ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

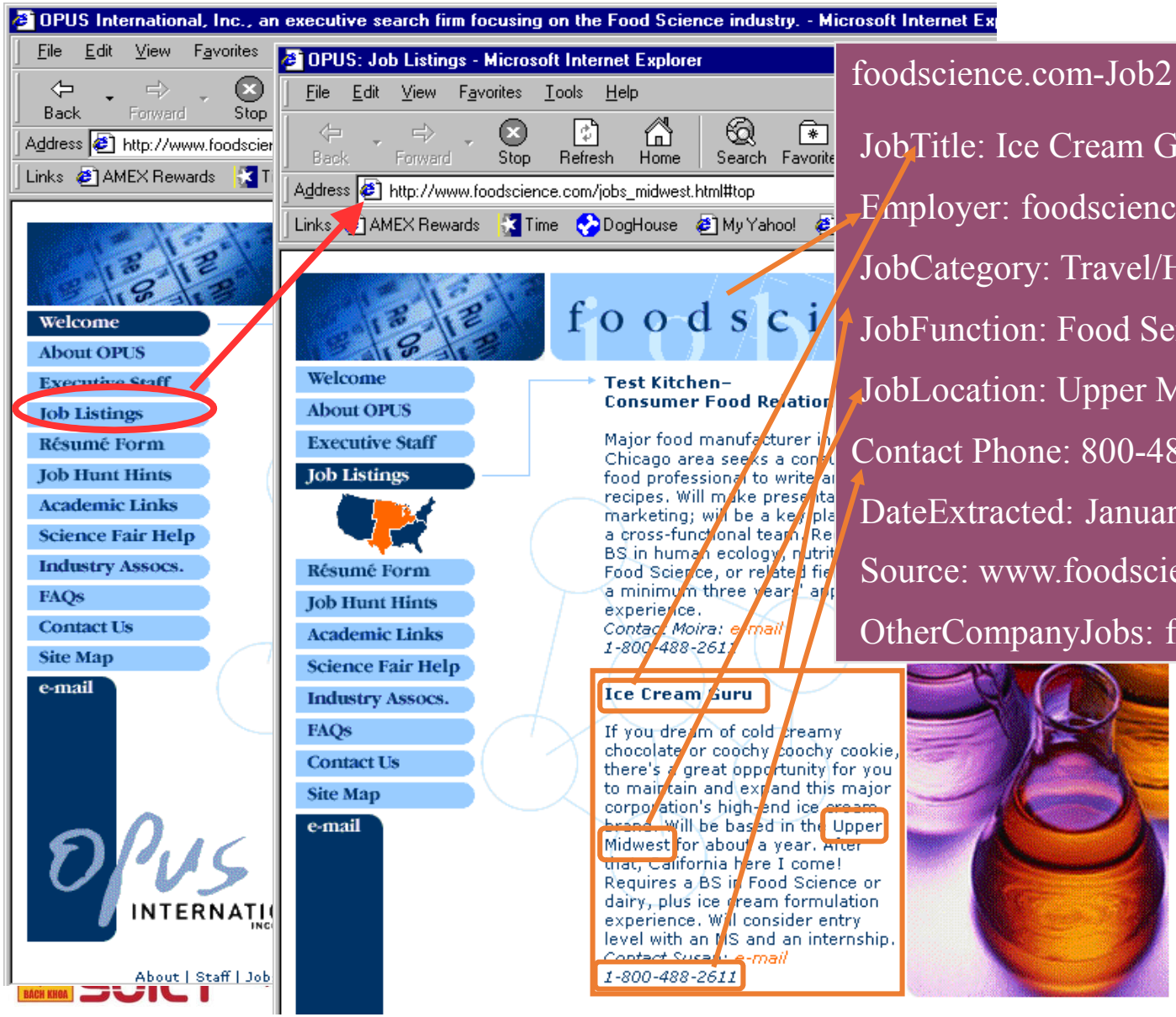
Trích rút thông tin

Viện CNTT &TT – Trường ĐHBKHN

Giới thiệu

- 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

Trích rút thông tin về quảng cáo việc làm từ Web



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

Quảng cáo nhà đất

- Các quảng cáo ở dạng văn bản
- Thêm các thẻ cơ bản: chỉ 70+ tờ 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

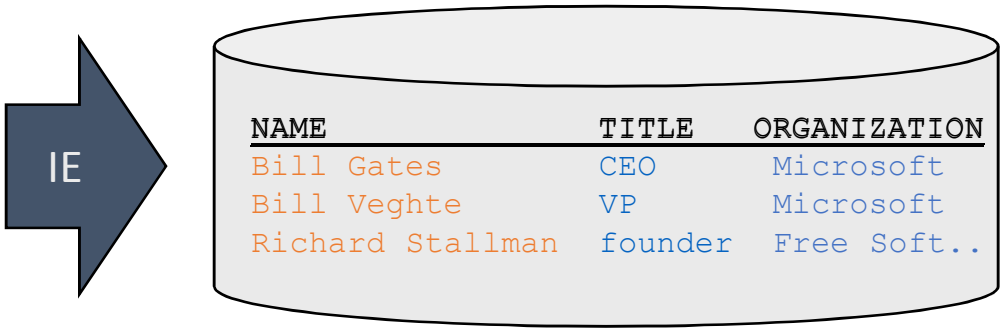
October 14, 2002, 4:00 a.m. PT

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…



“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

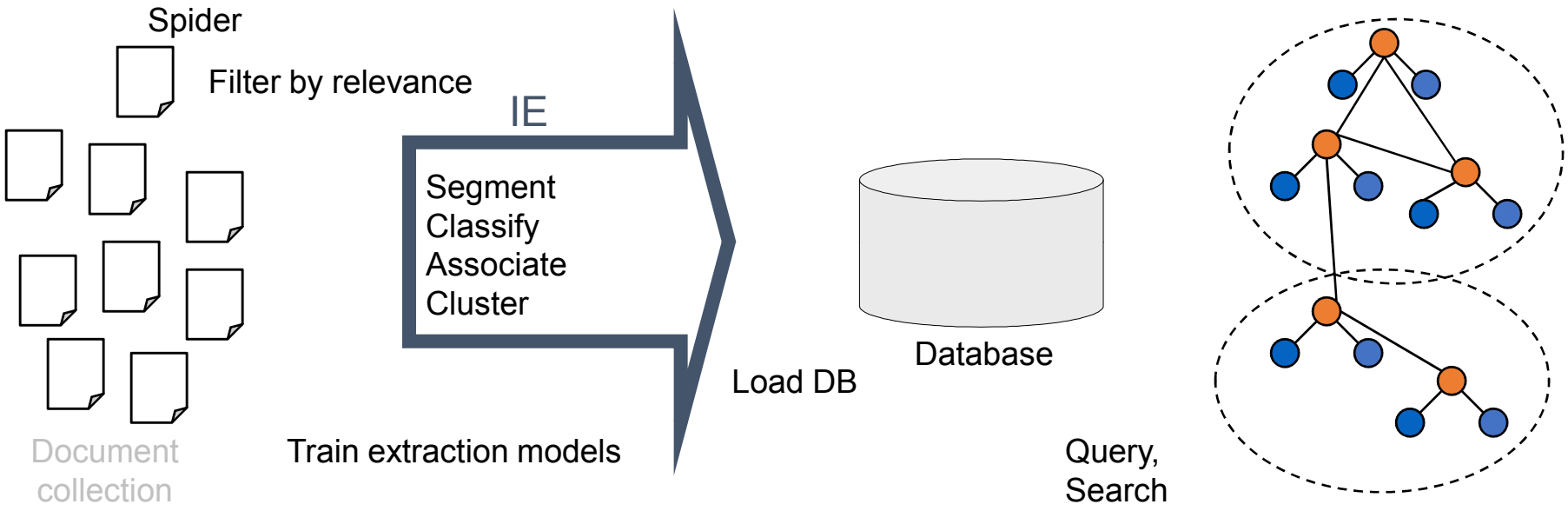
* Microsoft

* 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.

