语义网与知识图谱

上海大学计算机学院

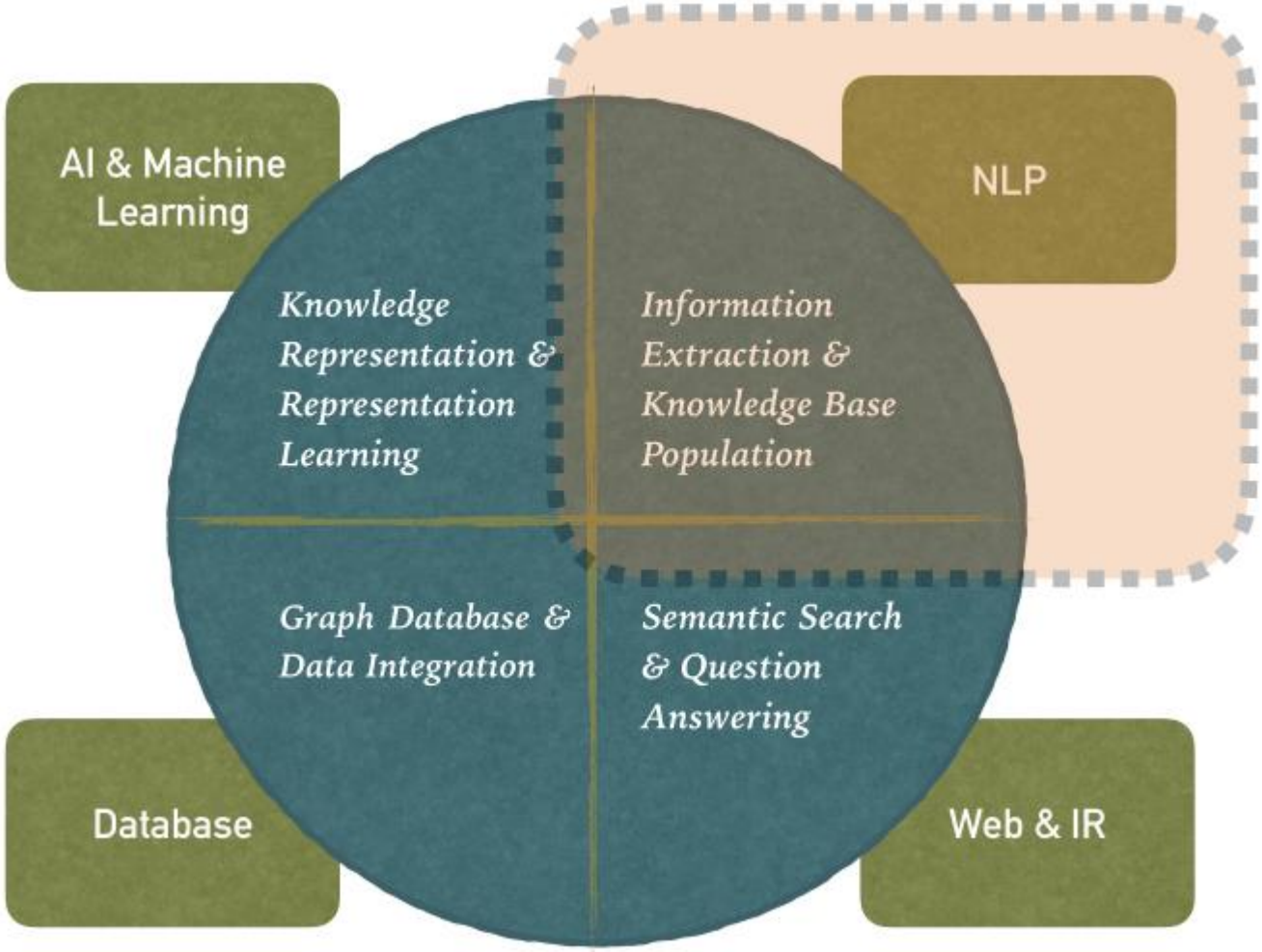
主讲： 刘 炜

知识抽取

上海大学计算机学院 刘炜

2020年10月

知识图谱是一个交叉研究领域



内容概览

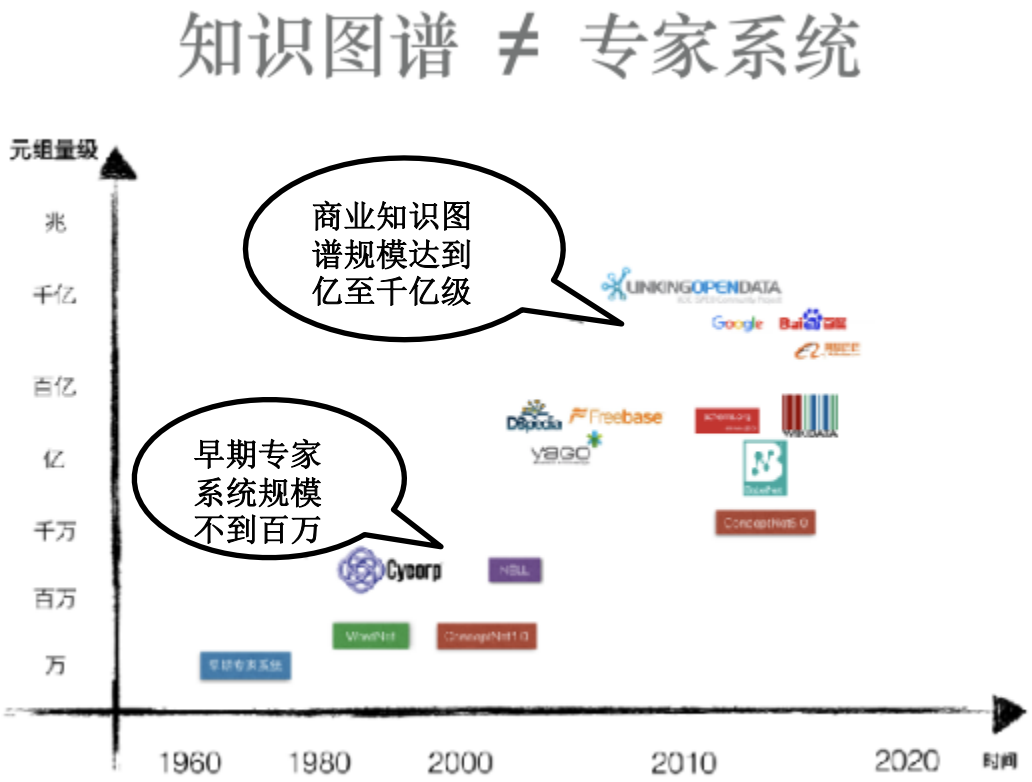
一、知识获取与知识图谱获取

二、面向结构化的知识抽取

三、面向非结构化的知识抽取：实体抽取

四、面向非结构化的知识抽取：关系抽取

知识图谱的获取与构建



冯诺依曼曾估计单个个体的大脑中的全量知识需要 $2.4*10^{20}$ 字节存储，知识工程的根本性科学问题是知识完备性问题，即规模化自动化知识获取与处理能力。

人工 — 高阶谓词逻辑

CYC - $5.71 per statement
Freebase - $2.25 per statement
NELL - 14.25₵ per statement
DBPedia - 1.85₵ per statement
Yago - 0.83₵ per statementure

How much is a Triple? ISWC2018

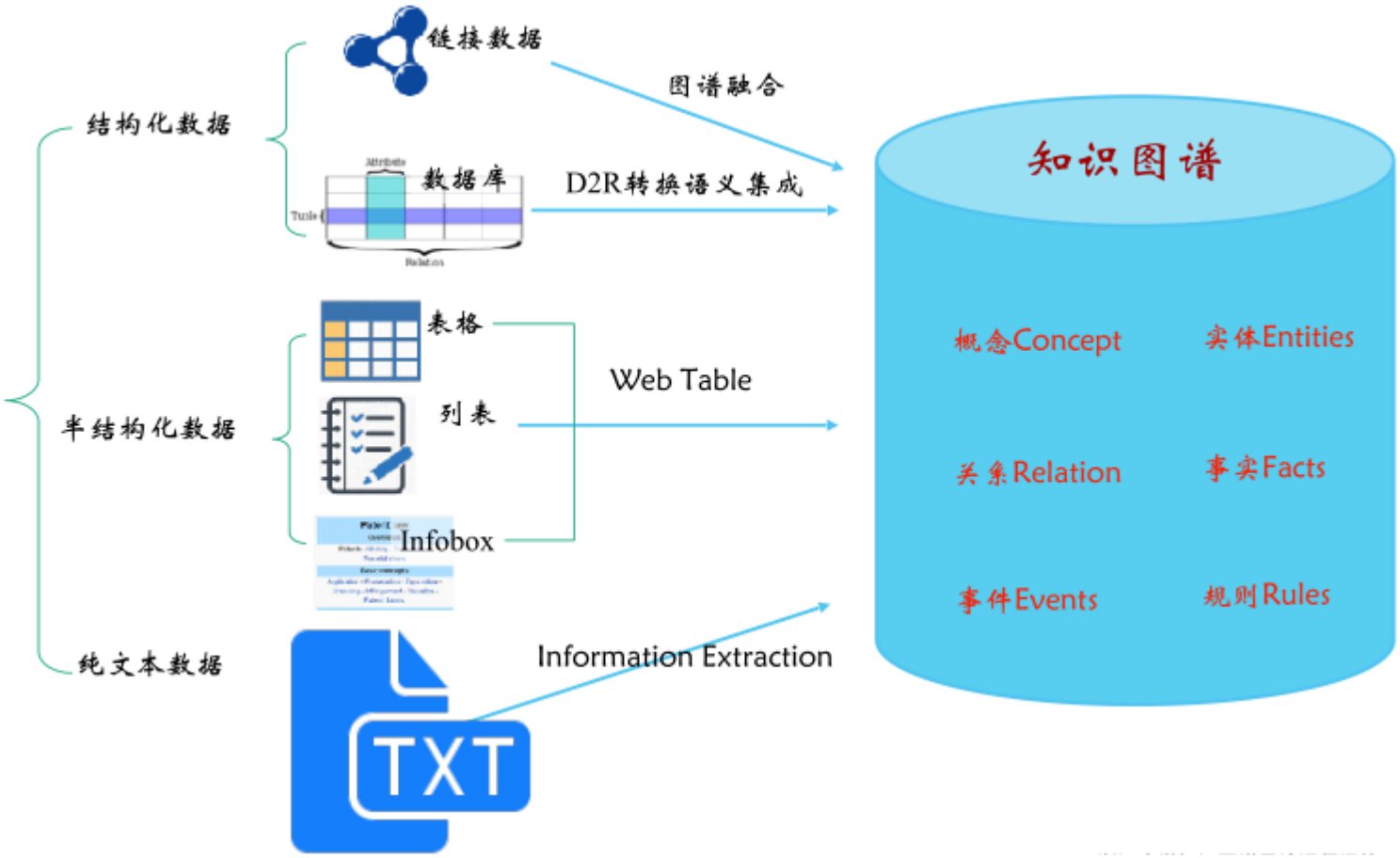
自动化 — 以三元组为主

知识图谱获取的内容


概念抽取
实体抽取
关系抽取
事件抽取
规则抽取
DBpedia
Freebase
yago
Cycorp
Frame Net
Google
Bai du 百度
知识抽取技术
数据库
知识库

