《知识图谱: 概念与技术》

第 3 讲
关系抽取

肖仰华

复旦大学

shawyh@fudan.edu.cn

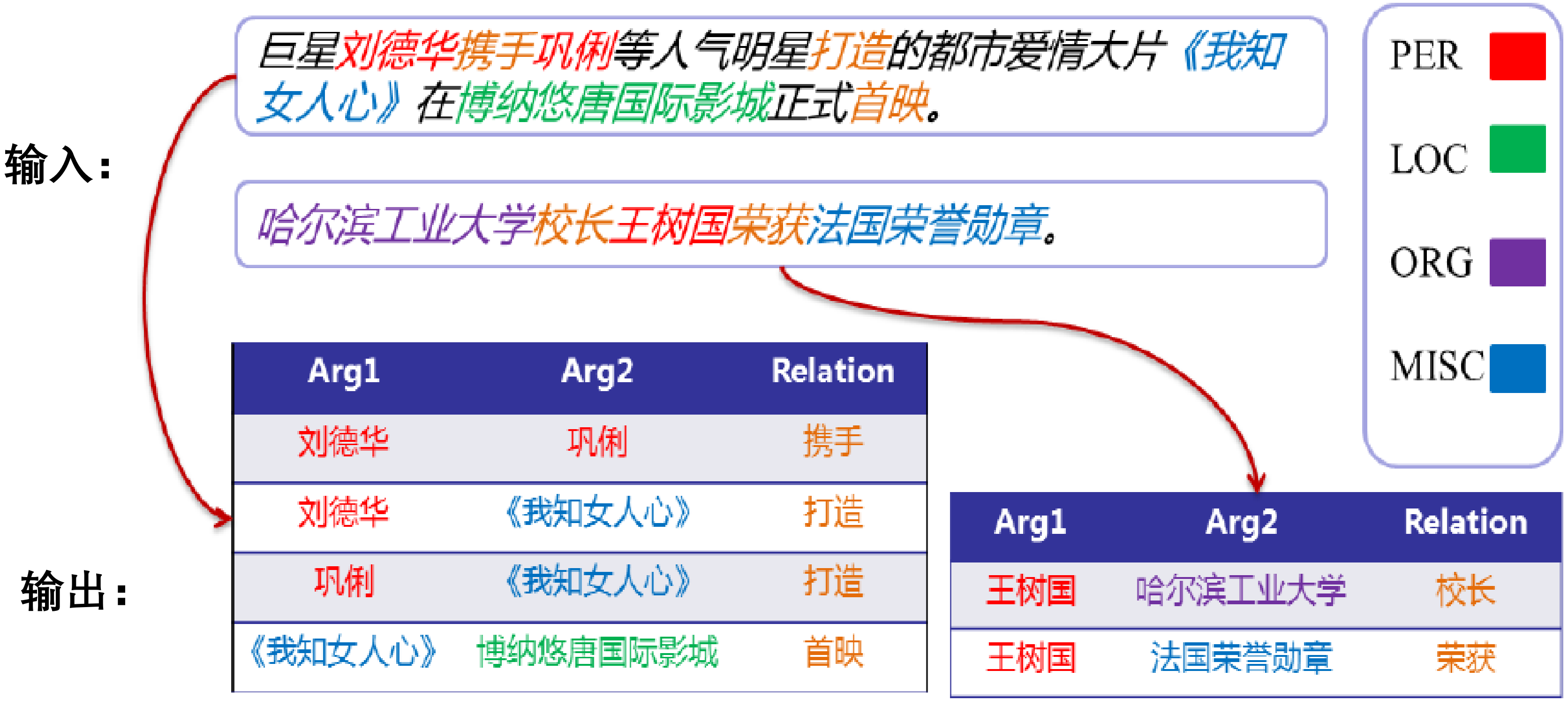
本章大纲

- 1、关系抽取概述
- 2、基于Rule的关系抽取
- 3、基于有监督的关系抽取
- 4、基于bootstrapping的关系抽取
- 5、基于远程监督的关系抽取
- 6、开放关系抽取
- 7、参考文献

1、关系抽取概述

1.1 关系抽取（relation extraction）定义

- 从文本抽取实体对之间的语义关系
  - 实体对可以预先给定或者基于NRE获得



输入：

巨星刘德华携手巩俐等人气明星打造的都市爱情大片《我知女人心》在博纳悠唐国际影城正式首映。

哈尔滨工业大学校长王树国荣获法国荣誉勋章。

PER	🟥
LOC	🟩
ORG	🟪
MISC	🟦

输出：

Arg1	Arg2	Relation
刘德华	巩俐	携手
刘德华	《我知女人心》	打造
巩俐	《我知女人心》	打造
《我知女人心》	博纳悠唐国际影城	首映

Arg1	Arg2	Relation
王树国	哈尔滨工业大学	校长
王树国	法国荣誉勋章	荣获

1.2 关系抽取的意义

- 关系抽取是信息抽取（information extraction）重要子任务
- 在知识库构建与补全的关键步骤之一

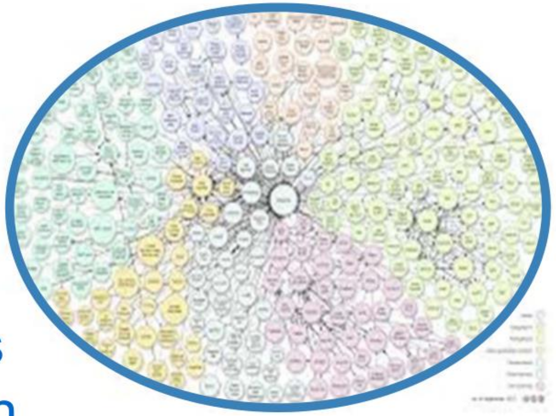
基于关系抽取的自动化知识
库构建获得了广泛的研究!

Freebase

yago
select knowledge

DBpedia

NELL: Never-Ending
Language Learning



OpenIE
(Reverb, OLLIE)

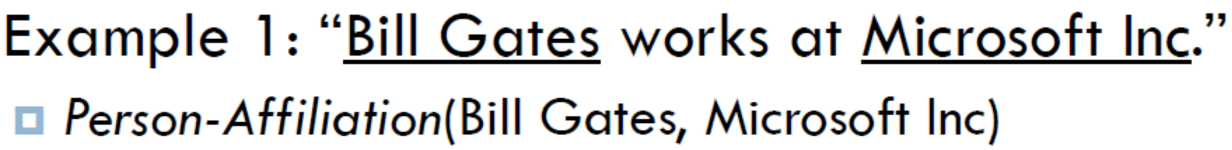
Facebook's
Entity Graph

Microsoft's
Satori

Google's
Knowledge Graph

1.3 关系抽取的任务分类

- 根据关系集合是否预选给定，将关系抽取分为两类：
  - **关系分类**
    - 将关系抽取转化为对候选实体对的分类问题
  - **开放关系抽取（OpenIE）**
    - 直接从文本中抽取出结构化文本关系（textual relation）
    - 规范化：对文本关系映射到知识库的规范关系



Example 1: “Bill Gates works at Microsoft Inc.”

- *Person-Affiliation*(Bill Gates, Microsoft Inc)

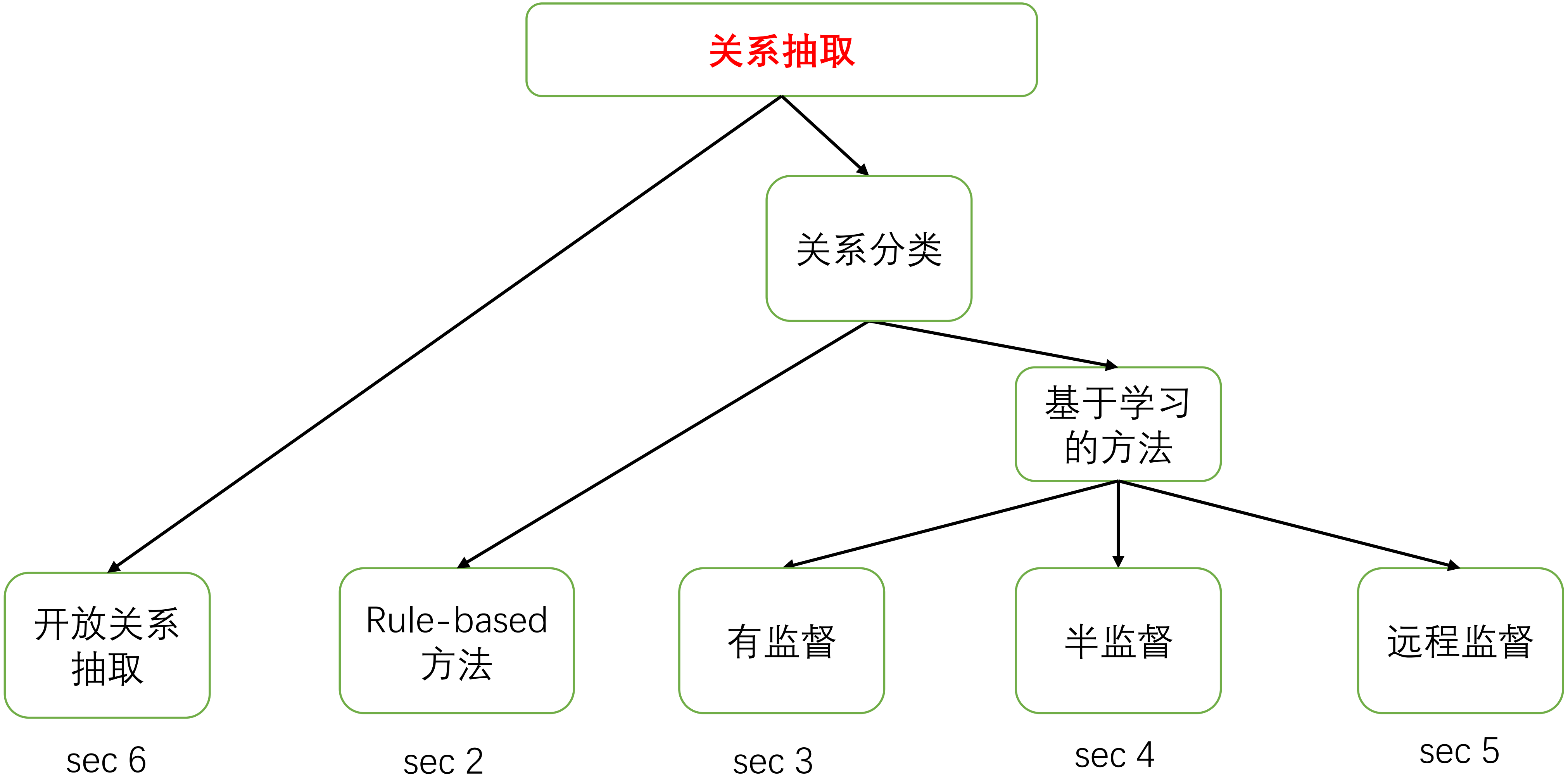
关系分类。其中，Person-Affiliation是预先给定的关系。

- **Hudson was born in Hampstead, which is a suburb of London.**

- (Hudson, *was born in*, Hampstead)
- (Hampstead, *is a suburb of*, London)

