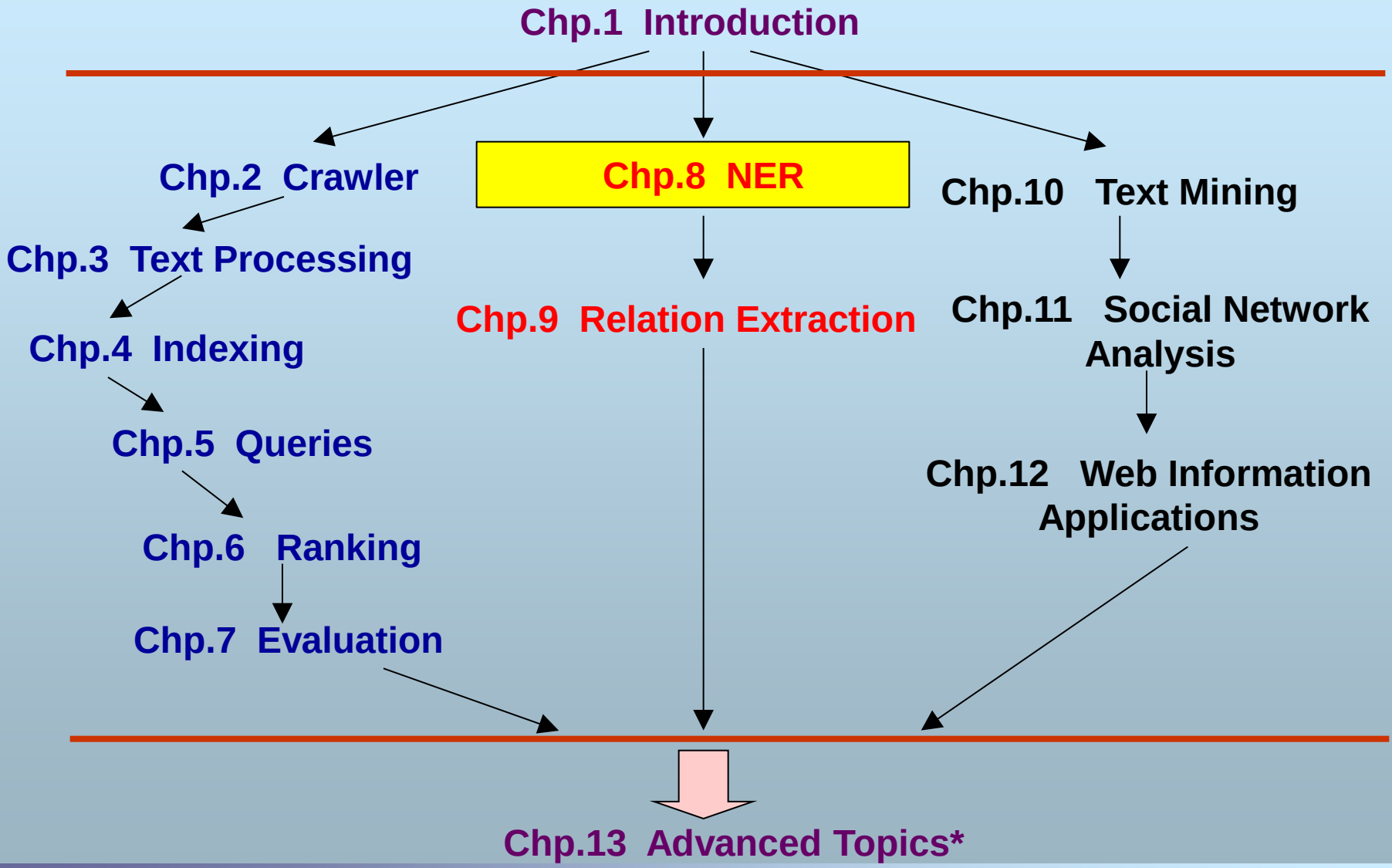
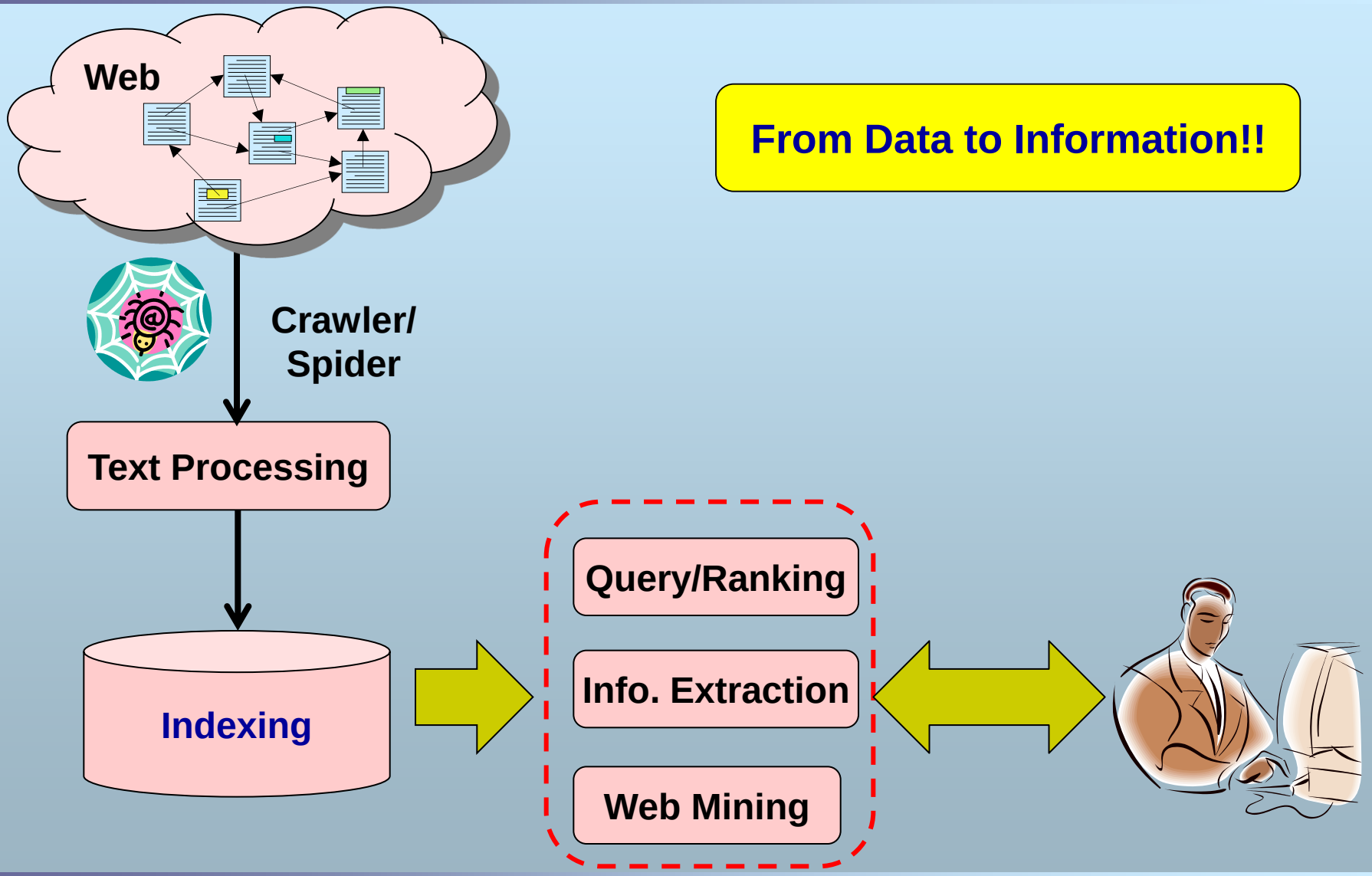
Named Entity Recognition

