SOICT

Information extraction

Lê Thanh Hương

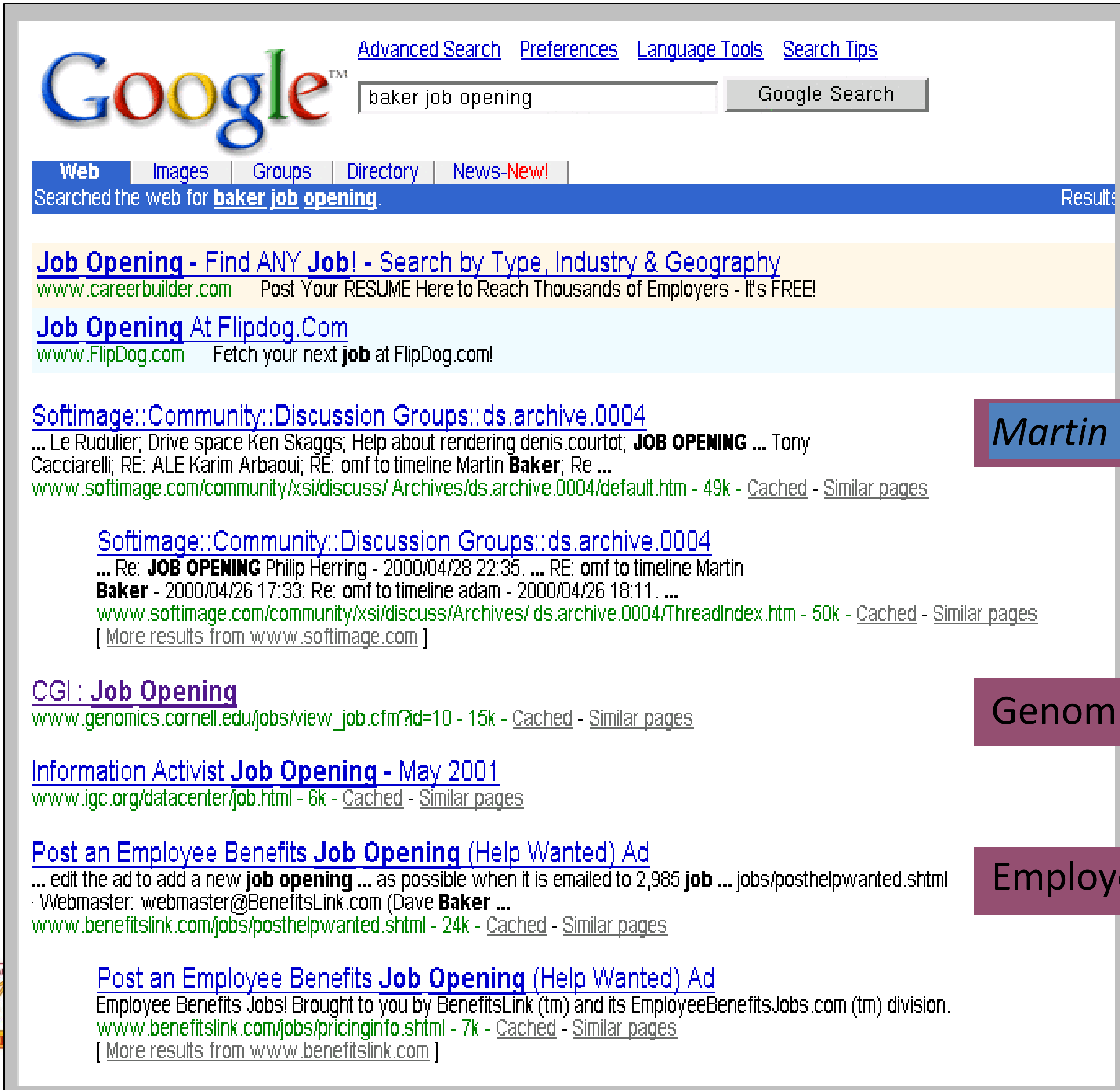
School of Information and Communication Technology

Email: huonglt@soict.hust.edu.vn

NLP in IR

- IR rarely considers semantics, e.g:
  - Search “Micheal Jordan” (basketball, machine learning)
  - Search “laptop”, not “notebook”
- Focus on common short queries and news

