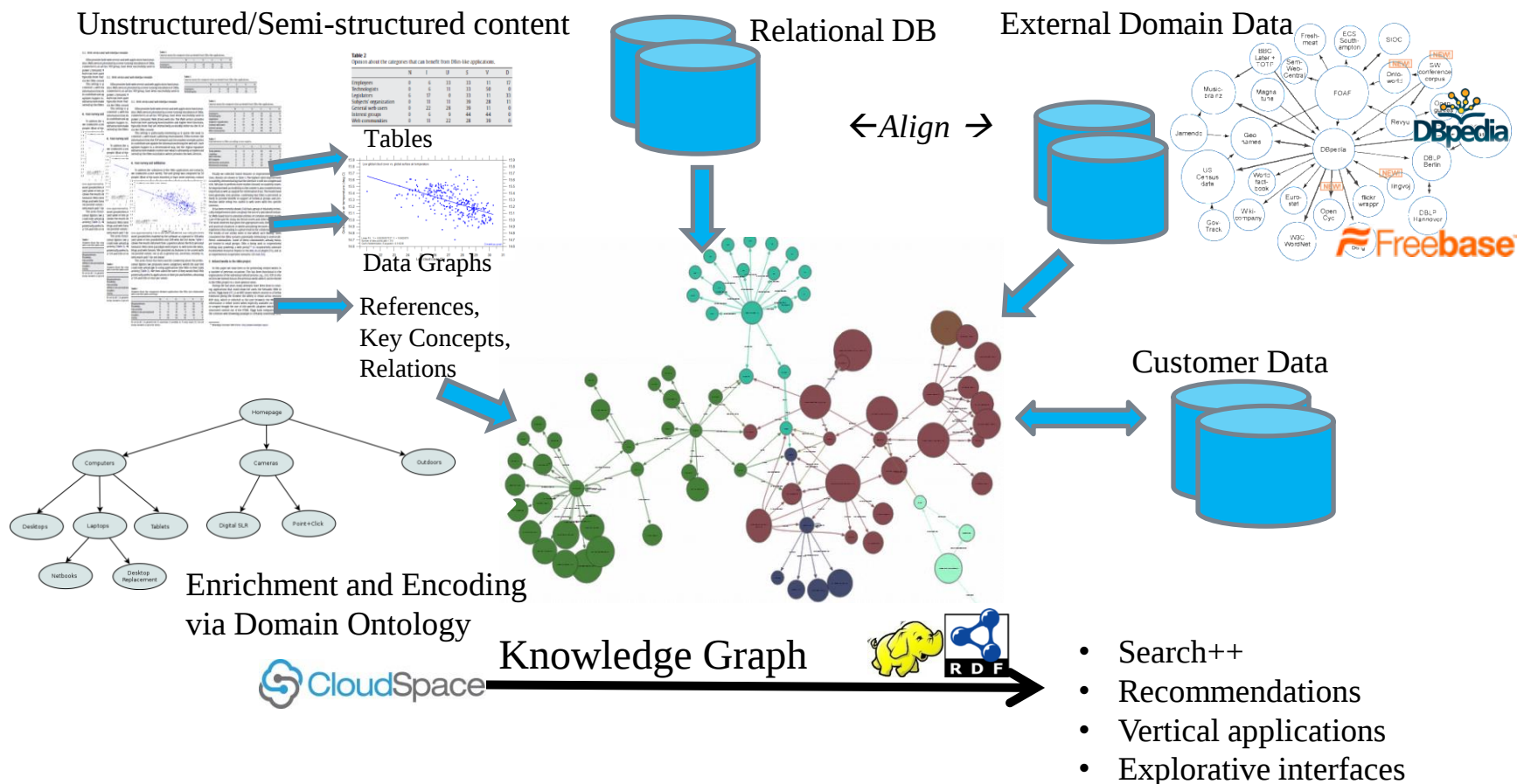


知识抽取技术

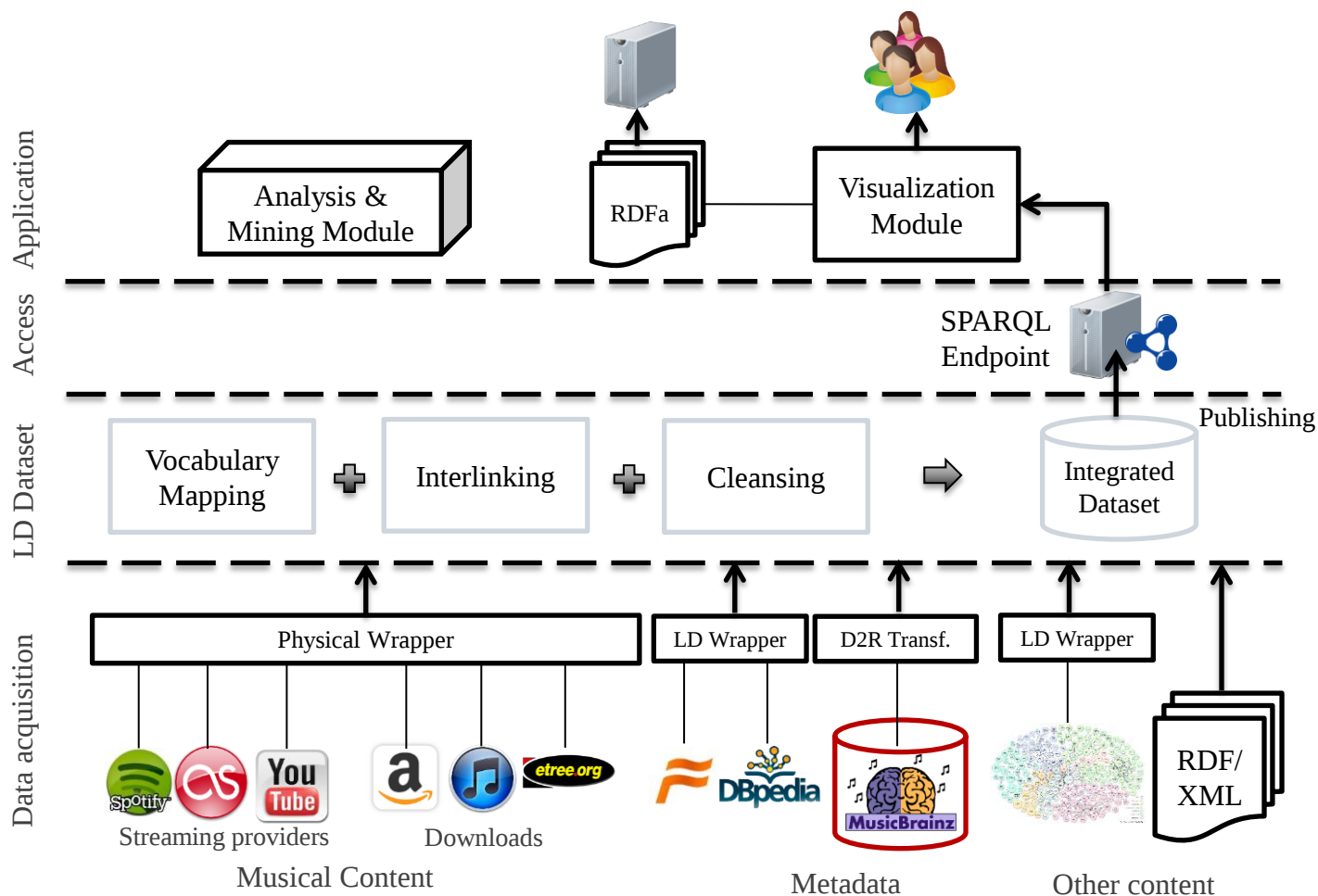
- 实体抽取
- 关系抽取
- 事件抽取



(行业) 知识图谱数据来源

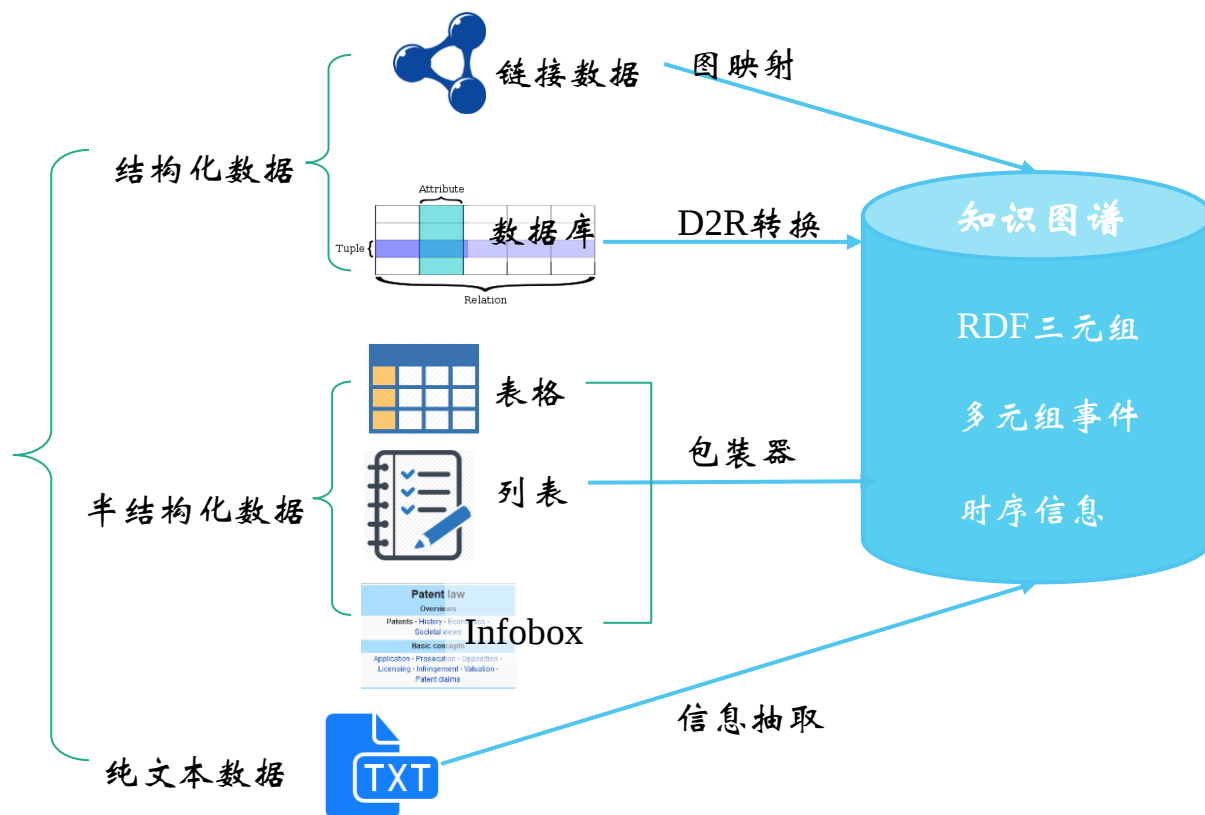


示例应用：音乐领域



知识抽取

从不同来源、不同结构的数据中进行知识提取，形成知识存入到知识图谱。



知识获取关键技术与难点

- 从结构化数据库中获取知识：D2R
 - 难点：复杂表数据的处理
- 从链接数据中获取知识：图映射
 - 难点：数据对齐
- 从半结构化（网站）数据中获取知识：使用包装器
 - 难点：方便的包装器定义方法，包装器自动生成、更新与维护
- 从文本中获取知识：信息抽取
 - 难点：结果的准确率与覆盖率

什么是知识抽取

- 20世纪70年代后期出现在NLP领域
- 自动化地从文本中发现和抽取相关信息
- 从多个文本碎片中合并信息
- 通常应用在特定领域
- 将非结构化转化为结构化数据
 - Schemas
 - Relations
 - Knowledge base
 - RDF triples

苹果公司将于西部时间9月12日上午10点(北京时间9月13日凌晨1点)举行新品发布会，这一次的发布会地点是全新建造的史蒂夫·乔布斯剧院。根据目前的消息，这次发布会上苹果将会发布iPhone 8(命名不确定，暂且称之为iPhone 8)、iPhone 7s、iPhone 7s Plus、Apple Watch 3以及全新Apple TV。



事件类型	发布会
公司	苹果公司
时间	西部时间9月12日上午10点
地点	史蒂夫·乔布斯剧院
产品	iPhone8、iPhone7s、iPhone7s plus、apple watch 3、apple TV

子任务

● 命名实体识别

- 检测：  库克非常兴奋。 [库克]: 实体
- 分类：  库克非常兴奋。 [库克]: 人物

● 术语抽取

从语料中发现多个单词组成的相关术语。

● 关系抽取

王思聪是万达集团董事长王健林的独子。  [王健林] <父子关系> [王思聪]

● 事件抽取

据路透社消息，英国当地时间9月15日早8时15分，位于伦敦西南地铁线District Line的Parsons Green地铁站发生爆炸，目前已确定有多人受伤，具体伤亡人数尚不明确。目前，英国警方已将此次爆炸与起火定性为恐怖袭击。

- 恐怖袭击事件 →

触发词：	发生爆炸
时间：	当地时间9月15日早8时15分
地点：	Parsons Green地铁站
攻击者：	-
伤亡人数：	-

● 共指消解

[美国总统特朗普]否决了一家有中资背景的私募基金对美国莱迪思半导体公司的收购案，[他]在多个专家小组的建议下做出该决定。



相关竞赛与数据集

Message Understanding Conference (MUC)

由美国DARPA启动并资助的项目，目的是鼓励 and 开发更好的信息抽取方法

<i>Conference</i>	<i>Year</i>	<i>Text Source</i>	<i>Topic (Domain)</i>
MUC-1	1987	Mil. reports	Fleet Operations
MUC-2	1989	Mil. reports	Fleet Operations
MUC-3	1991	News reports	Terrorist activities in Latin America
MUC-4	1992	News reports	Terrorist activities in Latin America
MUC-5	1993	News reports	Corporate Joint Ventures, Microelectronic production
MUC-6	1995	News reports	Negotiation of Labor Disputes and Corporate Management Succession
MUC-7	1997	News reports	Airplane crashes, and Rocket/Missile Launches

https://en.wikipedia.org/wiki/Message_Understanding_Conference

Message Understanding Conference (MUC)

□ 命名实体识别 (Named Entity Recognition, NER)

- 2017年10月31日, 宋仲基和宋慧乔大婚
 - 日期: <TIMEX TYPE="DATE"> 2017年10月31日 </TIMEX>
 - 人名: <ENAMEX TYPE="PERSON"> 宋仲基 </ENAMEX>
 - 人名: <ENAMEX TYPE="PERSON"> 宋慧乔 </ENAMEX>

□ 共指消解 (Co-reference Resolution, CR)

- 中国江苏省会南京简称宁
 - <COREF ID="100" MIN="南京市"> 南京 </COREF>
 - <COREF ID="101" TYPE="IDENT" REF="100"> 宁 </COREF>

Automatic Content Extraction (ACE)

ACE对MUC定义的任务进行了融合、分类和细化；
主要分为五大任务，包含英语、阿拉伯语和汉语；

❑ 实体检测与识别 (Entity Detection And Recognition, EDR)

- such as: persons, organizations, locations, facilities, weapons, vehicles, and geo-political entities
- [俄罗斯检方]已经要求将涉嫌间谍罪的[[美国嫌犯]][波普]判处20年徒刑

❑ 数值检测与识别 (Value Detection And Recognition, VAL)

- 百分比: [25/100]
- 钱: [20 元]
- 邮箱: example@domain.com
- 时间: 美国商会中国分会[近日]派出一个25人组成的代表团，在华盛顿向国会和白宫展开为期[一周]的游说活动

Automatic Content Extraction (ACE)

❑ 时间表达检测与识别 (Time Detection And Recognition, TERN)

- [1995年10月], 中国长江三峡工程开发...
- [公元前2100年左右], 美索布达米亚人已有了乘法表。

❑ 关系检测与识别 (Relation Detection And Recognition, RDR)

- such as: person A is the manager of company B. Relation types include: role, part, located, near, and social
- 张三出席了上海的一个国际会议 位于(张三, 上海)

❑ 事件检测与识别 (Event Detection And Recognition, VDR)

- such as: interaction, movement, transfer, creation and destruction
- 布什总统周六晚上离开华盛顿前往巴黎同 欧盟领导会谈
 - 移动事件 (人物: 布什总统 起点: 华盛顿 目的地: 巴黎 时间: 周六晚上)
 - 会议事件 (人物: 布什总统, 欧盟领导 地点: 巴黎 时间: 周六晚上)

TAC Knowledge Base Population (KBP)

KBP对ACE定义的任务进一步修订, 适合现代知识抽取的需求
主要分为四个独立任务和一个整合任务 <https://tac.nist.gov/2017/KBP/>

❑ 实体发现与链接 (Entity Discovery and Linking, EDL)

- person (PER), organization (ORG), geopolitical entity (GPE), location (LOC), and facility (FAC) entities mentioned in the documents, and to link each mention to its KB node

❑ 槽填充 (Slot Filling, SF)

- to fill in values for specific attributes ("slots") for specific entities
- 姚明, 1980年9月12日出生于上海市徐汇区, 祖籍江苏省苏州市吴江区震泽镇

主语	谓语	宾语
姚明	出生日期	1980年9月12日
姚明	出生地	上海市徐汇区
姚明	祖籍	江苏省苏州市吴江区震泽镇

TAC Knowledge Base Population (KBP)

❑ 事件抽取 (Event)

- Event Nugget (EN) to detect event nuggets (i.e., mentions of events in text), and Event Argument (EAL) to extract event arguments and link arguments that belong to the same event.

❑ 信念和情感 (Belief and Sentiment, BeSt)

- detects belief and sentiment of an entity toward another entity, relation, or event
- 联合国安理会谴责埃及恐怖袭击事件
 - 谴责 (发起方: 联合国安理会 承受方: 埃及恐怖袭击事件)

❑ 端到端冷启动知识构建

- build a KB from scratch, using a predefined KB schema and a collection of unstructured text

Semantic Evaluation(SemEval)

由ACL-SIGLEX组织的国际权威的词义消歧评测，目标是增进人们对词义与多义现象的理解

<https://en.wikipedia.org/wiki/SemEval>

会议	部分任务
Senseval-1	Word Sense Disambiguation(WSD)、Lexical Sample WSD
Senseval-2	WSD、Translation WSD
Senseval-3	Logic Form Transformation、Machine Translation、Semantic Role Labelling(SRL)、WSD
SemEval2007	WSD、Time Expression、Sentiment Analysis、Frame Extraction、Information Extraction 等
SemEval2010	Coreference、Cross-lingual、Noun Compounds、Semantic Relations、SRL、Textual Entailment等
SemEval2012	Common Sense Reasoning、Lexical Simplification、Semantic and Textual Similarity等
SemEval2013	Temporal Annotation、Cross and Multilingual WSD、BioMedical Texts、Textual Similarity等
SemEval2014	Compositional Distributional Semantic、Cross-Level Semantic Similarity、Sentiment Analysis等
SemEval2015	Text Similarity and Question Answering、Learning Semantic Relations、Time and Space等
SemEval2016	Textual Similarity and Question Answering、Sentiment Analysis、Semantic Taxonomy等
SemEval2017	Semantic comparison for words and texts、Detecting sentiment, humor, and truth等

应用领域

- 恐怖袭击
- 企业联合经营
- 空难事故
- 疾病爆发
- 研讨会公告
- 生物医药
- ...



实体抽取定义

任务定义：抽取文本中的原子信息元素

- 人名
- 组织/机构名
- 地理位置
- 时间/日期
- 字符值
- 金额值

实体抽取举例

北京时间10月25日，骑士后来居上，在主场以119-112击退公牛。

↑ ↑ ↑ ↑

地点 时间 组织 组织

中新社华盛顿10月24日电 美国众议院三个委员会24日宣布将分别展开两项与希拉里·克林顿有关的调查，国会民主党人称这是共和党人试图转移注意力。

↑

人物

序列标注方法

人工特征

词本身的特征

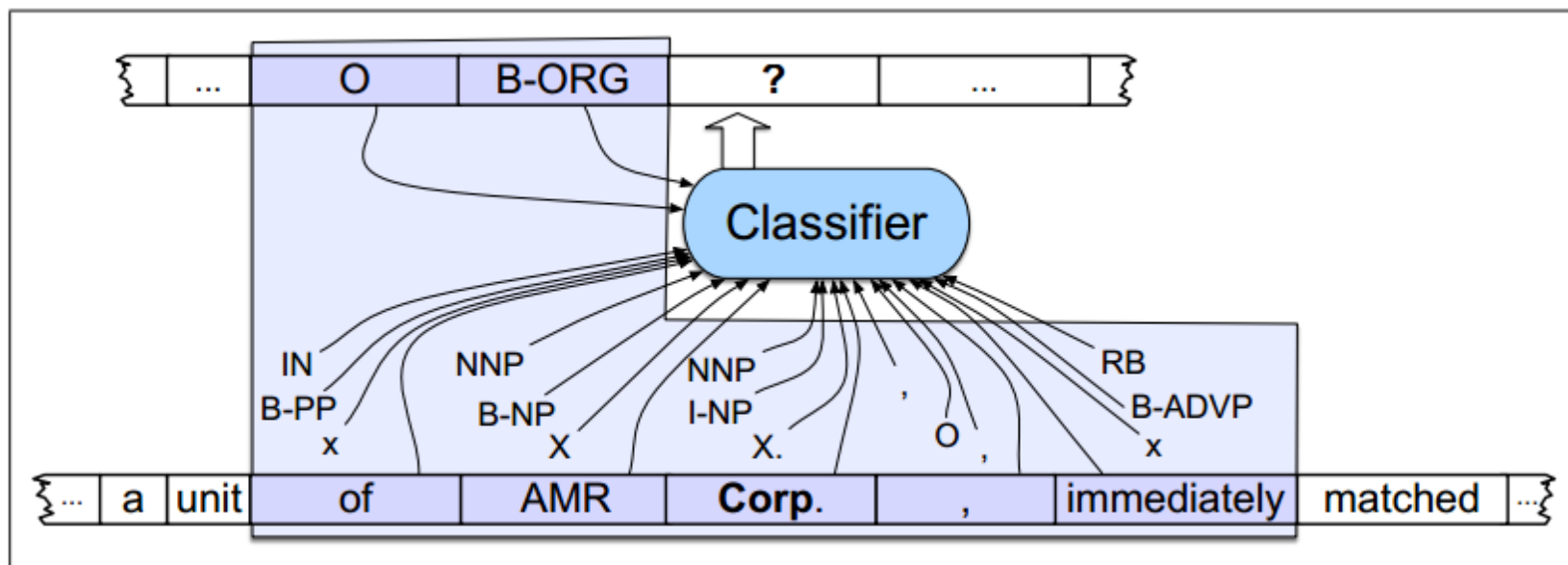
- 边界特征：边界词概率
- 词性
- 依存关系

前后缀特征

- 姓氏：李XX、王X
- 地名：XX省、XX市

字本身的特征

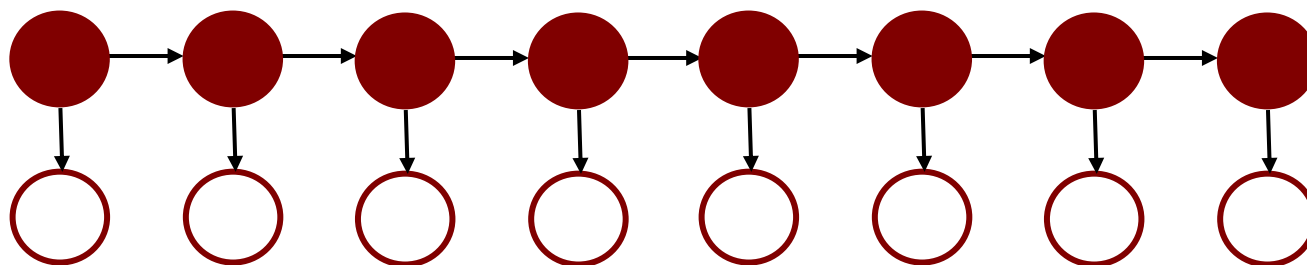
- 是否是数字
- 是否是字符



实体识别的序列标注实例

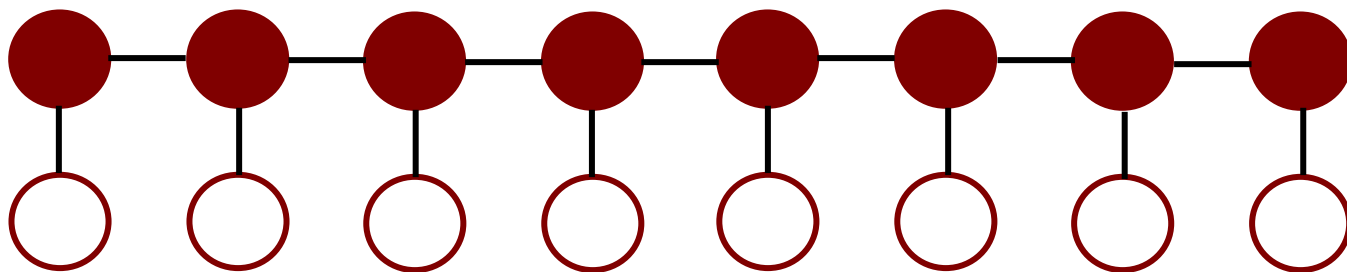
	IOB标注体系	IO标注体系		IOB标注体系	IO标注体系
由	O	O	印	B-ORG	I-ORG
中	B-ORG	I-ORG	尼	I-ORG	I-ORG
国	I-ORG	I-ORG	女	I-ORG	I-ORG
女	I-ORG	I-ORG	队	I-ORG	I-ORG
队	I-ORG	I-ORG	的	O	O
的	O	O	王	B-PER	I-PER
叶	B-PER	I-PER	莲	I-PER	I-PER
钊	I-PER	I-PER	香	I-PER	I-PER
颖	I-PER	I-PER			
迎	O	O			
战	O	O			

HMM (隐马尔可夫模型)



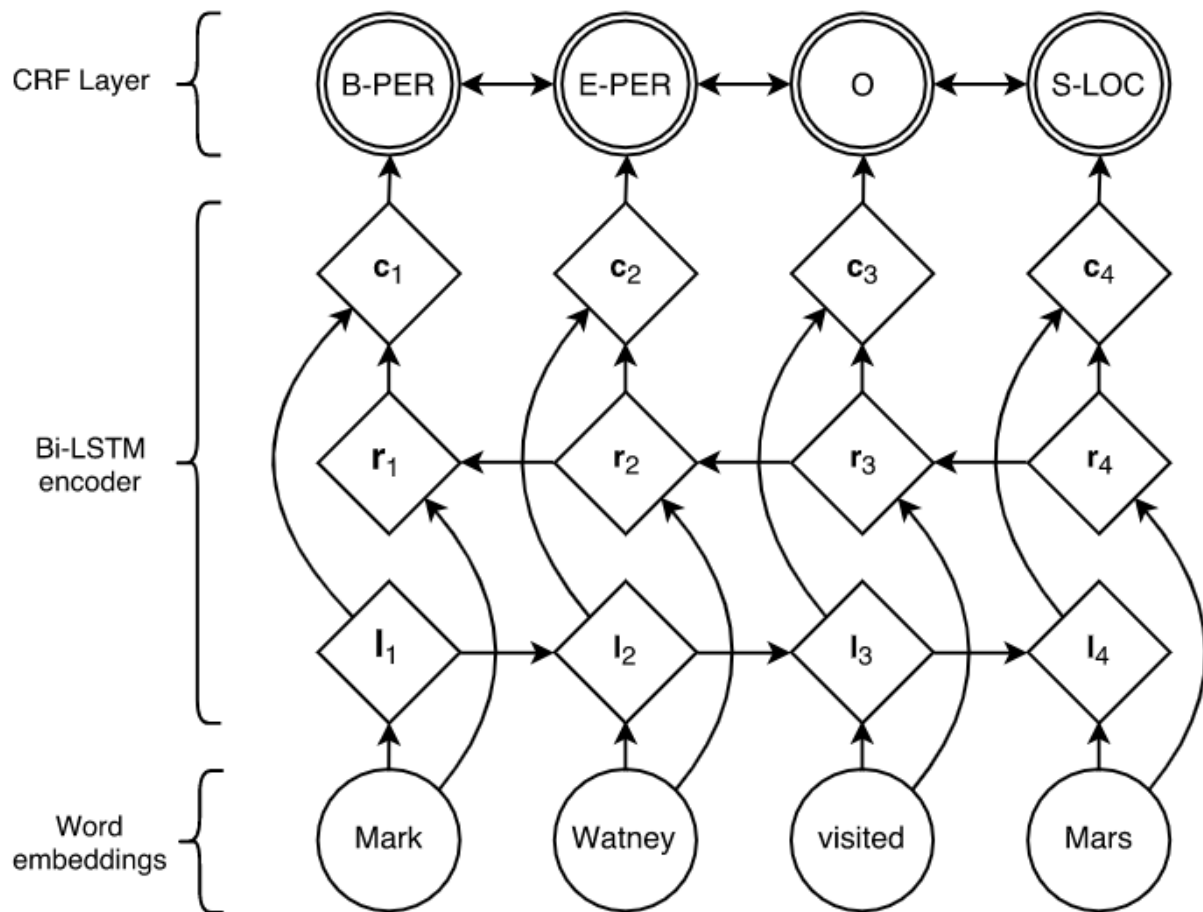
- 有向图模型
- 生成式模型
 - 找到使 $P(X,Y)$ 最大的参数
- 假设特征之间是独立的