知识图谱工程

知识抽取是实现自动化构建大规模知识图谱的重要技术。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。文本一般不作为知识图谱构建的初始来源，而多用来做知识图谱补全。



从文本获取——文本知识抽取任务

□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(概念抽取)

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]

从文本获取——文本知识抽取任务

口事件抽取

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： -

口规则抽取：IF this THEN that

信息抽取相关的竞赛与数据集

消息理解会议 (Message Understanding Conference，MUC)

由美国DARPA启动并资助的项目，目的是鼓励和开发更好的信息抽取方法

Conference	*Year*	*Text Source*	*Topic (Domain)*
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

在MUC的评测中，召回率(Recall)和精确率(Precision)是评价信息抽取系统性能的两个重要评价指标。召回率是系统抽取的正确结果占标准结果的比例；精确率是系统抽取的正确结果占其抽取的所有结果的比例。为了综合两个方面的因素考量系统的性能，通常基于召回率和准确率计算F1值。

MUC实体识别实例

- 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>

自动内容抽取(Automatic Content Extraction, ACE)

美国国家标准技术所(NIST)组织的测评会，从1999-2008举办了8次，ACE对MUC定义的任务进行了融合、分类和细化；主要分为三大任务，包含英语、阿拉伯语和汉语；

(1)实体检测和跟踪。这是 ACE 最基础和核心的任务，该任务要求识别文本中的实体，实体类型包括人物(Person,PER)、组织(Organization,ORG)、设施(Facility, FAC)、地缘政治实体(Geographical Political Entity,GPE)和位置(Location,LOC)等。

(2)关系检测与表征。该任务要求识别和表征实体间的关系，关系被分为五大类，包括角色(role)关系、部分整体(part-whole)关系、位于(at)关系、邻近(near)关系和社会(social)关系，每个大类关系又被进一步细分，总共有24种类型。

Automatic Content Extraction (ACE)

(3) 事件检测与表征。该任务要求识别实体参与的五类事件，包括交互(interaction)、移动(movement)、转移(transfer)、创建(creation)和销毁(destruction)事件。任务要求自动标注每个事件的文本提及或锚点，并按类型和子类型对其进行分类；最后，还需要根据类型特定的模板进一步确定事件参数和属性。

- 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)

知识库填充(KBP)

TAC Knowledge Base Population(KBP，知识库填充)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，从2009-2017，举办九届。 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

知识库填充(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

端到端冷启动知识库构建任务基于给定的知识库模式(KB schema)从文本中获取以下信息：实体，在实体发现与链接任务中定义的实体和实体提及；槽关系，在槽填充中涉及的实体属性(“槽”)；事件，在事件跟踪任务中的事件和事件块；事件参数，在事件跟踪任务中的事件参数；情绪，信念和情感任务中源实体向目标实体的情绪。

语义测评会议(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。1998-2017举办了11届。

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

早期评测比较关注词义消歧问题，后来出现了更多文本语义理解的任务，包括语义角色标注、情感分析、跨语言语义分析等。

内容概览

一、知识获取与知识图谱获取

二、面向结构化的知识抽取

三、面向非结构化的知识抽取：实体抽取

四、面向非结构化的知识抽取：关系抽取

直接映射方法

- 直接映射规范定义了一个从关系数据库到 RDF 图数据的简单转换，为定义和比较更复杂的转换提供了基础。
- 直接映射将关系数据库表结构和数据直接转换为RDF图，关系数据库的数据结构直接反映在RDF图中。直接映射的基本规则包括：
  - 数据库中的表映射为RDF类；
  - 数据库中表的列映射为RDF属性；
  - 数据库表中每一行映射为一个资源或实体，创建IRI；
  - 数据库表中每个单元格的值映射为一个文字值(Literal Value)；如果单元格的值对应一个外键，则将其替换为外键值指向的资源或实体的IRI。

直接映射实例

```
CREATE TABLE "Addresses" (
"ID" INT, PRIMARY KEY("ID"),
"city" CHAR(10),
"state" CHAR(2)
)

CREATE TABLE "People" (
"ID" INT, PRIMARY KEY("ID"),
"fname" CHAR(10),
"addr" INT,
FOREIGN KEY("addr")REFERENCES "Addresses"("ID")
)

INSERT INTO "Addresses" ("ID", "city", "state")VALUES (18, 'Cambridge',
'MA')
INSERT INTO "People" ("ID", "fname", "addr")VALUES (7, 'Bob', 18)
INSERT INTO "People" ("ID", "fname", "addr")VALUES (8, 'Sue', NULL)
```

People 表

PK		→ Address(ID)
ID	**fname**	**Address(ID)**
7	Bob	18
8	Sue	**NULL**

Address 表

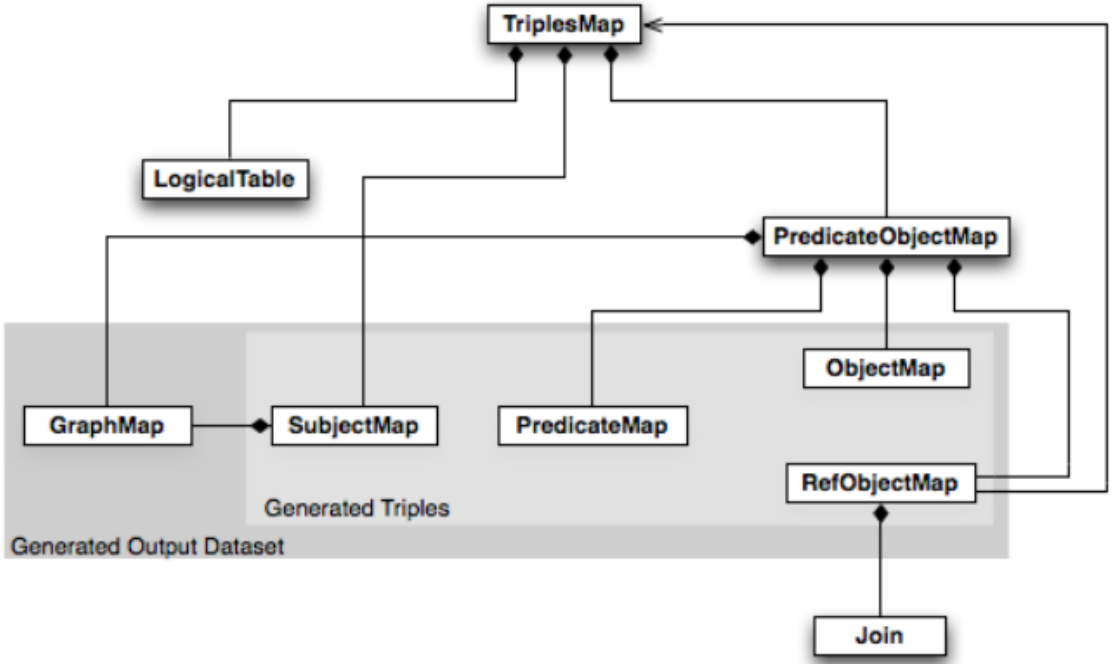
PK		
ID	**city**	**state**
18	Cambridge	MA

```
<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方法

- R2RML映射语言是一种用于表示从关系数据库到RDF数据集的自定义映射的语言。这种映射提供了在 RDF 数据模型下查看现有关系型数据的能力，并且可以基于用户自定义的结构和目标词汇表示原有的关系型数据。
- 在数据库的直接映射中，生成的 RDF 图的结构直接反映了数据库的结构，目标 RDF 词汇直接反映数据库模式元素的名称，结构和目标词汇都不能改变。
- R2RML 映射是通过逻辑表(Logic Tables)从数据库中检索数据的。一个逻辑表可以是数据库中的一个表、视图或有效的 SQL 语句查询。



R2RML实例

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

DEPT（部门）		
DEPTNO INTEGER PRIMARY KEY	DNAME VARCHAR(30)	LOC 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.
```

Example R2RML mapping

```
@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" ];
    ].
```

