

Trích rút thông tin

Lê Thanh Hương
Viện CNTT & TT – Trường ĐHBKHN
Email: huonglt@soict.hust.edu.vn

XLNNTN trong tìm kiếm thông tin

- IR ít quan tâm đến ngữ nghĩa, vd:
 - Tìm “Micheal Jordan” (cầu thủ bóng rổ, nhà nghiên cứu về học máy)
 - Tìm “laptop”, không tìm “notebook”
- Có nhiều cải tiến dựa trên phân tích liên kết
- Phân tích liên kết là một dạng của phân tích thực chứng: con người nghĩ cái gì là chính xác và quan trọng
- Sử dụng trí tuệ con người luôn thắng trí tuệ nhân tạo: con người luôn có thể duyệt trên tập kết quả
- Tập trung vào các truy vấn ngắn thông dụng và các bản tin

Ví dụ - Công cụ tìm kiếm

The screenshot shows a Google search interface with the query 'baker job opening'. The search results are displayed in a list format. The first result is a yellow banner for 'Job Opening - Find ANY Job! - Search by Type, Industry & Geography' from careerbuilder.com. The second result is a light blue banner for 'Job Opening At Flipdog.Com' from FlipDog.com. The third result is a blue link to a Softimage discussion group thread titled 'Softimage::Community::Discussion Groups::ds.archive.0004' with a snippet mentioning 'JOB OPENING ... Tony Cacciarelli; RE: ALE Karim Arbaoui; RE: omf to timeline Martin Baker; Re ...'. The fourth result is another blue link to the same Softimage discussion group thread with a snippet mentioning 'Philip Herring - 2000/04/28 22:35. ... RE: omf to timeline Martin Baker - 2000/04/26 17:33; Re: omf to timeline adam - 2000/04/26 18:11. ...'. The fifth result is a blue link to 'CGI: Job Opening' from genomics.cornell.edu. The sixth result is a blue link to 'Information Activist Job Opening - May 2001' from igc.org. The seventh result is a blue link to 'Post an Employee Benefits Job Opening (Help Wanted) Ad' from benefitslink.com. The eighth result is another blue link to 'Post an Employee Benefits Job Opening (Help Wanted) Ad' from benefitslink.com.

Google™ [Advanced Search](#) [Preferences](#) [Language Tools](#) [Search Tips](#)

[Web](#) [Images](#) [Groups](#) [Directory](#) [News-New!](#)

Searched the web for **baker job opening**. Results

[Job Opening - Find ANY Job! - Search by Type, Industry & Geography](#)
[www.careerbuilder.com](#) Post Your RESUME Here to Reach Thousands of Employers - It's FREE!

[Job Opening At Flipdog.Com](#)
[www.FlipDog.com](#) Fetch your next **job** at FlipDog.com!

[Softimage::Community::Discussion Groups::ds.archive.0004](#)
... Le Rudulier; Drive space Ken Skaggs; Help about rendering denis.courtot; **JOB OPENING** ... Tony Cacciarelli; RE: ALE Karim Arbaoui; RE: omf to timeline Martin **Baker**; Re ...
[www.softimage.com/community/xsi/discuss/Archives/ds.archive.0004/default.htm](#) - 49k - [Cached](#) - [Similar pages](#)

[Softimage::Community::Discussion Groups::ds.archive.0004](#)
... Re: **JOB OPENING** Philip Herring - 2000/04/28 22:35. ... RE: omf to timeline Martin **Baker** - 2000/04/26 17:33; Re: omf to timeline adam - 2000/04/26 18:11. ...
[www.softimage.com/community/xsi/discuss/Archives/ds.archive.0004/ThreadIndex.htm](#) - 50k - [Cached](#) - [Similar pages](#)
[[More results from www.softimage.com](#)]

[CGI: Job Opening](#)
[www.genomics.cornell.edu/jobs/view_job.cfm?id=10](#) - 15k - [Cached](#) - [Similar pages](#)

[Information Activist Job Opening - May 2001](#)
[www.igc.org/datacenter/job.html](#) - 6k - [Cached](#) - [Similar pages](#)

[Post an Employee Benefits Job Opening \(Help Wanted\) Ad](#)
... edit the ad to add a new **job opening** ... as possible when it is emailed to 2,985 **job** ... [jobs/posthelpwanted.shtml](#)
- Webmaster: [webmaster@BenefitsLink.com](#) (Dave **Baker** ...
[www.benefitslink.com/jobs/posthelpwanted.shtml](#) - 24k - [Cached](#) - [Similar pages](#)

[Post an Employee Benefits Job Opening \(Help Wanted\) Ad](#)
Employee Benefits Jobs! Brought to you by BenefitsLink (tm) and its EmployeeBenefitsJobs.com (tm) division.
[www.benefitslink.com/jobs/pricinginfo.shtml](#) - 7k - [Cached](#) - [Similar pages](#)
[[More results from www.benefitslink.com](#)]

Martin Baker, a person

Genomics job

Employers job posting form

Ví dụ: một giải pháp

job search find employment careers @ FlipDog.com free! - Microsoft Internet Explorer

Address <http://www.flipdog.com/home.html> Go File Edit View Favorites Tools Help Links



Home Find Jobs Your Account Resource Center Support Employers

Job Search at FlipDog.com: Employment & Career Management



647,514
Job Opportunities
from **53,641** Employers

Find a Job!

Post Your Resume

Employers
click here for
Products & Services



Job Seekers: Find your dream job!

- ▶ Check our 'Best Places to Find a Job' [January report](#).
- ▶ Open your [FREE account](#) and put your [resume online](#).
- ▶ Search 24x7 with our FREE automatic [JobHunters™](#).
- ▶ Research our database of over [50,000 employers](#).
- ▶ Get [expert advice](#) at our new [Resource Center](#).
- ▶ Access [salary surveys/calculators](#), [relocation tools](#), [networking opportunities](#), & [training/testing](#) tools.
- ▶ Use FlipDog.com to search jobs right from your desktop! Download [Snippets](#) today!

Pigskin Places

- Health Care in NY [2,770](#)
- Health Care in MD [1,262](#)
- Sales in NY [3,751](#)
- Sales in MD [958](#)
- Computing in NY [8,050](#)
- Computing in MD [4,114](#)

Jobs for Sports Fans

- [Head Football Coach](#)
- [Football Coach](#)
- [Asst. Football Coach](#)
- [High School Football Coach](#)
- [Univ. Asst. Football Coach](#)

Job Seeker Newsletter

Enter your e-mail address:

[Sign Me Up!](#)

Showcase Jobs



We provide total staffing solutions in the areas of Human Resources, Compensation, Web-based HR self-service, and Customer Management Systems.

[Learn More](#)



Looking for a Vice President of Academic Affairs to oversee planning, operation and evaluation of the college's academic programs.

[Learn More](#)

powered by **WhizBang!**



"Top 100 Web Sites"
PC Magazine, Nov. 2000



"Top 10 Career Web Site"
Media Matrix, Sept. 2000



"Top 10 Job Site"

Internet

Start Microsoft PowerPoint - [sta... job search find employem... 12:12 AM

Trích rút thông tin về quảng cáo việc làm từ Web

OPUS International, Inc., an executive search firm focusing on the Food Science industry. - Microsoft Internet Explorer

File Edit View Favorites

Back Forward Stop

Address <http://www.foodscier>

Links AMEX Rewards T

OPUS: Job Listings - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorite

Address http://www.foodscience.com/jobs_midwest.html#top

Links AMEX Rewards Time DogHouse My Yahoo!

Welcome

About OPUS

Executive Staff

Job Listings

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

OPUS INTERNATIONAL INC.

About | Staff | Job

Welcome

About OPUS

Executive Staff

Job Listings

Résumé Form

Job Hunt Hints

Academic Links

Science Fair Help

Industry Assocs.

FAQs

Contact Us

Site Map

e-mail

**Test Kitchen-
Consumer Food Relation**

Major food manufacturer in Chicago area seeks a con food professionals to write recipes. Will make presenta marketing; will be a key pla a cross-functional team. Re BS in human ecology, nutrit Food Science, or related fie a minimum three years' exp experience.
Contact: Moira: e-mail
1-800-488-2611

Ice Cream Guru

If you dream of cold creamy chocolate or coochy boochy cookie, there's a great opportunity for you to maintain and expand this major corporation's high-end ice cream brand. Will be based in the Upper Midwest for about a year. After that, California here I come! Requires a BS in Food Science or dairy, plus ice cream formulation experience. Will consider entry level with an MS and an internship.
Contact: Susan: e-mail
1-800-488-2611

foodscience.com-Job2

JobTitle: Ice Cream Guru

Employer: foodscience.com

JobCategory: Travel/Hospitality

JobFunction: Food Services

JobLocation: Upper Midwest

Contact Phone: 800-488-2611

DateExtracted: January 8, 2001

Source: www.foodscience.com/jobs_midwest.htm

OtherCompanyJobs: foodscience.com-Job1



Job Openings: Category = Food Services Keyword = Baker Location = Continental U.S.

FlipDog.com Fetch Your Next Job Here™

Home Find Jobs Your Account Resource Center

Return to Results | Modify Search | New Search

The University Alliance A BISK EDUCATION NETWORK
Degrees Online

Learn While You Earn **MBA, BA, AA Degrees** Online & **Project Mgt.**

Click here to e-mail your resume to 1000's of Head Hunters with **ResumeZapper.com**

Breakthrough ebook shows why most people are **WRONG** about how to apply for jobs.

how to easily **DOUBLE your chances** when applying **FOR JOBS!**

1 - 25 of 47 jobs shown below 1 2 Next >

Search these results for: GO! [Search tips](#) **Show Jobs Posted:** For all time periods

View: [Brief](#) | [Detailed](#)

Web Jobs: FlipDog technology has found these jobs on thousands of employer Web sites.