CRF (条件随机场)



- ☐ 无向图模型
- ☐ 判别式模型
 - 找到使 $P(Y|X)$ 最大的参数
- ☐ 没有关于特征之间是独立的假设

LSTM+CRF



- ☐ 判别式
- ☐ 端到端
- ☐ Dropout
- ☐ 特征自动提取

Guillaume Lample et al. NAACL-HIT. 2016

方法效果比较

System	accuracy
Combination of HMM, Maxent etc. (Florian et al., 2003)	87.76
MaxEnt classifier (Chieu., 2003)	88.31
Semi-supervised model combination (Ando and Zhang., 2005)	89.31
Conv-CRF (Collobert et al., 2011)	81.47
Conv-CRF (Senna + Gazetteer) (Collobert et al., 2011)	89.59
CRF with Lexicon Infused Embeddings (Passos et al., 2014)	90.90
BI-LSTM-CRF (ours)	84.26
BI-LSTM-CRF (Senna + Gazetteer) (ours)	90.10

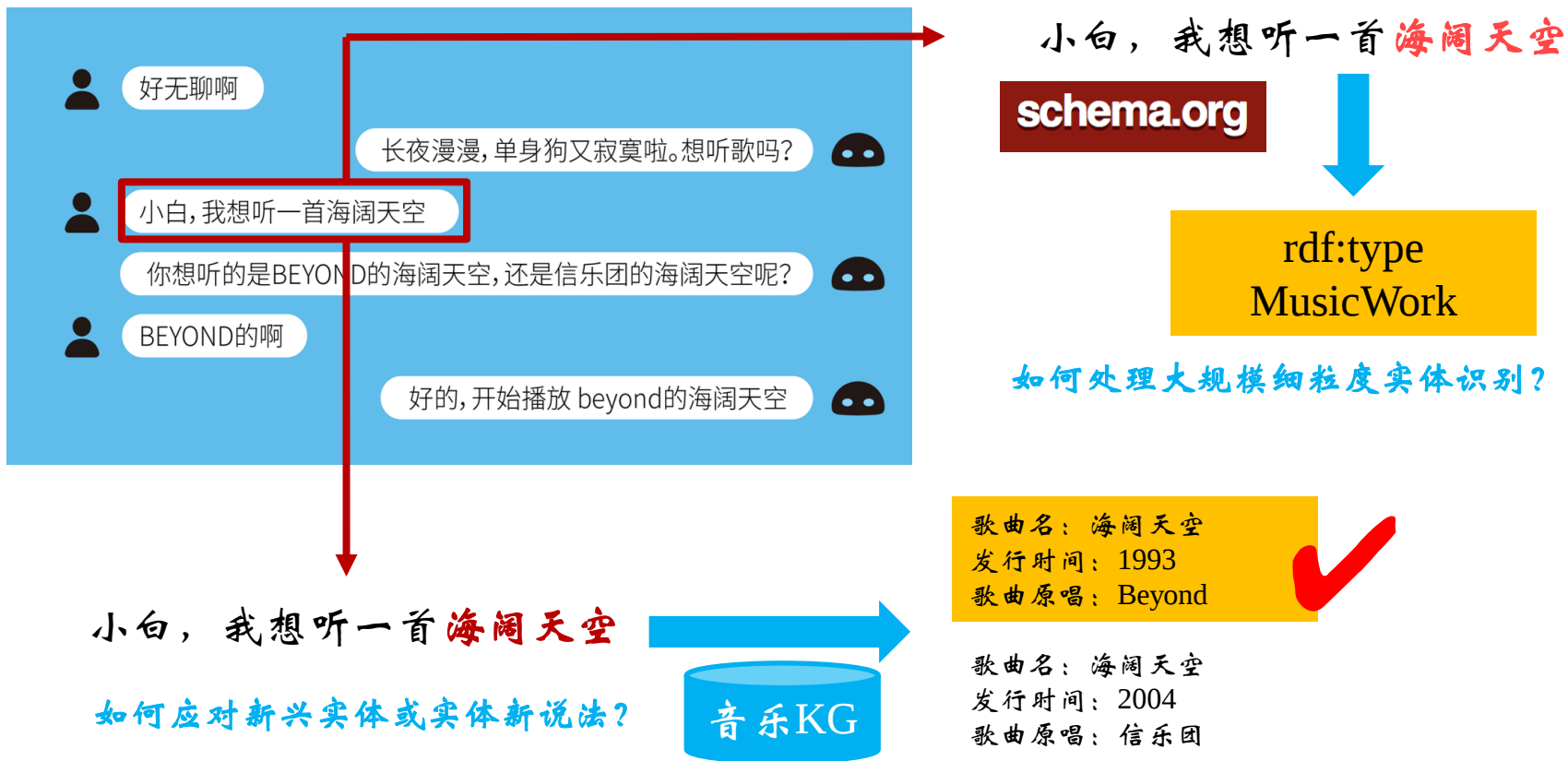
不同模型效果对比 (F1值)

Guillaume Lample, et al. CoRR. 2015

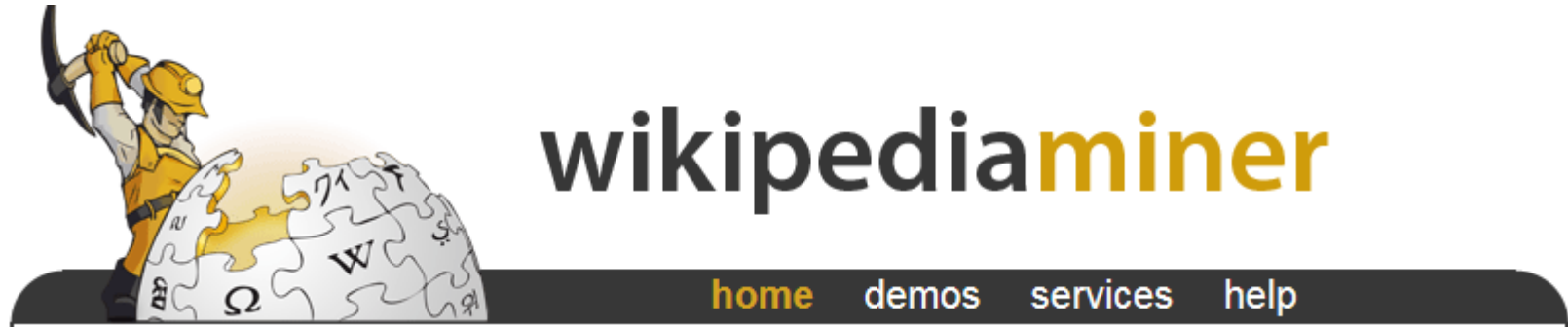
实体抽取—参考文献

- hiheng Huang, Wei Xu, Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging. CoRR. 2015
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyer. Neural Architectures for Named Entity Recognition. The 2016 Conference of the North American Chapter of the Association for Computational Linguistics. 2016: 260-270

实体识别与链接



Wikipedia Miner



- ☐ Open source
- ☐ (Public) web service
 - Java
 - Hadoop preprocessing pipeline
- ☐ Lexical matching + machine learning
- ☐ Target KB: Wikipedia
- ☐ See <http://wikipedia-miner.cms.waikato.ac.nz>

introduction

the demos

search

compare

annotate

Text to annotate

Wikipedia is a free, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation. Its name is a portmanteau of the words wiki (a technology for creating collaborative websites, from the Hawaiian word wiki, meaning 'fast') and encyclopedia.

Wikipedia's 12 million articles (2.77 million in English) have been written collaboratively by volunteers around the world, and almost all of its articles can be edited by anyone who can access the Wikipedia website. Launched in January 2001 by Jimmy Wales and Larry Sanger, it is currently the most popular general reference work on the Internet.

[show options](#)**Annotate**

Annotated text

MediaWiki Markup**Detected Topics**

[[Wikipedia]] is a free, multilingual encyclopedia project supported by the [[Nonprofit organization|non-profit]] [[Wikimedia Foundation]]. Its name is a [[portmanteau]] of the words [[wiki]] (a technology for creating collaborative websites, from the [[Hawaiian language|Hawaiian word]] wiki, meaning 'fast') and encyclopedia.

Wikipedia's 12 million articles (2.77 million in English) have been written collaboratively by volunteers around the world, and almost all of its articles can be edited by anyone who can access the Wikipedia website. Launched in January 2001 by [[Jimmy Wales]] and [[Larry Sanger]], it is currently the most popular general reference work on the Internet.

no definition available

59% probability of being a link

DBpedia Spotlight



- ☐ Open source
- ☐ Public web service
- ☐ Disambiguation in local context
 - vector-space model using bag-of-words and cosine similarity
 - (actually, Lucene)
- ☐ Target KB: DBpedia
- ☐ See <http://spotlight.dbpedia.org>

Confidence:

0.5

Language: English



☐ n-best candidates

SELECT TYPES...

ANNOTATE

Wikipedia is a free, multilingual encyclopedia project supported by the non-profit [Wikimedia Foundation](#). Its name is a [portmanteau](#) of the words wiki (a technology for creating collaborative websites, from the [Hawaiian](#) word wiki, meaning 'fast') and encyclopedia.

Wikipedia's 12 million http://dbpedia.org/resource/Hawaiian_language written collaboratively by volunteers around the world, and almost all of its articles can be edited by anyone who can access the Wikipedia website. Launched in January 2001 by [Jimmy Wales](#) and [Larry Sanger](#), it is currently the most popular general reference work on the [Internet](#).

http://dbpedia.org/resource/Jimmy_Wales

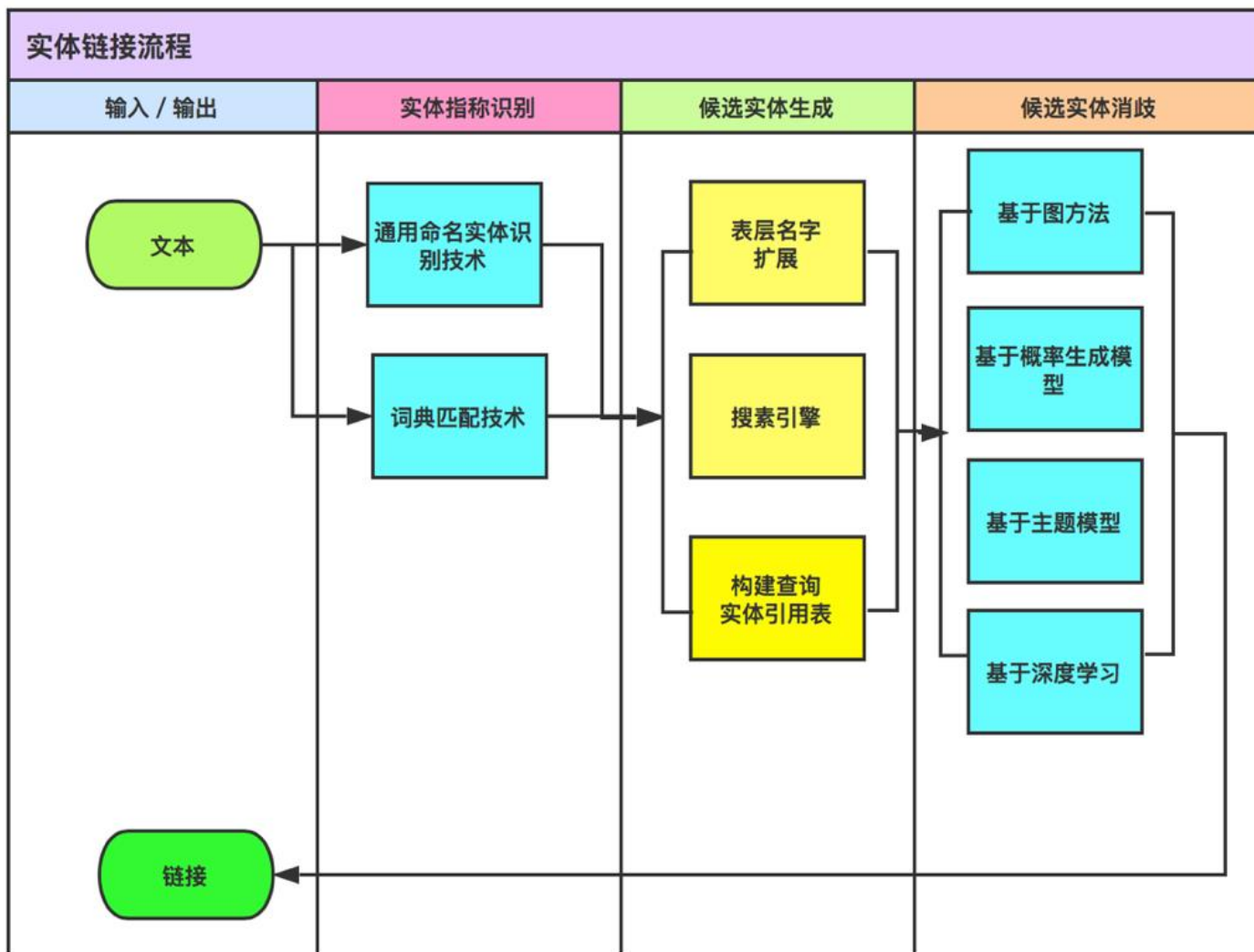
BACK TO TEXT

Demo: <http://dbpedia-spotlight.github.io/demo/>

OpenCalais

- ❑ Only on public content
 - does not keep a copy of content
 - keeps a copy of the metadata it extracts
- ❑ Free for up to 50,000 documents per day
- ❑ Early adopters:
 - CBS Interactive / CNET, Huffington Post, Al Jazeera, The White House
 - more than 30,000 developers && 50 publishers
- ❑ Target KB: Calais
- ❑ See <http://www.opencalais.com/>

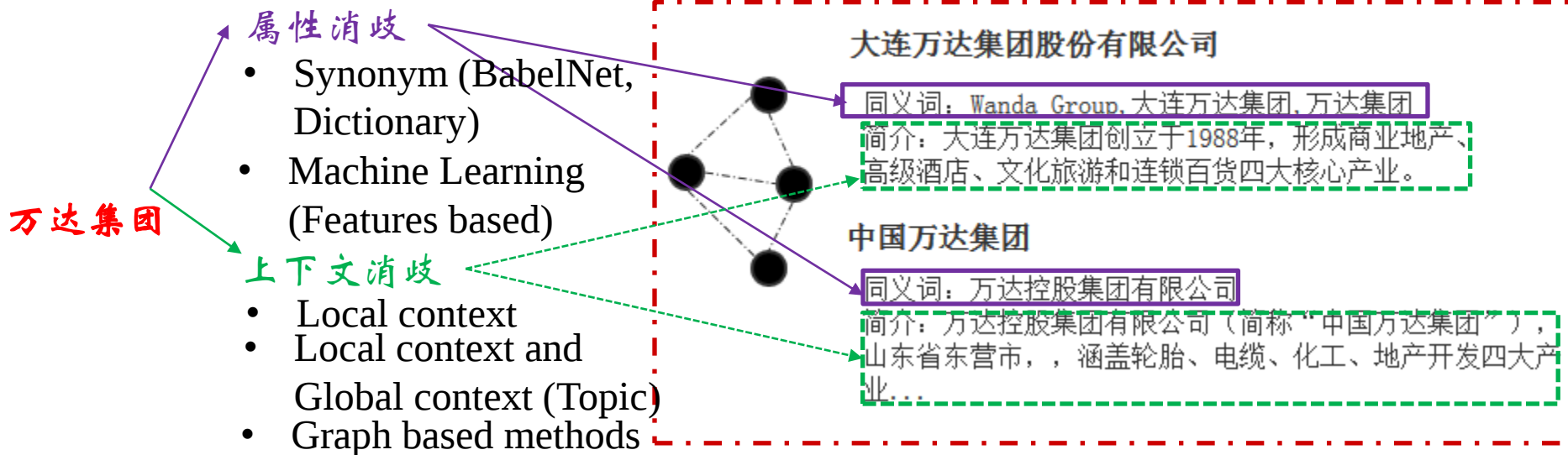




实体链接

中国证券网讯（记者 王雪青）中国证券网记者今日获悉，**万达集团**的文化产业版图将再添世界级新军——传奇影业，具体收购情况或于下周二正式发布。

KG



开源系统

- 【1】 <http://acube.di.unipi.it/tagme/>
- 【2】 <https://github.com/parthatalukdar/junto>
- 【3】 <http://orion.tw.rpi.edu/~zhengj3/wod/wikify.php>
- 【4】 <https://github.com/yahoo/FEL>
- 【5】 <https://github.com/yago-naga/aida>
- 【6】 <http://www.nzdl.org/wikification/about.html>
- 【7】 <http://aksw.org/Projects/AGDISTIS.html>
- 【8】 <https://github.com/dalab/pboh-entity-linking>

关系抽取简介

什么是关系抽取？

- 信息抽取 (Information Extraction) 研究领域的任务之一
- 从文本中抽取两个或者多个实体之间的语义关系

举例：

王健林谈儿子王思聪：我期望他稳重一点。

➡ 父子 (王健林, 王思聪)

中国证券网讯（记者 严政）卓翼科技3月4日晚公告称，公司于近日收到中兴通讯股份有限公司发出的《中标结果》通知，公司在中兴通讯EPON产品、GPON产品招标中均获第二名的份额，预计本次中标总金额约为1.914亿元（不含税），占公司2011年度经审计的营业收入的15.46%。

其中，公司本次中标中兴通讯四个EPON产品类别，中兴通讯预计该四个EPON产品类别2013年全年总量为270万台，根据招标结果，公司中标份额的比例分别为20%、20%、30%、30%，对应的产品总量约为57万台，预计金额为8800万元，占公司2011年度经审计的营业收入的7.11%。



公司A	公司B	关系 (A是B的)	时间	来源
中兴通讯	卓翼科技 (002369)	客户	2013.03.05	中国证券网 公司公告
中兴康讯	Acacia	客户	2015.12.28	OFweek光通讯网 行业新闻



Acacia自2011年开始出货高速高效节能产品，目前已拥有20家客户，包括ADVA、中兴康讯和阿尔卡特朗讯。Acacia与阿尔卡特朗讯关系良好——其CEO Raj Shanmugarai曾是阿尔卡特朗讯美国公司的光网络事业部的业务发展副总裁。

中国联通与中兴通讯正式签署SDN/NFV战略合作协议

2016-03-23

中兴通信中标联通规模最大100G项目

作者：李雁争 来源：上海证券报

2015-06-12 09:56 | 评论 |

近日，中国联通和中兴通讯共同宣布，双方正式签署SDN/NFV战略合作协议。双方旨在通过协同创新更好地把握以SDN/NFV为代表的网络发展趋势，实现SDN/NFV的落地应用，推动中国联通新一代网络的构建，快速满

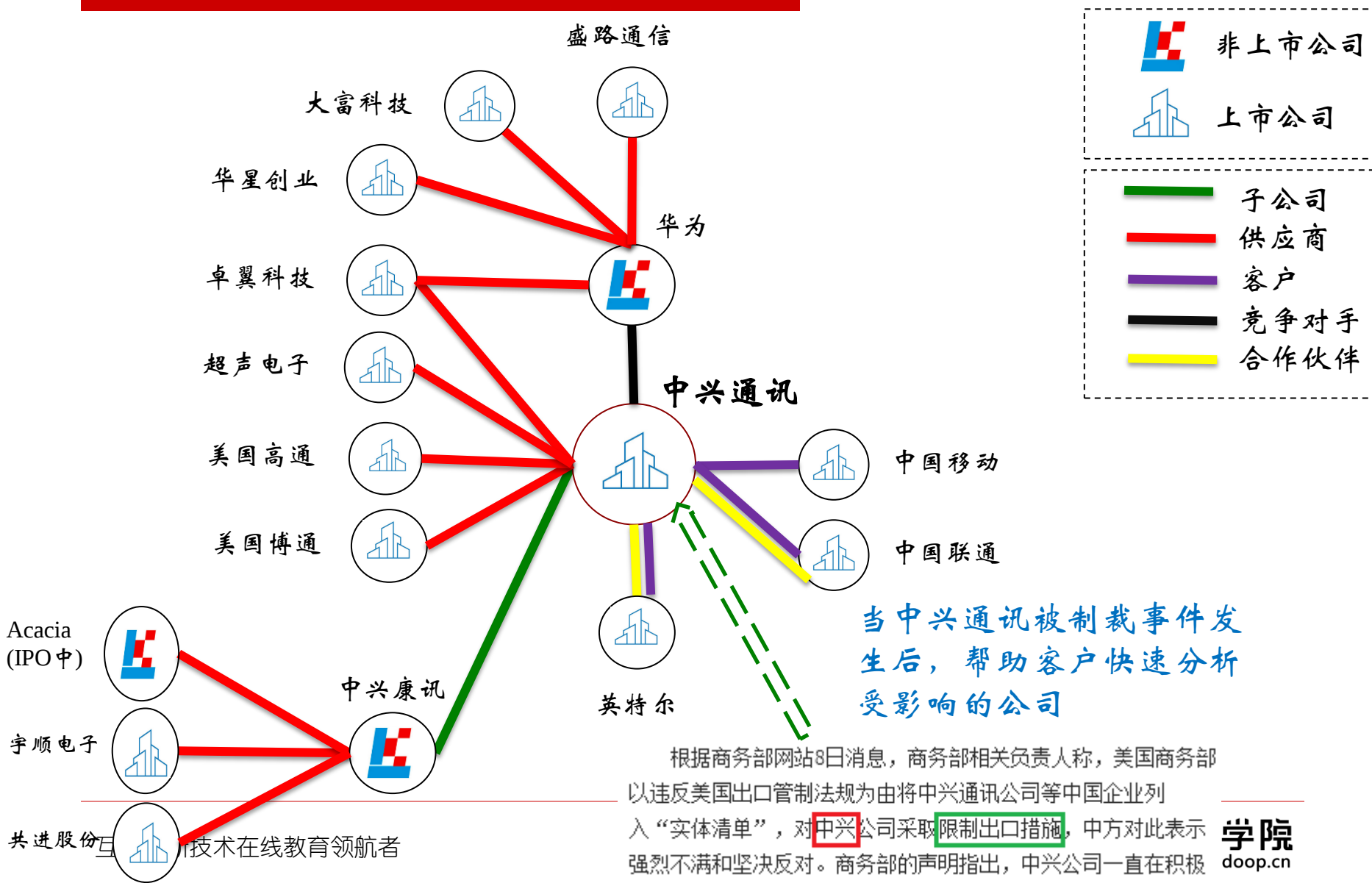
公司A	公司B	关系 (A是B的)	时间	来源
中兴通讯	中国联通	合作伙伴	2016.03.23	公司新闻
中兴通讯	中国联通	客户	2015.06.12	公司新闻
中兴通讯	英特尔(INTC)	合作伙伴	2013.01.16	公司新闻

[首页](#) » [新闻资讯](#) » [新闻动态](#)

中兴通讯携手英特尔重推CoCore E9000超高密刀片服务器

2013-01-16

Enterprise Knowledge Graph



方法分类

□ 基于模板的方法

- 基于触发词的Pattern
- 基于依存句法分析的Pattern

□ 监督学习方法

- 机器学习方法
- 深度学习方法

□ 弱监督学习方法

- 远程监督
- Bootstrapping

基于模板的方法——基于触发词的Pattern

姚明 老婆 叶莉
徐峥 老婆 陶虹

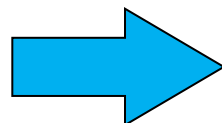


X 老婆 Y

黄晓明 妻子 杨颖



X 妻子 Y



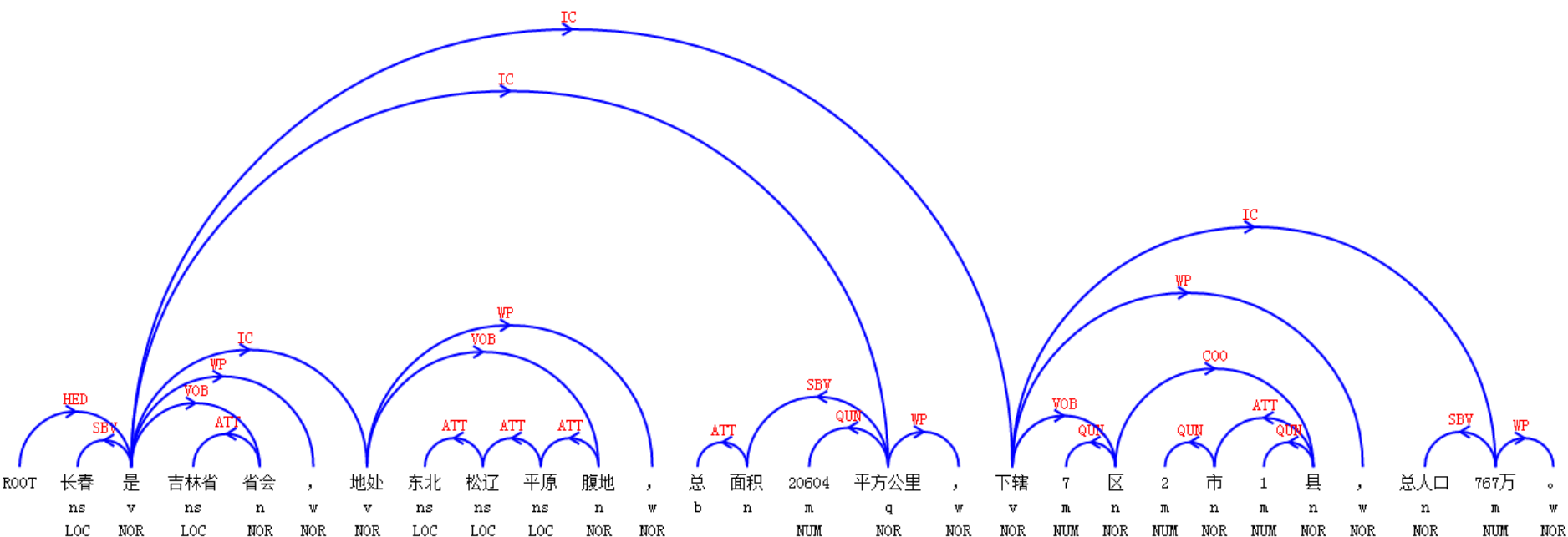
夫妻关系 (X, Y)

刘德华 配偶 朱丽倩



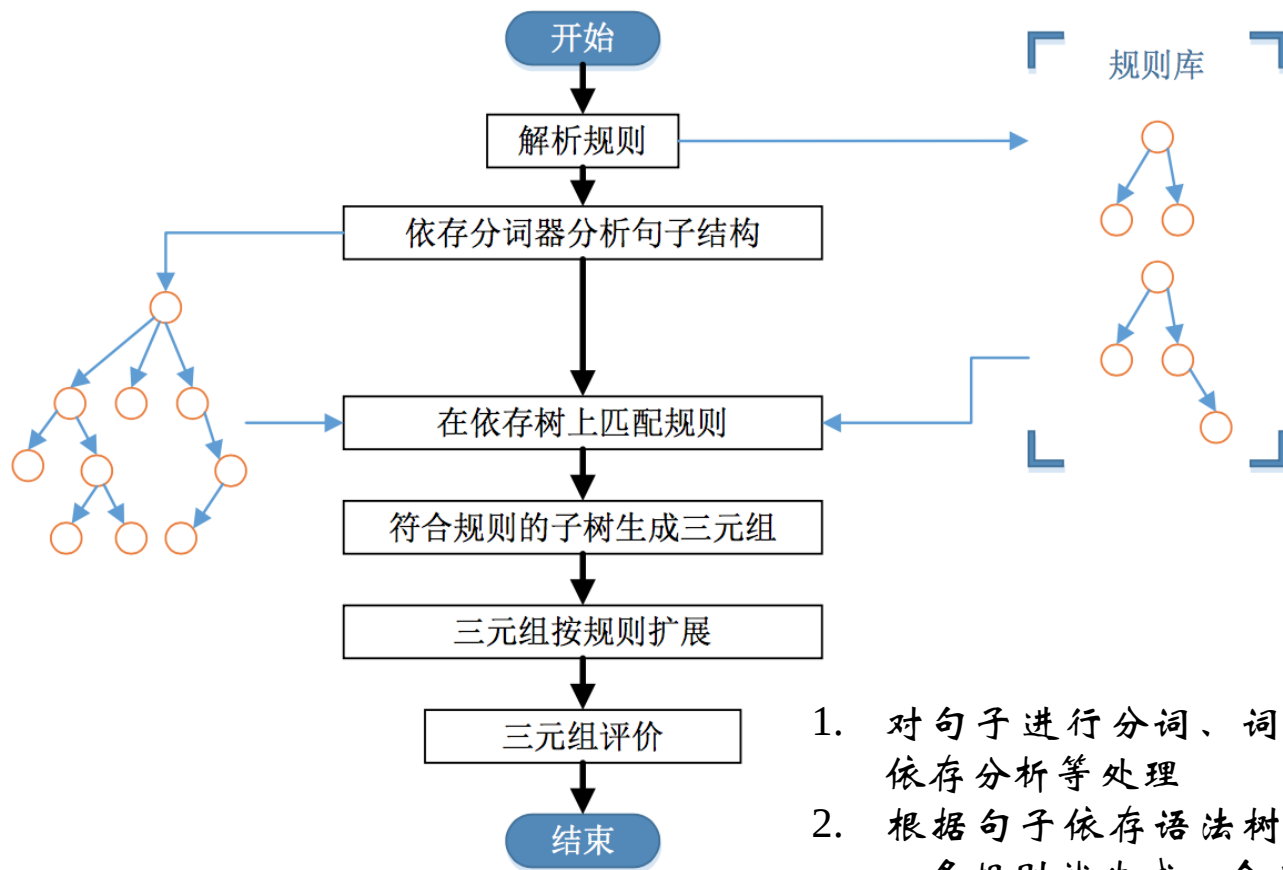
X 配偶 Y

基于模板的方法—基于依存句法分析的Pattern



- 依存句法分析句子的句法结构
- 以动词为起点，构建规则，对节点上的词性和边上的依存关系进行限定

基于模板的方法—基于依存句法分析的Pattern



1. 对句子进行分词、词性标注、命名实体识别、依存分析等处理
2. 根据句子依存语法树结构上匹配规则，每匹配一条规则就生成一个三元组
3. 根据扩展规则对抽取到的三元组进行扩展
4. 对三元组实体和触发词进一步处理抽取出关系