Example output data

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

内容概览

一、知识获取与知识图谱获取

二、面向结构化的知识抽取

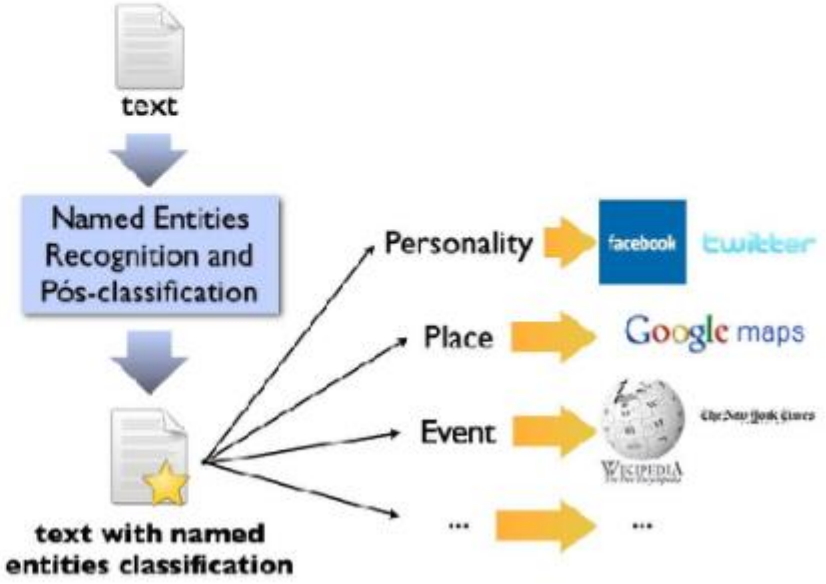
三、面向非结构化的知识抽取：实体抽取

四、面向非结构化的知识抽取：关系抽取

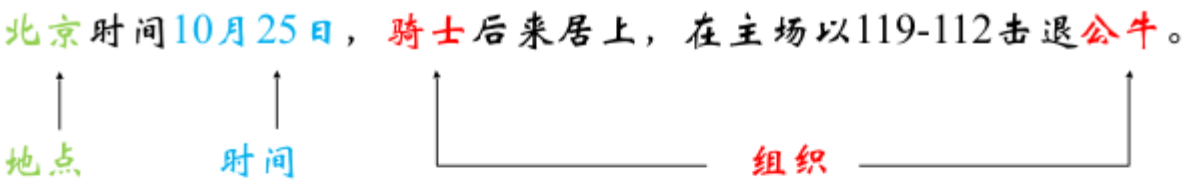
NER：从文本中识别实体

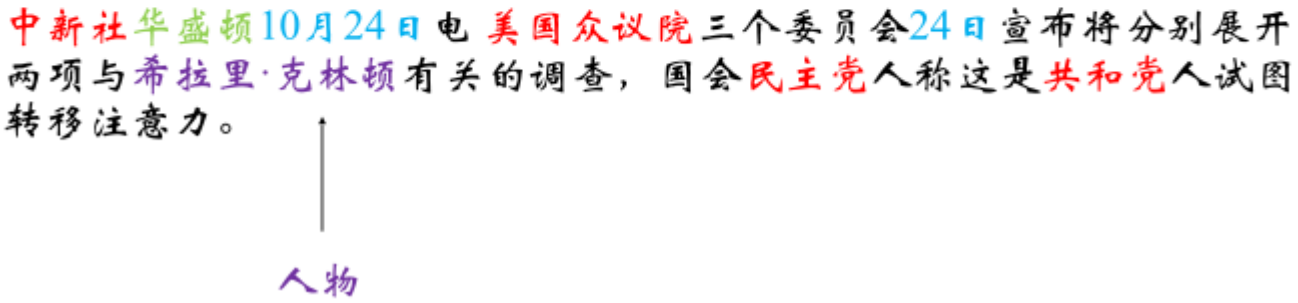
实体抽取又称命名实体识别，实体抽取是解决很多自然语言处理问题的基础，也是知识抽取中最基本的任务。

- 任务定义：抽取文本中的原子信息元素，是进一步实现关系抽取的基础
- 通用领域实体类型举例
  - 人名
  - 组织/机构名
  - 地理位置
  - 时间/日期
  - 字符值
  - 金额值



实体识别举例


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


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

实体识别常用方法

- 基于模板和规则的方法
- 基于统计模型的方法
- 基于深度学习的方法

基于模板与规则的方法

- 将文本与规则进行匹配来识别出命名实体
  - “***说”、“***老师”;
  - “***大学”、“***医院”

- 优点:
  - 准确，有些实体识别只能依靠规则抽取
- 缺点:
  - 需要大量的语言学知识
  - 需要谨慎处理规则之间的冲突问题;
  - 构建规则的过程费时费力、可移植性不好。

基于统计模型的方法

- 基于统计模型的方法利用完全标注或部分标注的语料进行模型训练，主要采用的模型包括隐马尔可夫模型(Hidden Markov Model)、条件马尔可夫模型(Conditional Markov Model)、最大熵模型(Maximum Entropy Model)以及条件随机场模型(Conditional Random Fields)。
- 该类方法将命名实体识别作为序列标注问题处理。与普通的分类问题相比，序列标注问题中当前标签的预测不仅与当前的输入特征相关，还与之前的预测标签相关，即预测标签序列是有强相互依赖关系的。从自然文本中识别实体是一个典型的序列标注问题。
- 基于统计模型构建命名实体识别方法主要涉及：(1)训练语料标注、(2)特征定义和 (3)模型训练三个方面。

(1) 训练语料标注

- IOB 标注体系中：文本中的每个词被标记为实体名称的起始词(B)、实体名称的后续词(I)或实体名称的外部词(O)。
- IO标注体系中：文本中的词被标记为实体名称内部词(I)或实体名称外部词(O)。

表4-1 IOB和IO实体标注示例

标注体系	苹	果	公	司	是	一	家	美	国	的	跨	国	公	司
IOB 标注	B-ORG	I-ORG	I-ORG	I-ORG	O	O	O	B-ORG	I-ORG	O	O	O	O	O
IO 标注	I-ORG	I-ORG	I-ORG	I-ORG	O	O	O	I-ORG	I-ORG	O	O	O	O	O

	IOB标注体系	IO标注体系		IOB标注体系	IO标注体系
由	O	O	清	B-ORG	I-ORG
浙	B-ORG	I-ORG	华	I-ORG	I-ORG
江	I-ORG	I-ORG	大	I-ORG	I-ORG
大	I-ORG	I-ORG	学	I-ORG	I-ORG
学	I-ORG	I-ORG	的	O	O
的	O	O	李	B-PER	I-PER
张	B-PER	I-PER	大	I-PER	I-PER
小	I-PER	I-PER	大	I-PER	I-PER
小	I-PER	I-PER			
迎	O	O			
战	O	O			

序列标签体系

角色	意义	例子
A	上文	参与亚太经合组织的活动
B	下文	中央电视台报道
X	连接词	北京电视台和天津电视台
C	特征词的一般性前缀	北京电影学院
F	特征词的人名前缀	何镜堂纪念馆
G	特征词的地名性前缀	交通银行北京分行
K	特征词的机构名、品牌名前缀	中共中央顾问委员会 美国摩托罗拉公司
I	特征词的特殊性前缀	中央电视台、中海油集团
J	特征词的简称性前缀	巴政府
D	机构名的特征词	国务院侨务办公室
S	开始标志	始##始

(2) 特征定义

- **在训练模型之前，统计模型需要计算每个词的一组特征作为模型的输入。这些特征具体包括单词级别特征、词典特征和文档级特征等。**

□ 单词级别特征
- 是否首字母大写、是否以句点结尾
- 是否包含数字
- 词性、词的n-gram等

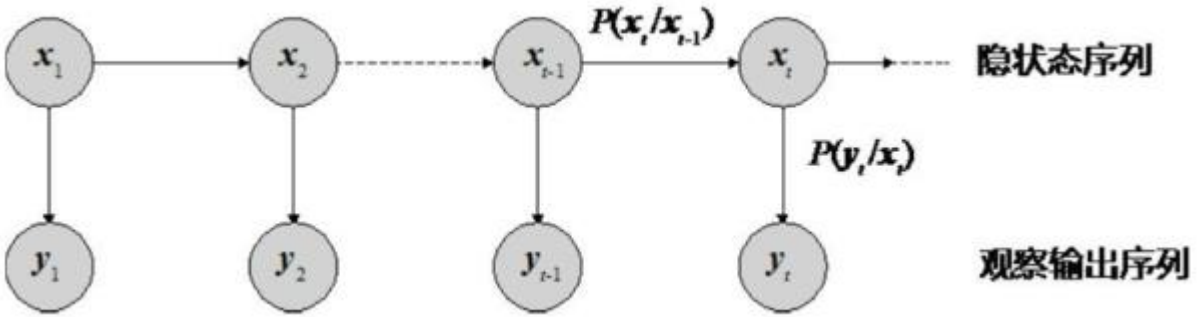
□ 词典特征
- 依赖外部词典定义
- 预定义的词表
- 地名列表等

□ 文档级特征
- 基于整个语料文档集计算
- 例如文档集中的词频、共现词等

- Stanford NER 模型中定义的特征包括当前词、当前词的前一个词、当前词的后一个词、当前词的字符n-gram、当前词的词性、当前词上下文词性序列、当前词的词形、当前词上下文词形序列、当前词左侧窗口中的词(窗口大小为4)、当前词右侧窗口中的词(窗口大小为4)。
- 定义何种特征对于命名实体识别结果有较大的影响，因此不同命名实体识别算法使用的特征有所不同。

(3) 模型训练

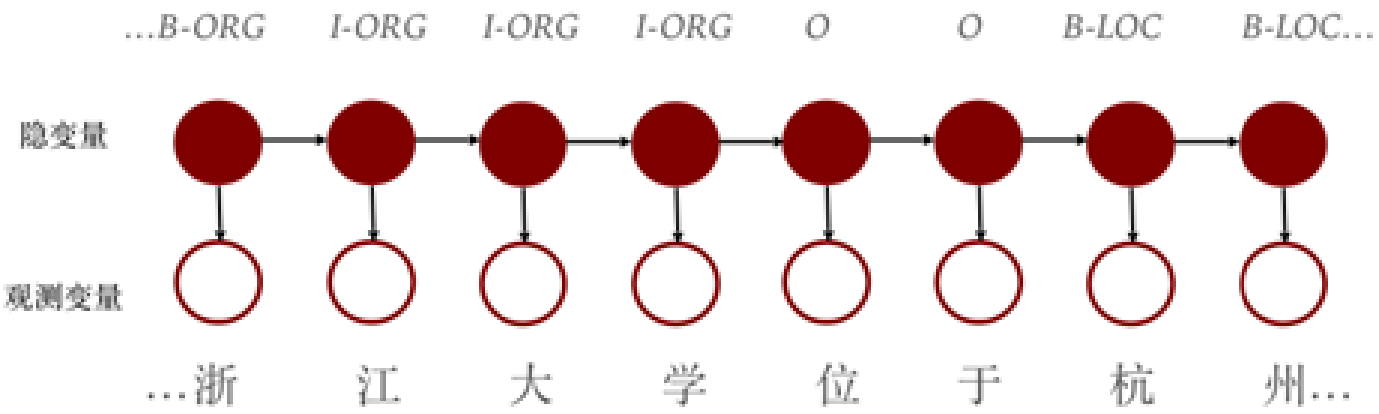
- 隐马尔可夫模型(Hidden Markov Model,HMM)和条件随机场(Conditional Random Field,CRF)是两个常用于标注问题的统计学习模型，也被广泛应用于实体抽取问题。
- HMM 是一种有向图概率模型，模型中包含了隐藏的状态序列和可观察的观测序列。每个状态代表了一个可观察的事件，观察到的事件是状态的随机函数。



- 在任意 t 时刻的状态只依赖于其前一时刻的状态，与其他观测及状态无关，即 $P(x_t|x_{t-1},x_{t-2},\ldots,x_1,y_{t-1},y_{t-2},\ldots,y_1)=P(x_t|x_{t-1})$;
- 任意时刻的观测只依赖于该时刻的马尔可夫链的状态，与其他观测及状态无关，即:$P(y_t|x_t,x_{t-1},x_{t-2},\ldots,x_1,y_{t-1},y_{t-2},\ldots,y_1)=P(y_t|x_t)$