开放关系抽取。其中，”took”为抽取的文本关系。

关系抽取的方法分类


关系抽取
关系分类
基于学习的方法
开放关系抽取
Rule-based方法
有监督
半监督
远程监督
sec 6
sec 2
sec 3
sec 4
sec 5

1.4 关系抽取的挑战

- 实体对的关系通常在文本中隐式表达
  - *<特朗普，任职总统，美国>*
  - *特朗普执掌着美国的行政大权*

- 实体对的关系在文本中存在多样化的表达
  - *<特朗普，任职总统，美国>*
  - *特朗普是美国白宫工作*
  - *特朗普“新政”确实有效提振了美国经济*
  - *自从特朗普上任后，美国公民反应激烈*
- 对于学习模型，高质量的训练样本极少
  - 人工标注成本高

1.6 常用数据集

- 人工构造数据集
  - ACE 2005数据集：包含与新闻和电子邮件相关的599个文档，并包含7个主要类型的关系
  - SemEval-2010 Task 8：包含10,717个样本，包含9种有序关系类型

- 基于远程监督构造的数据集
  - NYT数据集：对齐Freebase和纽约时报，包含53种具体关系和1种NA关系；
  - KBP数据集：对齐Wikipedia infoboxes和KBP共享任务语料和Wikipedia语料；

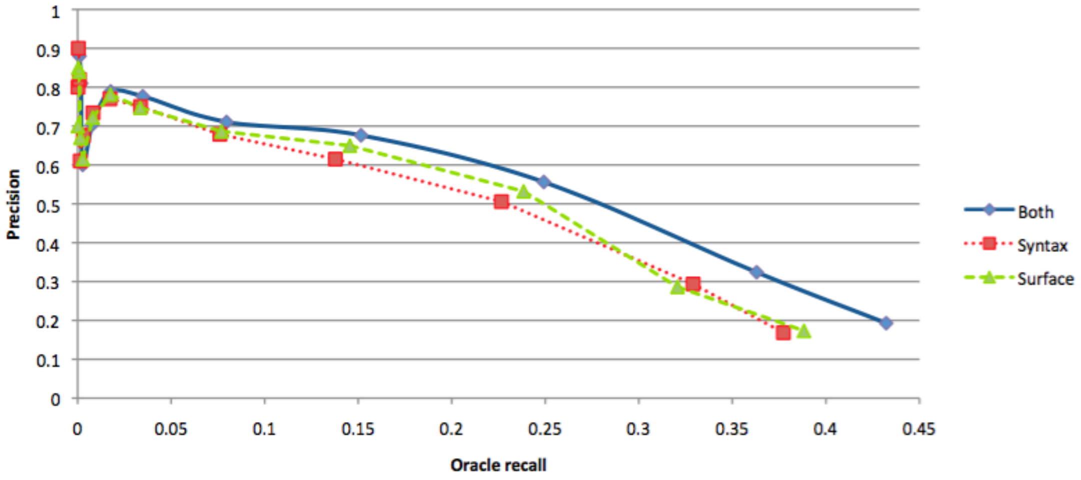
1.7 评估方法

- 自动评估（held-out evaluation）
  - 比较模型预测的结果和测试集中的标准值来判断对错
- 人工评估（human evaluation）
  - 通过多数投票的方法对预测的关系进行评估
- 度量标准
  - 精确率（precision），准确率（accuracy）和召回率（recall）和F1值
  - precision- recall曲线

度量指标

- 评估对象
  - 模型对测试实体对的预测关系集合
- 假定
  - 测试集中的关系实例数量: *N*
  - 模型预测的关系实例数量: *E*
  - 模型预测的正确的关系实例数量: *C*
- 度量指标
  - Recall = *C*/*N*
  - Precision = *C*/*E*
  - F-Measure = Harmonic mean of recall and precision

$$F = \frac{2 \bullet \text{Precision} \bullet \text{Recall}}{\text{Precision} + \text{Recall}}$$



Precision-recall曲线 [Mintz et al 2009]

2、基于Rule的关系抽取

2.1 概述

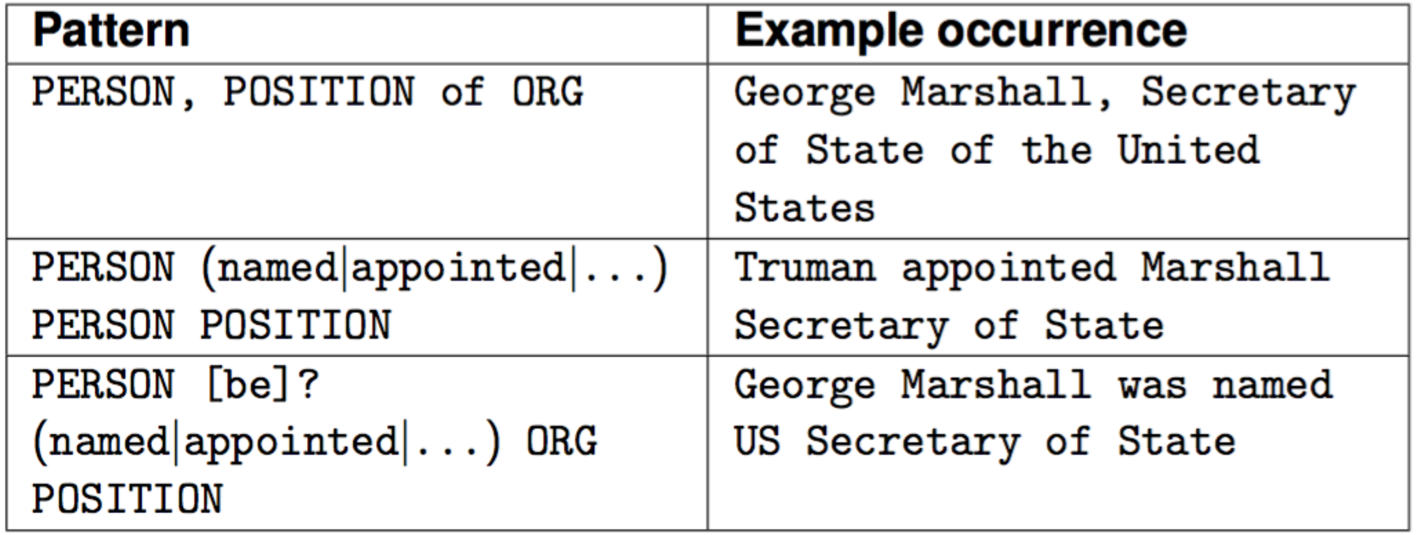
- 通过手工编写规则匹配文本，实现关系抽取
  - 手动编写词汇句法模式
  - 编写规则以识别文本中的模式

  - 例子：founder-of（jobs，apple）
  - Text：**Jobs** is the new CEO of **Apple** in 1976
  - rule： is the new CEO of
  - New text: **Mayer** is the new CEO of **Yahoo**!
  - New entity pair: (Mayer ,Yahoo)

[Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora, Fourteenth International Conference on Computational Linguistics, Nantes, France, 1992.]

命名实体标签

- 动机：关系往往在特定类型的实体对之间成立
  - *Located_in* (ORGANIZATION, LOCATION)
  - *Founded* (PERSON, ORGANIZATION)
  - *Cures* (DRUG, DISEASE)
  - *Serves_as* (PERSON, POSITION)
- 命名实体标签帮助关系分类：



Pattern	**Example occurrence**
PERSON, POSITION of ORG	George Marshall, Secretary of State of the United States
PERSON (named|appointed|...) PERSON POSITION	Truman appointed Marshall Secretary of State
PERSON [be]? (named|appointed|...) ORG POSITION	George Marshall was named US Secretary of State

人工规则的优缺点

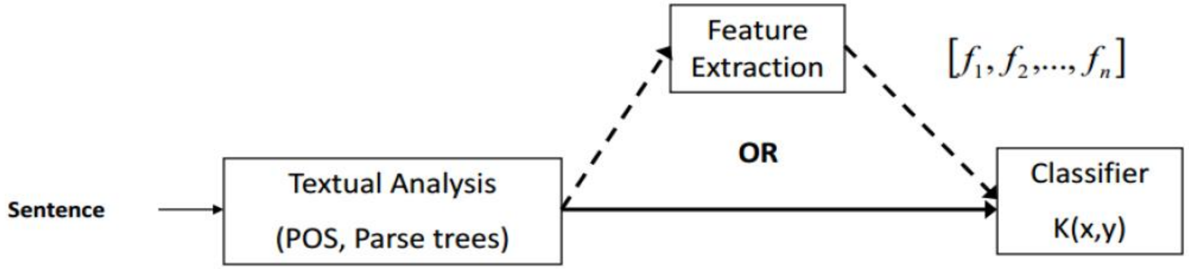
- 优点
  - 人工规则往往是高精度的
  - 可以针对特定领域进行定制
- 缺点
  - 人工规则往往导致低召回率
  - 人工成本高、代价大

3、基于有监督的关系抽取

3.1 概述

- 主流方法：将关系实例转换成高维空间中的特征向量或直接用离散结构来表示，在标注语料库上使用学习器来生成分类模型，然后再抽取语义关系。
  - 基于特征向量方法：最大熵模型(Kambhatla 2004)和支持向量机(Zhao等2005；Zhou等2005; Jiang等2007)等;
  - 基于核函数的方法：浅层树核（Zelenko 等 2003)、依存树核（Culotta 等 2004)、最短依存树核（Bunescu等 2005)、卷积树核（Zhang等 2006; Zhou 等 2007）。