Example – Search engine

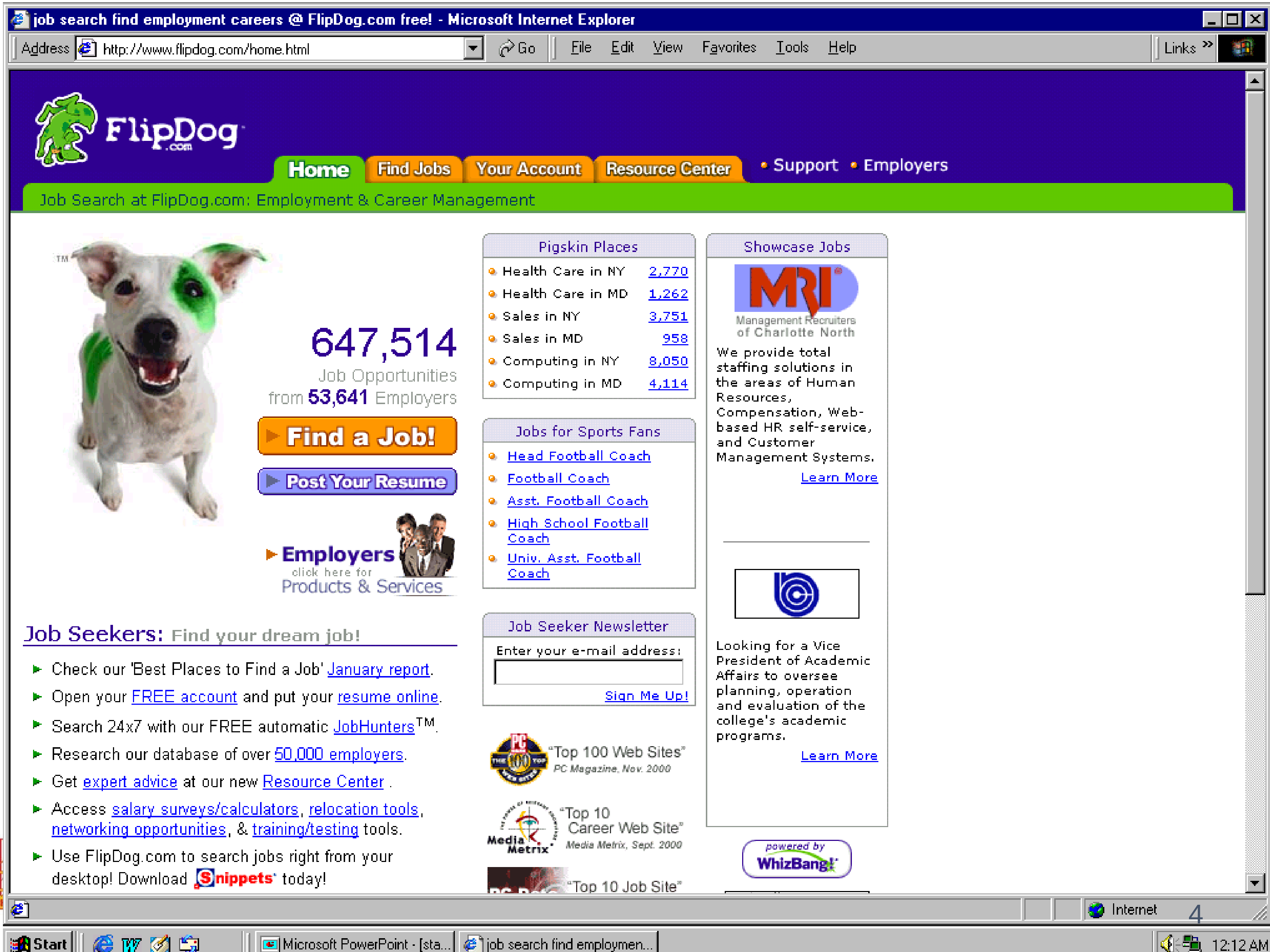


Martin Baker, a person

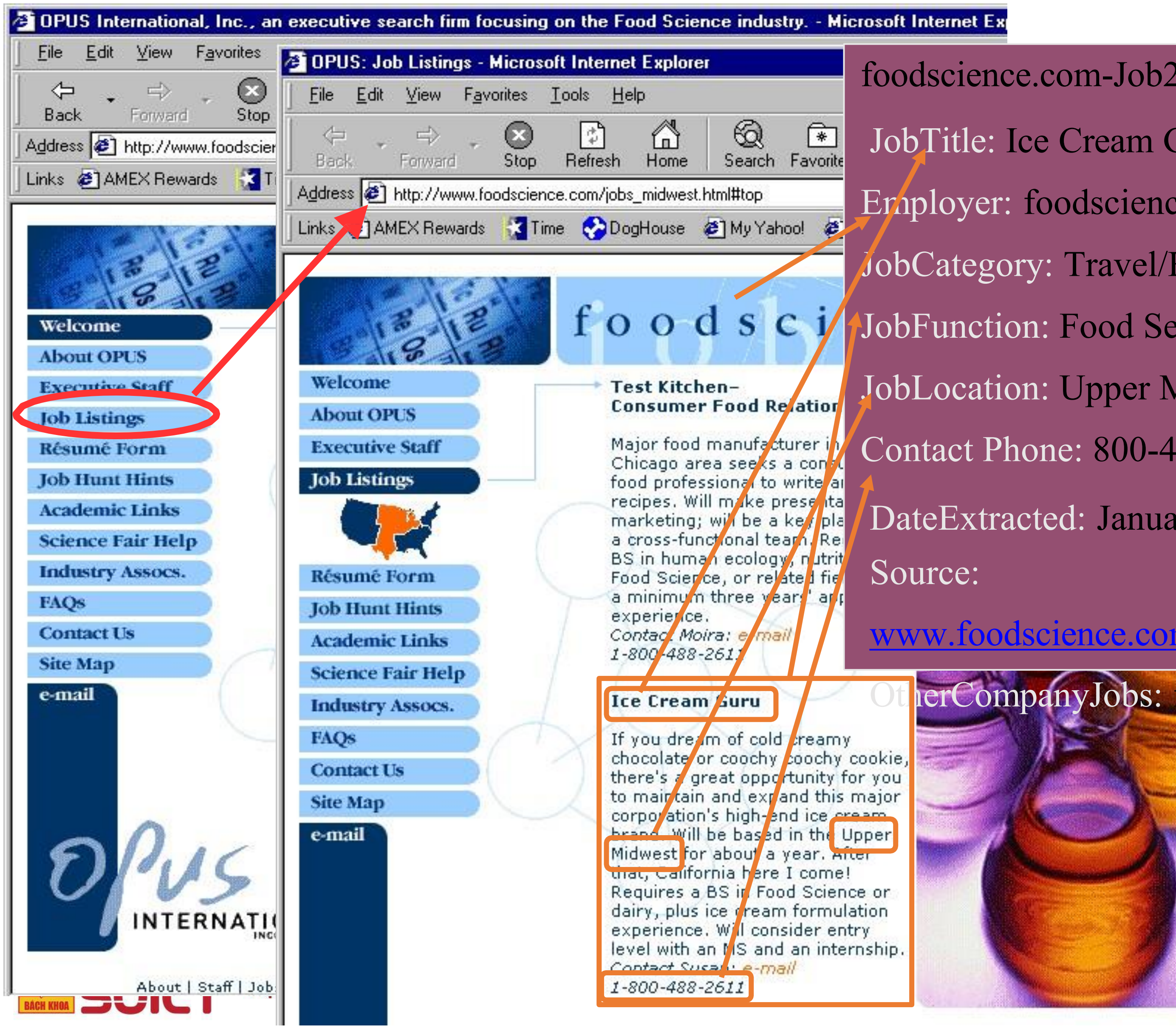
Genomics job

Employers job posting form

Example: a solution



IE on job ads from 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

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

Information extraction

- IE systems:
  - Detect and understand parts of the document
    - Explicit information (who did what to whom when?)
  - Construct a structural representation of relevant information, similarly to relations in DBs
  - Combine domain and linguistic knowledge
  - Automatically extract required information
- Example
  - Collect information on revenue from reports
  - Learn drug-gene interactions from medical studies
  - Generate smart tags (Microsoft) in documents

Real-estate ads

- Textual ads
- Add basic tags: just 70+ news agencies and 20+ publishers can do

<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>

Why (document) search engine cannot?

- Search information about real-estate ads:
  - Location:
    - Phrase: only 45 minutes from Parramatta