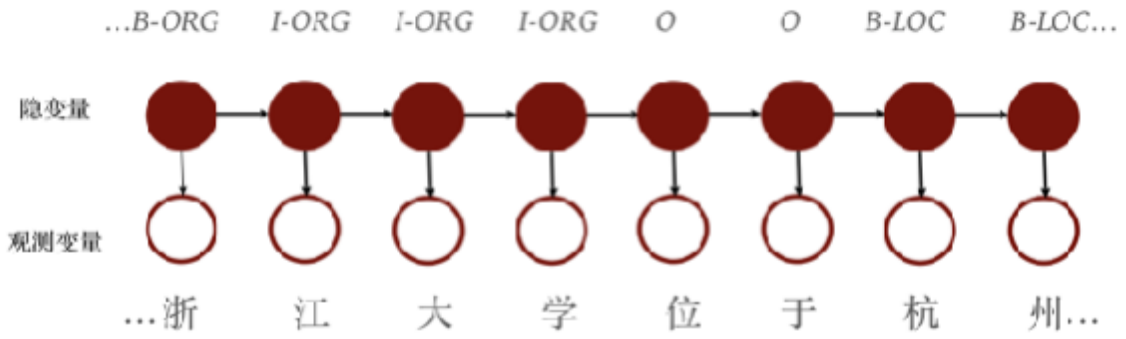
(3) 模型训练

- HMM模型中的状态对应词的的标记，标注问题可以看做是对给定的观测序列进行序列标注。



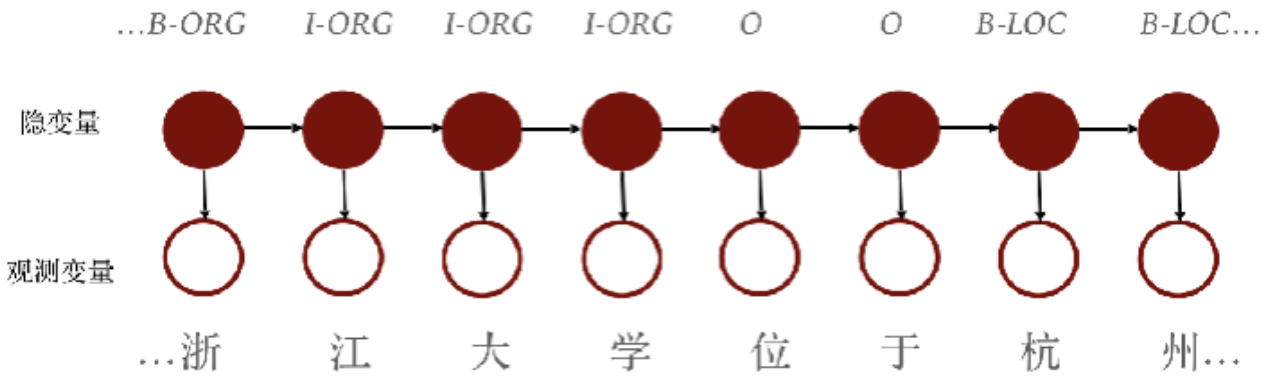
HMM 的要素定义

- 隐藏状态$Q=\{q_1,q_2,\ldots,q_N\}$，这里对应所有可能的标签集合，N是隐藏状态数。
- 观测状态$V=\{v_1,v_2,\ldots,v_M\}$，这里对应所有可能的词的集合，M是观测状态数。
- 对于一个长度为T的序列，I对应状态序列（标签序列），O对应观测序列（词序列），即：
  - $I=\{i_1,i_2,\ldots,i_T\}$，$O=\{o_1,o_2,\ldots,o_T\}$，其中$i_t \in Q$，$o_t \in V$
- 状态转移概率矩阵$A=[a_{ij}]_{N*N}$：转移概率是指某一个隐藏状态（标签）转移到下一个隐藏状态（标签）的概率，A记录所有状态转移的概率。
  - 依据齐次马科夫链假设，$a_{ij}=P(i_{t+1}=q_j|i_t=q_i)$
- 发射概率矩阵$B=[b_j(k)]_{N*M}$：是指在某个隐藏状态（标签，如“B-Per”）下，生成某个观测状态（词，如“陈”）的概率。
  - 依据观测独立性假设：$b_j(k)=P(o_t=v_k|i_t=q_i)$
- 隐藏状态的初始分布$\Pi=[\pi(i)]_N$，其中$\pi(i)=P(i_1=q_i)$，这里指的是标签的先验概率分布。



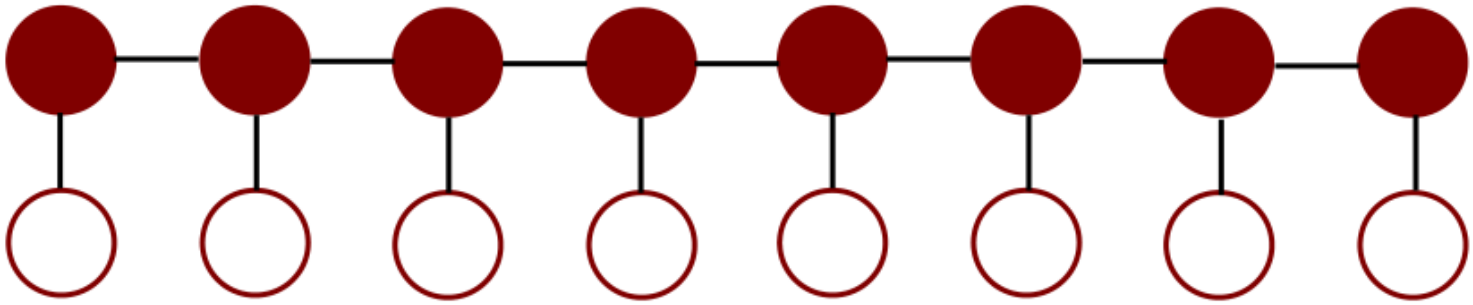
HMM 的计算问题

- 评估观察序列概率。即给定模型$\lambda=(A,B,\Pi)$ 和观测序列$O=\{o_1,o_2,\ldots,o_T\}$（如一句话“浙江大学位于杭州”），计算在模型λ下观测序列O出现的概率$P(O|\lambda)$。
- 模型参数学习问题。即给定观测序列$O=\{o_1,o_2,\ldots,o_T\}$，估计模型$\lambda=(A,B,\Pi)$的参数，使该模型下观测序列的条件概率$P(O|\lambda)$最大。这个问题的求解需要用到基于EM算法的鲍姆-韦尔奇算法。
- 预测问题，也称为解码问题。即给定模型$\lambda=(A,B,\Pi)$和观测序列$O=\{o_1,o_2,\ldots,o_T\}$，求给定观测序列条件下，最可能出现的对应的隐藏状态序列（标签序列），这个问题的求解需要用到基于动态规划的维特比算法。



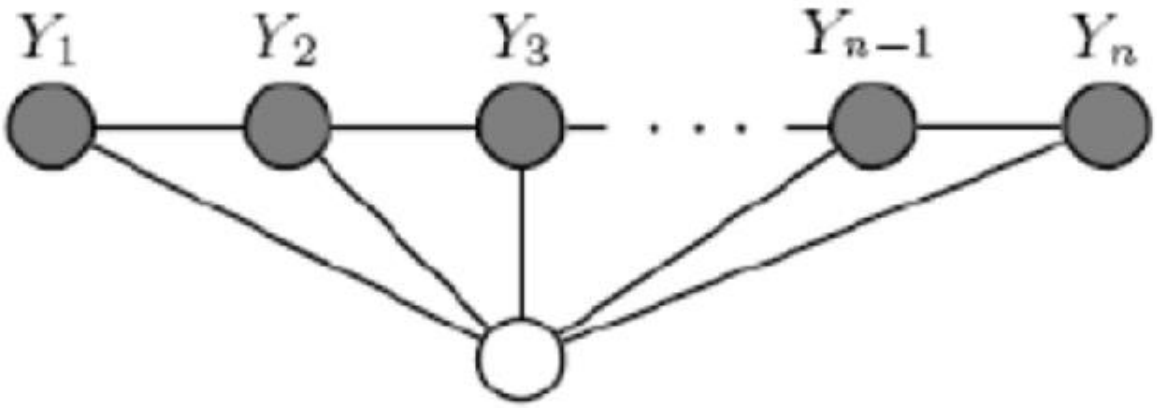
常见模型：CRF (条件随机场)

条件随机场无向图模型



- 随机场是由若干个位置组成的整体，当给每一个位置中按照某种分布随机赋予一个值之后，其全体就叫做随机场。
- 马尔科夫随机场是随机场的特例，它假设随机场中某一个位置的赋值仅与和它相邻的位置的赋值有关，和与其不相邻的位置的赋值无关。
- 条件随机场是马尔科夫随机场的特例，它假设马尔科夫随机场中只有X和Y两种变量，X一般是给定的，而Y一般是在给定X的条件下的输出。
- 例如：实体识别任务要求对一句话中的十个词做实体类型标记，这十个词可以从可能实体类型标签中选择，这就形成了一个随机场。如果假设某个词的标签只与其相邻的词的标签有关，则形成马科夫随机场，同时由于这个随机场只有两种变量，令X为词，Y为实体类型标签，则形成一个条件随机场。

线性链条件随机场——LinearCRF



设$X=(X1,X2,...Xn),Y=(Y1,Y2,...Yn)$均为线性链表示的随机变量序列，在给定随机变量序列$X$的情况下，随机变量$Y$的条件概率分布$P(Y|X)$构成条件随机场，即满足马尔科夫性：

