文本挖掘技术(2015春)

第十章：

信息抽取

杨建武

北京大学计算机科学技术研究所

Email:yangjw@pku.edu.cn

What is Information Extraction?

- **Information Retrieval**
  - You have an information need, but what you get back isn't *information* but *documents*, which you hope have the information
- **Information extraction**
  - **It is *one* approach to going further for a special case:**
    - There's some relation you're interested in
    - Your query is for elements of that relation
    - A limited form of natural language understanding
- The goal of Information extraction (IE) is transform text into a structured format (e.g. database records) according to its content

Information Extraction of Seminar Announcements

From: teruko+@cs.cmu.edu
To: lti-seminar@cs.cmu.edu, lti-faculty-all@cs.cmu.edu
Subject: LTI seminar, Sept 28 Fri at 2:00pm
Date: Tue, 25 Sep 2001 10:20:14 -0400

Date: Sept 28, Friday
Time: 2:00pm
Place: 3002 NSH
Host: Teruko Mitamura

A New Approach to Automatic Speech Summarization
Chiori Hori
Tokyo Institute of Technology

Abstract: This work is an investigation of an automatic

Information Extraction of Seminar Announcements

TEMPLATE SLOTS	EXTRACTED TEXT
SEMINAR NAME	LTI Seminar
DATE	Sept. 28, Friday, 2001
TIME	2:00pm
LOCATION	3002 NSH
HOST	Teruko Mitamura
TITLE	A New Approach to Automatic Speech Summarization
SPEAKER	Chiori Hori
INSTITUTION	Tokyo Institute of Technology
ABSTRACT	This work is an investigation of automatic speech …

Information Extraction As An Annotation Task

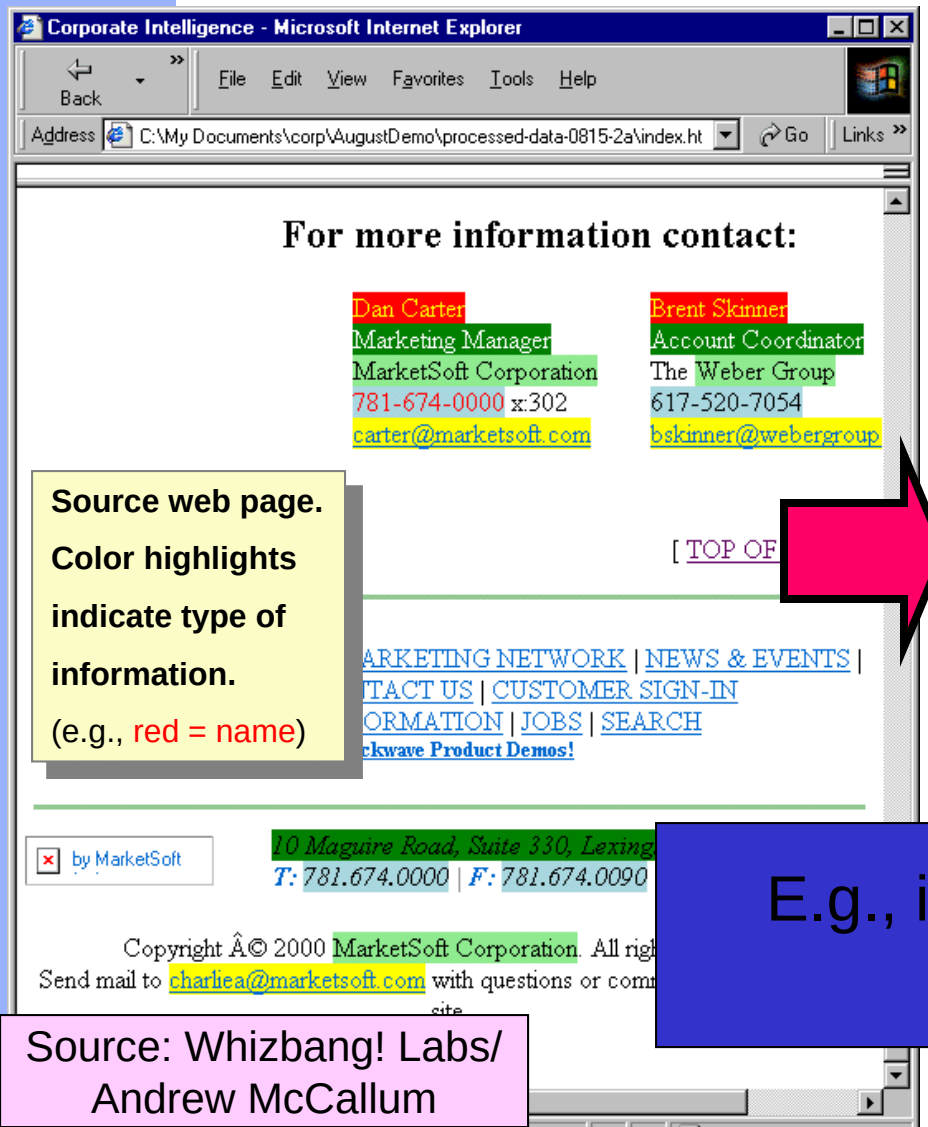
```
From:  teruko+@cs.cmu.edu
To:  lti-seminar@cs.cmu.edu, lti-faculty-all@cs.cmu.edu
Subject:  <SEMINAR NAME> LTI seminar </SEMINAR NAME> ,
          Sept 28 Fri at 2:00pm
Date:  Tue, 25 Sep 2001 10:20:14 -0400

   Date:  <DATE> Sept 28, Friday </DATE>
   Time:  <TIME> 2:00pm </TIME>
   Place:  <LOCATION> 3002 NSH </LOCATION>
   Host:  <HOST> Teruko Mitamura </HOST>

  <TITLE> A New Approach to Automatic Speech Summarization </TITLE>
              <SPEAKER> Chiori Hori </SPEAKER>
        <INSTITUTION> Tokyo Institute of Technology </INSTITUTION>

Abstract:  <ABSTRACT> This work is an investigation of an  ...
```

Extracting Corporate Information



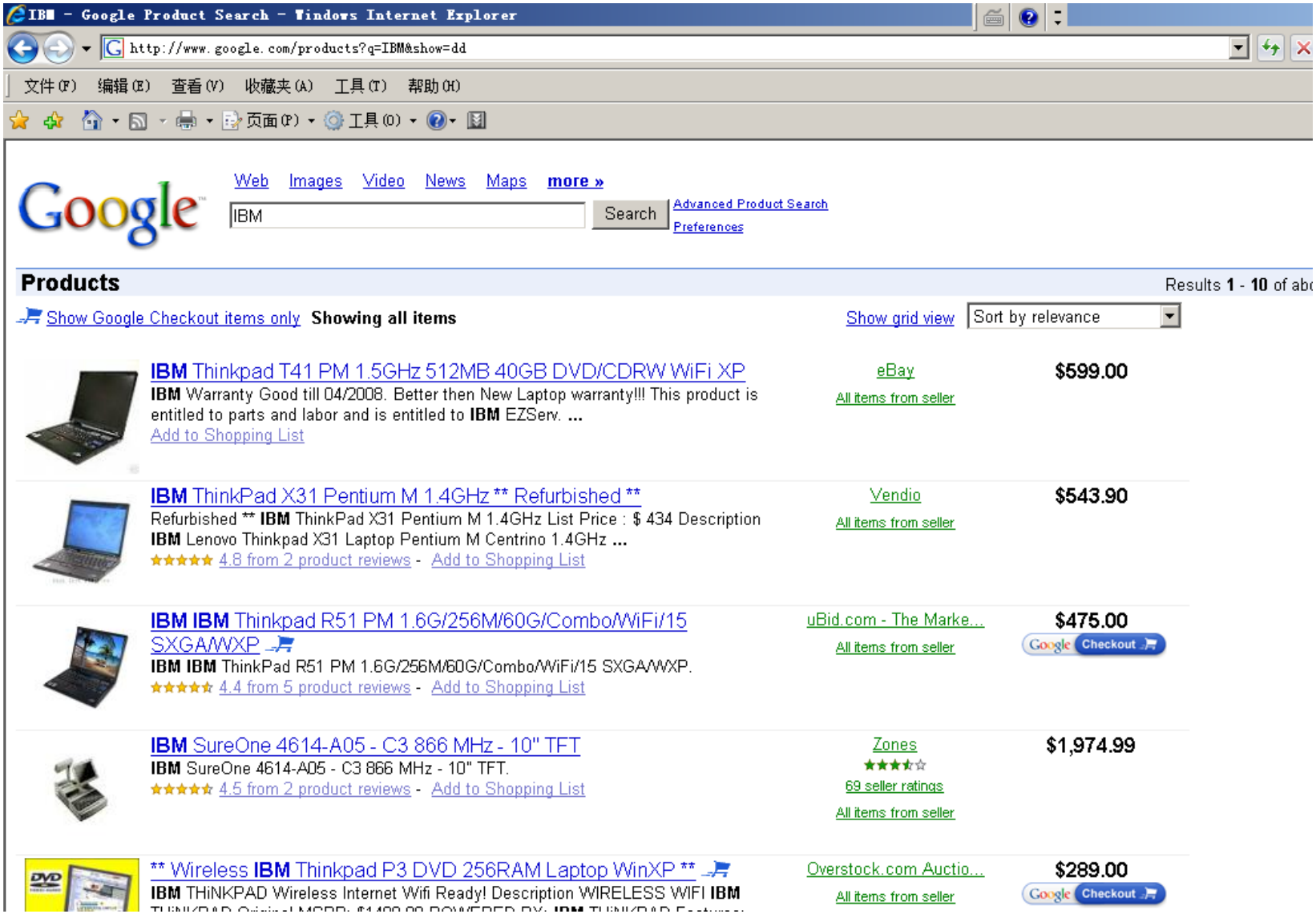
Source web page. Color highlights indicate type of information. (e.g., red = name)

Data automatically extracted from marketsoft.com

E.g., information need: Who is the CEO of MarketSoft?

Source: Whizbang! Labs/ Andrew McCallum

Product information



Landscape of IE Tasks

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity

Person: **Jack Welch**

Person: **Jeffrey Immelt**

Location: **Connecticut**

Binary relationship

Relation: **Person-Title**
Person: **Jack Welch**
Title: **CEO**

Relation: **Company-Location**
Company: **General Electric**
Location: **Connecticut**

N-ary record

Relation: **Succession**
Company: **General Electric**
Title: **CEO**
Out: **Jack Welsh**
In: **Jeffrey Immelt**

难 点

- Textual inconsistency

例：digital cameras

  - Image Capture Device: 1.68 million pixel 1/2-inch CCD sensor
  - Image Capture Device Total Pixels Approx. 3.34 million Effective Pixels Approx. 3.24 million
  - Image sensor Total Pixels: Approx. 2.11 million-pixel
  - Imaging sensor Total Pixels: Approx. 2.11 million 1,688 (H) x 1,248 (V)
  - CCD Total Pixels: Approx. 3,340,000 (2,140[H] x 1,560 [V])
  - Effective Pixels: Approx. 3,240,000 (2,088 [H] x 1,550 [V])
  - Recording Pixels: Approx. 3,145,000 (2,048 [H] x 1,536 [V])

- *These all came off the* same manufacturer's *website!!*
- And this is a very technical domain.

评 价

- Template Measure for each test document:
  - Total number of correct extractions in the solution template: *N*
  - Total number of slot/value pairs extracted by the system: *E*
  - Number of extracted slot/value pairs that are correct (i.e. in the solution template): *C*
- Compute average value of metrics adapted from IR:
  - Recall = *C*/*N*
  - Precision = *C*/*E*
  - F-Measure = Harmonic mean of recall and precision

Three generations of IE systems

- Hand-Built Systems–Knowledge Engineering [1980s–]
  - Rules written by hand
  - Require experts who understand both the systems and the domain
  - Iterative guess-test-tweak-repeat cycle
- Automatic, Trainable Rule-Extraction Systems[1990s–]
  - Rules discovered automatically using predefined templates
  - Require huge, labeled corpora (effort is just moved!)
- Machine Learning (Sequence) Models [1997 –]
  - One decodes a statistical model that classifies the words of the text, using HMMs, random fields or statistical parsers
  - Learning usually supervised; may be partially unsupervised

包装器Wrappers

“Wrappers”

- If we think of things from the database point of view
  - We want to be able to database-style queries
  - But we have data in some horrid textual form/content management system that doesn’t allow such querying
  - We need to “wrap” the data in a component that understands database-style querying
- Many people have “wrapped” many web sites
  - Commonly something like a Perl script
  - Often easy to do as a one-off
- But handcoding wrappers in Perl isn’t very viable
  - Sites are numerous, and their surface structure mutates (变异) rapidly (around 10% failures each month)

Amazon Book Description

….
The Age of Spiritual Machines : When Computers Exceed Human Intelligence

by <a href="/exec/obidos/search-handle-url/index=books&field-author=
Kurzweil%2C%20Ray/002-6235079-4593641">Ray Kurzweil

<img src="http://images.amazon.com/images/P/0140282025..gif" width=90 height=140 align=left border=0>

List Price: $14.95
Our Price: $11.96
You Save: $2.99
(20%)

<p>
…

Extracted Book Template

Title: The Age of Spiritual Machines :
When Computers Exceed Human Intelligence
Author: Ray Kurzweil
List-Price: $14.95
Price: $11.96
:
:

Template Types

- Slots in template typically filled by a substring from the document.
- Some slots may have a fixed set of pre-specified possible fillers (可能的填充值) that may not occur in the text itself.
  - Terrorist act: threatened, attempted, accomplished.
  - Job type: clerical, service, custodial, etc.
  - Company type: SEC code
- Some slots may allow multiple fillers.
  - Programming language
- Some domains may allow multiple extracted templates per document.
  - Multiple apartment listings in one ad

Wrappers:Simple Extraction Patterns

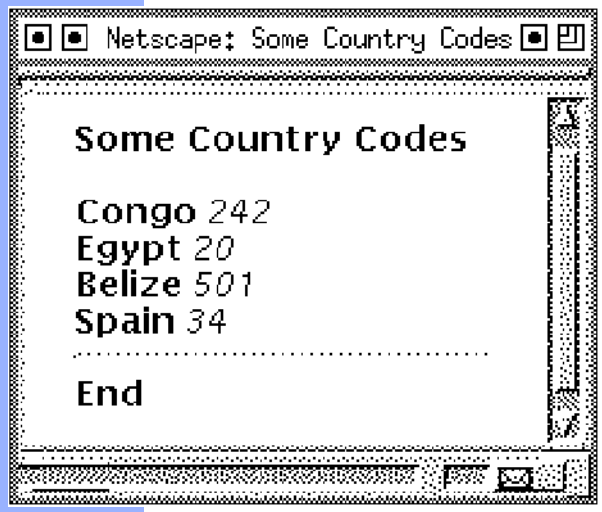
- Specify an item to extract for a slot using a regular expression pattern.
  - Price pattern: "\b\$\d+(\.\d{2})?\b"
- May require preceding (pre-filler) pattern to identify proper context.
  - Amazon list price:
    - Pre-filler pattern: "List Price: "
    - Filler pattern: "\$\d+(\.\d{2})?\b"
- May require succeeding (post-filler) pattern to identify the end of the filler.
  - Amazon list price:
    - Pre-filler pattern: "List Price: "
    - Filler pattern: ".+"
    - Post-filler pattern: ""

Simple Template Extraction

- Extract slots in order, starting the search for the filler of the *n*+1 slot where the filler for the *n*th slot ended. Assumes slots always in a fixed order.
  - Title
  - Author
  - List price
  - …
- Make patterns specific enough to identify each filler always starting from the beginning of the document.