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

- Press

Contact

- General information
- Directions maps

Non-grammatical snippets, rich formatting & links

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

Tables

8:30 - 9:30 AM	**Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty** *Joseph Y. Halpern, Cornell University*
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 *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas, and Bert Van Nuffelen*	758: Title Generation for Machine-Translated Documents *Rong Jin and Alexander G. Hauptmann*	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories *Marco Cadoli, Thomas Eiter, and Georg Gottlob*	179: Knowledge Extraction and Comparison from Local Function Networks *Kenneth McGarry, Stefan Wermter, and John MacIntyre*	71: Iterative Widening *Tristan Cazenave*
549: Online-Execution of ccGolog Plans *Henrik Grosskreutz and Gerhard Lakemeyer*	131: A Comparative Study of Logic Programs with Preference *Torsten Schaub and Kewen*	246: Dealing with Dependencies between Content Planning and Surface Realisation in a Pipeline Generation	470: A Perspective on Knowledge Compilation *Adnan Darwiche and Pierre Marquis*	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):
Miều dữ liệu xử lý

Web site specific

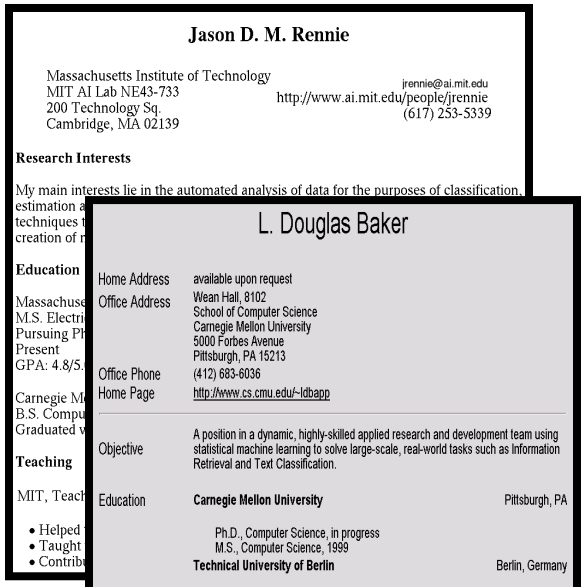
Formatting

Amazon.com Book Pages

Genre specific

Layout

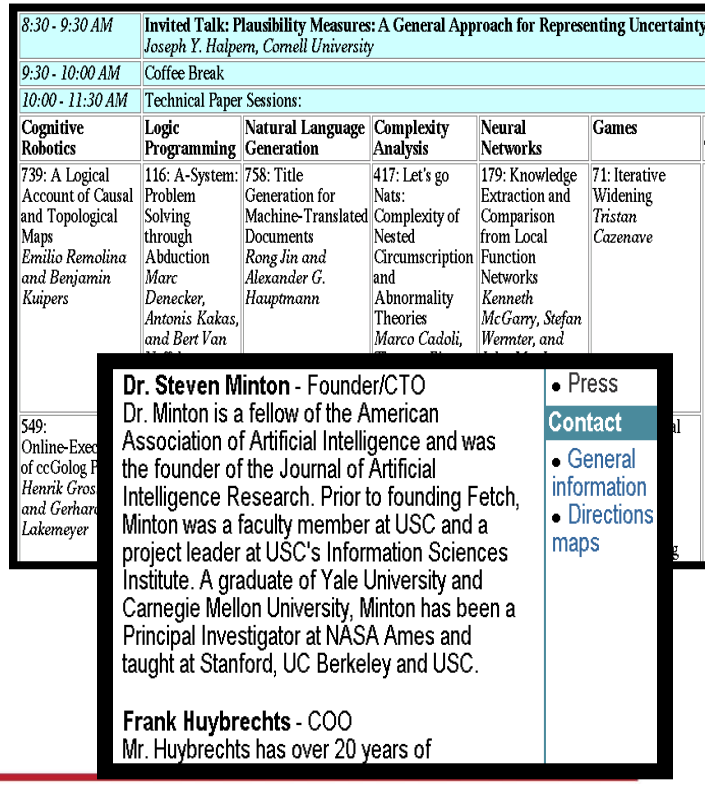
Resumes



Wide, non-specific

Language

University Names



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…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

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

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

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	Binary relationship	N-ary record
Person: Jack Welch	*Relation:* Person-Title *Person:* Jack Welch *Title:* CEO	*Relation:* Succession *Company:* General Electric *Title:* CEO *Out:* Jack Welsh *In:* Jeffrey Immelt
Person: Jeffrey Immelt		
Location: Connecticut	*Relation:* Company-Location *Company:* General Electric *Location:* Connecticut	

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% ≤ F1 ≤ 95%
- Trích rút quan hệ giữa các thực thể
  - Contained-in (Location1, Location2)
    Member-of (Person1, Organization1)
  - 60% ≤ F1 < 90%

Các bài toán trong 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:

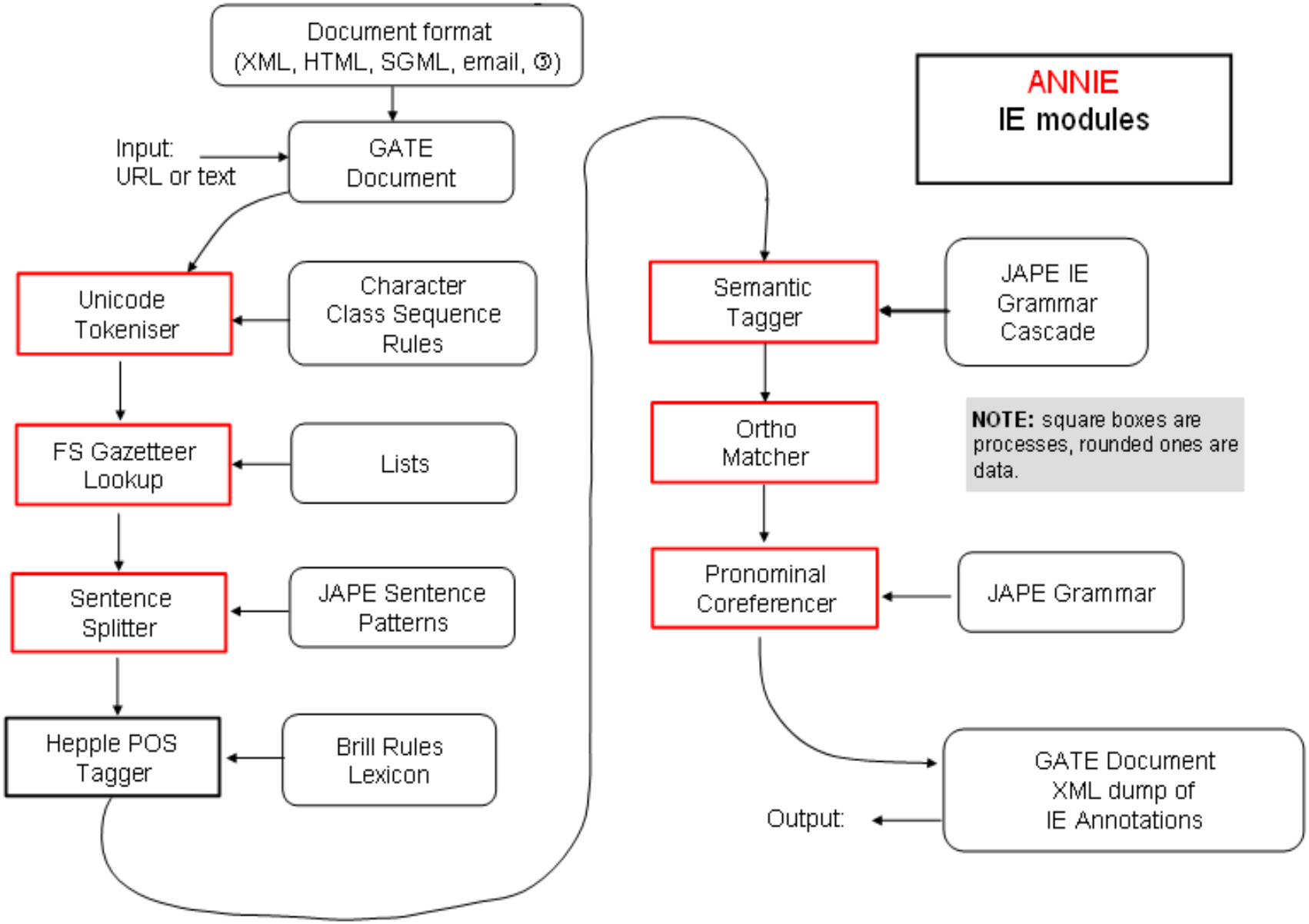
- "Dr. Peter"
  - ({DictionaryLookup = Titles}{String = "."}{Orthography type = capitalized word}) → Person Name.
  - Từ điển Titles gồm các từ "Prof", "Dr", "Mr", …
- "The XYZ Corp." hoặc "ABC Ltd."
  - ({String="The"}? {Orthography type = All capitalized} {Orthography type = Capitalized word, DictionaryType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

