



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:yangjianwu@icst.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

Corporate Intelligence - Microsoft Internet Explorer

Back File Edit View Favorites Tools Help

Address C:\My Documents\corp\AugustDemo\processed-data-0815-2a\index.htm Go Links

For more information contact:

Dan Carter
Marketing Manager
MarketSoft Corporation
781-674-0000 x:302
carter@marketsoft.com

Brent Skinner
Account Coordinator
The Weber Group
617-520-7054
bskinner@webergroup

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

[TOP OF]

ARKETING NETWORK | NEWS & EVENTS |
CONTACT US | CUSTOMER SIGN-IN
FORMATION | JOBS | SEARCH
ckwave Product Demos!

10 Maguire Road, Suite 330, Lexington MA 02421-3112
T: 781.674.0000 | F: 781.674.0090

Copyright © 2000 MarketSoft Corporation. All rights reserved.
Send mail to charliea@marketsoft.com with questions or comments.

Source: Whizbang! Labs/
Andrew McCallum

MarketSoft Corporation (?)

<http://marketsoft.com>

Street address: [10 Maguire Road, Suite 330](#)
City: [Lexington](#)
State: [MA](#)
Zip code: [02421-3112](#)
Telephone: [781-674-0000 \(?\)](#)
Fax: [\(212\) 924-0240 \(???\)](#)
Email: info@marketsoft.com
SIC code: 7372 [Prepackaged software] (???)

Data automatically
extracted from
[marketsoft.com](#)

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

[10011-6901](#)

Digital Equ



Product information

IBM - Google Product Search - Windows Internet Explorer
http://www.google.com/products?q=IBM&show=dd

文件(F) 编辑(E) 查看(V) 收藏夹(A) 工具(T) 帮助(H)
☆ + 家 页面(P) 工具(O) ?

Google Web Images Video News Maps more »
IBM Search Advanced Product Search Preferences

Products Results 1 - 10 of about 1,000,000

Show Google Checkout items only Showing all items Show grid view Sort by relevance

Product	Price	Seller
 IBM Thinkpad T41 PM 1.5GHz 512MB 40GB DVD/CDRW WiFi XP IBM Warranty Good till 04/2008. Better then New Laptop warranty!!! This product is entitled to parts and labor and is entitled to IBM EZServ. ... Add to Shopping List	\$599.00	eBay All items from seller
 IBM ThinkPad X31 Pentium M 1.4GHz ** Refurbished ** Refurbished ** IBM ThinkPad X31 Pentium M 1.4GHz List Price : \$ 434 Description IBM Lenovo Thinkpad X31 Laptop Pentium M Centrino 1.4GHz ... ★★★★★ 4.8 from 2 product reviews - Add to Shopping List	\$543.90	Vendio All items from seller
 IBM IBM Thinkpad R51 PM 1.6G/256M/60G/Combo/WiFi/15 SXGAMWXP IBM IBM ThinkPad R51 PM 1.6G/256M/60G/Combo/WiFi/15 SXGAMWXP. ★★★★★ 4.4 from 5 product reviews - Add to Shopping List	\$475.00	uBid.com - The Market... All items from seller
 IBM SureOne 4614-A05 - C3 866 MHz - 10" TFT IBM SureOne 4614-A05 - C3 866 MHz - 10" TFT. ★★★★★ 4.5 from 2 product reviews - Add to Shopping List	\$1,974.99	Zones ★★★★★ 69 seller ratings All items from seller
 ** Wireless IBM Thinkpad P3 DVD 256RAM Laptop WinXP ** IBM THINKPAD Wireless Internet WiFi Ready! Description WIRELESS WIFI IBM THINKPAD OUTLINE STICKER 14.1" INCH POWERBOOK DV IBM THINKPAD EXC	\$289.00	Overstock.com Auction... All items from seller



Product information

IBM Thinkpad R60 - CNET Reviews - Windows Internet Explorer
http://reviews.cnet.com/4244-5_7-0.html?query=IBM+Thinkpad+R60&tag=srch&target=

文件(?) 编辑(?) 查看(?) 收藏夹(?) 工具(?) 帮助(?) 链接 >

☆ + 家 页面(?) 工具(?) ?

c|net REVIEWS

Search: Reviews

Today on CNET **Reviews** News Downloads Tips & Tricks CNET TV Compare Prices **VIDEO** Explore music in Firefox

Cell phones | Desktops | Digital cameras | Laptops | MP3 players | TVs | All Categories

THE LENOVO N100 WIDESCREEN New World. New Thinking.[™]

\$949 PN 0768DCU (\$999 before \$50 mail-in-rebate)

Intel Core 2 Duo Processor Dual-core. Do more.

15.4" WXGA VibrantView TFT display • Upgrade offer for Microsoft Vista[™] Home Premium
• Intel[®] Core[™]2 Duo Processor • Free shipping [Learn more](#)

Dell Home HP Home Lenovo HP Biz Intel Dell Biz

Narrow by category Laptops

Top matching products for "IBM Thinkpad R60"

Sort by: Relevance | Most popular | Rating | Review date | Price

COMPARE Select two or more checkboxes and click Compare to see products side by side

Lenovo ThinkPad R60 9461 (Core Duo 2.67 GHz, 1 GB RAM, 100 GB HDD) Good Check prices \$973
Businesses seeking a sturdy, secure, portable workhorse should consider the ThinkPad R60. [Read review](#)

Lenovo ThinkPad R60 9457 (Core Duo 1.83 GHz, 512 MB RAM, 60 GB HDD) Check prices \$998 - \$1,750
Businesses seeking a sturdy, secure, portable workhorse should consider the ThinkPad R60. [Read review](#)

Lenovo ThinkPad R60 9460 (Core Duo 2.67 GHz, 1 GB RAM, 100 GB HDD) Check prices \$1,849 -

CNET.com
Keep your kids safe online

http://reviews.cnet.com/4244-5_7-0.html?query=IBM+Thinkpad+R60&tag=srch&target=

Internet | 保护模式: 禁用 100% 8



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



文本信息提取类型

- 实体提取
 - ❖ 上下文无关实体的提取
 - Context-Free Entity Extraction
 - ❖ 基于规则的实体提取
- 关系提取(Relational Extraction)