Wrapper induction

- Delimiter (分隔符)-based extraction



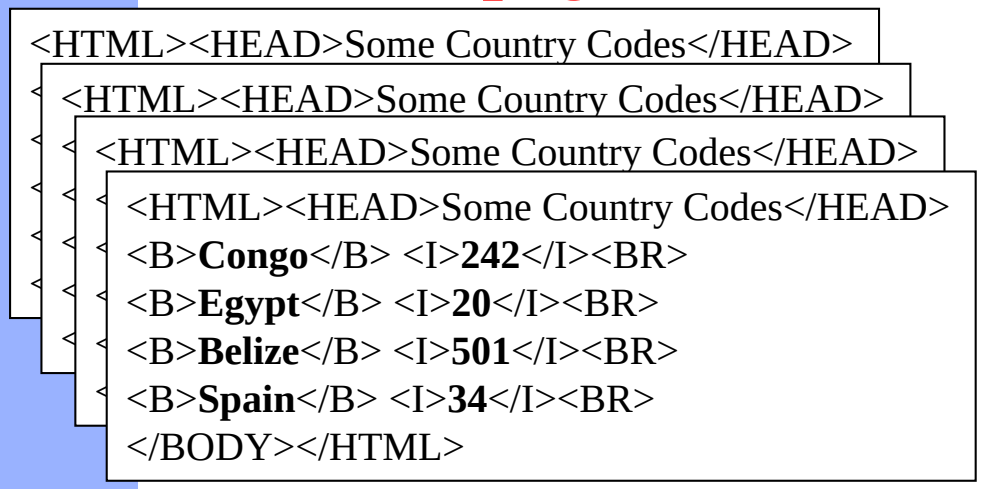
<HTML><TITLE>Some Country Codes</TITLE>
<B>Congo</B> <I>242</I>
<B>Egypt</B> <I>20</I>
<B>Belize</B> <I>501</I>
<B>Spain</B> <I>34</I>
</BODY></HTML>

Use **<B>**, **</B>**, **<I>**, **</I>** for extraction

Wrapper induction

- Learning LR wrappers

labeled pages



wrapper

$\langle l_1, r_1, \ldots, l_K, r_K \rangle$

Example: Find **4** strings

⟨<B>, </B>, <I>, </I>⟩

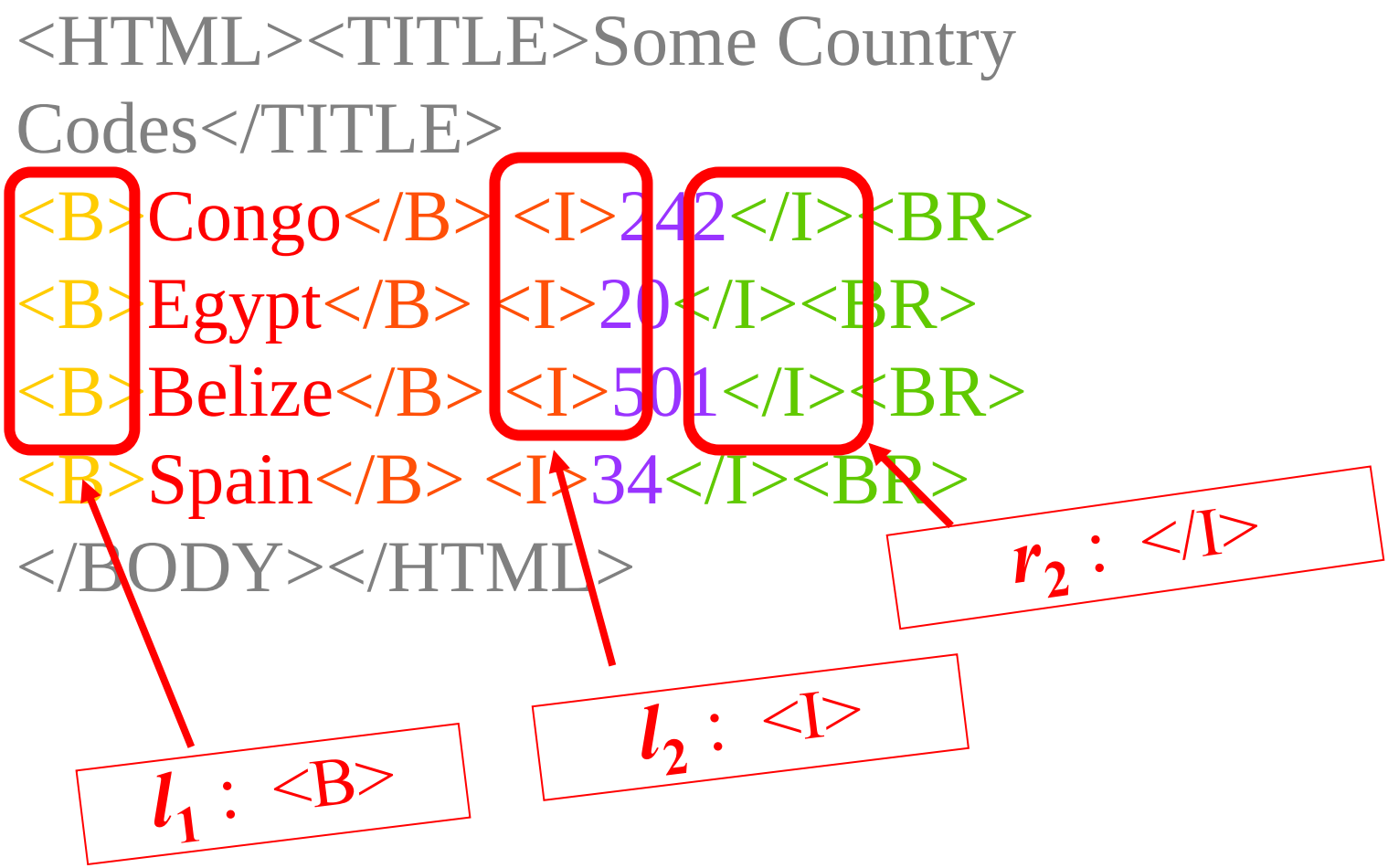
$\langle l_1, r_1, l_2, r_2 \rangle$

LR: Finding r_1

<HTML><TITLE>Some Country Codes</TITLE>
<B>Congo</B> <I>242</I>
<B>Egypt</B> <I>20</I>
<B>Belize</B> <I>501</I>
<B>Spain</B> <I>34</I>
</BODY></HTML>

r_1 : </B>

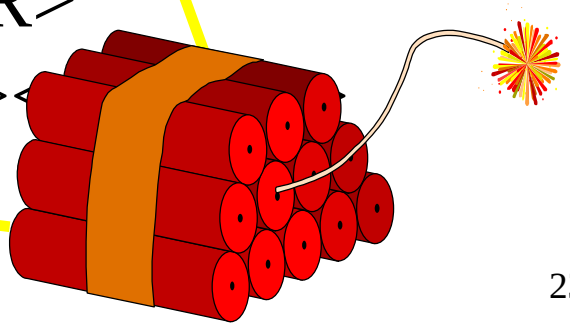
LR: Finding l_1, l_2 and r_2



A problem with LR wrappers

Distracting text in head and tail

```
<HTML><TITLE>Some Country
  Codes</TITLE>
   <BODY><B>Some Country Codes</B><P>
   <B>Congo</B> <I>242</I><BR>
   <B>Egypt</B> <I>20</I><BR>
   <B>Belize</B> <I>501</I><BR>
   <B>Spain</B> <I>34</I><BR>
   <HR><B>End</B></BODY><
```



One (of many) solutions: HLRT

*Ignore page's **head** and **tail***

<HTML><TITLE>Some Country Codes</TITLE>
<BODY><B>Some Country Codes</B><P>
<B>Congo</B> <I>242</I>
<B>Egypt</B> <I>20</I>
<B>Belize</B> <I>501</I>
<B>Spain</B> <I>34</I>
<HR><B>End</B></BODY></HTML>

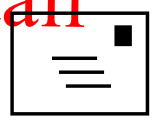
end of head

head

body

tail

start of tail



***H*ead-*L*eft-*R*ight-*T*ail wrappers**

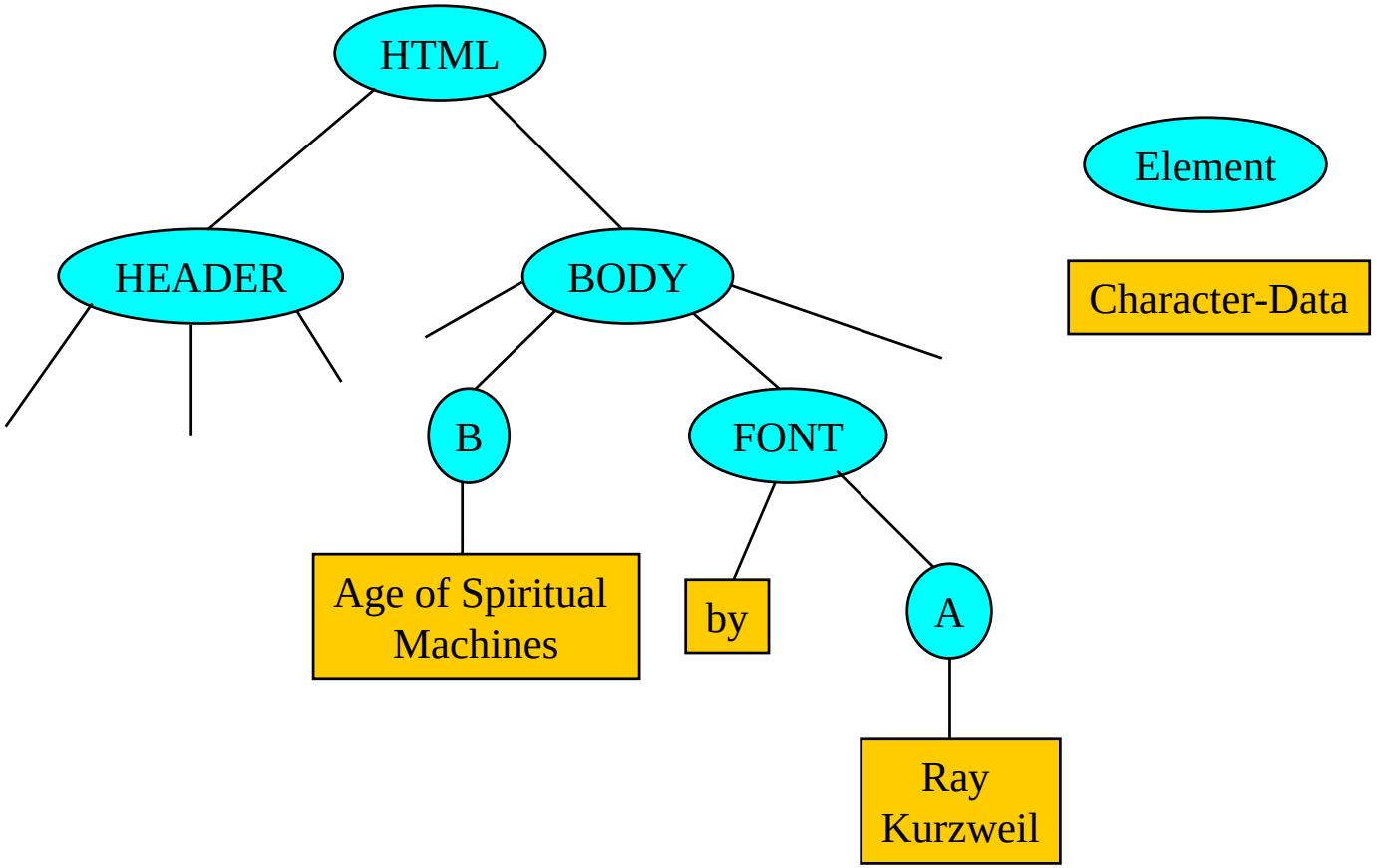
More sophisticated wrappers

- LR and HLRT wrappers are extremely simple
  - Though applicable to many tabular patterns
- more expressive wrapper classes:
  - Disjunctive delimiters
  - Multiple attribute orderings
  - Missing attributes
  - Multiple-valued attributes
  - Hierarchically nested data
  - Wrapper verification and maintenance

Web Extraction using DOM Trees

- Web extraction may be aided by first parsing web pages into DOM trees.
- Extraction patterns can then be specified as paths from the root of the DOM tree to the node containing the text to extract.
- May still need regex patterns to identify proper portion of the final CharacterData node.