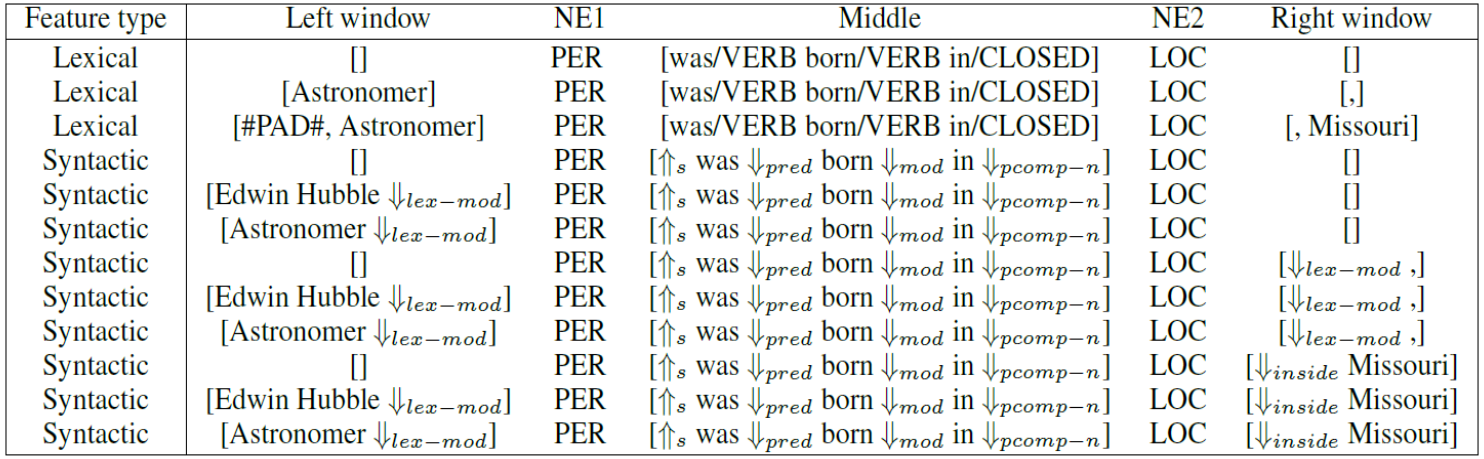
- 流程图



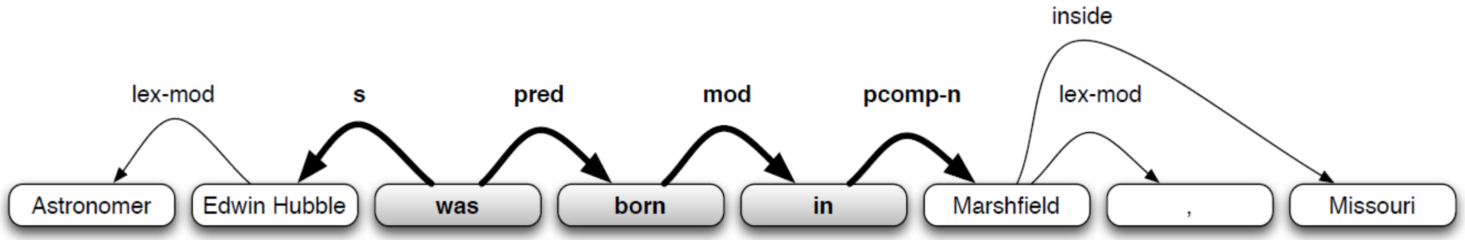
- 传统有监督的关系抽取的核心问题是有效特征抽取，包括实体对上下文中的各种词法、句法、语义等信息，或者背景知识等等；

3.2 常用特征

- 特征选取：从自由文本及其语法结构中抽取出各种表面特征以及结构化特征的平面形式。
    - 实体词汇及其上下文特征
    - 实体类型及其组合特征
    - 基本短语块特征
    - 依存树特征
    - 句法树特征

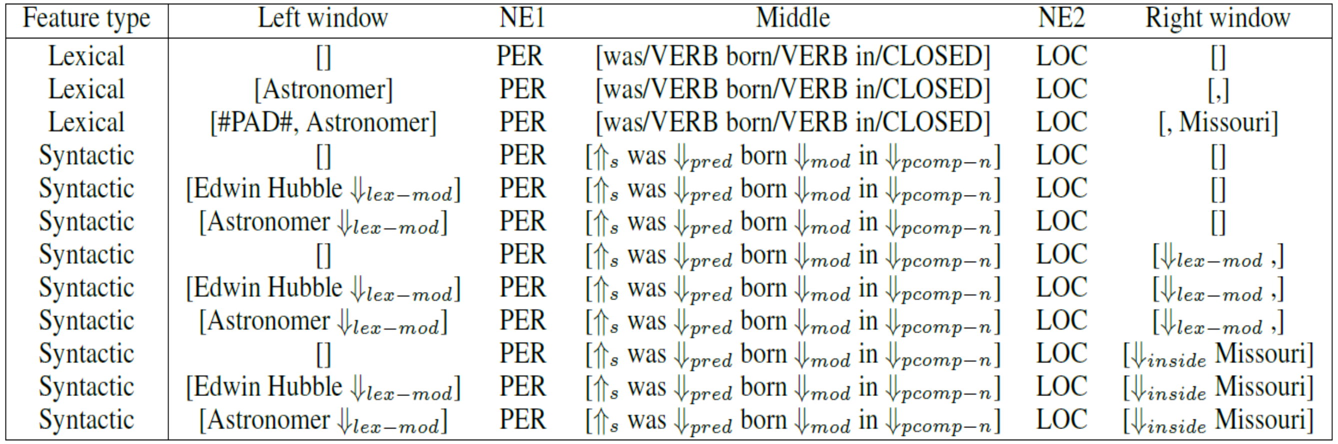


Feature type	Left window	NE1	Middle	NE2	Right window
Lexical	[]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[]
Lexical	[Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[,]
Lexical	[#PAD#, Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[, Missouri]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]



词汇特征

- 词汇特征
  - 主要指实体对之间或周围的特定的词汇
    - （1）两个实体之间的词袋信息；
    - （2）词袋的词性标注结果信息；
    - （3）实体对在句子中的顺序标志信息；
    - （4）左实体的窗口大小为k的词袋及其词性标注信息；
    - （5）右实体的窗口大小为k的词袋及其词性标注信息；
  - 例子

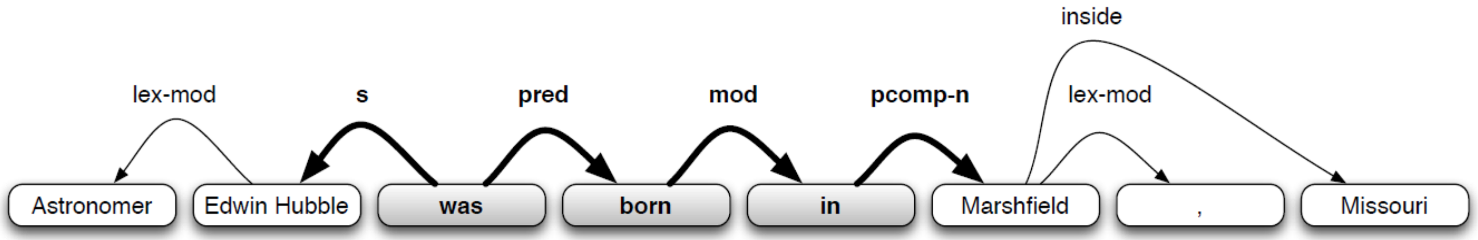


Feature type	Left window	NE1	Middle	NE2	Right window
Lexical	[]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[]
Lexical	[Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[,]
Lexical	[#PAD#, Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[, Missouri]
Syntactic	[]	PER	[$\Uparrow_{s}$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_{s}$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_{s}$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[]	PER	[$\Uparrow_{s}$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_{s}$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_{s}$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[]	PER	[$\Uparrow_{s}$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_{s}$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_{s}$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]

句子“Astronomer Edwin Hubble was born in Marshfield, Missouri”的词汇和句法特征组合[Mintz et al. (2009]

句法特征和其他特征

- 句法特征
  - 通过依存分析器获得句子的句法解析结果
  - 例子（参考上一页例子）



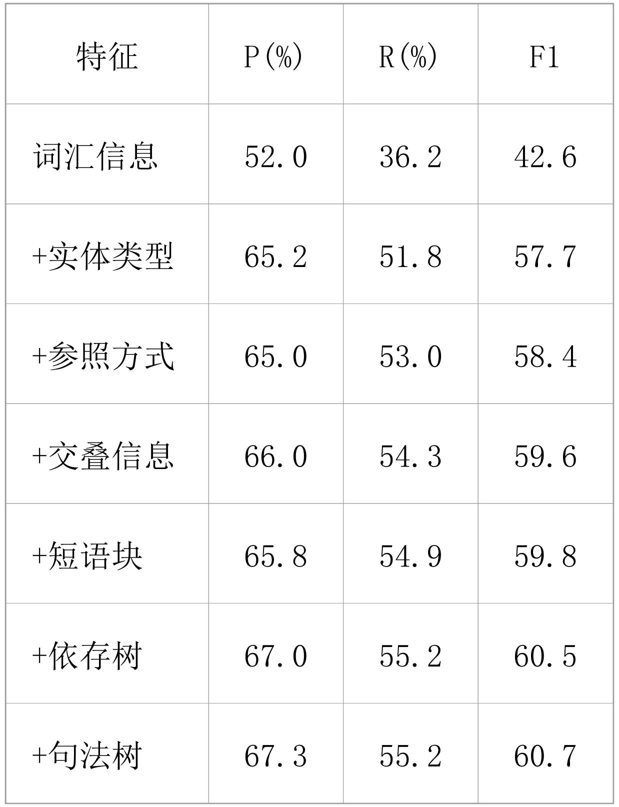
图：句子“Astronomer Edwin Hubble was born in Marshfield, Missouri”依存分析结果。
粗体部分表示实体对《Edwin Hubble，Marshfield》之间的直接依赖路径。[Mintz et al. (2009]

- 其他特征
  - 实体类型、概念、背景知识（如wordnet），位置信息等等;

3.2 基于特征向量的关系抽取

- 特征选取：从自由文本及其语法结构中抽取出各种表面特征以及结构化特征的平面形式。
  - 实体词汇及其上下文特征
  - 实体类型及其组合特征
  - 实体参照方式
  - 交叠特征
  - 基本短语块特征
  - 依存树特征
  - 句法树特征

特征的有效性



特征	P(%)	R(%)	F1
词汇信息	52.0	36.2	42.6
+实体类型	65.2	51.8	57.7
+参照方式	65.0	53.0	58.4
+交叠信息	66.0	54.3	59.6
+短语块	65.8	54.9	59.8
+依存树	67.0	55.2	60.5
+句法树	67.3	55.2	60.7

ACE RDC 2004关系大类

- 数据分析
  - 词汇信息、实体类型信息等特征在ACE 2004上的语义关系抽取中比较有效；
  - 实体参照方式、交叠信息等特征有一定作用；
  - 其它结构化特征仅能略微提高关系抽取的性能。
- 实验结论
  - 基于特征向量的方法可以使用一些成本较低的特征达到一定的性能；
  - 结构化信息在基于特征的方法中不能很好被利用，并非是它们本身没有作用。
  - 因此结构化信息的探索和利用成为关系抽取的研究重点。

3.3 基于树核函数的关系抽取

- 卷积核函数：用两个结构之间的公共子结构的数目来衡量它们之间的相似度。
  - 句法树核（Collins and Duffy et al. 2001）
  - 字符串核（Lodhi et al.2002）
  - 图形核（Suzuki et al. 2003）
- 卷积树核函数
  - 优点：能有效捕获离散数据对象中的结构化信息，在自然语言处理领域中取得了广泛的应用，如语义角色标注、关系抽取和指代消解等。
  - 缺点：计算效率较低。