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



有限状态机方法



有限状态机方法 识别命名实体



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

President Clinton

Mickey Kantor

Ms. Xuemei Peng

Mr. Hideto Suzuki

Mr. Langford

PLACES

Singapore

China

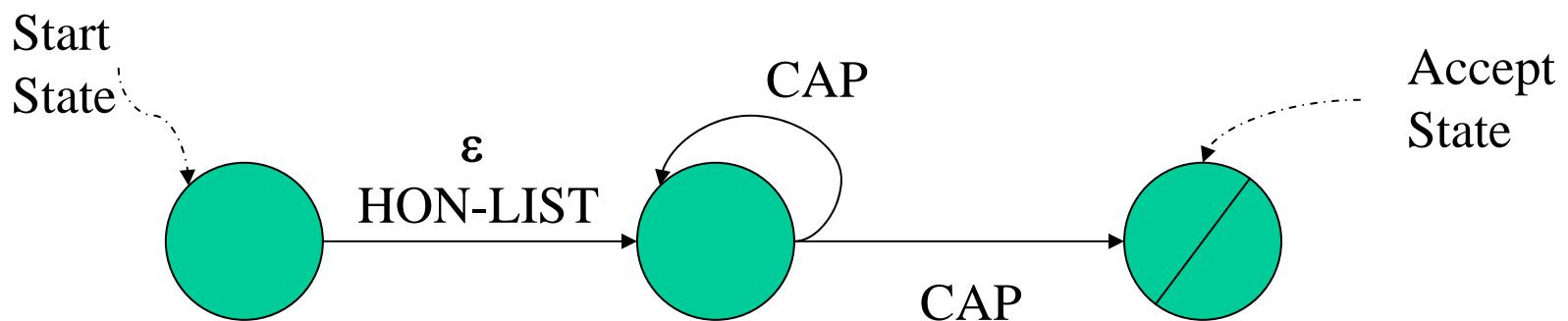
Japan

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）

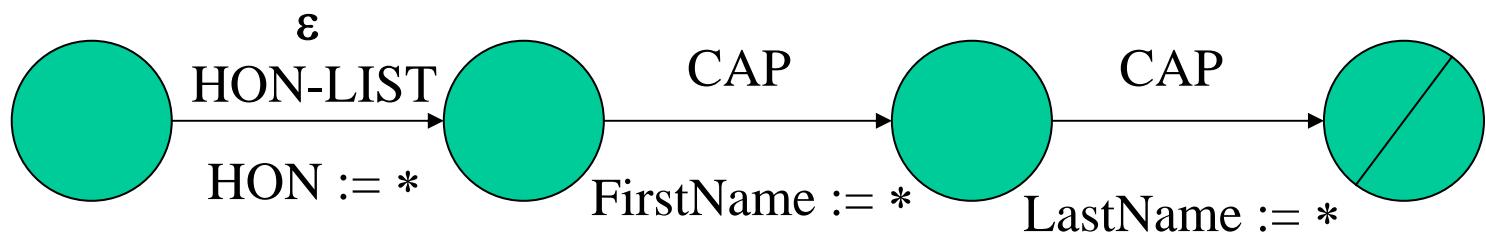
e.g. CAP matches any capitalized word
 - ❖ 表成员匹配（list-membership， 例如称呼）

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
```



有限状态机方法 识别关系信息



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"]

]



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** rapidly (around 10% failures each month)



Amazon Book Description

....

<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence

by Ray Kurzweil

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 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-based extraction



```
<HTML><TITLE>Some Country Codes</TITLE>
<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>
</BODY></HTML>
```



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



Wrapper induction

➤ Learning LR wrappers

labeled pages

```
<HTML><HEAD>Some Country Codes</HEAD>
```

```
  <HTML><HEAD>Some Country Codes</HEAD>
```

```
    <HTML><HEAD>Some Country Codes</HEAD>
```

```
      <HTML><HEAD>Some Country Codes</HEAD>
```

```
        <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>
```

```
      </BODY></HTML>
```

wrapper

→ $\langle l_1, r_1, \dots, l_K, r_K \rangle$

Example: Find 4 strings

$\langle <\!\!<\!\!B\!\!>, <\!\!<\!\!/\!\!B\!\!>, <\!\!<\!\!I\!\!>, <\!\!<\!\!/\!\!I\!\!> \rangle$

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



LR: Finding r_1

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

$r_1 : $



LR: Finding l_1 , l_2 and r_2

<HTML><TITLE>Some Country
Codes</TITLE>

```
<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>R2: </I>
</BODY></HTML>
```

$l_2 : <I>$

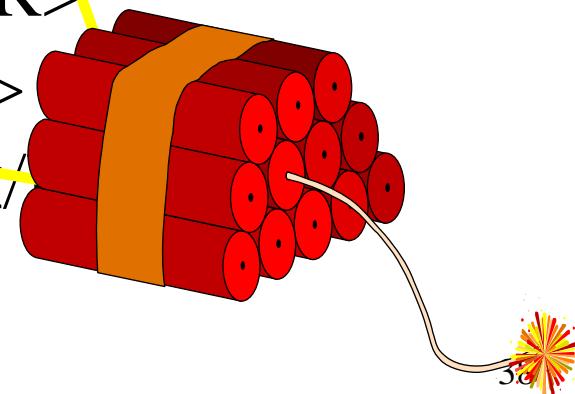
$l_1 : $



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></HTML>
```