23

Kiến trúc IE trong GATE



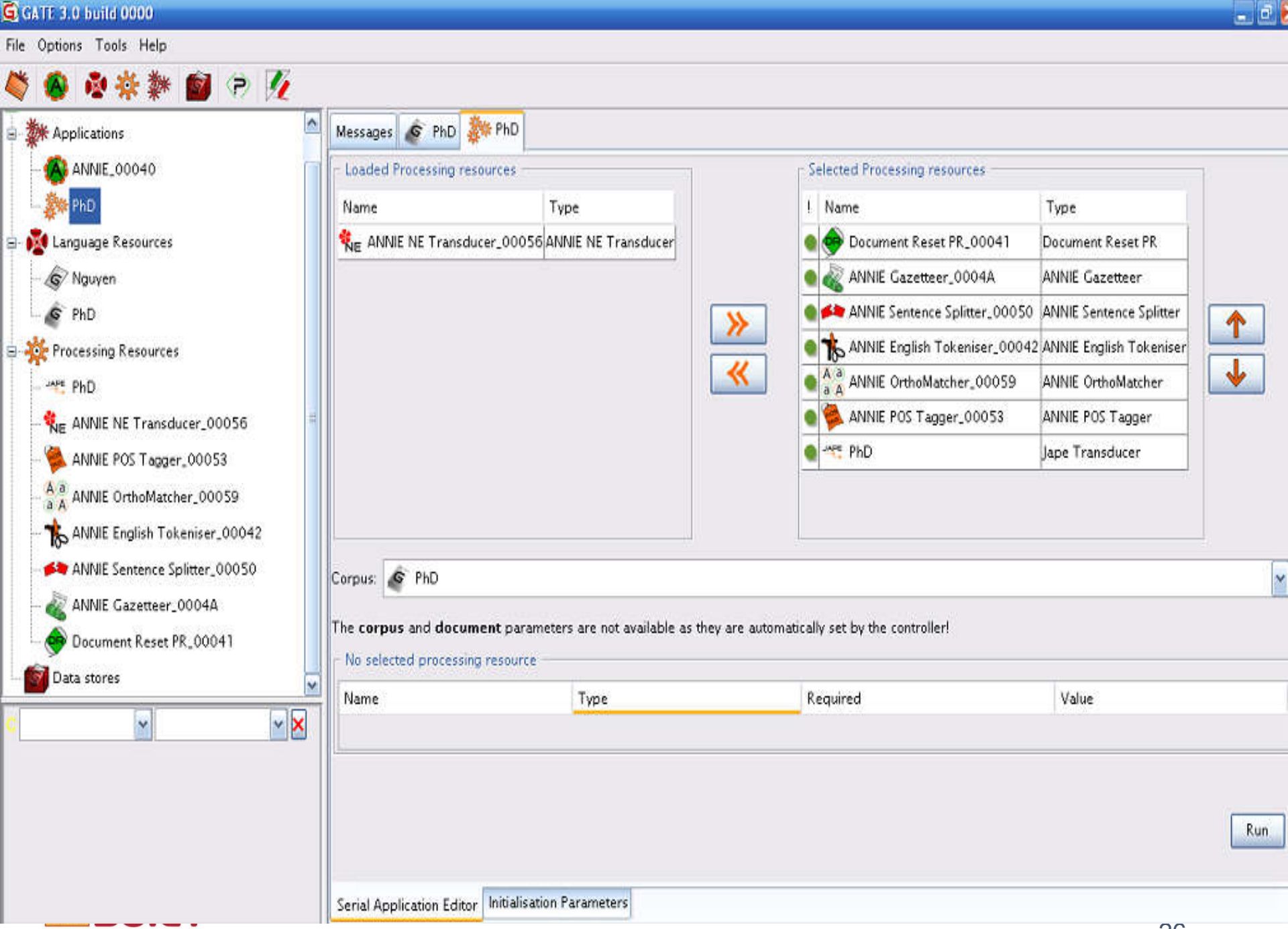
Some NER rules in GATE

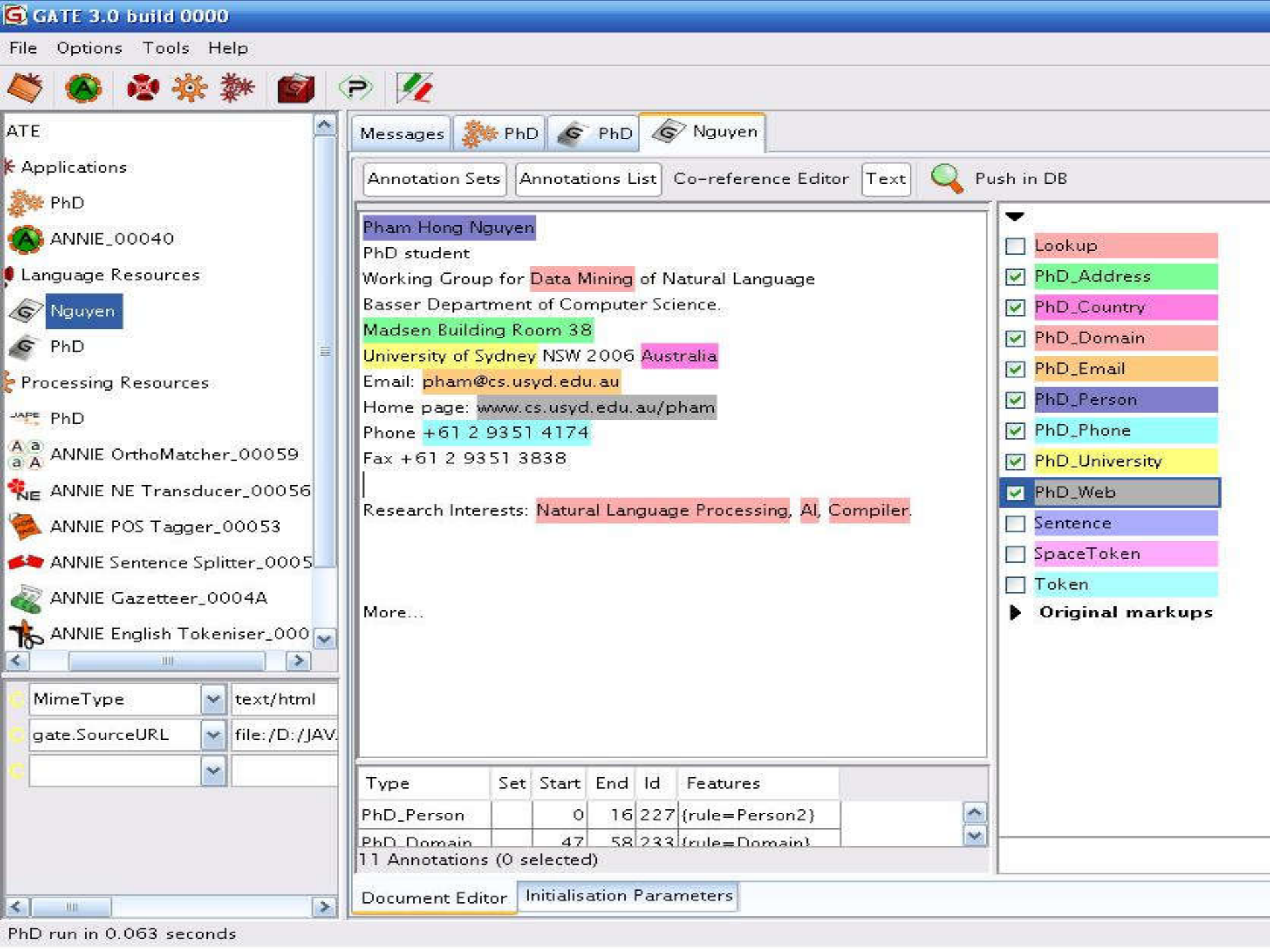
Rule: TheGazOrganization
Priority: 50
// Matches "The <in list of company names>"
({Part of speech = DT | Part of speech = RB} {DictionaryLookup = organization})
→ Organization

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


GATE 3.0 build 0000
File Options Tools Help
Applications
ANNIE_00040
PhD
Language Resources
Nguyen
PhD
Processing Resources
PhD
ANNIE NE Transducer_00056
ANNIE POS Tagger_00053
ANNIE OrthoMatcher_00059
ANNIE English Tokeniser_00042
ANNIE Sentence Splitter_00050
ANNIE Gazetteer_0004A
Document Reset PR_00041
Data stores
Messages
PhD
PhD
Loaded Processing resources
Name
Type
ANNIE NE Transducer_00056
ANNIE NE Transducer
Selected Processing resources
!
Name
Type
Document Reset PR_00041
Document Reset PR
ANNIE Gazetteer_0004A
ANNIE Gazetteer
ANNIE Sentence Splitter_00050
ANNIE Sentence Splitter
ANNIE English Tokeniser_00042
ANNIE English Tokeniser
ANNIE OrthoMatcher_00059
ANNIE OrthoMatcher
ANNIE POS Tagger_00053
ANNIE 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


GATE 3.0 build 0000
File Options Tools Help
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