金培权（jpq@ustc.edu.cn）

课程知识结构


Chp.1 Introduction
Chp.2 Crawler
Chp.3 Text Processing
Chp.4 Indexing
Chp.5 Queries
Chp.6 Ranking
Chp.7 Evaluation
Chp.8 NER
Chp.9 Relation Extraction
Chp.10 Text Mining
Chp.11 Social Network Analysis
Chp.12 Web Information Applications
Chp.13 Advanced Topics*

本章讨论的问题


Web
Crawler/
Spider
Text Processing
Indexing
Query/Ranking
Info. Extraction
Web Mining
From Data to Information!!

本章主要内容

- 信息抽取 Information Extraction
- 命名实体识别 Named Entity Recognition

一、信息抽取

- 传统的信息检索实际上“文档检索”，其结果是“文档”的集合，并非信息

一、信息抽取

- 大多数情况下用户想要的是“信息”
  - 中国科学技术大学计算机学院院长，李向阳

1、信息抽取含义

- 从语料中抽取指定的事件、事实等信息，形成结构化的数据
  - 从语料中抽取用户感兴趣的事件、实体和关系
  - 被抽取的信息以结构化的形式描述
  - 为情报分析、检测、比价购物、自动文摘、文本分类等各种应用提供服务

Web时代的企业——“数据富翁”，“信息穷人”

2、信息抽取应用

- 企业竞争情报

竞争对手发现

雷蒙德：BP集团副总裁、BP中国首席执行官及总裁
孙振耀：惠普全球副总裁兼中国惠普公司总裁
陈永正：微软中国有限公司总裁(原摩托罗拉中国区总裁)
庞德明：通用电气（中国）有限公司中国区总裁
关志华：巴斯夫中华区总裁
周伟焜：IBM大中华区董事长兼首席执行总裁
林正刚：思科系统（中国）网络技术有限公司中国区总裁
路易普：ABB公司中国区总裁
陈永正：微软公司副总裁、微软大中华区首席执行官
何庆源：诺基亚（中国）投资有限公司总裁
小泽秀树：佳能中国区总裁兼佳能亚洲营销集团总裁
高瑞彬：摩托罗拉（中国）电子有限公司总裁
罗宏斐：宝洁（中国）有限公司总裁
朱华熙：百事（中国）投资有限公司总裁

职位关系抽取

2、信息抽取应用

- 其它领域的应用
  - 灾害预防部门从自然灾害的新闻报道中抽取出灾害的类型、时间、地点、人员伤亡、经济损失等情况
  - 从病人的医疗记录中抽取出症状、诊断记录和检验结果
  - 税务分析不同企业交税记录、发现异常模式和趋势

3、信息抽取与文本理解

- 信息抽取需要一定程度的理解
  - 只关心有限的感兴趣的事实信息
  - 不关心文本意义的细微差别
  - 不关心作者的写作意图等深层理解问题
- 信息抽取只能算一种浅层的文本理解
- 信息抽取可以看作信息检索的进一步深化