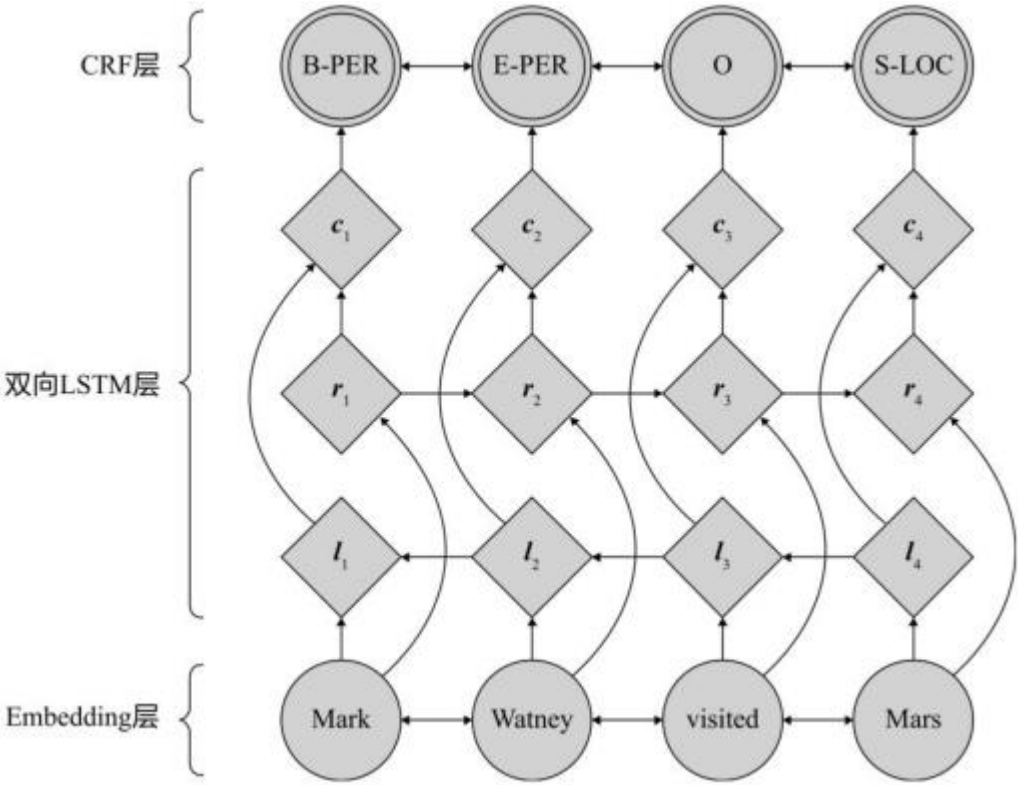
$$P(Yi|X,Y1,Y2,...Yn)=P(Yi|X,Yi-1,Yi+1)$$

则称$P(Y|X)$为线性链条件随机场。

基于深度学习的方法

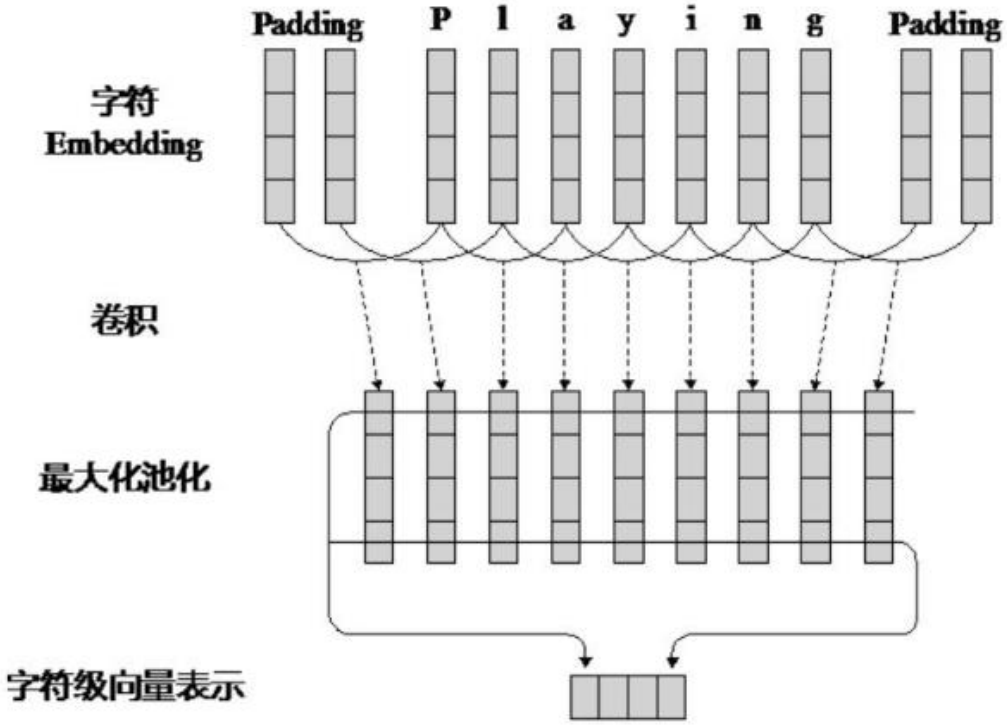
- 与传统统计模型相比，基于深度学习的方法直接以文本中词的向量为输入，通过神经网络实现端到端的命名实体识别，不再依赖人工定义的特征。
- 目前，用于命名实体识别的神经网络主要有卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)以及引入注意力机制(Attention Mechanism)的神经网络。
- 一般地，不同的神经网络结构在命名实体识别过程中**扮演编码器的角色**，它们基于初始输入以及词的上下文信息，得到每个词的新向量表示；最后再通过模型输出对每个词的标注结果。

常见模型：LSTM-CRF



- 使用长短时记忆神经网络与CRF相结合进行命名实体识别。
- 三层：**词嵌层、双向LSTM层和CRF层**。
  - 词嵌层是句子中词的向量表示，作为B-LSTM的输入，通过词向量学习模型获得。
  - 双向LSTM通过一个正向LSTM和一个反向LSTM，分别计算每个词考虑左侧或右侧时对应的向量，然后将每个词的两个向量进行连接，形成词的向量输出。
  - CRF层以双向LSTM的输出作为输入，对句子中的命名实体进行标注。

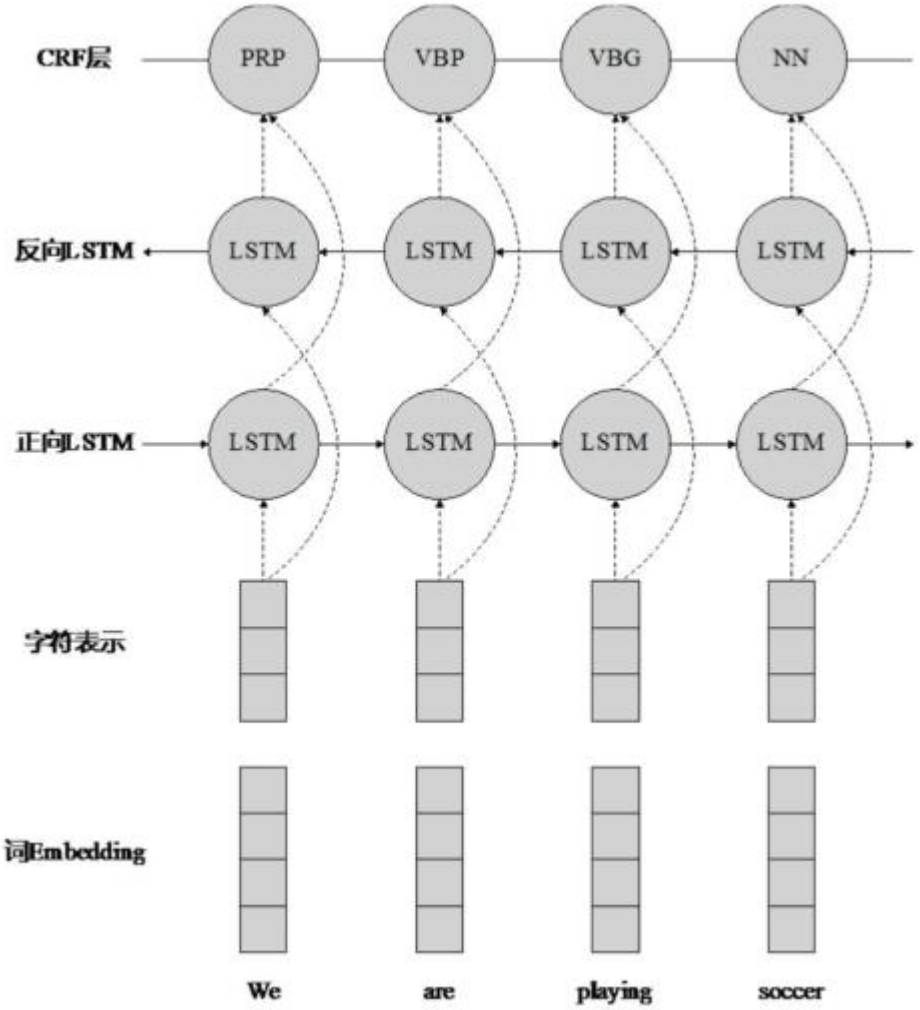
常见模型：LSTM-CNNs-CRF



获取词语字符级向量表示的CNN模型

- 将双向LSTM、CNN 和 CRF 相结合的序列标注模型
- 该模型与 LSTM-CRF 模型十分相似，不同之处**是在 Embedding 层中加入了每个词的字符级向量表示。**
- 利用CNN 模型获取词语字符级向量表示的，该模型可以有效地获取词的形态信息，如前缀、后缀等。模型Embedding层中每个词的向量输入**由预训练获得的词向量和 CNN 获得的字符级向量连接而成**，通过双向 LSTM 和 CRF 层获得词的标注结果。

常见模型：LSTM-CNNs-CRF



LSTM-CNNs-CRF序列标注模型框架

模型Embedding层中每个词的向量输入由预训练获得的词向量和 CNN 获得的字符级向量连接而成，通过双向 LSTM 和 CRF 层获得词的标注结果。

小结：中文实体识别

- 中文没有自然分词，词的概念很模糊，也不具备英文中的字母大小写等形态特征。
- 中文用字变化多，有些实体不能脱离上下文语境，同一实体在不同语境可能是不同的实体类型。
- 中文多嵌套实体，如“复旦大学附属第一医院”。
- 中文简化表达现象严重，如“上大”、“新时代”，“伟长学院”。

内容概览

一、知识获取与知识图谱获取

二、面向结构化的知识抽取

三、面向非结构化的知识抽取：实体抽取

四、面向非结构化的知识抽取：关系抽取

实体关系抽取

- 从文本中抽取出两个或者多个实体之间的语义关系
- 信息抽取 (Information Extraction)研究领域的任务之一
- 从文本获取知识图谱三元组的主要技术手段

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)

实体关系抽取

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。

实体关系抽取举例

中兴通讯与中国联通战略合作,开展部署全球首批5G规模商用-贤集网...

2019年6月27日 - 上海世界移动大会期间,中兴通讯与中国联通签署"5G智慧场馆"战略合作协议。双方宣布正式启动"5G智慧场馆"全方位合作,充分发挥各自优势组建联合运作团队,...