  - Price: $\$120K < M < \$200K$
    - Multi-constraint: before $155K, now $145
  - Bedroms: synonyms (br, bdr, beds, B/R)

Information extraction

Objective: **Extract information from documents and fill in DBs**

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…

IE

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

What is "Information extraction"?

Set of tools

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"

What is “Information extraction”?

Set of tools

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

What is “Information extraction”?

Set of tools

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

What is “Information extraction”?

Set of tools

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…

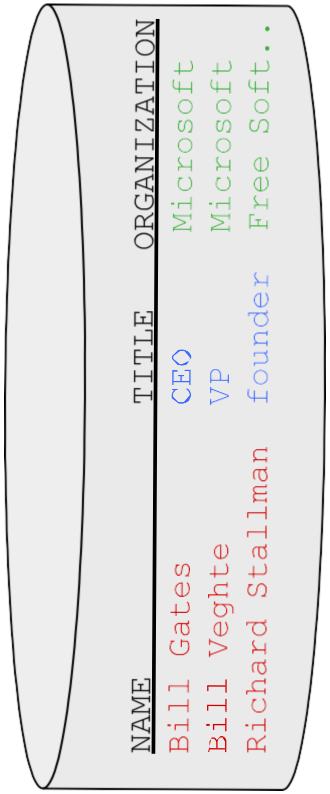
* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates

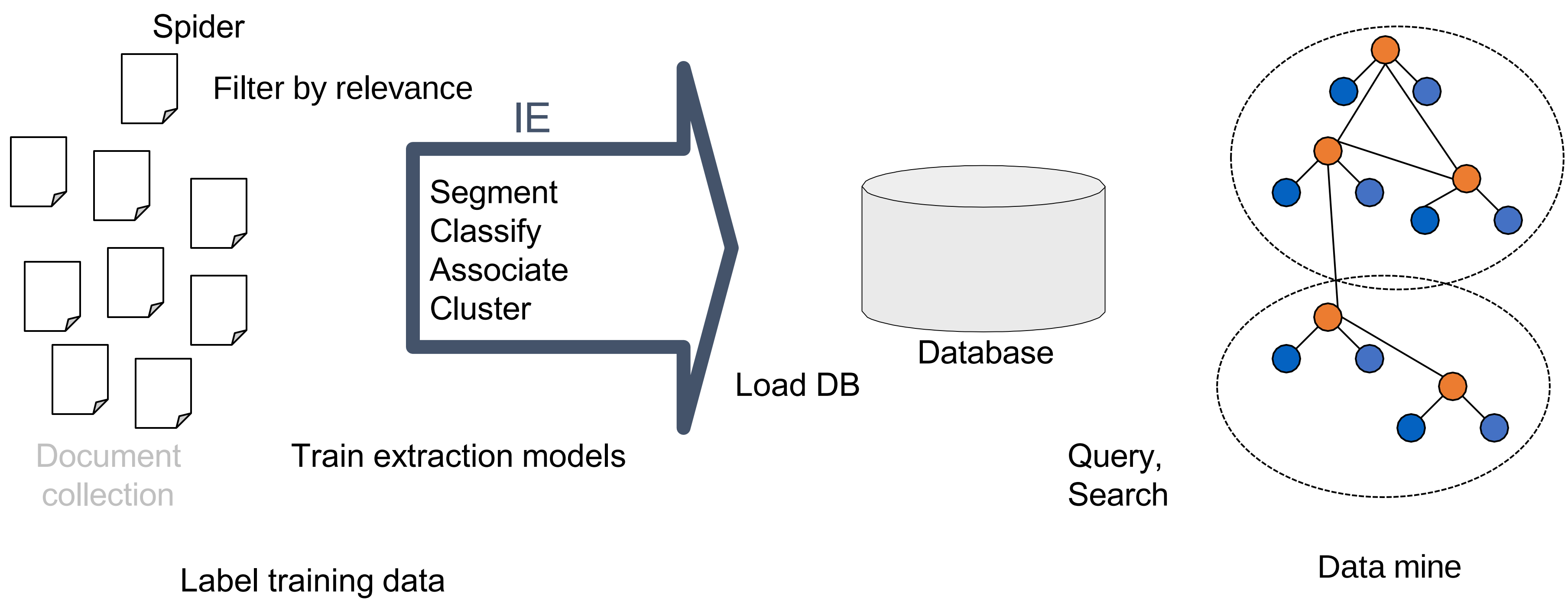
* Microsoft
Bill Veghte

* Microsoft
VP

Richard Stallman
founder
Free Software Foundation



Information extraction



Challenges in IE (1/4): Text format

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.