基于模板的方法——基于依存句法分析的Pattern

董卿现身国家博物馆看展优雅端庄大方

依存分析结果

词顺序	词	词性	依存关系路径	依存关系
0	董卿	人名	1	定语
1	现身	动词	-1	核心词
2	国家博物馆	地名	1	宾语
3	看	动词	1	顺承
4	展	动词	3	补语
5	优雅	形容词	7	定语
6	端庄	形容词	7	定语
7	大方	形容词	4	宾语

规则抽取结果

(董卿, 现身, 国家博物馆)  位于(董卿, 国家博物馆)

基于模板的方法——优劣

□ 优点

- 在小规模数据集上容易实现
- 构建简单

□ 缺点

- 特定领域的模板需要专家构建
- 难以维护
- 可移植性差
- 规则集合小的时候，召回率很低

监督学习

确定实体对的情况下，根据句子上下文对实体关系进行预测，构建一个监督学习应该怎么做？

- 预先定义好关系的类别
- 人工标注一些数据
- 设计特征表示
- 选择一个分类方法 (SVM、NN、Naive Bayes)
- 评估结果

监督学习-特征

□ 轻量级特征

- 实体前后的词
- 实体的类型
- 实体之间的距离

□ 中等量级特征

- Chunk序列

□ 重量级特征

- 实体间的依存关系路径
- 实体间树结构的距离
- 特定的结构信息

监督学习-特征

机器学习方法特征设计

Example 117: Forward [motion] _{E1} of the vehicle through the air caused a [suction] _{E2} on the road draft tube.	
Feature Set	Feature Values
Lexical	e1Word=motion, e2Word=suction, e1OrE2Word={motion,suction}, wordsBetween={of, the, vehicle, through, the, air, caused, a}, posE1=NN, posE2=NN, posE1orE2=NN posBetween=I_D_N_I_D_N_V_D, distance=8, wordsOutside={Forward, on}, prefix5Between={air, cause, a, of, the, vehic, throu, the}
Dependency	depPathLen1={caused→nsubj→<E1>, caused→dobj→<E2>,...} depPathLen1VerbNet={vn:27→nsubj→<E1>, vn:27→dobj→<E2>,...} depPathLen2VerbNet={<E1>←nsubj←vn:27→dobj→<E2>}, depPathLen2Location={<E1>←nsubj←BETWEEN→dobj→<E2>}
PropBank	pbPredStem=caus, pbVerbNet=27, pbE1CoarseRole=ARG0, pbE2CoarseRole=ARG1, pbE1orE2CoarseRole={ARG1,ARG2}, pbNumPredToks=1, pbE1orE2PredHyper={cause#v#1, create#v#1}
FrameNet	fnAnyLU={cause.v, vehicle.n, road.n}, fnAnyTarget={cause,vehicle,road}, fnE2LU=cause.v, fnE1OrE2LU=cause.v
Hypernym	hyperE1={gesture#n#2, communication#n#2, entity#n#1, ...}, hyperE2={suction#n#1, phenomenon#n#1, entity#n#1,...}, hyperE1orE2={gesture#n#2, communication#n#2, entity#n#1, suction#n#1, phenomenon#n#1, ...}, hyperBetween={quality#n#1, cause#v#1, create#v#1, ...}
NomLex-Plus	<i>Features did not fire</i>
NGrams	knnE1={motion, amendment, action, appeal, decision}, knnE2={suction, hose, pump, vacuum, nozzle}, knnE1Role=Message, knnE2Role=Component
TextRunner	trE1_E2={may result from, to contact, created, moves, applies, causes, falls below, corresponds to which}, trE2_E1={including, are moved under, will cause, according to, are effected by, repeats, can match}, trE1_E2Hyper={be#v#6, agree#v#3, cause#v#1, ensue#v#1, contact#v#1, apply#v#1, ...}

Bryan Rink et. al ACL 2016

监督学习—特征

深度学习 方法 特征设计

- Position embeddings
- Word embeddings
- Knowledge embeddings
- ...

监督学习—深度学习方法

◆ Pipeline

- 识别实体和关系分类是完全分离的两个过程，不会相互影响，关系的识别依赖于实体识别的效果

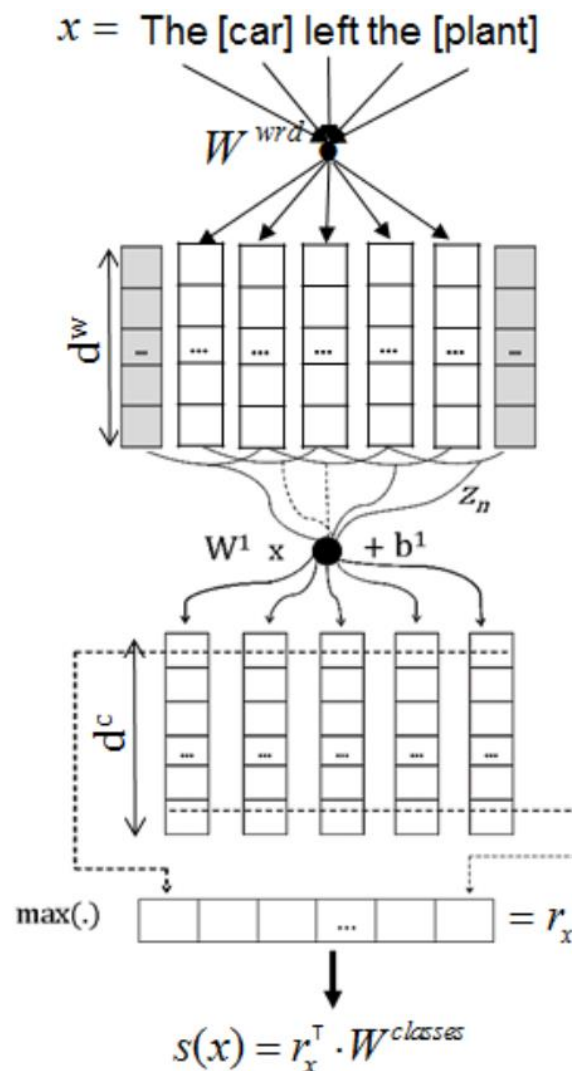
◆ Joint Model

- 实体识别和关系分类的过程是共同优化的

监督学习—深度学习方法

Pipeline方法-CR-CNN模型

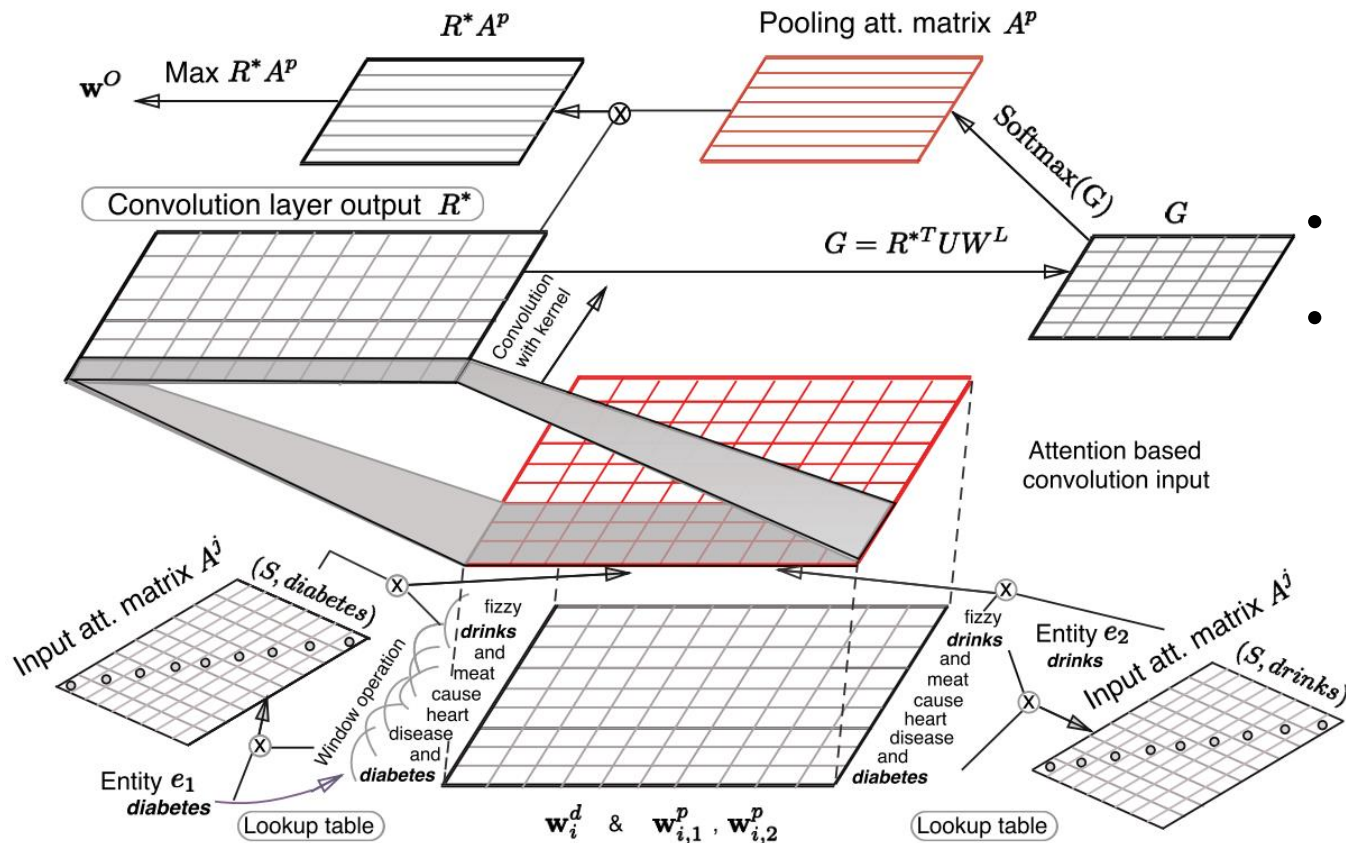
- 仅使用词向量和位置向量作为输入
- F1值84.1，超过目前最好的非深度学习
方法



Santos et. al Computer Science 2015

监督学习—深度学习 方法

Pipeline 方法-Att-CNN模型

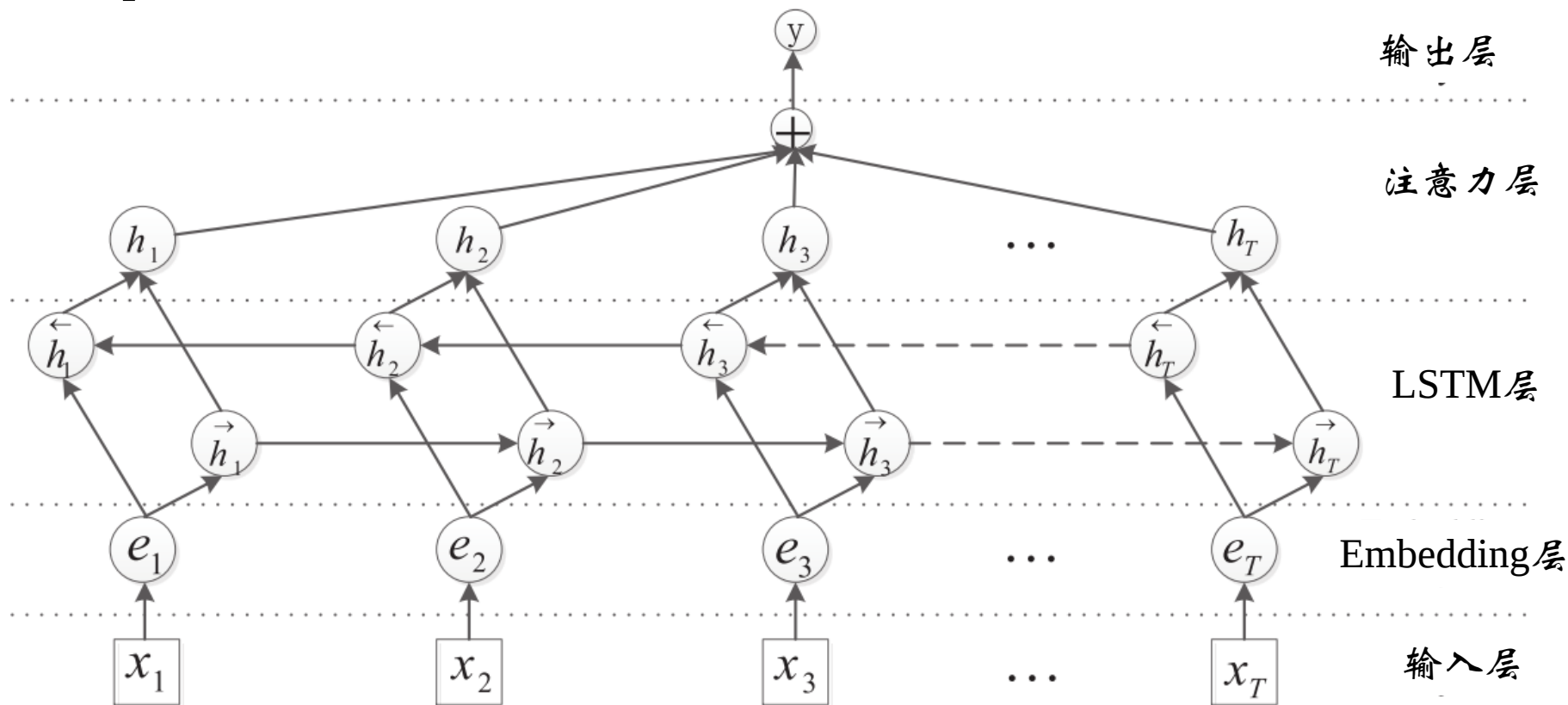


- 应用注意力机制
- 目前最好的方法 (F1值88.0)

Zhiyuan Liu et. al ACL 2016

监督学习—深度学习方法

Pipeline方法-Att-BLSTM模型



Peng Zhou et. al ACL 2016

监督学习-效果对比

Pipeline 方法

Classifier	F1
<i>Manually Engineered Methods</i>	
SVM (Rink and Harabagiu, 2010)	82.2
<i>Dependency Methods</i>	
RNN (Socher et al., 2012)	77.6
MVRNN (Socher et al., 2012)	82.4
FCM (Yu et al., 2014)	83.0
Hybrid FCM (Yu et al., 2014)	83.4
SDP-LSTM (Xu et al., 2015b)	83.7
DRNNs (Xu et al., 2016)	85.8
SPTree (Miwa and Bansal, 2016)	84.5

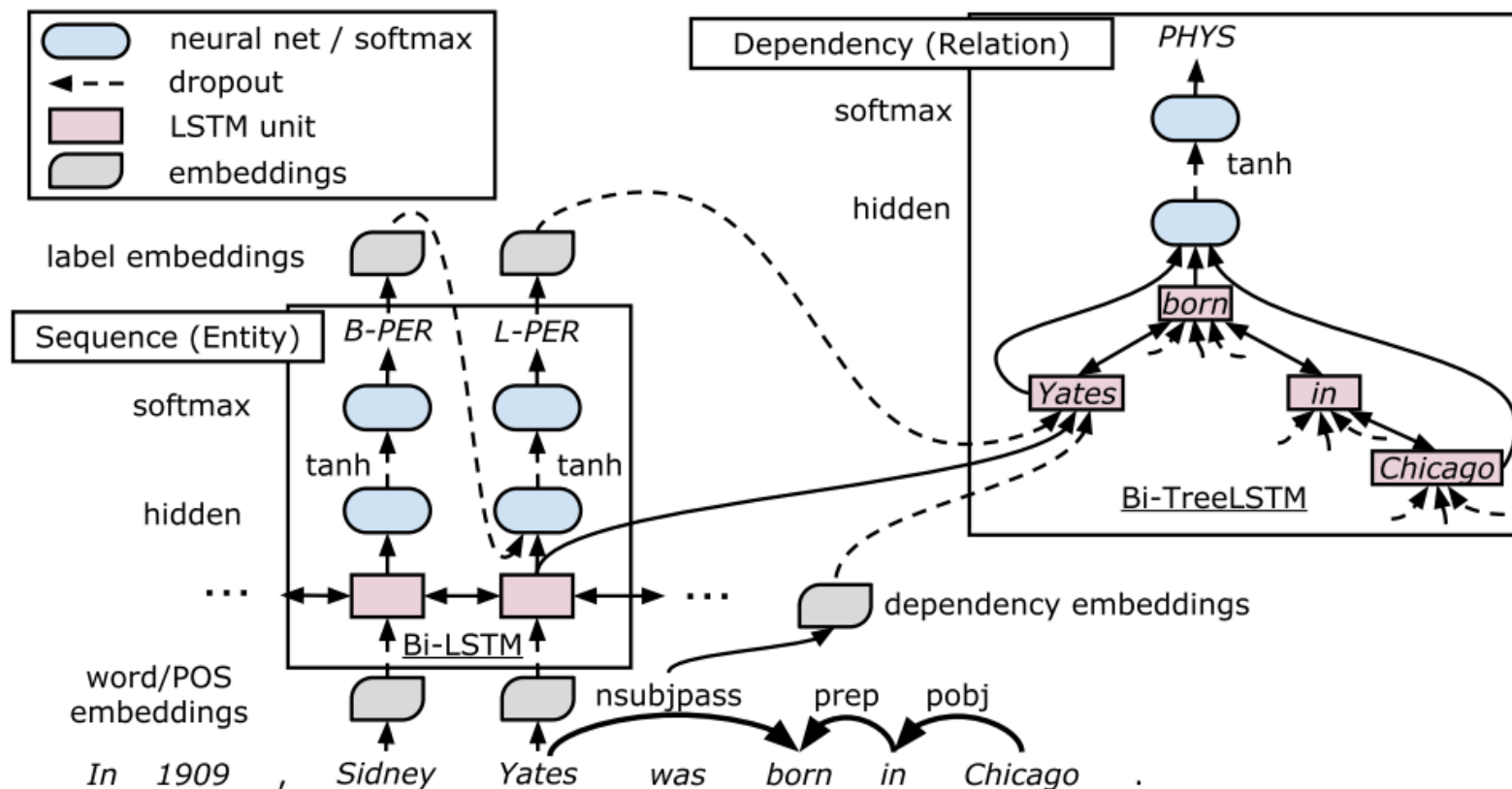
<i>End-To-End Methods</i>	
CNN+ Softmax (Zeng et al., 2014)	82.7
CR-CNN (dos Santos et al., 2015)	84.1
DepNN (Liu et al., 2015)	83.6
depLCNN+NS (Xu et al., 2015a)	85.6
STACK-FORWARD*	83.4
VOTE-BIDIRECT*	84.1
VOTE-BACKWARD*	84.1
<i>Our Architectures</i>	
Att-Input-CNN	87.5
Att-Pooling-CNN	88.0

BLSTM	WV (Turian et al., 2010) (dim=50) + PI	80.7
Att-BLSTM	WV (Turian et al., 2010) (dim=50) + PI	82.5
BLSTM	WV (Pennington et al., 2014) (dim=100) + PI	82.7
Att-BLSTM	WV (Pennington et al., 2014) (dim=100) + PI	84.0

不同模型在SemEval-2010 Task 8 数据集上的效果比较

监督学习—深度学习方法

Joint Model - LSTM-RNNs 模型



Miwa et. al ACL 2016

监督学习—深度学习方法

Joint Model

Settings	Macro-F1
No External Knowledge Resources	
Our Model (SPTree)	0.844
dos Santos et al. (2015)	0.841
Xu et al. (2015a)	0.840
+WordNet	
Our Model (SPTree + WordNet)	0.855
Xu et al. (2015a)	0.856
Xu et al. (2015b)	0.837

不同模型在SemEval-2010 Task 8 dataset 上的效果比较

监督学习—优劣

□ 优点

- 准确率高，标注数据越多越准确

□ 缺点

- 标注数据成本太高
- 不能扩展新的关系

半监督学习

- 没有足够多标注数据的情况下，怎么办？
- 数据量特别大的情况下，如何抽取实体间关系？

➤ 远程监督方法

知识库与非结构化文本对齐来自动构建大量训练数据，减少模型对人工标注数据的依赖，增强模型跨领域适应能力

➤ Bootstrapping

通过在文本中匹配实体对和表达关系短语模式，寻找和发现新的潜在关系三元组

半监督学习——远程监督

两个实体如果在知识库中存在某种关系，则包含该两个实体的非结构化句子均能表示出这种关系。

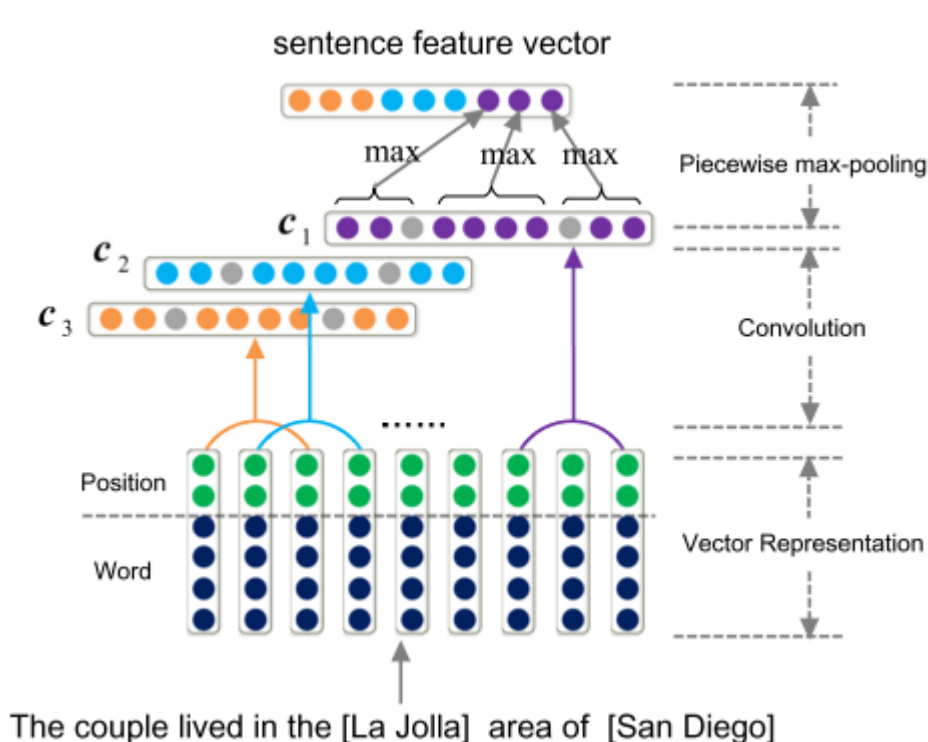
在某知识库中存在： 创始人(乔布斯, 苹果公司)

则可构建训练正例：乔布斯是苹果公司的联合创始人和CEO

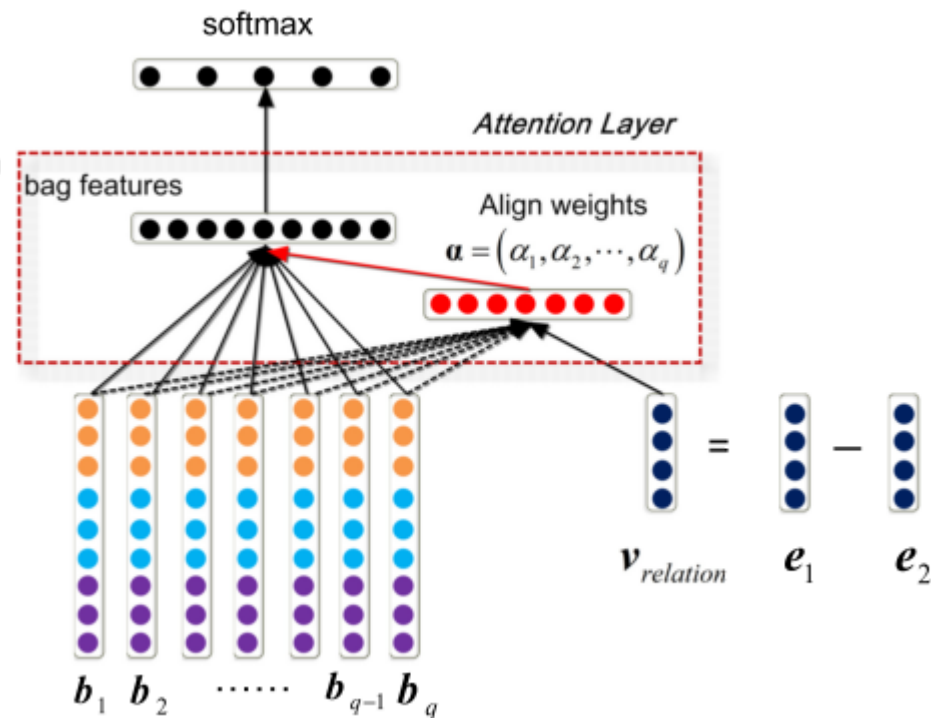
具体步骤

1. 从知识库中抽取存在关系的实体对
2. 从非结构化文本中抽取含有实体对的句子作为训练样例

半监督学习—远程监督



PCNNs 模型



句子级别的Attention

Kang Liu et.al AI 2017

远程监督—优劣

□ 优点

- 可以利用丰富的知识库信息，减少一定的人工标注

□ 缺点

- 假设过于肯定，引入大量噪声，存在语义漂移现象
- 很难发现新的关系

半监督学习—Bootstrapping

给定种子集合，如：<姚明, 叶莉>

1. 从文档中抽取出包含种子实体的新闻，如

- 姚明 老婆 叶莉 简历身高曝光
X 老婆 Y 简历身高曝光
- 姚明 与妻子 叶莉 外出赴约
X 与妻子 Y 外出赴约
- 姚明 携爱妻 叶莉 外出赴约
X 携爱妻 Y 外出赴约

2. 将抽取出的Pattern去文档集中匹配

- 小猪 与妻子 伊万 外出赴约

3. 根据Pattern抽取出的新文档如种子库，迭代多轮直到不符合条件

Bootstrapping—优劣

□ 优点

- 构建成本低，适合大规模构建
- 可以发现新的关系（隐含的）

□ 缺点

- 对初始给定的种子集敏感
- 存在语义漂移问题
- 结果准确率较低
- 缺乏对每一个结果的置信度的计算

关系抽取—参考文献

- Rink, Bryan, and S. Harabagiu. UTD: Classifying semantic relations by combining lexical and semantic resources. *International Workshop on Semantic Evaluation* Association for Computational Linguistics. 2010:256-259
- Santos, Cicero Nogueira Dos, B. Xiang, and B. Zhou. Classifying Relations by Ranking with Convolutional Neural Networks. *Computer Science*. 2015:132-137
- Linlin Wang, Zhu Cao, Gerard de Melo, Zhiyuan Liu. Relation Classification via Multi-Level Attention CNNs. *Meeting of the Association for Computational Linguistics*. 2016:1298-1307
- Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, Bo Xu. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 2016:207-212
- Makoto Miwa, Mohit Bansal. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 2016
- Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao. Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions. *Proceedings of the Thirty-First Conference on Artificial Intelligence*. 2017:3060-3066

事件抽取

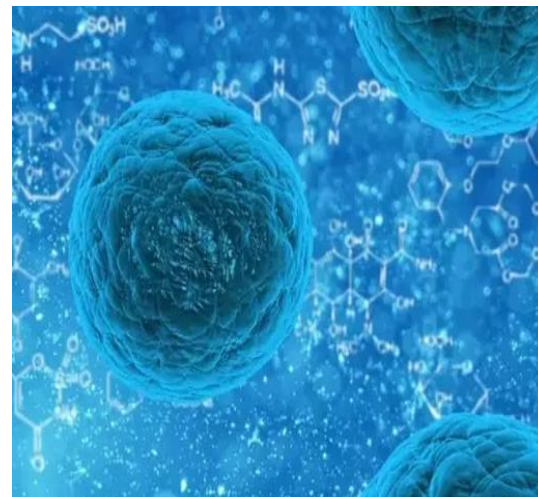
事件是指发生的事情，通常具有时间、地点、参与者等属性，事件的发生可能因为一个动作的产生或者系统状态的改变。



美国攻打伊拉克



股票上涨



细胞病变

事件抽取的定义

从自然语言中抽取出用户感兴趣的事件信息，并以结构化的形式呈现出来，例如事件发生的时间、地点、发生原因、参与着等。

事件抽取的相关术语

- 事件描述 (Event Mention): 描述事件的词组或句子
- 事件触发 (Event Trigger): 表明事件出现的主要词汇
- 事件元素 (Event Argument): 事件的重要信息
- 元素角色 (Argument Role): 元素在句子中的语义角色