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><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></HTML>
```

end of head

} head

} body

} tail

start of tail

Head-Left-Right-Tail wrappers



More sophisticated wrappers

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



Boosted wrapper induction

- 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

Natural Language Processing-based Information Extraction



- If extracting from automatically generated web pages, simple regex 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

Example of IE from FASTUS (1993)

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

(高尔夫球棍)

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

Example of IE from FASTUS (1993)

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs 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



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.



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



机器学习方法

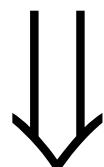


Learning for IE

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)

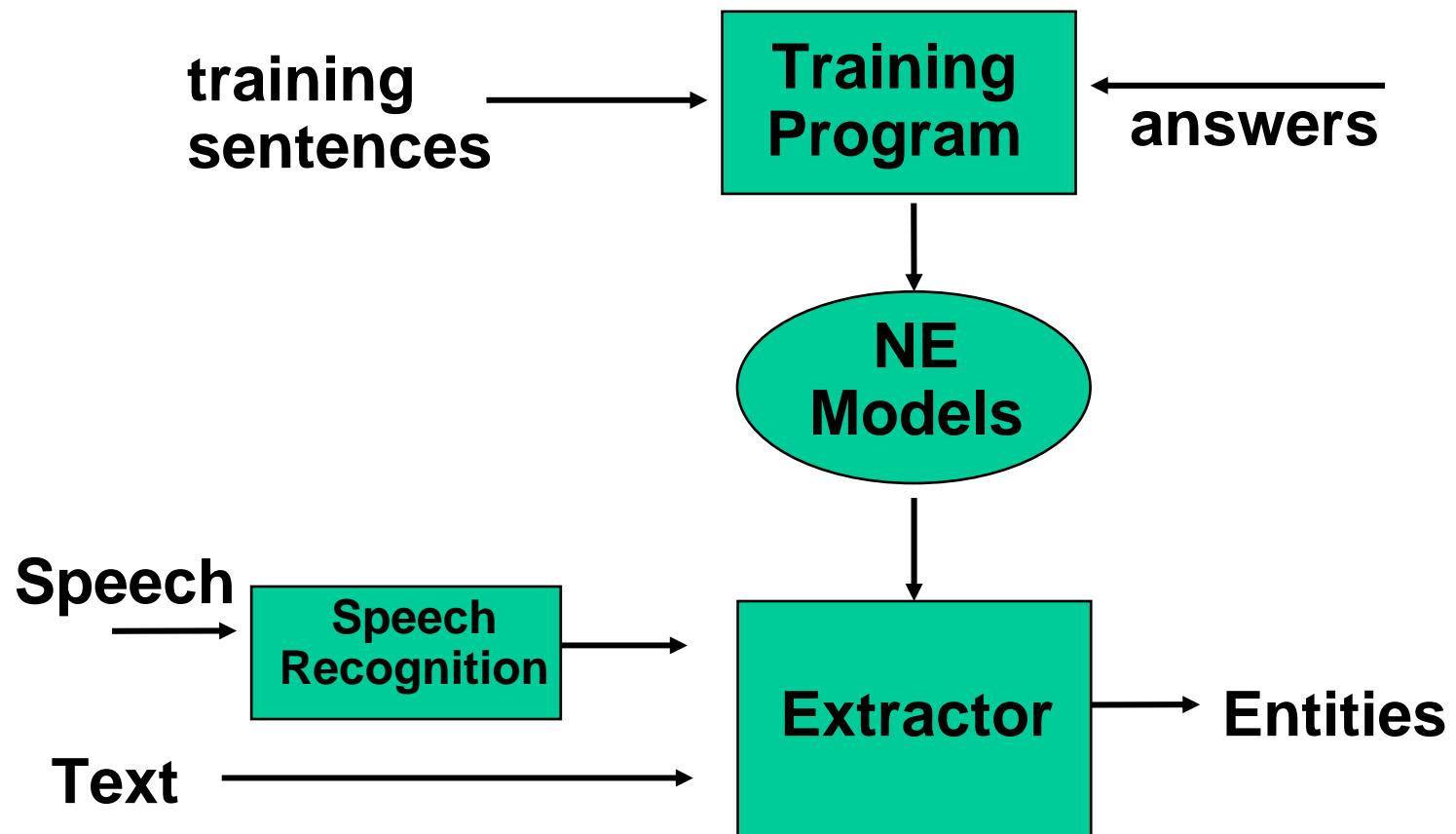


Statistical generative models

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



Name Extraction via HMMs



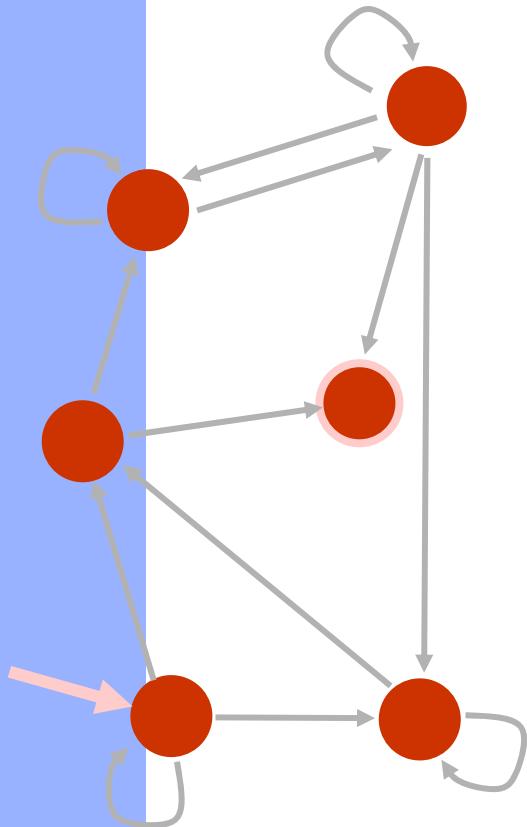


Applying HMMs to IE

- **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



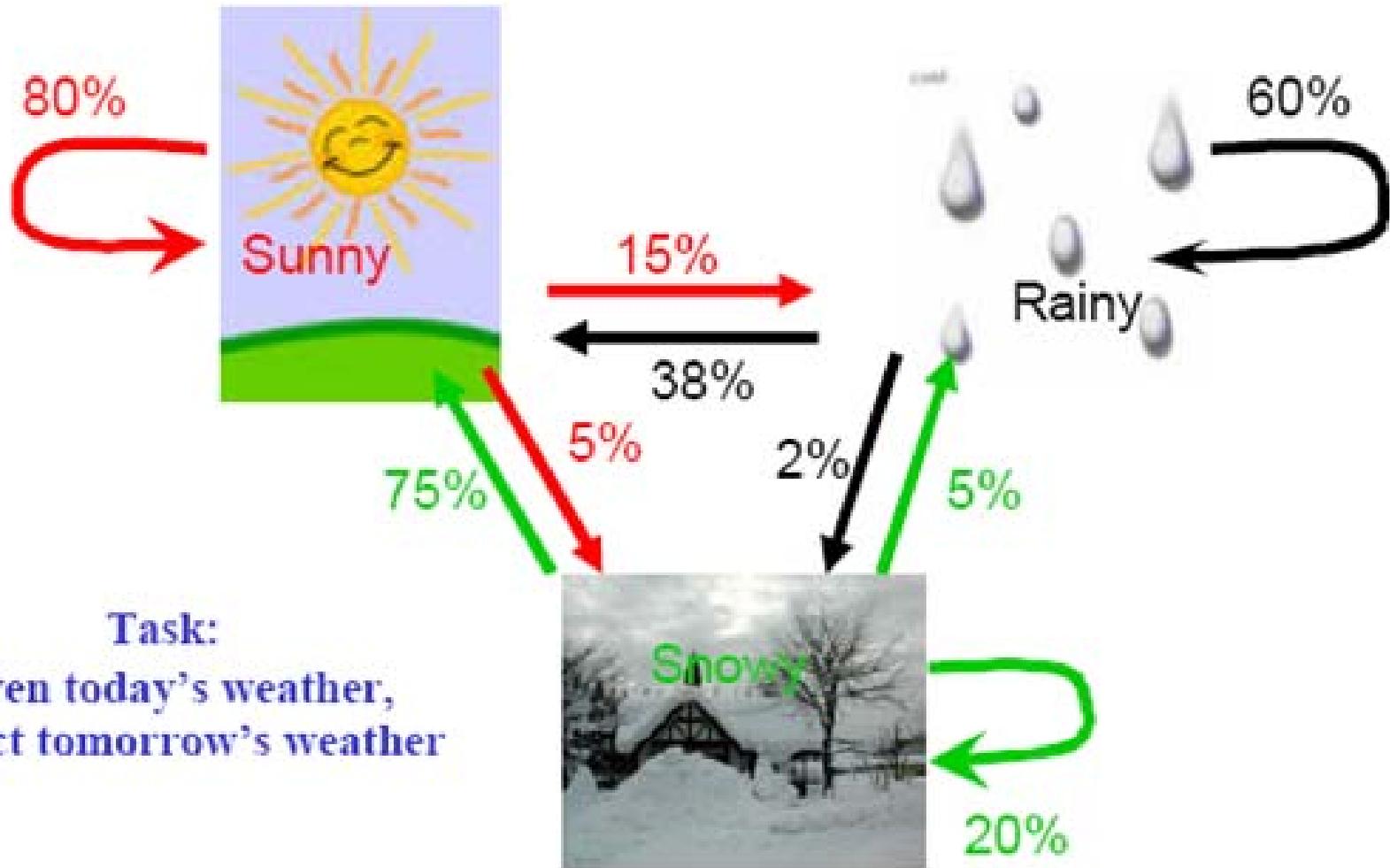
HMM = probabilistic FSA

HMM = states s_1, s_2, \dots
(special start state s_1
special end state s_n)
token alphabet a_1, a_2, \dots
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].

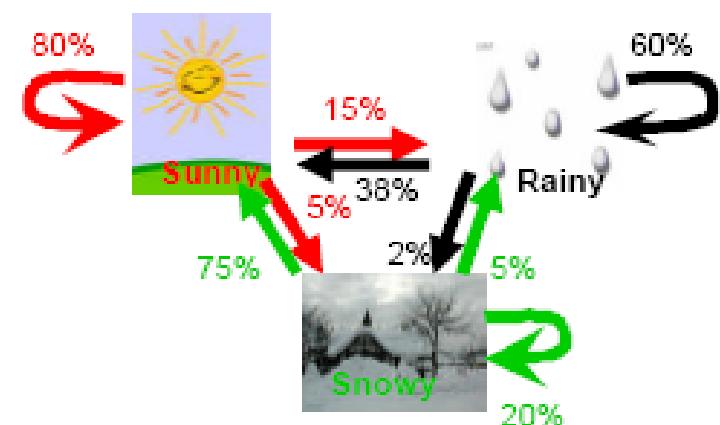
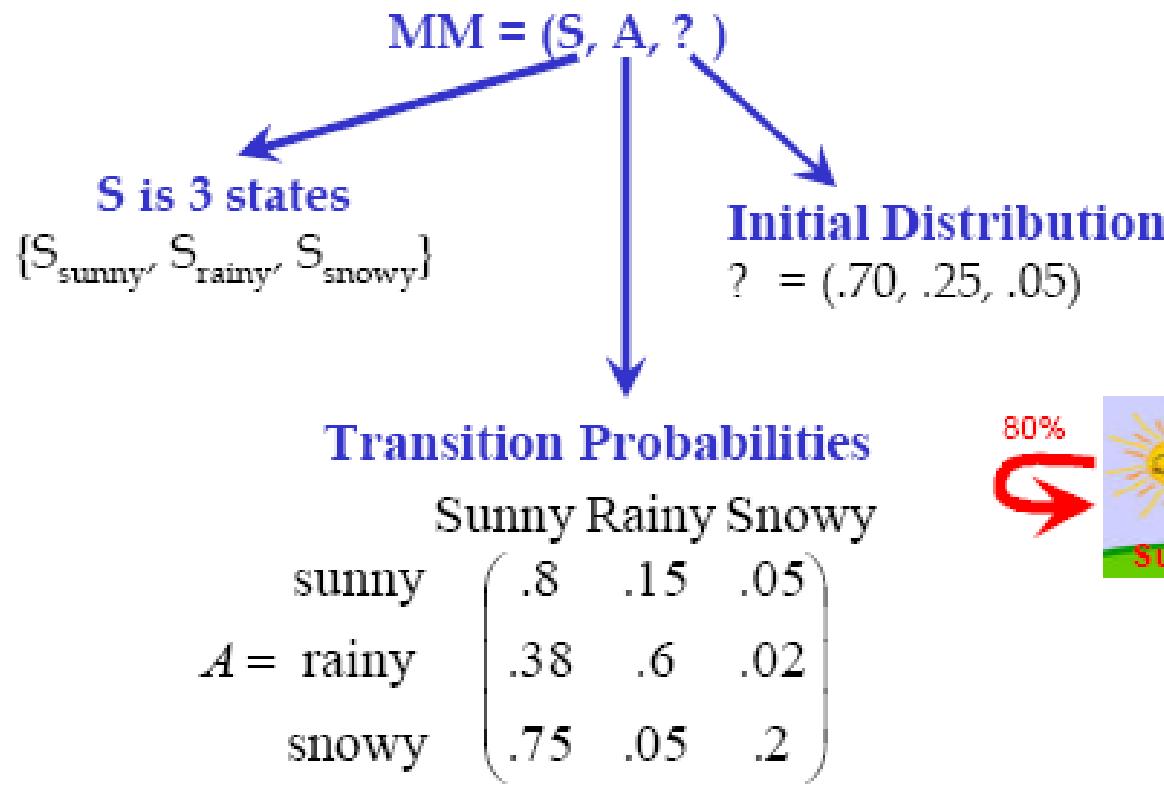


A Markov Model : Weather



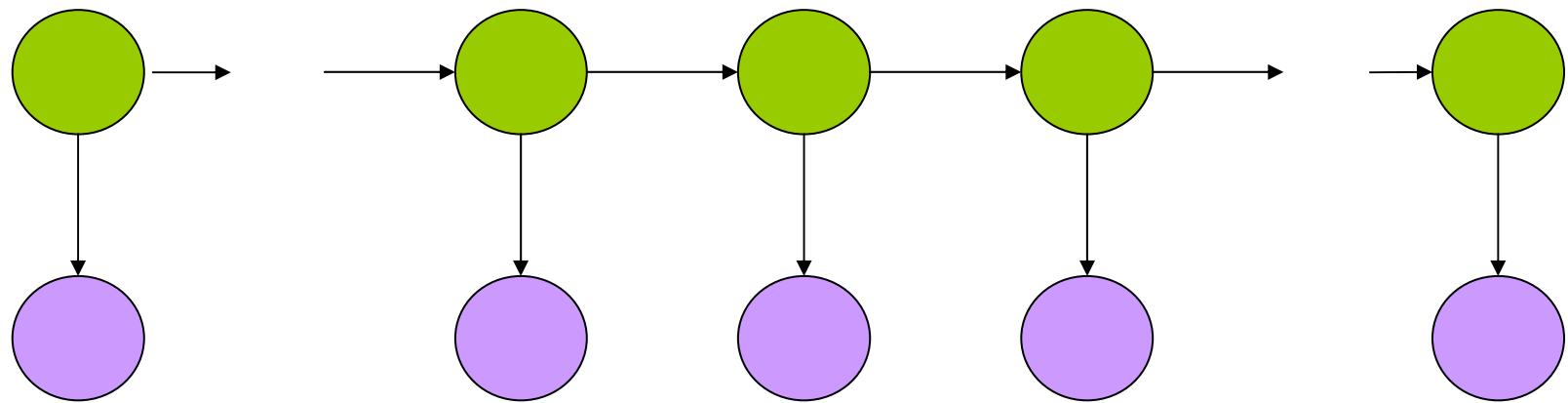


A Markov Model : Weather





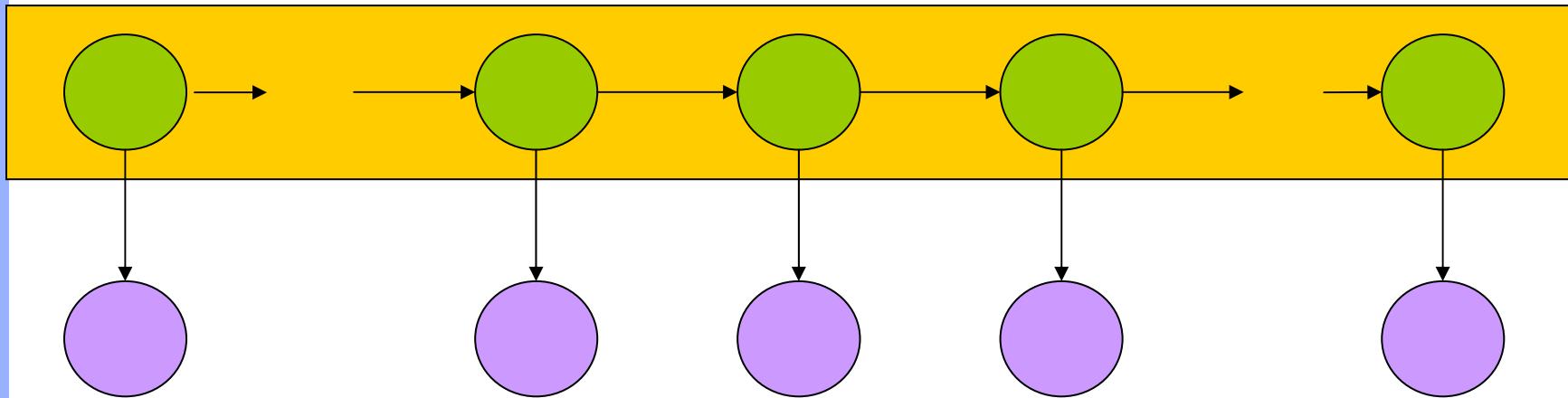
What is an HMM?



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

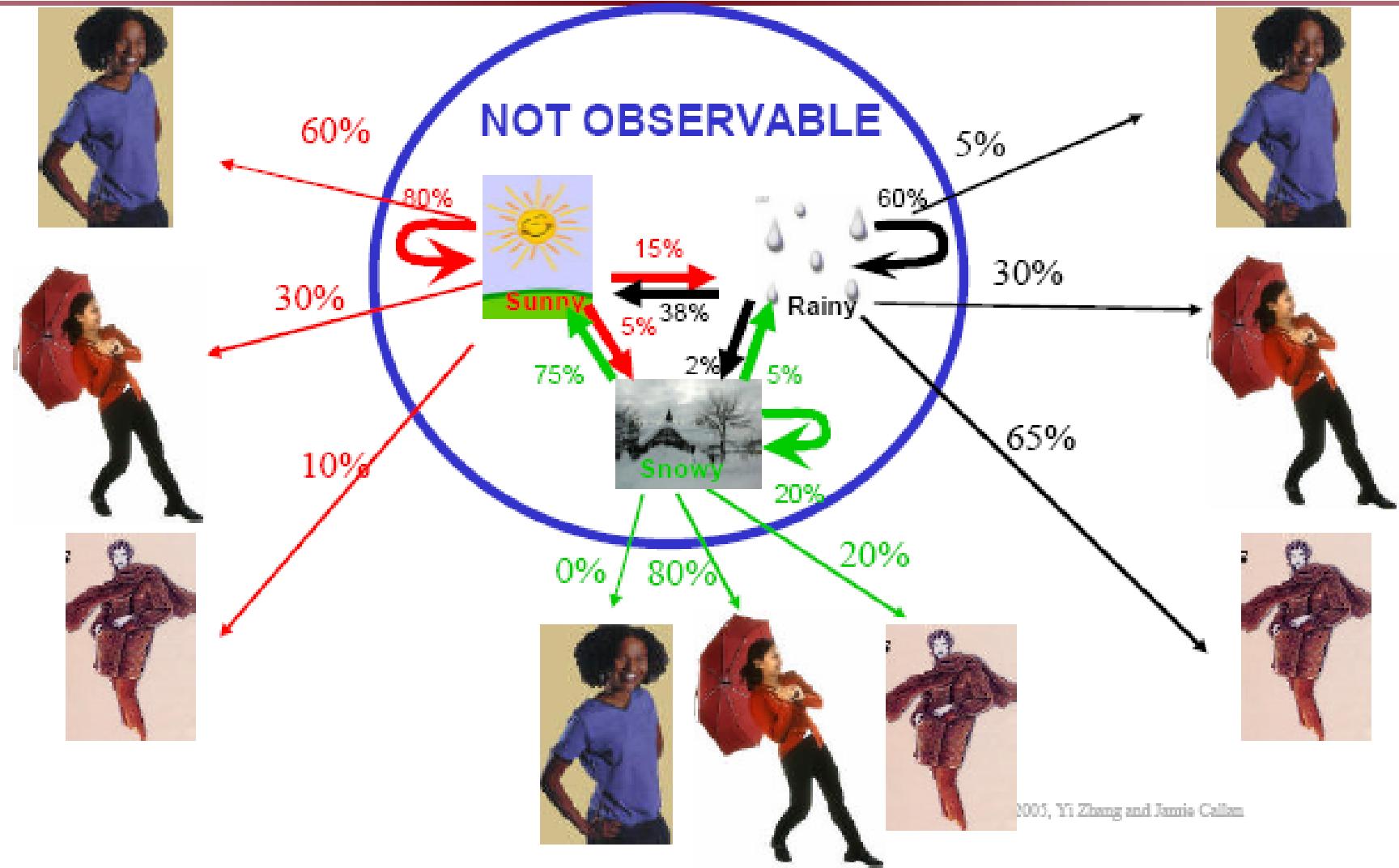


What is an HMM?