Food Pantry Workers at Lutheran Social Services	October 11, 2002	Archbold, OH
Cooks at Lutheran Social Services	October 11, 2002	Archbold, OH
Bakers Assistants at Fine Catering by Russell Morin	October 11, 2002	Attleboro, MA
Baker's Helper at Bird-in-Hand	October 11, 2002	United States
Assistant Baker at Gourmet To Go	October 11, 2002	Maryland Heights, MO
Host/Hostess at Sharis Restaurants	October 10, 2002	Beaverton, OR
Cooks at Alta's Rustler Lodge	October 10, 2002	Alta, UT
Line Attendant at Sun Valley Coporation	October 10, 2002	Huntsville, UT
Food Service Worker II at Garden Grove Unified School District	October 10, 2002	Garden Grove, CA
Night Cook / Baker at SONOCO	October 10, 2002	Houma, LA
Cooks/Prep Cooks at GrandView Lodge	October 10, 2002	Nisswa, MN
Line Cook at Lone Mountain Ranch	October 10, 2002	Big Sky, MT
Production Baker at Whole Foods Market	October 08, 2002	Willowbrook, IL
Cake Decorator/Baker at Mandalay Bay Hotel and Casino	October 08, 2002	Las Vegas, NV
Shift Supervisors at Brueggers Bagels	October 08, 2002	Minneapolis, MN

Trích rút thông tin

- Các hệ thống Trích rút thông tin:
 - Tìm và hiểu một số phần trong văn bản
 - Các thông tin rõ ràng (who did what to whom when?)
 - Xây dựng một cách biểu diễn có cấu trúc các thông tin liên quan, như các quan hệ trong CSDL
 - Kết hợp tri thức về ngôn ngữ và miền ứng dụng
 - Tự động trích rút các thông tin mong muốn
- Vd
 - Thu thập thông tin về lợi nhuận từ các báo cáo của công ty
 - Học các tương tác giữa thuốc và gen từ các nghiên cứu y học
 - Tạo ra các thẻ thông minh “Smart Tags” (Microsoft) trong các tài liệu

Quảng cáo nhà đất

- Các quảng cáo ở dạng văn bản
- Thêm các thẻ cơ bản: chỉ 70+ từ báo với 20+ nhà xuất bản có thể làm được

<ADNUM> 2067206v1

</ADNUM>

<DATE>March, 02 </DATE>

<ADTITLE> MADDINGTON

\$89,000</ADTITLE>

<ADTEXT>OPEN 1.00-

1.45
 U 11/10 BERTRAM

ST
 NEW TO MARKET

Beautiful
 3brm

freestanding
 villa, close

to shops & bus
 ideally

suit 1st home

buyer,
investor & 55 and

over.
 </ADTEXT>

Tại sao các công cụ tìm kiếm tài liệu không làm được

- Tìm thông tin về quảng cáo nhà đất :
 - Vị trí:
 - Các cụm từ: only 45 minutes from Parramatta
 - Giá: \$120K < M < \$200K
 - Nhiều giá: trước \$155K, giờ \$145
 - Số phòng ngủ: các từ đồng nghĩa (br, bdr, beds, B/R)

Trích rút thông tin

Nhiệm vụ:

Lấy thông tin từ văn bản và điền vào CSDL

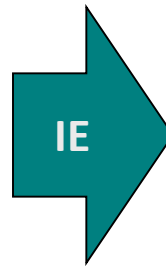
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



<u>NAME</u>	<u>TITLE</u>	<u>ORGANIZATION</u>
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

“Trích rút thông tin” là gì?

Là 1 họ các
công cụ:

Information Extraction =
segmentation + classification + clustering + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Microsoft Corporation

CEO

Bill Gates

Microsoft

Gates

Microsoft

Bill Veghte

Microsoft

VP

Richard Stallman

founder

Free Software Foundation

“named entity
extraction”

“Trích rút thông tin” là gì?

Là 1 họ các
công cụ:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, [Microsoft Corporation](#) [CEO](#) [Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, [Microsoft](#) claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. [Gates](#) himself says [Microsoft](#) will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft](#) [VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...

[Microsoft Corporation](#)
[CEO](#)

[Bill Gates](#)

[Microsoft](#)
[Gates](#)

[Microsoft](#)
[Bill Veghte](#)

[Microsoft](#)
[VP](#)

[Richard Stallman](#)
[founder](#)

[Free Software Foundation](#)

“Trích rút thông tin” là gì?

Là 1 họ các
công cụ:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, [Microsoft Corporation](#) [CEO](#) [Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, [Microsoft](#) claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. [Gates](#) himself says [Microsoft](#) will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft](#) [VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...

[Microsoft Corporation](#)
[CEO](#)
[Bill Gates](#)

[Microsoft](#)
[Gates](#)

[Microsoft](#)
[Bill Veghte](#)
[Microsoft](#)
[VP](#)

[Richard Stallman](#)
[founder](#)
[Free Software Foundation](#)

“Trích rút thông tin” là gì?

Là 1 họ các
công cụ:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, [Microsoft Corporation](#) [CEO](#) [Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, [Microsoft](#) claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. [Gates](#) himself says [Microsoft](#) will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft](#) [VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...

* [Microsoft Corporation](#)
[CEO](#)
[Bill Gates](#)

* [Microsoft](#)
[Gates](#)

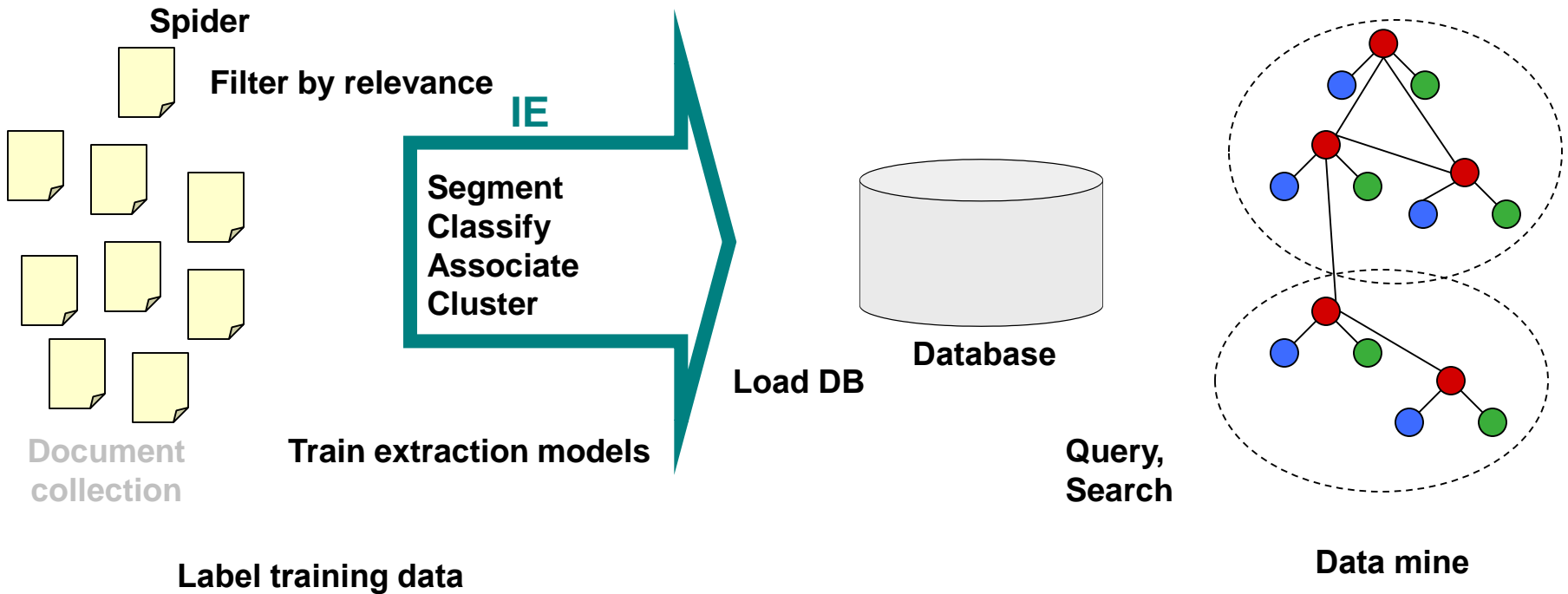
* [Microsoft](#)
[Bill Veghte](#)

* [Microsoft](#)
[VP](#)

[Richard Stallman](#)
[founder](#)
[Free Software Foundation](#)

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft...

Các nội dung của IE



Các khó khăn trong IE (1/4): Định dạng văn bản

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.