事件抽取的定义

mention

trigger

苹果公司将于西部时间9月12日上午10点(北京时间9月13日凌晨1点)举行新品发布会,这一次的发布会地点是全新建造的史蒂夫·乔布斯剧院。根据目前的消息,这次发布会上苹果将会发布iPhone 8(命名不确定,暂且称之为iPhone 8)、iPhone 7s、iPhone 7s Plus、Apple Watch 3以及全新Apple TV。

Argument
role

事件类型	发布会
公司	苹果公司
时间	西部时间9月12日上午10点
地点	史蒂夫·乔布斯剧院
产品	iPhone8、iPhone7s、iPhone7s plus、apple watch 3、apple TV

arguments

更多事件抽取示例

外媒称中印边界对峙开始撤军 披露双方对峙缘由

正文

我来说两句(5961人参与)

扫描二维码 扫描到手机

2014-09-29 08:38:32 来源: 环球时报 作者: 王盼盼

手机看新闻 | 保存到博客 | 图 | 图

【环球军事报道】据《印度快报》27日报道，印度外交部长斯瓦拉吉周五在纽约表示，她与**中国外长王毅**会谈后决定，**中印双方周五**开始从边界撤军，到下周二撤完，双方军队将恢复到9月1日所在的位置，“糟糕的时期即将结束”。《纽约时报》称，至此，持续近三周的中印边界对峙宣告结束，双方兵力



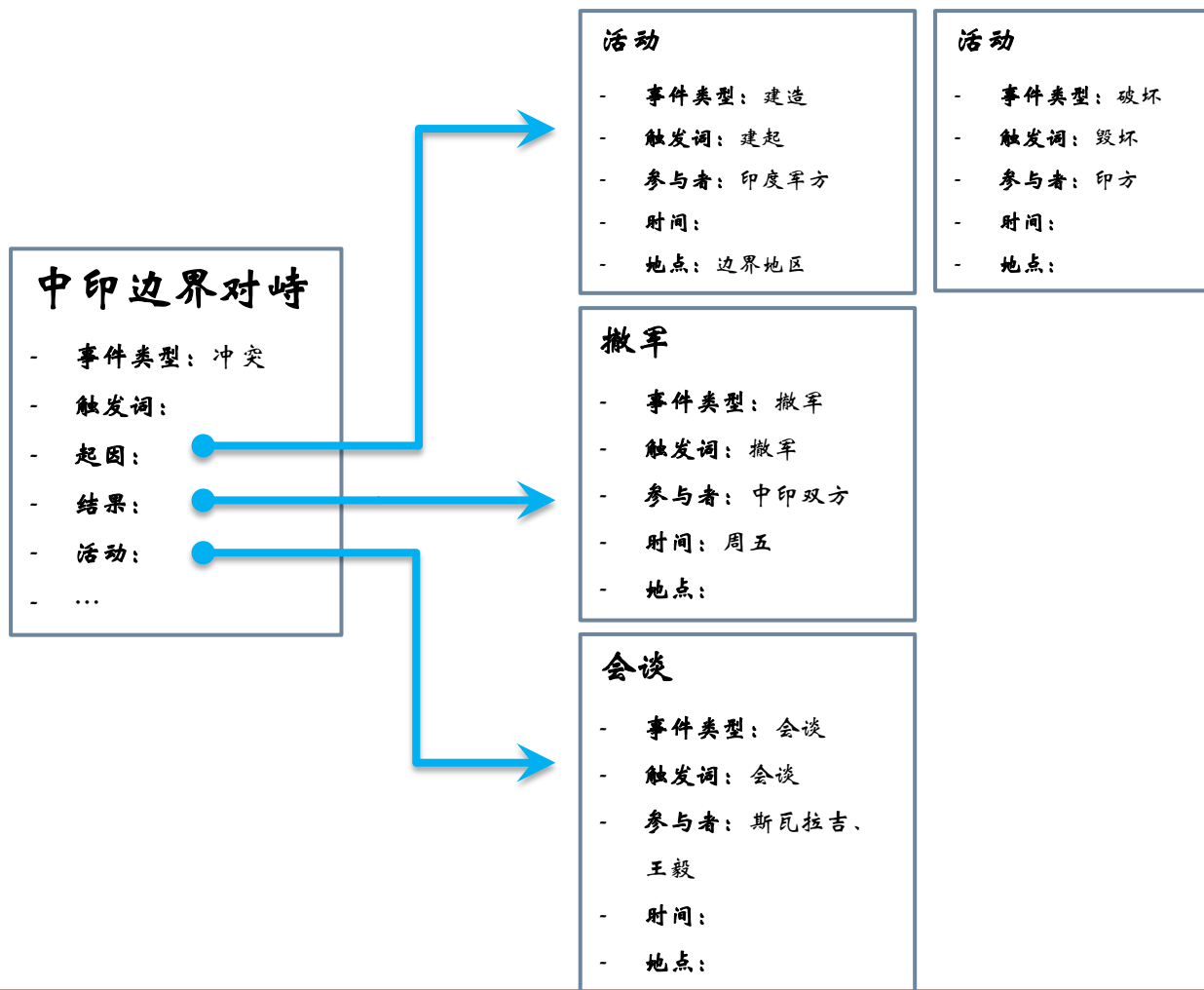
多家西方媒体披露了此次对峙的起因。路透社25日发文称，**本月初**，**印度军方**在**边界地区**建起一个观察屋，从这里可以观察对面中国士兵的举动。这一行动激怒了中方，中国士兵之后开始铺设通往印度声索主权领土的公路，并要求印方拆除观察屋。**印方**拒绝了中国的要求，**毁坏了**中国建造的部分**公路**，并增加了在这一地区的驻兵。“莫迪曾在竞选中承诺更加强硬的国家安全政策，中印上千名士兵在边界地区对峙说明，即便是强势的中国，也并非无所不能。”《纽约时报》26日称，拉达克地区官员坚持称，观察屋仅用于民用，但建设活动已经停止。



简单事件抽取

事件类型	会谈
触发词	会谈
参与者	斯瓦拉吉、王毅
时间	-
地点	-
事件类型	撤军
触发词	撤军
参与者	中印双方
时间	周五
事件类型	建造
触发词	建起
参与者	印度军方
时间	-
事件类型	破坏
触发词	毁坏
参与者	印方
时间	-
地点	-

事件嵌套



事件抽取任务

事件抽取任务最基础的部分包括：

- 识别事件触发词及事件类型
- 抽取事件元素同时判断其角色
- 抽出描述事件的词组或句子

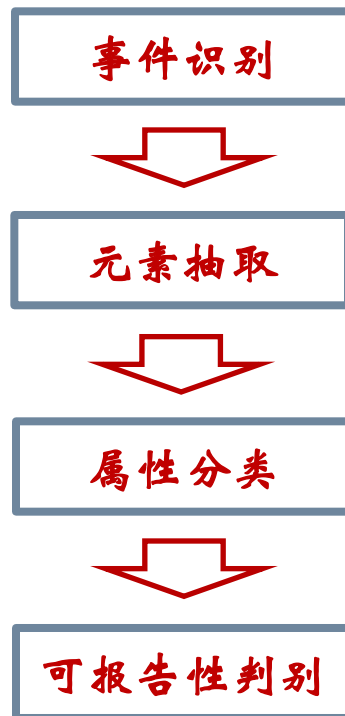
此外，事件抽取任务还包括：

- 事件属性标注
- 事件共指消解

事件抽取的pipeline方法

有监督的事件抽取方法的标准流程一种pipeline的方法，将事件抽取任务转化为多阶段的分类问题，需要的分类器包括：

- 事件触发次分类器 (Trigger Classifier)
 - 用于判断词汇是否是事件触发词，以及事件的类别
- 元素分类器 (Argument Classifier)
 - 判别词组是否是事件的元素
- 元素角色分类器 (Role Classifier)
 - 判定元素的角色类别
- 属性分类器 (attribute classifier)
 - 判定事件的属性
- 可报告性分类器 (Reportable-Event Classifier)
 - 判定是否存在值得报告的事件实例



分类器模型可以是机器学习方法中的各种分类器模型，比如MaxEnt、SVM等。
(Ahn, ARTE 2006) (Ji and Grishman, 2008, ACL 2008)

典型的分类特征

■ 触发词标注

■ 词汇

- Trigger单词和上下文单词的Token和POS标签

■ 字典

- Trigger列表、同义词字典

■ 句法

- Trigger在句法树中的深度
- Trigger到句法树根节点的路径
- 由Trigger的父节点展开的词组结构
- Trigger的词组类型

■ 实体

- 句法上距离trigger最近的实体的类型
- 句子中距离trigger物理距离最近的实体的类型

■ 元素标注

■ 事件类型和触发词

- Trigger的token
- 事件类型和子类型

■ 实体

- 实体类型和子类型
- Head word of the entity mention

■ 上下文

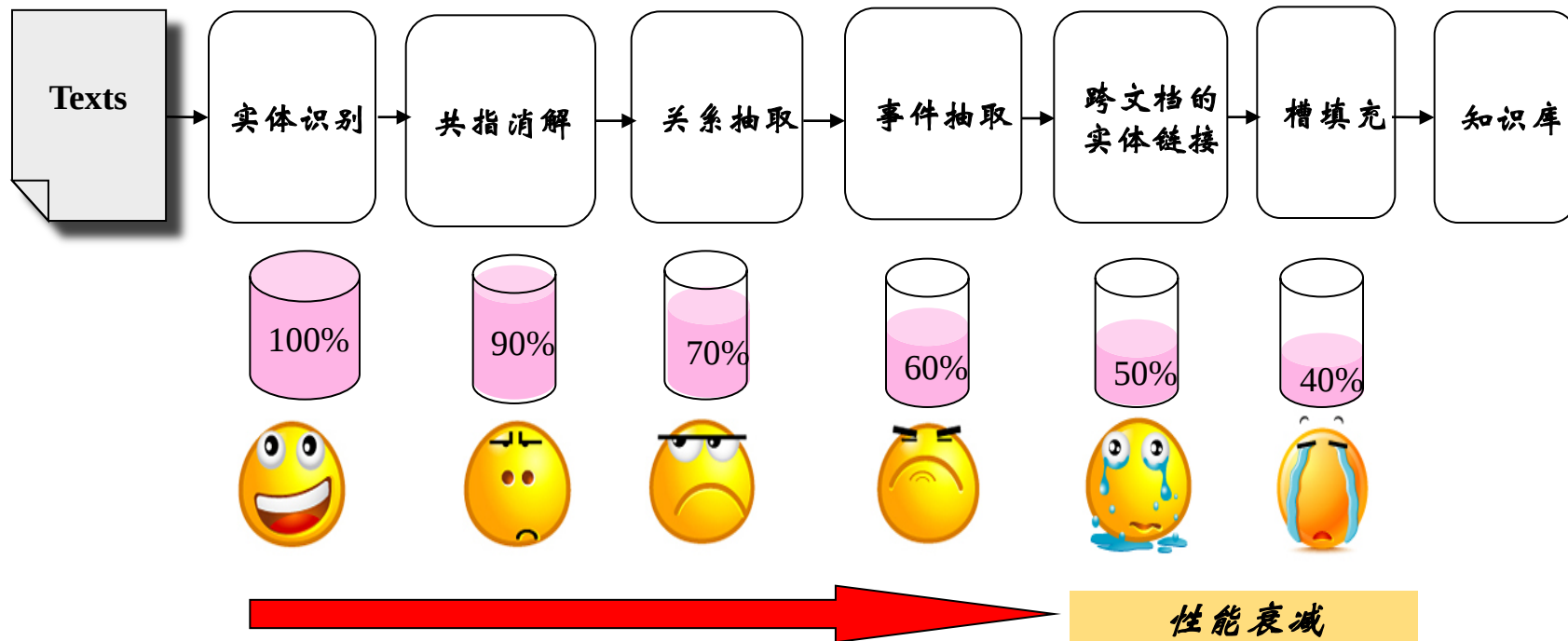
- 候选元素的上下文单词

■ 句法

- 扩展trigger父节点的词组结构
- 实体和trigger的相对位置(前或后)
- 实体到触发词的最短路径
- 句法树中实体到触发词的最短长度

事件抽取的联合抽取方法

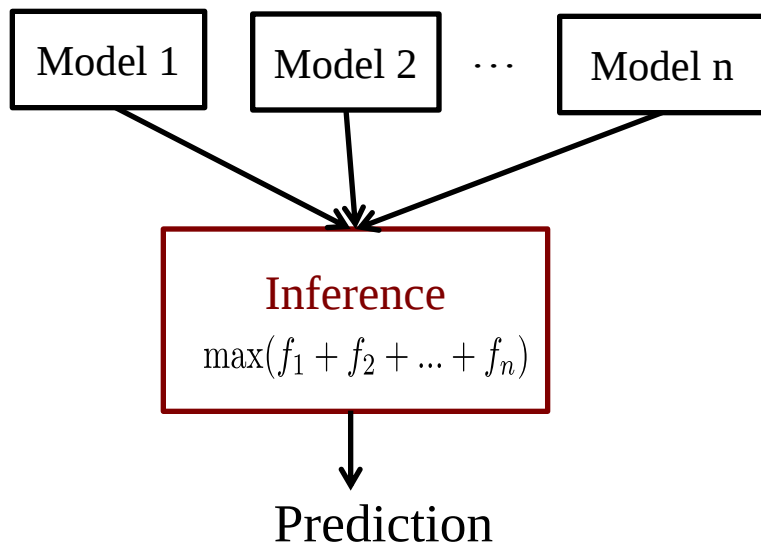
● 传统的Pipelined方法存在的问题



- 误差从前面的环节传播到后面的环节，使得性能急剧衰减
- 各个环节的预测任务是独立的，之间没有互动
- 无法处理全局的依赖关系

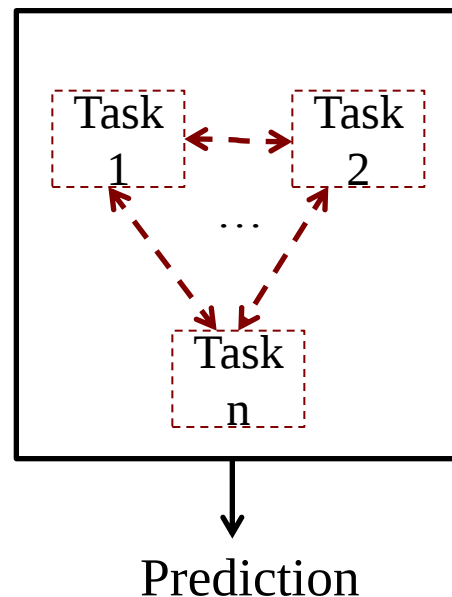
一般的解决方法：

Joint Inference



- Constrained Conditional Models
- ILP Re-ranking
- Dual decomposition

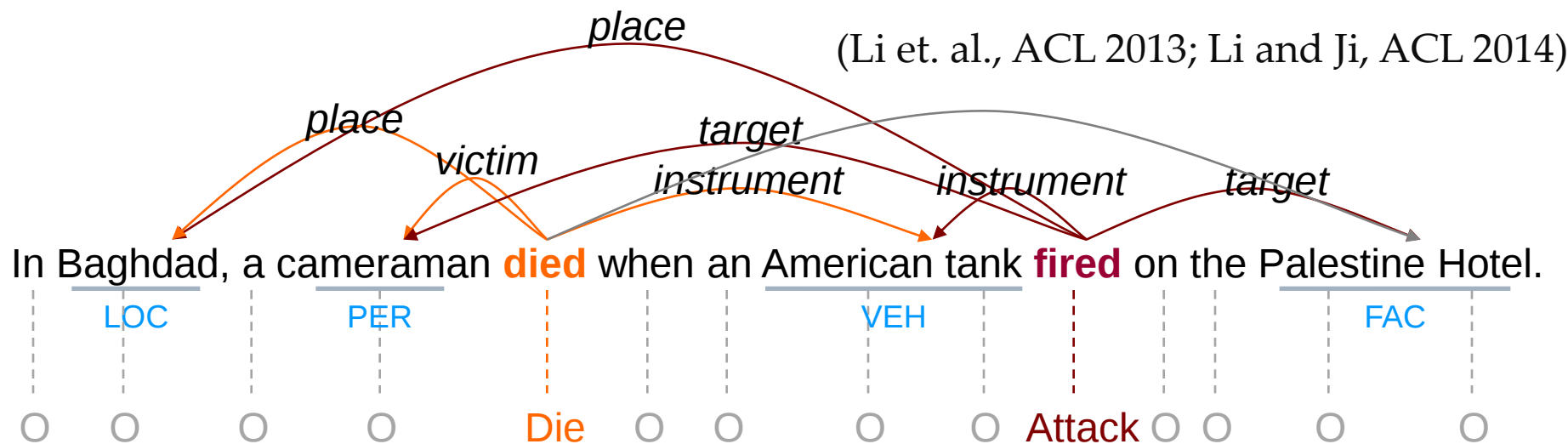
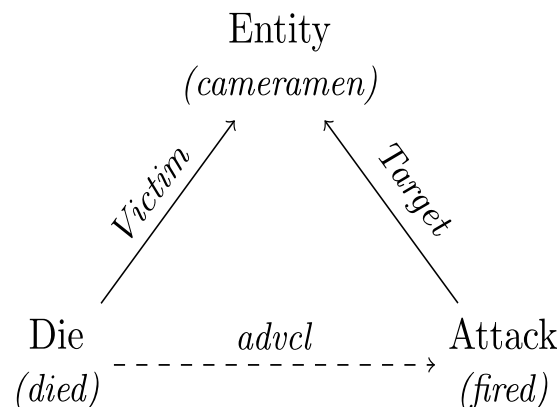
Joint Modeling



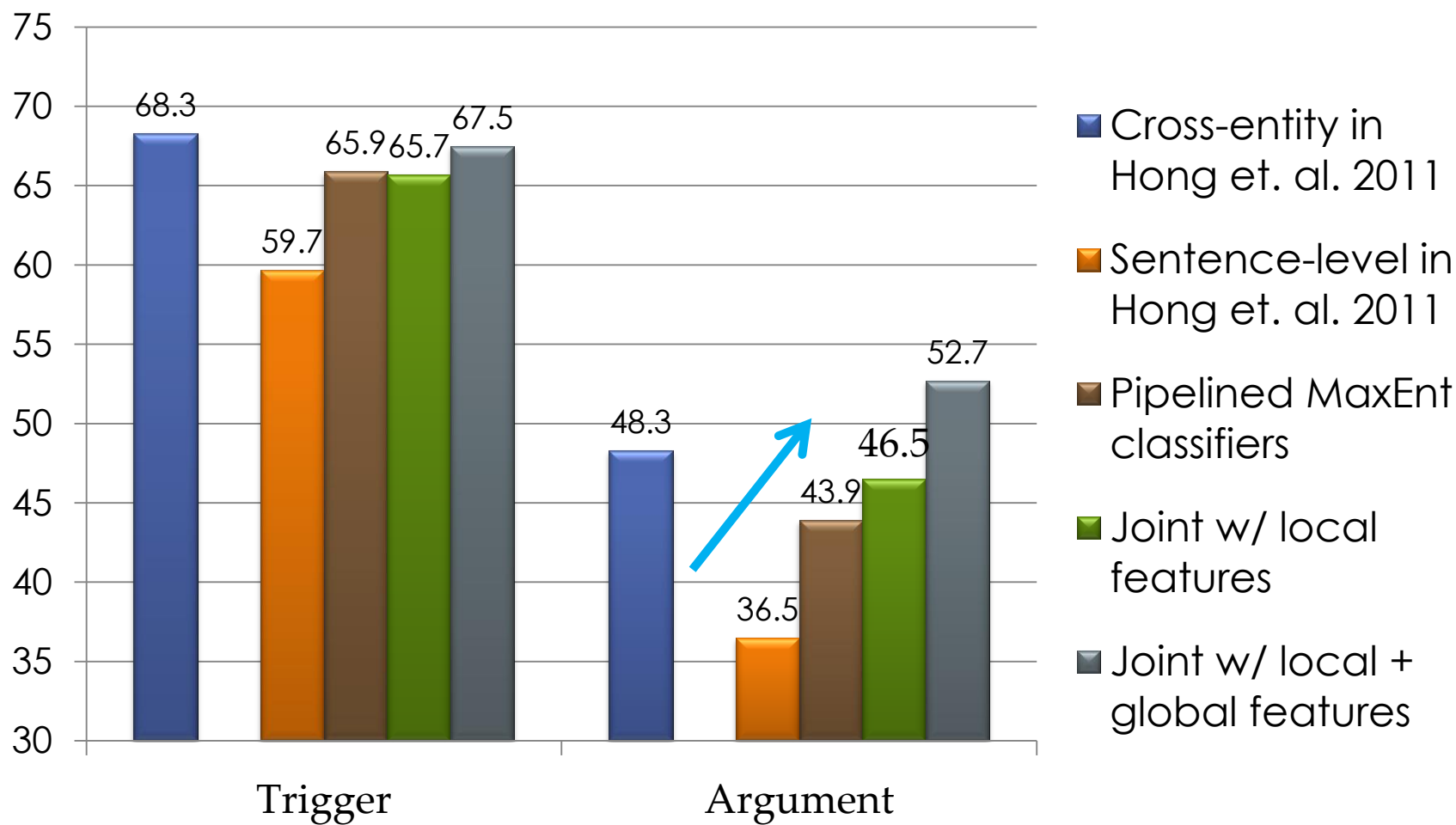
- Probabilistic Graphical Models
- Markov logic networks

Joint Modeling with Structured Prediction

- 使用一个模型同时抽取出所有的信息的联合。
- 将问题建模成结构预测问题，使用搜索方法进行求解。
- 避免了误差传播导致的性能下降。
- 全局特征可以从整体的结构中学习到，从而使用全局的信息来提升局部的预测。



几种方法的Trigger和argument抽取效果



基于深度学习的事件抽取方法

■ 传统方法的缺陷：

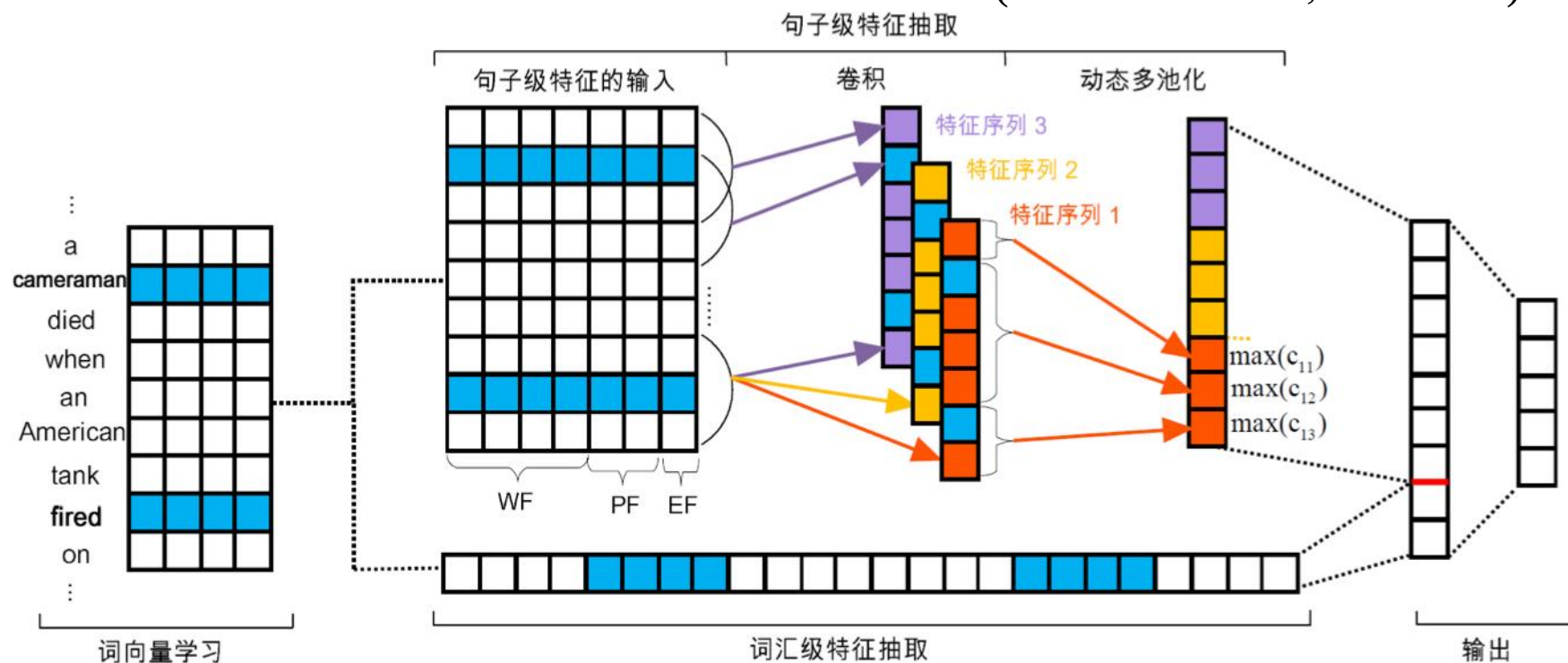
- 需要借助外部的NLP工具
 - 导致了误差的累积、传播
 - 有些语言或者领域缺少NLP工具
- 需要人工设计特征

■ 深度学习方法的优势：

- 减少了对外部NLP工具的依赖，甚至不依赖NLP工具，建模成端对端的系统。
- 使用词向量作为输入，词向量蕴含了丰富的语言特征。
- 神经网络具有自动提取句子特征的能力，避免了人工设计特征的繁琐工作。

基于动态多池化卷积神经网络的事件抽取方法

(Yubo Chen et. al., ACL 2015)



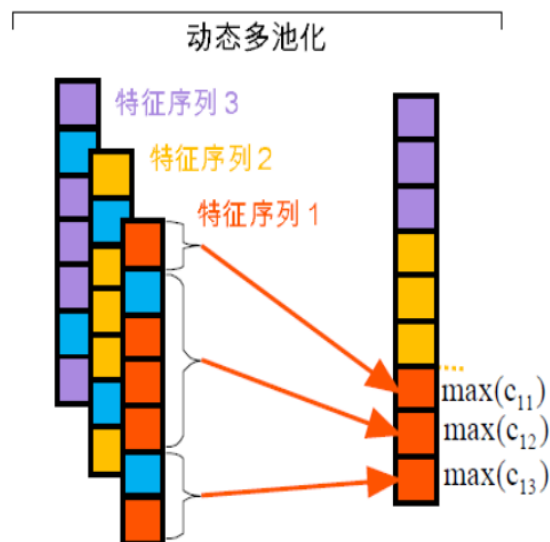
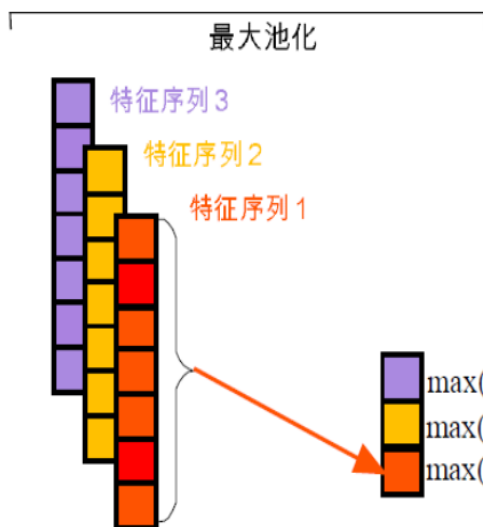
系统主要包括四个部分:

1. 词向量学习; 2. 词汇级别特征提取; 3. 句子级别特征提取; 4. 分类输出

动态多池化层：

- 传统的神经网络利用最大池化层
- 事件抽取中，一个句子可能包含多个事件

S: In Baghdad, a cameraman died when an American tank fired on the Palestine Hotel.