Sample DOM Tree Extraction



Title: HTML→BODY→B→CharacterData

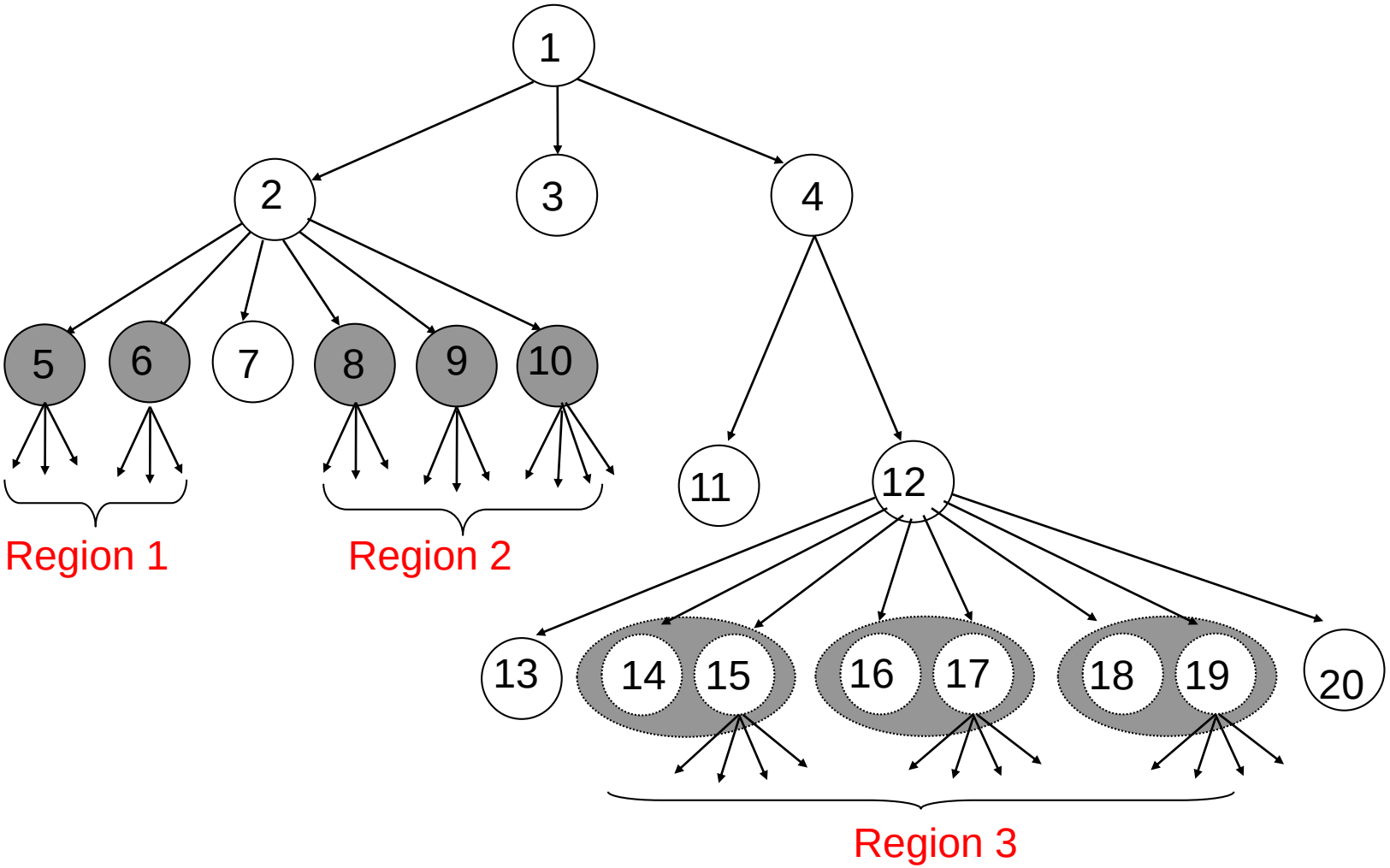
Author: HTML→ BODY→FONT→A→ CharacterData

Web Extraction using DOM Trees

Given a page, three steps:

- Building the HTML Tag Tree
  - Erroneous tags, unbalanced tags, etc
  - Some problems are hard to fix
- Mining Data Regions
  - String matching or tree matching
- Identifying Data Records
- Extract Data from Data Records

Mining Data Regions



Identify Data Records

- A generalized node may not be a data record.
- Extra mechanisms are needed to identify true atomic objects.
- Some highlights:
  - Contiguous
  - non-contiguous data records.

Name 1 Description of object 1	Name 2 Description of object 2
Name 3 Description of object 3	Name 4 Description of object 4

Name 1	Name 2
Description of object 1	Description of object 2
Name 3	Name 4
Description of object 3	Description of object 4

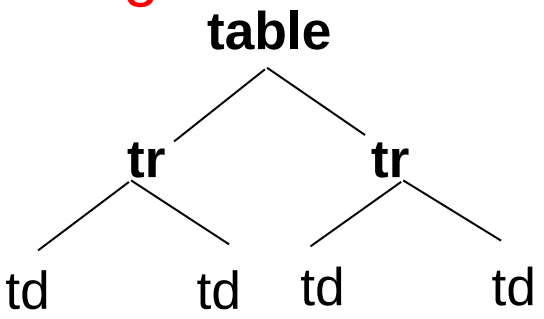
Extract Data from Data Records

- Once a list of data records are identified, we can align and extract data items in them.
- Approaches (align multiple data records):
  - Multiple string alignment
    - Many ambiguities due to pervasive use of table related tags.
  - Multiple tree alignment (partial tree alignment)
    - Together with visual information is effective
- Most multiple alignment methods work like hierarchical clustering,
  - Not effective, and very expensive

Building tree based on visual cues

Rendering 呈现 (or visual 视觉) information is very useful in the whole process

The tag tree




```
1  <table>
2        <tr>
3              <td> … </td>
4              <td> … </td>
5        </tr>
6        <tr>
7              <td> … </td>
8              <td> … </td>
9        </tr>
10</table>
```

left	right	top	bottom
100	300	200	400
100	300	200	300
100	200	200	300
200	300	200	300
100	300	300	400
100	200	300	400
200	300	300	400

Natural Language Processing-based Information Extraction

Natural Language Processing-based Information Extraction

- Wrapper induction is only ideal for rigidly-structured machine-generated HTML…
- … or is it?!
- Can we use simple patterns to extract from natural language documents?
  - http://www.smi.ucd.ie/bwi/

… **Name:** Dr. Jeffrey D. Hermes …
… **Who:** Professor Manfred Paul …
... will be given by Dr. R. J. Pangborn …

… Ms. Scott **will be speaking …**
… Karen Shriver**, Dept. of ...**
… Maria Klawe**, University of ...**

Natural Language Processing-based Information Extraction

- If extracting from automatically generated web pages, simple regex (正则表达式) patterns usually work.
- If extracting from more natural, unstructured, human-written text, some NLP may help.
  - Part-of-speech (POS) tagging （词性）
    - Mark each word as a noun, verb, preposition, etc.
  - Syntactic parsing （句法分析）
    - Identify phrases: NP, VP, PP
  - Semantic word categories (e.g. from WordNet)
    - KILL: kill, murder, assassinate, strangle, suffocate
- Extraction patterns can use POS or phrase tags.
  - Crime victim:
    - Prefiller: [POS: V, Hypernym: KILL]
    - Filler: [Phrase: NP]

MUC: the NLP genesis of IE

- DARPA funded significant efforts in IE in the early to mid 1990's.
- Message Understanding Conference (MUC) was an annual event/competition where results were presented.
- http://www-nlpir.nist.gov/related_projects/muc/
- Focused on extracting information from news articles:
  - Terrorist events
  - Industrial joint ventures
  - Company management changes
- Information extraction is of particular interest to the intelligence community.

Rule-based Extraction Examples

Determining which person holds what office in what organization

- [person] , [office] *of* [org]
  - Vuk Draskovic, leader of the Serbian Renewal Movement
- [org] (named, appointed, etc.) [person] P [office]
  - NATO appointed Wesley Clark as Commander in Chief

Determining where an organization is located

- [org] *in* [loc]
  - NATO headquarters in Brussels
- [org] [loc] (*division*, *branch*, *headquarters*, etc.)
  - KFOR Kosovo headquarters

Example of IE from FASTUS (1993)

Bridgestone Sports Co. said Friday it had set up a joint venture(合资) in Taiwan with a local concern and a Japanese trading house(商行) to produce golf clubs (高尔夫球棍) to be supplied to Japan.
The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP (合伙)
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Example of IE from FASTUS (1993)

Bridgestone Sports Co. said Friday it had set up a joint venture（合资）in Taiwan with a local concern and a Japanese trading house（商行）to produce golf clubs（高尔夫球棍）to be supplied to Japan.
The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

FASTUS

Based on finite state automata (FSA) transductions

set up
new Taiwan dollars

a Japanese trading house
had set up

production of
20, 000 iron and
metal wood clubs

[company]
[set up]
[Joint-Venture]
with
[company]

1.Complex Words:

Recognition of multi-words and proper names

2.Basic Phrases:

Simple noun groups, verb groups and particles

3.Complex phrases:

Complex noun groups and verb groups

4.Domain Events:

Patterns for events of interest to the application

Basic templates are to be built.

5. Merging Structures:

Templates from different parts of the texts are merged if they provide information about the same entity or event.

有限状态机方法

命名实体的识别

- **Named Entity Identification**
- **目的（回答下面这样的问题）：**
  - 在这100篇文章中提到了哪些人？
  - 在这2000篇网页中提到了哪些地点？
  - 在这些专利申请表中提到了哪些公司？
  - 今年的消费者报告评估了什么产品？
- 注意：
  - 并不是给定X，问哪些文档含有X。
  - 需要有一定的语法分析能力（词汇表+有限状态机）。

命名实体的识别

Example

President Clinton decided to send special trade envoy Mickey Kantor to the special Asian economic meeting in Singapore this week. Ms. Xuemei Peng, trade minister from China, and Mr. Hideto Suzuki from Japan's Ministry of Trade and Industry will also attend. Singapore, who is hosting the meeting, will probably be represented by its foreign and economic ministers. The Australian representative, Mr. Langford, will not attend, though no reason has been given. The parties hope to reach a framework for currency stabilization.

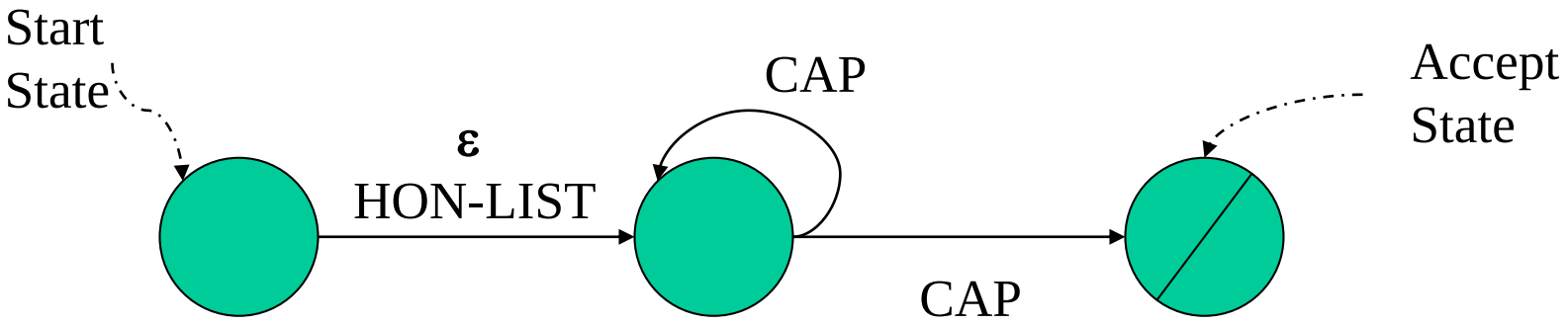
命名实体的识别

Extracted Named Entities (NEs)

PEOPLE	PLACES
President Clinton	
Mickey Kantor	Singapore
Ms. Xuemei Peng	China
Mr. Hideto Suzuki	Japan
Mr. Langford	Australia

命名实体的识别–有限状态机

- 有限状态接收器Finite State Acceptor (FSA)的定义
  - FSA是一个有向图
  - 它有一个起始节点，"start" node
  - 它至少有一个接收节点，"accepting" nodes
  - 有一个输入源（例如，string of words）
  - 在节点上可能输出"YES" or "NO"



  - *CAP matches any capitalized word*
  - *HON-LIST :=* 称呼*(Mr, Ms, Dr, President, ...)*