Grammatical sentences and some formatting & links

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

- Press
- **Contact**
- General information
- Directions maps

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

Non-grammatical snippets, rich formatting & links

Barto, Andrew G. Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.	(413) 545-2109	barto@cs.umass.edu	CS276	 
Berger, Emery D. Assistant Professor.	(413) 577-4211	emery@cs.umass.edu	CS344	 
Brock, Oliver Assistant Professor.	(413) 577-0334	oli@cs.umass.edu	CS246	 
Clarke, Lori A. Professor. Software verification, testing, and analysis; software architecture and design.	(413) 545-1328	clarke@cs.umass.edu	CS304	 
Cohen, Paul R. Professor. Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.	(413) 545-3638	cohen@cs.umass.edu	CS278	 

Tables

8:30 - 9:30 AM	Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty <i>Joseph Y. Halpern, Cornell University</i>				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
Cognitive Robotics	Logic Programming	Natural Language Generation	Complexity Analysis	Neural Networks	Games
739: A Logical Account of Causal and Topological Maps <i>Emilio Remolina and Benjamin Kuipers</i>	116: A-System: Problem Solving through Abduction <i>Marc Denecker, Antonis Kakas, and Bert Van Nuffelen</i>	758: Title Generation for Machine-Translated Documents <i>Rong Jin and Alexander G. Hauptmann</i>	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories <i>Marco Cadoli, Thomas Eiter, and Georg Gottlob</i>	179: Knowledge Extraction and Comparison from Local Function Networks <i>Kenneth McGarry, Stefan Wermter, and John MacIntyre</i>	71: Iterative Widening <i>Tristan Cazenave</i>
549: Online-Execution of ccGolog Plans <i>Henrik Grosskreutz and Gerhard Lakemeyer</i>	131: A Comparative Study of Logic Programs with Preference <i>Torsten Schaub and Kewen</i>	246: Dealing with Dependencies between Content Planning and Surface Realisation in a Pipeline Generation	470: A Perspective on Knowledge Compilation <i>Adnan Darwiche and Pierre Marquis</i>	258: Violation-Guided Learning for Constrained Formulations in Neural-Network Time-Series	353: Temporal Difference Learning Applied to a High Performance Game-Playing

Các khó khăn trong IE (2/4): Miền dữ liệu xử lý

Web site specific

Formatting

Amazon.com Book Pages

The screenshot shows two overlapping Amazon.com book pages. The top page is for 'Machine Learning by Tom M. Mitchell' and the bottom page is for 'Learning in Graphical Models by Michael Irwin Jordan (Editor)'. The pages exhibit various formatting problems, including overlapping text, inconsistent button styles, and cluttered layouts.

Genre specific

Layout

Resumes

The screenshot displays two resumes side-by-side. The top resume is for Jason D. M. Rennie, and the bottom one is for L. Douglas Baker. Both resumes show inconsistent formatting, such as misaligned text, overlapping information, and unclear section headers.

Wide, non-specific

Language

University Names

The screenshot shows a university schedule and contact information page. The schedule table lists various sessions and speakers, while the contact section provides details for Dr. Steven Minton and Frank Huybrechts. The page contains a mix of English and non-English text, illustrating challenges in language and university name extraction.

8:30 - 9:30 AM	Invited Talk: Plausibility Measures: A General Approach <i>Joseph Y. Halpern, Cornell University</i>		
9:30 - 10:00 AM	Coffee Break		
10:00 - 11:30 AM	Technical Paper Sessions:		
Cognitive Robotics	Logic Programming	Natural Language Generation	Complexity Analysis
739: A Logical Account of Causal and Topological Maps <i>Emilio Remolina and Benjamin Kuipers</i>	116: A-System: Solving Problem through Abduction <i>Marc Denecker, Antonis Kakas, and Bert Van</i>	758: Title Generation for Machine-Translated Documents <i>Rong Jin and Alexander G. Hauptmann</i>	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories <i>Marco Cadoli, W</i>

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

- Press
- General information
- Directions maps

Các khó khăn trong IE (3/4): Độ phức tạp

E.g. word patterns:

Closed set

U.S. states

He was born in Alabama...

The big Wyoming sky...

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Ambiguous patterns, needing context and many sources of evidence

Person names

...was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

Các khó khăn trong IE (4/4): Trường dữ liệu/bản ghi

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

Trích rút thực thể (“Named entity” extraction)

Đánh giá hệ thống trích rút thực thể

Đúng:

Michael Kearns and Sebastian Seung will start Monday's tutorial, followed by Richard M. Karpe and Martin Cooke.

Dự đoán:

Michael Kearns and Sebastian Seung will start Monday's tutorial, followed by Richard M. Karpe and Martin Cooke.

$$\text{Precision} = \frac{\text{\# correctly predicted segments}}{\text{\# predicted segments}} = \frac{2}{6}$$

$$\text{Recall} = \frac{\text{\# correctly predicted segments}}{\text{\# true segments}} = \frac{2}{4}$$

$$\text{F1} = \text{Harmonic mean of Precision \& Recall} = \frac{1}{((1/P) + (1/R)) / 2}$$

Các kết quả trên thế giới

- Nhận dạng thực thể từ các bản tin
 - Person, Location, Organization, ...
 - $85\% \leq F1 \leq 95\%$
- Trích rút quan hệ giữa các thực thể
 - Contained-in (Location1, Location2)
 - Member-of (Person1, Organization1)
 - $60\% \leq F1 < 90\%$

Trích rút thông tin

- **Nhận dạng thực thể (Named Entity Recognition):** định vị và phân loại các thành phần đơn vị trong văn bản thành các loại được định nghĩa trước như tên riêng (tên người, tổ chức, nơi chốn), thời gian, ...
- **Trích rút quan hệ (Relation Extraction):** trích rút mối quan hệ giữa các thực thể

Nhận dạng thực thể

Vào: văn bản chưa gán nhãn, tập nhãn

Ra: văn bản đã gán nhãn

VD:

Hi. My name is **<Person>** Hang Dinh **</Person>**. I am currently attending the **<Domain>** Computer Science **</Domain>** PhD program at the **<University>** University of Connecticut **</University>**.

Nhận dạng thực thể

- Hướng tiếp cận
 - Dùng luật thủ công: Quan sát qui luật của dữ liệu
 - Ưu điểm: Độ chính xác cao
 - Nhược điểm: không xử lý được trường hợp chưa đề cập trong luật.
 - Sinh luật dựa trên học máy : học để tạo mô hình phân loại dữ liệu từ dữ liệu mẫu.
 - Ưu điểm: đáp ứng được tập dữ liệu mới
 - Nhược điểm: cần tập dữ liệu lớn đã gán nhãn

NER - Luật tạo thủ công

- ***Biểu diễn luật:*** Contextual Pattern → Action
- Mẫu nhận dạng gồm các mẫu gán nhãn để lưu các đặc trưng của thực thể và nội dung của nó
- Các đặc trưng của 1 token:
 - từ
 - từ loại
 - định dạng từ: viết hoa, số, ...
 - tiền tố, hậu tố, ...
- Hành động: gán nhãn thực thể cho 1 chuỗi các token

NER - Luật tạo thủ công

- Các luật NER có 3 dạng:
 - Nội dung trước 1 thực thể
 - Nội dung trong 1 thực thể
 - Nội dung sau 1 thực thể

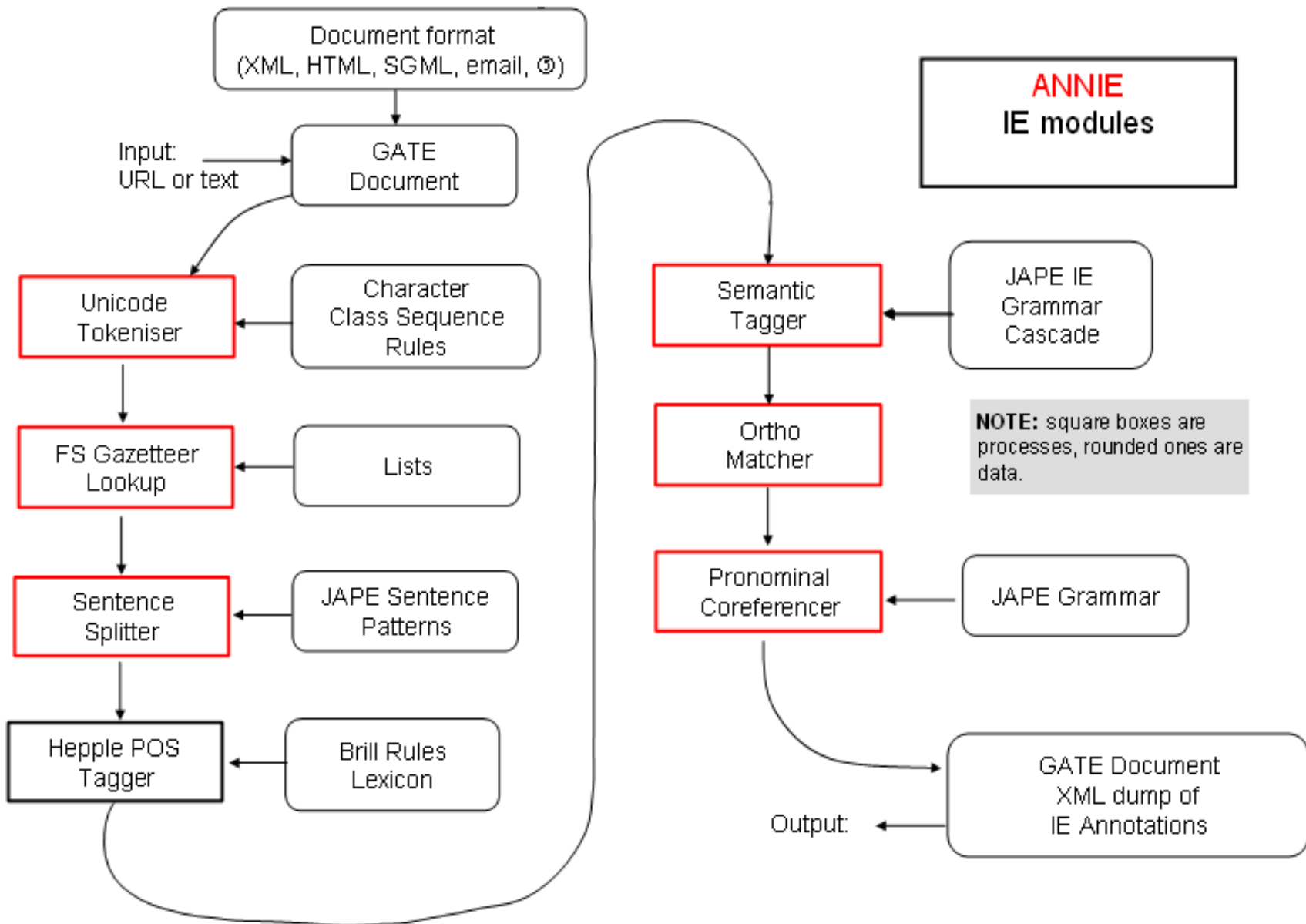
VD:

- “Dr. Peter”
 - ({DictionaryLookup = Titles}{String = “.”}{Orthography type = capitalized word}) → Person Name.
 - Từ điển Titles gồm các từ “Prof”, “Dr”, “Mr”, ...
- “The XYZ Corp.” hoặc “ABC Ltd.”
 - ({String=“The”}? {Orthography type = All capitalized}
 - {Orthography type = Capitalized word, DictionaryType =
 - Company end}) → Company name.