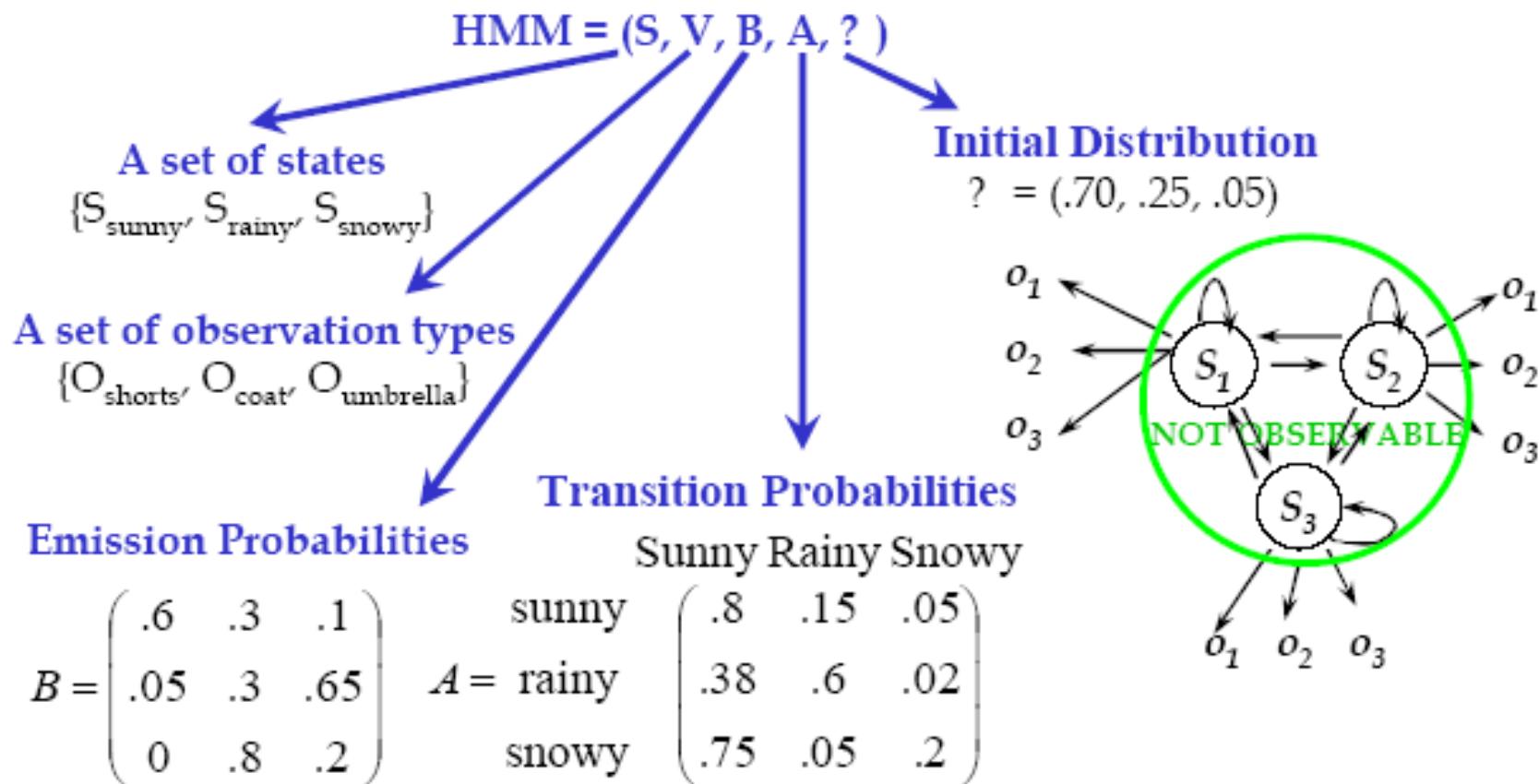
- Green circles are *hidden states*
- Dependent only on the previous state: Markov process
- “The past is independent of the future given the present.”

Hidden Markov Models: Inferring The Weather From What People Wear

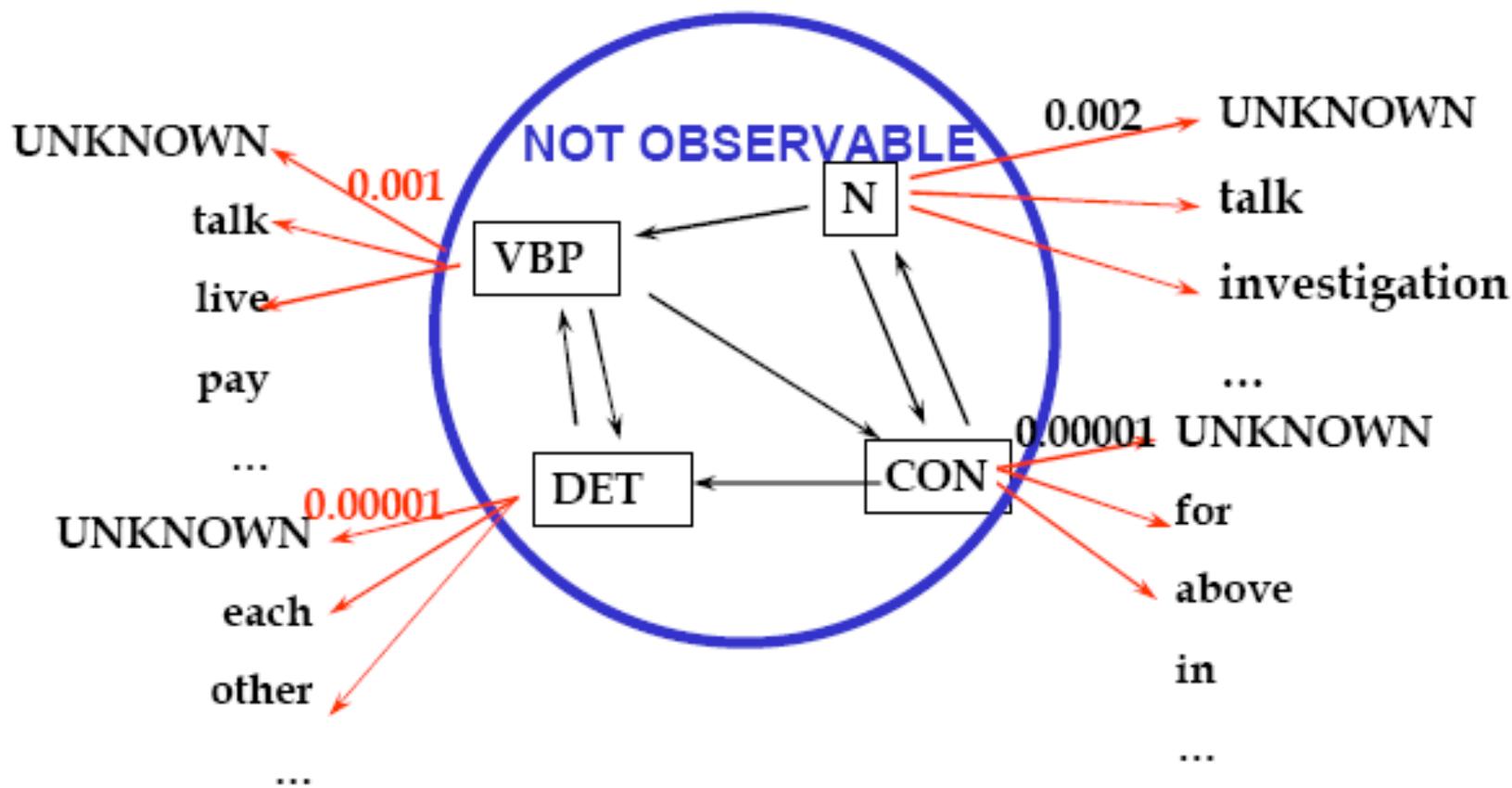


2005, Yi Zhang and Jamie Callan

Hidden Markov Models: Inferring The Weather From What People Wear



Hidden Markov Models of a Simple Part of Speech Tagger



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

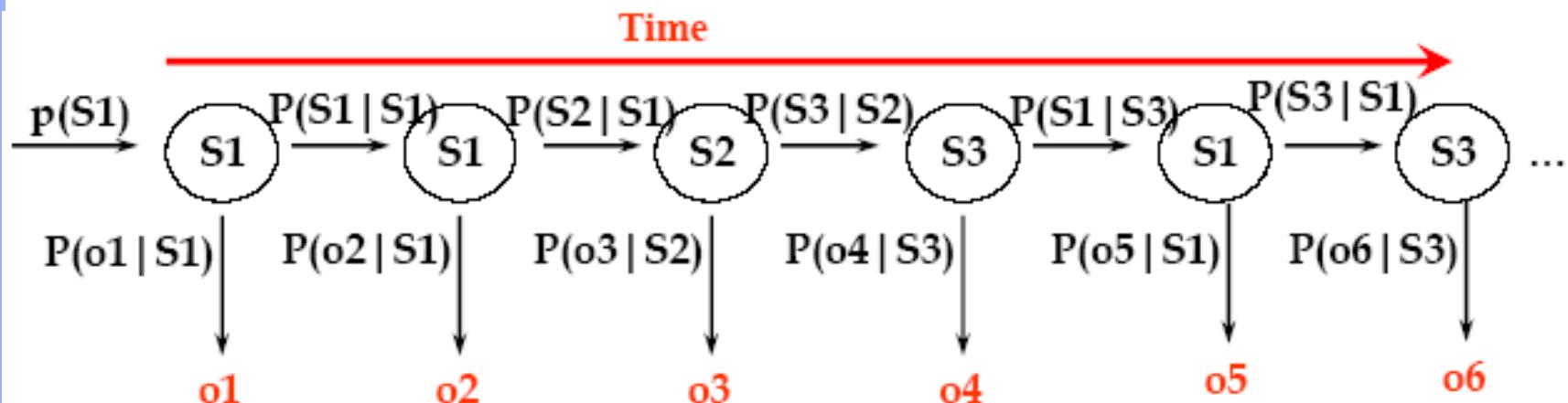
Initial distribution: $(N, VBP, CON, DET) = (0.1, 0.2, 0.1, 0.6)$

© 2005, Yi Zhang and Jamie Callan



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

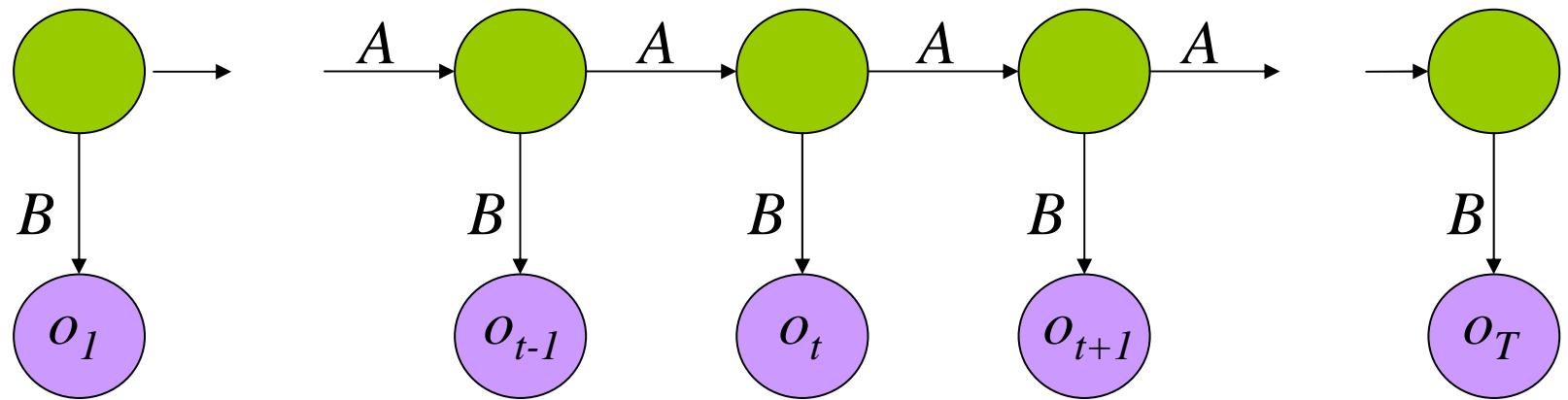


Probability (state sequence, observation sequence)

$$= p(S1) P(o1|S1) P(S1|S1) P(o2|S1) P(S2|S1) \dots$$

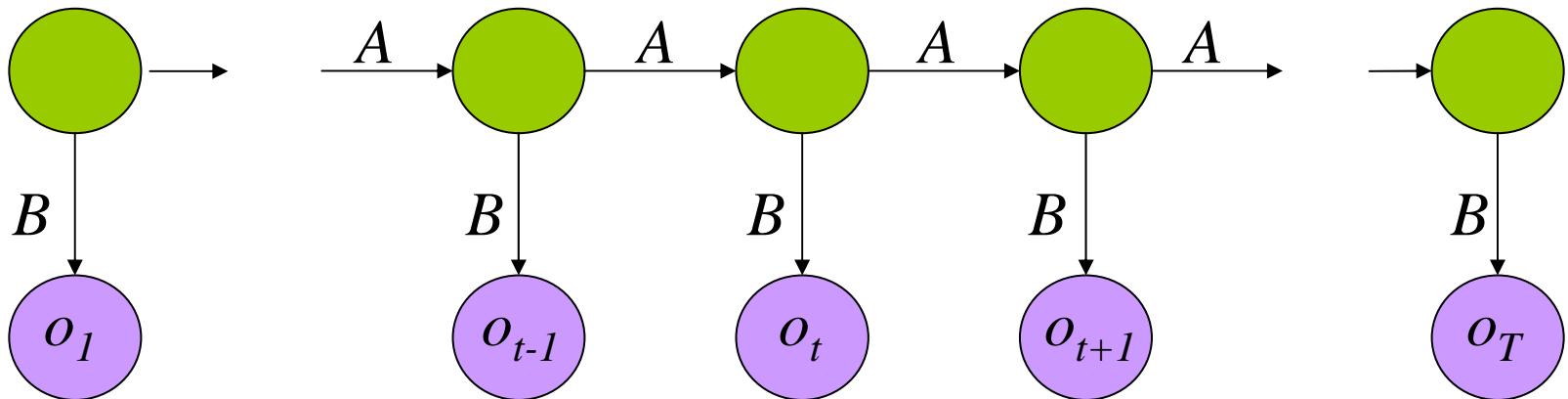


Learning = Parameter Estimation



- 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