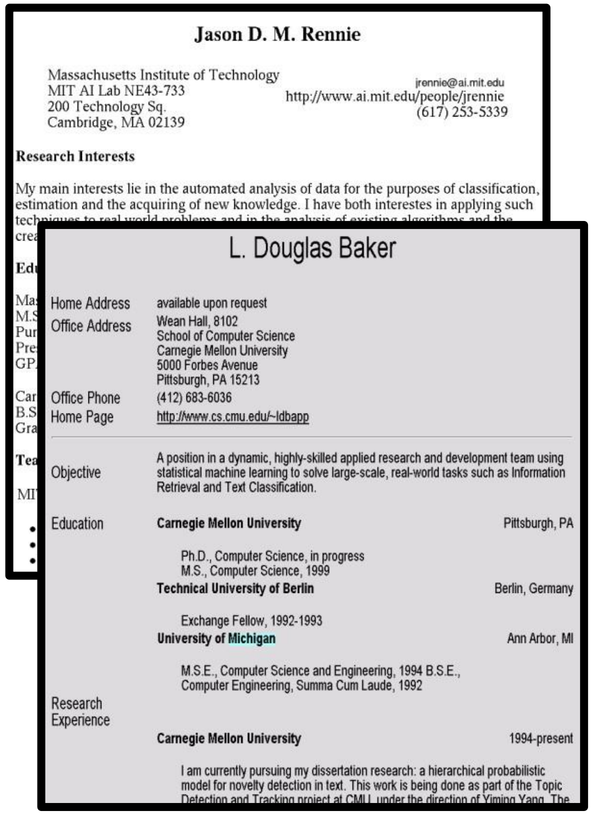
Cohen, Paul R. (413) 545-3638 cohen@cs.umass.edu CS278
Professor.
Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.

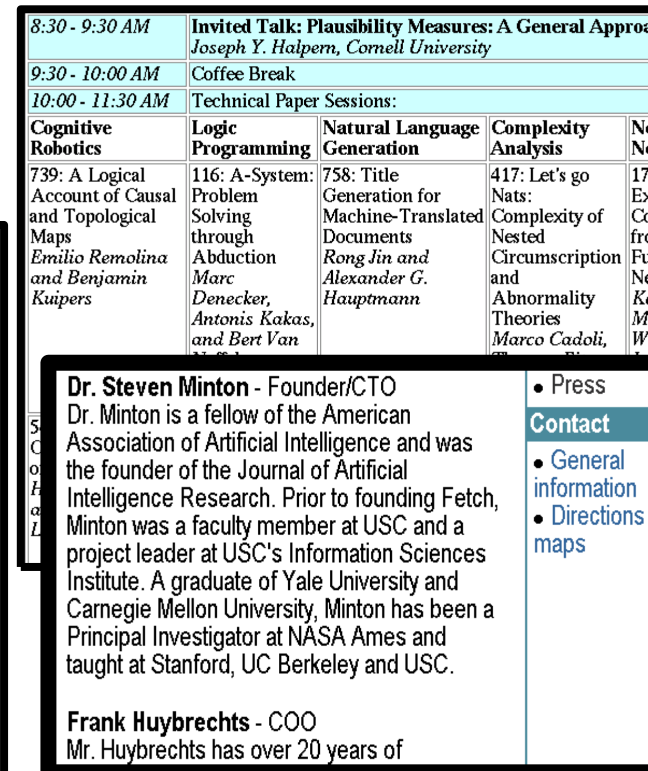
Tables

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	

Challenges in IE (2/4): Domain

Web site specific	Genre specific	Wide, non-specific
Formatting	Layout	Language
Amazon.com Book Pages	Resumes	University Names





Challenges in IE (3/4): Complexity

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.

Challenges in IE (4/4):
Data fields/records

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

<u>Single entity</u>

Person: Jack Welch

Person: Jeffrey Immelt

Location: Connecticut

"Named entity" extraction

<u>Binary relationship</u>

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

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

<u>N-ary record</u>

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

Evaluation

Golden:

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

Prediction:

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

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

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

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

State-of-the-art

- NER from news
  - Person, Location, Organization, …
  - $85\% \leq F1 \leq 95\%$

- Relation extraction
  - Contained-in (Location1, Location2)
    Member-of (Person1, Organization1)
  - $60\% \leq F1 < 90\%$

Information extraction

- Named Entity Recognition: recognize and classify unit elements in the document (person, organization, location, time)

- Relation Extraction: extract relations between entities

NER

Input: Raw document, tag set

Ra: Tagged document

Example:

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>.

NER

- Approach
  - Manual rule: Observe data patterns
    - Pro: Accurate
    - Cons: rules coverage
  - ML-generated rules: learn classifiers from annotated data
    - Pro: accurate
    - Cons: requires annotation

NER – Manual rules

- *Rule:* Contextual Pattern → Action
- Token features:
  - word
  - POS
  - format: capitalization, digit, ...
  - prefix, suffix, ...