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)
Document Editor
Initialisation Parameters
PhD run in 0.063 seconds

Bài tập

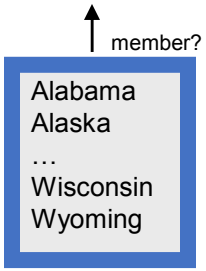
Hãy phát hiện loại thực thể và đề xuất luật nhận dạng thực thể đó:

- 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

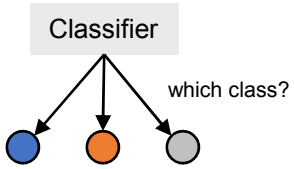
Lexicons

Abraham Lincoln was born in Kentucky.



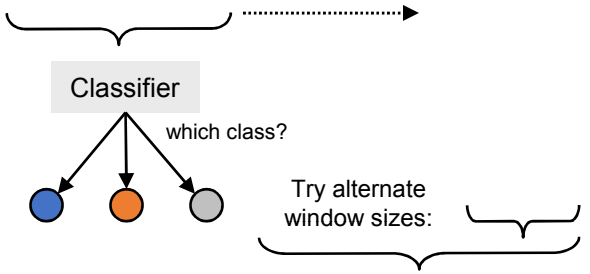
Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.

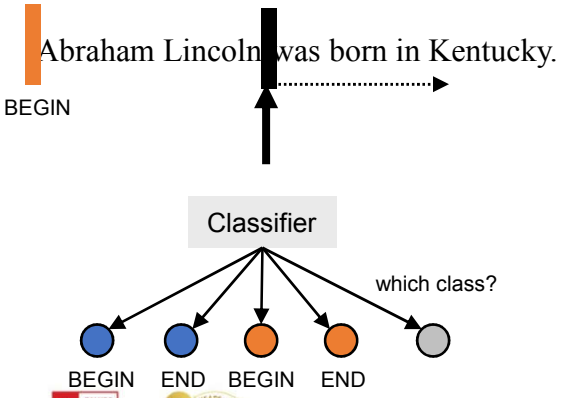


Sliding Window

Abraham Lincoln was born in Kentucky.

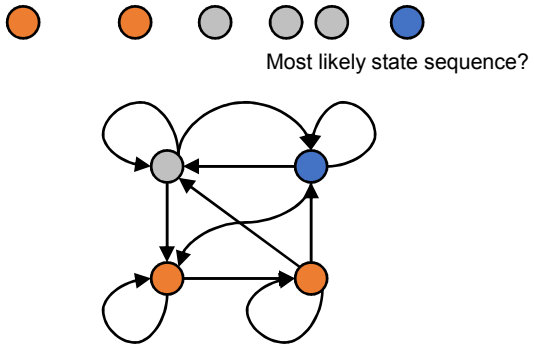


Boundary Models



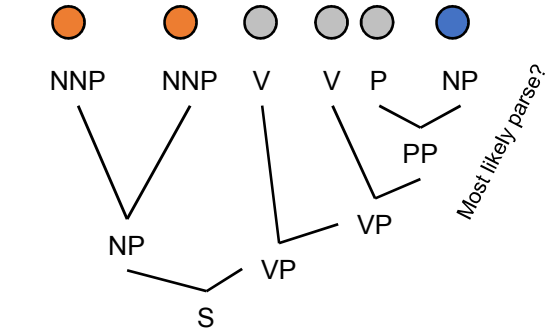
Finite State Machines

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 cửa sổ 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

Trích rút dùng cửa sổ 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

Trích rút dùng cửa sổ 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

Trích rút dùng cửa sổ 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]

… 00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thrun …

w_{t-m} w_{t-1} w_t w_{t+n} w_{t+n+1} w_{t+n+m}

prefix contents suffix

Đánh giá Pr(LOCATION|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("Place" in prefix|LOCATION)