命名实体的识别–有限状态机

- 节点之间的链接标记和输入项的匹配
  - 精确匹配，exact-match links labels

    e.g. "China" matching only "China"
  - 通配符（?）匹配

    e.g. "?" matches "100" or "China" or ...
  - 特征匹配（feature-match）

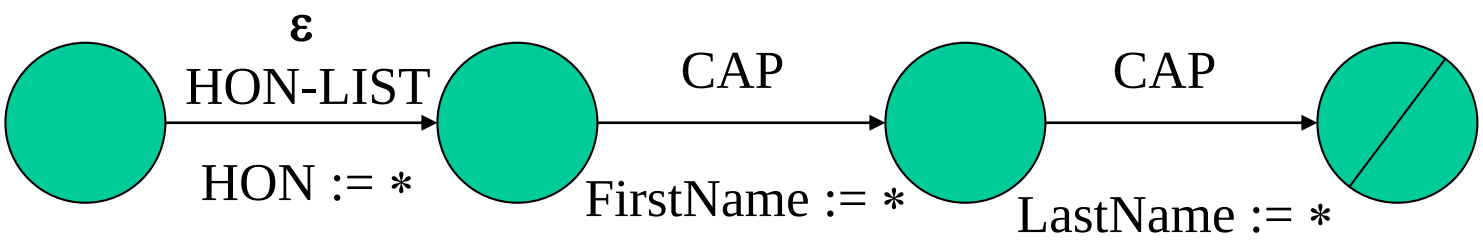
    e.g. CAP matches any capitalized word
  - 表成员匹配（list-membership，例如称呼）

    e.g. if HON-LIST := (Mr, Ms, Dr, President, ...)

    it would match any of those words in the input

命名实体的识别–有限状态机

- **有限状态变换器，A Finite State Transducer (FST)**
  - 带有变量绑定的FSA
  - 在输出“NO”或“YES”的同时给出特定变量的绑定，从而可以给出对具体实体的识别
  - *e.g. "YES <firstname Hideto> <lastname Suzuki>"*



  - *CAP matches any capitalized word*
  - *HON-LIST :=* 称呼*(Mr, Ms, Dr, President, ...)*

带有角色信息的命名实体

Motivation

- 知道命名实体的角色常常是有用的，例如：
  - 谁参加了经济会议？
  - 谁主持了这个会议？
  - 在这经济会上讨论了谁的情况？
  - 这次经济会议谁缺席了？

带有角色信息的命名实体

如何确定实体的角色？

- 一个FSM不够了，考虑用三个FSMs
  - <left-context-FSA><entity-FSM><right-context-FSA>
  - 其中左边和右边的上下文帮助确定中间实体的角色

例子（根据左右内容的含义）

```
If <right-context> =
        <? "not" ("attend" | "participate")>
   Then entity.role = ABSENT
If <left-context> =
        <("meet" | "meeting") ("in" | "at")>
   Then entity.role = HOST
```

关系信息的提取

- **目的：**想知道谁对谁做了什么。
- **Example**

  "John Snell reporting for Wall Street. Today Flexicon Inc. announced a tender offer for Supplyhouse Ltd. for $30 per share, representing a 30% premium over Friday's closing price. Flexicon expects to acquire Supplyhouse by Q4 2001 without problems from federal regulators"

关系信息提取

提取系统可以看成是若干FSMs构成的一个模板，其设计根据具体应用确定

[Corporate-acquisition（公司收购）

[acquirer <company-FSM> <r-acquirer-FSM>]

[acquiree <l-acquiree-FSM> <company-FSM)]

[share-price <money-FSM> <r-stock-FSM>]

[date <l-event-date-FSM> <date-FSM>]

]

关系信息提取

输出就是FSM的实例化

[Corporate-acquisition
    [acquirer "Flexicon Inc."]
    [acquiree "Supplyhouse Ltd."]
    [share-price "30 USD"]
    [date "Q4 2001"]
]

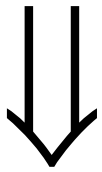
机器学习方法

Popular Machine Learning Methods

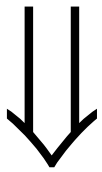
- Naive Bayes
- SRV [Freitag 1998], Inductive Logic Programming
- Rapier [Califf and Mooney 1997]
- Hidden Markov Models [Leek 1997]
- Maximum Entropy Markov Models [McCallum et al. 2000]
- Conditional Random Fields [Lafferty et al. 2001]

Learning for IE

Highly regular source documents



Relatively simple extraction patterns



Efficient learning algorithm

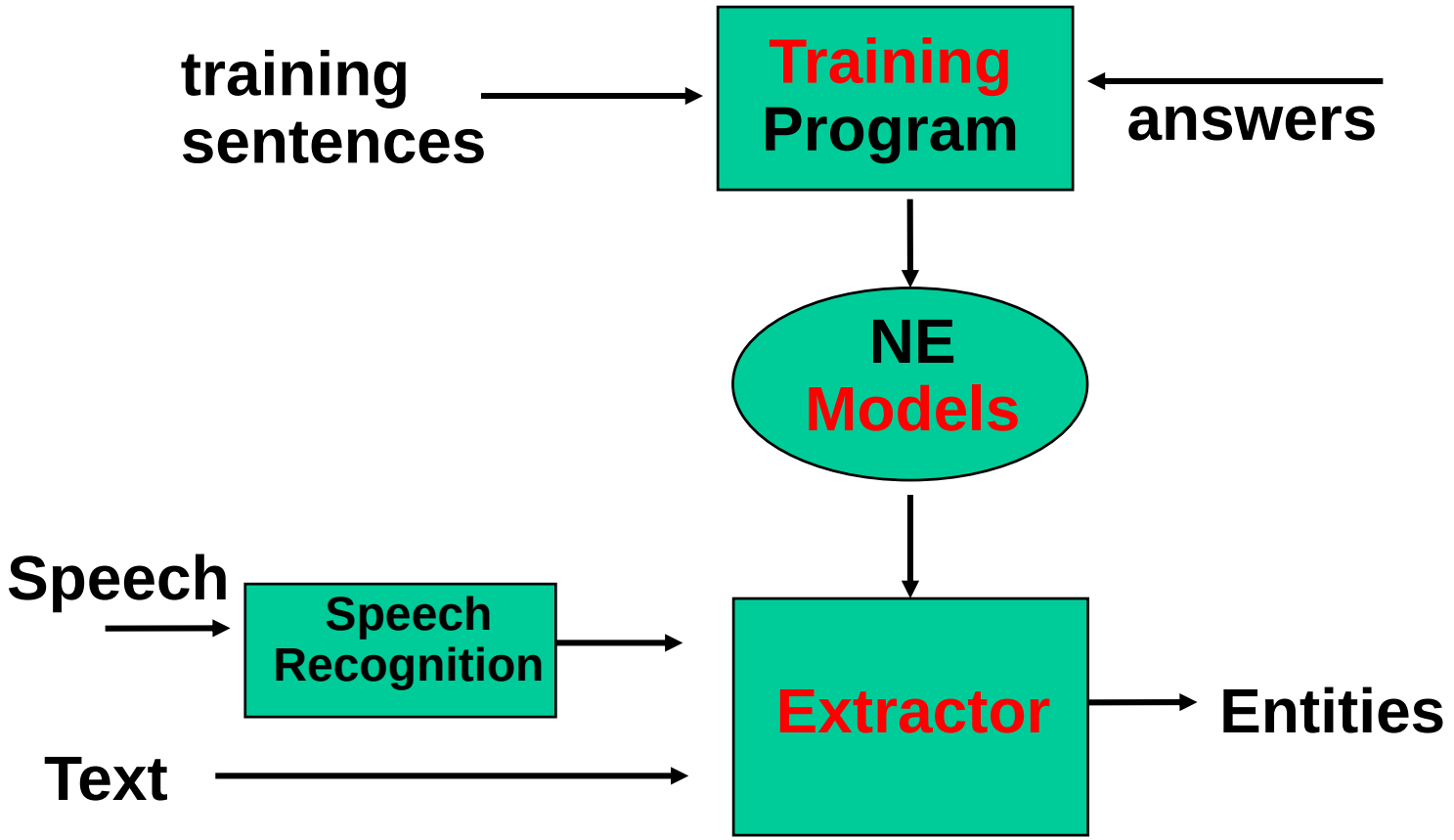
- Writing accurate patterns for each slot for each domain (e.g. each web site) requires laborious software engineering.
- Alternative is to use machine learning:
  - Build a training set of documents paired with human-produced filled extraction templates.
  - Learn extraction patterns for each slot using an appropriate machine learning algorithm.

Hidden Markov Models (HMMs)

Statistical generative models

- Sequence Models are statistical models of whole token sequences that effectively label subsequences
  - Best known case is generative Hidden Markov Models (HMMs)
- Pros（优）:
  - Well-understood underlying statistical models make it easy to used wide range of tools from statistical decision theory
  - Portable, broad coverage, robust, good recall
- Cons（劣）:
  - Range of features and patterns usable may be limited
  - Not necessarily as good for complex multi-slot patterns

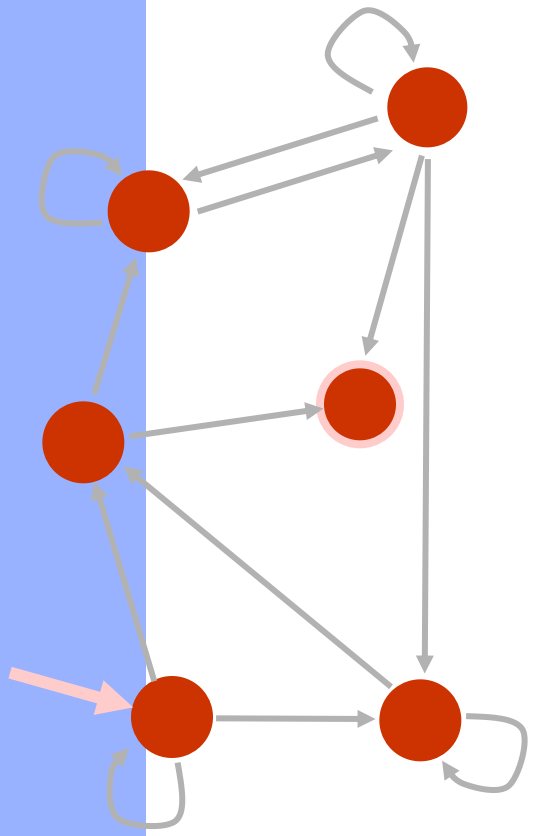
Name Extraction via HMMs



Applying HMMs to IE

- **Document** ⇒ generated by a stochastic process modelled by an HMM
- **Token** ⇒ word
- **State** ⇒ "reason/explanation" for a given token
  - '*Background*' state emits tokens like '*the*', '*said*', …
  - '*Money*' state emits tokens like '*million*', '*euro*', …
  - '*Organization*' state emits tokens like '*university*', '*company*', …
- **Extraction**: via the Viterbi algorithm, a dynamic programming technique for efficiently computing the most likely sequence of states that generated a document.

HMM formalism

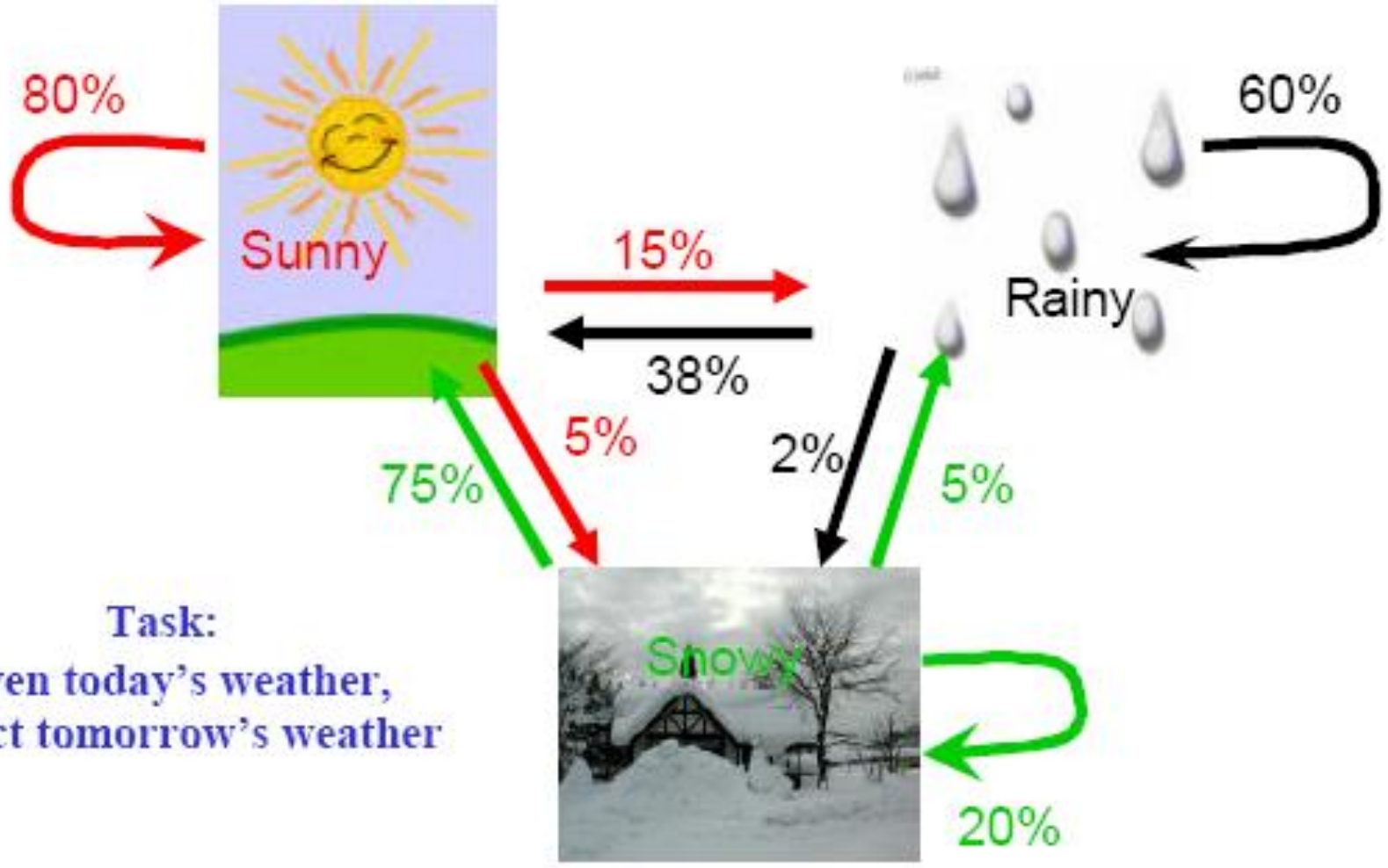


HMM = probabilistic FSA

HMM = states $s_1, s_2, \ldots$
(special start state s_1
special end state s_n)
token alphabet $a_1, a_2, \ldots$
state transition probs $P(s_i|s_j)$
token emission probs $P(a_i|s_j)$

Widely used in many language processing tasks, *e.g.*, speech recognition [Lee, 1989], POS tagging [Kupiec, 1992], topic detection [Yamron *et al*, 1998].