GATE

- GATE - General Architecture for Text Engineering
- GATE hỗ trợ các nhà phát triển phần mềm trên 3 khía cạnh:
 - Kiến trúc phần mềm
 - Bộ khung
 - Môi trường phát triển phần mềm
- GATE có 3 dạng tài nguyên, gọi là CREOLE (Collection of REusable Object for Language Engineering).
 - tài nguyên ngôn ngữ (Language Resource)
 - tài nguyên xử lý (Processing Resource)
 - tài nguyên hiển thị (Visual Resource)

Kiến trúc IE trong GATE



Some NER rules in GATE

Rule: TheGazOrganization

Priority: 50

// Matches “The <in list of company names>”

```
( {Part of speech = DT | Part of speech = RB} {DictionaryLookup = organization})  
→ Organization
```

Rule: LocOrganization

Priority: 50

// Matches “London Police”

```
({DictionaryLookup = location | DictionaryLookup = country} {DictionaryLookup  
= organization} {DictionaryLookup = organization}? ) → Organization
```

Rule: INOrgXandY

Priority: 200

// Matches “in Bradford & Bingley”, or “in Bradford & Bingley Ltd”

```
( {Token string = “in”} )
```

```
({Part of speech = NNP}+ {Token string = “&”} {Orthography type =  
upperInitial}+ {DictionaryLookup = organization end}? ):orgName → Organiza-  
tion=:orgName
```

Rule: OrgDept

Priority: 25

// Matches “Department of Pure Mathematics and Physics”

```
({Token.string = “Department”} {Token.string = “of”} {Orthography type = up-  
perInitial}+ ({Token.string = ”and”} {Orthography type = upperInitial}+)?) →  
Organization
```



Applications

- ANIE_00040
 - PhD
- Language Resources
 - Nguyen
 - PhD
- Processing Resources
 - JAPE PhD
 - ANIE NE Transducer_00056
 - ANIE POS Tagger_00053
 - ANIE OrthoMatcher_00059
 - ANIE English Tokeniser_00042
 - ANIE Sentence Splitter_00050
 - ANIE Gazetteer_0004A
 - Document Reset PR_00041
- Data stores

Messages PhD

Loaded Processing resources

Name	Type
ANIE NE Transducer_00056	ANIE NE Transducer



Selected Processing resources

!	Name	Type
	Document Reset PR_00041	Document Reset PR
	ANIE Gazetteer_0004A	ANIE Gazetteer
	ANIE Sentence Splitter_00050	ANIE Sentence Splitter
	ANIE English Tokeniser_00042	ANIE English Tokeniser
	ANIE OrthoMatcher_00059	ANIE OrthoMatcher
	ANIE POS Tagger_00053	ANIE POS Tagger
	PhD	Jape Transducer



Corpus: PhD

The **corpus** and **document** parameters are not available as they are automatically set by the controller!

No selected processing resource

Name	Type	Required	Value

Run

Serial Application Editor Initialisation Parameters



ATE

- Applications
 - PhD
 - ANNIE_00040
- Language Resources
 - Nguyen
 - PhD
- Processing Resources
 - PhD
 - ANNIE OrthoMatcher_00059
 - ANNIE NE Transducer_00056
 - ANNIE POS Tagger_00053
 - ANNIE Sentence Splitter_0005
 - ANNIE Gazetteer_0004A
 - ANNIE English Tokeniser_000

MimeType: text/html

gate.SourceURL: file:/D:/JAV

Document Editor: Initialisation Parameters

Messages PhD PhD Nguyen

Annotation Sets Annotations List Co-reference Editor Text Push in DB

Pham Hong Nguyen
 PhD student
 Working Group for Data Mining of Natural Language
 Basser Department of Computer Science.
 Madsen Building Room 38
 University of Sydney NSW 2006 Australia
 Email: pham@cs.usyd.edu.au
 Home page: www.cs.usyd.edu.au/pham
 Phone +61 2 9351 4174
 Fax +61 2 9351 3838

Research Interests: Natural Language Processing, AI, Compiler.

More...

- Lookup
- PhD_Address
- PhD_Country
- PhD_Domain
- PhD_Email
- PhD_Person
- PhD_Phone
- PhD_University
- PhD_Web
- Sentence
- SpaceToken
- Token
- Original markups

Type	Set	Start	End	Id	Features
PhD_Person		0	16	227	{rule=Person2}
PhD_Domain		47	58	233	{rule=Domain}

11 Annotations (0 selected)

Du lịch Hạ Long 1 Ngày



khởi hành từ Hà Nội

Thời gian: 1 Ngày

Giá tour: 695.000đ

Giá KM: 599.000đ

Phương tiện: Ôtô + thuyền

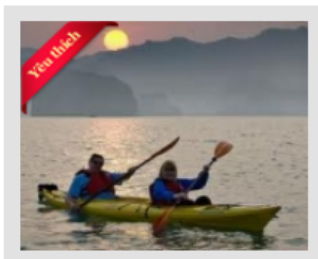
Khởi hành ngày: Hàng ngày

Giới thiệu tour: Hành trình du lịch Hạ Long 1 Ngày từ Hà Nội sẽ cùng quý khách đến với kỳ quan thiên nhiên thế giới tại Việt Nam. Từ trên cao nhìn xuống Vịnh Hạ Long như một bức tranh thủy mặc khổng lồ vô cùng sống động. Đó là những tác phẩm tạo hình tuyệt mỹ, tài hoa của tạo hoá, của thiên nhiên ...

[Đặt tour](#)

[xem tiếp](#)

Du Lịch Hạ Long 2 Ngày (Ngủ Đêm Trên Du Thuyền 3 Sao Halong Dolphin)



khởi hành từ Hà Nội

Thời gian: 2 Ngày 1 Đêm

Giá tour: 2.650.000đ

Giá KM: 1.795.000đ

Phương tiện: Ôtô + Du thuyền

Khởi hành ngày: Hàng ngày

Giới thiệu tour: Hành trình tour du lịch Hạ Long 2 Ngày 1 đêm sẽ đưa quý khách thưởng thức vẻ đẹp kỳ bí của Vịnh Hạ Long trên du thuyền 3 sao Hạ Long Dolphin. Với dáng vẻ của tàu gỗ truyền thống, con tàu dài 32 mét, rộng 8 mét được làm từ chất liệu gỗ tốt nhất, được bao người nghệ nhân dày công chạm khắc. Chuyển đi ...

[Đặt tour](#)

[xem tiếp](#)

Du lịch Hạ Long 3 Ngày (2 Đêm Trên Du Thuyền 3 Sao Halong Dolphin)



khởi hành từ Hà Nội

Thời gian: 3 Ngày 2 Đêm

Giá tour: 3.938.000đ

Giá KM: 2.950.000đ

Phương tiện: Ôtô + thuyền

Khởi hành ngày: Hàng ngày

Giới thiệu tour: Đến với Vịnh Hạ Long như một bức tranh thủy mặc khổng lồ vô cùng sống động. Với tour du lịch Hạ Long 3 Ngày giúp quý khách cảm nhận được những tác phẩm tạo hình tuyệt mỹ, tài hoa của tạo hoá, của thiên nhiên biến hàng ngàn đảo đá vô tri tĩnh lặng kia trở nên những tác phẩm điêu khắc, hội họa ...

[Đặt tour](#)

[xem tiếp](#)

Du lịch Hạ Long - Đảo Cát Bà 3 Ngày (1 đêm ngủ tàu + 1 đêm tại ks trên đảo Cát Bà)



khởi hành từ Hà Nội

Thời gian: 3 Ngày 2 Đêm

Giá tour: 3.570.000đ

Giá KM: 2.956.000đ

Phương tiện: Ô tô + thuyền

Khởi hành ngày: Hàng ngày

Giới thiệu tour: Cát Bà với vẻ đẹp nguyên sơ và hùng vĩ, Cát Bà được mệnh danh là Hòn Ngọc của Vịnh Bắc Bộ. Với tour du lịch Hạ Long Cát Bà 3 ngày 2 đêm này, Du lịch Việt Nam sẽ đưa quý khách đến với đảo Cát Bà - nơi có những bãi tắm mịn màng, phẳng lặng, có vườn Quốc Gia rộng 600 ha tạo ...