If P("Wean Hall Rm 5409" = LOCATION) > θ , 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.
```

Field	F1
Person Name:	30%
Location:	61%
Start Time:	98%

BWI: Học phát hiện biên

[Freitag & Kushmerick, AAAI 2000]

- Học 3 bộ phân lớp dựa trên xác suất:
  - *START(i)* = Prob(vị trí *i* là bắt đầu một trường)
  - *END(j)* = Prob(vị trí *j* là kết thúc một trường)
  - *LEN(k)* = Prob(trường trích rút có độ dài k)
- Tính điểm khả năng trích rút *(i,j):*

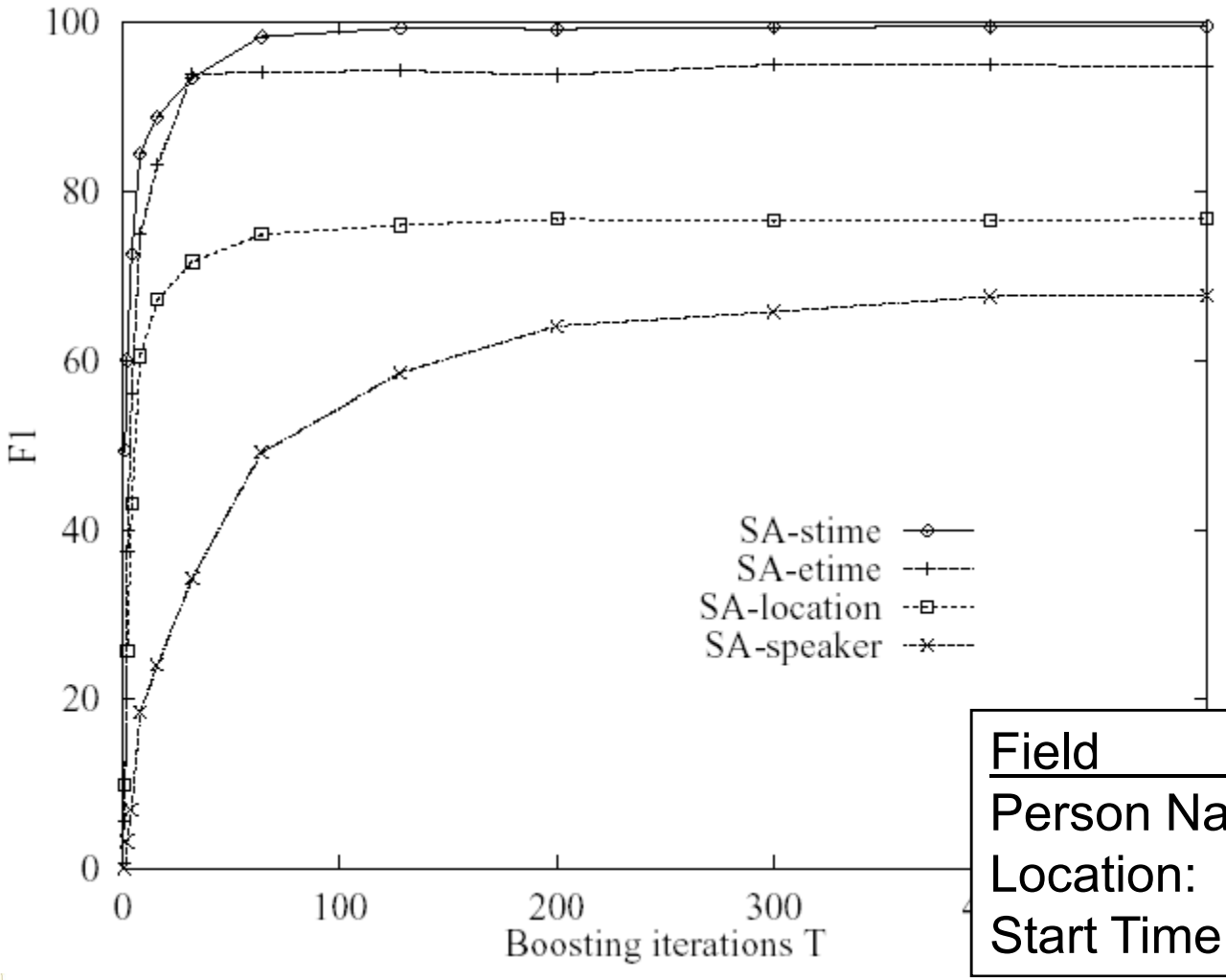
  *START(i) * END(j) * LEN(j-i)*

- *LEN(k)* được ước lượng dựa trên histogram từ tập luyện.

BWI: Học phát hiện biên

- BWI sử dụng **boosting** để tìm các bộ học mẫu để xác định điểm bắt đầu và kết thúc đ/v thực thể.
- Mỗi bộ học mẫu yếu có 1 mẫu *BEFORE* và 1 mẫu *AFTER* (các thẻ trước/sau vị trí *i).*
- Mỗi mẫu là 1 chuỗi các thẻ và/hoặc các ký tự như: anyAlphabeticToken, anyToken, anyUpperCaseLetter, anyNumber, …
- Mỗi bộ học mẫu yếu sử dụng tìm kiếm tham lam (+ nhìn trước) để lặp lại việc mở rộng các mẫu *BEFORE, AFTER* (khởi tạo rỗng) trước đó.

BWI: Học phát hiện biên



Field	F1
Person Name:	30%
Location:	61%
Start Time:	98%

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

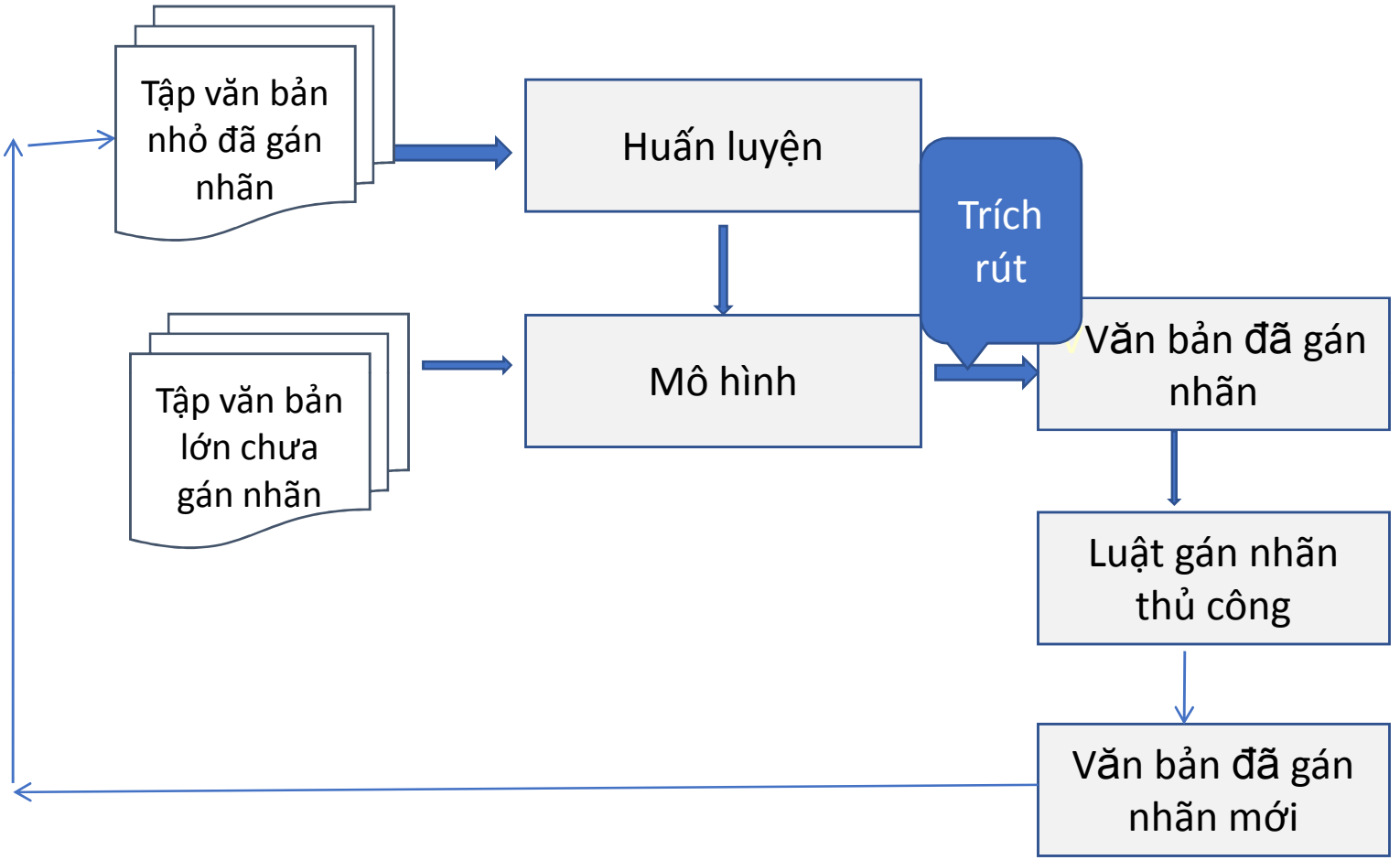
- Các quyết định về các từ bên cạnh độc lập với nhau.
  - Naïve Bayes Sliding Window có thể tiên đoán "seminar end time" trước "seminar start time".
  - Trong hệ thống tìm biên, bước tìm biên trái độc lập với bước tìm biên phải.

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 Đạo 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

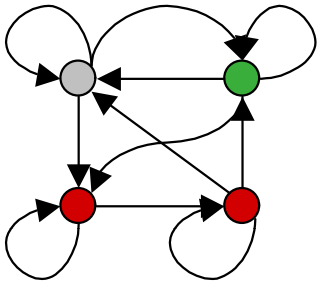
9. *k* chữ cuối của hai tên giống nhau, phần đầu của N_2 là viết tắt phần đầu của N_1, với điều kiện N_2 có $k + 1$ chữ, ví dụ: "*Công ty HP VN*" và "*Cty HP VN*".
10. Các chữ viết tắt của N_2 đều là viết tắt các cụm từ trong N_1 và các chữ còn lại trong N_2 đều xuất hiện trong N_1, ví dụ: "*Công ty 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)