A Markov Model : Weather



Task:
Given today's weather, predict tomorrow's weather

A Markov Model : Weather

MM = (S, A, ?)

S is 3 states

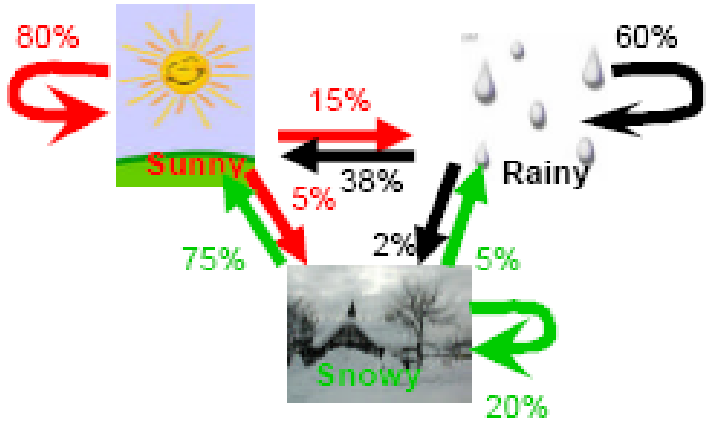
$\{S_{sunny}, S_{rainy}, S_{snowy}\}$

Initial Distribution

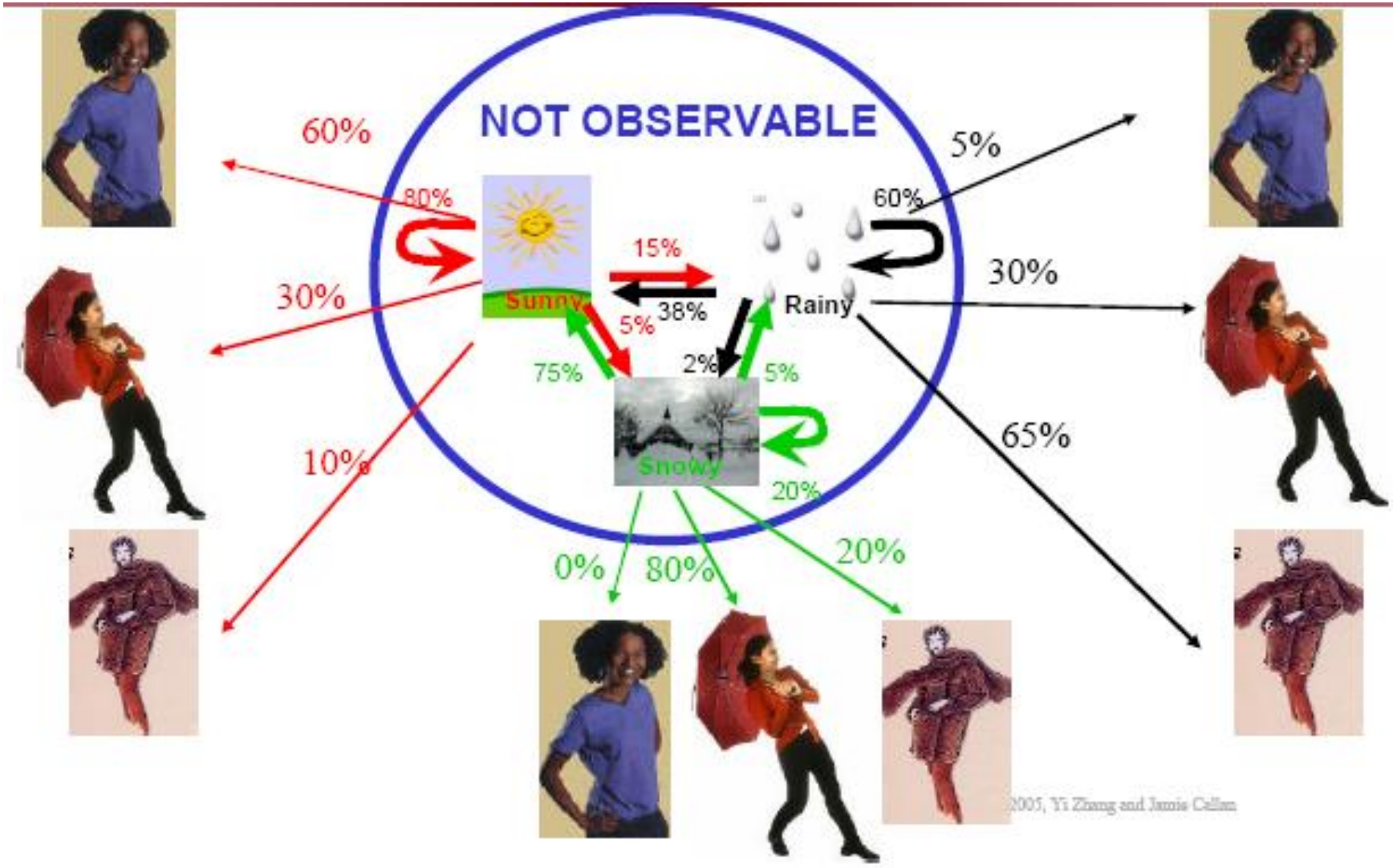
? = (.70, .25, .05)

Transition Probabilities

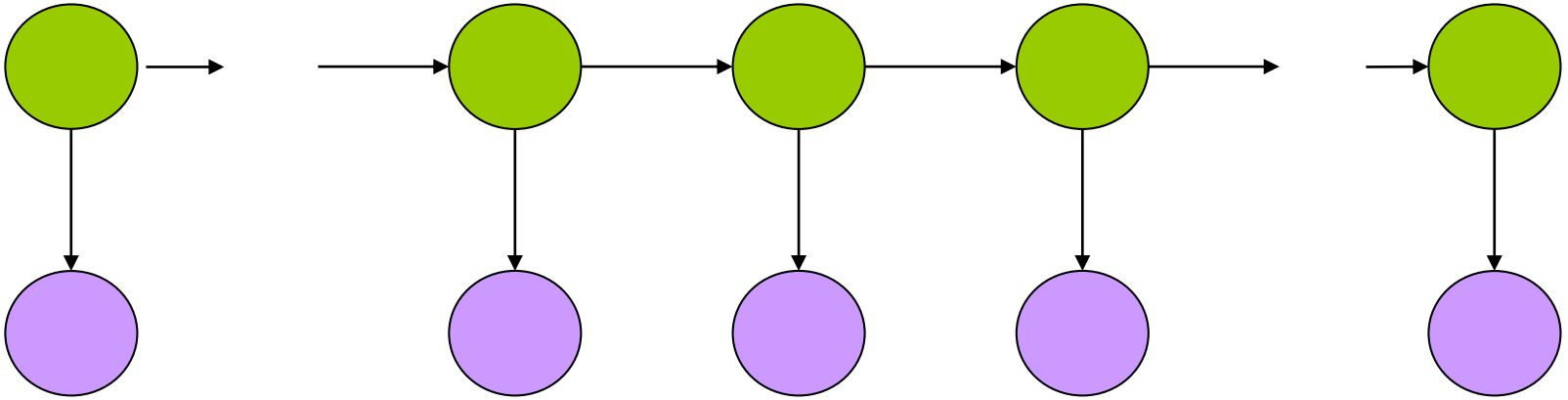
	Sunny	Rainy	Snowy
sunny	.8	.15	.05
$A =$ rainy	.38	.6	.02
snowy	.75	.05	.2



Hidden Markov Models:
Inferring The Weather From What People Wear

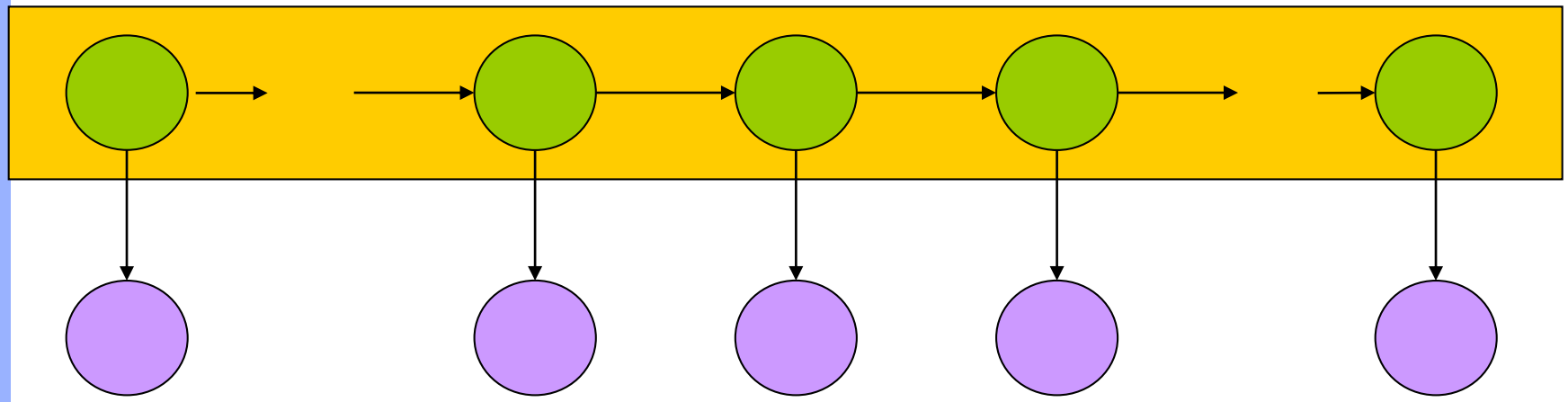


What is an HMM?



- Graphical Model Representation: Variables by time
- Circles indicate states
- Arrows indicate probabilistic dependencies between states

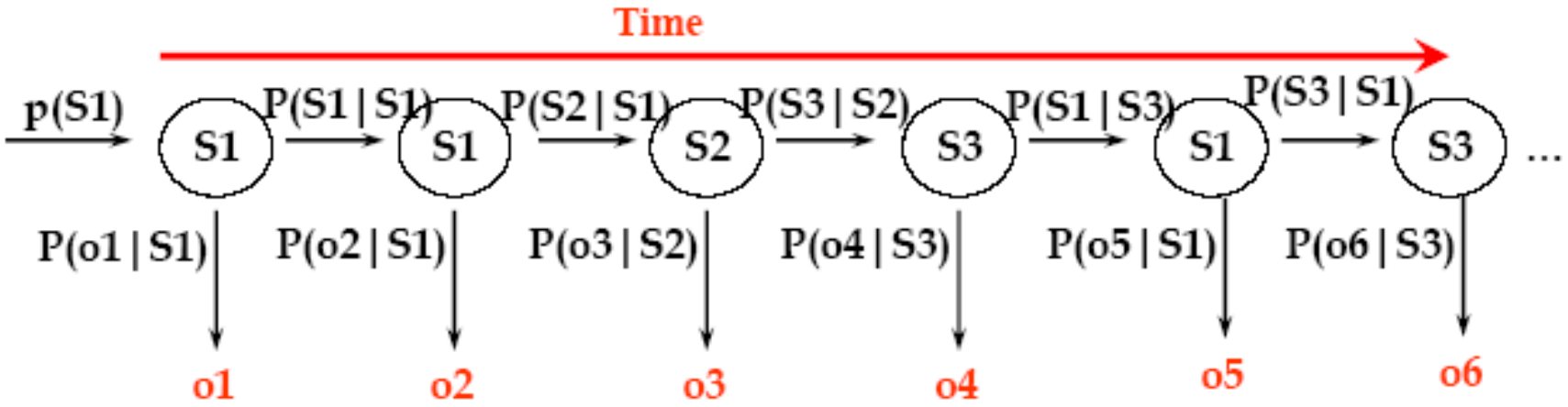
What is an HMM?



- Green circles are ***hidden states***
- Dependent only on the previous state: Markov process
- "The past is independent of the future given the present."

How Does an HMM Generate Data?

- 1. Pick an initial state
- 2. Given the state, pick an emission
- 3. Given the state, pick a transition to a next state
- 4. Go to step 2



Probability (state sequence, observation sequence)

= p(S1) P(o1|S1) P(S1|S1) P(o2|S1) P(S2|S1)…

Hidden Markov Models:
Inferring The Weather From What People Wear

HMM = (S, V, B, A, ?)