4、信息抽取 vs. 信息检索

- 密切相关但又存在差异
  - 功能不同
    - 检索：从文档集合中找文档子集
    - 抽取：从文本中获取用户感兴趣的事实信息
  - 处理技术不同
    - 检索：通常利用统计与关键词等技术
    - 抽取：借助于自然语言处理技术
  - 使用领域不同
    - 检索：通常领域无关
    - 抽取：通常领域相关

5、信息抽取的任务

- MUC会议 Message Understanding Conference
  - 美国国防高级研究计划委员会(DARPA)资助
  - 评测信息抽取系统
  - 87-98进行了7次，MUC-1, …, MUC-7
- MUC-7定义了5类信息抽取任务，分别进行评测
  - 命名实体NE
  - 模板元素TE
  - 共指关系CR
  - 模板关系TR
  - 背景模板ST

5、信息抽取的任务

- 1、命名实体 NE　（实体抽取）
  - 最主要的任务
  - 命名实体是文本中基本的信息元素，是正确理解文本的基础
  - 狭义：指现实世界中具体或抽象的实体
    - 如 人、组织、地点等
    - “中国科学技术大学/Org 校长 包信和/Person”
  - 广义：还可以包含日期和时间、数量表达式等
  - 具体含义由应用来确定

5、信息抽取的任务

- 2、模板元素TE　　（属性抽取）
  - 模板元素又称为实体的属性
  - 通过槽（Slots）描述了命名实体的基本信息
  - 为命名实体建立各种属性槽从而更加清楚地描述命名实体
  - 槽Slots：名称、类别、描述符、种类等

5、信息抽取的任务

- 3、共指关系 CR　　（实体间的共指关系）
  - 不同的命名实体表达了相同的含义，这些实体之间的关系就是共指，也称为等价概念
  - 共指任务在于抽取关于共指表达的信息
  - 包括那些已在命名实体和模板元素任务中作了标记的对于某个命名实体的所有表述

5、信息抽取的任务

- 4、模板关系TR　　（关系抽取）
  - 实体之间的各种关系，又称为事实
    - 雇佣关系(employee_of)、生产关系(product_of) ...
    - 如：*post_of*(校长, 包信和),
      employee_of(中国科学技术大学, 包信和)

5、信息抽取的任务

- 场景模板 ST　　（事件抽取）
  - 又称事件，是指实体发生的事件
  - 例如
    - 会议(Time<…>, Spot<…>, Convener<…>, Topic<…>)
  - 新闻事件 5W1H
    - Who、When、Where、What、Why、How

6、信息抽取示例

- 人民日报1998-01-07

19980107-06-016-001意大利总理普罗迪４日说，欧洲国家将采取行动，共同对付库尔德难民涌入问题。普罗迪４日晚召开了由意外长、内政和国防部长参加的紧急会议，商讨应付库尔德难民问题的对策。会前，普罗迪说，“在经过最初的混乱后，欧洲国家的行动已经大大加强”，今后几天内将在此问题上进行系统合作。

6、信息抽取示例

- **NE实体抽取结果示例**

```
<NamedEntities>
        <PersonList>
                库尔德    (occurrence: 1/1/15; 1/2/19;)
                普罗迪    (occurrence: 1/1/3; 1/2/0; 1/3/2;)
        </PersonList>
        <OrgList>
        </OrgList>
</NamedEntities>
```

- **TR关系抽取结果示例**

```
<EntityRelations>
                post_of(意大利总理,普罗迪)
</EntityRelations>
```