中兴中标中国联通数据设备集中采购 尽最大努力保护全球用户利益

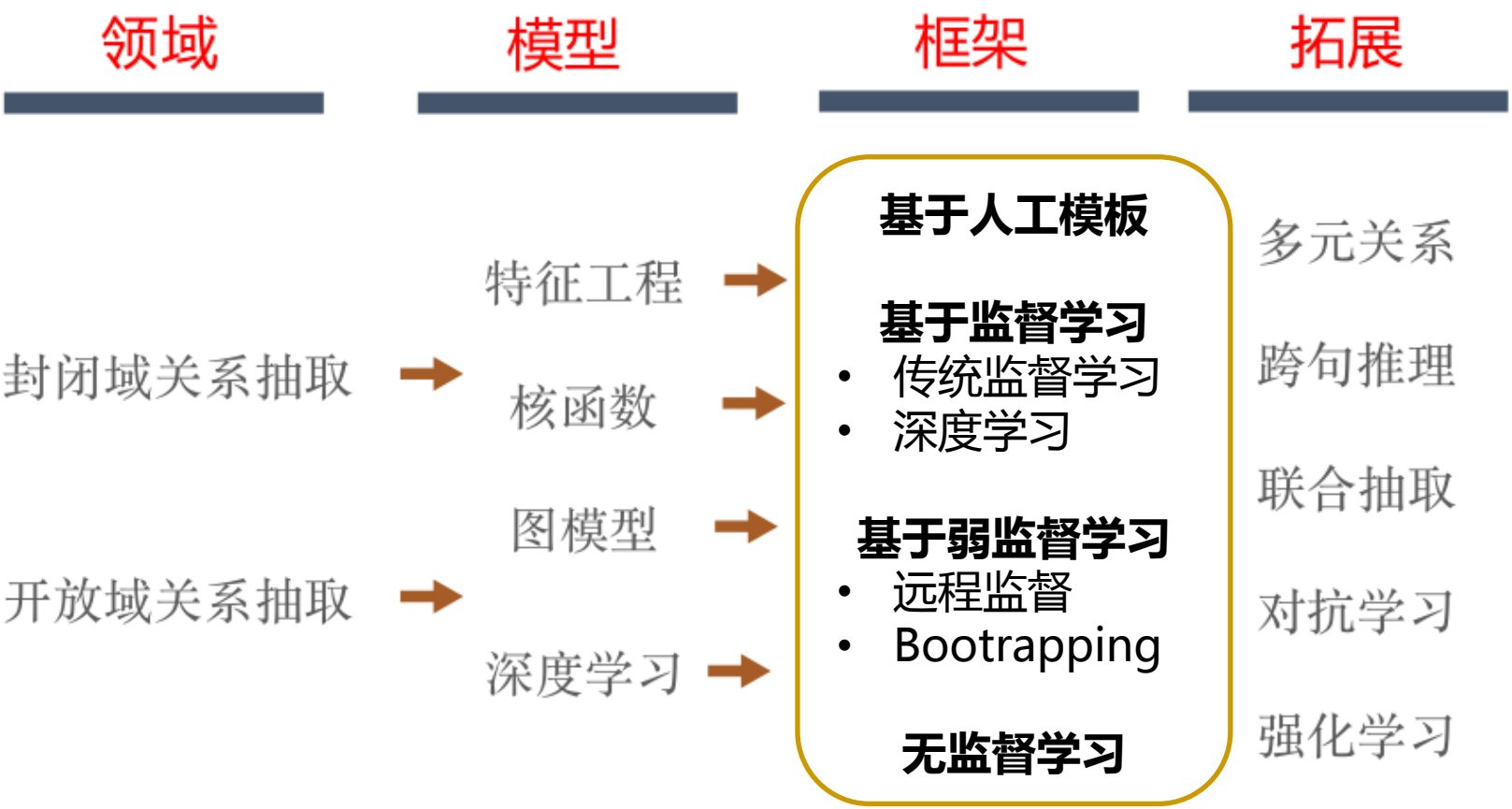
2018年4月24日 - 据来自中国联通的官方消息显示,2017-2018年中国联通数据设备集中采购已经开标,中标候选人为华为、中兴通讯、新华三和诺基亚贝尔四家设备商。 据C114了解,本次包括2017...

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	2019-6	公司新闻
中兴通讯	中国联通	客户	2018-4	公司新闻
中兴通讯	英特尔(INTC)	合作伙伴	2019-2	公司新闻

中兴通讯与英特尔联合发布了Light Cloud方案 - 移动..._电子发烧友

2019年2月26日 - 中兴通讯与英特尔联合发布了Light Cloud方案-Light Cloud是一种NFVI(网络功能虚拟化基础设施)解决方案,采用基于英特

实体关系抽取概览


领域
模型
框架
拓展
封闭域关系抽取
开放域关系抽取
特征工程
核函数
图模型
深度学习
基于人工模板
基于监督学习
传统监督学习
深度学习
基于弱监督学习
远程监督
Bootrapping
无监督学习
多元关系
跨句推理
联合抽取
对抗学习
强化学习

基于模板的方法—基于触发词的Pattern

[姚明]与妻子[叶莉]还有女儿姚沁蕾并排坐在景区的游览车上，画面十分温馨

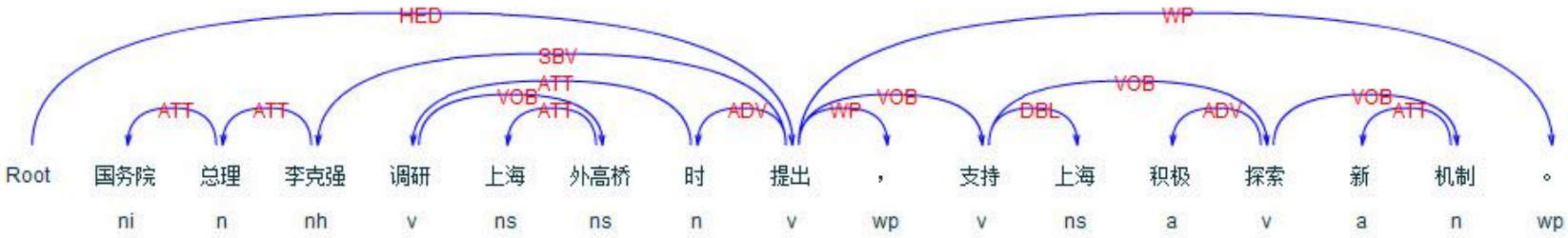
模板1:[X]与妻子[Y]

夫妻关系(X,Y)

[徐峥]老婆[陶虹]晒新写真

模板2:[X]老婆[Y]

基于模板的方法—基于依存句法分析的Pattern

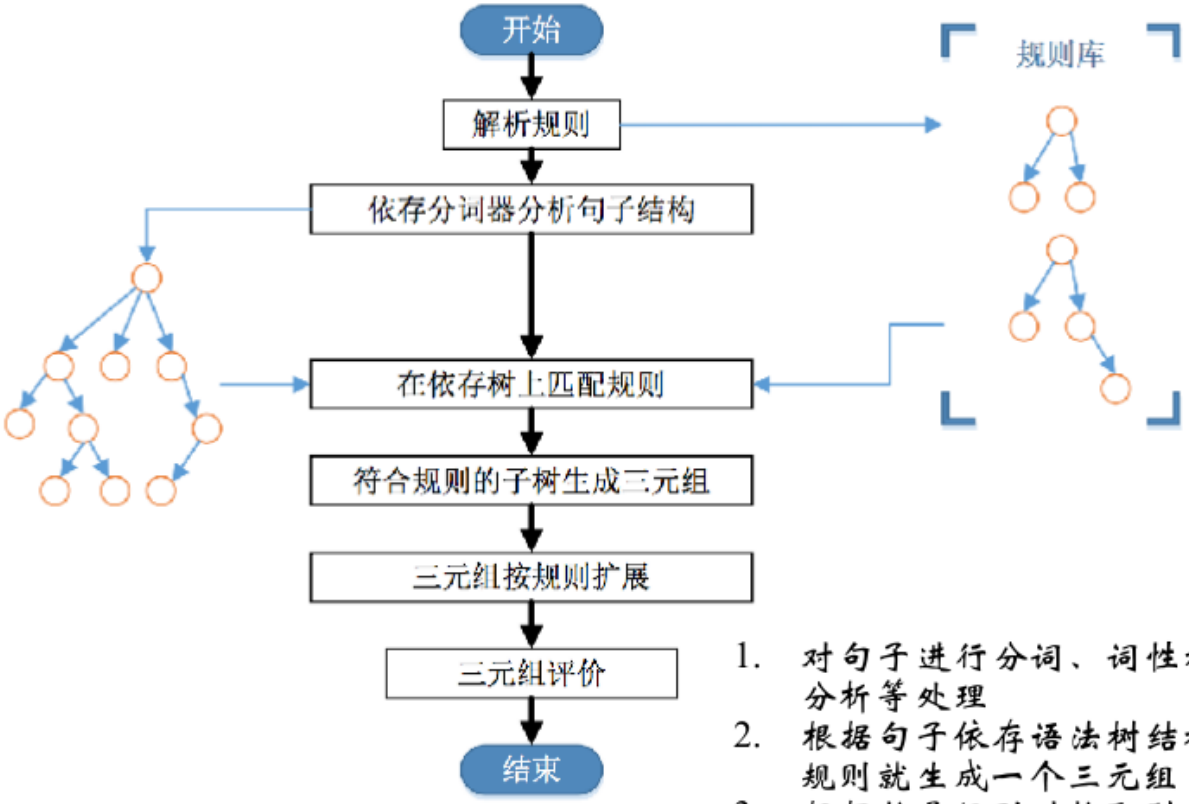


- 依存句法分析句子的句法结构
- 以动词为起点，构建规则，对节点上的词性和边上的依存关系进行限定

依存句法分析标注关系 (共14种)

关系类型	Tag	Description	Example
主谓关系	SBV	subject-verb	我送她一束花 (我 <-- 送)
动宾关系	VOB	直接宾语，verb-object	我送她一束花 (送 --> 花)
间宾关系	IOB	间接宾语，indirect-object	我送她一束花 (送 --> 她)
前置宾语	FOB	前置宾语，fronting-object	他什么书都读 (书 <-- 读)
兼语	DBL	double	他请我吃饭 (请 --> 我)
定中关系	ATT	attribute	红苹果 (红 <-- 苹果)
状中结构	ADV	adverbial	非常美丽 (非常 <-- 美丽)
动补结构	CMP	complement	做完了作业 (做 --> 完)
并列关系	COO	coordinate	大山和大海 (大山 --> 大海)
介宾关系	POB	preposition-object	在贸易区内 (在 --> 内)
左附加关系	LAD	left adjunct	大山和大海 (和 <-- 大海)
右附加关系	RAD	right adjunct	孩子们 (孩子 --> 们)
独立结构	IS	independent structure	两个单句在结构上彼此独立
核心关系	HED	head	指整个句子的核心

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

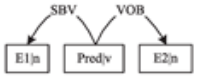
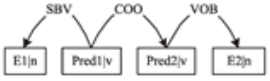
Number	Logical Expression and Graph Expression	Relation triples
DSNF1	 [E1-ATT-AttWord, AttWord-ATT-E2)]	(E1, AttWord{1,2}+, E2)
DSNF2	 [E1-SBV-Pred, E2-VOB-Pred]	(E1, Pred, E2)
DSNF3	 [E1-SBV-Pred, Dobj-VOB-Pred, E2-POB-Prep, Prep-ADV-Pred]	(E1, Pred-[Dobj]?+, E2)
DSNF4	 [E1-SBV-Pred, E2-POB-Prep, Prep-CMP-Pred]	(E1, Pred-Prep, E2)
DSNF5	 [Conj-LAD-E1, E1-COO-E2, E2-SBV-Pred, E3-VOB-Pred]	(E2, Pred, E3) (E1, Pred, E3)
DSNF6	 [E2-SBV-Pred, Conj-LAD-E1, E1-COO-E3, E3-VOB-Pred]	(E2, Pred, E3) (E2, Pred, E1)
DSNF7	 [E1-SBV-Pred1, E2-VOB-Pred2, Pred2-COO-Pred1]	(E1, Pred2, E2)

Fig. 5. Two kinds of definitions of DSNFs and the triples are available to extract from DSNFs. – denotes the combination of two words. {1, 2}+ indicates the word occurring once or twice. []?+ means the word occurring once or not.