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

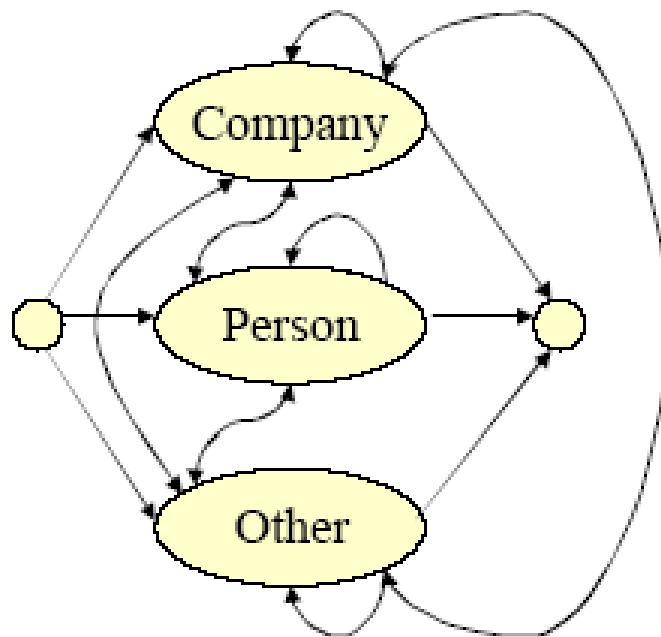
Probability of
traversing an arc

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

Probability of
being in state i



Example: Name Entity Extraction



P (w | Company)

Apple	0.0100
apple	0.0001
Clinton	0.0001

: :

P (w | Person)

Apple	0.00010
apple	0.00001
Clinton	0.01000

: :

Emission probabilities for S_{company}

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

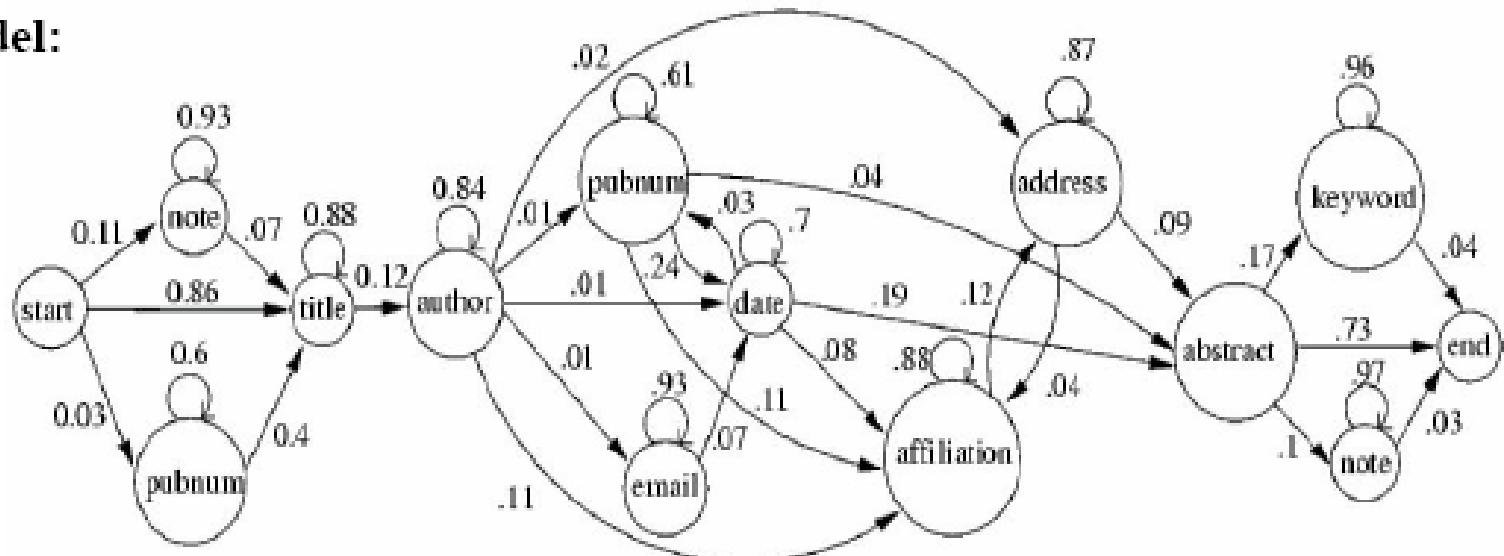
Andrew McCallum[‡]
mccallum@justresearch.com

Ronald Rosenfeld[†]
roni@cs.cmu.edu

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

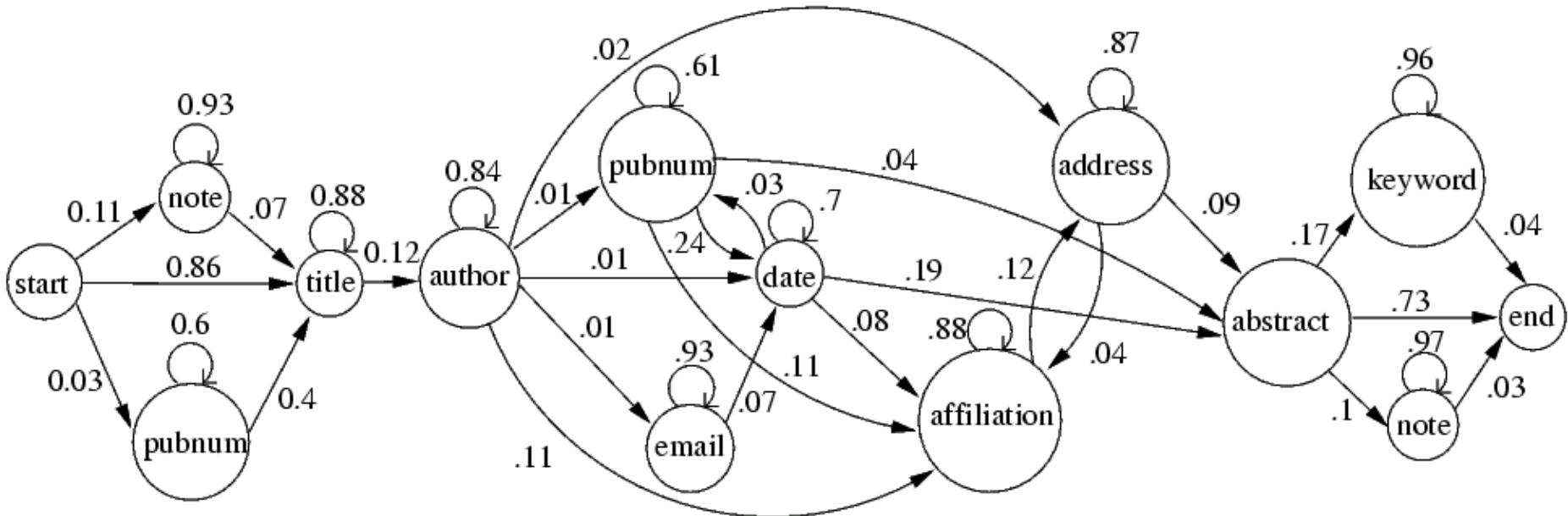
[‡]Just Research
4616 Henry Street
Pittsburgh, PA 15213

Model:





Example: transitions [Seymore *et al.*, 99]



Boosted Wrapper Induction

<p>Dayne Freitag Just Research Pittsburgh, PA, USA dayne@cs.cmu.edu</p>	<p>Nicholas Kushmerick Department of Computer Science University College Dublin, Ireland nich@cs.tcd.ie</p>