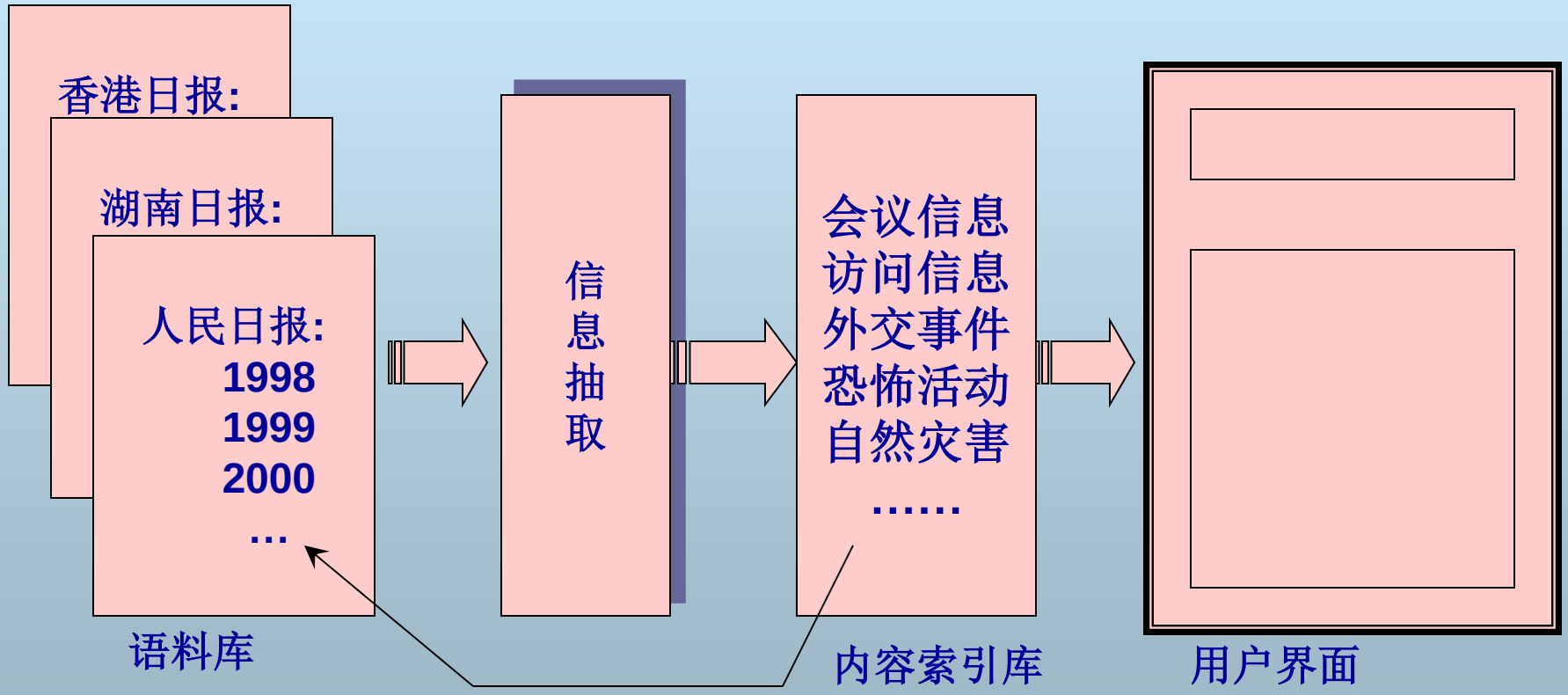
6、信息抽取示例

- **ST事件抽取示例**

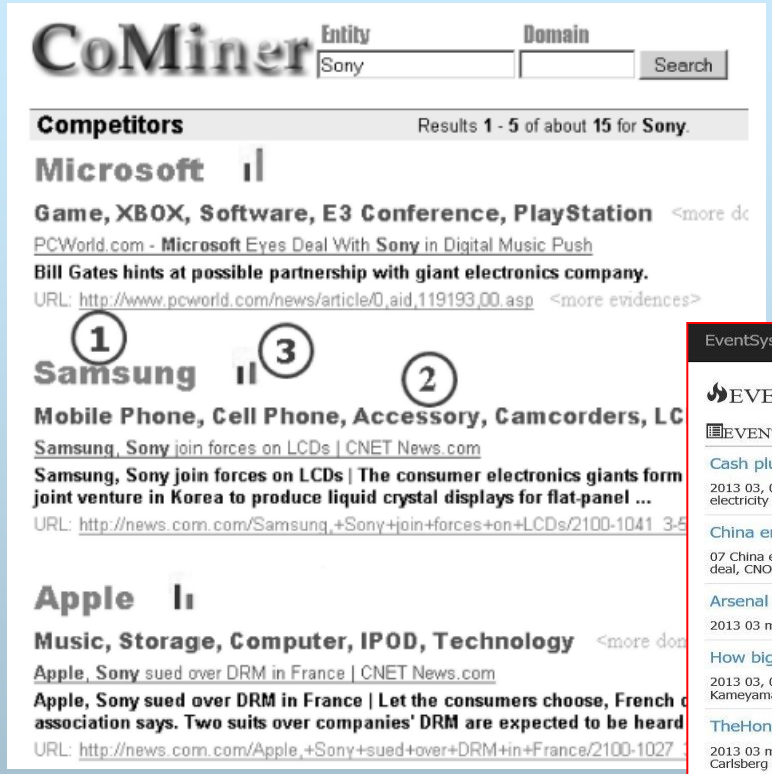
```
<EventTemplateInstatnces>
<ConferenceInfo>
        <Time> 4 日晚 (1998-01)</Time>
        <Spot>意大利</Spot>
        <Converner>普罗迪</Converner>
        <Title>由意外长、内政和国防部长参加的紧急会议
</Title>
</ConferenceInfo>
</EventTemplateInstatnces>
```

会议时间 Time	4 日晚 (1998-01)	
会议地点 Spot	意大利	
召集人 Convener	姓名/团体名称 Name	普罗迪
	机 构 、 职 位 Org/Post	意大利总理
会 议 名 / 标 题 Conf-Title	由意外长、内政和国防部长参加的紧急会议	