– Action: entity tagging for a token sequence

NER – Manual rule

- NER rules have three types:
  - Content before an entity
  - Content in an entity
  - Content after an entity

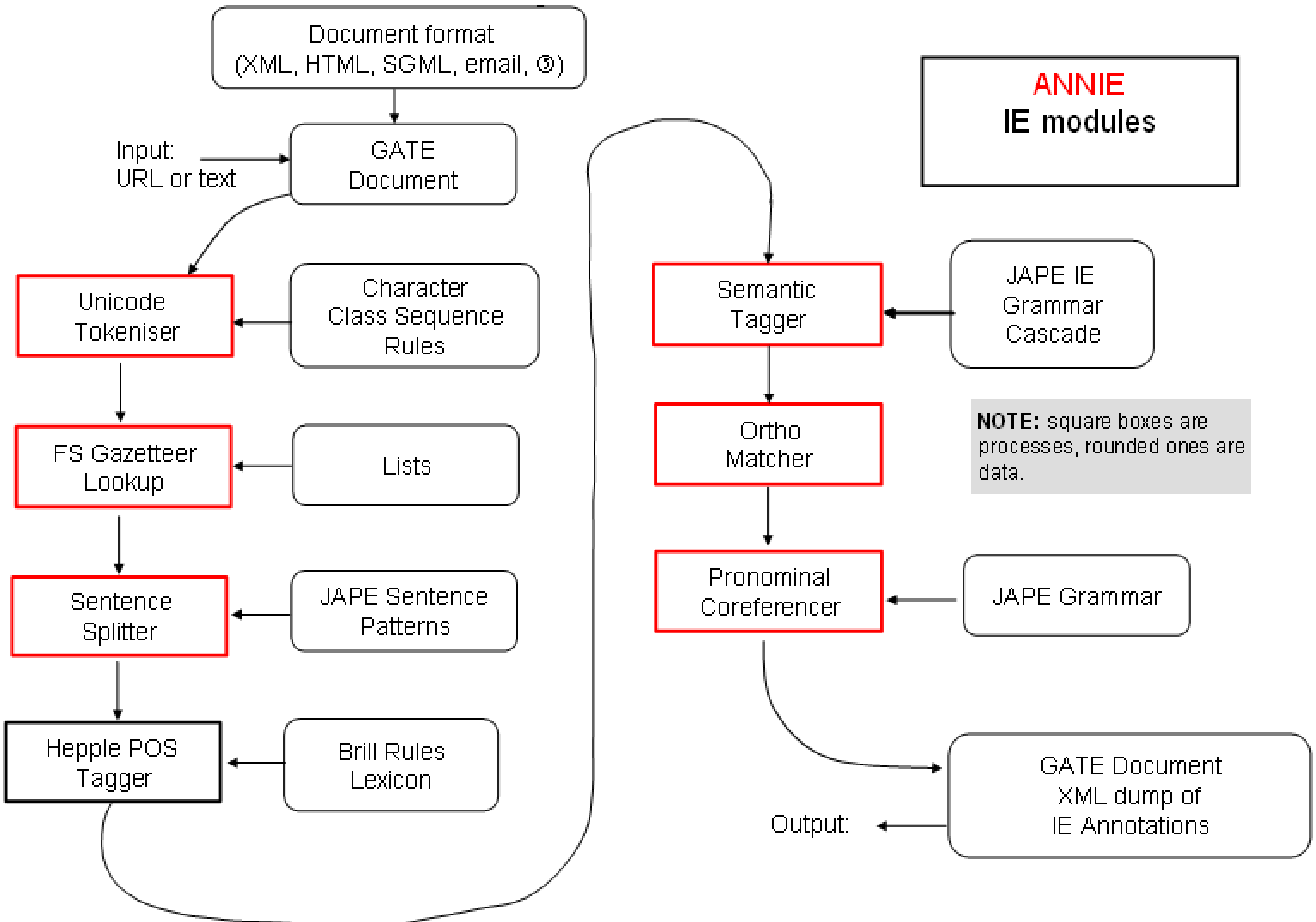
Eg:

- “Dr. Peter”
  - ({DictionaryLookup = Titles}{String = “.”}{Orthography type = capitalized word}) → Person Name.
  - Titles dictionary includes “Prof”, “Dr”, “Mr”, …
- “The XYZ Corp.” or “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 supports:
  - Software architecture
  - Framework
  - Software development environment
- GATE has three resources, called CREOLE (Collection of REusable Object for Language Engineering).
  - Language Resource
  - Processing Resource
  - Visual Resource