--	--

Abstract

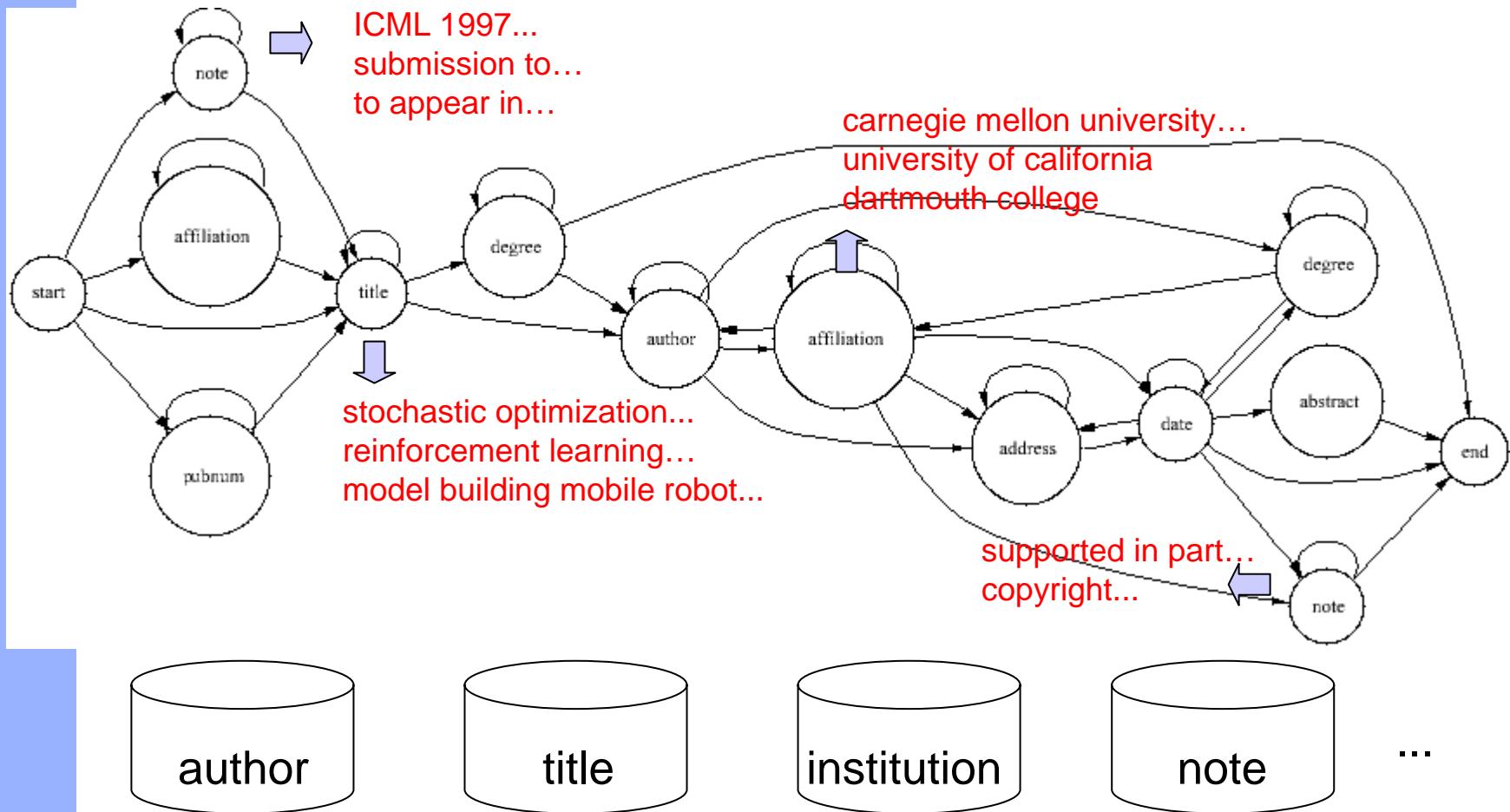
Recent work in machine learning for *information extraction* has focused on two distinct sub-problems: the computational problem of filling template slots from natural language texts, and the problem of *wrapper induction*, which extracts structured extractors ("wrappers") for highly structured text such as Web pages generated by CGI scripts. For the first regularities exist, existing learning algorithms can efficiently learn wrappers that are simple and highly accurate, but the regularity bias of these algorithms makes them unsuitable for most common information extraction tasks. In this paper we report on the performance of a simple machine learning algorithm by repeatedly applying it to our training set with different example weightings. We describe an algorithm that learns simple low-level contextual extraction patterns, which we then apply to conventional information extraction