3.3 基于树核函数的关系抽取（续）

- 卷积数核函数

计算两棵树T_1和T_2之间的相似度为两者之间的公共子树的目。

$$K_{CTK}(T_1,T_2)=\sum_{n_1\in N_1,n_2\in N_2}\Delta(n_1,n_2)$$

其中N1和N2分别为T1和T2的结点集合，$\Delta(n_1,n_2)$ 用来计算以n1和n2为根结点的两棵子树之间的相似度，它可以通过下列递归的方法得出：

- 1) 如果和的产生式（采用上下文无关文法）不同，则 $\Delta(n_1,n_2)=0$； 否则转2;
- 2) 如果和是词性（POS）标记，则 $\Delta(n_1,n_2)=1\times\lambda$ ； 否则转3;
- 3) 递归计算：

$$\Delta(n_1,n_2)=\lambda\prod_{k=1}^{\#ch(n_1)}(1+\Delta(ch(n_1,k),ch(n_2,k))$$

其中 (0< λ <1) 则是衰减因子，λ用来防止子树的相似度过度依赖于子树的大小。

3.3 基于树核函数的关系抽取（续）

- 标准卷积树核函数（CTK）
  - 在计算两棵子树的相似度时，只考虑子树本身，不考虑子树的上下文信息。
- 上下文相关卷积树核函数（CS-CTK）
  - 在计算子树相似度量，同时考虑子树的祖先信息，如子树根结点的父结点、祖父结点信息，并对不同祖先的子树相似度加权平均。

$$K_{CSCTK}(T_1,T_2)=\sum_{\substack{n_1\in N_1\\ n_2\in N_2}}\sum_{i=1}^{m}w_i\cdot\Delta^i(n_1,n_2)$$

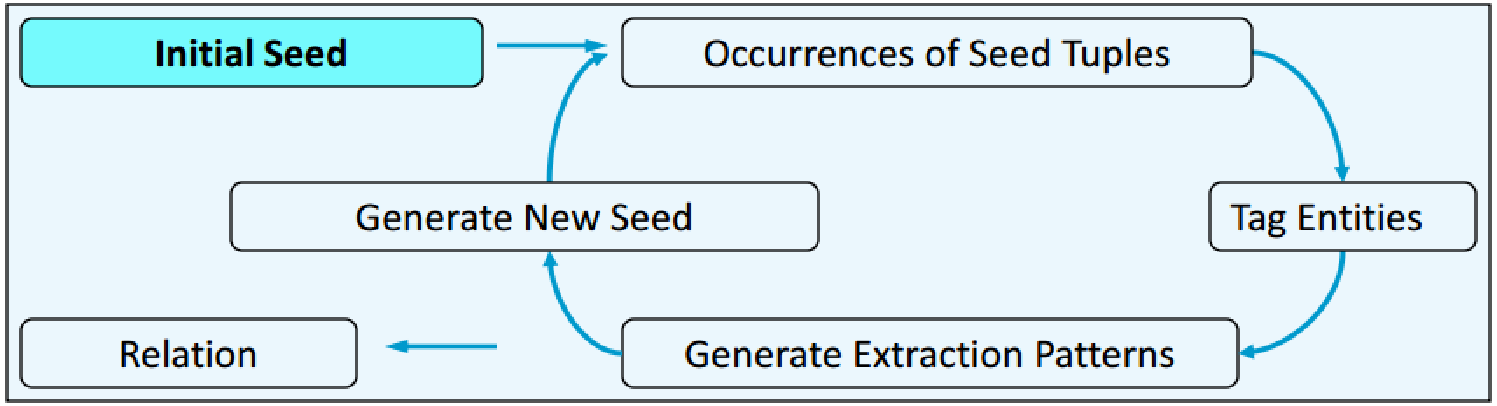
4、基于Bootstrapping的关系抽取

4.1 动机

- 如果没有足够的标注数据，该怎么办？
  - 但是可能有一些种子实体对；
  - 或者可能有一些非常好的pattern；
- 能否用种子实体对或pattern做关系抽取？
  - Bootstrapping通过迭代抽取的方式获得更多的种子实体对和种子pattern；

4.2 基本思想

- 为每种关系标注少量**种子实体**对，基于这些实体对在文本语料库中抽取相关**句子集合**，基于这些句子抽取表达关系的**模式**（pattern），以此循环**迭代**，这个过程也被称之为“滚雪球”（snowball）

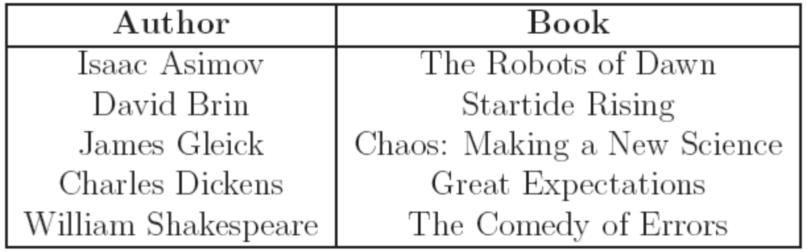


图：基于bootstrapping的关系抽取流程[]

[Eugene Agichtein and Luis Gravano. Snowball: Extracting Relations from Large Plain-Text Collections, Fifth ACM Conference on Digital Libraries. San Antonio, TX, USA, 2000. Page 3.]

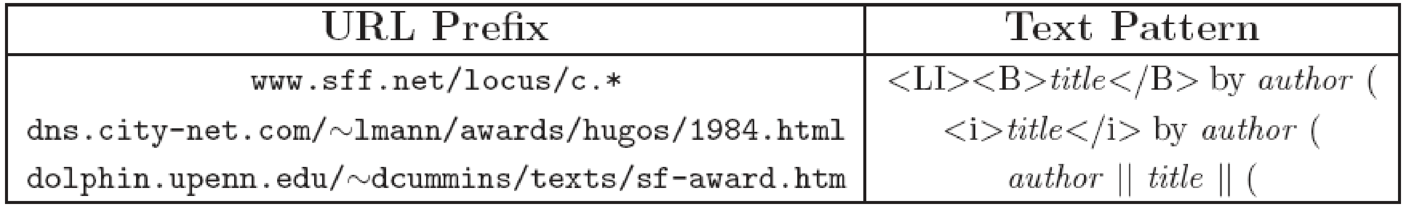
示例

- 英文例子：抽取 <author, book>，引用Dipre (Brin 1998)
  - 以5个种子实体对开始



Author	Book
Isaac Asimov	The Robots of Dawn
David Brin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickens	Great Expectations
William Shakespeare	The Comedy of Errors

  - 使用这些pattern获得更多的实例和pattern



URL Prefix	Text Pattern
`www.sff.net/locus/c.*`	<LI><B>*title*</B> by *author* (
`dns.city-net.com/~lmann/awards/hugos/1984.html`	<i>*title*</i> by *author* (
`dolphin.upenn.edu/~dcummins/texts/sf-award.htm`	*author* || *title* || (

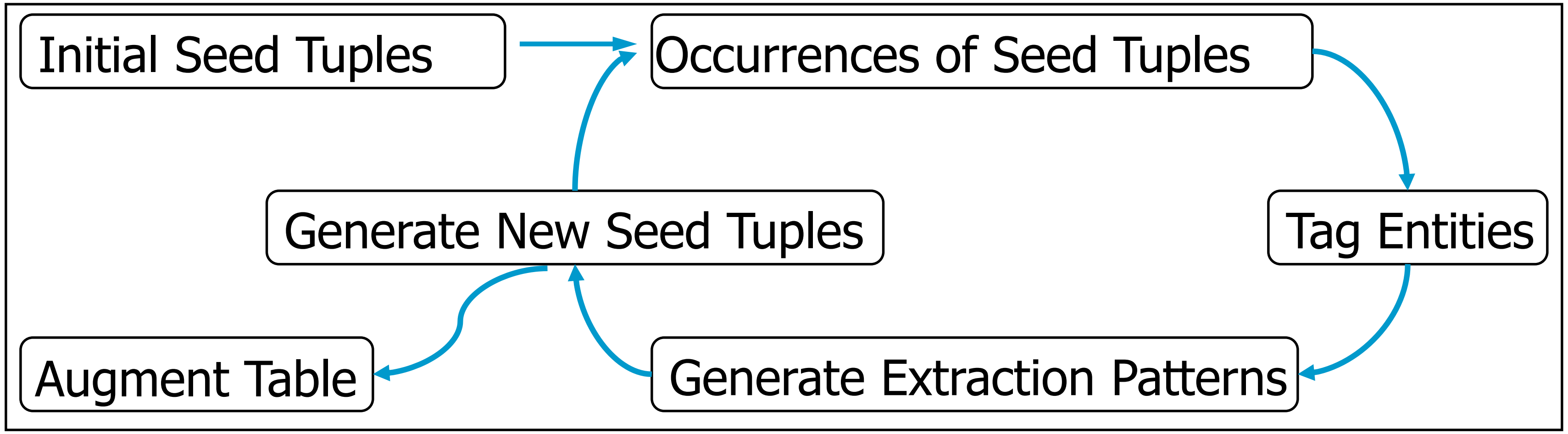
示例（二）

- 基于中文例子的迭代过程
  - Step1：给定关系“出生于”、种子实体对《周杰伦，台湾》和《林丹，福建》
  - Step2：抽取出句子集合：{“周杰伦，出生于台湾省新”，“周杰伦在台湾…”，“林丹小时候在福建学球”}
  - Step3：得到关系“出生于”的描述模式{“，出生于”，“在”，“小时候在”}
  - Step4：基于该模式，抽得句子“林俊杰，出生于新加坡的一个音乐世家”，从而得到实体对《林俊杰，新加坡》
- 代表性系统
  - DIPRE系统 (Brin, 1998)、Snowball系统 (Agichtein, 2000)、KnowItAll系统 (Etzioni et al. 2005)、TextRunner系统 (Banko et al. 2007)

示例：Snowball [Agichtein & Gravano 2000]

- 探索pattern和实体对之间的对偶性
  - 发现匹配给定pattern集合的实体对集合
  - 发现匹配实体对集合的pattern集合

→ *bootstrapping* approach



第一步：基于种子实体对抽取对齐的句子集合

ORGANIZATION	***LOCATION***
MICROSOFT	REDMOND
IBM	ARMONK
BOEING	SEATTLE
INTEL	SANTA CLARA

种子实体对

Computer servers at **Microsoft**’s headquarters in **Redmond**…

In mid-afternoon trading, share of **Redmond**-based **Microsoft** fell…

The **Armonk**-based **IBM** introduced a new line…

The combined company will operate from **Boeing**’s headquarters in **Seattle**.