Bài tập

Trích rút sự kiện từ đoạn sau:

- Police sources have reported that unidentified individuals planted a **bomb** in front of a **Mormon Church** in **Talcahuano District**. The bomb, which exploded and caused **property damage worth 50,000 pesos**, was placed at a chapel of the Church of Jesus Christ of Latter-Day Saints located at **No 3856 Gomez Carreno Street**.
- Prosecutor Juan Carbone Herrera requested the **25 years imprisonment** for **General Rolando Cabezas Alarcon** of the **Republican Guard** for ordering the shooting of 124 of the San Pedro prison inmates.
- Last night in San Clemente District, 9 km north of Pisco, a group of terrorists dynamited machinery belonging to Albolones Peruanos, Inc.

Cho biết các vấn đề có thể xảy ra khi phân tích từ vựng và NER. Vd:

1. Cho ví dụ các thông tin có trong các câu trên.
2. Cho ví dụ về các tên và các dạng đặc biệt khác . Viết luật để tìm ra chúng.

Bài tập (tiếp)

- Bây giờ sử dụng Wordnet để xử lý đoạn trên
- Đưa các từ ở ví dụ trên vào WordNet xem có thể tận dụng được gì không?
- Các vấn đề khi sử dụng WordNet cho IE?

Bài tập

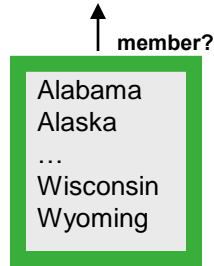
Hãy phát hiện loại thực thể và đề xuất luật nhận dạng thực thể đó:

- Hôm nay, chị **Nguyễn Chi Mai** đi thành phố **Hồ Chí Minh**
- Ông **Võ Nguyên Giáp**
- Công ty TNHH nhà đất **Đại Nam**, **Hà Nội**
- Đường **Tạ Quang Bửu**
- **Andrew Grove** là một giám đốc công ty
- **Vinamilk**, công ty sữa lớn nhất **Việt Nam**, được thành lập năm 1976.

Các kỹ thuật IE: các mô hình

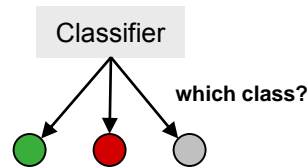
Lexicons

Abraham Lincoln was born in Kentucky.



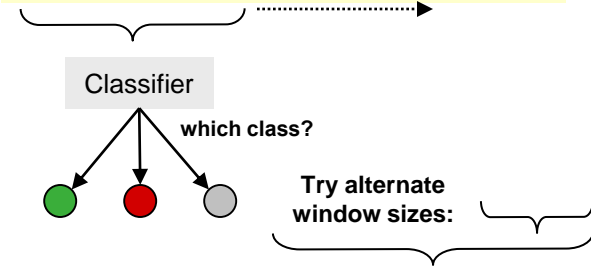
Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.



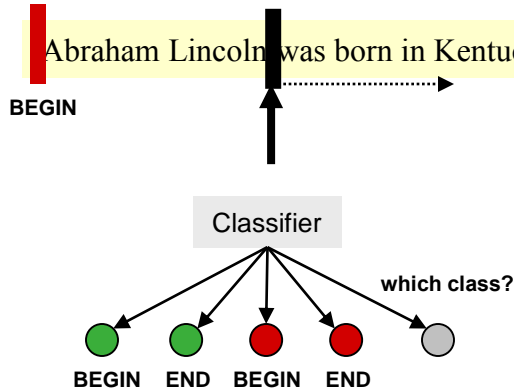
Sliding Window

Abraham Lincoln was born in Kentucky.



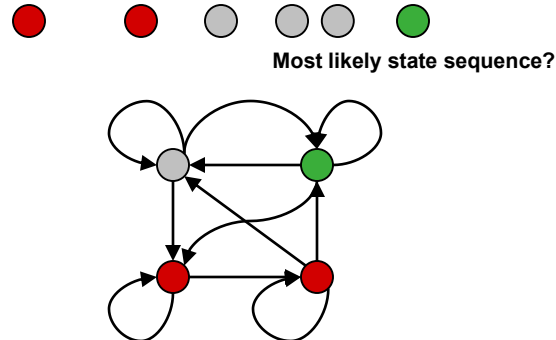
Boundary Models

Abraham Lincoln was born in Kentucky.



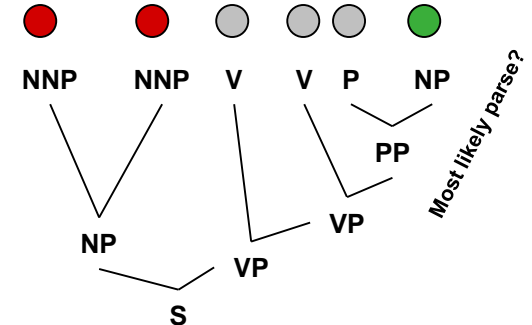
Finite State Machines

Abraham Lincoln was born in Kentucky.



Context Free Grammars

Abraham Lincoln was born in Kentucky.



Any of these models can be used to capture words, formatting or both.

Sliding Windows

Trích rút dùng của sô trượt

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell
School of Computer Science
Carnegie Mellon University

3:30 pm
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**E.g.
Looking for
seminar
location**

CMU UseNet Seminar Announcement

Trích rút dùng của sô trượt

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell
School of Computer Science
Carnegie Mellon University

3:30 pm
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**E.g.
Looking for
seminar
location**

CMU UseNet Seminar Announcement

Trích rút dùng của sổ trượt

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell
School of Computer Science
Carnegie Mellon University

3:30 pm
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

**E.g.
Looking for
seminar
location**

CMU UseNet Seminar Announcement

Trích rút dùng của số trượt

**E.g.
Looking for
seminar
location**

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell
School of Computer Science
Carnegie Mellon University

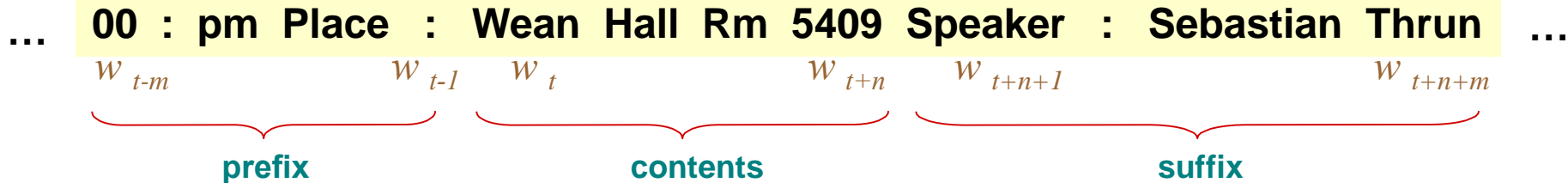
3:30 pm
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

CMU UseNet Seminar Announcement

Mô hình cửa sổ trượt “Naïve Bayes”

[Freitag 1997]



Đánh giá $\Pr(\text{LOCATION}|\text{window})$ sử dụng luật Bayes

Thử tất cả các cửa sổ trượt hợp lý (chiều dài và vị trí thay đổi)

Sử dụng giả thiết độc lập với độ dài, tiền tố, hậu tố, từ nội dung

Đánh giá từ dữ liệu: $\Pr(\text{“Place” in prefix}|\text{LOCATION})$

If $P(\text{“Wean Hall Rm 5409”} = \text{LOCATION}) > \theta$, extract it.

Các phương pháp khác: cây quyết định trên các từ đơn và ngữ cảnh của nó

Mô hình cửa sổ trượt “Naïve Bayes”: kết quả

Domain: CMU UseNet Seminar Announcements

GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell
School of Computer Science
Carnegie Mellon University

3:30 pm
7500 Wean Hall

Machine learning has evolved from obscurity in the 1970s into a vibrant and popular discipline in artificial intelligence during the 1980s and 1990s. As a result of its success and growth, machine learning is evolving into a collection of related disciplines: inductive concept acquisition, analytic learning in problem solving (e.g. analogy, explanation-based learning), learning theory (e.g. PAC learning), genetic algorithms, connectionist learning, hybrid systems, and so on.