"中国国家主席习近平访问韩国，并在首尔大学发表演讲“

- (中国，国家主席，习近平)
- (习近平，访问，韩国)
- (习近平，发表演讲，首尔大学)

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

基于监督学习的实体关系抽取

- 基于监督学习的关系抽取方法将关系抽取转化为分类问题，在大量标注数据的基础上，训练有监督学习模型进行关系抽取。
- 利用监督学习方法进行关系抽取的一般步骤包括：
  - 1、预定义关系的类型；
  - 2、人工标注数据；
  - 3、设计关系识别所需的特征，一般根据实体所在句子的上下文计算获得；
  - 4、选择分类模型(如支持向量机、神经网络和朴素贝叶斯等)，基于标注数据训练模型；
  - 5、对训练的模型进行评估。

监督学习-特征

关系抽取**特征的定义**对于抽取的结果具有较大的影响，因此大量的研究工作围绕关系抽取特征的设计展开。根据计算特征的复杂性，可以将常用的特征分为轻量级、中等量级和重量级三大类。

- **轻量级：**
  - **基于实体和词的特征，例如句子中实体前后的词、实体的类型以及实体之间的距离等。**
- **中等量级：**
  - **基于句子中语块序列的特征。**
- **重量级：**
  - **实体间的依存关系路径、实体间依存树结构的距离以及其他特定的结构信息。**

例如，对于句子“Forward [motion] of the vehicle through the air caused a [suction] on the road draft tube”

轻量级的特征可以是：实体[motion]和[suction]、实体间的词{of,the,vehicle,through,the,air,caused,a}等；

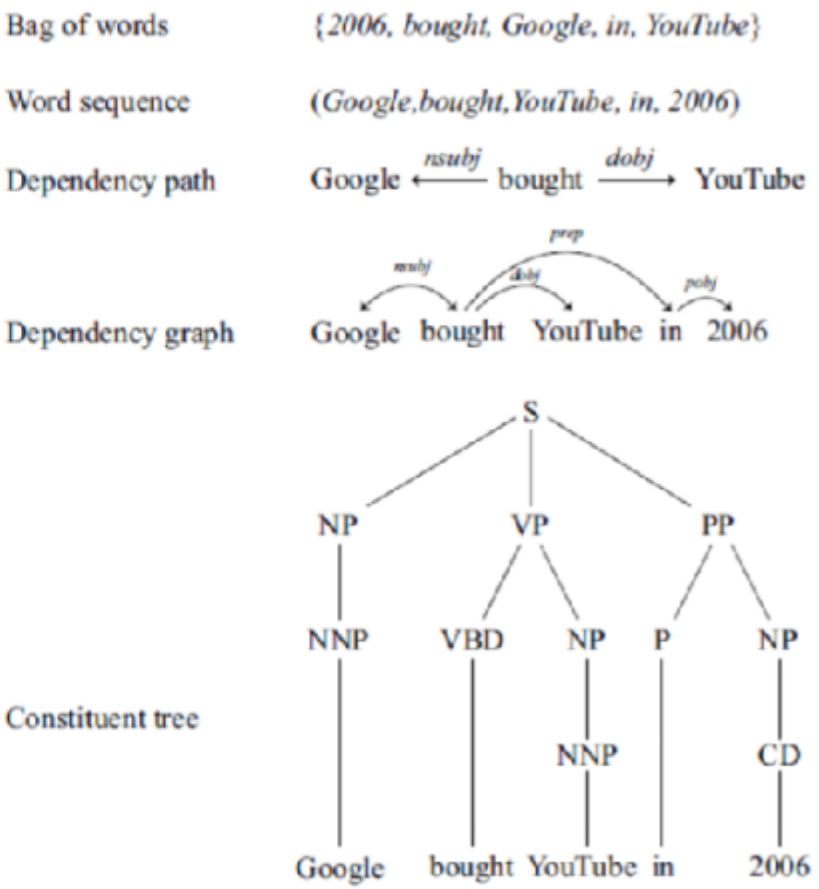
重量级的特征可以包括依存树中的路径“caused→nsubj→实体1”“caused→dobj→实体2”等。

实体特征

- 基本实体特征
  - String values
  - Examples:
    - string value
    - words/stems/lemmata
  - PROs: 特征信号强
  - CONs: 过于稀疏
- 实体的上下文特征
  - 语法信息——Syntactic information, e.g., grammatical role
  - 语义信息——Semantic information, e.g., semantic class
  - 词的共现特征, e.g., dog and cat
  - 分布式表示
  - 词向量
  - 引入外部语义关系, e.g.,
    - ACE entity types
    - WordNet features

关系特征

- 基本关系特征
  - 实体之间的词：words between the two arguments
  - 窗口：words from a fixed window on either side of the arguments
  - 依存关系：a dependency path linking the arguments
  - 整个依存关系图：an entire dependency graph
  - 最小子树：the smallest dominant subtree

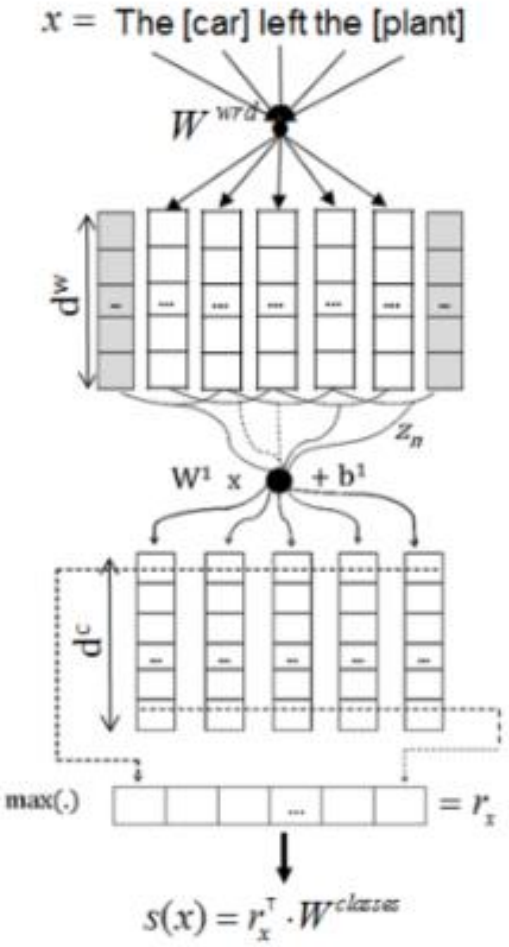


基于深度学习的关系抽取方法

- 传统的基于监督学习的关系抽取是一种依赖特征工程的方法，近年来有多个基于深度学习的关系抽取模型被研究者们提出。深度学习的方法不需要人工构建各种特征，其输入一般只包括句子中的词及其位置的向量表示。
- 目前，已有的基于深度学习的关系抽取方法主要包括**流水线方法**和**联合抽取方法**两大类。
  - 流水线方法将识别实体和关系抽取作为两个分离的过程进行处理，两者不会相互影响；关系抽取在实体抽取结果的基础上进行，因此关系抽取的结果也依赖于实体抽取的结果。
  - 联合抽取方法将实体抽取和关系抽取相结合，在统一的模型中共同优化；联合抽取方法可以避免流水线方法存在的错误积累问题。

基于深度学习的流水线关系抽取方法

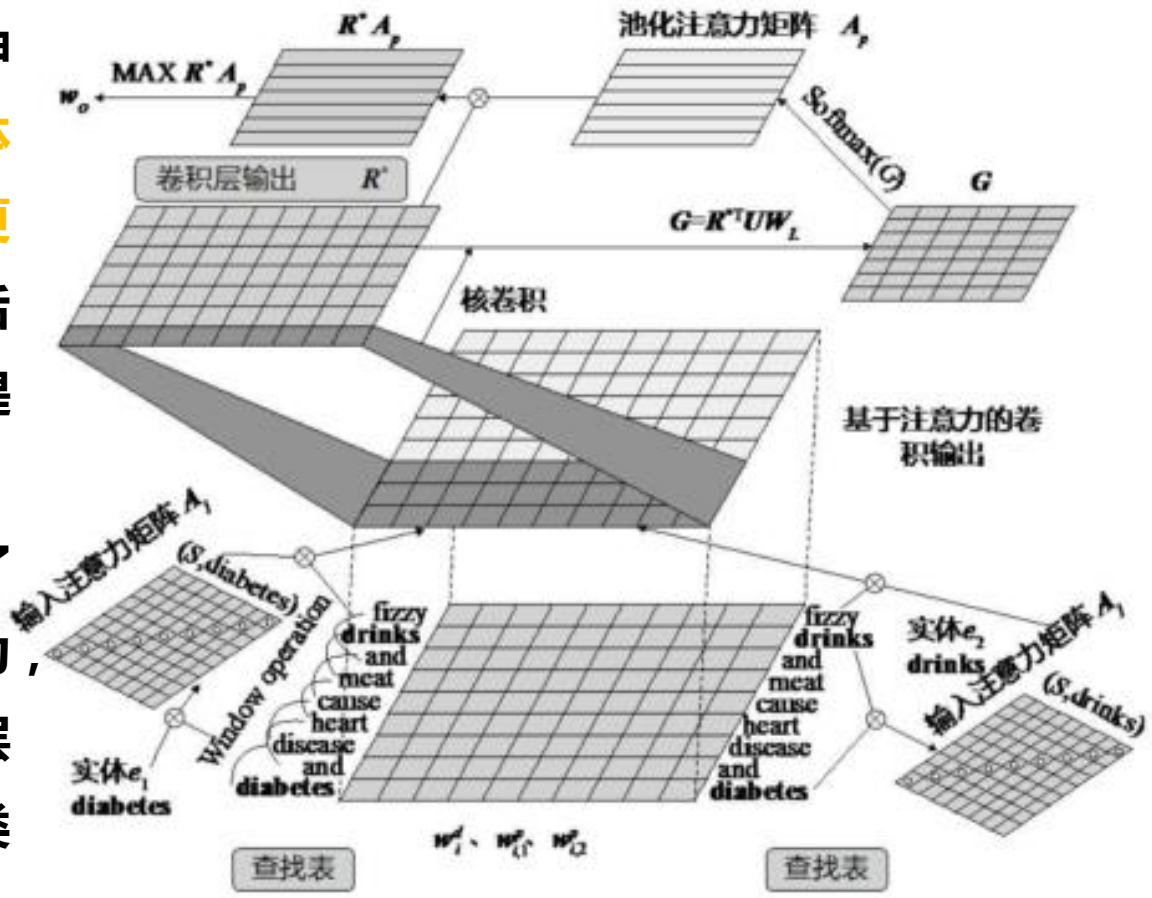
- CR-CNN模型
  - ✓ 给定输入的句子，CR-CNN模型首先将句子中的词映射到长度为d_w的低维向量，每个词的向量包含了词向量和位置向量两部分。
  - ✓ 然后，模型对固定大小滑动窗口中的词的向量进行卷积操作，为每个窗口生成新的长度为d_c的特征向量；对所有的窗口特征向量求最大值，模型最终得到整个句子的向量表示d_x.
  - ✓ 在进行关系分类时，CR-CNN模型计算句子向量和每个关系类型向量的点积，得到实体具有每种预定关系的分值。