实验结果

- 测试集：ACE2005英文数据集的篇新闻
- 开发集：ACE2005英文数据集的30篇标注文档
- 训练集：剩下的529篇标注文档

Methods	Trigger Identification(%)			Trigger Identification + Classification(%)			Argument Identification(%)			Argument Role(%)		
	P	R	F	P	R	F	P	R	F	P	R	F
Li's baseline	76.2	60.5	67.4	74.5	59.1	65.9	74.1	37.4	49.7	65.4	33.1	43.9
Liao's cross-event	N/A			68.7	68.9	68.8	50.9	49.7	50.3	45.1	44.1	44.6
Hong's cross-entity	N/A			72.9	64.3	68.3	53.4	52.9	53.1	51.6	45.5	48.3
Li's structure	76.9	65.0	70.4	73.7	62.3	67.5	69.8	47.9	56.8	64.7	44.4	52.7
DMCNN model	80.4	67.7	73.5	75.6	63.6	69.1	68.8	51.9	59.1	62.2	46.9	53.5

DMCNN与传统方法的比较

Stage	Method	1/1	1/N	all
		F_1	F_1	F_1
Trigger	Embedding+T	68.1	25.5	59.8
	CNN	72.5	43.1	66.3
	DMCNN	74.3	50.9	69.1
Argument	Embedding+T	37.4	15.5	32.6
	CNN	51.6	36.6	48.9
	DMCNN	54.6	48.7	53.5

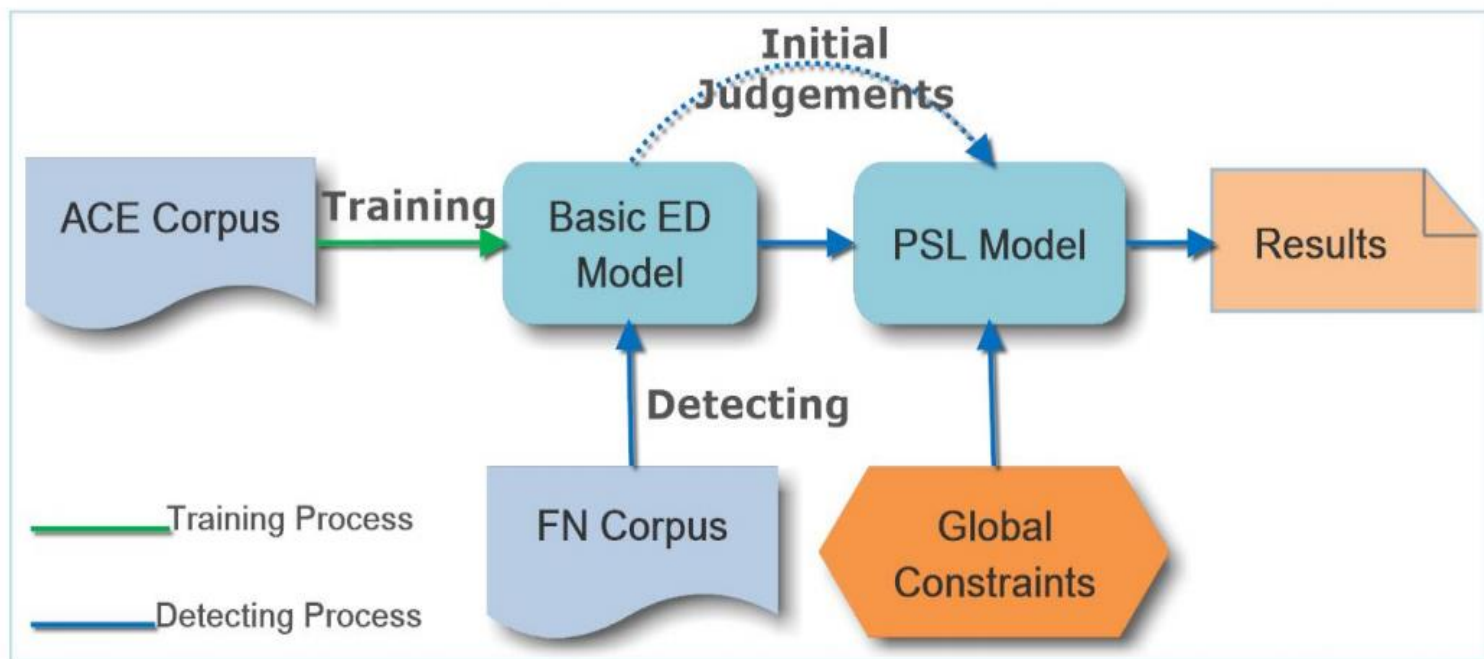
DMCNN与其他CNN方法的比较

扩充语料的方法

使用FrameNet扩展语料 Shulin Liu, ACL 2016

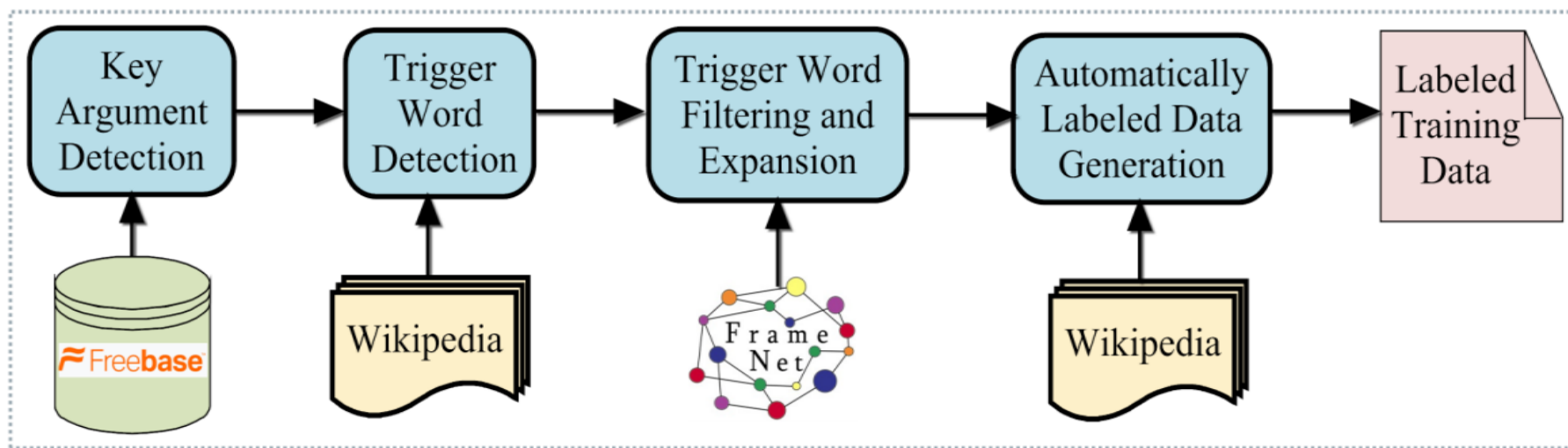
FrameNet

- 语言学家定义及标注的语义框架资源
- 层级的组织结构
- 1000+ 框架、10000+ 词法单元、150000+ 标注例句



系统框架图

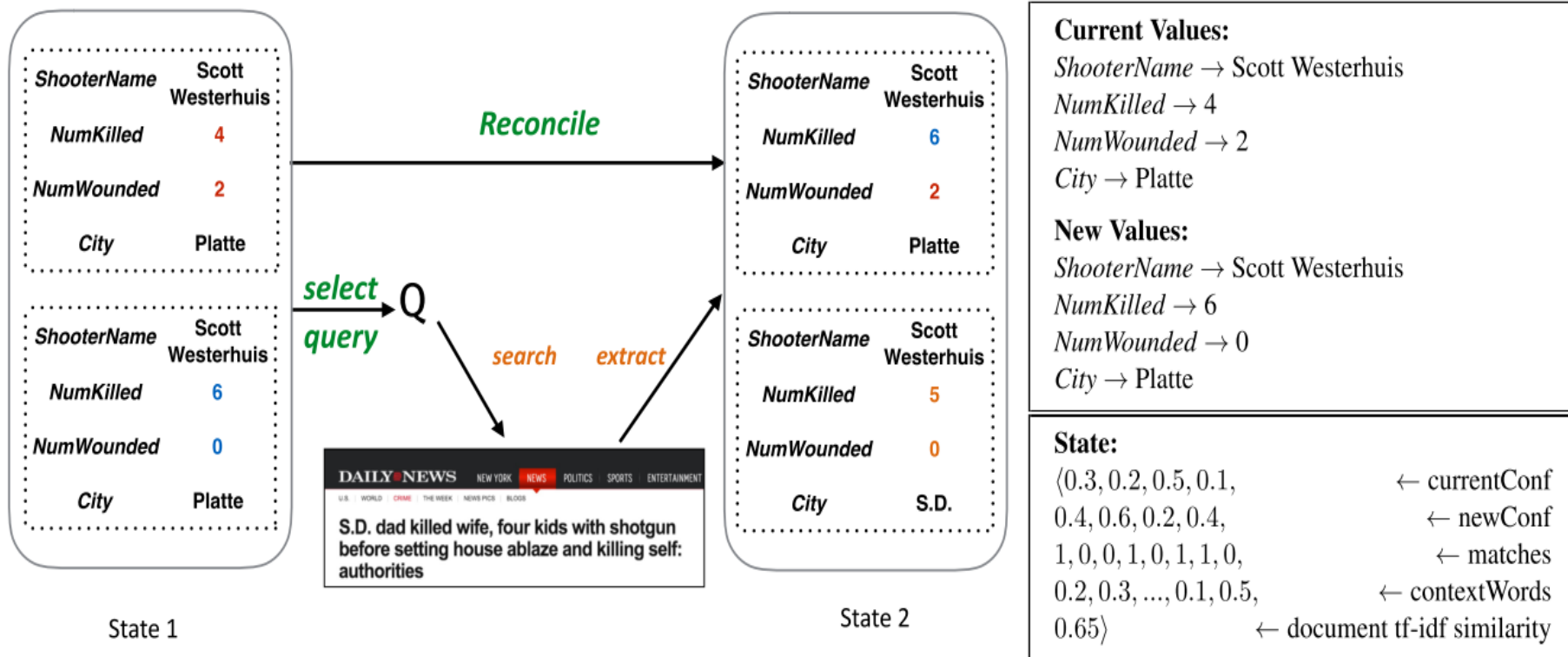
运用结构化的知识库自动生成语料 Yubo Chen, ACL 2017



- 利用世界知识和语言知识
- 自动生成大规模事件语料

从网络获取事件信息

(Karthik Narasimhan et. al., EMNLP 2016)



- 从网络获取同一事件的不同报道
- 使用强化学习方法，做信息融合的决策

参考文献

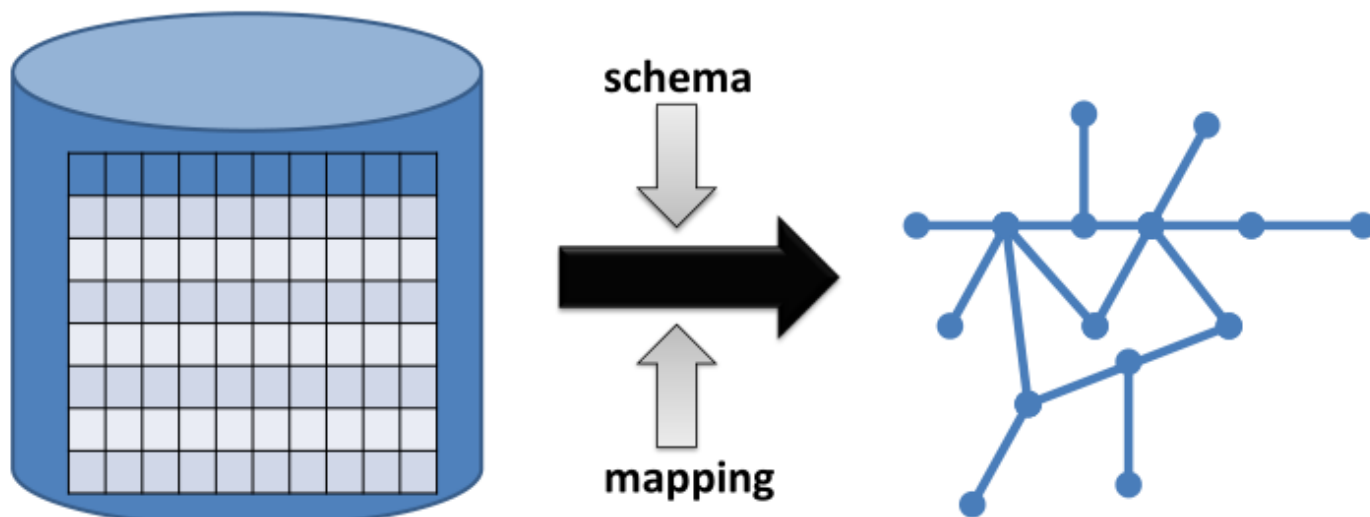
- Ahn, D. (University of A. (2006). The stages of event extraction. ARTE '06 Proceedings of the Workshop on Annotating and Reasoning about Time and Events, (July), 1–8. <https://doi.org/10.3115/1629235.1629236>
- Ji H, Grishman R. Refining Event Extraction Through Cross-document Inference[C]// ACL 2008, Proceedings of the, Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, Usa. DBLP, 2008:254-262.
- Li Q, Ji H, Huang L. Joint Event Extraction via Structured Prediction with Global Features[C]// Meeting of the Association for Computational Linguistics. 2013:73-82.
- Li Q, Ji H. Incremental Joint Extraction of Entity Mentions and Relations[C]// Meeting of the Association for Computational Linguistics. 2014:402-412.
- Chen Y, Xu L, Liu K, et al. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks[C]// The, Meeting of the Association for Computational Linguistics. 2015.

-
- Liu S, Chen Y, He S, et al. Leveraging FrameNet to Improve Automatic Event Detection[C]// Meeting of the Association for Computational Linguistics. 2016:2134-2143.
 - Narasimhan CSAIL, K., Yala, A., & Barzilay, R. (2016). Improving Information Extraction by Acquiring External Evidence with Reinforcement Learning. Emnlp, 2355–2365.
 - Chen, Y., Xu, L., Liu, K., Zeng, D., & Zhao, J. (2015). Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. Proceedings ACL 2015, 167–176.

目录

- 面向非结构化数据的知识抽取
- **面向结构化数据的知识抽取**
- 面向半结构化数据的知识抽取
- 实践：基于百科数据的知识抽取

垂直领域的知识往往来源于支撑企业业务系统的关系数据库



R2RML

a standard transformation of a
relational database in RDF

标准与工具

Mapping languages

- Standards by RDB2RDF working group (W3C)
 - Direct Mapping
 - R2RML
- Proprietary

Tools

Free: D2R, Virtuoso, Morph, r2rml4net, db2triples, ultrawrap, Quest

- Commercial: Virtuoso, ultrawrap, Oracle SW

Direct Mapping Idea

People

PK		→ Addresses(ID)
ID	fname	addr
7	Bob	18
8	Sue	NULL

Addresses

PK		
ID	City	State
18	Cambridge	Ma

Convert relational data into RDF, by making explicit the semantics encoded in the relational schema.

Create URIs following some simple rules: Map

- ☐ table to class
- ☐ column to property
- ☐ row to resource
- ☐ cell to literal value
- ☐ in addition cell to URI
 - if there is a foreign key constraint

Direct Mapping: Consequences

People

PK		→ Addresses(ID)
ID	fname	addr
7	Bob	18
8	Sue	NULL

Addresses

PK		
ID	City	State
18	Cambridge	Ma

We need IRIs for identifying

- ❑ the resource **class** corresponding to a **table**
- ❑ the **resources** represented by the **table rows**
- ❑ the **properties** of the resources corresponding to **table cells**
- ❑ the references due to **foreign keys**

Direct Mapping: Encoding Principles

People

PK		→ Addresses(ID)
ID	fname	addr
7	Bob	18
8	Sue	NULL

Addresses

PK		
ID	City	State
18	Cambridge	Ma

- ❑ **Base IRI** for the whole graph/dataset, e.g.
@base <http://foo.example/DB/> .
- ❑ Table name → Class name, e.g.
People → <People>
- ❑ Row with PK → Resource with PK, e.g.,
<People/ID=7>
- ❑ Table row → Property, e.g.,
<People#ID>
<People#addr>
- ❑ Table cells: what if NULL?
- ❑ Foreign key reference → additional property, e.g.,
<People#ref-addr>

Example: RDF Resulting from Direct Mapping

People

PK		→ Addresses(ID)
ID	fname	addr
7	Bob	18
8	Sue	NULL

Addresses

PK		
ID	City	State
18	Cambridge	Ma

Provide a base IRI `http://foo.example/DB/` !

@base <http://foo.example/DB/> .

@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<People/ID=7> rdf:type <People> .

<People/ID=7> <People#ID> 7 .

<People/ID=7> <People#fname> "Bob" .

<People/ID=7> <People#addr> 18 .

<People/ID=7> <People#ref-addr> <Addresses/ID=18> .

<People/ID=8> rdf:type <People> .

<People/ID=8> <People#ID> 8 .

<People/ID=8> <People#fname> "Sue" .

<Addresses/ID=18> rdf:type <Addresses> .

<Addresses/ID=18> <Addresses#ID> 18 .

<Addresses/ID=18> <Addresses#city> "Cambridge" .

<Addresses/ID=18> <Addresses#state> "MA" .

R2RML概述

R2RML is a language for specifying **mappings** from relational to RDF data.

A mapping takes as input a **logical table**, i.e.,

- ❑ a database table
- ❑ a database view, or
- ❑ an SQL query

(called an “R2RML view” because it is like an SQL view but does not modify the database)

A logical table is mapped to a set of triples by a rule called

- ❑ **triples map.**

<http://www.w3.org/TR/r2rml/>

Triples Maps

Idea: triples are produced by:

- ☐ subject maps
- ☐ predicate maps
- ☐ object maps.

EMP			
EMPNO	ENAME	JOB	DEPTNO
INTEGER PRIMARY KEY	VARCHAR(100)	VARCHAR(20)	INTEGER REFERENCES DEPT (DEPTNO)
7369	SMITH	CLERK	10

Example

- ☐ The subject IRI is generated from the empno column by the [template](#)
`http://data.example.com/employee/{empno}`
- ☐ The predicate IRI is the [constant](#) `ex:name`
- ☐ The object is the [literal](#) "SMITH", that is copied from the ENAME column

示例

EMP

EMPNO INTEGER PRIMARY KEY	ENAME VARCHAR(100)	JOB VARCHAR(20)	DEPTNO INTEGER REFERENCES DEPT (DEPTNO)
7369	SMITH	CLERK	10

Relational tables

DEPT

DEPTNO INTEGER PRIMARY KEY	DNAME VARCHAR(30)	LOC VARCHAR(100)
10	APPSERVER	NEW YORK

Set of RDF triples

```
<http://data.example.com/employee/7369> rdf:type ex:Employee.  
<http://data.example.com/employee/7369> ex:name "SMITH".  
<http://data.example.com/employee/7369> ex:department <http://data.example.com/department/10>.  
  
<http://data.example.com/department/10> rdf:type ex:Department.  
<http://data.example.com/department/10> ex:name "APPSERVER".  
<http://data.example.com/department/10> ex:location "NEW YORK".  
<http://data.example.com/department/10> ex:staff 1.
```

数据库表映射

EMP

EMPNO	ENAME	JOB	DEPTNO
INTEGER PRIMARY KEY	VARCHAR(100)	VARCHAR(20)	INTEGER REFERENCES DEPT (DEPTNO)
7369	SMITH	CLERK	10

DEPT

DEPTNO	DNAME	LOC
INTEGER PRIMARY KEY	VARCHAR(30)	VARCHAR(100)
10	APPSERVER	NEW YORK

结果

```
<http://data.example.com/employee/7369> rdf:type ex:Employee.  
<http://data.example.com/employee/7369> ex:name "SMITH".  
<http://data.example.com/employee/7369> ex:department  
    <http://data.example.com/department/10>.
```

```
<http://data.example.com/department/10> rdf:type ex:Department.  
<http://data.example.com/department/10> ex:name "APPSERVER".  
<http://data.example.com/department/10> ex:location "NEW YORK".  
<http://data.example.com/department/10> ex:staff 1.
```

```
@prefix rr: <http://www.w3.org/ns/r2rml#>.
```

```
@prefix ex: <http://example.com/ns#>.
```

```
<#TriplesMap1>
```

```
  rr:logicalTable [ rr:tableName "EMP" ];
```

```
  rr:subjectMap [
```

```
    rr:template "http://data.example.com/employee/{EMPNO}";
```

```
    rr:class ex:Employee;
```

```
  ];
```

```
  rr:predicateObjectMap [
```

```
    rr:predicate ex:name;
```

```
    rr:objectMap [ rr:column "ENAME" ];
```

```
  ].
```

R2RML Views

EMP

EMPNO	ENAME	JOB	DEPTNO
INTEGER PRIMARY KEY	VARCHAR(100)	VARCHAR(20)	INTEGER REFERENCES DEPT (DEPTNO)
7369	SMITH	CLERK	10

DEPT

DEPTNO	DNAME	LOC
INTEGER PRIMARY KEY	VARCHAR(30)	VARCHAR(100)
10	APPSERVER	NEW YORK

```
<http://data.example.com/employee/7369> rdf:type ex:Employee.  
<http://data.example.com/employee/7369> ex:name "SMITH".  
<http://data.example.com/employee/7369> ex:department  
<http://data.example.com/department/10>.
```

```
<http://data.example.com/department/10> rdf:type ex:Department.  
<http://data.example.com/department/10> ex:name "APPSERVER".  
<http://data.example.com/department/10> ex:location "NEW YORK".  
<http://data.example.com/department/10> ex:staff 1.
```

```
<#DeptTableView> rr:sqlQuery ""  
SELECT DEPTNO,  
      DNAME,  
      LOC,  
      (SELECT COUNT(*) FROM EMP WHERE EMP.DEPTNO=DEPT.DEPTNO)  
      AS STAFF  
FROM DEPT;  
"".
```

视图定义

视图映射

EMP

EMPNO	ENAME	JOB	DEPTNO
INTEGER PRIMARY KEY	VARCHAR(100)	VARCHAR(20)	INTEGER REFERENCES DEPT (DEPTNO)
7369	SMITH	CLERK	10

DEPT

DEPTNO	DNAME	LOC
INTEGER PRIMARY KEY	VARCHAR(30)	VARCHAR(100)
10	APPSERVER	NEW YORK

Result

<http://data.example.com/employee/7369> rdf:type ex:Employee.
<http://data.example.com/employee/7369> ex:name "SMITH".
<http://data.example.com/employee/7369> ex:department <http://data.example.com/department/10>.

<http://data.example.com/department/10> rdf:type ex:Department.
<http://data.example.com/department/10> ex:name "APPSERVER".
<http://data.example.com/department/10> ex:location "NEW YORK".
<http://data.example.com/department/10> ex:staff 1.

<#TriplesMap2>

rr:logicalTable <#DeptTableView>;

rr:subjectMap [

rr:template "http://data.example.com/department/{DEPTNO}";

rr:class ex:Department;

];

rr:predicateObjectMap [

rr:predicate ex:name;

rr:objectMap [rr:column "DNAME"];

];

rr:predicateObjectMap [

rr:predicate ex:location;

rr:objectMap [rr:column "LOC"];

];

rr:predicateObjectMap [

rr:predicate ex:staff;

rr:objectMap [rr:column "STAFF"];

].

Mapping to a View
Definition

Linking Two Logical Tables

EMP

EMPNO	ENAME	JOB	DEPTNO
INTEGER PRIMARY KEY	VARCHAR(100)	VARCHAR(20)	INTEGER REFERENCES DEPT (DEPTNO)
7369	SMITH	CLERK	10

DEPT

DEPTNO	DNAME	LOC
INTEGER PRIMARY KEY	VARCHAR(30)	VARCHAR(100)
10	APPSERVER	NEW YORK

结果

```
<http://data.example.com/employee/7369> rdf:type ex:Employee.  
<http://data.example.com/employee/7369> ex:name "SMITH".  
<http://data.example.com/employee/7369> ex:department  
<http://data.example.com/department/10>.
```

```
<http://data.example.com/department/10> rdf:type ex:Department.  
<http://data.example.com/department/10> ex:name "APPSERVER".  
<http://data.example.com/department/10> ex:location "NEW YORK".  
<http://data.example.com/department/10> ex:staff 1.
```

```
<#TriplesMap1>  
rr:predicateObjectMap [  
  rr:predicate ex:department;  
  rr:objectMap [  
    rr:parentTriplesMap <#TriplesMap2>;  
    rr:joinCondition [  
      rr:child "DEPTNO";  
      rr:parent "DEPTNO";  
    ];  
  ];  
].
```



-ontop- is a platform to query databases as Virtual RDF Graphs using SPARQL. It's extremely fast and is packed with features.

Features:

- Query language: SPARQL 1.0 (and part of 1.1) support Mapping languages:
 - Intuitive Ontop mapping language
 - W3C R2RML
- Ontology language: OWL 2 QL (DL-Lite)
- Database: Support for free and commercial DBMS
 - PostgreSQL, MySQL, H2, DB2, ORACLE, MS SQL SERVER, TEIID, ADP
- Java library/providers for Sesame and OWLAPI
 - Sesame: “a de-facto standard framework for processing RDF data”
 - OWLAPI: “Java API and reference implementation for OWL Ontologies”
- Integrated with Protege 4.x
- SPARQL end-point (via Sesame Workbench)
- Apache License

Feature matrix of SPARQL query answering systems

Type	System	Reasoning	Mapping support	License	Starting year
Triplestore	Virtuoso	RDFS *	Native, R2RML*	GPL 2, Commercial	1999
	GraphDB	OWL 2 RL	–	Commercial	2005
	Stardog	OWL 2 */ SWRL*	Native, R2RML	Commercial	2012
	RDFox	OWL 2 RL / SWRL / Datalog	–	Academic	2013
OBDA	D2RQ	No	D2RQ Mapping, R2RML*	Apache 2	2004
	Mastro	OWL 2 QL	R2RML*	Academic	2006
	Ultrawrap	RDFS-Plus	Native, R2RML	Commercial	2012
	Morph-RDB	No	R2RML	Apache 2	2013
	<i>Ontop</i>	OWL 2 QL / SWRL*	<i>Ontop</i> Mapping, R2RML	Apache 2	2010

(* indicates limited support)

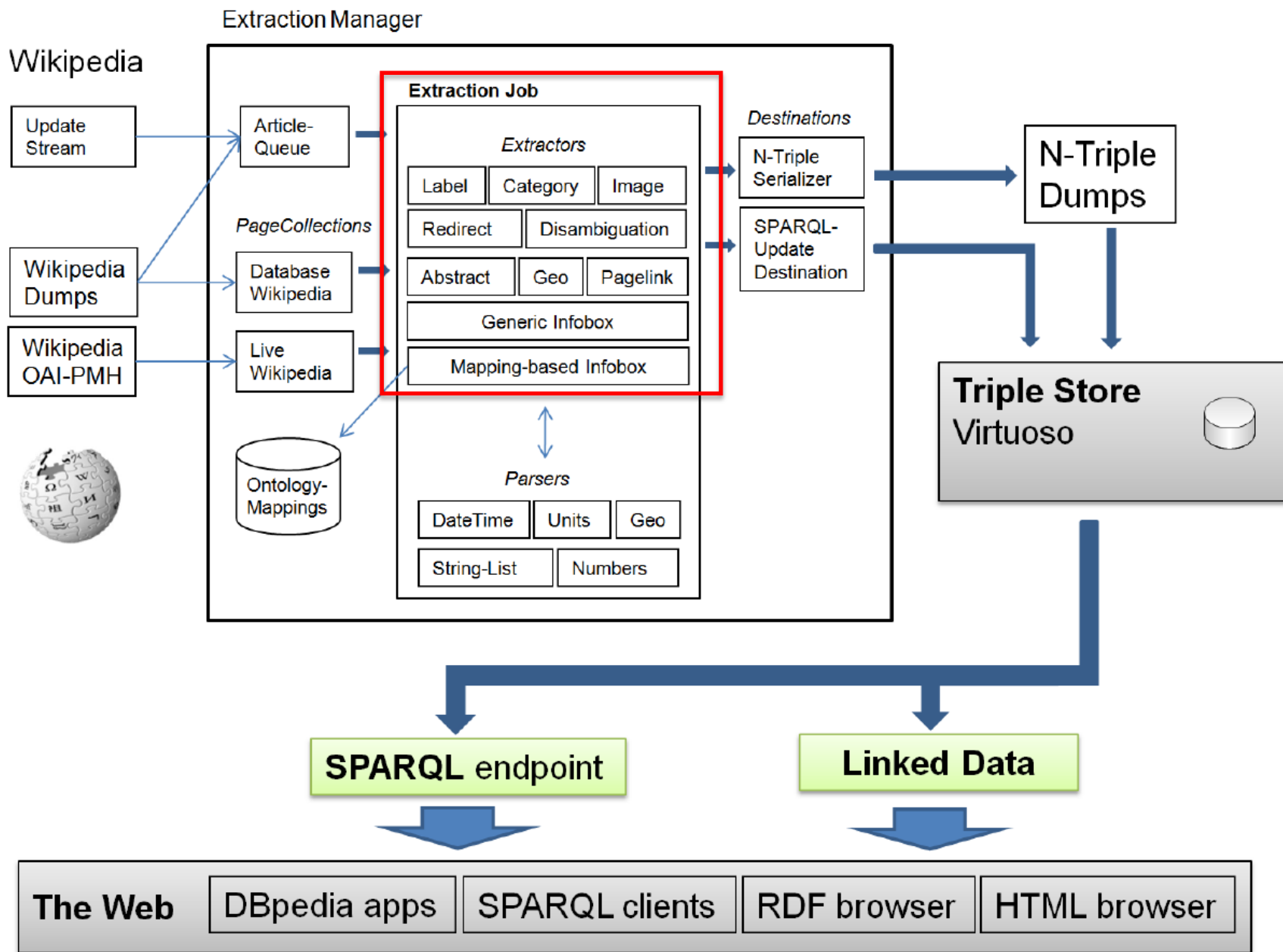
<https://github.com/ontop/ontop-examples/tree/master/ontop-v3-tutorial>

目录

- 面向非结构化数据的知识抽取
- 面向结构化数据的知识抽取
- 面向半结构化数据的知识抽取**
- 实践：基于百科数据的知识抽取

百科类知识抽取

- ❑ 大规模多语言百科知识图谱，维基百科的结构化版本，linked data核心数据集
- ❑ 覆盖127种语言，两千八百万个实体，数亿三元组，支持数据集的完全下载
- ❑ 固定模式对实体信息进行抽取，包括abstract, infobox, category, page link等



article titles=labels, (rdfs: label)

The screenshot shows the Wikipedia page for Tim Berners-Lee. The title 'Tim Berners-Lee' is highlighted with a red box. Below it, the first paragraph of the article is also highlighted with a red box. A red arrow points from the title to the text 'article titles=labels, (rdfs: label)' above. Another red arrow points from the first paragraph to the text 'article abstracts: short abstracts (first paragraph, rdfs: comment)' below.