7、信息抽取的应用架构

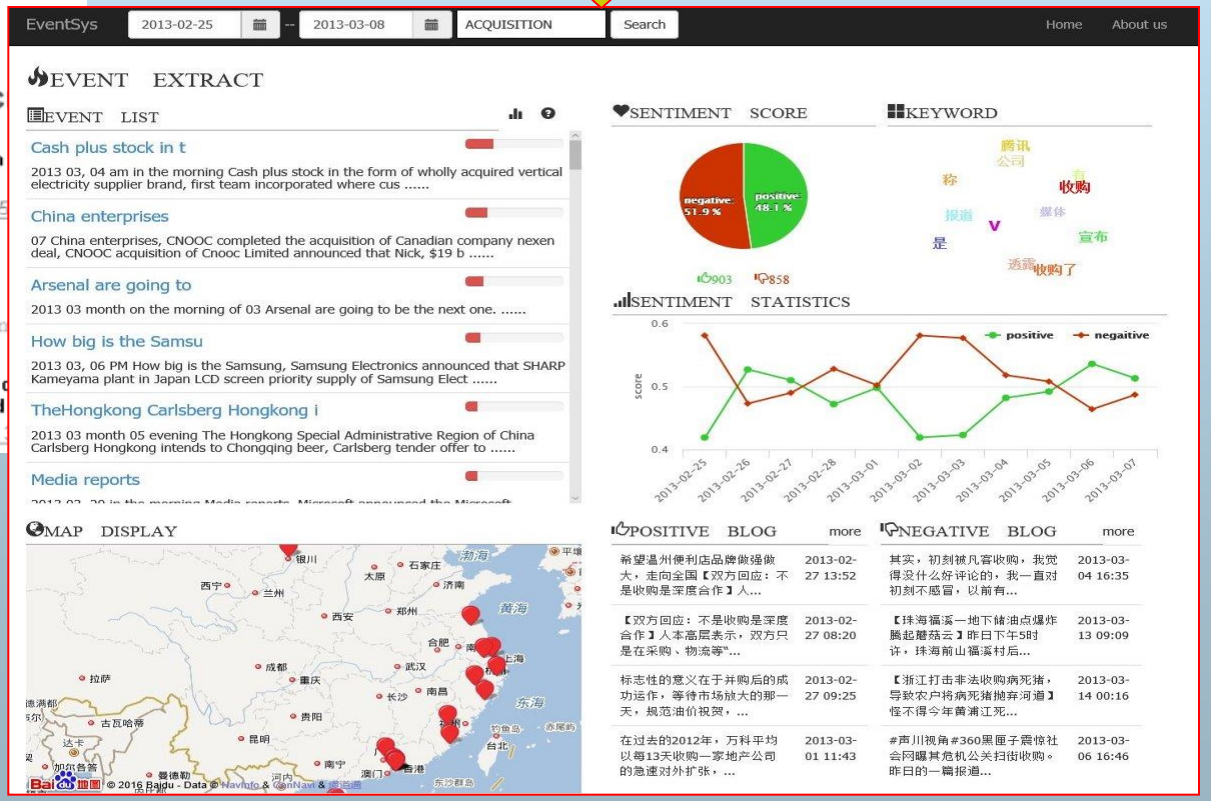

香港日报:
湖南日报:
人民日报:
1998
1999
2000
…
语料库
信
息
抽
取
会议信息
访问信息
外交事件
恐怖活动
自然灾害
……
内容索引库
用户界面