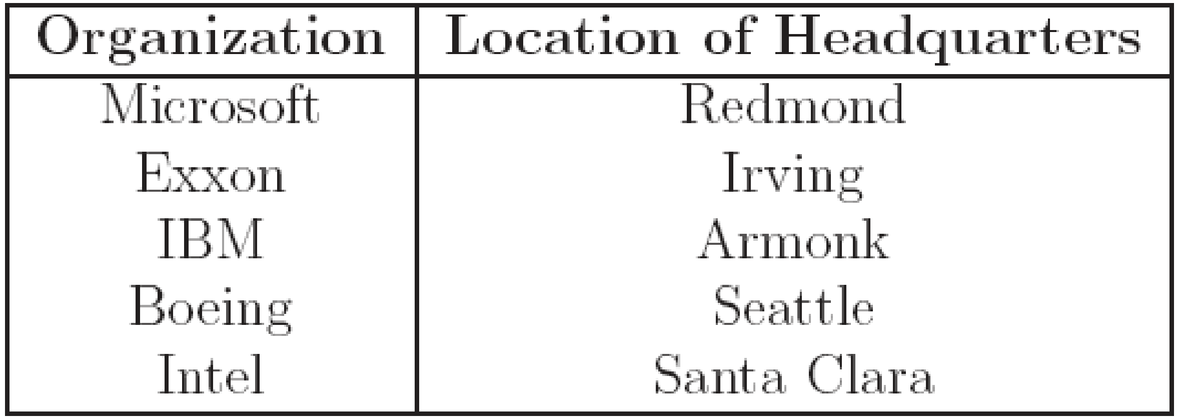
Intel, **Santa Clara**, cut prices of its Pentium processor.

包含种子实体对的句子集合

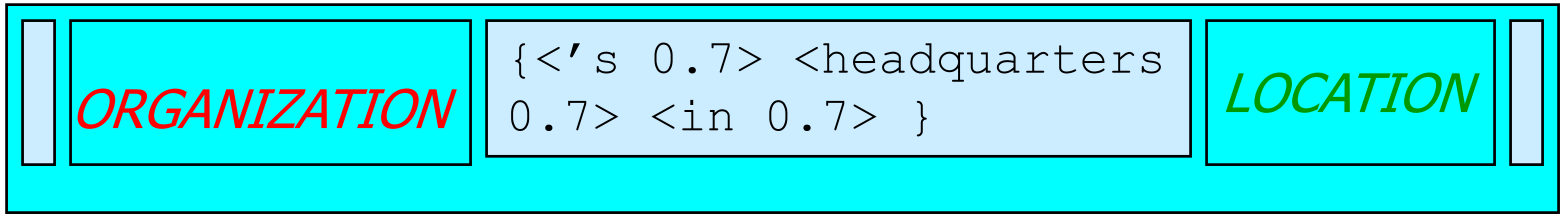
Slides from Dan Jurafsky, Rion Snow, Jim Martin, Chris Manning and William Cohen

第二步：基于对齐的句子集合挖掘候选pattern

- 要求X和Y都是特定类型的命名实体



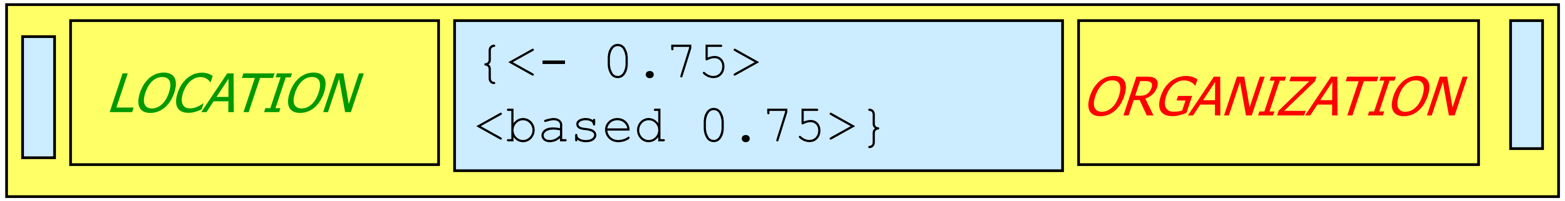
Organization	Location of Headquarters
Microsoft	Redmond
Exxon	Irving
IBM	Armonk
Boeing	Seattle
Intel	Santa Clara



ORGANIZATION

{<’s 0.7> <headquarters 0.7> <in 0.7> }

LOCATION



LOCATION

{<- 0.75> <based 0.75>}

ORGANIZATION

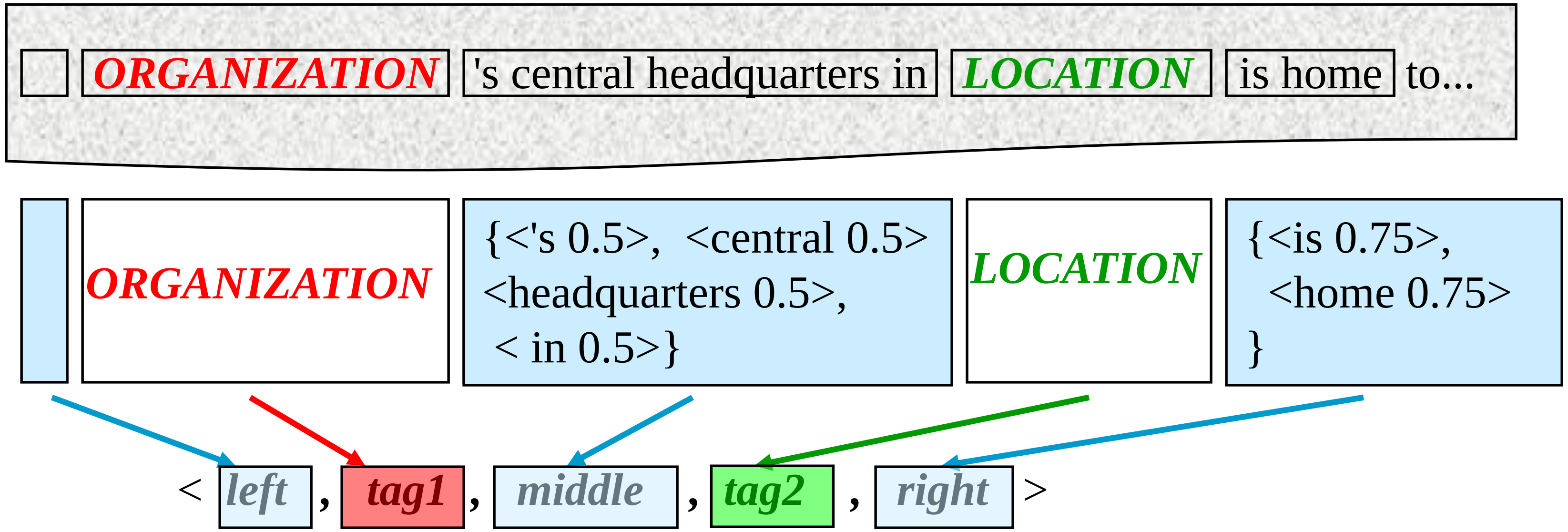
Slides from Dan Jurafsky, Rion Snow, Jim Martin, Chris Manning and William Cohen

第二步：pattern的表示

抽取的pattern具有形式<left, tag1, middle, tag2, right>,

where **tag1**, **tag2** are named-entity tags

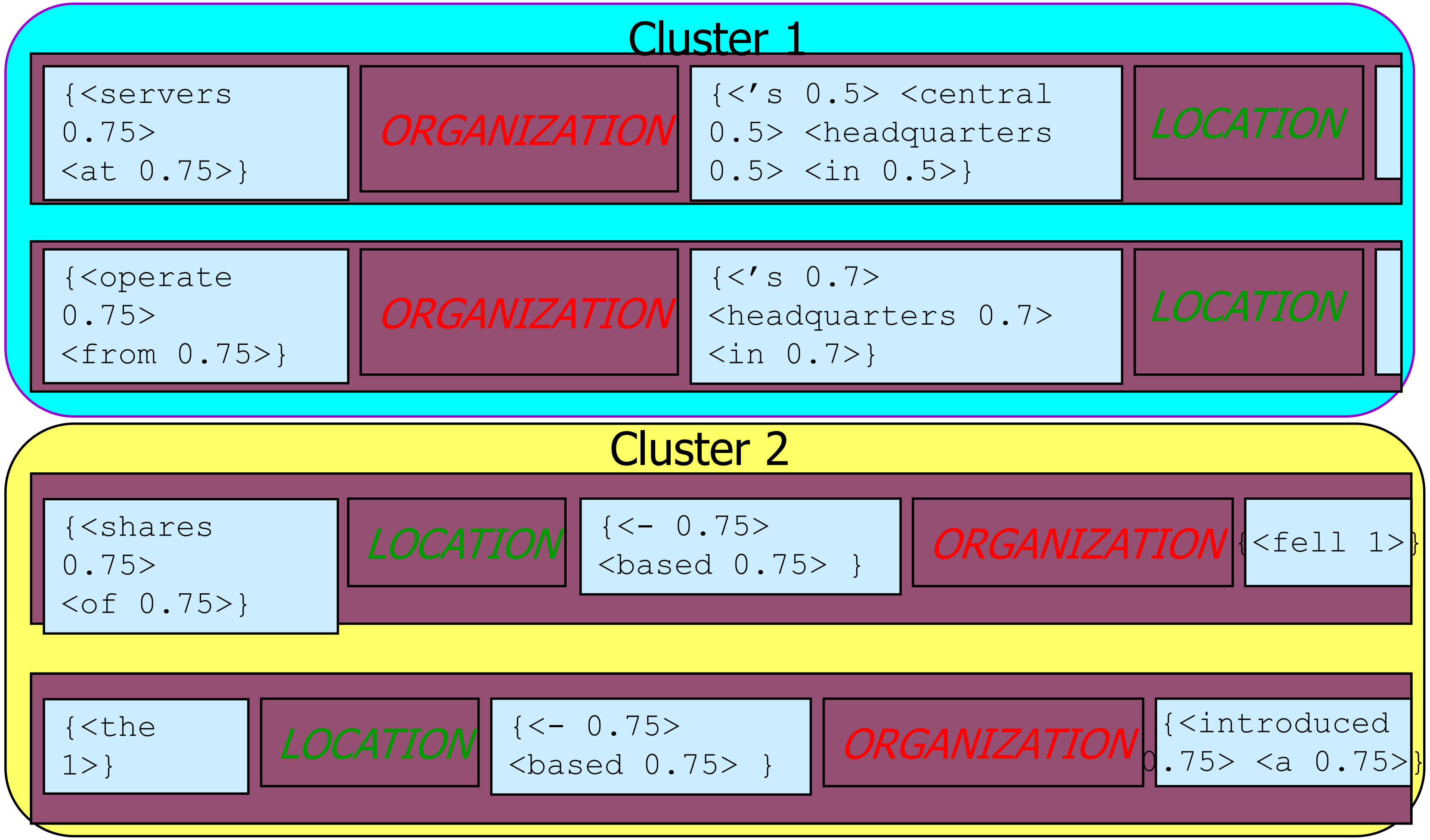
left, **middle**, and **right** are vectors of weighted terms



- patterns derived directly from occurrences are too specific

第三步： 候选pattern聚类

cluster patterns, cluster centroids define patterns



Cluster 1

{<servers 0.75> <at 0.75>}	ORGANIZATION	{<’s 0.5> <central 0.5> <headquarters 0.5> <in 0.5>}	LOCATION	
{<operate 0.75> <from 0.75>}	ORGANIZATION	{<’s 0.7> <headquarters 0.7> <in 0.7>}	LOCATION	