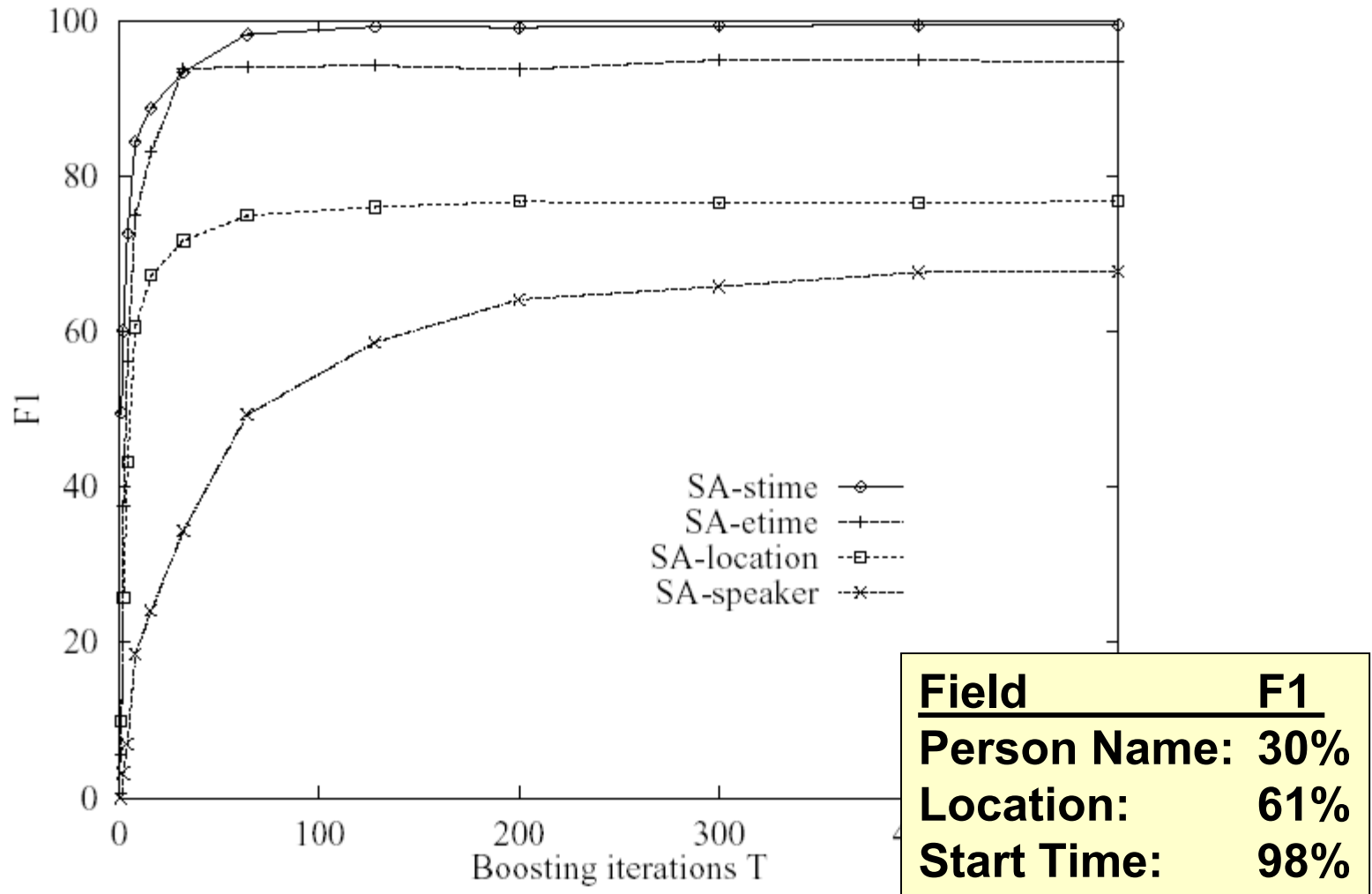
<u>Field</u>	<u>F1</u>
Person Name:	30%
Location:	61%
Start Time:	98%

BWI: Học phát hiện biên

[Freitag & Kushmerick, AAI 2000]

- Học 3 lớp dựa trên xác suất:
 - $START(i) = \text{Prob}(\text{vị trí } i \text{ là bắt đầu một trường})$
 - $END(j) = \text{Prob}(\text{vị trí } j \text{ là kết thúc một trường})$
 - $LEN(k) = \text{Prob}(\text{trường trích rút có độ dài } k)$
- Tính điểm khả năng trích rút (i, j) :
 $START(i) * END(j) * LEN(j-i)$
- $LEN(k)$ được ước lượng dựa trên histogram

BWI: Học phát hiện biên



Các vấn đề với cửa sổ trượt và học phát hiện biên

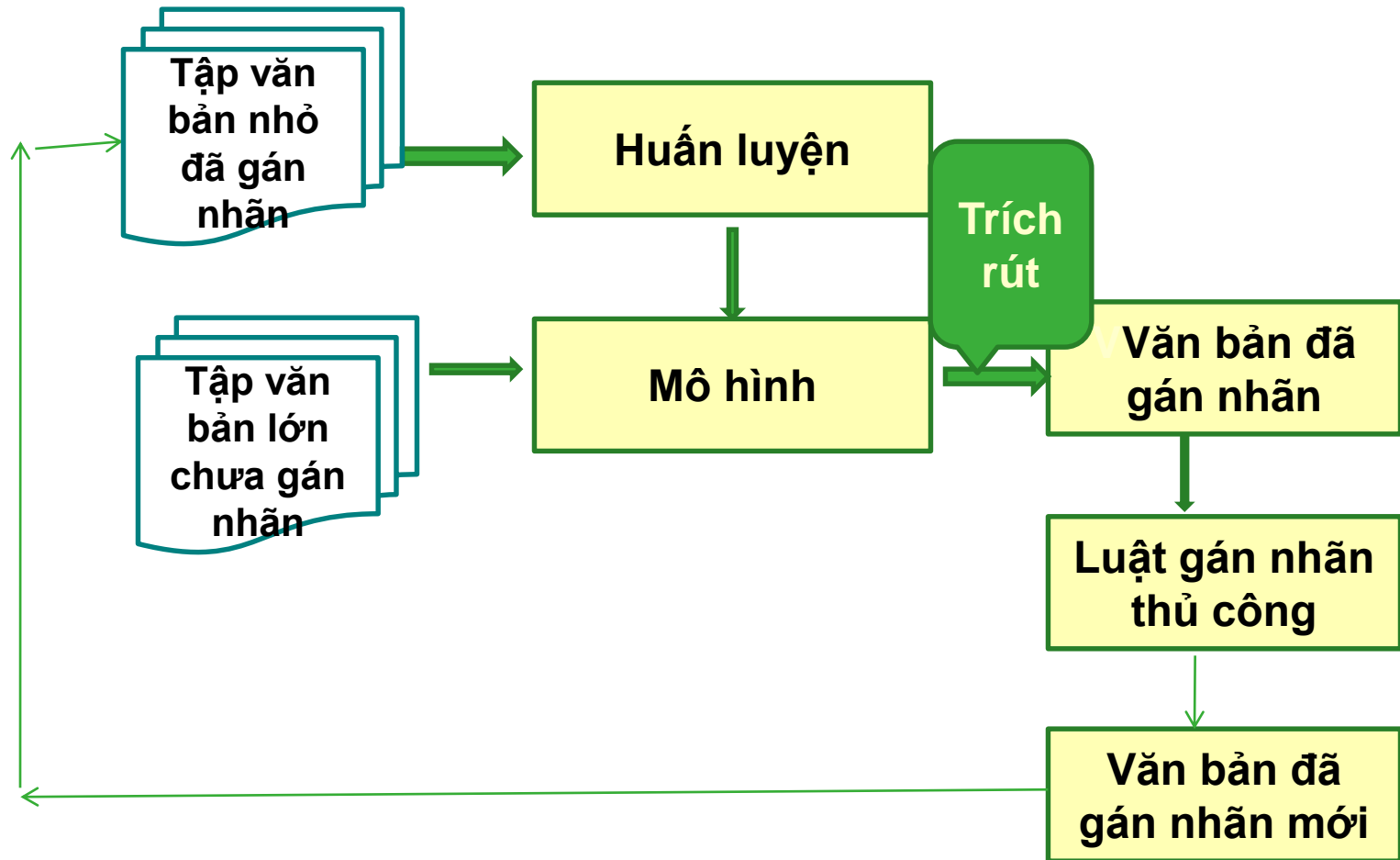
- Các quyết định về các từ bên cạnh độc lập với nhau.
 - Naïve Bayes Sliding Window có thể tiên đoán “seminar end time” trước “seminar start time”.
 - Trong hệ thống tìm biên, bước tìm biên trái độc lập với bước tìm biên phải.

Sam Chanrathany (2014)

Nhận xét

- Hệ thống NER chỉ nhận dạng được kiểu thực thể của dữ liệu có ngữ cảnh trong tập dữ liệu huấn luyện.
- Tên có thể xuất hiện nhiều lần trong văn bản dưới nhiều dạng khác nhau (các tên đồng tham chiếu) → có thể có cùng một kiểu thực thể.
- Các tên đồng tham chiếu này có thể xuất hiện nhiều lần trong văn bản trong các ngữ cảnh khác nhau.

Sam Chanrathany (2014)



Các luật đồng tham chiếu về tên trong văn bản tiếng Việt

Hai tên (N_1 và N_2) là đồng tham chiếu nếu:

1. Hai tên giống nhau
2. Một tên là phần tên của tên còn lại, ví dụ: “*Mai Liêm Trực*” và “*Trực*”.
3. Một tên là bí danh của tên khác, ví dụ: “*Sài Gòn*” và “*TP Hồ Chí Minh*”.
4. Một tên là viết tắt của tên khác, ví dụ: “*IBM*” và “*International Bussiness Machines*”.
5. k chữ đầu và m chữ cuối của hai tên giống nhau, với điều kiện $k + m$ là số chữ của N_2 , ví dụ: “*Công ty Cổ phần Đại An*” và “*Công ty Đại An*”.

Các luật đồng tham chiếu về tên trong văn bản tiếng Việt

6. Ngoại trừ phần tiền tố, tất cả các chữ của N_2 đều xuất hiện trong N_1 và phần tiền tố của N_2 hoặc là giống tiền tố của N_1 hoặc là viết tắt phần tiền tố của N_1 , ví dụ: “*Công ty TNHH Apave Việt Nam*”, “*Cty Apave Việt Nam*”, “*Công ty Apave*” cùng là tên của một công ty.
7. Một tên là phần cuối của tên còn lại, ví dụ: “*Trịnh Chân Trâu*” và “*Chân Trâu*”.
8. Phần cuối của một tên là viết tắt kí tự đầu của các chữ trong phần cuối của tên kia, phần còn lại của hai tên giống nhau, ví dụ, với “*Bộ Giáo dục và Đào tạo*” và “*Bộ GD & ĐT*” thì “*GD & ĐT*” là viết tắt kí tự đầu của “*Giáo dục và Đào tạo*”.