WIKIPEDIA
The Free Encyclopedia

Article Talk

Read View source View history Search

Tim Berners-Lee

From Wikipedia, the free encyclopedia

Sir Timothy John Berners-Lee OM KBE FRS FREng FRSA FRCGS (born 8 June 1955),^[1] also known as **TimBL**, is an English [engineer](#) and [computer scientist](#), best known as the inventor of the [World Wide Web](#). He made a proposal for an information management system in March 1989,^[3] and he implemented the first successful communication between a [Hypertext Transfer Protocol \(HTTP\)](#) client and [server](#) via the internet in mid-November the same year.^{[4][5][6][7][8]}

Berners-Lee is the director of the [World Wide Web Consortium \(W3C\)](#), which oversees the continued development of the Web. He is also the founder of the [World Wide Web Foundation](#) and is a senior researcher and holder of the [founders chair](#) at the [MIT Computer Science and Artificial Intelligence Laboratory \(CSAIL\)](#).^[9] He is a director of the [Web Science Research Initiative \(WSRI\)](#),^[10] and a member of the advisory board of the [MIT Center for Collective Intelligence](#).^{[11][12]} In 2011, he was named as a member of the board of trustees of the [Ford Foundation](#).^[13]

In 2004, Berners-Lee was [knighted](#) by Queen Elizabeth II for his pioneering work.^{[14][15]} In April 2009, he was elected a foreign associate of the [United States National Academy of Sciences](#).^{[16][17]} Named in [Time](#) magazine's list of the [100 Most Important People of the 20th century](#), Berners-Lee has received a [number of other accolades](#) for his invention.^[18] He was honoured as the "Inventor of the World Wide Web" during the [2012 Summer Olympics opening ceremony](#), in which he appeared in person, working with a vintage [NeXT Computer](#) at the [London Olympic Stadium](#).^[19] He [tweeted](#) "This is for everyone",^[20] which instantly was spelled out in [LCD](#) lights attached to the chairs of the 80,000 people in the audience.^[19] Berners-Lee received the 2016 [Turing Award](#) "for inventing the World Wide Web, the first web browser, and the fundamental protocols and algorithms allowing the Web to scale".^[21]

article abstracts: short abstracts (first paragraph, rdfs: comment)

long abstracts (dbpedia: abstract)



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

In other projects
Wikimedia Commons
Wikisource

Article [Talk](#)

Tim Berners-Lee

From Wikipedia, the free encyclopedia

Sir Timothy John Berners-Lee OM KBE F

He made a proposal for an information ma
internet in mid-November the same year. [

Berners-Lee is the director of the World
researcher and holder of the founders ck
member of the advisory board of the MIT

In 2004, Berners-Lee was knighted by Que
Named in *Time* magazine's list of the 100
World Wide Web" during the 2012 Summer C
everyone", [20] which instantly was spell
Web, the first web browser, and the fund

Contents [hide]

- 1 Early life and education
- 2 Career
- 3 Current work
- 4 Awards and honours
- 5 Personal life
- 6 See also
- 7 References
- 8 Further reading
- 9 External links

Early life and education

Berners-Lee was born in London, England,
Ferranti Mark 1. He attended Sheen Mount
independent school in 1975. [1][14] A kee
where he received a first-class bachelor

*Wikipedia Commons Images=images,
(foaf: depiction)*

Tim Berners-Lee

[View Wikimedia Comm](#)

From Wikimedia Commons, the free media repository

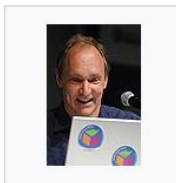
English: [Tim Berners-Lee](#)

Français : [Tim Berners-Lee](#)

Italiano: [Tim Berners-Lee](#)



Tim Berners-Lee in
2009



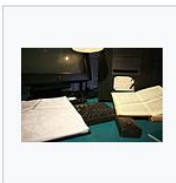
Tim Berners-Lee in
2009



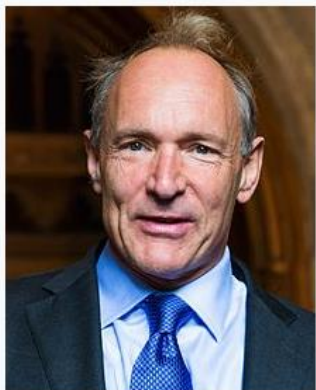
Tim Berners-Lee in
2005



First web server



Tim Berners-Lee's
computer at CERN



Berners-Lee in 2014

infobox template names=instance types, (rdf: type)

Born	Timothy John Berners-Lee 8 June 1955 (age 62) ^[1] London, England, UK
Other names	TimBL TBL
Education	Emanuel School
Alma mater	The Queen's College, Oxford (BA)
Occupation	Computer scientist
Spouse(s)	Rosemary Leith (m. 2014) Nancy Carlson (m. 1990; div. 2011)
Children	2
Parent(s)	Conway Berners-Lee Mary Lee Woods
Awards	Turing Award (2016) Queen Elizabeth Prize (2013) OM (2007) KBE (2004) FRS (2001) ^[2] FREng (2001) FRSA (2001) DFBCS (1995) See full list of honours
Website	www.w3.org/People/ /Berners-Lee
Scientific career	
Institutions	World Wide Web Consortium University of Oxford University of Southampton Plessey

```

{{pp-move-indef}}
{{pp-semi-vandalism|small=yes}}
{{Infobox person
name = Sir Tim Berners-Lee
honorific_suffix = {{postnominals|country=GBR|OM|KBE|FRS|FREng|FRSA|FBCS}}
image = Sir Tim Berners-Lee (cropped).jpg
image_size = 220px
caption = Berners-Lee in 2014
alt = blond man in his fifties wearing a blue suit, light blue shirt, and blue
birth_name = Timothy John Berners-Lee
birth_date = {{birth date and age|1955|6|8|df=y}}<ref name="whoswho"/>
birth_place = [[London]], England, UK
education = [[Emanuel School]]
alma_mater = [[The Queen's College, Oxford]] (BA)
awards = {{Plainlist|
* [[Turing Award]] (2016)
* [[Queen Elizabeth Prize]] (2013)
* [[Member of the Order of Merit|OM]] (2007)
* [[Knight Commander of the Order of the British Empire|KBE]] (2004)
* [[Fellow of the Royal Society|FRS]] (2001)<ref name=frs/>
* [[Fellow of the Royal Academy of Engineering|FREng]] (2001)
* [[Fellow of the Royal Society of Arts|FRSA]] (2001)
* [[Distinguished Fellow of the British Computer Society|DFBCS]] (1995)
* [[Awards and honours presented to Tim Berners-Lee|See full list of honours]]}}
spouse = {{Plainlist|

```

*infobox properties=instance properties,
(dbpedia:property/[propertyName])*



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes

page links=internal links, (dbpedia: wikilink)

Not lo

Article Talk

Read View source View history Search

Tim Berners-Lee

From Wikipedia, the free encyclopedia

Sir Timothy John Berners-Lee OM KBE FRS FREng FRSA FRCS (born 8 June 1955),^[1] also known as **TimBL**, is an English [engineer](#) and [computer scientist](#), best known as the inventor of the [World Wide Web](#). He made a proposal for an information management system in March 1989,^[3] and he implemented the first successful communication between a [Hypertext Transfer Protocol](#) (HTTP) client and [server](#) via the internet in mid-November the same year.^{[4][5][6][7][8]}

Berners-Lee is the director of the [World Wide Web Consortium](#) (W3C), which oversees the continued development of the Web. He is also the founder of the World Wide Web Foundation and is a senior researcher and holder of the [founders chair](#) at the [MIT Computer Science and Artificial Intelligence Laboratory](#) (CSAIL).^[9] He is a director of the [Web Science Research Initiative](#) (WSRI),^[10] and a member of the advisory board of the [MIT Center for Collective Intelligence](#).^{[11][12]} In 2011, he was named as a member of the board of trustees of the Ford Foundation.^[13]

In 2004, Berners-Lee was [knighted](#) by Queen Elizabeth II for his pioneering work.^{[14][15]} In April 2009, he was elected a foreign associate of the [United States National Academy of Sciences](#).^{[16][17]} Named in *Time* magazine's list of the [100 Most Important People of the 20th century](#), Berners-Lee has received a [number of other accolades](#) for his invention.^[18] He was honoured as the "Inventor of the World Wide Web" during the [2012 Summer Olympics opening ceremony](#), in which he appeared in person, working with a vintage [NeXT Computer](#) at the [London Olympic Stadium](#).^[19] He [tweeted](#) "This is for everyone",^[20] which instantly was spelled out in LCD lights attached to the chairs of the 80,000 people in the audience.^[19] Berners-Lee received the 2016 [Turing Award](#) "for inventing the World Wide Web, the first web browser, and the fundamental protocols and algorithms allowing the Web to scale".^[21]

Categories: [1955 births](#) | [Living people](#) | [People from Barnes, London](#) | [People educated at Emanuel School](#) | [Alumni of The Queen's College, Oxford](#) | [Members of the Department of Computer Science, University of Oxford](#) | [People associated with CERN](#) | [English computer scientists](#) | [English expatriate](#) | [Fellows of the American Academy of Arts and Sciences](#) | [Fellows of the British Computer Society](#) | [Fellows of The Queen's College, Oxford](#) | [Fellow Hypertext Transfer Protocol](#) | [Internet pioneers](#) | [Knights Commander of the Order of the British Empire](#) | [MacArthur Fellows](#) | [Massachusetts Institute of Technology](#) | [Members of the United States National Academy of Engineering](#) | [Turing Award laureates](#) | [Members of the United States National Academy of Sciences](#) | [Unitarian Universalists](#) | [Former Anglicans](#) | [Webby Award winners](#) | [World Wide Web Consortium](#) | [Freemen of the City of London](#)

*article categories=topical concepts,
(http://purl.org/dc/terms/subject)*

Category:Computer scientists

From Wikipedia, the free encyclopedia



Pages in this category **should be moved to subcategories where applicable**. This category may require frequent maintenance if it contains very few, if any, pages and should mainly contain subcategories.



Category:Computer scientists are people who do "*research*" in the field of computer science, which is distinct from who would be better placed in [Category:Software engineers](#) or [Category:Computer programmers](#).

This category of [Computer scientists](#) consists of people who do research in the field of [computer science](#): university professors, researchers employed by

Contents

[Top](#) · [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

Subcategories

This category has the following 13 subcategories, out of 13 total.

- [Computer scientists by award](#) (7 C)
- [Computer scientists by field of research](#) (23 C)
- [Computer scientists by nationality](#) (62 C, 1 P)
- [Computer specialists by nationality](#) (17 C)
- [Women computer scientists](#) (368 P)

► [Computer architects](#) (2 P)

F

- [Fellows of the Association for Computing Machinery](#) (572 P)
- [Fictional computer scientists](#) (6 P)

I

- [IBM Research computer scientists](#) (31 P)

A

L

Wikipedia category system=categories, (skos: boarder/ skos: narrower)

Albert Einstein (disambiguation)

From Wikipedia, the free encyclopedia

Albert Einstein was a noted German-born theoretical physicist.

Albert Einstein may also refer to other people with the same name and to many things named in his honor:

People [[edit](#)]

- Albert Lawrence Einstein, the birth name of actor, comedian, director, and writer [Albert Brooks](#)

Educational institutions [[edit](#)]

- [Albert Einstein College of Medicine](#), the graduate medical school of Yeshiva University in New York City
- [Albert Einstein Institute](#), also known as the [Max Planck Institute for Gravitational Physics](#), located in Golm, Potsdam, Germany (theoretic
- [Albert Einstein Institution](#), East Boston, Massachusetts, a non-profit organization dedicated to the study of methods of non-violent resis
- [Albert Einstein Medical Center](#), Philadelphia, Pennsylvania, a tertiary-care teaching hospital
- [Albert Einstein School \(disambiguation\)](#)

Honors [[edit](#)]

- [Albert Einstein Award](#), with honorees selected by a committee of the Institute for Advanced Study, awarded sporadically from 1951 to 1979
- [UNESCO Albert Einstein medal](#)
- [Albert Einstein Medal](#), presented by the Albert Einstein Society in Bern, Switzerland, first awarded in 1979
- [Albert Einstein Peace Prize](#), awarded by the Chicago-based Albert Einstein Peace Prize Foundation, first presented in 1980
- [Albert Einstein World Award of Science](#), given by the World Cultural Council, first awarded in 1984

Arts and entertainment [[edit](#)]

- *[Albert Einstein](#)* (album), a 2013 album by rappers Prodigy and The Alchemist
- *[Albert Einstein: The Practical Bohemian](#)*, a 1978 one-man stage play written and performed by actor-writer Ed Metzger
- Albert Einstein, a character in the television series *[Alien Nation](#)*

Other uses [[edit](#)]

- [Albert Einstein Hospital](#), São Paulo, Brazil
- [Albert Einstein Science Park](#), Potsdam, Germany
- [Albert Einstein ATV](#), a European unmanned cargo resupply spacecraft
- [Albert Einstein Society](#), based in Bern, Switzerland

homonymies=*ambiguous instances*, (dbpedia: **disambiguates**)

New York City

From Wikipedia, the free encyclopedia
(Redirected from [NYC](#))

"NYC" and "New York, New York" redirect here. For other uses, see [New York City \(disambiguation\)](#); [NYC \(disambiguation\)](#); and [New York, New York \(disambiguation\)](#).

The **City of New York**, often called **New York City** or simply **New York**, is the [most populous city](#) in the United States.^[9] With an estimated 2016 population of 8, a land area of about 302.6 square miles (784 km²),^{[10][11]} New York City is also the [most densely populated](#) major city in the United States.^[12] Located at the southe [York](#), the city is the center of the [New York metropolitan area](#), one of the most populous [urban agglomerations](#) in the world.^{[13][14]} A [global power city](#),^[15] New York impact upon commerce, finance, media, art, fashion, research, technology, education, and entertainment, its fast pace^{[16][17]} defining the term *New York minute*.^[18] [the United Nations](#),^[19] New York is an important center for [international diplomacy](#)^[20] and has been described as the cultural, financial, and [media capital](#)^{[21][22]} o world.^{[23][24][25][26][27]}

synonyms=redirects, (owl:sameAs)

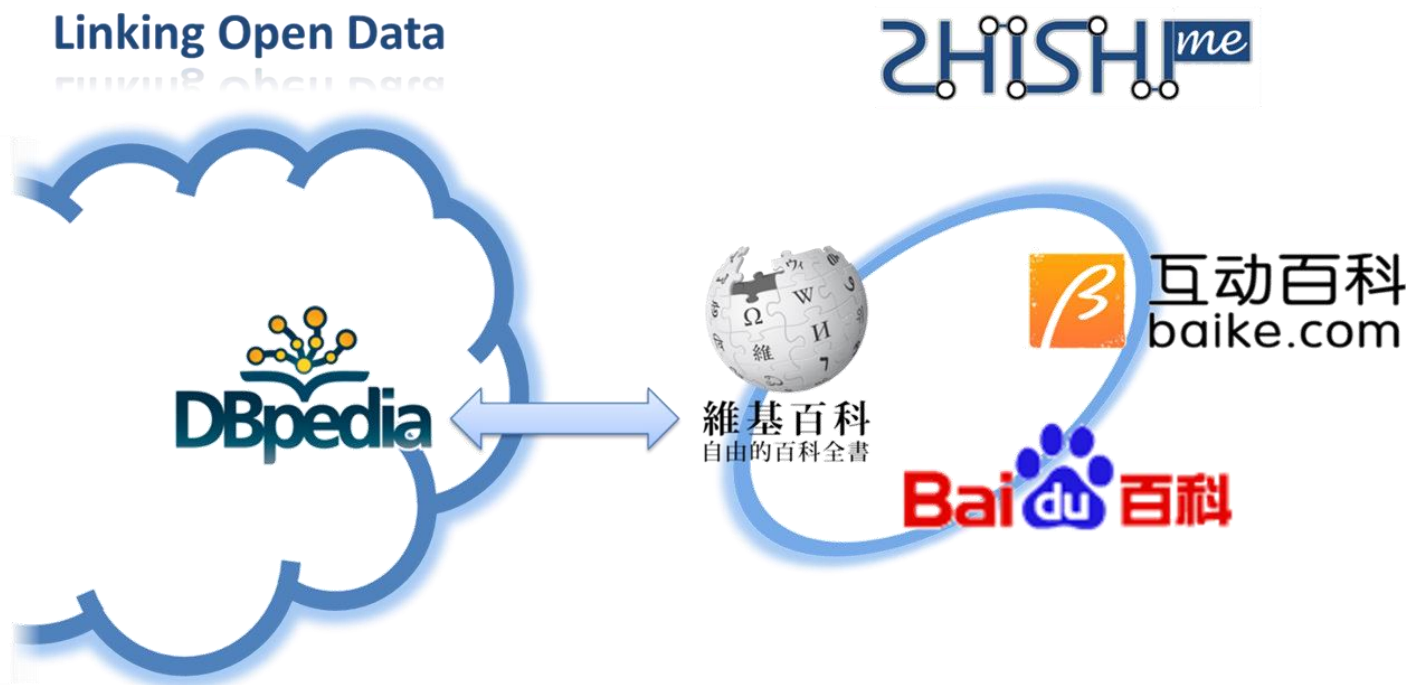
Generic Infobox Extraction

同义属性名不做映射，保持Wikipedia中原有内容，如：birthdate与dateOfBirth.

Mapping-based Infobox Extraction

定义dbpedia ontology，将属性做好对齐（二次处理, 人工定义规则），目前共有2795 properties.

1) 第一份中文大规模开放链接数据 (Chinese Linking Open Data);



2) 1000万实体，2亿三元组，提供关键字查询服务，API调用，SPARQL Endpoint

Country	Year	Value
Algeria	2000	0.00
Algeria	2001	0.00
Algeria	2002	0.00
Algeria	2003	0.00
Algeria	2004	0.00
Algeria	2005	0.00
Algeria	2006	0.00
Algeria	2007	0.00
Algeria	2008	0.00
Algeria	2009	0.00
Algeria	2010	0.00
Algeria	2011	0.00
Algeria	2012	0.00
Algeria	2013	0.00
Algeria	2014	0.00
Algeria	2015	0.00
Algeria	2016	0.00
Algeria	2017	0.00
Algeria	2018	0.00
Algeria	2019	0.00
Algeria	2020	0.00
Algeria	2021	0.00
Algeria	2022	0.00
Algeria	2023	0.00
Algeria	2024	0.00
Algeria	2025	0.00
Algeria	2026	0.00
Algeria	2027	0.00
Algeria	2028	0.00
Algeria	2029	0.00
Algeria	2030	0.00
Algeria	2031	0.00
Algeria	2032	0.00
Algeria	2033	0.00
Algeria	2034	0.00
Algeria	2035	0.00
Algeria	2036	0.00
Algeria	2037	0.00
Algeria	2038	0.00
Algeria	2039	0.00
Algeria	2040	0.00
Algeria	2041	0.00
Algeria	2042	0.00
Algeria	2043	0.00
Algeria	2044	0.00
Algeria	2045	0.00
Algeria	2046	0.00
Algeria	2047	0.00
Algeria	2048	0.00
Algeria	2049	0.00
Algeria	2050	0.00
Algeria	2051	0.00
Algeria	2052	0.00
Algeria	2053	0.00
Algeria	2054	0.00
Algeria	2055	0.00
Algeria	2056	0.00
Algeria	2057	0.00
Algeria	2058	0.00
Algeria	2059	0.00
Algeria	2060	0.00
Algeria	2061	0.00
Algeria	2062	0.00
Algeria	2063	0.00
Algeria	2064	0.00
Algeria	2065	0.00
Algeria	2066	0.00
Algeria	2067	0.00
Algeria	2068	0.00
Algeria	2069	0.00
Algeria	2070	0.00
Algeria	2071	0.00
Algeria	2072	0.00
Algeria	2073	0.00
Algeria	2074	0.00
Algeria	2075	0.00
Algeria	2076	0.00
Algeria	2077	0.00
Algeria	2078	0.00
Algeria	2079	0.00
Algeria	2080	0.00
Algeria	2081	0.00
Algeria	2082	0.00
Algeria	2083	0.00
Algeria	2084	0.00
Algeria	2085	0.00
Algeria	2086	0.00
Algeria	2087	0.00
Algeria	2088	0.00
Algeria	2089	0.00
Algeria	2090	0.00
Algeria	2091	0.00
Algeria	2092	0.00
Algeria	2093	0.00
Algeria	2094	0.00
Algeria	2095	0.00
Algeria	2096	0.00
Algeria	2097	0.00
Algeria	2098	0.00
Algeria	2099	0.00
Algeria	2100	0.00
Algeria	2101	0.00
Algeria	2102	0.00
Algeria	2103	0.00
Algeria	2104	0.00
Algeria	2105	0.00
Algeria	2106	0.00
Algeria	2107	0.00
Algeria	2108	0.00
Algeria	2109	0.00
Algeria	2110	0.00
Algeria	2111	0.00
Algeria	2112	



Label

foaf:page <http://www.hudong.com/wiki/%E5%8C%97%E4%...1%E5%8C%97%E4%BA%ACzhishi:resourceID757341>

Index:

- [zhishi:abstract](#)
- [infobox](#)
- [dcterm:subject](#)
- [zhishi:thumbnail](#)
- [zhishi:relatedPage](#)
- [zhishi:externalLink](#)
- [zhishi:internalLink](#)

- ☒ hudong:北京 (this)
- ☒ baidu:北京
- ☒ zhwiki:北京

MERGE PAGE

Abstract

北京是中华人民共和国的首都，四个直辖市之一，是中国的政治、文化、科技、教育和国际交往中心，也是中国最大的航空交通枢纽和中国北方的经济中心。北京市位于华北平原北部，东南隅地区与天津市相连，其余为河北省所环绕。以市区人口数相比，北京为次于上海的中国第二大城市。北京是一座有三千年历史的文化名城，历史上共有八个朝代曾在此建都，北京荟萃了源远流长的中华文化，拥有诸多名胜古迹和人文景观。北京于2008年成功举办了第29届夏季奥运会（2008年8月8日至8月24日），及北京奥运会开幕式（2008年8月8日19时）。

>IOE

所属地区	中国
GDP	11,865.8亿元人民币(2009年)
主要街道	长安街、平安大街
面积	16800平方公里
中文名	北京
下辖地区	海淀区 西城区 朝阳区 大兴区
友好城市	纽约、东京、芝加哥
知名企业	首钢集团、北京科技园集团有限公司、北京亨宝公司、北京味群公司
特产	烤鸭、瑞昌贡、艾蒿菜、豆汁
人口	1,765万人(2009年)
时区/时差	010
英文名	Municipality of Beijing

[»»TOP](#)

[水游地名](#)
[文化](#)
[首都](#)
[各国首都](#)
[植物](#)
[历史](#)
[互联网](#)
[地理](#)
[中国](#)
[中国城市](#)

>>TOP

A collage of 20 small images representing various aspects of China. The images include: a lake with mountains, a blue pagoda, a map of China, a person in a green field, a stone bridge, a modern building, a night cityscape, a red lantern, a white building, a modern building, a traditional Chinese gate, a lake with mountains, a traditional Chinese gate, a group of people, a traditional Chinese gate, a night cityscape, a traditional Chinese gate, a group of people, a traditional Chinese gate, a night cityscape, and a traditional Chinese gate.

>>TOP

SHOW MORE (10)

[»»TOP](#)

<http://www.cq6.cc/index.php%3Fdoc-innerlink-%25E5%258C%2597%25E4%25BA%25AC>
SHOW MORE (14)

2010

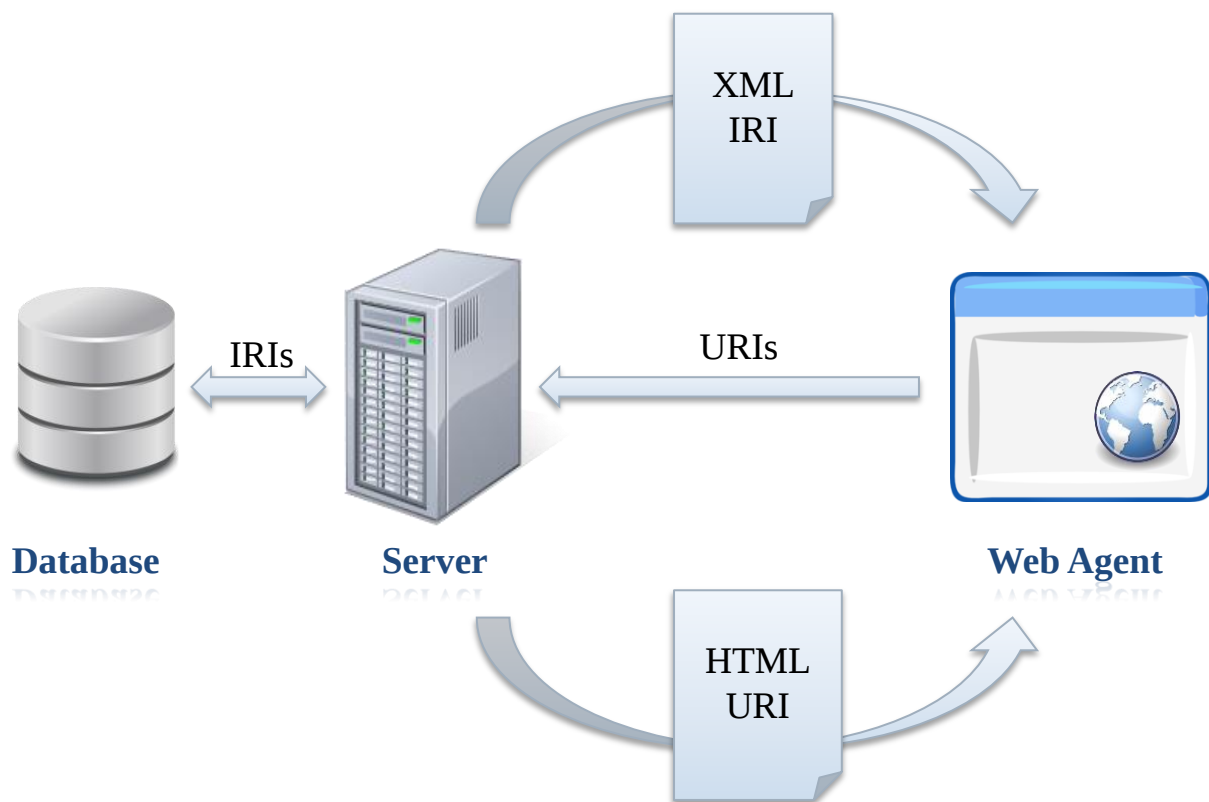
福建省 [SHOW MORE \(798\)](#)

>>TOF

Other properties

五种访问Zhishi.me的方式:

1) Linked data



五种访问Zhishi.me的方式:

2) Lookup service (<http://zhishi.me/lookup/>);



太平洋

SUBMIT

- ☒ baidu:太平洋
- ☐ hudong:太平洋
- ☐ hudong:《太平洋》
- ☒ hudong:太平洋[海洋]
- ☐ hudong:太平洋[股票]
- ☒ zhwiki:太平洋

MERGE PAGE

3) SPARQL endpoint (<http://zhishi.me/sparql/>);

五种访问 Zhishi.me 的方式:

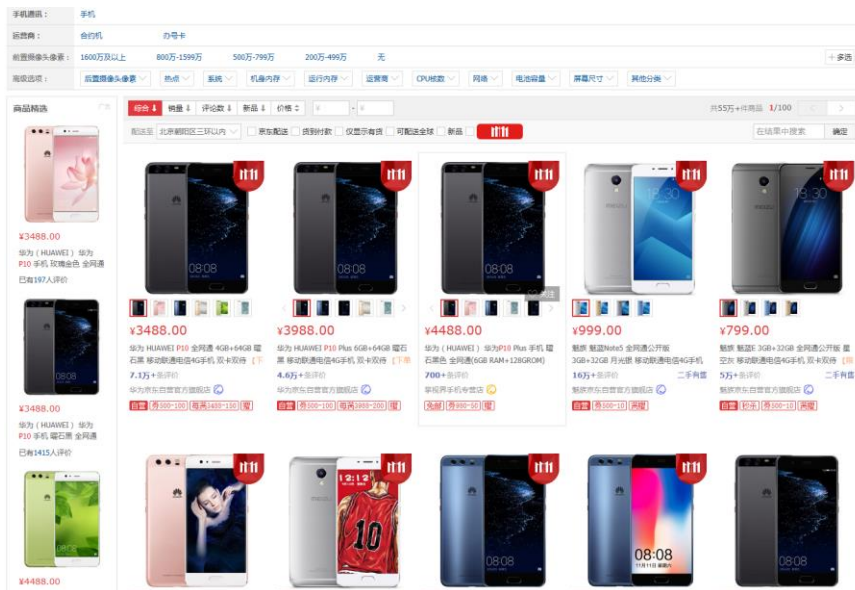
4) APIs (<http://zhishi.me/api>);