7、信息抽取的应用架构



面向网页的竞争对手抽取

面向微博的突发事件追踪

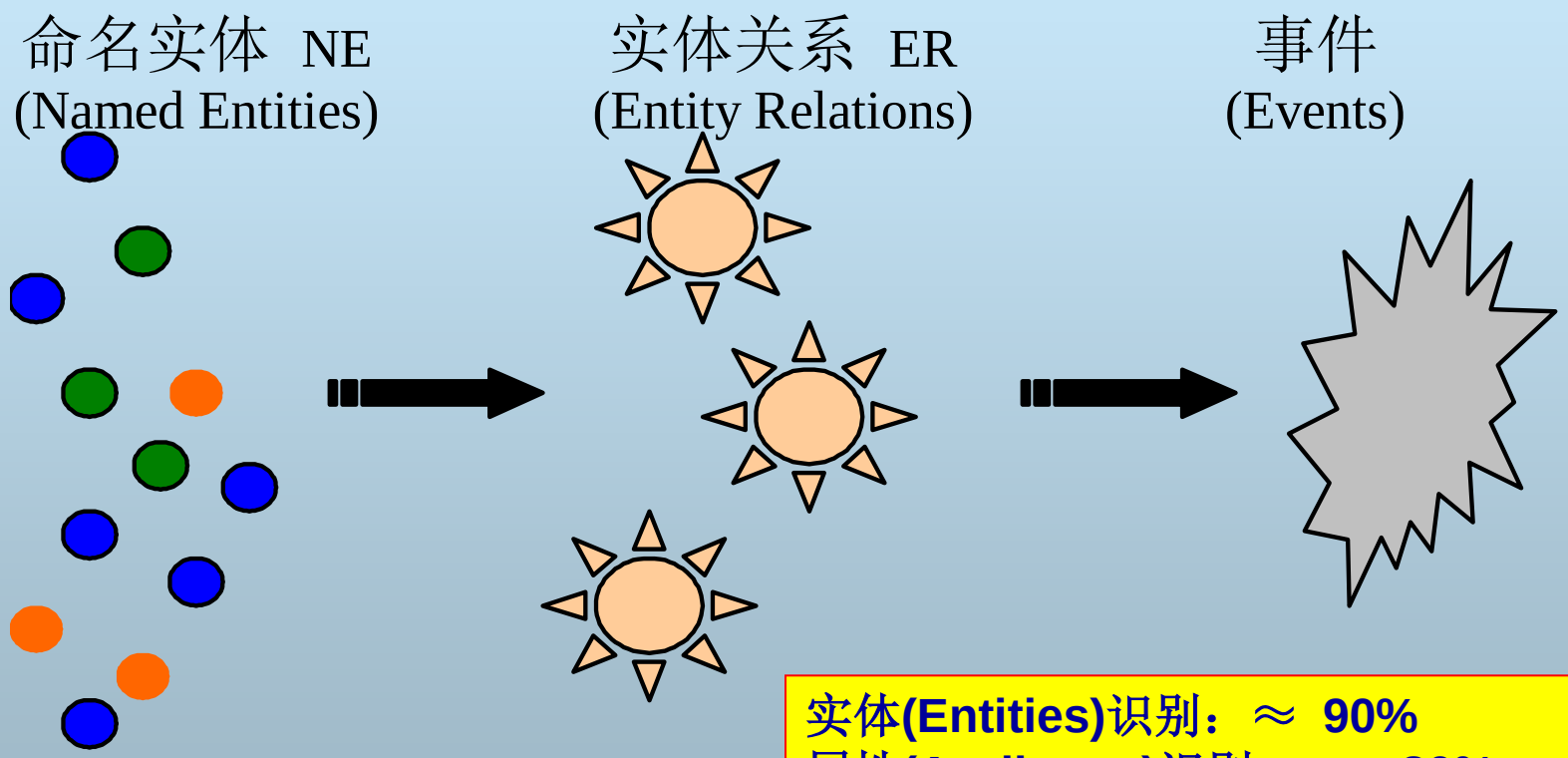


8、信息抽取的内容

- 实体
  - 即命名实体，指文本中的基本构成块，如人、机构等
- 属性
  - 实体的特征，如人的年龄、机构的类型等
- 关系
  - 实体之间存在的联系，也称事实（fact），如公司和地址之间的位置关系、公司与人之间的雇佣关系
- 事件
  - 实体的行为或实体参与的活动，如恐怖袭击（911）、刘翔退赛、公司收购等

9、信息抽取的关键

- 8字方针“抽取实体，确定关系”



实体(Entities)识别： ≈ 90%
属性(Attributes)识别： ≈ 80%
关系(Relations)识别： ≈ 70%
事件(Events)识别： ≈ 60%

二、命名实体识别

- NER
  - 识别出文本中的人名、地名等专有名称和有意义的时间、日期等数量短语并加以归类
  - 信息抽取中的核心任务
- 发展历史
  - 1991，Lisa F. Rau, Extracting Company Names from Text, 7th IEEE Conf. Artificial Intelligence Applications
  - 1996，成为MUC-6的信息抽取评测子任务
  - 后来也成为多个会议的评测任务
    - IEER'99、CoNLL'02-03、LREC等

1、NER的抽取内容

- 一般按照MUC-7的定义（3大类7小类）
  - 实体类
    - 人名、地名、机构名
  - 时间类
    - 日期、时间
  - 数值类
    - 货币、百分比

ACE (Automatic Content Extraction)定义中的NER任务：
人名（Person）、机构名（Organization）、地名（Location）、设备名（Facility）、武器名（Weapon）、交通工具名（Vehicle）和地理政治实体（Geo-Political Entity）

- 哪些不是命名实体？
  - 人造物：如Wall Street Journal、MTV
  - 重复指代的普通名词：如飞机、公司等
  - 人的团体名称以及以人命名的法律、奖项等：如共和国、诺贝尔奖等
  - 从名词派生出来的形容词：如Chinese、American等
  - 非时间、日期、货币、百分比的数字

2、NER的难点

- 命名实体类型多样
  - e.g. John Smith, Mr Smith, John.
- 不断有新的命名实体涌现
  - 如新的人名、地名等，难以建立大而全的姓氏库、名字库、地址库等数据库
- 命名实体的歧义
  - John Smith (company vs. person)
  - May (person vs. month)
  - Washington (person vs. location)
  - 1945 (date vs. time)
- 命名实体构成结构复杂
  - 别名、缩略词等问题，没有严格的规律可以遵循；人名中也存在比较长的少数民族人名或翻译过来的外国人名，没有统一的构词规范
  - 如USTC、Univ. Sci. & Techno. China

3、NER的性能评价

- 正确率P
  - Option 1
    - Correct answer / total answer
  - Option 2
    - [Correct + (1/2) partial correct] / total answer
    - E.g., "Sebastian */person* Karpe"
- 召回率R
  - Correct answer / total correct answer
  - [Correct + (1/2) partial correct] / total correct and partial correct answer
- F值
  - 2PR / (P+R)

3、NER的一般方法

- Baseline：list lookup
- 基于规则的方法
- 基于统计的方法
- 混合方法

List Lookup

- 预先构建一个命名实体词典（gazetteer）
- 出现在词典中的词汇即识别为命名实体
- 词典的构建
  - Person/Organizations：可以利用黄页、电话簿等
  - Locations：可利用现有的一些lists
    - US GEOnet Names Server (GNS) data – 3.9 million locations with 5.37 million names
    - UN site: http://unstats.un.org/unsd/citydata
    - World Gazetteer, http://www.world-gazetteer.com

List Lookup

- 优点
  - 方法简单、快速，与具体语境无关，容易部署和更新（只需更新词典）
- 缺点
  - 大部分情况下很难枚举所有的命名实体名
  - 构建和维护词典的代价较大
  - 难以有效处理实体歧义

基于规则的方法

- 采用手工构造规则模板，选用特征包括统计信息、标点符号、关键字、指示词和方向词、位置词（如尾字）、中心词等方法，以模式和字符串相匹配为主要手段
- 多数参加MUC-7（1997）会议评测的系统，都采用了此方法
- 例如

[ORGANIZATION]'s headquarter in [LOCATION]

e.g. We visited Microsoft /org's headquarter in Seattle /loc.