Các luật đồng tham chiếu về tên trong văn bản tiếng Việt

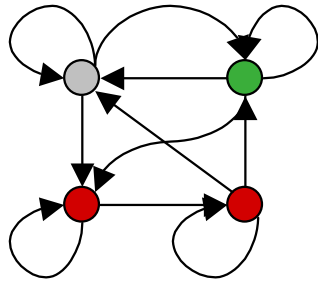
9. k chữ cuối của hai tên giống nhau, phần đầu của N_2 là viết tắt phần đầu của N_1 , với điều kiện N_2 có $k + 1$ chữ, ví dụ: “*Công ty HP VN*” và “*Cty HP VN*”.
10. Các chữ viết tắt của N_2 đều là viết tắt các cụm từ trong N_1 và các chữ còn lại trong N_2 đều xuất hiện trong N_1 , ví dụ: “*Công ty TNHH Hewlett Packard Việt Nam*”, “*Cty HP VN*”, “*HP VN*”, “*HP Việt Nam*” và “*Công ty HP Việt Nam*”
11. Hai tên xuất hiện liên tiếp trong văn bản theo dạng $N_1(N_2)$, với điều kiện N_2 chỉ có một chữ và thực thể tương ứng thuộc lớp tổ chức. Ví dụ: “*Phòng Thương mại và Công nghiệp Việt Nam (VCCI)*”, hoặc “*Liên đoàn Bóng đá Việt Nam (VFF)*”, hoặc “*Tổng công ty Cao su VN (Geruco)*”.

Máy trạng thái hữu hạn (Finite State Machines)

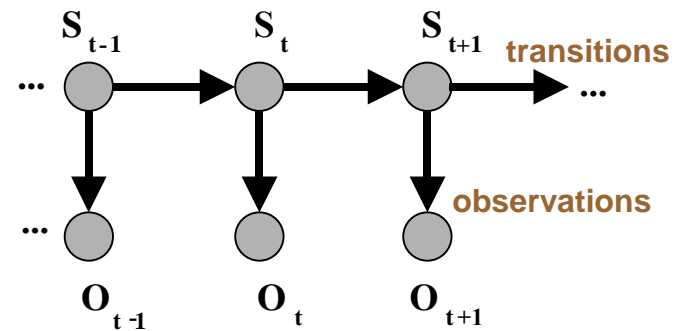
Hidden Markov Models

HMMs là công cụ mô hình hóa chuỗi chuẩn, sử dụng trong xử lý tiếng nói, XLNNTN, xử lý âm nhạc, vv

Finite state model

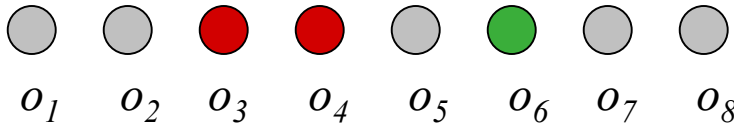


Graphical model



Generates:

State sequence
Observation sequence



$$P(\vec{s}, \vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t | s_{t-1}) P(o_t | s_t)$$

Parameters: for all states $S = \{s_1, s_2, \dots\}$

Start state probabilities: $P(s_t)$

Transition probabilities: $P(s_t | s_{t-1})$

Observation (emission) probabilities: $P(o_t | s_t)$ Usually a multinomial over atomic, fixed alphabet

Training:

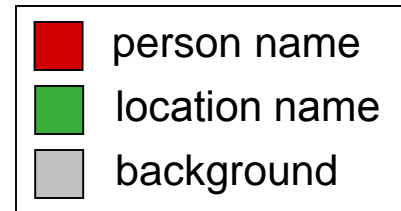
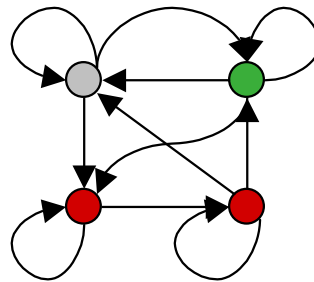
Maximize probability of training observations (w/ prior)

IE với HMM

Cho chuỗi văn bản

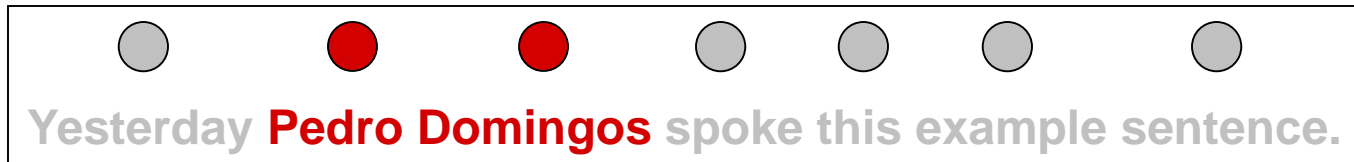
Yesterday Pedro Domingos spoke this example sentence.

Và 1 mô hình huấn luyện HMM



Tìm chuỗi trạng thái phù hợp nhất

$$\arg \max_{\bar{s}} P(\bar{s}, \bar{o})$$



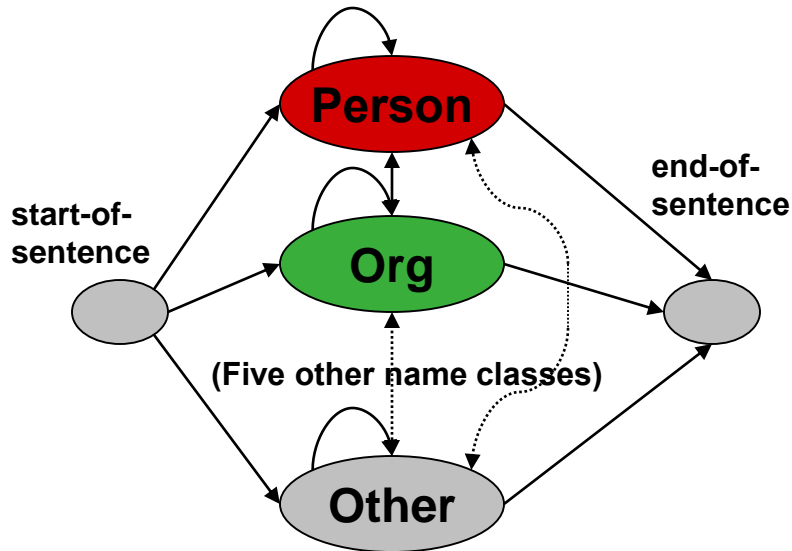
Các từ được sinh bởi mô hình nhận dạng “person name” được trích rút là person name:

Person name: **Pedro Domingos**

Ví dụ HMM : “Nymble”

[Bikel, et al 1998],
[BBN “IdentiFinder”]

Nhiệm vụ: Named Entity Extraction



Transition probabilities

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

Back-off to:

$$P(s_t | s_{t-1})$$

$$P(s_t)$$

Observation probabilities

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

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

Back-off to:

$$P(o_t | s_t)$$

$$P(o_t)$$

Luyện trên ~500k từ từ văn bản tin tức

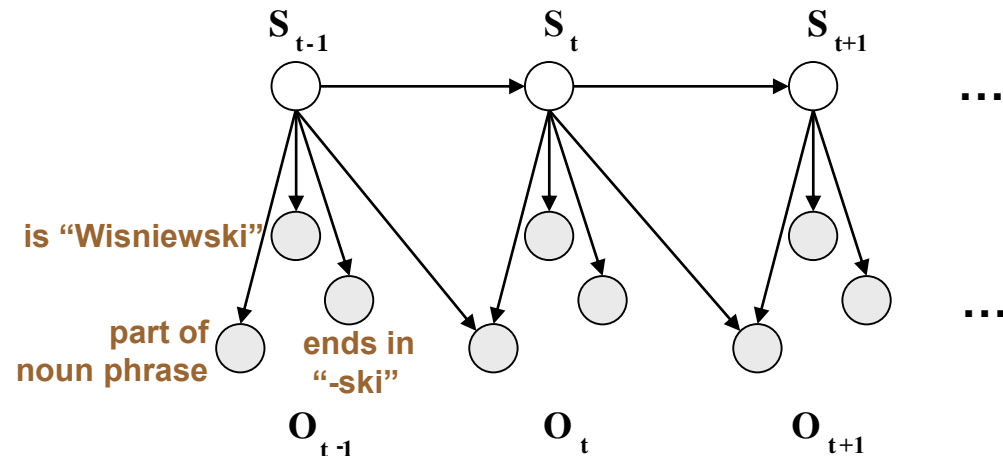
Kết quả:

<u>Case</u>	<u>Language</u>	<u>F1 .</u>
Mixed	English	93%
Upper	English	91%
Mixed	Spanish	90%

Mô hình phức tạp hơn

Các đặc trưng có thể chồng nhau

- identity of word
- ends in “-ski”
- is capitalized
- is part of a noun phrase
- is in a list of city names
- is under node X in WordNet
- is in bold font
- is indented
- is in hyperlink anchor
- last person name was female
- next two words are “and Associates”



Vấn đề

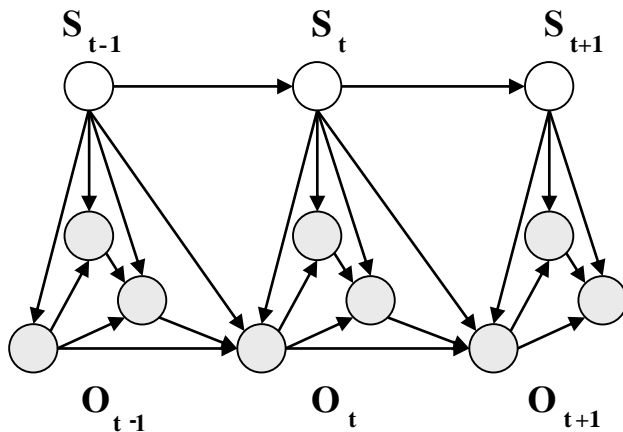
Các đặc trưng không độc lập

- Nhiều mức đơn vị cơ sở: ký tự, từ, đoạn
- Nhiều mô hình: từ, định dạng từ, các khuôn dạng

Hai lựa chọn:

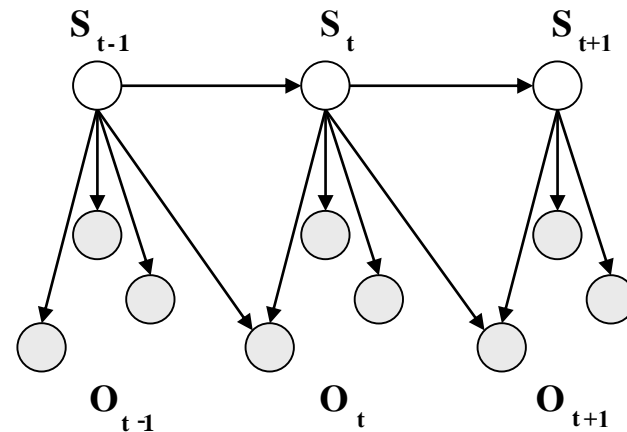
Mô hình hóa sự phụ thuộc.

Mỗi trạng thái có 1 mạng Bayes riêng. Nhưng ta thiếu dữ liệu luyện



Bỏ qua các phụ thuộc.

Gây ra việc đếm lặp lại các sự kiện (naïve Bayes). Là vấn đề lớn khi kết hợp các dữ kiện



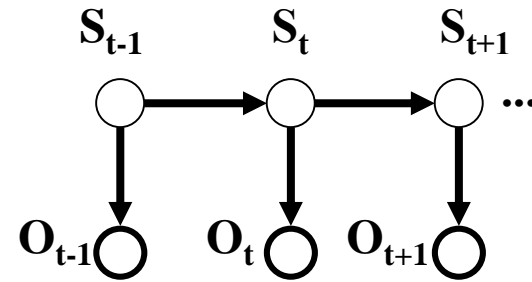
Mô hình chuỗi có điều kiện (Conditional Sequence Models)

- Mô hình luyện để tối đa xác suất có điều kiện thay vì xác suất kết hợp
 $P(\bar{s}|\bar{o})$ thay vì $P(\bar{s},\bar{o})$:
 - Có thể kiểm tra các đặc trưng, nhưng không sinh ra chúng
 - Không thể mô hình hóa các ràng buộc của chúng một cách tường minh

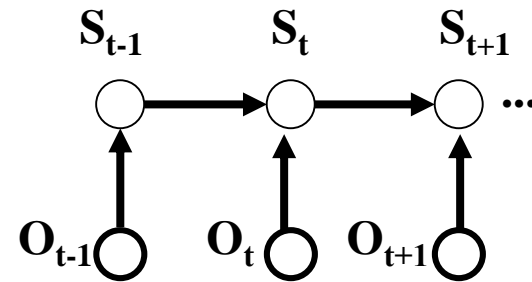
Conditional Markov Models (CMMs) vs HMMS

HMM

$$\Pr(s, o) = \prod_i \Pr(s_i | s_{i-1}) \Pr(o_i | s_i)$$



$$\Pr(s | o) = \prod_i \Pr(s_i | s_{i-1}, o_i)$$



Có rất nhiều cách để đánh giá $\Pr(y | x)$

Conditional Finite State Sequence Models

[McCallum, Freitag & Pereira, 2000]

[Lafferty, McCallum, Pereira 2001]

Từ HMMs đến CRFs

$$\vec{s} = s_1, s_2, \dots, s_n \quad \vec{o} = o_1, o_2, \dots, o_n$$

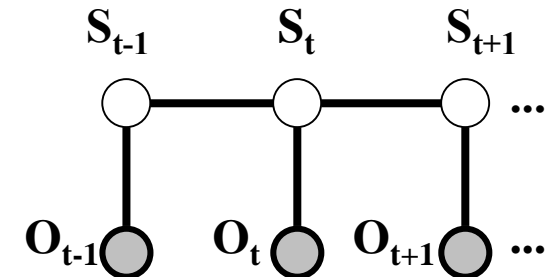
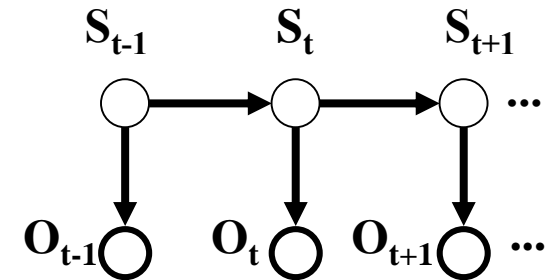
Joint
$$P(\vec{s}, \vec{o}) = \prod_{t=1}^{|\vec{o}|} P(s_t | s_{t-1}) P(o_t | s_t)$$

Conditional

$$\begin{aligned} P(\vec{s} | \vec{o}) &= \frac{1}{P(\vec{o})} \prod_{t=1}^{|\vec{o}|} P(s_t | s_{t-1}) P(o_t | s_t) \\ &= \frac{1}{Z(\vec{o})} \prod_{t=1}^{|\vec{o}|} \Phi_s(s_t, s_{t-1}) \Phi_o(o_t, s_t) \end{aligned}$$

Trong đó
$$\Phi_o(t) = \exp\left(\sum_k \lambda_k f_k(s_t, o_t)\right)$$

Các đặc trưng ngẫu nhiên của s, o , và t



(Một trường hợp đặc biệt của Conditional Random Fields.)

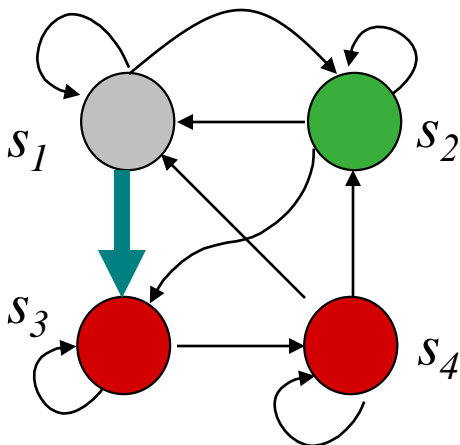
Các hàm đặc trưng

Vd. $f_k(s_t, s_{t-1}, \vec{o}, t)$:

$$f_{\langle \text{Capitalized}, s_i, s_j \rangle}(s_t, s_{t-1}, \vec{o}, t) = \begin{cases} 1 & \text{if } \text{Capitalized}(o_t) \wedge s_i = s_{t-1} \wedge s_j = s_t \\ 0 & \text{otherwise} \end{cases}$$

$\vec{o} =$ **Yesterday Pedro Domingos spoke this example sentence.**

o_1 o_2 o_3 o_4 o_5 o_6 o_7



$$f_{\langle \text{Capitalized}, s_1, s_2 \rangle}(s_2, s_1, \vec{o}, 2) = 1$$

Học tham số của CRFs

Cho tập dữ liệu luyện D , tối đa log-likelihood của các tham số $\Lambda = \{\lambda_k\}$

$$L = \sum_{\langle s, \bar{o} \rangle \in D} \log \left(\frac{1}{Z(\bar{o})} \prod_{t=1}^{|\bar{o}|} \exp \left(\sum_k \lambda_k f_k(s_t, s_{t-1}, \bar{o}, t) \right) \right) - \sum_k \frac{\lambda_k^2}{2\sigma^2}$$

Log-likelihood gradient:

$$\frac{\partial L}{\partial \lambda_k} = \sum_{\langle \bar{s}, \bar{o} \rangle \in D} \#_k(\bar{s}, \bar{o}) - \sum_i \sum_{\bar{s}'} P_{\Lambda}(\bar{s}' | \bar{o}^{(i)}) \#_k(\bar{s}', \bar{o}^{(i)}) - \frac{\lambda_k}{\sigma^2}$$

$$\text{where } \#_k(\bar{s}, \bar{o}) = \sum_t f_k(s_{t-1}, s_t, \bar{o}, t)$$

Phương pháp:

- iterative scaling (quite slow – 2000 iterations from good start)
- gradient, conjugate gradient (faster)
- limited-memory quasi-Newton methods (“super fast”)

[Sha & Pereira 2002] & [Malouf 2002]

Làm việc với dữ liệu IE

- Một số đặc trưng của IE:
 - Dựa trên việc trích rút từ văn bản
 - Dữ liệu có nhiễu (thiếu sự kiện, các giá trị thực thể chưa chuẩn hóa)
 - Có thể cần làm sạch dữ liệu trước
- Dữ liệu nhiễu, chưa chuẩn hóa thì có thể làm gì?
 - Khai phá dữ liệu
 - Truy vấn trực tiếp dựa trên các ngôn ngữ có thể xử lý mềm dẻo các từ/cụm từ gần giống chứ không dựa trên từ khóa. *[Cohen 1998]*
 - Sử dụng nó để xây dựng các đặc trưng cho bộ học *[Cohen 2000]*