A set of states

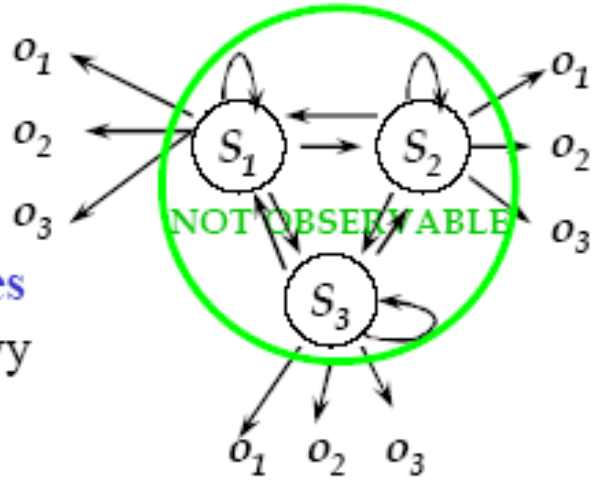
$\{S_{sunny}, S_{rainy}, S_{snowy}\}$

A set of observation types

$\{O_{shorts}, O_{coat}, O_{umbrella}\}$

Initial Distribution

? = (.70, .25, .05)



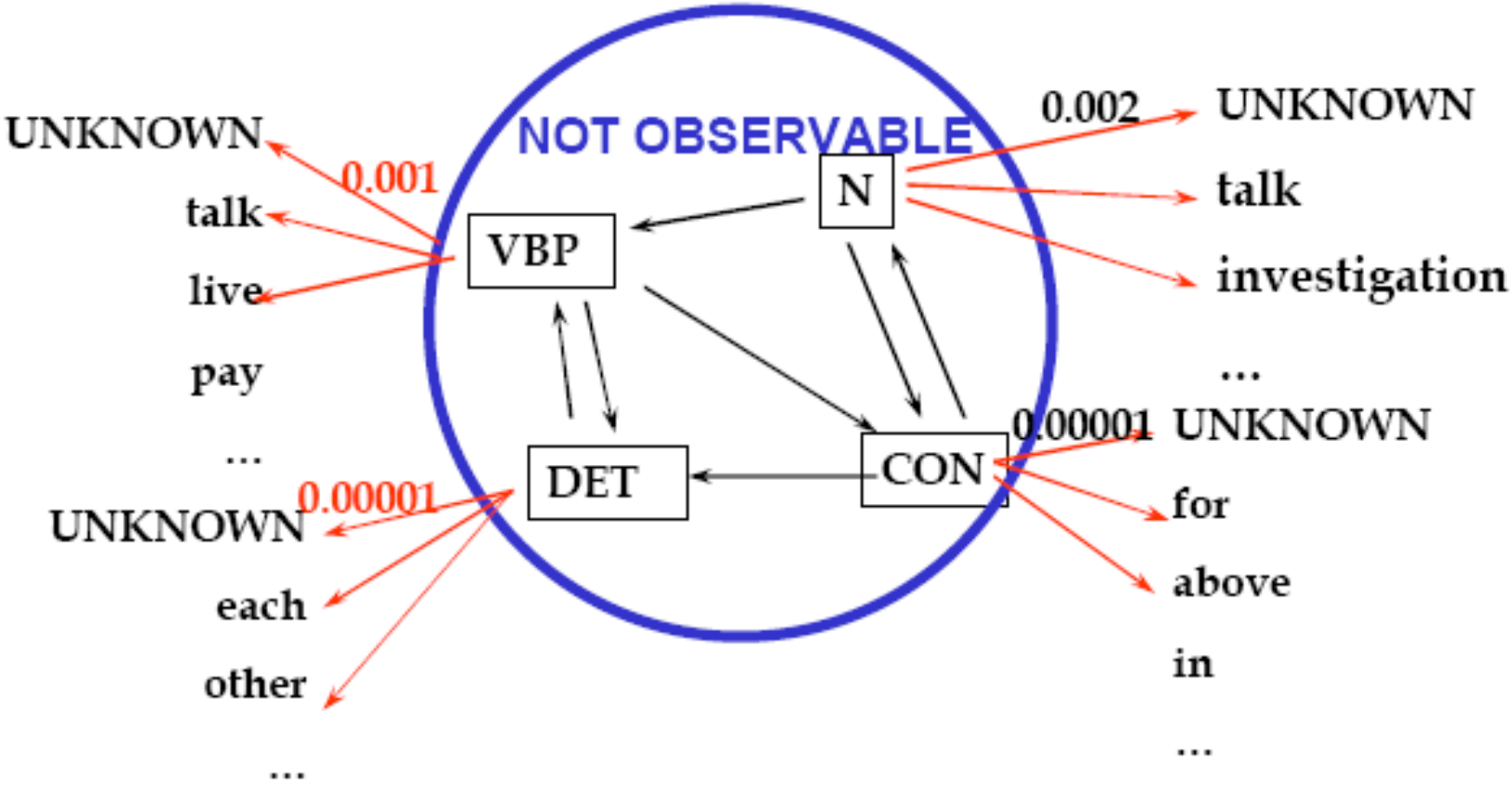
Emission Probabilities

$$B = \begin{pmatrix} .6 & .3 & .1 \\ .05 & .3 & .65 \\ 0 & .8 & .2 \end{pmatrix}$$

Transition Probabilities

$$A = \begin{array}{c} \\ \text{sunny} \\ \text{rainy} \\ \text{snowy} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Sunny} & \text{Rainy} & \text{Snowy} \end{array} \\ \begin{pmatrix} .8 & .15 & .05 \\ .38 & .6 & .02 \\ .75 & .05 & .2 \end{pmatrix} \end{array}$$

Hidden Markov Models of a Simple Part of Speech Tagger



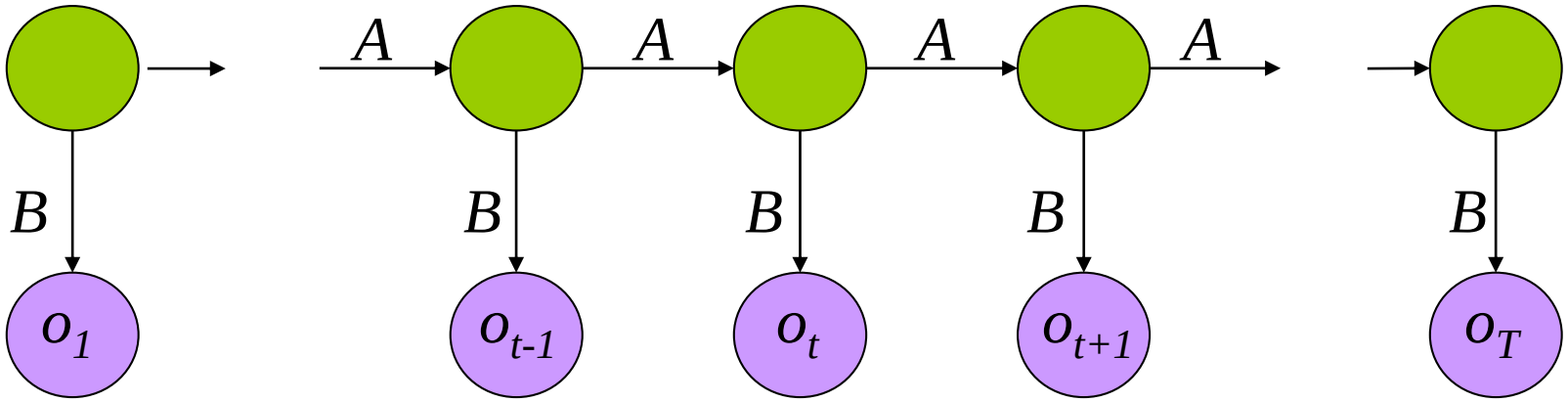
4 hidden states: VBP (verb) CON (conjunction) N (noun) DET (determiner)

Initial distribution: (N, VBP, CON, DET) = (0.1, 0.2, 0.1, 0.6)

隐马尔可夫模型

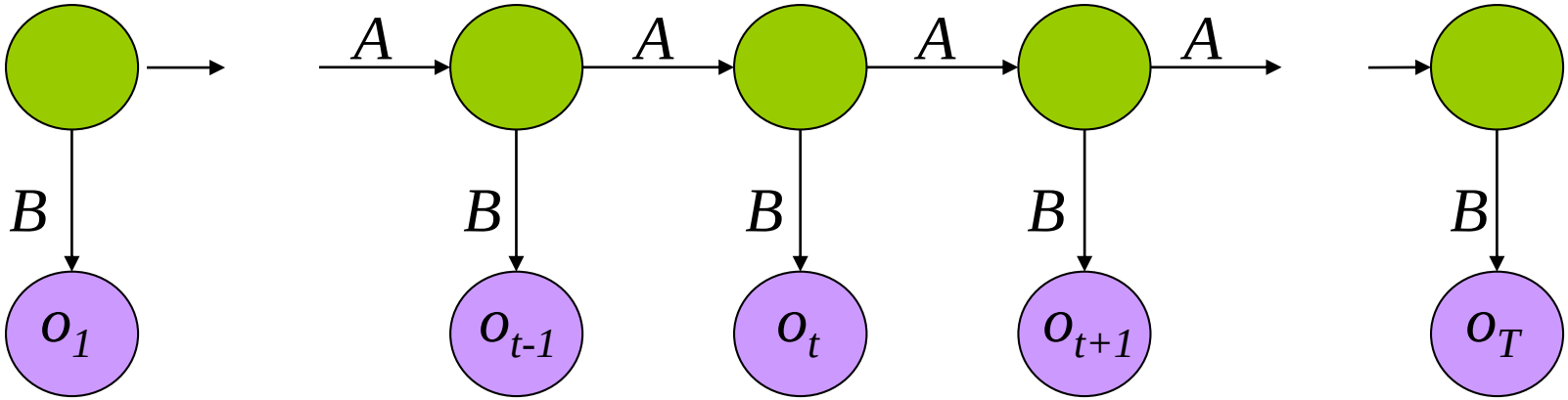
- 三个基本问题
  - 1）给定一个模型，如何高效的计算某一观察序列的概率；
  - 2）给定一个模型和一个观察序列，如何找到产生这一观察序列概率最大的状态序列；
  - 3）给定一个模型和一个观察序列，如何调整模型的参数使得产生这一序列的概率最大。
- 三个独立性假设
  - *t*时刻的状态只依赖于*t-1*时刻的状态；
  - *t*时刻所生成的观察值只依赖于*t*时刻的状态值；
  - 状态与具体时间无关。

Learning = Parameter Estimation



- Given an observation sequence, find the model that is most likely to produce that sequence.
- No analytic method, so:
- Given a model and observation sequence, update the model parameters to better fit the observations.

Parameter Estimation: Baum-Welch or Forward-Backward



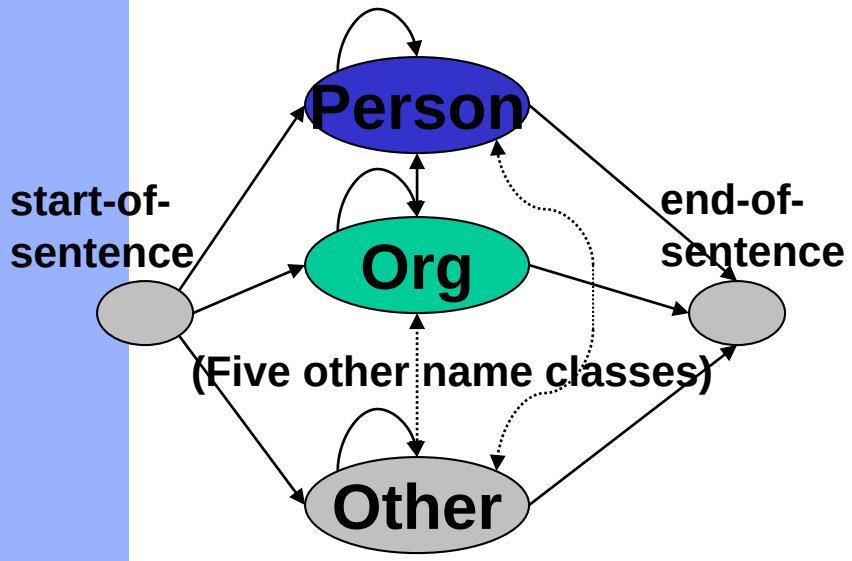
$$p_t(i,j) = \frac{\alpha_i(t) a_{ij} b_{jo_{t+1}} \beta_j(t+1)}{\sum_{m=1\ldots N} \alpha_m(t)\beta_m(t)}$$

Probability of traversing an arc

$$\gamma_i(t) = \sum_{j=1\ldots N} p_t(i,j)$$

Probability of being in state *i*

Example: Name Entity Extraction



Transition probabilities

$P(s_t \mid s_{t-1}, o_{t-1})$

Back-off to:

$P(s_t \mid s_{t-1})$

$P(s_t)$

Observation probabilities

$P(o_t \mid s_t, s_{t-1})$

or $P(o_t \mid s_t, o_{t-1})$

Back-off to:

$P(o_t \mid s_t)$

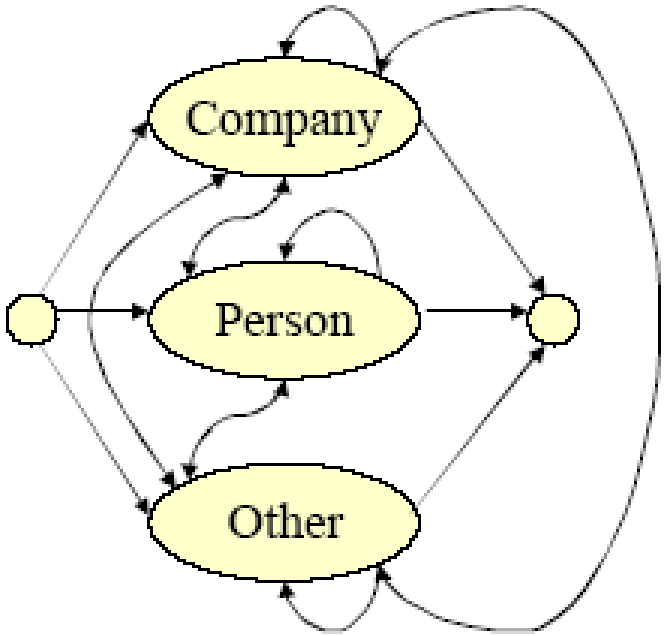
$P(o_t)$

Train on ~500k words of news wire text.