problems using boosting. The results is state-of-the-art in terms of both precision and recall, with a strong precision bias and F1 performance better than state-of-the-art techniques in many domains.

Introduction

Information Extraction (IE) is a key component of many systems that process unstructured text. IE systems typically take a document as input and extract specific pieces of information from it. These pieces of information are often represented as structured data, such as tables or XML documents. IE systems can be used for a variety of applications, such as news aggregation, spam filtering, and search engines.



Example: emissions [Seymore *et al.*, 99]

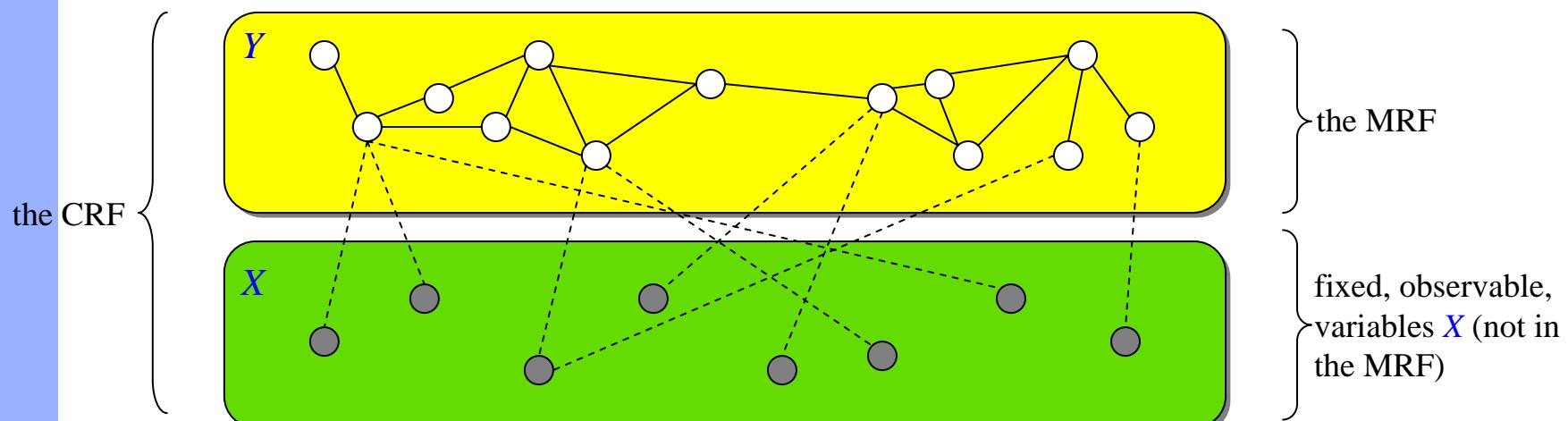


Trained on 2 million words of BibTeX data from the Web



其他学习方法

- A CRF (Conditional Random Fields) is a Markov random field of unobservables which are globally conditioned on a set of observables





小结

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



Any Question?