IE architecture in GATE



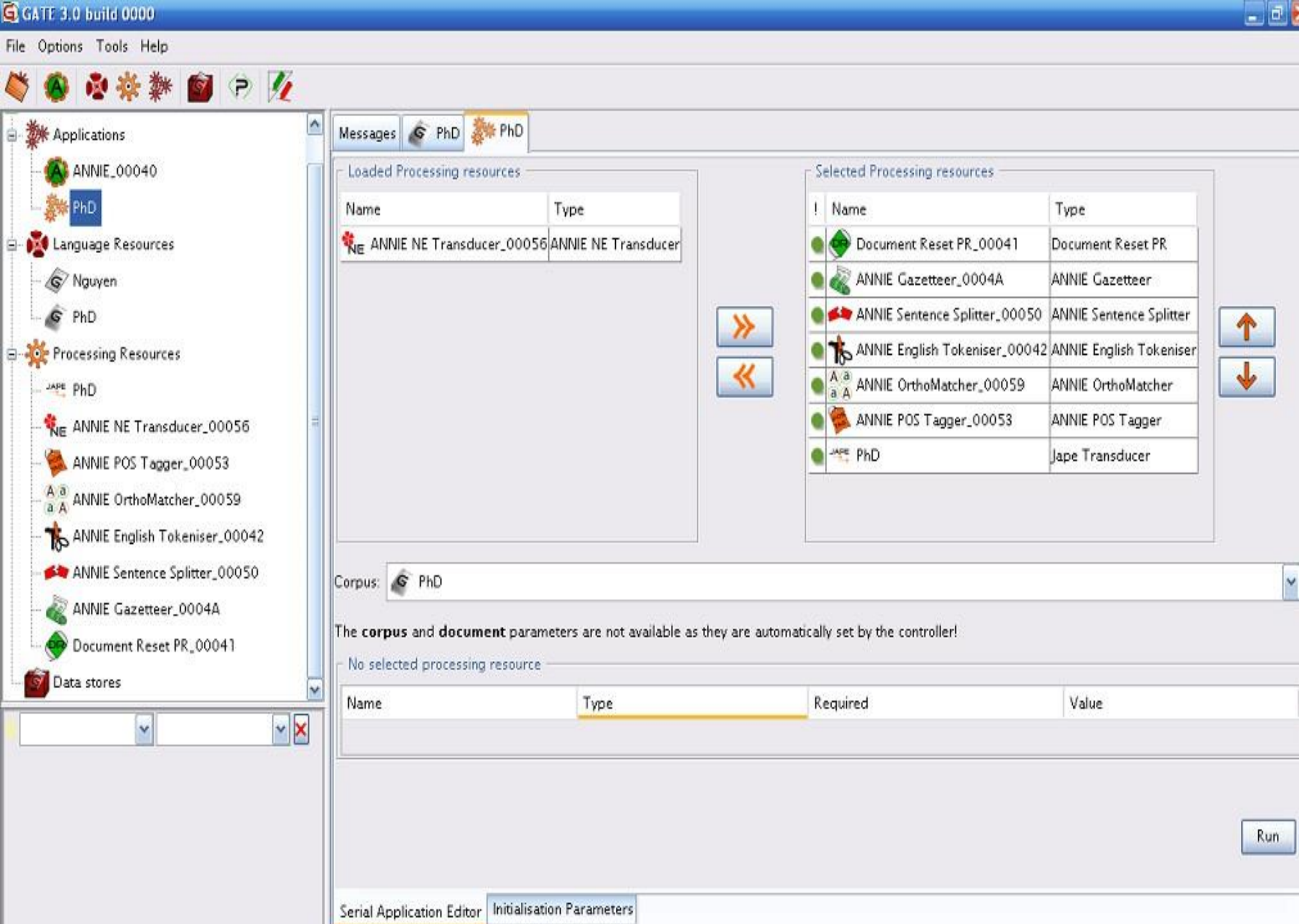
Some NER rules in GATE

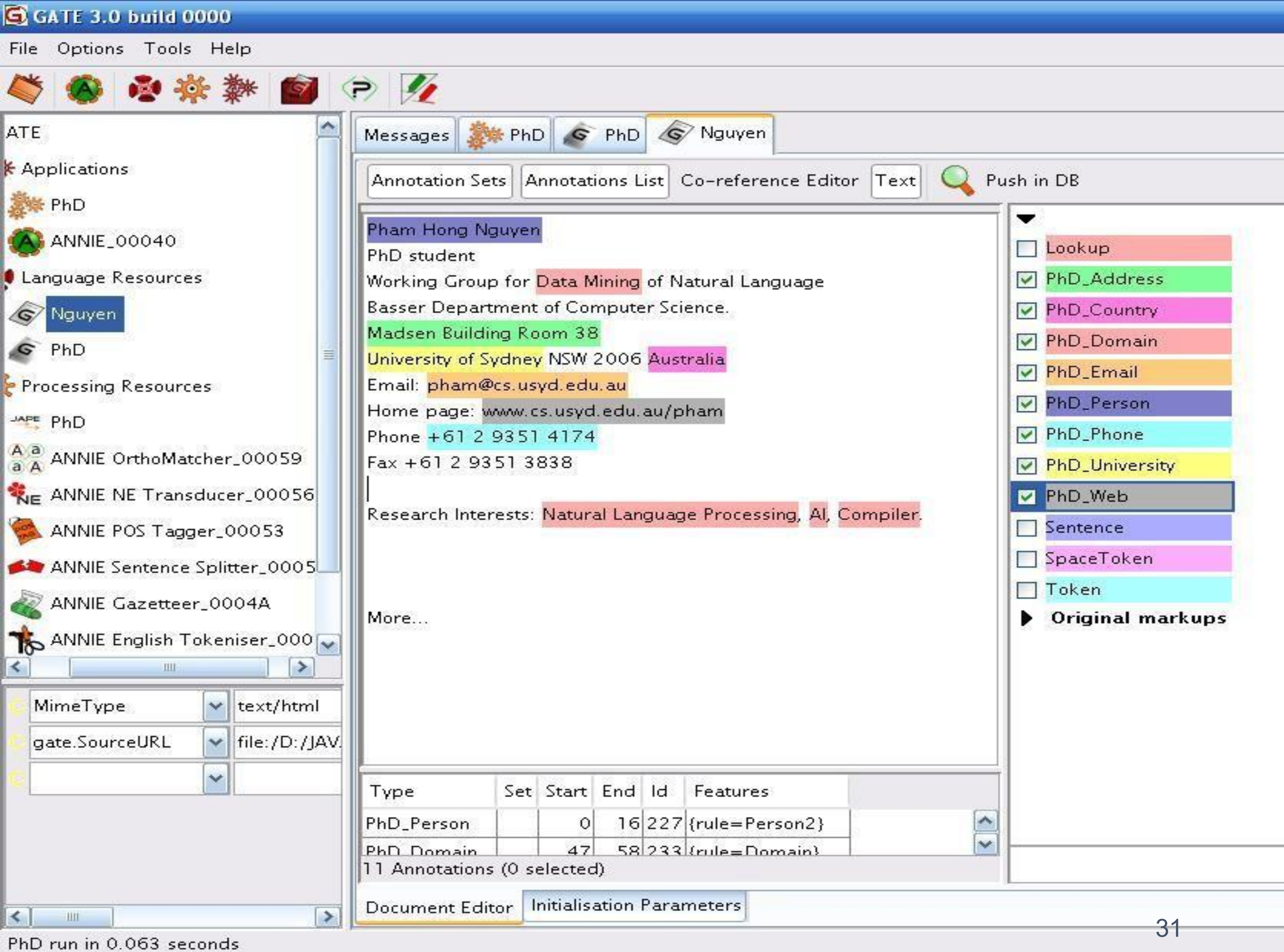
Rule: TheGazOrganization
Priority: 50
// Matches "The <in list of company names>"
({Part of speech = DT | Part of speech = RB} {DictionaryLookup = organization})
→ Organization