F1值达到84.1%

基于深度学习的流水线关系抽取方法

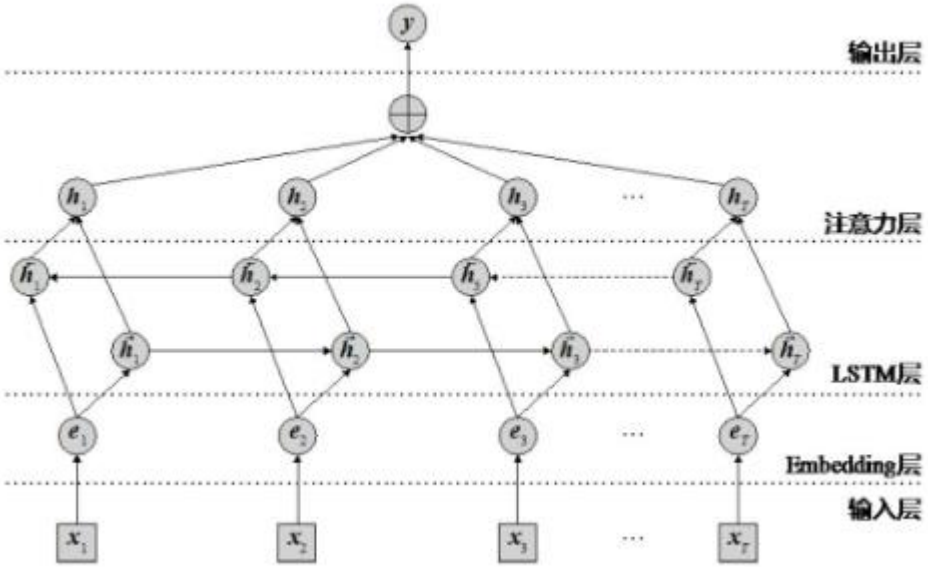
- Attention CNNs模型
  - ✓ 将注意力机制引入到神经网络中，对反映实体关系更重要的词赋予更大的权重，借助改进后的目标函数提高关系提取的效果。
  - ✓ 在输入层，模型引入了词与实体相关的注意力，同时还在池化和混合层引入了针对目标关系类别的注意力。



F1值达到88%

基于深度学习的流水线关系抽取方法

- Attention BLSTM模型
  - ✓ 包含两个LSTM网络，从正向和反向处理输入的句子，从而得到每个词考虑左边和右边序列背景的状态向量；词的两个状态向量通过元素级求和产生词的向量表示。
  - ✓ 在双向LSTM产生的词向量基础上，该模型通过注意力层组合词的向量产生句子向量，进而基于句子向量将关系分类。
  - ✓ 注意力层首先计算每个状态向量的权重，然后计算所有状态向量的加权和得到句子的向量表示。



基于深度学习的流水线关系抽取方法

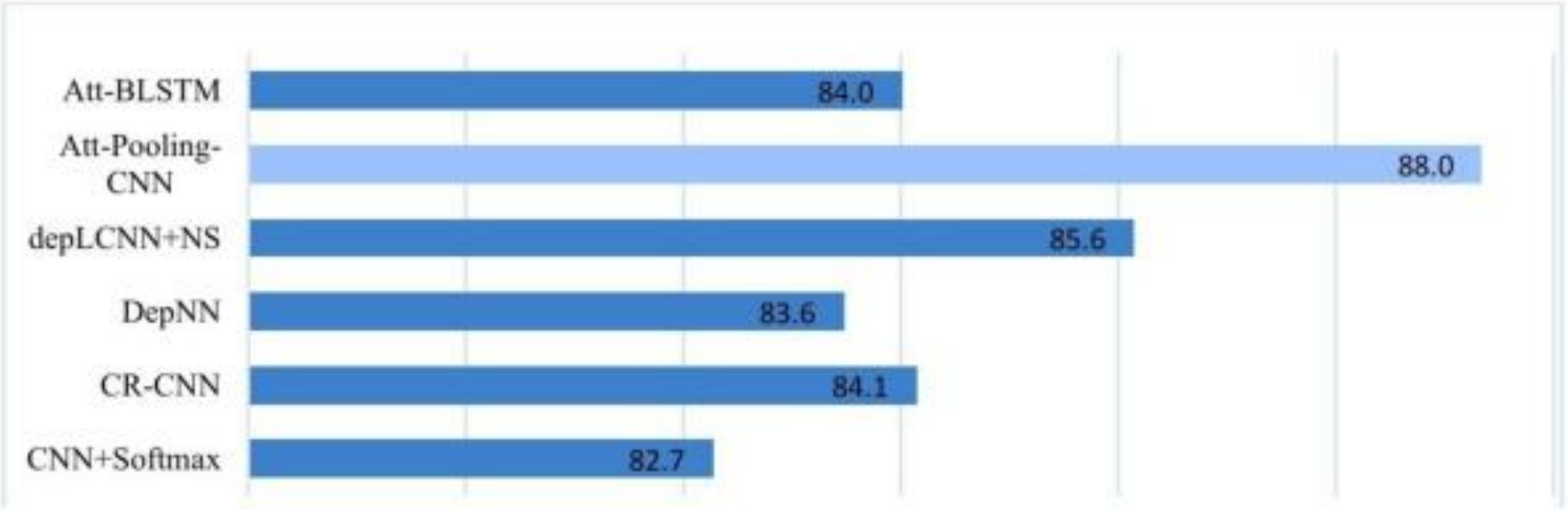
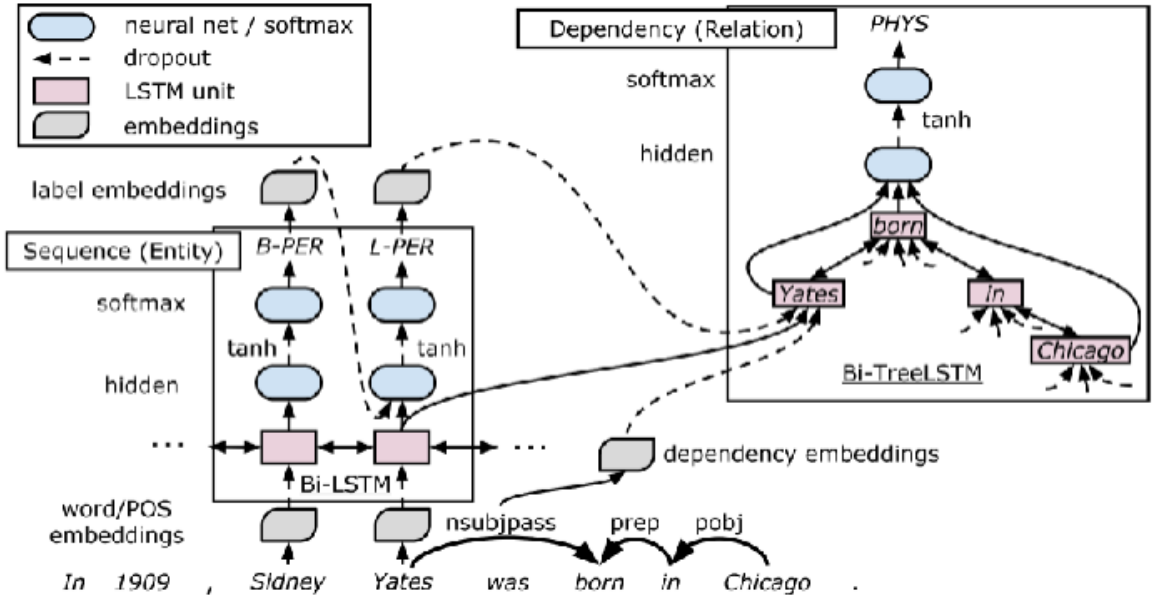


图4-12 关系抽取模型在SemEval-2010 Task 8数据集F1值对比（%）

基于深度学习的联合关系抽取

将实体抽取和关系抽取相结合。模型包含三层：词嵌入层、基于单词序列的LSTM-RNN层(序列层)以及基于依赖性子树的LSTM-RNN层(依存关系层)，在解码过程中，模型在序列层上构建从左到右的实体识别，并实现依存关系层上的关系分类，其中每个基于子树的LSTM-RNN对应于两个被识别实体之间的候选关系。



总结

- 知识抽取是实现自动化构建大规模知识图谱的重要技术。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。文本一般不作为知识图谱构建的初始来源，而多用来做知识图谱补全。
- 知识抽取分为面向结构化数据的知识抽取和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包括直接映射和R2RML映射语言等方法。
- 面向非结构化数据的知识抽取分为实体抽取、关系抽取、事件抽取和规则抽取等。
- 实体抽取常用的方法包括基于模板和规则的方法、基于统计模型的方法、基于深度学习的方法。
- 关系抽取常用的方法包括基于人工模板基于监督学习(传统监督学习、深度学习)、基于弱监督学习(迁移学习、Bootstrapping)和无监督学习方法。