Cluster 2

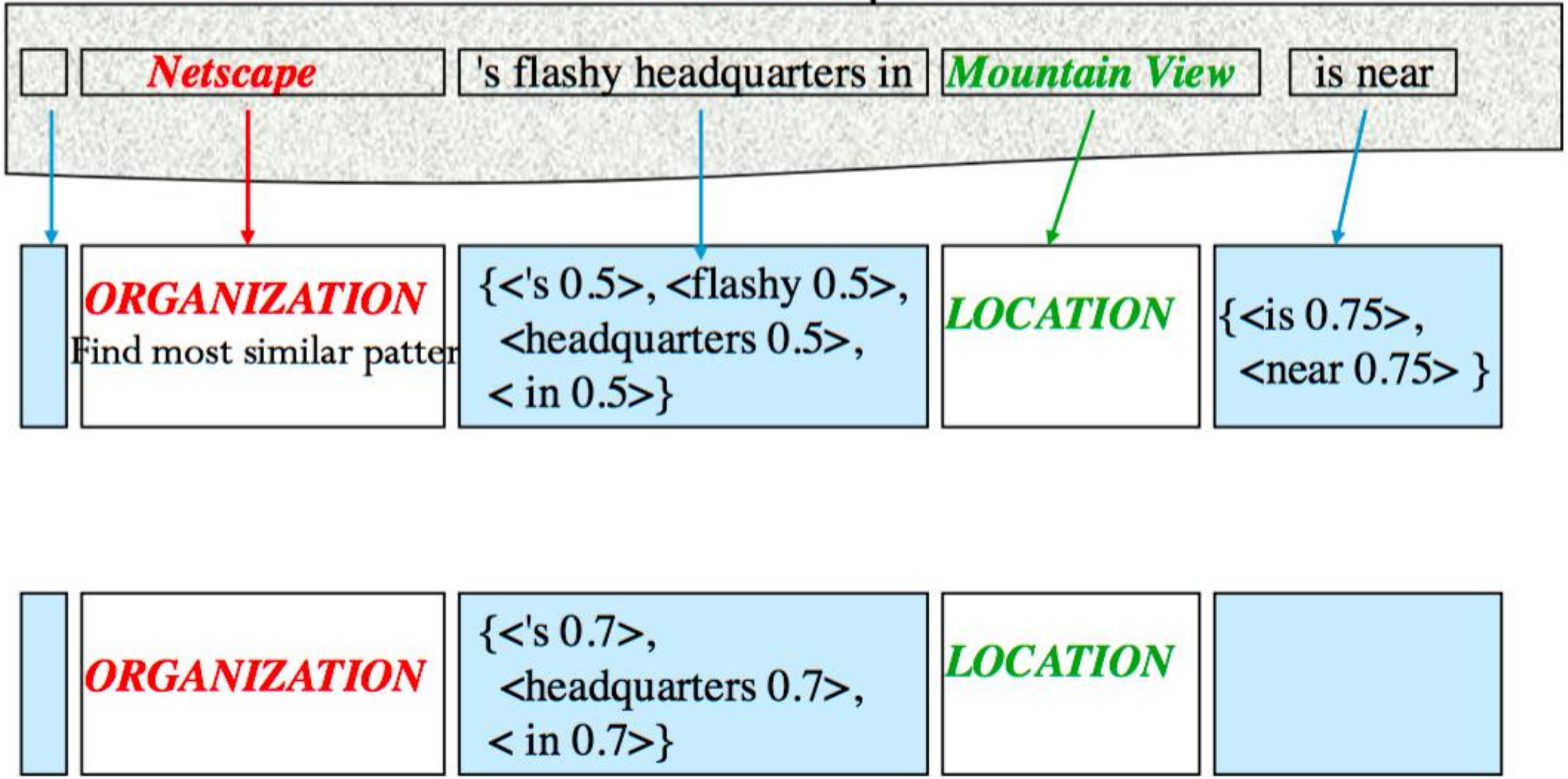
{<shares 0.75> <of 0.75>}	LOCATION	{<- 0.75> <based 0.75> }	ORGANIZATION	{<fell 1>}
{<the 1>}	LOCATION	{<- 0.75> <based 0.75> }	ORGANIZATION	{<introduced 0.75> <a 0.75>}

The pattern generation uses a simple single-pass clustering method to group similar tuples and generate a corresponding new pattern.

Slides from Dan Jurafsky, Rion Snow, Jim Martin, Chris Manning and William Cohen

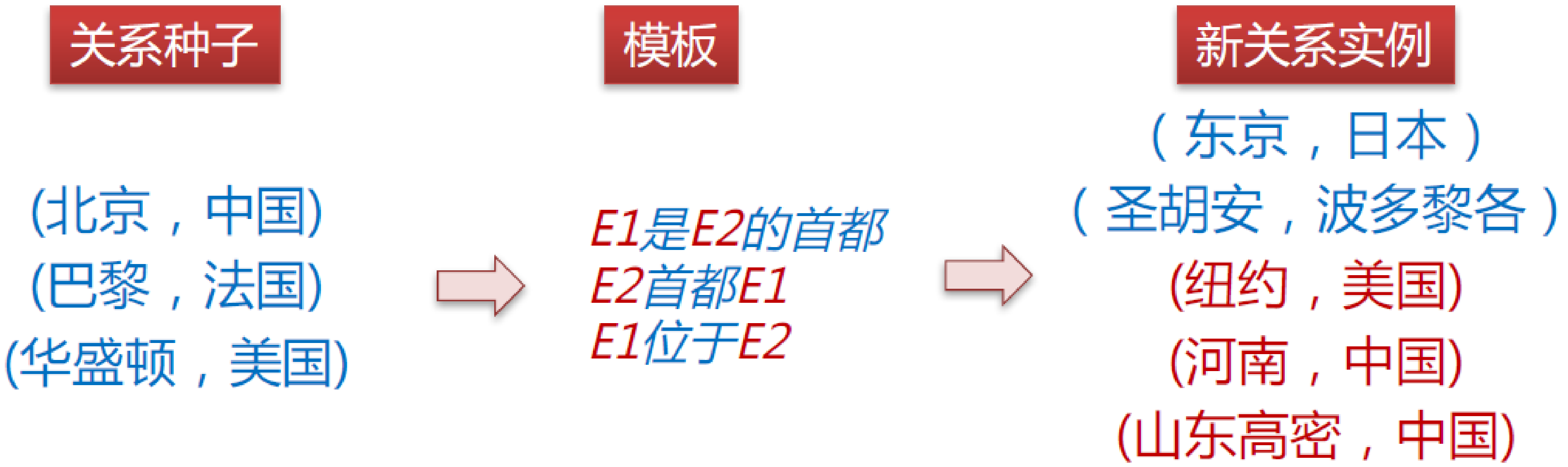
第四步：基于发现的新pattern抽取实体对

Represent each new text segment in the collection as the context 5-tuple:


Netscape
's flashy headquarters in
Mountain View
is near
ORGANIZATION
Find most similar patter
{<'s 0.5>, <flashy 0.5>, <headquarters 0.5>, < in 0.5>}
LOCATION
{<is 0.75>, <near 0.75> }
ORGANIZATION
{<'s 0.7>, <headquarters 0.7>, < in 0.7>}
LOCATION

4.3 语义漂移（Semantic drift）

- 迭代会引入噪音实例和噪音模板
- 例子



关系种子

(北京，中国)
(巴黎，法国)
(华盛顿，美国)

模板

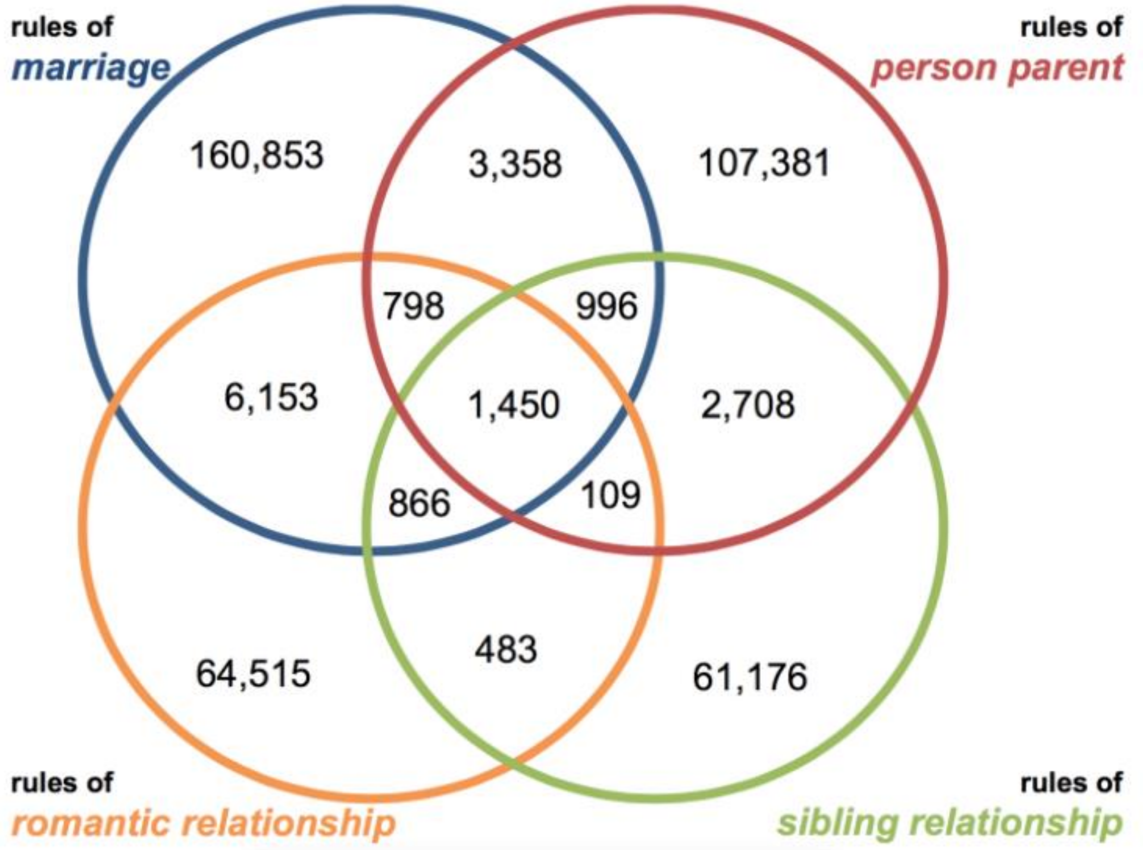
E1是E2的首都
E2首都E1
E1位于E2

新关系实例

（东京，日本）
（圣胡安，波多黎各）
(纽约，美国)
(河南，中国)
(山东高密，中国)