提供了4中不同的查询方式. 用户可通过HTTP请求查询知识图谱。查询结果将会以JSON字符串的方式返回。

5) Data Dump (于 OpenKG 中开放, <http://openkg.cn/dataset/zhishi-me-dump>);

WEB网页数据抽取： 包装器生成

网页中的信息



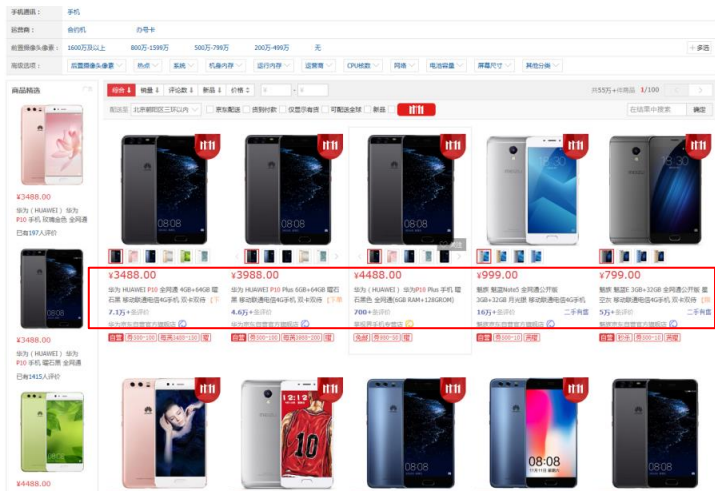
列表页数据



详情页数据

互联网中的网页含有丰富的数据。例如电商网站中的商品数据，黄页网站中的公司数据等等。我们获取网页中的数据经过加工后可以丰富我们的知识图谱。

如何获取网页信息？—手工方法



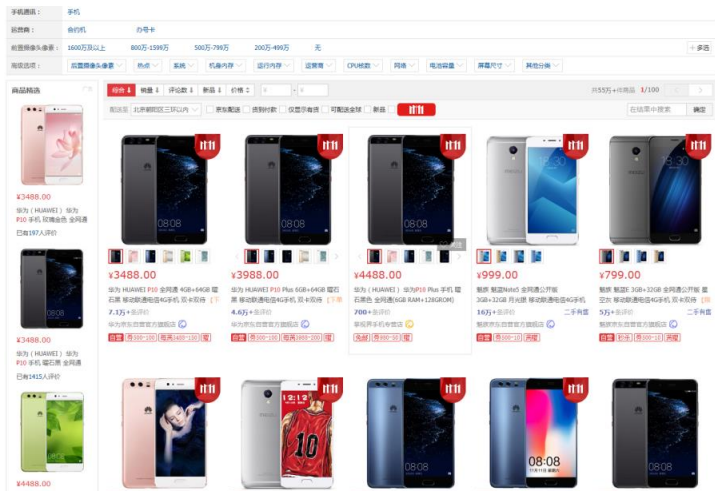
某电商网站搜索结果页面

```
<div id="J_goodsList" class="goods-list-v2 gl-type-3 J-goods-list">
  <ul class="gl-wrap clearfix" data-tpl="3">
    <li class="gl-item" data-sku="4483094" data-spu="4483110" data-pid="4483110">
      <div class="gl-i-wrap">
        <div class="p-img">...</div>
        <div class="p-scroll">...</div>
        <div class="p-price">
          <strong class="J_4483094" data-done="1">
            <em>¥</em>
            <i>3488.00</i>
          </strong>
        </div>
        <div class="p-name p-name-type-2">
          <a target="_blank" title="【下单减150+白条6期免息】限量领券再减100！成交价3238！不憋着，现在买。如遇11月12号前降价，京东补差价！ 猛戳我！" href="//item.jd.com/4483094.html" onclick="searchlog(1,4483094,0,1,'','flagsCtk=1245713032')">
            <em>
              "华为 HUAWEI "
              <font class="skcolor_ljg">P10</font>
              "全网通 4GB+64GB 曜石黑 移动联通电信4G手机 双卡双待"
            </em>
            <i class="promo-words" id="J_AD_4483094">【下单减150+白条6期免息】限量领券再减100！成交价3238！不憋着，现在买。如遇11月12号前降价，京东补差价！ 猛戳我！</i>
          </a>
        </div>
        <div class="p-commit">...</div>
        <div class="p-focus">...</div>
        <div class="p-shop" data-selfware="1" data-score="5" data-reputation="96" data-done="1">...</div>
        <div class="p-icons" id="J_pro_4483094" data-done="1">...</div>
      </div>
    </li>
  </ul>
</div>
```

某电商网站搜索结果页面HTML代码

手工方法需要查看网页结构和代码，通过人工分析，手工写出适合这个网站的表达式，这个表达式的形式可以是XPath表达式，也可以是CSS选择器的表达式等。

如何获取网页信息? --手工方法



某电商网站搜索结果页面

```
<div id="J_goodsList" class="goods-list-v2 gl-type-3 J-goods-list">
  <ul class="gl-warp clearfix" data-tpl="3">
    <li class="gl-item" data-sku="4483094" data-spu="4483110" data-pid="4483110">
      <div class="gl-i-wrap">
        <div class="p-img"></div>
        <div class="p-scroll"></div>
        <div class="p-price">
          <strong class="J_4483094" data-done="1">
            <em>¥</em>
            <i>3488.00</i>
          </strong>
        </div>
        <div class="p-name p-name-type-2">
          <a target="_blank" title="【下单减150+白条6期免息】限量领券再减100! 成交价3238! 不憋着, 现在买。如遇11月12号前降价, 京东补差价! 猛戳我!" href="//item.jd.com/4483094.html" onclick="searchlog(1,4483094,0,1,'','flagsCtk=1245713032')">
            <em>
              "华为 HUAWEI "
              <font class="skcolor_ljg">P10</font>
              "全网通 4GB+64GB 曜石黑 移动联通电信4G手机 双卡双待"
            </em>
            <i class="promo-words" id="J_AD_4483094">【下单减150+白条6期免息】限量领券再减100! 成交价3238! 不憋着, 现在买。如遇11月12号前降价, 京东补差价! 猛戳我!</i>
          </a>
        </div>
        <div class="p-commit"></div>
        <div class="p-focus"></div>
        <div class="p-shop" data-selfware="1" data-score="5" data-reputation="96" data-done="1"></div>
        <div class="p-icons" id="J_pro_4483094" data-done="1"></div>
      </div>
    </li>
  </ul>
</div>
```

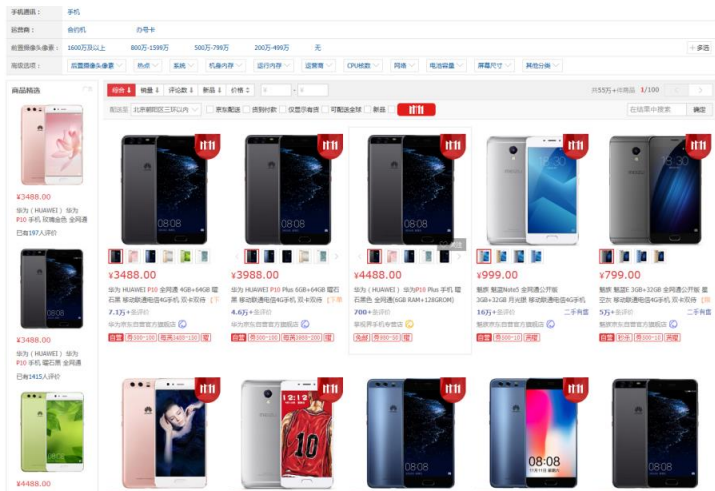
某电商网站搜索结果页面HTML代码

• 何为XPath?

XPath即为XML路径语言, 它是一种用来确定XML (标准通用标记语言的子集) 文档中某部分位置的语言。借助它可以获取网页中元素的位置, 从而获取需要的信息。分析上页搜索结果页面, 价格信息的XPath为:

```
//*[ @id="J_goodsList"]/ul/*/div/div[3]/strong
```

如何获取网页信息? -- 手工方法



某电商网站搜索结果页面

```
<div id="J_goodsList" class="goods-list-v2 gl-type-3 J-goods-list">
  <ul class="gl-warp clearfix" data-tpl="3">
    <li class="gl-item" data-sku="4483094" data-spu="4483110" data-pid="4483110">
      <div class="gl-i-wrap">
        <div class="p-img"></div>
        <div class="p-scroll"></div>
        <div class="p-price">
          <strong class="J_4483094" data-dome="1">
            <em>¥</em>
            <i>3488.00</i>
          </strong>
        </div>
        <div class="p-name p-name-type-2">
          <a target="_blank" title="【下单减150+白条6期免息】限量领券再减100! 成交价3238! 不憋着, 现在买。如遇11月12号前降价, 京东补差价! 猛戳我!" href="//item.jd.com/4483094.html" onclick="searchlog(1,4483094,0,1, '', 'flagsCtk=1245713032')">
            <em>
              "华为 HUAWEI "
              <font class="skcolor_ljg">P10</font>
              "全网通 4GB+64GB 曜石黑 移动联通电信4G手机 双卡双待"
            </em>
            <i class="promo-words" id="J_AD_4483094">【下单减150+白条6期免息】限量领券再减100! 成交价3238! 不憋着, 现在买。如遇11月12号前降价, 京东补差价! 猛戳我!</i>
          </a>
        </div>
        <div class="p-commit"></div>
        <div class="p-focus"></div>
        <div class="p-shop" data-selfware="1" data-score="5" data-reputation="96" data-dome="1"></div>
        <div class="p-icons" id="J_pro_4483094" data-dome="1"></div>
      </div>
    </li>
  </ul>
</div>
```

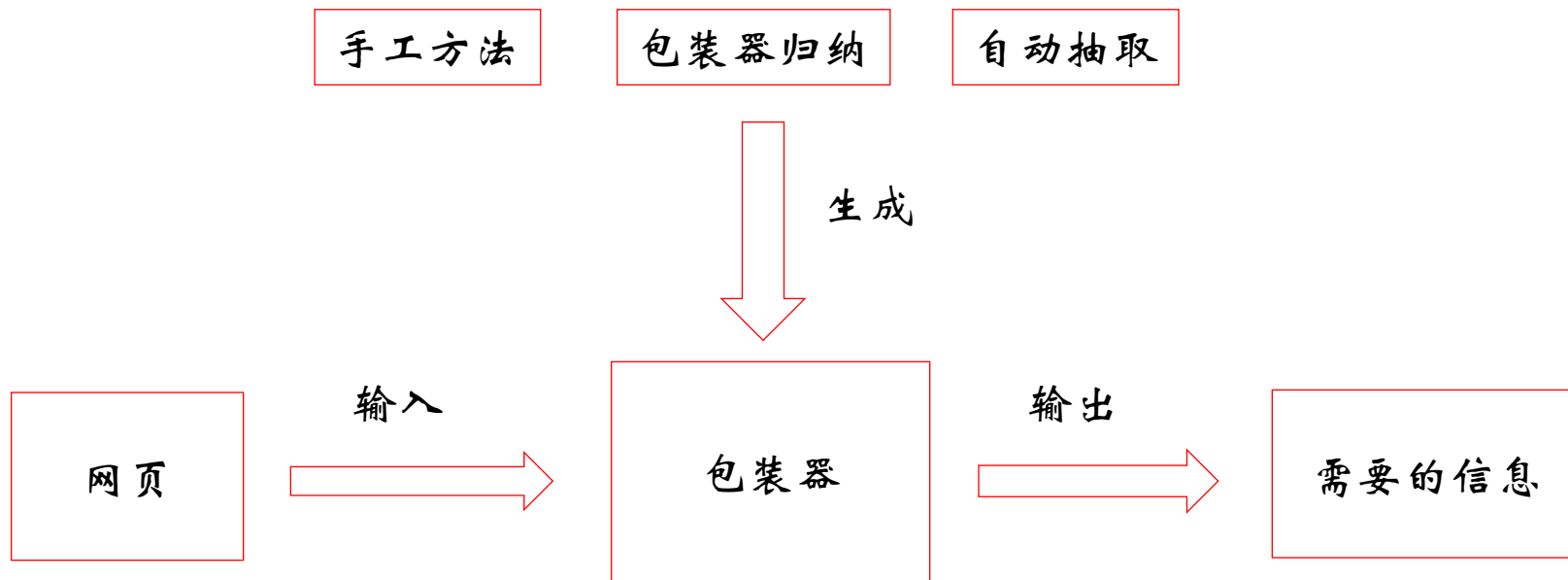
某电商网站搜索结果页面HTML代码

• 何为CSS选择器表达式?

通过CSS元素实现对网页中元素的定位, 获取元素的信息。分析上页搜索结果页面, 价格信息的CSS选择器表达式为:

```
#J_goodsList > ul > li:nth-child(1) > div > div.p-price > strong
```

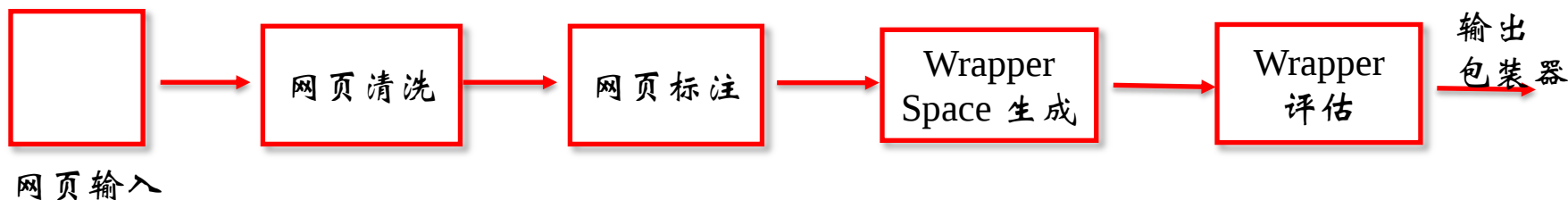
包装器简介及描述方式



包装器定义：

包装器是一个能够将数据从HTML网页中抽取出来，并且将它们还原为结构化的数据的软件程序。

包装器归纳



包装器归纳是基于有监督学习的，他从标注好的训练样例集合中学习数据抽取规则，用于从其他用相同标记或相同网页模板抽取目标数据。

包装器归纳

- 网页清洗

有些网页结构不规范，例如前后标签不对成，没有结束标签符。不规范的网页结构容易在抽取的过程中产生噪声。清洗可以用Tidy来完成。

- 网页标注

网页标注是在网页上标注你需要抽取数据的过程。标注的过程可以是给网页中的某个位置打上特殊的标签表明这是需要抽取的数据。

例如我们要抽取上面举例的“华为P10”搜索页面的商品信息和价格信息，就可以在通过在他们所在的标签里打上一个特殊的标记作为标注。

- 包装器空间的生成

对标注的数据生成XPath集合空间，对生成的集合进行归纳，形成若干个子集。归纳的规则是在子集中的XPath能够覆盖多个标注的数据项，具有一定的泛化能力。

包装器归纳

- 包装器评估

评估规则一：准确率。将筛选出来的包装器对原先训练的网页进行标注，统计与人工标注的相同项的数量，除以当前标注的总数量。准确率越高评分越高。

评估规则二：召回率。将筛选出来的包装器对原先训练的网页进行标注，统计与人工标注的相同项的数量，除以人工标注的总数量。召回率越高评分越高。

- 包装器归纳结果

经过前面一系列的工作之后，“华为P10”搜索结果页面最后价格信息的XPath的路径为：

```
//*[ @id="J_goodsList"]/ul/*/div/div[3]/strong
```

包装器归纳：一个简单的例子

Name	Address	Mobile	Zip
n1	a1	z1	p5
n2	a2	z2	p2
n3	a3	z3	p3
n4	a4	z4	p4
n5	a5	z5	p5

- 包装器空间生成：

假如我们标注的信息同时标注上了n1,n2。他们的XPath分别是：

n1的XPath：/html[1]/body[1]/table[1]/tbody[1]/tr[2]/td[1]

n2的XPath：/html[1]/body[1]/table[1]/tbody[1]/tr[3]/td[1]

假如将其归纳到一个集合，然后泛化之后的XPath可能是：

a. /html[1]/body[1]/table[1]/tbody[1]/*td[1]

b. //*/td

- 包装器评估：

评估规则1：包装器a的准确率比包装器b的高，所以 a优于b

评估规则2：包装器a的召回率和包装器b的一样高，所以 a等于b

综上所述，最后选择包装器a

为什么要有自动抽取？

- 有监督学习包装器有缺陷

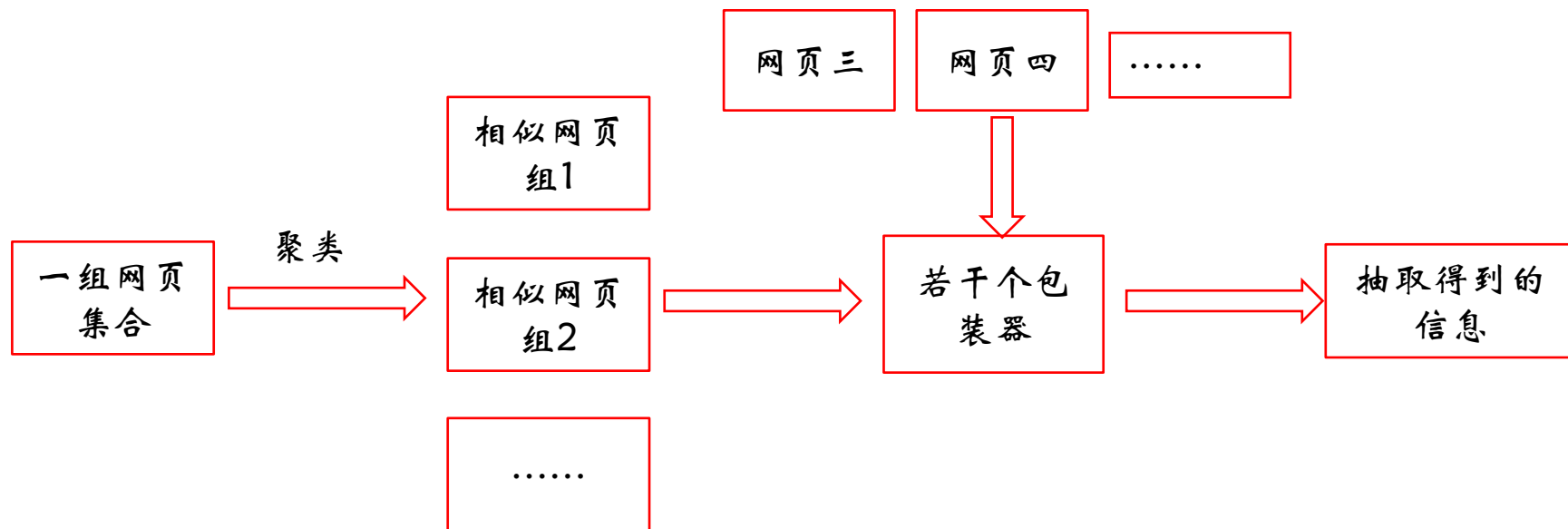
1. 由于需要手工标注的工作，它不适合对大量站点的抽取。
2. 包装器维护的开销会很大。例如网站改变了其已有的模板，之前生成的包装器将会无效。

- 自动抽取是可行的

网站中的数据通常是用很少的一些模板来编码的，通过挖掘多个数据记录中的重复模式来寻找这些模板是可能的。



自动抽取



- 包装器训练

自动抽取网页中的信息不需要任何的先验知识和人工数据的标注。将一组网页通过聚类将相似的网页分成若干个组，每组相似的网页将获得不同的包装器。

- 包装器应用

将需要抽取的网页与之前生成包装器的网页进行比较，在某个分类下则使用该分类下的包装器来获取网页中的信息。

自动抽取: 两个相似网页的比较过程

-Sample(Page1)

1: <HTML>

2: Books of:

3:

4: Paul Smith

5:

6:

7:

8-10: <I>Title:</I>

11: DB Primer

12:

13:

14-16: <I>Title</I>

17: comp.Sys.

18:

19:

20: </HTML>

算法进行之后的包装器

<HTML>Books of:#PCDATA

(?)

(<I>Title:</I>#PCDATA)+

</HTML>

文本字符串失配项
(#PCDATA)

标签失配项 (?) 可选数据项

文本字符串失配项
(#PCDATA)

文本字符串失配项
(#PCDATA)

标签失配项 (+) 迭代器

向上迭代, 看有没有相似的代码结构

-Sample(Page2)

1: <HTML>

2: Books of:

3:

4: Paul Jones

5:

6:

7:

8:

9-11: <I>Title:</I>

12: XML at Work

13:

14:

15-17: <I>Title</I>

18: HTML Scripts

19:

20:

21-23: <I>Title</I>

24: Javascript

25:

26:

27: </HTML>

自动抽取

RoadRunner是自动抽取的典型代表，其抽取的内容如下图所示

The screenshot displays the RoadRunner web application interface, which has extracted data from various websites. The interface is divided into several sections:

- Automotive Industry:** Features a sidebar with categories like Aerospace & Defense, Automotive, Industry Compliance, Food & Beverage, General Manufacturing, Medical & Pharmaceuticals, and Packaging. The main content area shows featured clients like Baldor and Honeywell, along with a 'Take Control' button and a 'REGISTER FOR A LIVE DEMO' button.
- Aerospace & Defense Industry:** Similar to the Automotive section, it features a sidebar with categories and a main content area with featured clients like Honeywell and Lockheed Martin, along with a 'Take Control' button and a 'REGISTER FOR A LIVE DEMO' button.
- Right Sidebar:** Contains various widgets including 'Take Control', 'Webinars', 'Product', and 'Related Blog Posts'. It also includes a 'Request a Free Software Trial' button and a 'Launch Product Demo' button.

The extracted data includes company names, logos, and descriptions. For example, the Baldor Electric Company is described as a manufacturer of industrial electric motors, drives, and generators. The Honeywell company is described as a Fortune 100 diversified technology and manufacturing leader. The Lockheed Martin company is described as a global security company that employs about 132,000 people worldwide.

抽取的结果

需要抽取的网页

RoadRunner 项目地址: <http://www.dia.uniroma3.it/db/roadRunner/>

小结

三种实现方法比较

	手工方法	包装器归纳	自动抽取
优点	- 对于任何一个网页都是通用的，简单快捷 - 能抽取到用户感兴趣的数据	- 需要人工标注训练集 - 能抽取到用户感兴趣的数据 - 可以运用到规模不大网站的信息抽取	- 无监督的方法，无需人工进行数据的标注 - 可以运用到大规模网站的信息抽取
缺点	- 需要对网页数据进行标注，耗费大量的人力 - 维护成本高 - 无法处理大量站点的情况	- 可维护性比较差 - 需要投入大量的人力去做标注	- 需要相似的网页作为输入 - 抽取的内容无法可能达不到预期，会抽取出一些无关信息

面临的挑战

- 现在网页是用过动态生成的，很多信息无法直接通过网页得到。
- 包装器的实时性越来越高，包装器的维护的成本依旧很大

WEB TABLE抽取简介

为了解释Web table中隐含的语义，一些工作将其中的内容标注为RDF三元组。

这种标注的第一步就是实体链接 (**entity linking**), 即将表格中各单元格的字符串映射到给定知识库的实体上。

什么是表格实体链接?

➤ 即将表格中各单元格的字符串映射到给定知识库的实体上;

球队	城市	拥有者	球场
洛杉矶快船队	洛杉矶	史蒂夫·鲍尔默	斯台普斯球馆
达拉斯小牛队	达拉斯	马克·库班	美航中心球馆
...	...	...	...
夏洛特黄蜂队	夏洛特	迈克尔·乔丹	时代华纳有线球馆
芝加哥公牛队	芝加哥	杰里·雷因斯多夫	联合中心球馆



表格实体链接的步骤

1) 候选生成.

针对表格单元格中的每个字符串，从给定的知识库中识别候选实体。
(*token*匹配(字符串匹配, 同义词匹配,.....))

例子.

字符串: “夏洛特”

候选实体

“夏洛特(北卡罗来纳)”

“夏洛特(伊利诺伊)”

“夏洛特黄蜂”

...

2) 实体消歧.

从给定字符串所对应的实体集中选择唯一的一个实体作为链接实体。

实体消歧

位于相同行或者列的字符串可能相关

球队	城市	拥有者	球场
洛杉矶快船队	洛杉矶	史蒂夫·鲍尔默	斯台普斯球馆
达拉斯小牛队	达拉斯	马克·库班	美航中心球馆
...	...	...	...
夏洛特黄蜂队	夏洛特	迈克尔·乔丹	时代华纳有线球馆
芝加哥公牛队	芝加哥	杰里·雷因斯多夫	联合中心球馆

 “迈克尔·乔丹”的相关字符串

换句话说，出现在同一表格中的任意两个字符串都存在某种潜在的关联。

使用图算法对给定表格中的所有字符串进行联合消歧。

实体消歧—构建实体消歧图

对每张给定的表格建立一个实体消歧图

每个图由如下的元素构成：

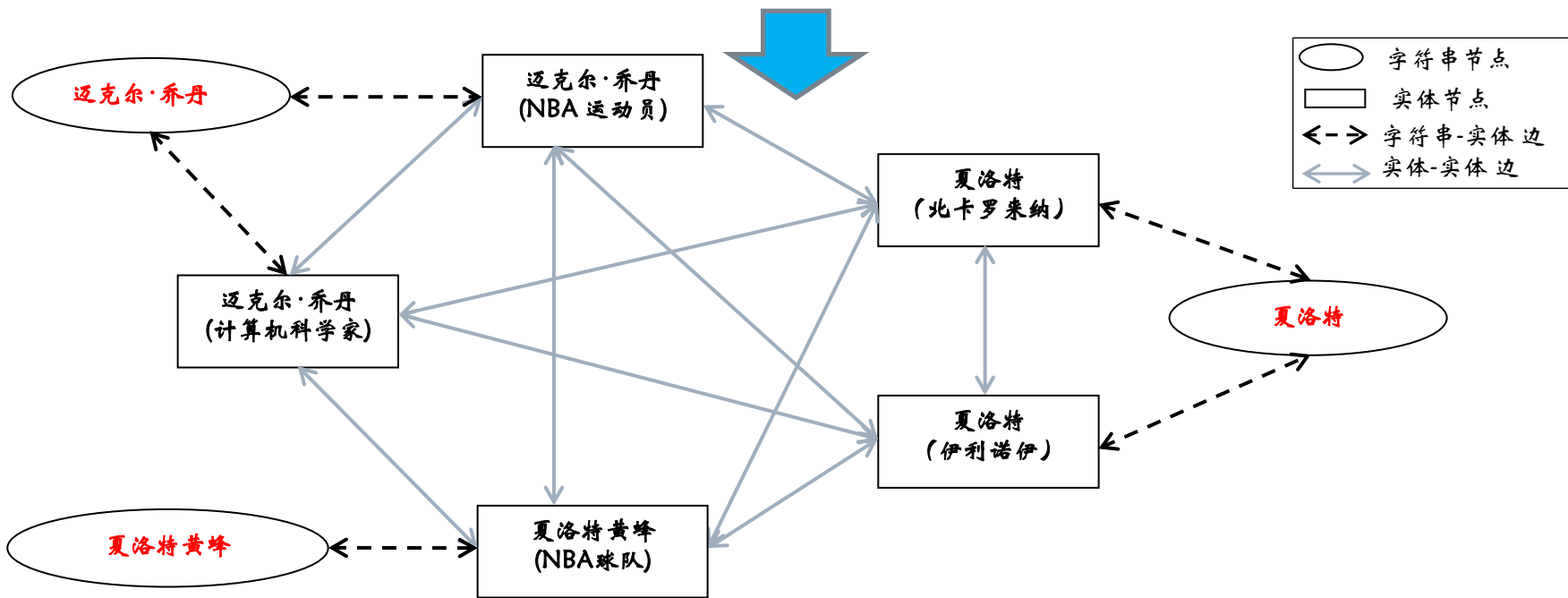
字符串节点, 实体节点

字符串-实体边: 字符串与候选实体间的无向边,

实体-实体边: 实体间的无向边。

例子.

球队	城市	拥有者	球场
洛杉矶快船队	洛杉矶	史蒂夫·鲍尔默	斯台普斯球馆
达拉斯小牛队	达拉斯	马克·库班	美航中心球馆
...	...	...	...
夏洛特黄蜂队	夏洛特	迈克尔·乔丹	时代华纳有线球馆
芝加哥公牛队	芝加哥	杰里·雷因斯多夫	联合中心球馆



实体消歧——计算实体链接影响因子

两类实体链接影响因子：

- 1) 每个字符串的初始重要性(*importance* of each mention);
- 2) 不同节点间的语义相关度(*semantic relatedness* between different nodes).

实体消歧算法——PageRank.

PageRank算法 (Iterative probability propagation) 用来整合不同的实体链接影响因子从而做出最终的实体链接决定.