For location extraction:

Capital Word + {City, Forest, Center}

e.g. Salt Lake City

Capital Word + {Street, Boulevard, Avenue, Crescent, Road}

e.g. *Portobello Street*

基于规则的方法

- 优点
  - 当提取的规则能较精确地反映语言现象时，性能较好
- 缺点
  - 规则往往依赖于具体语言、领域和文本风格
    - 例如考虑规则：“A [/LOC] …… 公司” / ORG
  - 代价太大，系统建设周期长、移植性差而且需要建立不同领域知识库

基于统计的方法

- 采用机器学习方法，利用人工标注的语料进行训练后进行命名实体识别
- 目前主流的NER方法
  - CoNLL'03上参与评测的16个系统全部采用了基于统计的方法

基于统计的方法

- 以机构名识别为例，常见的内部特征包括
  - 单词特征、核心词特征、词性特征、语义特征等

标注	类型	示例
F	机构特征词	北京搜狐畅游时代网络技术*有限公司*
R	机构名中的人名	法国*马蒂尼埃*集团
NR	其它人名	*俞昊然*创立了“计蒜客”
S	机构名中的地名	*北京市*文化局相关领导表示
NS	其它地名	在前不久的*中国*游戏行业年会上
O	常见机构名	*中国人民银行*
E	机构名中的其它词	侵犯*腾讯*公司相关游戏著作权一案
L	机构名之间的连接词	中国移动*和*中国联通慢慢掌控了很多版权
P	职位名	友达*董事长*李焜耀
Z	其它词	
……	……	……

基于统计的方法

- 用于机构名识别的上下文特征示例

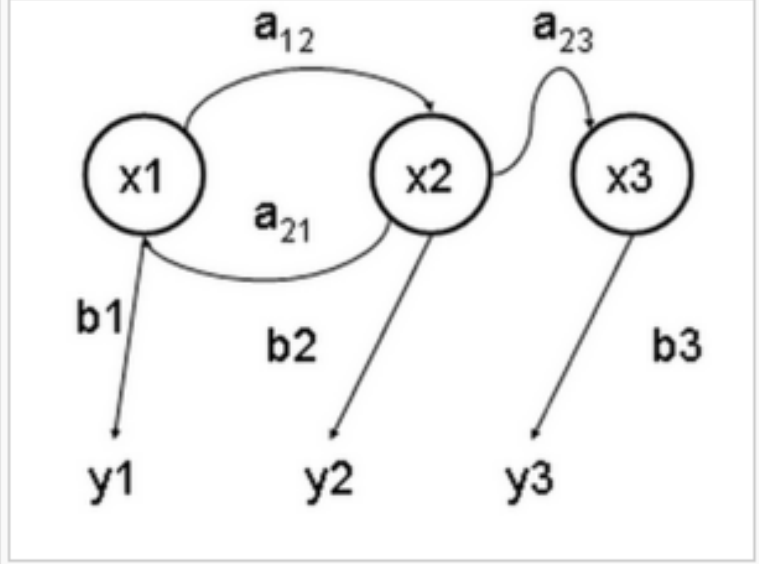
标注	类型	示例
M	修饰词	*国内知名厂商*长虹
C	中心词	华为*市场份额*
W	谓语动词	诺基亚终于*发布*了其第一 款TD产品
N	主谓之间的词	诺基亚*终于*发布了其第一 款TD产品
K	谓宾之间的词	中国联通联合*了*中国电信
J	介词	*在*央视广告招标中
B	机构名上文的前一个词	瑞典*正是*爱立信的总部所在地。
A	机构名下文的后一个词	北京市文化局相关领导*表示*
……	……	……

基于统计的方法

- 常用的方法包括
  - 隐马尔可夫模型(Hidden Markov Model, HMM)
  - 最大熵 (Maximum Entropy, ME)
  - 支持向量机 (Support Vector Machine, SVM)
  - 条件随机场 (Conditional Random Field, CRF)

Hidden Markov Model, HMM

- **HMM是从可观察的输出中确定马尔可夫过程的隐含状态的统计模型**
  - **马尔可夫过程是一个随机过程，但它的未来状态仅依赖于现在状态及所有过去状态**



隐马尔可夫模型状态变迁图（例子）
x — 隐含状态
y — 可观察的输出
a — 转换概率（transition probabilities）
b — 输出概率（output probabilities）

Hidden Markov Model, HMM

- 通过在样本数据集上训练得到HMM模型
  - 基于自定义的特征集
- 一个HMM模型包含两组状态集合和三组概率集合(pi,A,B)：
  - 隐藏状态：一个系统的（真实）状态，可以由一个马尔科夫过程进行描述（例如，天气）。
  - 观测状态：在这个过程中‘可视’的状态（例如，海藻的湿度）。
  - 初始向量Pi：在初始时间t=0时，隐藏状态的初始概率。
  - 转移矩阵A：包含了一个隐藏状态到另一个隐藏状态的概率
  - 混淆矩阵B：包含了给定隐马尔科夫模型的某一个特殊的隐藏状态，观察到的某个观察状态的概率。
- 然后应用到测试数据集中，得到最大转换概率的隐含状态序列