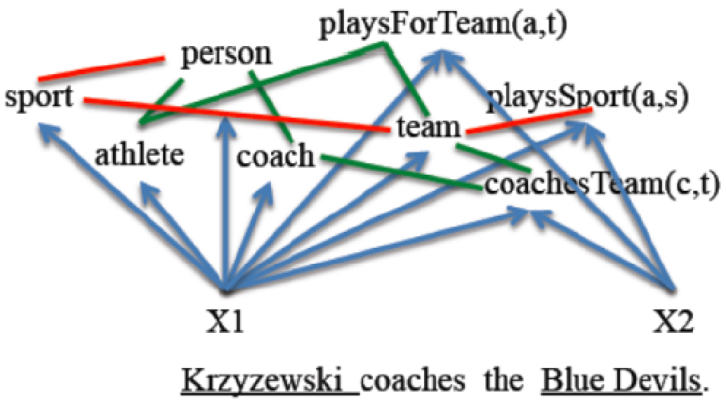
pattern的交叉度量

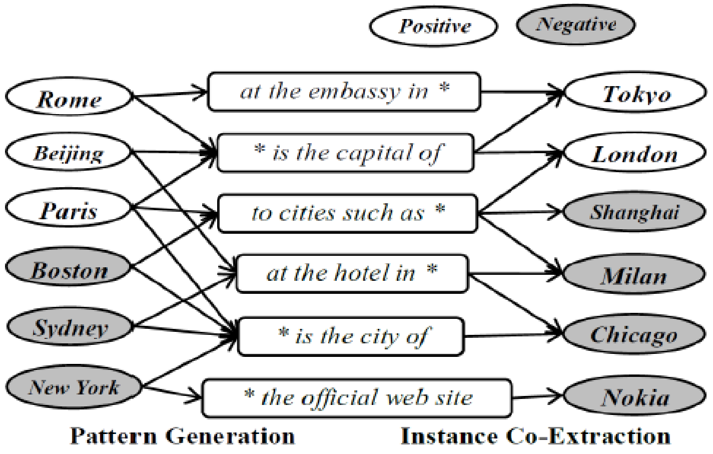
- 根据[Krause et al.,2012]，人物之间四种关系Pattern的交叉程度



语义漂移的解决方案

- Bootstrapping-语义漂移解决方案
  - Mutual exclusive Bootstrapping (McIntosh et al., 09)：同时扩展多个互斥类别，一个实体对只能属于一个类别；
  - Coupled training（Carlson et al., 10）：建模不同抽取关系之间的约束，寻找最大化满足这些约束的抽取结果；
  - 关系之间的约束，寻找最大化满足这些约束的抽取结果 Co-Bootstrapping (Shi et al. 14)：引入负实例来限制语义漂移；



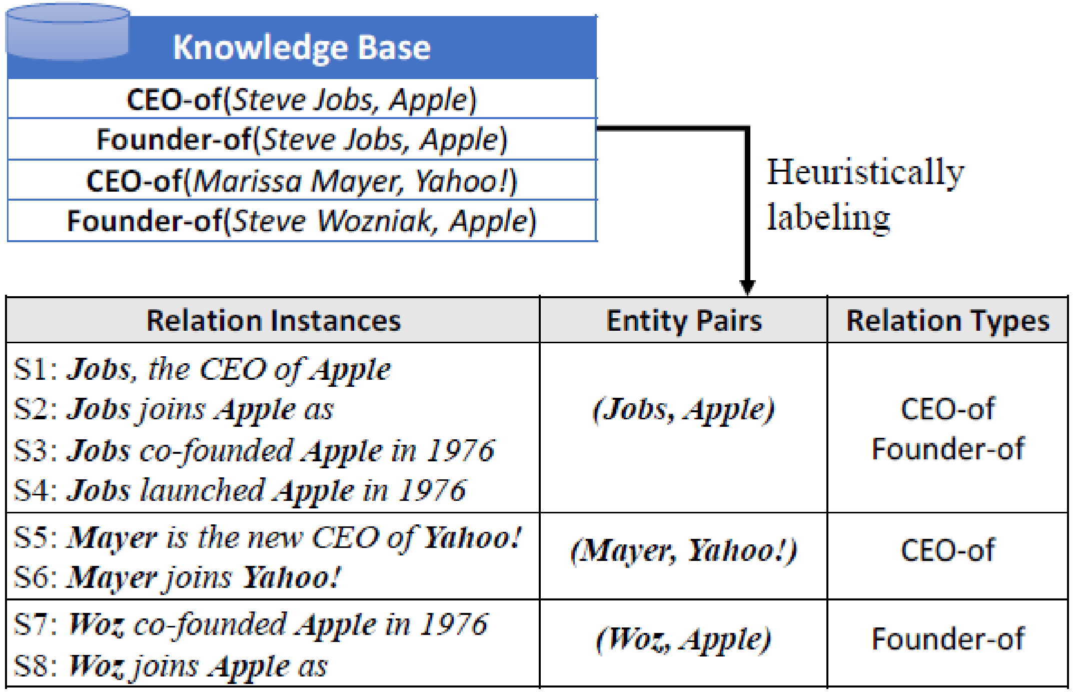


5、基于远程监督的关系抽取

5.1 远程监督概述

基本假设：

若一个实体对在知识库中存在某个关系，那么包含该实体对的所有句子都以某种方式表达该关系。



Knowledge Base
CEO-of(*Steve Jobs, Apple*)
Founder-of(*Steve Jobs, Apple*)
CEO-of(*Marissa Mayer, Yahoo!*)
Founder-of(*Steve Wozniak, Apple*)

Relation Instances	Entity Pairs	Relation Types
S1: *Jobs, the CEO of Apple* S2: *Jobs joins Apple as* S3: *Jobs co-founded Apple in 1976* S4: *Jobs launched Apple in 1976*	*(Jobs, Apple)*	CEO-of Founder-of
S5: *Mayer is the new CEO of Yahoo!* S6: *Mayer joins Yahoo!*	*(Mayer, Yahoo!)*	CEO-of
S7: *Woz co-founded Apple in 1976* S8: *Woz joins Apple as*	*(Woz, Apple)*	Founder-of

Mintz, Bills, Snow, Jurafsky. 2009. Distant supervision for relation extraction without labeled data. ACL09

远程监督的动机

- Combine bootstrapping with supervised learning
  - Instead of 5 seeds,
    - Use a large database to get huge # of seed examples
  - Create lots of features from all these examples
  - Combine in a supervised classifier

Snow, Jurafsky, Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. NIPS 17
Fei Wu and Daniel S. Weld. 2007. Autonomously Semantifying Wikipeida. CIKM 2007
Mintz, Bills, Snow, Jurafsky. 2009. Distant supervision for relation extraction without labeled data. ACL09

5.2 基于远程监督的数据集构造过程

- Step 1：从知识库（如Freebase）中为目标关系类别抽取尽可能多的实体对;
- Step 2：对于每个实体对，基于实体链接技术从大规模文本中抽取包含该实体对mention的句子集合，并为每个句子标注相应的关系;
- Step 3：包含实体对的句子集合和关系类型构成关系抽取的数据集，即实体对的训练数据为相应的句子，标签为知识库中的关系类型。

5.3 远程监督：优点与缺点

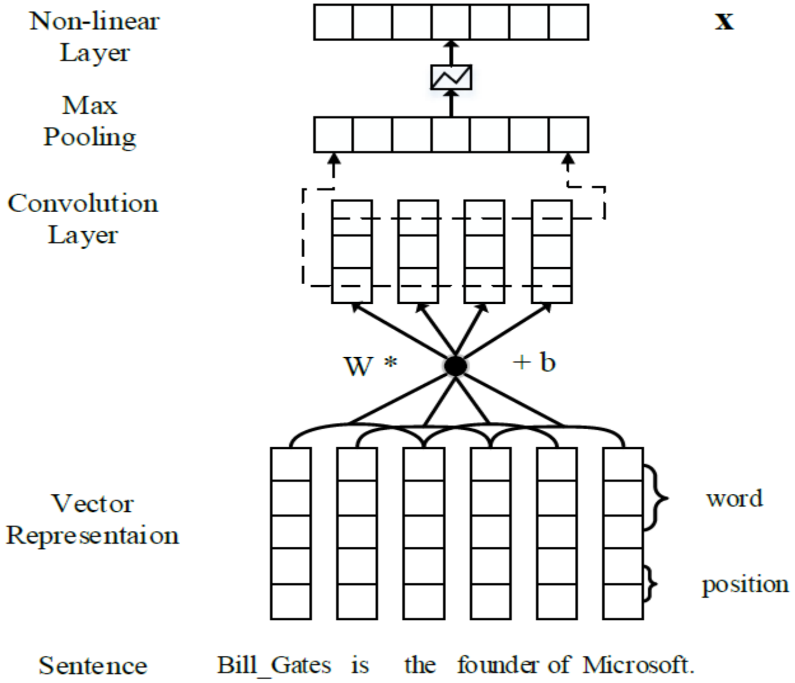
- 优点
  - 减少人工标注代价
  - 可扩展性：可以使用大量未标记的数据
- 缺点
  - 训练语料库含有大量错标的噪声，包含实体对的句子可能没有表达目标语义关系；
  - 没有严格合理的负样本用于训练

5.4 基于远程监督的关系抽取方法

- 传统的基于特征抽取的方法
  - POS, WordNet, FrameNet, 依存分析、句法分析
  - 特征抽取容易造成错误累计，影响分类性能
  - 无法充分利用训练数据的隐式语义信息
- 基于深度学习的方法
  - 自动学习句子的语义
  - 容易实现端到端的抽取

示例：基于卷积神经网络的关系抽取

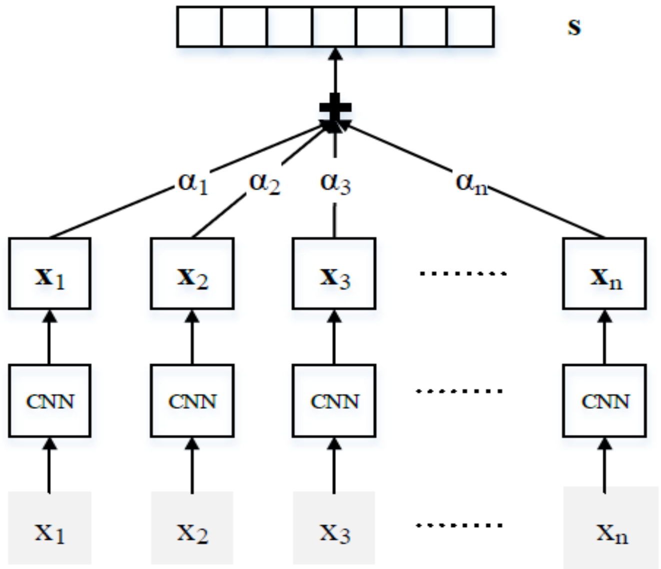
- 卷积神经网络（CNN）被用于句子建模取得了很大的成功;
- 基本思路：利用CNN学习句子语义表示，将表示结果用于关系分类;
- 基本框架：