生成三元组

罗纳尔迪尼奥	巴西	巴塞罗那
法比奥·卡纳瓦罗	意大利	尤文图斯
卡卡	巴西	AC米兰
里奥·梅西	阿根廷	巴塞罗那

两列之间任意同一行的两个单元格中的字符串之间的关系：
“国籍”

在给定某个知识库的情况下，进行实体链接后，我们发现：

里奥·梅西 国籍 阿根廷

卡卡 国籍 巴西

罗纳尔迪尼奥 国籍 巴西



推理

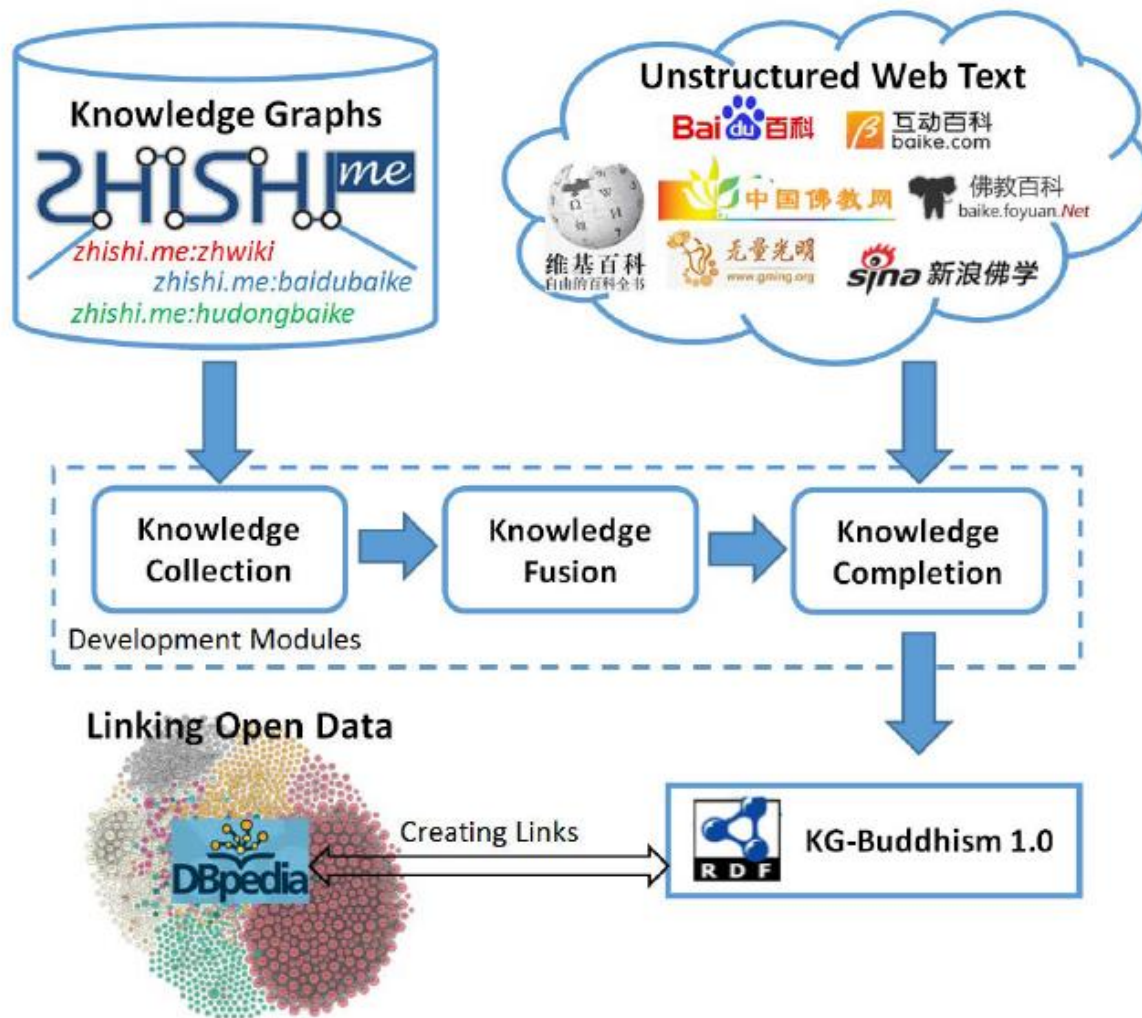
法比奥·卡纳瓦罗 国籍 意大利

目录

- 面向非结构化数据的知识抽取
- 面向结构化数据的知识抽取
- 面向半结构化数据的知识抽取
- **实践：基于百科数据的知识抽取**

在线百科知识抽取技术应用--佛学知识图谱构建

抽取框架



Knowledge Collection (Buddhist Figures)

佛

- 佛教之最 (1个页面)
- 佛教修行法門 (29个页面)
- 佛教历史 (8个分类, 90个页面)
- 佛教咒語 (15个页面)
- 佛教哲學 (3个分类, 6个页面)
- 佛教圣者 (3个分类)
- 佛教宗派 (10个分类, 11个页面)
- 佛教建筑 (1个分类, 1个页面)
- 佛教戒律 (2个分类, 23个页面)
- 佛教政治 (2个分类, 1个页面)
- 佛教文化 (12个分类, 15个页面)
- 佛教文学 (8个页面)

- 佛教文献 (6个分类, 2个页面)
- 佛教相關人物 (5个分类, 1个页面)
- 佛教研究 (2个分类, 6个页面)
- 佛教神祇 (8个分类, 37个页面)
- 佛教神話 (14个页面)
- 佛教組織 (5个分类, 8个页面)
- 佛教與其他宗教 (2个页面)
- 佛教與女性 (1个分类, 1个页面)
- 佛教节日 (3个分类, 5个页面)
- 佛教術語 (7个分类, 449个页面)
- 佛教題材作品 (5个分类, 12个页面)

反

- 反佛教 (1个分类, 13个页面)

• Category 方法

- 人工观察百科中与佛教人物相关的分类
- 抽取佛教人物分类下所有文章对应的实体

• 命名规则方法

例: “. + 菩萨” “. + 禅师”

- 维基百科“佛教头衔”分类下的所有实体
- 已抽取出的实体名中高频的公共字符串

Knowledge Fusion (Buddhist Figures)

□ 主语融合

- 实体的“别名”属性和重定向作为实体的别名集合
- 不同来源的实体存在一个完全匹配的别名则认为相同实体
- 人工检查相同实体数多于三个的映射

例：

Entities:

百度百科：	<code><http://zhishi.me/hudongbaike/resource/额尔德尼·确吉坚赞></code>
互动百科：	<code><http://zhishi.me/baidubaike/resource/班禅额尔德尼·确吉坚赞></code>
维基百科：	<code><http://zhishi.me/zhwiki/resource/十世班禅></code>



Alias Sets:

{十世班禅, 班禅额尔德尼·确吉坚赞, 第十世班禅额尔德尼}

{确吉坚赞, 班禅额尔德尼·确吉坚赞, 罗桑赤烈伦珠}

{班禅额尔德尼·确吉坚赞, 额尔德尼·确吉坚赞, 罗桑赤烈伦珠确吉坚赞}



Result:

`<http://www.kg-Buddhism.com/entity/额尔德尼·确吉坚赞>`

Knowledge Fusion (Buddhist Temples)

□ 主语融合 (寺庙知识图谱中的应用)

问题

—同名不同实体

<baidubaike/龙泉寺>: {龙泉寺、北京龙泉寺}

<zhwiki/龙泉寺>: {龙泉寺、南京龙泉寺}

—同实体不同名

<baidubaike/龙泉寺>: {龙泉寺、北京龙泉寺}

<hudongbaike/龙泉禅寺>: {龙泉禅寺、北京凤凰岭龙泉寺}

<zhwiki/龙泉寺_(海淀区)>: {龙泉寺_(海淀区)、海淀区龙泉寺}

Knowledge Fusion (Buddhist Temples)

□ 主语融合 (寺庙知识图谱中的应用) 方案

- 多个相同别名判断实体是否相同
- 实体的“地址”、“建造时间”属性是否冲突

例:

Entities: 百度百科: <http://zhishi.me/baidubaike/resource/三祖寺>
互动百科: <http://zhishi.me/hudongbaike/resource/乾元禅寺>
维基百科: <http://zhishi.me/zhwiki/resource/三祖寺>



Alias Sets: {三祖寺, 山谷寺}
{三祖寺, 山谷寺, 乾元禅寺}
{三祖寺, 三祖山谷乾元禅寺, 山谷寺, 乾元禅寺}
安徽省潜山县城西北9公里处的谷口凤形山上

Address: 安徽省潜山县天柱山风景区
安徽省安庆市天柱山南



Result: <http://www.kg-Buddhism.com/entity/三祖寺>

Knowledge Fusion

- 谓语融合

- Infobox属性

保留选定的15个佛学人物子属性与9个佛学寺庙子属性，人工总结每个属性在现有知识图谱中存在的谓语形式

- 其它属性

直接替换谓语的命名空间

- 宾语融合

- 单值属性

精确性原则：日期、地点等类型的属性值出现冲突时选择最精确的一个。

大多数原则：不同来源的属性值出现冲突时，选择出现次数最多的值。

- 多值属性

直接合并去重

Knowledge Fusion

triple 1:

<http://www.kg-Buddhism.com/entity/释净空>

<http://www.kg-Buddhism.com/property/出生日期>

“1927-03-28”@zh → 最精确

triple 2:

<http://www.kg-Buddhism.com/entity/释净空>

<http://www.kg-Buddhism.com/property/出生日期>

“1927”@zh

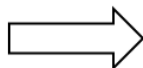
triple 3:

<http://www.kg-Buddhism.com/entity/释净空>

<http://www.kg-Buddhism.com/property/出生日期>

“1927”@zh

精确性原则



KG-Buddhism 1.0:

<http://www.kg-Buddhism.com/entity/释净空>

<http://www.kg-Buddhism.com/property/出生日期>

“1927-03-28”@zh

□ 宾语融合

■ 单值属性

精确性原则：日期、地点等类型的属性值出现冲突时选择最精确的一个。

大多数原则：不同来源的属性值出现冲突时，选择出现次数最多的值。

■ 多值属性

直接合并去重

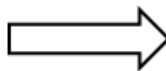
Knowledge Fusion

triple 1:
<http://www.kg-buddhism.com/entity/灵谷寺>
<http://www.kg-buddhism.com/property/地址>
"中山陵以东约1.5千米处"@zh

triple 2:
<http://www.kg-buddhism.com/entity/灵谷寺>
<http://www.kg-buddhism.com/property/地址>
"南京市中山陵东面1.5公里处"@zh

triple 3:
<http://www.kg-buddhism.com/entity/灵谷寺>
<http://www.kg-buddhism.com/property/地址>
"江苏省南京市"@zh

精确性原则



KG-Buddhism:
<http://www.kg-buddhism.com/entity/灵谷寺>
<http://www.kg-buddhism.com/property/地址>
"南京市中山陵东面1.5公里处"@zh

最精确

□ 宾语融合

■ 单值属性

精确性原则：日期、地点等类型的属性值出现冲突时选择最精确的一个。

大多数原则：不同来源的属性值出现冲突时，选择出现次数最多的值。

■ 多值属性

直接合并去重

Knowledge Fusion

triple 1:

<http://www.kg-Buddhism.com/entity/释星云>
<http://www.kg-Buddhism.com/property/出生日期>
“1927-09-19”@zh

triple 2:

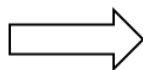
<http://www.kg-Buddhism.com/entity/释星云>
<http://www.kg-Buddhism.com/property/出生日期>
“1927-08-19”@zh

triple 3:

<http://www.kg-Buddhism.com/entity/释星云>
<http://www.kg-Buddhism.com/property/出生日期>
“1927-08-19”@zh

大多数

大多数原则



KG-Buddhism 1.0:

<http://www.kg-Buddhism.com/entity/释星云>
<http://www.kg-Buddhism.com/property/出生日期>
“1927-08-19”@zh

□ 宾语融合

■ 单值属性

精确性原则：日期、地点等类型的属性值出现冲突时选择最精确的一个。

大多数原则：不同来源的属性值出现冲突时，选择出现次数最多的值。

■ 多值属性

直接合并去重

Knowledge Completion (Buddhist Figures)

□ 对infobox属性进行补全

- 人工编写规则从非结构化文本中抽取属性值
- 依照知识融合方法将属性值对转换为三元组

例：

Text：

title: 李叔同

李叔同，谱名文涛，幼名成蹊，学名广侯，字息霜，别号漱筒；出家后法名演音，号弘一，晚号晚晴老人。

Pattern：

.*(法名|法号)(为|曰|称|叫|即){0,1}([\S]+?)(字|祖籍|,|。){0,1}.*

Property-value pair：

法名：（李叔同，演

Triple：

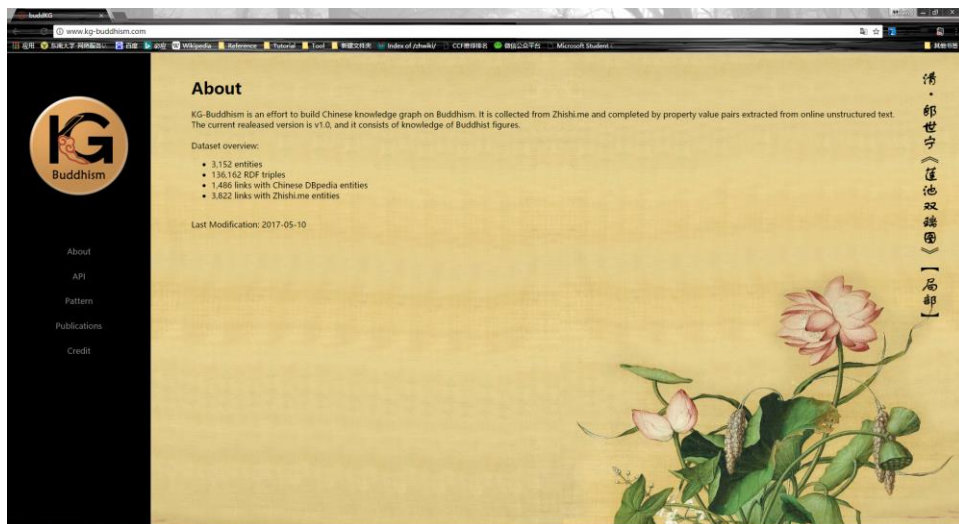
<http://www.kg-buddhism.com/entity/李叔同>
<http://www.kg-buddhism.com/property/法名>
"演音"@zh

Application

□ 在线接口

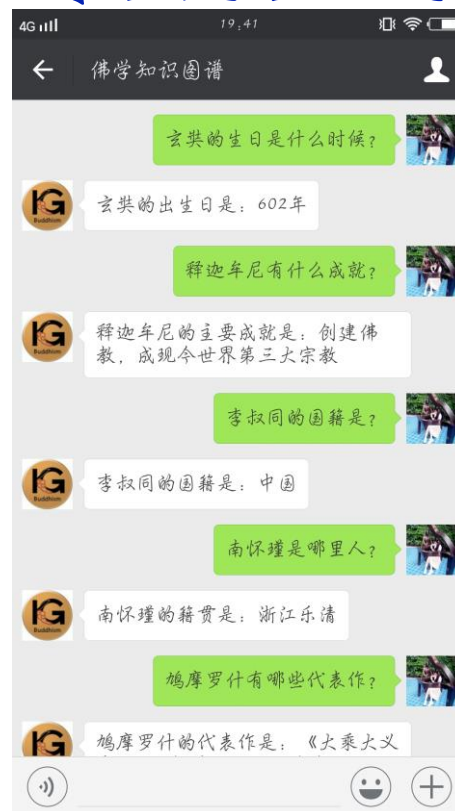
■ 数据访问地址:

www.kg-buddhism.com



■ 应用

问答微信公众号



实践任务一

□ 属性-值 抽取

Description:

实体的Abstract属性值是一段描述实体的文本，其中包含实体的一些基本属性。观察abstract文本，人工总结模板，使用模板从abstract文本中抽取指定的属性值。

可选属性：别名、始建时间、宗派、建筑风格

Input:

文件	说明
abstracts.ttl	DBpedia中佛教寺庙实体的abstract属性

```
<http://zh.dbpedia.org/resource/一柱寺> <http://dbpedia.org/ontology/abstract> "一柱寺（越南语：Chùa Một Cột / 一柱），正式名称延祐寺（越南语：Diên Hựu tự / 延祐寺），或莲花台（越南语：Liên Hoa Đài / 莲台）"
<http://zh.dbpedia.org/resource/一烟寺> <http://dbpedia.org/ontology/abstract> "医王山 一烟寺为位于日本岛根县出云市小境町之临济宗妙心寺派寺院，也是一烟药师教团的总本山。通称为一烟药师。以「目之药师様」
<http://zh.dbpedia.org/resource/一粟庵> <http://dbpedia.org/ontology/abstract> "一粟庵为旧时上海名庵之一，与清莲庵、铎庵合称为清代上海“三大名庵”。一粟庵的旧址明代时为徐光启私家花园。康熙七年（1666年）
<http://zh.dbpedia.org/resource/七塔寺> <http://dbpedia.org/ontology/abstract> "七塔寺全称七塔报恩禅寺，是浙江省宁波市一座禅宗佛寺，位于江东区百丈路，始建于唐大中十二年（858年），是宁波城区唯一保存完
<http://zh.dbpedia.org/resource/万佛寺> <http://dbpedia.org/ontology/abstract> "万佛寺（Ten Thousand Buddhas Monastery）位于香港沙田区排头村后万佛山上，由月溪法师在1949年建设，于1957年落成，寺院占地65
<http://zh.dbpedia.org/resource/万佛寺_成都> <http://dbpedia.org/ontology/abstract> "万佛寺是中国成都一座已经不存在的佛教寺庙，遗址位于市一环路北二段与白马寺街交叉路口北侧，即成都老城西门外。"@zh .
<http://zh.dbpedia.org/resource/万佛寺_雾峰区> <http://dbpedia.org/ontology/abstract> "万佛寺是位于台湾台中市雾峰区的佛教寺庙，是中台湾著名的佛教寺院。原寺在921大地震中倒塌，至今仍在重建中。"@zh .
<http://zh.dbpedia.org/resource/万固寺> <http://dbpedia.org/ontology/abstract> "万固寺，位于山西省运城市永济市蒲州镇胜利庄南300米，是中华人民共和国山西省文物保护单位之一。1965年5月24日，授予山西省文物
<http://zh.dbpedia.org/resource/万寿兴隆寺> <http://dbpedia.org/ontology/abstract> "万寿兴隆寺是清代故宫外八庙之一，在今北京市西城区北长街路西，邻近紫禁城西华门，与福佑寺隔街相对。"@zh .
<http://zh.dbpedia.org/resource/万寿寺_北京> <http://dbpedia.org/ontology/abstract> "万寿寺位于北京市海淀区万寿寺路121号，是一座汉传佛教寺院，为全国重点文物保护单位，现辟为北京艺术博物馆。"@zh .
<http://zh.dbpedia.org/resource/万寿寺_张掖> <http://dbpedia.org/ontology/abstract> "万寿寺位于中国甘肃省张掖市甘州区县府街，又称木塔寺，1958年被列为张掖县文物保护单位，1993年被列为甘肃省文物保护单位
<http://zh.dbpedia.org/resource/万寿弥陀寺> <http://dbpedia.org/ontology/abstract> "万寿弥陀寺是位于中国北京市西城区鼓楼西大街63号的一座汉传佛教寺院。"@zh .
```

主语（实体）：

<http://zh.dbpedia.org/resource/万佛寺_(成都)>

实践任务一

□ 属性-值 抽取

Description:

实体的Abstract属性值是一段描述实体的文本，其中包含实体的一些基本属性。观察abstract文本，人工总结模板，使用模板从abstract文本中抽取指定的属性值。

可选属性：别名、始建时间、宗派、建筑风格

Input:

文件	说明
abstracts.ttl	DBpedia中佛教寺庙实体的abstract属性

```
<http://zh.dbpedia.org/resource/一柱寺> <http://dbpedia.org/ontology/abstract> "一柱寺（越南语：Chùa Một Cột / 一柱），正式名称延祐寺（越南语：Diên Hựu tự / 延祐寺），或莲花台（越南语：Liên Hoa Đài / 莲台）"
<http://zh.dbpedia.org/resource/一烟寺> <http://dbpedia.org/ontology/abstract> "医王山 一烟寺为位于日本岛根县出云市小境町之临济宗妙心寺派寺院，也是一烟药师教团的总本山。通称为一烟药师。以「目之药师様」
<http://zh.dbpedia.org/resource/一粟庵> <http://dbpedia.org/ontology/abstract> "一粟庵为旧时上海名庵之一，与清莲庵、铎庵合称为清代上海“三大名庵”。一粟庵的旧址明代时为徐光启私家花园。康熙七年（1666年）
<http://zh.dbpedia.org/resource/七塔寺> <http://dbpedia.org/ontology/abstract> "七塔寺全称七塔报恩禅寺，是浙江省宁波市一座禅宗佛寺，位于江东区百丈路，始建于唐大中十二年（858年），是宁波城区唯一保存完
<http://zh.dbpedia.org/resource/万佛寺> <http://dbpedia.org/ontology/abstract> "万佛寺（Ten Thousand Buddhas Monastery）位于香港沙田区排头村后万佛山上，由月溪法师在1949年建设，于1957年落成，寺院占地65
<http://zh.dbpedia.org/resource/万佛寺（成都）> <http://dbpedia.org/ontology/abstract> "万佛寺是中国成都一座已经不存在的佛教寺庙，遗址位于市一环路北二段与白马寺街交叉路口北侧，即成都老城西门外。"@zh .
<http://zh.dbpedia.org/resource/万佛寺（雾峰区）> <http://dbpedia.org/ontology/abstract> "万佛寺是位于台湾台中市雾峰区的佛教寺庙，是中台湾著名的佛教寺院。原寺在921大地震中倒塌，至今仍在重建中。"@zh .
<http://zh.dbpedia.org/resource/万固寺> <http://dbpedia.org/ontology/abstract> "万固寺，位于山西省运城市永济市蒲州镇胜利庄南300米，是中华人民共和国山西省文物保护单位之一。1965年5月24日，授予山西省文物
<http://zh.dbpedia.org/resource/万寿兴隆寺> <http://dbpedia.org/ontology/abstract> "万寿兴隆寺是清代故宫外八庙之一，在今北京市西城区北长街路西，邻近紫禁城西华门，与福佑寺隔街相对。"@zh .
<http://zh.dbpedia.org/resource/万寿寺（北京）> <http://dbpedia.org/ontology/abstract> "万寿寺位于北京市海淀区万寿寺路121号，是一座汉传佛教寺院，为全国重点文物保护单位，现辟为北京艺术博物馆。"@zh .
<http://zh.dbpedia.org/resource/万寿寺（张掖）> <http://dbpedia.org/ontology/abstract> "万寿寺位于中国甘肃省张掖市甘州区县府街，又称木塔寺，1958年被列为张掖县文物保护单位，1993年被列为甘肃省文物保护单位
<http://zh.dbpedia.org/resource/万寿弥陀寺> <http://dbpedia.org/ontology/abstract> "万寿弥陀寺是位于中国北京市西城区鼓楼西大街63号的一座汉传佛教寺院。"@zh .
```

谓语句（属性名）：

<http://dbpedia.org/ontology/abstract>



小象学院
ChinaHadoop.cn

实践任务一

□ 属性-值 抽取

Description:

实体的Abstract属性值是一段描述实体的文本，其中包含实体的一些基本属性。观察abstract文本，人工总结模板，使用模板从abstract文本中抽取指定的属性值。

可选属性：别名、始建时间、宗派、建筑风格

Input:

文件	说明
abstracts.ttl	DBpedia中佛教寺庙实体的abstract属性

```
<http://zh.dbpedia.org/resource/一柱寺> <http://dbpedia.org/ontology/abstract> "一柱寺（越南语：Chùa Một Cột / 一柱），正式名称延祐寺（越南语：Diên Hựu tự / 延祐寺），或莲花台（越南语：Liên Hoa Đài / 莲花台），是一座位于越南河内市的佛教寺庙。该寺始建于1049年，由李朝皇帝李日尊下令建造。该寺是一座独特的寺庙，其结构由一根巨大的石柱支撑，石柱上雕刻着精美的佛像。该寺是越南最重要的佛教寺庙之一，也是世界文化遗产。" @zh
<http://zh.dbpedia.org/resource/一烟寺> <http://dbpedia.org/ontology/abstract> "一烟寺（日语：一煙寺）是位于日本岛根县出云市小境町之临济宗妙心寺派寺院。也是一烟药师教团的总本山。通称为一烟药师。以「目之药师様」而闻名。" @zh
<http://zh.dbpedia.org/resource/一茶庵> <http://dbpedia.org/ontology/abstract> "一茶庵为旧时上海名庵之一，与清莲庵、静庵合称为清代上海“三人名庵”。一茶庵的旧址明代时为徐氏私家花园，康熙七年（1666年）被毁。一茶庵遗址现为静安寺公园。" @zh
<http://zh.dbpedia.org/resource/七塔报恩禅寺> <http://dbpedia.org/ontology/abstract> "七塔报恩禅寺，是浙江省宁波市一座禅宗佛寺，位于江东区百丈路，始建于唐大中十二年（858年），是宁波城区唯一保存完好的佛寺。" @zh
<http://zh.dbpedia.org/resource/万佛寺> <http://dbpedia.org/ontology/abstract> "万佛寺（Ten Thousand Buddhas Monastery）位于香港沙田区排头村后万佛山上，由月溪法师在1949年建设，于1957年落成，寺院占地65公顷。" @zh
<http://zh.dbpedia.org/resource/万佛寺（成都）> <http://dbpedia.org/ontology/abstract> "万佛寺是中国成都一座已经不存在的佛教寺庙。遗址位于市一环路北二段与白马寺街交叉路口北侧，即成都老城西门外。" @zh
<http://zh.dbpedia.org/resource/万佛寺（雾峰区）> <http://dbpedia.org/ontology/abstract> "万佛寺是位于台湾台中市雾峰区的佛教寺庙，是中台湾著名的佛教寺院。原寺在921大地震中倒塌，至今仍在重建中。" @zh
<http://zh.dbpedia.org/resource/万佛寺（万固寺）> <http://dbpedia.org/ontology/abstract> "万固寺，位于山西省运城市永济市蒲州镇胜利庄南300米，是中华人民共和国山西省文物保护单位之一。1965年5月24日，授予山西省文物保护单位。" @zh
<http://zh.dbpedia.org/resource/万寿兴隆寺> <http://dbpedia.org/ontology/abstract> "万寿兴隆寺是清代故宫外八庙之一，在今北京市西城区北长街路西，邻近紫禁城西华门，与福佑寺隔街相对。" @zh
<http://zh.dbpedia.org/resource/万寿寺（北京）> <http://dbpedia.org/ontology/abstract> "万寿寺位于北京市海淀区万寿寺路121号，是一座汉传佛教寺院，为全国重点文物保护单位，现辟为北京艺术博物馆。" @zh
<http://zh.dbpedia.org/resource/万寿寺（张掖）> <http://dbpedia.org/ontology/abstract> "万寿寺位于中国甘肃省张掖市甘州区县府街，又称木塔寺，1958年被列为张掖县文物保护单位，1993年被列为甘肃省文物保护单位。" @zh
<http://zh.dbpedia.org/resource/万寿弥陀寺> <http://dbpedia.org/ontology/abstract> "万寿弥陀寺是位于中国北京市西城区鼓楼西大街63号的一座汉传佛教寺院。" @zh
```

宾语（属性值）：

"万佛寺是中国成都一座已经不存在的佛教寺庙，遗址位于市一环路北二段与白马寺街交叉路口北侧，即成都老城西门外。"@zh。

实践任务一

□ 属性-值 抽取

Example:

目标属性	模板
地址	① 位于 $[\wedge, \wedge.]^+$ ② 坐落于 $[\wedge, \wedge.]^+$

主语（实体）：	<http://zh.dbpedia.org/resource/万佛寺_(成都)>
谓语（属性名）：	<http://dbpedia.org/ontology/abstract>
宾语（属性值）：	"万佛寺是中国成都一座已经不存在的佛教寺庙，遗址位于市一环路北二段与白马寺街交叉路口北侧，即成都老城西门外。"@zh.

↓

位于 $[\wedge, |\wedge.]^+$

↓

<http://zh.dbpedia.org/resource/万佛寺_(成都)>
位于市一环路北二段与白马寺街交叉路口北侧

实践任务一

□ 解决方案

目标属性	模板
别名	1. (俗称 原名 又名 亦称)[^, ^。]+
始建时间	1. (建造于 初建于 始建于 建于)[^, ^。]+ 2. (于 在).+(初 中叶 末 时代 年 年代)(创建 建成 修建)
宗派	1. (属于 是 为)[^, ^。]+(派 宗)
建筑风格	1. (属于 为 是 有 按)[^, ^。]+风格

实践任务二

□ 实体清理

Description:

在线百科对实体的分类通常存在错误，因此使用category属性抽取的实体中也会存在噪音。利用实体的多种属性去除非佛教寺庙实体。

Input:

文件	说明
entities.txt	使用category从DBpedia中抽取的佛教寺庙实体
abstracts.ttl	entities.txt文件中实体的abstract属性
aliases.ttl	entities.txt文件中实体的别名属性
categories.ttl	entities.txt文件中实体的类别属性
sections.ttl	entities.txt文件中实体的宗派属性

实践任务二

□ 解决方案

Hint	Example
实体名中的特征词	1. .*(礼拜 先知 清真 回教 大 女)寺 2. (题 宿 过 辞 别 赠 记 寄).+
实体abstract中的特征词	1. 官署 2. 清真、阿訇、穆斯林、伊斯兰教、 礼拜、天主教 3. 《实体名》、出版
实体的category中包含非佛教教派	1. 伊斯兰教 2. 天主教
实体section属性值非佛教宗派	1. 木教派 2. 虎非耶教派 3. 哲合林耶教派

谢谢大家！

本课程课件由OpenKG提供支持

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop