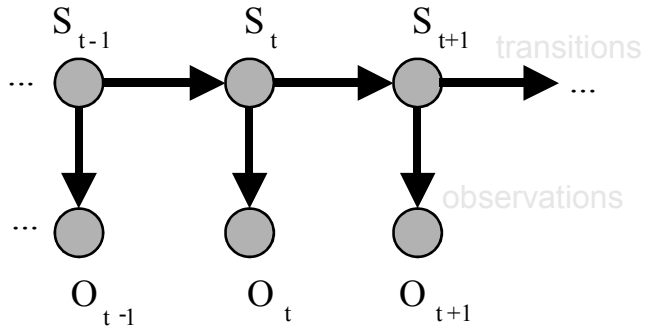
Hidden Markov Models

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



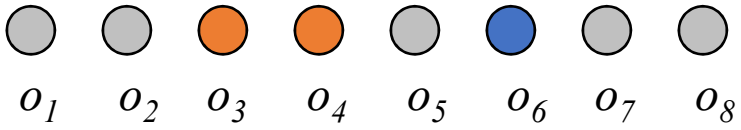
Graphical model



Generates:

State sequence

Observation sequence



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

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

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:

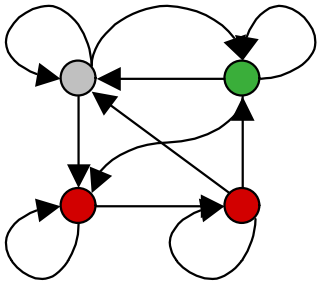
Maximize probability of training observations (w/ prior)

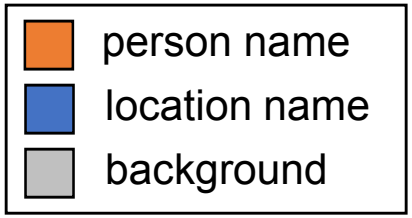
IE với HMM

Cho chuỗi văn bản

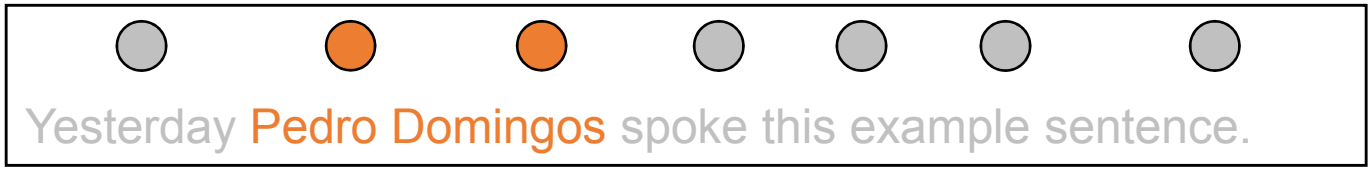
Yesterday Pedro Domingos spoke this example sentence.

Và 1 mô hình huấn luyện HMM





Tìm chuỗi trạng thái phù hợp nhất $\arg\max_{\vec{s}} P(\vec{s},\vec{o})$



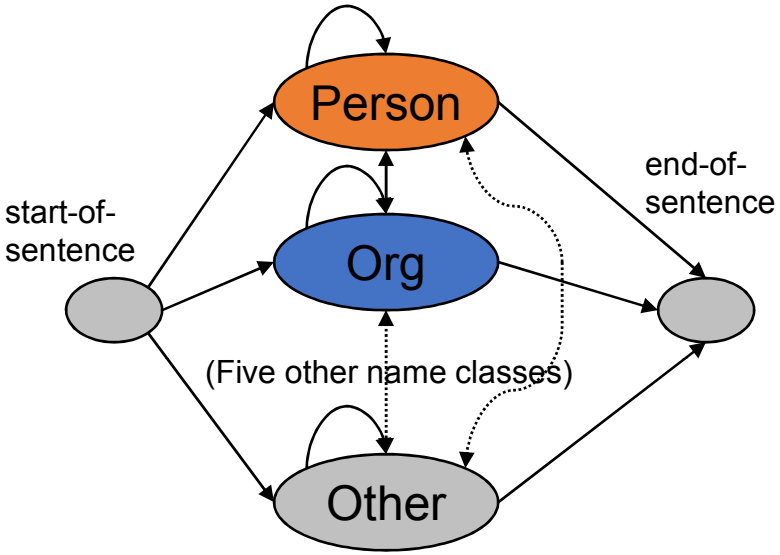
Các từ được sinh bởi mô hình nhận dạng "person name" được trích rút là person name:

Person name: Pedro Domingos

Ví dụ HMM : "Nymble"

[Bikel, et al 1998], [BBN "IdentiFinder"]

Nhiệm vụ: Named Entity Extraction



Transition probabilities

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

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:

$$P(o_t \mid s_t)$$

$$P(o_t)$$

Luyện trên ~500k từ từ văn bản tin tức

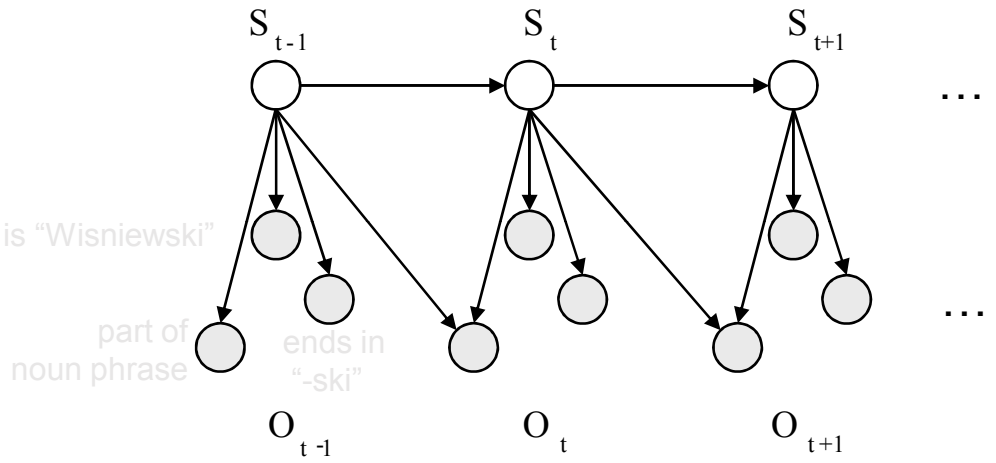
Kết quả:

Case	Language	F1 .
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”



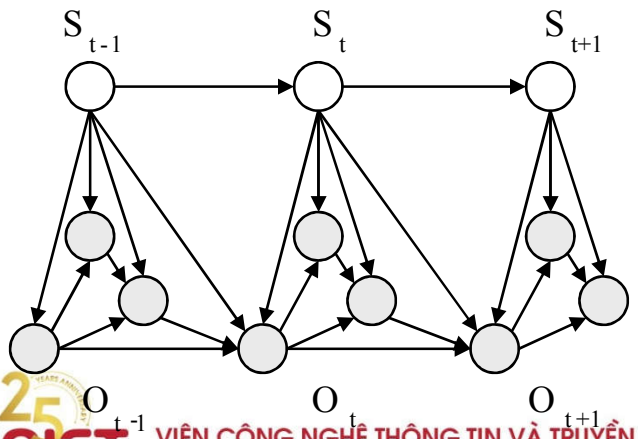
Vấn đề

Các đặc trưng không độc lập

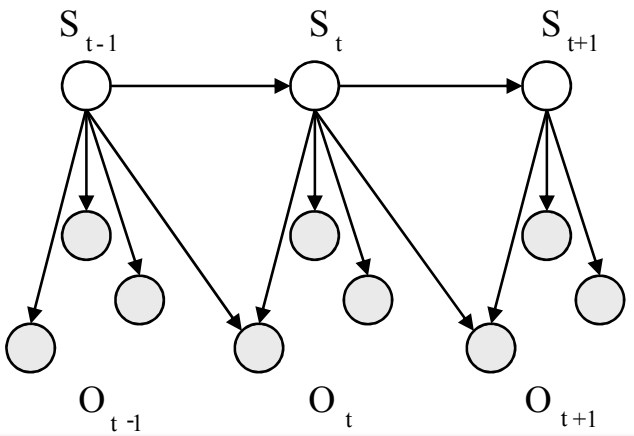
- Nhiều mức đơn vị cơ sở: ký tự, từ, đoạn
- Nhiều mô hình: từ, định dạng từ, các khuôn dạng

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)

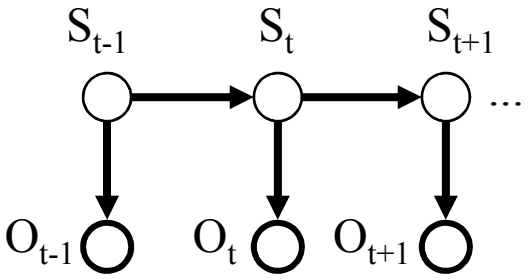
- Mô hình luyện để tối đa xác suất có điều kiện thay vì xác suất kết hợp
  $P(s|o)$ thay vì $P(s,o)$:

  - 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

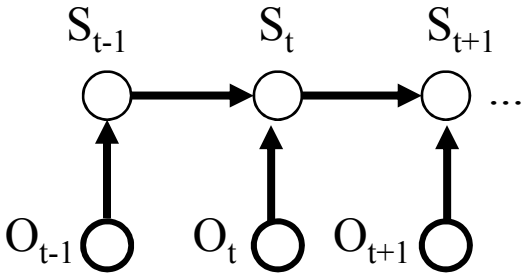
Conditional Markov Models (CMMs) vs HMMS

HMM

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