Results:

Case	Language	F1 .
Mixed	English	93%
Upper	English	91%
Mixed	Spanish	90%

Example: Name Entity Extraction



P (w | Company)	
Apple	0.0100
apple	0.0001
Clinton	0.0001
:	:

Emission probabilities for $S_{company}$

P (w | Person)	
Apple	0.00010
apple	0.00001
Clinton	0.01000
:	:

Emission probabilities for S_{person}

Text: President Clinton visited Apple Computer yesterday to announce ....
State: person person other company company other other other ....

Example: Research Paper

Observations:

Learning Hidden Markov Model Structure for Information Extraction

Kristie Seymore[†]
kseymore@ri.cmu.edu

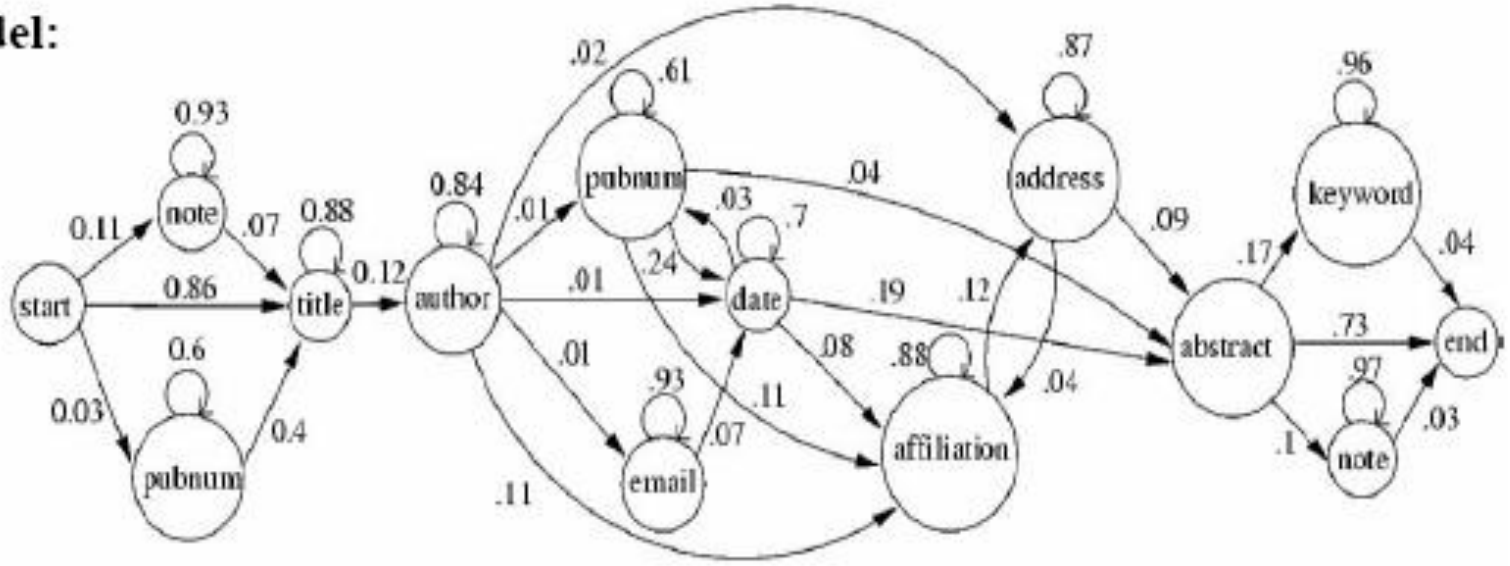
Andrew McCallum[†‡]
mccallum@justresearch.com

Ronald Rosenfeld[†]
roni@cs.cmu.edu

[†]School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[‡]Just Research
4616 Henry Street
Pittsburgh, PA 15213

Model:



隐马尔可夫模型

- 在隐马尔可夫模型中，假设*t*时刻的观察值只依赖于*t*时刻的状态，这确保了序列中的所有观察值互相独立。
- 但事实上，数据序列并不能完全地表示为一组独立的单元。
- 当序列中的数据元素存在长距离依赖时，允许这种长距离依赖并且使观察序列可以表示为非独立的交叉特征的模型才是比较合适的。

最大熵模型

- Maximum Entropy Models, MaxEnt
- 一个随机变量的不确定性是由熵体现的

$$H(X) \equiv -\sum_{x \in X} p(x) \log p(x)$$

- 熵最大时随机变量最不确定。
- 在已知部分知识前提下，关于未知分布最合理的推断是符合已知知识的最不确定（最随机）的推断。
- 将已知事件作为约束条件，求得可使熵最大化的概率分布作为正确的概率分布

最大熵模型

- 定义P为所有条件分布的集合

$$p(y \mid x) \in \mathrm{P}$$

- m个对模型真正有用的特征函数 f_i (体现统计数据的特性)
- 约束条件下所产生的集合C是P的一个子集

$$C = \left\{ p \in \mathrm{P} \mid p(f_i) = \tilde{p}(f_i), \quad i \in \{1,2,\cdots,m\} \right\}$$

- 对于任意给定的约束集C，能找到唯一的 $p^* \in C$ 使 $H(p)$ 取得最大值 $p^* = \arg\max_{p \in C} H(p)$

$$H(p) \equiv -\sum \tilde{p}(x) p(y \mid x) \log p(y \mid x)$$

最大熵模型

- 对模型中的每一个特征 f_i 都引入一个参数 λ_i，即Lagrange乘子
- 定义Lagrange函数(熵定义和约束条件)

$$\Lambda(p,\lambda) = H(p) + \sum_{i=1}^{m} \lambda_i (p(f_i) - \tilde{p}(f_i))$$

$$p_\lambda \equiv agr\max_{p \in \mathrm{P}} \Lambda(p,\lambda) \qquad \Psi(\lambda) \equiv \Lambda(p_\lambda,\lambda)$$

- 根据KT对偶定理，原问题：

$$p^* = \arg\max_{p \in C} H(p)$$

转换为对偶问题：

$$\lambda^* = \arg\max_{i \in \{1,2,\cdots,m\}} \Psi(\lambda_i)$$

最大熵模型

- 极值←→一阶偏导数等于零

$$\frac{\partial \Lambda(p,\lambda)}{\partial p} = -\sum_{x,y} \tilde{p}(x)(\log^{p(y|x)} + 1) + \sum_{x,y} \tilde{p}(x) \sum_{i} \lambda_i f_i(x,y)$$

←→

$$p_\lambda^* = \frac{1}{e} \exp(\sum_i \lambda_i f_i(x,y))$$

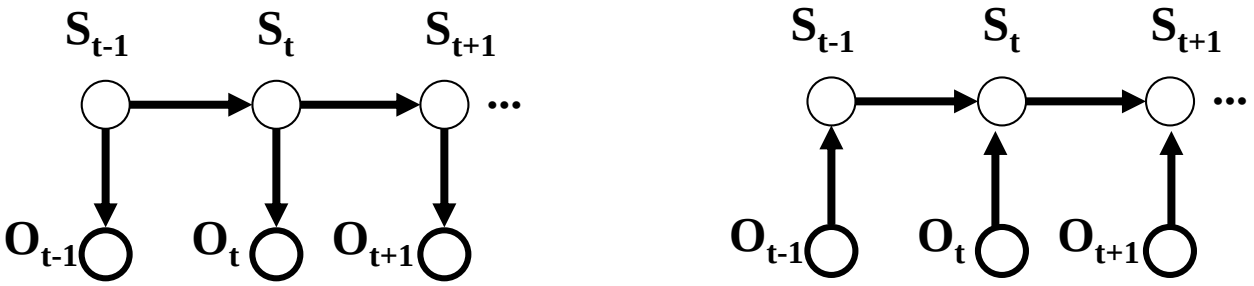
- 归一化可得：（$\sum_y p_\lambda^*(y|x) = 1$）

$$p_\lambda = \frac{1}{Z_\lambda(x)} \exp(\sum_i \lambda_i f_i(x,y))$$

$$Z_\lambda(x) = \sum_y \exp(\sum_i \lambda_i f_i(x,y))$$ （归一化因子）

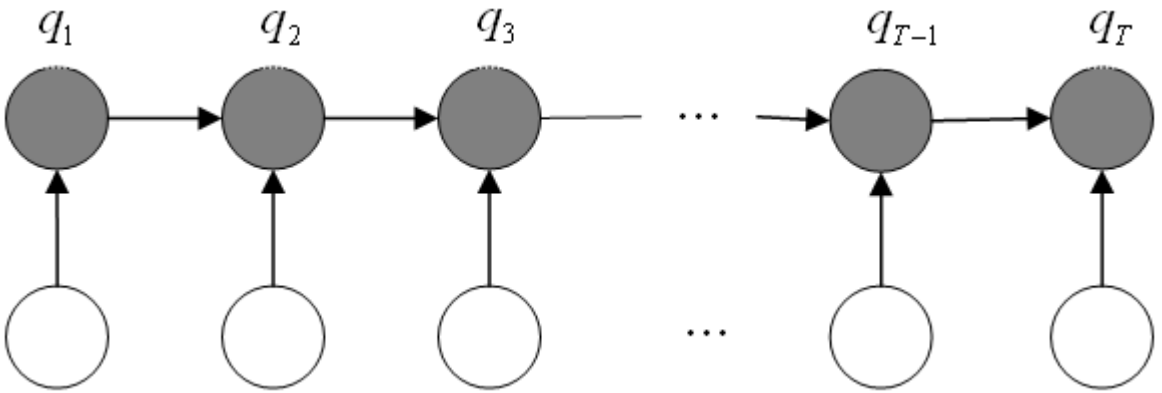
最大熵马尔可夫模型

- Maximum Entropy Markov Models, MEMMs
- 是Conditional Markov Models (CMMs)的特例
- 是HMM的改进，克服了HMM严格的独立假设；
- Idea: replace generative model in HMM with a maxent model, where state depends on observations and previous state history



最大熵马尔可夫模型

- MEMMs也是基于概率有限状态模型的概念，但与HMM不同的是，MEMMs将观察序列看作条件事件，而不是由状态生成的。
- HMM的转移函数和发射函数，在MEMMs中被一个分布函数 $p(q_t \mid q_{t-1}, O_t)$代替；
- 该分布函数表示在给定观察值 O_t条件下，由前一个状态 q_{t-1} 转移到当前状态 q_t 的概率。



最大熵马尔可夫模型

- 由最大熵模型可知：

$$p(q_t \mid q_{t-1}, \mathrm{O}_t) = \frac{1}{Z(q_{t-1}, \mathrm{O}_t)} \exp\left(\sum_i \lambda_i f_i(q_t, \mathrm{O}_t)\right)$$

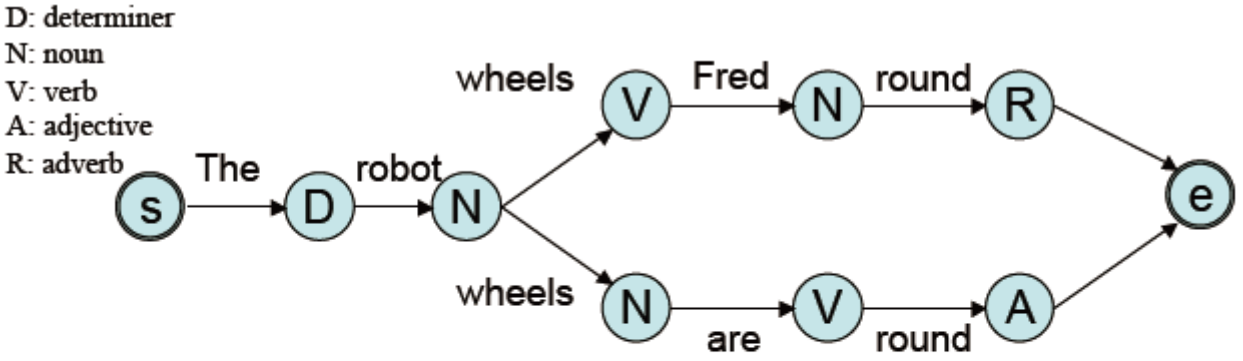
- 其中：
  - $Z(q_{t-1}, \mathrm{O}_t)$ 是归一化因子；
  - f_i 是当前观察值 O_t 和当前可能的状态 q_t 的二元特征函数；
  - λ_i 是 f_i 的权值，是需要估计的参数。

最大熵马尔可夫模型—优

- MEMMs结合了隐马尔可夫模型和最大熵模型的优点：
  - 它允许状态转移可以基于输入序列中的非独立特征。
  - 这使得最大熵马尔可夫模型性能优于隐马尔可夫模型。

最大熵马尔可夫模型—缺

- MEMMs和其它基于下一个状态分类的“判别式”有限状态模型，有个共同的缺点：标记偏置问题(Label Bias Problem)
  - States with low-entropy next-state distributions ignore observations

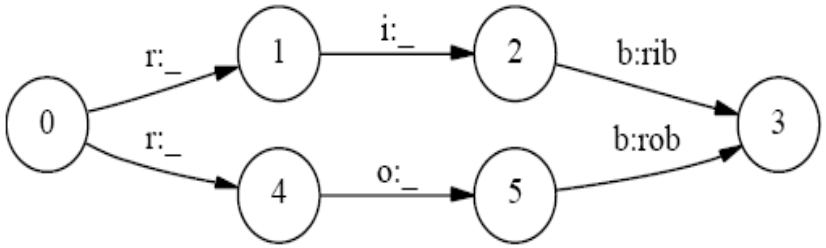


Obs: “The robot wheels are round.”