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 → Organization=:orgName

Rule: OrgDept
Priority: 25
// Matches "Department of Pure Mathematics and Physics"
({Token.string = "Department"} {Token.string = "of"} {Orthography type = upperInitial}+ ({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

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 thuỷ 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. Chuyến 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 thuỷ 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 ...

Exercise

Extract events from the following passages:

- 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.

What are the problems in POS tagging and NER. Vd:

1. Give examples on information in the text
2. Give examples on named entities. Find rules to extract them

Exercise (cont)

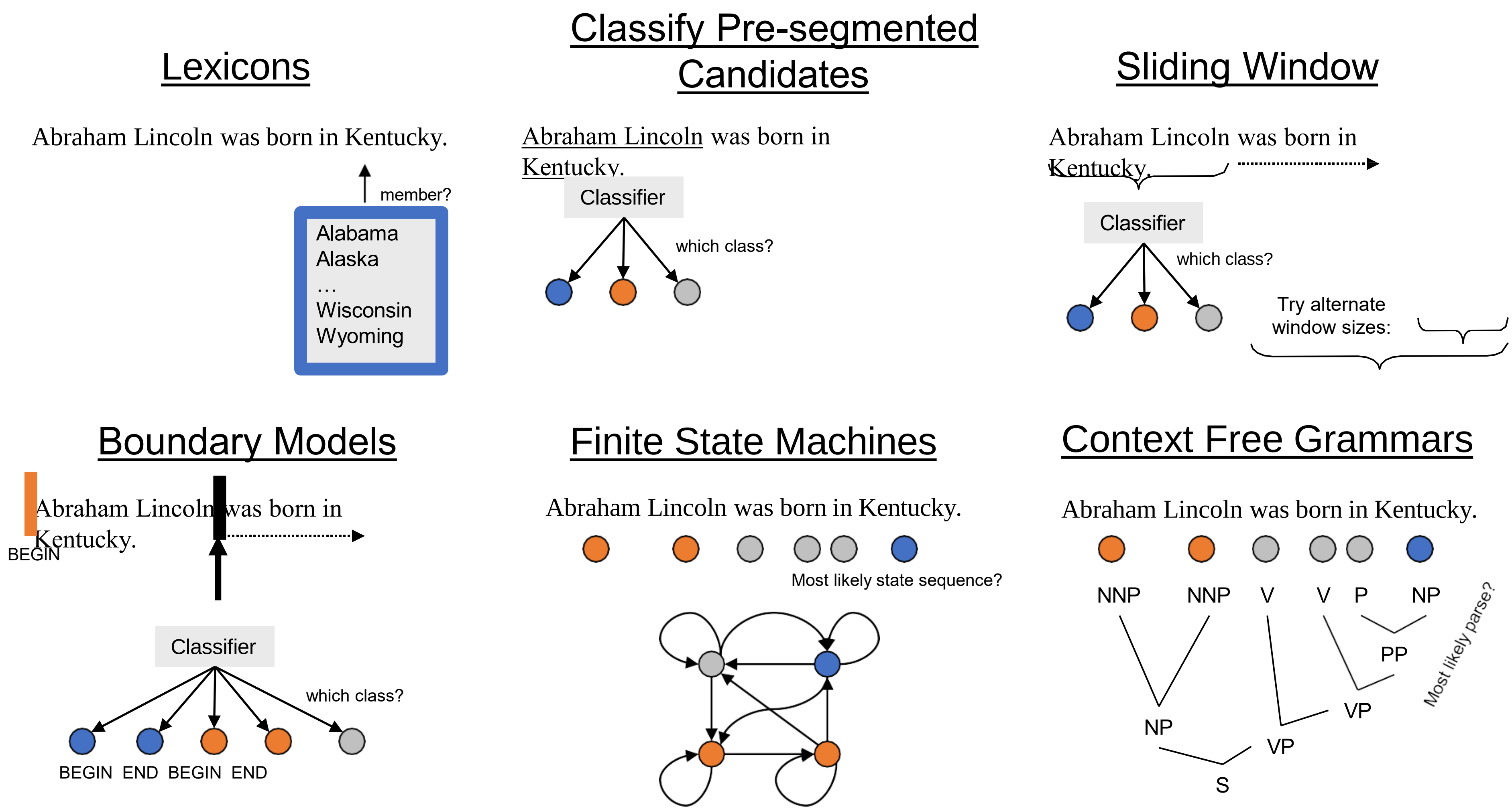
- Now use Wordnet to analyze words in the example
- Problems when using WordNet for IE?