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



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

Conditional Finite State Sequence Models

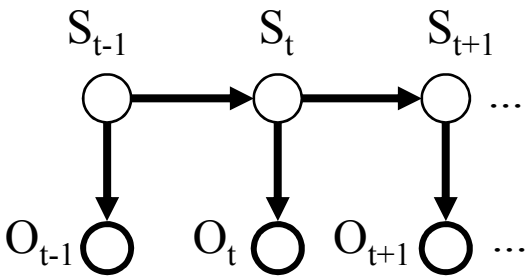
[McCallum, Freitag & Pereira, 2000]

[Lafferty, McCallum, Pereira 2001]

Từ HMMs đến CRFs

$\vec{s} = s_1, s_2, \ldots s_n \qquad \vec{o} = o_1, o_2, \ldots o_n$

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



Conditional

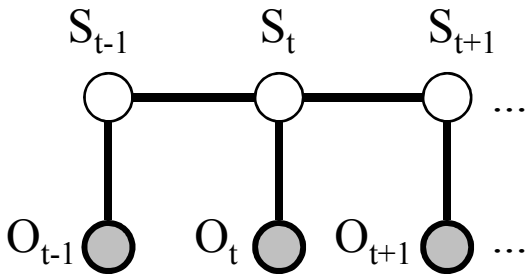
$$P(\vec{s} \mid \vec{o}) = \frac{1}{P(\vec{o})} \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}) P(o_t \mid s_t)$$

$$= \frac{1}{Z(\vec{o})} \prod_{t=1}^{|\vec{o}|} \Phi_s(s_t, s_{t-1}) \Phi_o(o_t, s_t)$$

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

Trong đó

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



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

Làm việc với dữ liệu IE

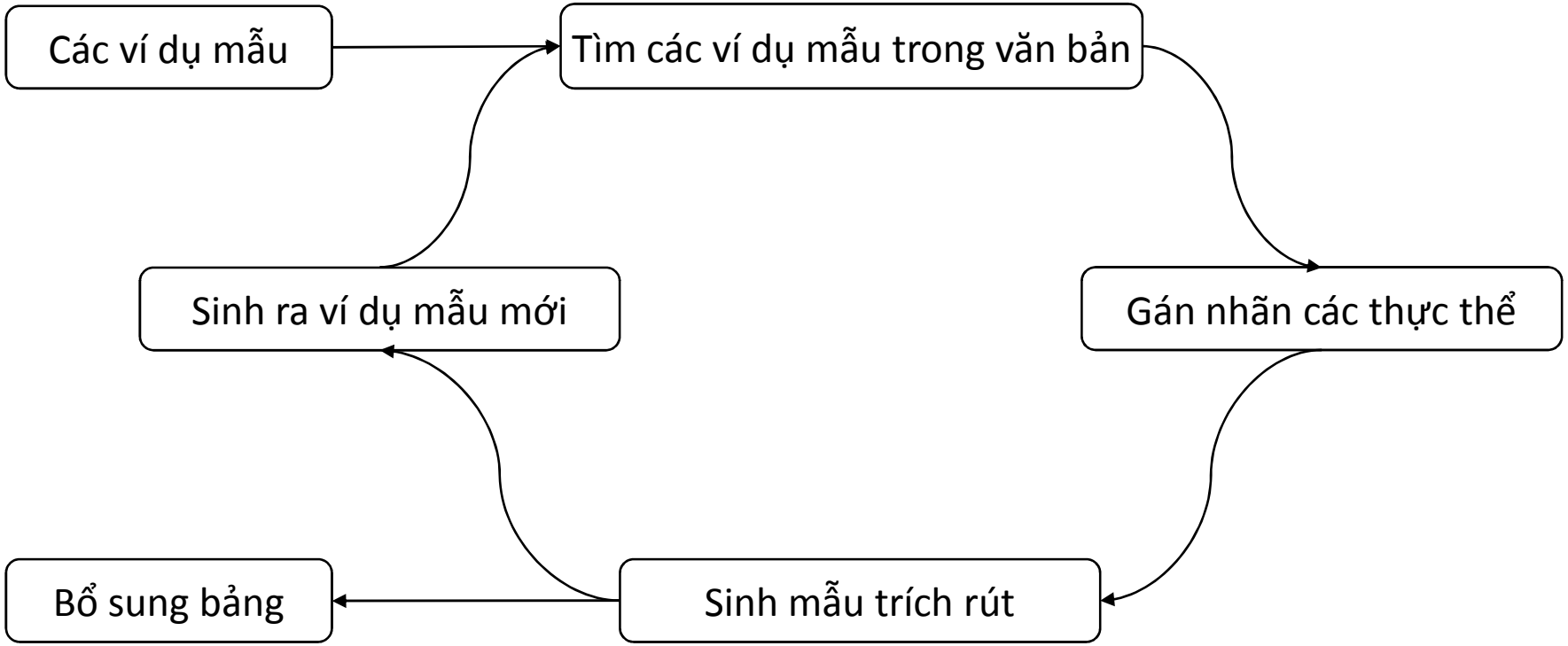
- Một số đặc trưng của IE:
  - Dựa trên việc trích rút từ văn bản
  - Dữ liệu có nhiễu (thiếu sự kiện, các giá trị thực thể chưa chuẩn hóa)
  - Có thể cần làm sạch dữ liệu trước

- 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]*

Trích rút quan hệ

- Học có giám sát có độ chính xác cao nhưng đòi hỏi DL huấn luyện
- Học không giám sát tận dụng được lượng DL lớn nhưng có độ chính xác thấp hơn
- Giám sát từ xa tận dụng được cơ sở tri thức và cải thiện độ chính xác so với học không giám sát

Học không giám sát : Snowball



Các ví dụ mẫu

- Do người dùng cung cấp
- Sau đó hệ thống tự động trích rút ra từ văn bản
- VD: Quan hệ <tập đoàn, trụ sở>
  - <Microsoft, Redmond>
  - <Exxon, Irving>

Tìm các ví dụ mẫu trong văn bản

- "Hệ thống máy chủ của **Microsoft** nằm ở trụ sở chính **Redmon**"
- "**Exxon**, **Irving** đang dần trở thành tập đoàn dầu khí..."
- "Tin đồn rút nhân viên khỏi Iraq đến từ trụ sở chính của **Exxon**, **Irving**..."

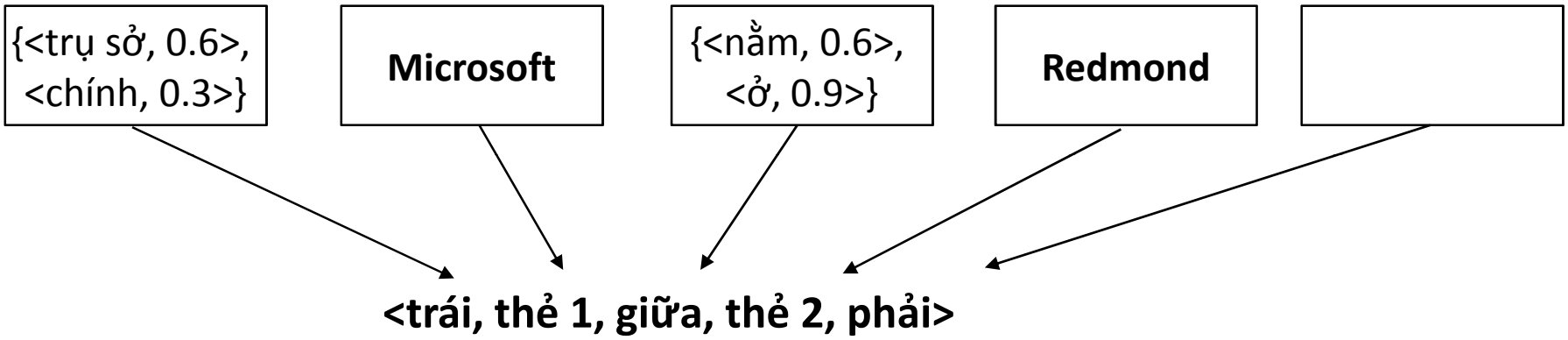
Gán nhãn thực thể

- "Hệ thống máy chủ của **<ORG>** nằm ở trụ sở chính **<LOC>**"