But if P(V|N,wheels) > P(N|N,wheels),
then upper path is chosen regardless of obs.

最大熵马尔可夫模型—缺

- MEMMs和其它基于下一个状态分类的“判别式”有限状态模型，有个共同的缺点：标记偏置问题(Label Bias Problem)
  - States with low-entropy next-state distributions ignore observations



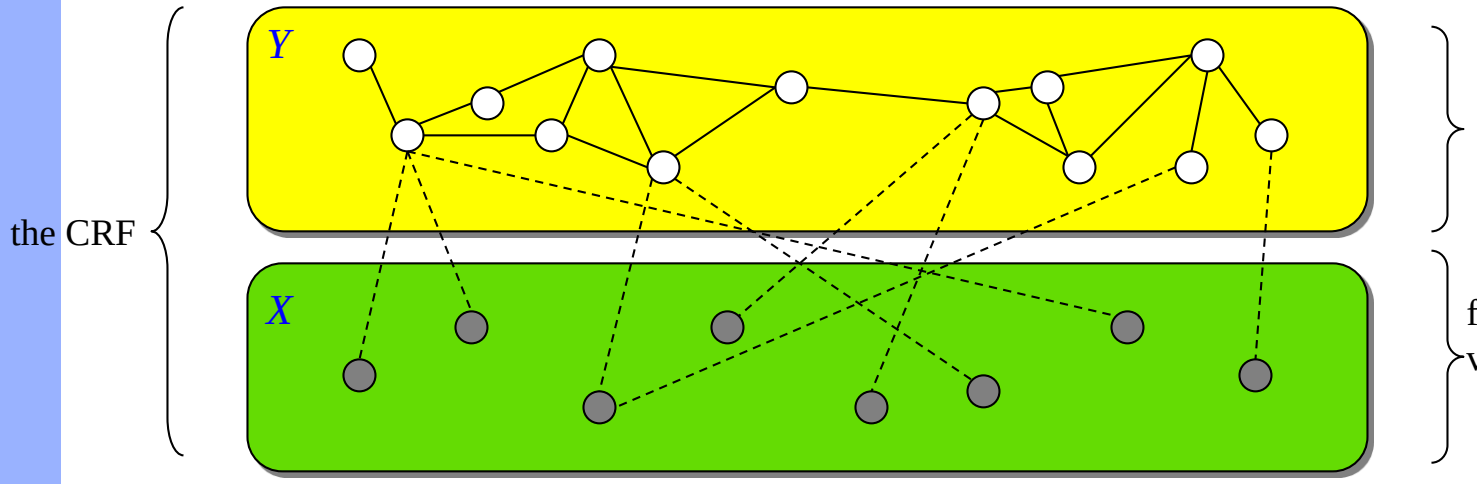
- 标记偏置的主要原因是局部归一化
- 标记偏置问题对MEMMs的应用带来很大的负面影响。

条件随机场(CRF)

- Conditional Random Fields, CRFs
- 由Lafferty等人于2001年提出；
- 模型思想的主要来源是最大熵模型；
- 模型的三个基本问题的解决用到了HMMs模型中提到的方法如forward-backward和Viterbi，而其参数训练部分有所不同；
- CRFs是在给定需要标记的观察序列的条件下，计算整个标记序列的联合概率，而不是在给定当前状态条件下，定义下一个状态的分布。

条件随机场(CRF)

- 标记序列(Label Sequence)的分布条件属性，可以让CRFs很好的拟合现实数据，而在这些数据中，标记序列的条件概率信赖于观察序列中非独立的、相互作用的特征，并通过赋予特征以不同权值来表示特征的重要程度。
- 是一种用来标记和切分序列化数据的统计模型。



CRF的概念

- 设无向图G=<V,E>中每个节点对应一个随机变量Y_v，当在条件X下，随机变量Y_v的条件概率分布服从图的马尔可夫属性

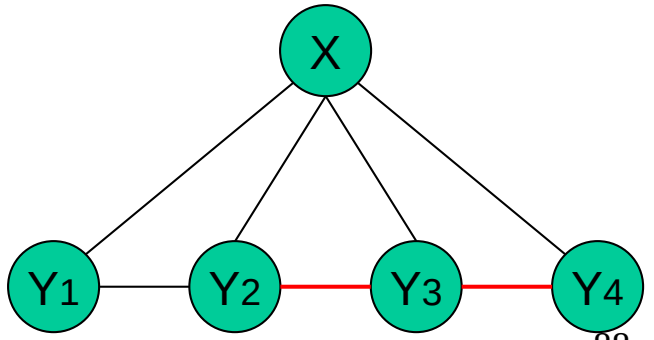
$$p(Y_v \mid X, Y_w, w \neq v) = p(Y_v \mid X, Y_w, w \sim v)$$

其中$w \sim v$表示(w,v)是无向图的边

则称(X,Y)为一个条件随机场。

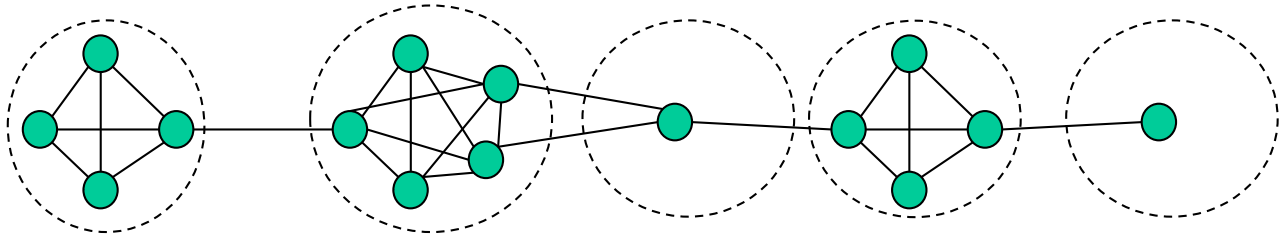
- 例如右图：

$$p(Y_3 \mid X, \text{其它}Y) = p(Y_3 \mid X, Y_2, Y_4)$$



CRF的势函数

- 随机变量联合概率



$$p(v_1, v_2, \cdots, v_N) = \frac{1}{Z} \prod_{c \in C} \Phi_{v_c}(v_c)$$

- $\{c\}$为G节点的划分，c是团(clique)
  (完全子图C是G的一个团<=>C不包含在G的更大的完全子图中);
- Ψc为团c的势函数(potential funciton)，代表了子图所施加的约束数量。
- 图的势越高，则整体约束越多，与全局优化是一致的。

CRF的势函数

- 定义每个势函数的形式如下：

$$\Phi_{y_c}(y_c) = \exp(\sum_k \lambda_k f_k(c, y \mid c, x))$$

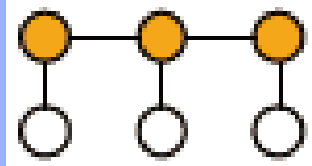
- CRF的分布函数：

$$p(y \mid x) = \frac{1}{Z(x)} \exp(\sum_{c \in C} \sum_k \lambda_k f_k(c, y \mid, x)$$

- 特例：两类不同特征：

$$p_\theta(y \mid x) = \frac{1}{Z(x)} \exp(\sum_{e \in E, k} \lambda_k f_k(e, y_e, x) + \sum_{v \in V, k} \mu_k g_k(v, y_v, x))$$

- 特例： HMM-like Chain graph



$$p_\theta(y \mid x) = \frac{1}{Z(x)} \exp(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \mu_k g_k(y_i, x))$$

CRF的参数估计

- 由最大熵模型可知，CRFs模型的参数估计实质是对概率的对数最大似然函数求最值

$$L(\lambda) = \sum_{x,y} \tilde{p}(x,y) \log p(y \mid x, \lambda)$$

  - 运用最优化理论循环迭代，直到函数收敛或达到给定的迭代次数

$$\lambda_k \leftarrow \lambda_k + \delta\lambda_k \qquad u_k \leftarrow u_k + \delta u_k$$

    - Initialize each λ_k
    - Do until convergence:

      Solve $\frac{dA(\theta', \theta)}{d\delta\lambda_k} = 0$ for each $\delta\lambda_k$

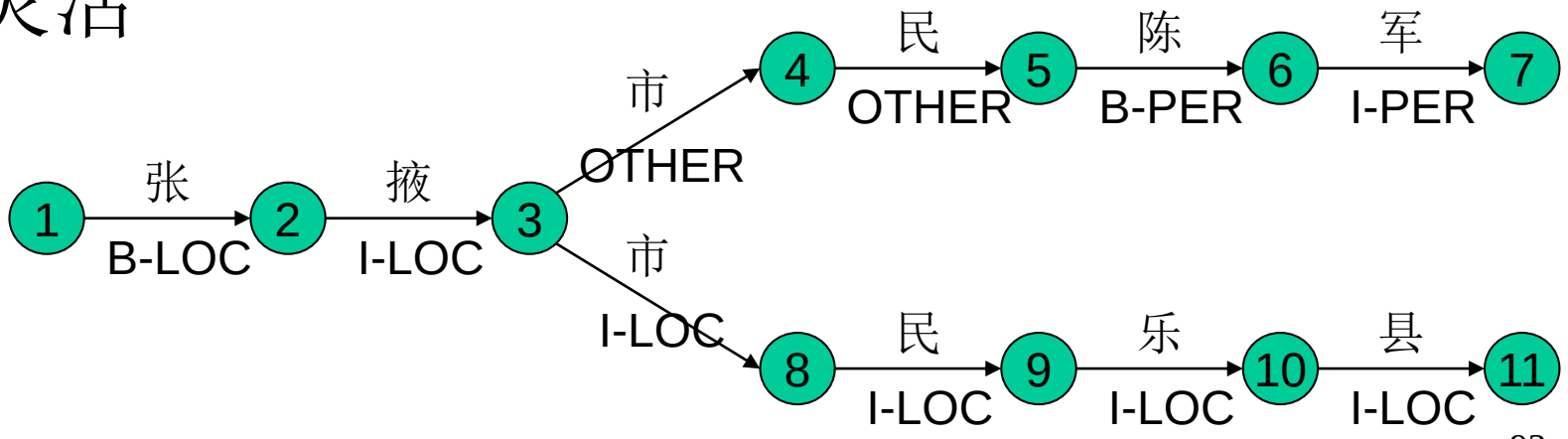
      Update parameter: $\lambda_k \leftarrow \lambda_k + \delta\lambda_k$

CRF与经典模型的对比

- 特点
  - 不要求独立性假设，特征灵活，全序列最优；
  - 没有标注偏置问题。
- 不足
  - 参数估计收敛慢；
  - 迭代的复杂度高：$O(L^2NTF)$
  - 对非链结构复杂度更高；
  - 链中的重现不能有效利用和识别。

CRF与经典模型的对比

- MaxEnt：较长的观察窗口，特征灵活
- HMM：较长的状态窗口，特定特征
- MEMM：较长的观察＋状态窗口，标注偏置
- CRF：全序列的（观察＋状态）窗口，特征灵活



CRF资源与工具

- http://www.inference.phy.cam.ac.uk/hmw26/crf/#software
- CRF++

  http://crfpp.sourceforge.net/
- Pocket CRF

  http://sourceforge.net/projects/pocket-crf-1

CRF的应用

- POS, Word Segmentation
- 英语NER
  - F1: 84.04% (CoNLL-2003 shared English)
- 语音停顿识别
  - 是HMM, MaxEnt和CRF中最好的模型

CRFs模型进行命名实体识别

- 利用CRFs进行词位标注
  - 俄/LB 罗/LM 斯/LE 总/N 统/N 叶/PB 利/PM钦/PE 2 /N 3 /N 日/N 在/N 莫/LB 斯/LM 科/LE 会/N 见/N 了/N 联/OB 邦/OM 委/OM 员/OM 会/OE 主/N 席/N 斯/PB 特/PM 罗/PM 耶/PM 夫/PE
- 寻找匹配模式
  - [俄/LB 罗/LM 斯/LE] [叶/PB 利/PM钦/PE]
  - [莫/LB 斯/LM 科/LE] [联/OB 邦/OM 委/OM 员/OM 会/OE] [斯/PB 特/PM 罗/PM 耶/PM 夫/PE]
- 最终结果
  - [L 俄罗斯]/总统/[P 叶利钦]/ 2 3 日在 /[L 莫斯科]/会见了/[O 联邦委员会]/主席/[P 斯特罗耶夫]/

CRFs模型进行命名实体识别
特征模板

模板类型	模板格式	说明
上下文词形特征	C_n, n=-2,-1,0,1,2	一元词形特征
	C_nC_{n+1}, n=-2,-1,0,1	二元词形特征
	$C_{-1}C_1$	二元词形特征
人名识别特征	IsSurname(C_0)	当前字是否为姓氏常用字
	IsNameChar(C_0)	当前字是否为人名常用字
	IsNameAffix(C_0)	当前字是否为人名常用前后缀的首字
地名识别特征	IsLN(C_0)	当前字是否为地名首字
	IsLNSuffix(C_0)	当前字是否为地名常用后缀的首字
机构名识别特征	IsON(C_0)	当前字是否为机构名首字
	IsONSuffix(C_0)	当前字是否为机构名常用后缀的首字

小结

- 文本信息抽取的概念
- 文本信息抽取的方法
  - Wrapper
  - 有限状态机
  - Hidden Markov Models
  - Conditional Random Fields

Any Question?