图：基于CNN的句子表示学习架构[Santos et al. 2015]

示例：基于注意力机制的关系抽取

- 动机
  - 基于远程监督得到的训练集噪声严重，严重影响模型性能
- 解决思路
  - 实体对的句子集合中，不同实例对于分类的贡献不一样，使用句子级别的注意力机制学习实例权重；
- 框架

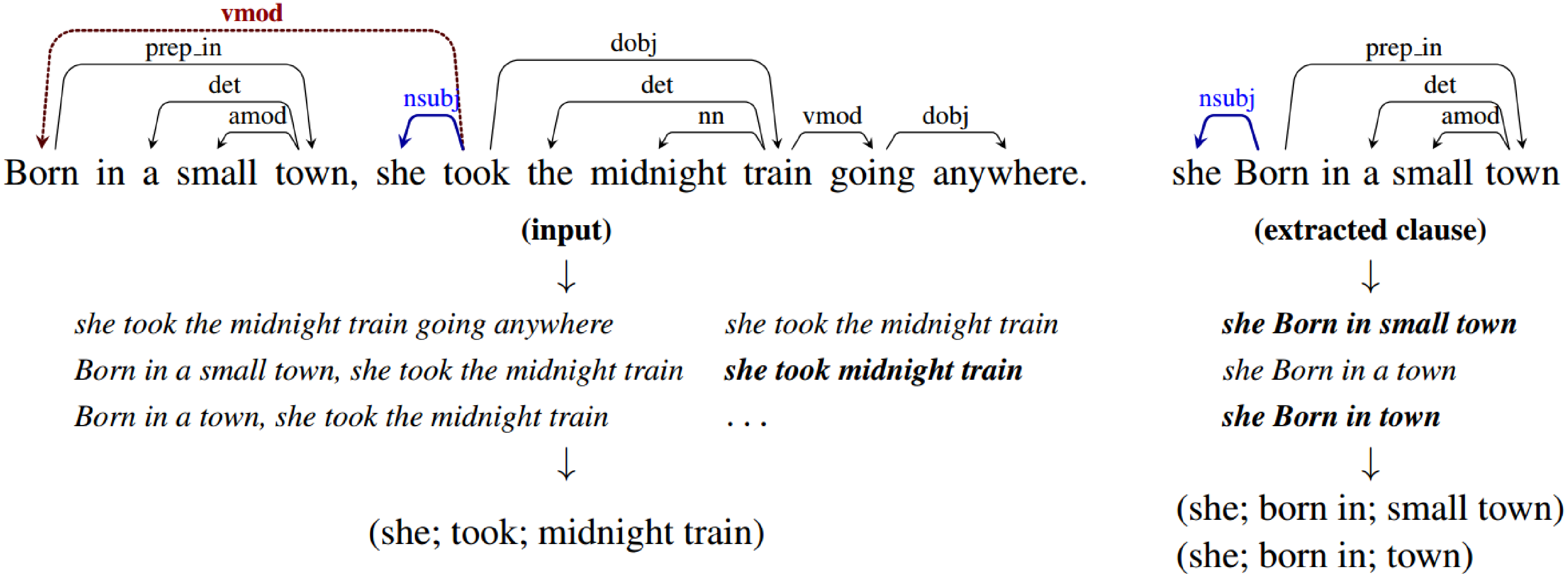


图：基于句子级别的注意力机制的关系抽取[Lin et al. 2016]

6、开放关系抽取

6.1 概述

- Open information extraction (open IE) refers to the extraction of relation tuples, typically binary relations, from plain text, such as *(Mark Zuckerberg; founded; Facebook)*.



Born in a small town, she took the midnight train going anywhere.

(input)

she took the midnight train going anywhere
Born in a small town, she took the midnight train
Born in a town, she took the midnight train

she took the midnight train
she took midnight train
. . .

(she; took; midnight train)

she Born in a small town

(extracted clause)

she Born in small town
she Born in a town
she Born in town

(she; born in; small town)
(she; born in; town)

https://nlp.stanford.edu/software/openie.html

6.2 基本方法

训练一个序列分类或标注模型（通常使用语法和POS特征）

Rel. Freq.	Category	Simplified Lexico- Syntactic Pattern	Example
37.8	Verb	E_1 Verb E_2	X established Y
22.8	Noun+Prep	E_1 NP Prep E_2	X settlement with Y
16.0	Verb+Prep	E_1 Verb Prep E_2	X moved to Y
9.4	Infinitive	E_1 to Verb E_2	X plans to acquire Y
5.2	Modifier	E_1 Verb E2 Noun	X is Y winner
1.8	Coordinate$_n$	E_1 (and|,|-|:) E_2 NP	X-Y deal
1.0	Coordinate$_v$	E_1 (and|,) E_2 Verb	X , Y merge
0.8	Appositive	E_1 NP (:|,)? E_2	X hometown : Y

开放关系的常见pattern

[Etzioni *et al.*. Open Information Extraction from the Web. Communications of the ACM, vol. 51 no. 12, Dec. 2008.]

[Banko and Etzioni. The Tradeoffs Between Open and Traditional Relation Extraction. Proc. of the ACL, Columbus, OH, USA, June 2008.]

基本过程

“三步”过程：

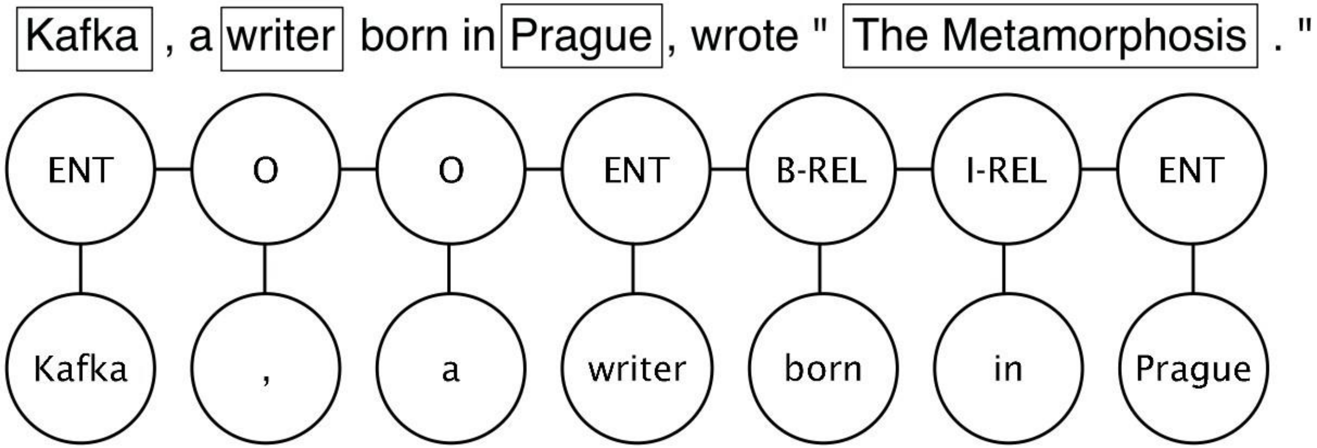
Label: Sentences are automatically labeled with extractions using heuristics or distant supervision.

Learn: A relation phrase extractor is learned, e.g. using a sequence-labeling graphical model (CRF).

Extract: The system takes a sentence as input, identifies a candidate pair of NP arguments (arg1, arg2) from the sentence, and then uses the learned extractor to label each word between the two arguments as part of the relation phrase or not.

[Fader *et al*.. Identifying Relations for Open Information Extraction. Proc. of EMNLP, Edinburgh, Scotland, UK, July 2011.]

示例：开放信息抽取的标注模型


Kafka , a writer born in Prague , wrote " The Metamorphosis . "
ENT
O
O
ENT
B-REL
I-REL
ENT
Kafka
,
a
writer
born
in
Prague

[Banko and Etzioni. The Tradeoffs Between Open and Traditional Relation Extraction. Proc. of the ACL, Columbus, OH, USA, June 2008.]

6.5 关系短语的归一化

- Shakespeare (*has written* | *wrote* | *was writing*) Hamlet.
- ->Shakespeare *write* Hamlet.

- Allow for minor variations in relation phrases.
  - Remove inflection
  - Remove auxiliary verbs, adjectives, adverbs

5、参考文献

- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. ACL, pages 1003–1011.
- Collins M, Duffy N, Park F. Parsing with a single neuron: Convolution kernels for natural language problems[J]. 2001.
- Lodhi H, Saunders C, Shawe-Taylor J, et al. Text classification using string kernels[J]. Journal of Machine Learning Research, 2002, 2(Feb): 419-444.
- [Eugene Agichtein and Luis Gravano. Snowball: Extracting Relations from Large Plain-Text Collections, Fifth ACM Conference on Digital Libraries. San Antonio, TX, USA, 2000. Page 3.]
- Sergey Brin, Extracting Patterns and Relations from the World Wide Web, Proc. of International Workshop on the Web and Databases, 1998.
- [Agichtein and Gravano, 2000] E. Agichtein and L. Gravano. Snowball: Extracting relations from large plain-text collections. In Proceedings of the Fifth ACM International Conference on Digital Libraries, 2000.
- [Downey et al., 2005] D. Downey, O. Etzioni, and S. Soderland. A Probabilistic Model of Redundancy in Information Extraction. In Proc. of IJCAI, 2005.
- [Cafarella et al., 2006] Michael J. Cafarella, Michele Banko, and Oren Etzioni. Relational web search. Technical Report 06-04-02, University of Washington, 2006.
- Sawyer S, Krause J, Guschanski K, et al. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA[J]. PloS one, 2012, 7(3): e34131.

- McIntosh K R, Cotsell J N, Cumpston J S, et al. An optical comparison of silicone and EVA encapsulants for conventional silicon PV modules: A ray-tracing study[C]//Photovoltaic Specialists Conference (PVSC), 2009 34th IEEE. IEEE, 2009: 000544-000549.
- Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning[C]//AAAI. 2010, 5: 3.
- Zhou Z H. A brief introduction to weakly supervised learning[J]. National Science Review, 2017, 5(1): 44-53.
- Snow, Jurafsky, Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. NIPS 17
- Fei Wu and Daniel S. Weld. 2007. Autonomously Semantifying Wikipeida. CIKM 2007
- Mintz, Bills, Snow, Jurafsky. 2009. Distant supervision for relation extraction without labeled data. ACL09
- Snow, Jurafsky, Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. NIPS 17
- Mintz, Bills, Snow, Jurafsky (2009) Distant supervision for relation extraction without labeled data. ACL-2009.
- Han X, Sun L. Distant Supervision via Prototype-Based Global Representation Learning[C]//AAAI. 2017: 3443-3449.
- Santos C N, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks[J]. arXiv preprint arXiv:1504.06580, 2015.
- Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, 1: 2124-2133.
- Banko and Etzioni. The Tradeoffs Between Open and Traditional Relation Extraction. Proc. of the ACL, Columbus, OH, USA, June 2008.