Hidden Markov Model, HMM

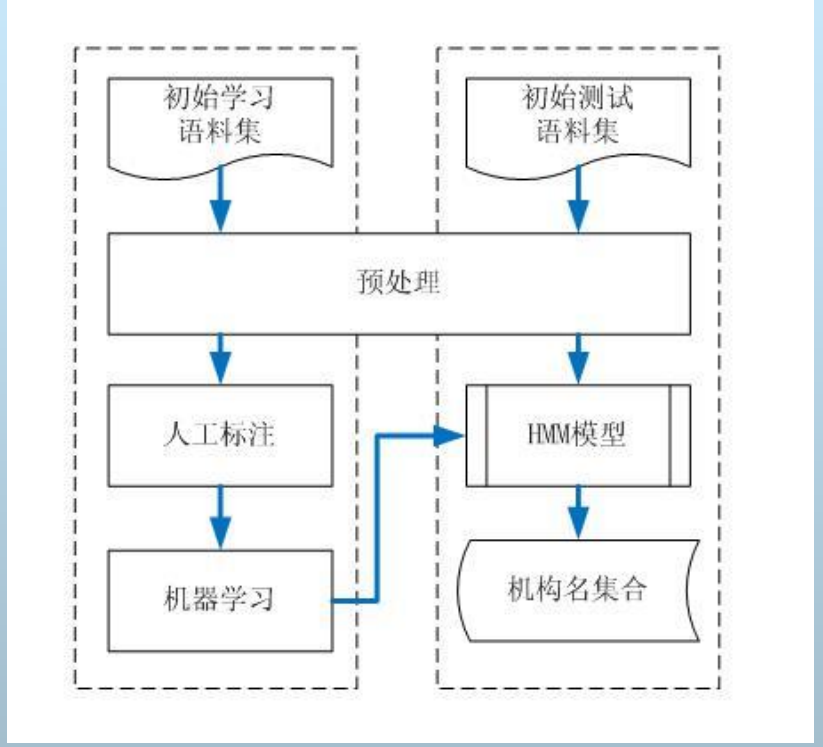
- 隐马尔可夫模型常用来解决三类问题：
  - 评估问题：给定模型，求某个观察值序列的概率。
  - 解码问题：给定模型和观察值序列，求可能性最大的状态序列。
  - 学习问题：给定一个观察值序列，调整模型参数，使观察值序列出现的概率最大。
- 信息检索与信息抽取领域中常用HMM模型来解决解码问题，如词性标注、命名实体识别等。

Hidden Markov Model, HMM

- 基于HMM的机构名识别

词序列 → 观察值序列

标注序列 → 状态序列



词序列：	中国	通信	设备	商	华为	在	北欧	两	次	斩	获	大	单
标注序列：	*M*	*M*	*M*	*M*	*E*	*A*	*Z*	*Z*	*Z*	*Z*	*Z*	*Z*	*Z*

Hidden Markov Model, HMM

- 预处理

人名 ---> 〈PER〉， 地名 ---> 〈LOC〉

常用机构名---> 〈ORG〉， 职位名---> 〈POS〉

人名、地名识别：ICTCLAS
常用机构名：企业黄页词典
职位名： 自定义算法抽取

- 人工标注

示例1	消息/Z 称/B 央/E 视/E 新/C 媒体/C 业务/C 3/A 年/Z 要/Z 占/Z 总收入/Z 25%/Z 以上/Z
示例2	<LOC>/S 搜狐/E 畅游/E 时代/E 网络/E 技术/E 有限公司/F <POS>/C <PER>/C 告诉/A 记者/Z
示例3	<LOC>/M 通信/M 设备/M 商/M 华为/E 将/N 持续/N 推动/W LTE/A 产业/Z 发展/Z
示例4	华为/E 再/N 获/W <LOC>/A 大/Z 单/Z ，/B 在/J 爱立信/E 老巢/C 击败/A 对手/Z

- 模型训练
  - 构造M个隐含状态之间的转移概率矩阵(M*M)，以及隐含状态和观测状态之间的混淆矩阵(M*N)，可使用第三方工具包
- 实体识别
  - 使用Viterbi算法得到最大转移概率的隐含状态序列，提取语料中特定标注（例如E/F/O等）的词序列作为机构名输出

基于统计的方法

- 对特征选取的要求较高，需要从文本中选择对NER有影响的特征来构建特征向量
- 通常做法是对训练语料所包含的语言信息进行统计和分析，从中挖掘出特征
- 对语料的依赖也较大，目前缺少通用的大规模语料
- 大部分需要人工标注

混合方法

- Gazetteer、规则与统计方法的混合
- 实践中往往采用混合的方法
- 例如，网页中的地名抽取
  - 基本地理名称
    - World Gazetteer, http://www.world-gazetteer.com
  - 隐式地名识别
    - 统计模型：CRF
  - 首要地名抽取
    - 规则

4、NER开源工具

- 英文NER
  - Stanford Named Entity Recognizer (NER)
  - http://www-nlp.stanford.edu/software/CRF-NER.shtml
  - CRF-based NER
- 中文NER
  - ICTCLAS http://ictclas.org/
    - HMM-based
  - LTP 哈工大
    - SVM-based

本章小结

- 信息抽取概述
  - 命名实体识别、属性识别、关系抽取、事件抽取
- 命名实体识别
  - 信息抽取中最主要的任务，也是关系抽取和事件抽取的基础
  - 基本方法
    - List lookup
    - 基于规则的方法
    - 基于统计的方法：HMM、CRF、ME、SVM
    - 混合方法

Relations Extraction