- "**<ORG>**, **<LOC>** đang dần trở thành tập đoàn dầu khí..."
- "Tin đồn rút nhân viên khỏi Iraq đến từ trụ sở chính của **<ORG>**, **<LOC>**..."

Sinh 5-tuple

- 5-tuple: <trái, thẻ 1, giữa, thẻ 2, phải>
- Trái: *k* từ ở bên trái cùng với véc-tơ trọng số
- Thẻ 1: thực thể thứ nhất
- Giữa: các từ ở giữa cùng với véc-tơ trọng số
- Thẻ 2: thực thể thứ hai
- Phải: *k* từ ở bên phải cùng với véc-tơ trọng số

Sinh 5-tuple (tiếp)



Sinh 5-tuple (tiếp)

{<trụ sở, 0.6>, <chính, 0.3>}	**ORG**	{<nằm, 0.6>, <ở, 0.9>}	**LOC**	
	ORG	{<',', 0.7>}	**LOC**	{<đang, 0.2>, <dần, 0.1>, <trở_thành, 0.15>}
{<trụ sở, 0.6>, <chính, 0.3>, <của, 0.5>}	**ORG**	{<',', 0.7>}	**LOC**	
{<trụ sở, 0.6>, <chính, 0.3>, <của, 0.5>}	**ORG**	{<ở, 0.95>}	**LOC**	

Sinh mẫu trích rút

- Cho 2 5-tuple có cùng thẻ tag_1 và tag_2:
  - $t = \{l, tag_1, m, tag_2, r\}$
  - $t' = \{l', tag_1, m', tag_2, r'\}$
- Độ tương đồng: $match(t, t') = l \cdot l' + m \cdot m' + r \cdot r'$
- Phân cụm các 5-tuple dựa trên độ tương đồng
- Với mỗi cụm, lấy centroid của *c* làm mẫu trích rút

$p = \{l_c, tag_1, m_c, tag_2, r_c\}$

Đánh giá mẫu

- Với mỗi ví dụ <org, loc>, phân loại
  - Positive nếu đã tồn tại ví dụ mẫu
  - Negative nếu tồn tại ví dụ mẫu <org, loc'>
  - Unknown nếu <org, *> chưa tồn tại
- Độ tin tưởng của mẫu P:

$$\text{conf}(P) = \frac{P.\text{positive}}{P.\text{positive} + P.\text{negative}}$$

- P.positive: số ví dụ positive khớp với P
- P.negative: số ví dụ negative khớp với P

Đánh giá ví dụ

- Độ tin tưởng của ví dụ T = {org, loc}

$$Conf(T) = 1 - \prod_{i=0}^{|P|} (1 - (Conf(P_i) \cdot Match(C_i, P_i)))$$

- $P = \{P_i\}$ là tập các mẫu sinh ra ví dụ T
- C_i là 5-tuple ứng với đoạn văn bản khớp với P_i với độ tương tự Match(C_i, P_i)
- Tập ví dụ mẫu = {T| Conf(T) > τ_t}

Ưu, nhược điểm

- Ưu điểm:
  - Tận dụng được dữ liệu không có nhãn
  - Chỉ cần một số ít ví dụ mẫu gốc
- Nhược điểm:
  - Vẫn yêu cầu gán nhãn thủ công từ người dùng
  - Quá trình lặp dẫn đến suy giảm chất lượng