Exercise

Recognize named entites and propose rules:

- 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.

IE techniques: models



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

Sliding Windows

Sliding window

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

Sliding window

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

Sliding window

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

Sliding window

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

Sliding windows with 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

Estimate Pr(LOCATION|window) using Bayes
rules
Try all possible sliding windows (change length and position)

Use independence assumption with length, prefix, suffix, and content words

Evaluate from data: Pr("Place" in prefix|LOCATION)

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

Other method: decision tree on single words and their contexts

Sliding windows with Naïve Bayes: performance

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%

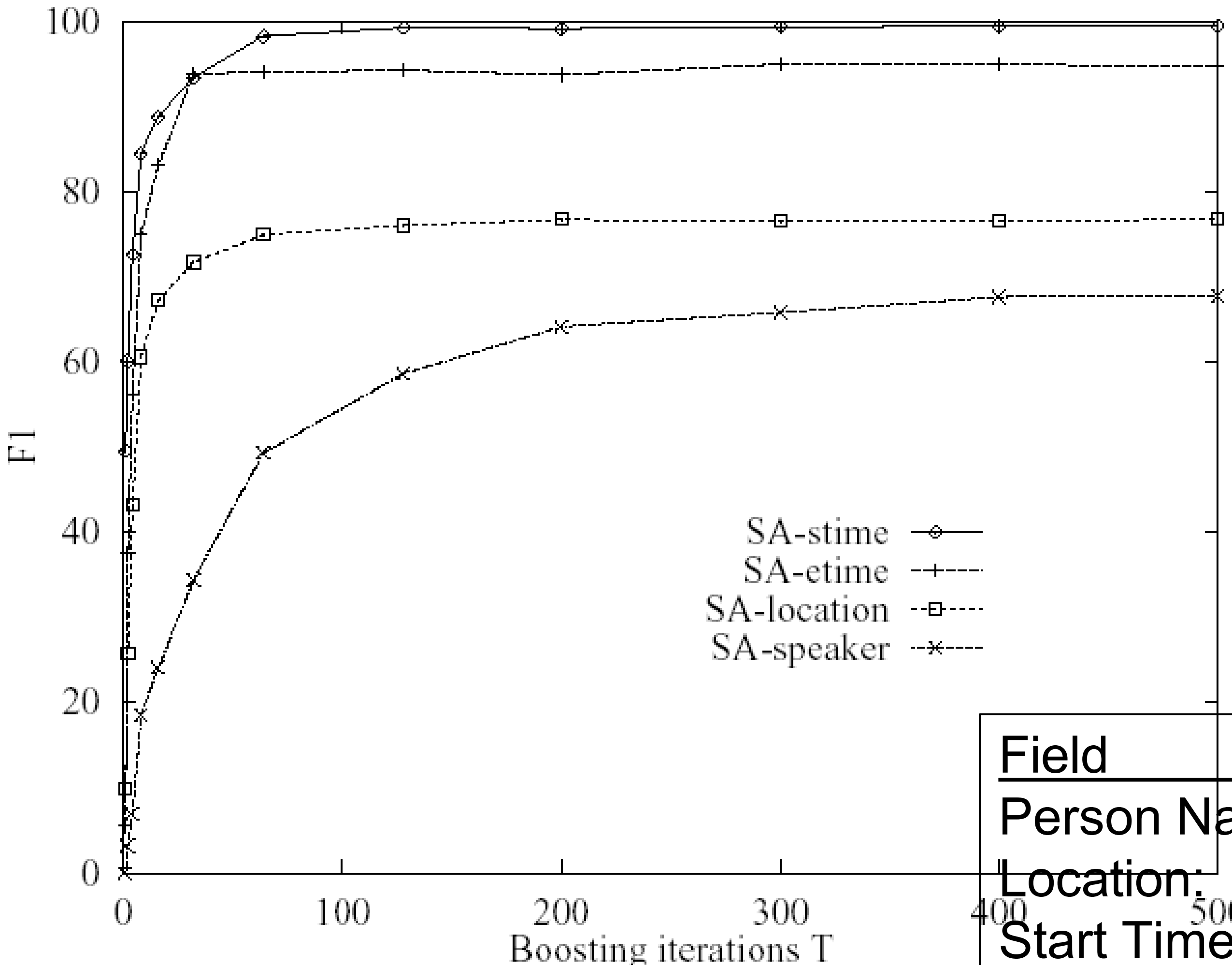
BWI

[Freitag & Kushmerick, AAAI 2000]

- Estimate probability for three classes:
  - *START(i)* = Prob(*i* is start of a field)
  - *END(j)* = Prob(*j* is end of a field)
  - *LEN(k)* = Prob(extract field with length k)
- Extraction score *(i,j):*
  *START(i) * END(j) * LEN(j-i)*

- *LEN(k)* is estimated based on histogram

BWI: Learn to detect boundary



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

Problems with sliding window and BWI

- Decisions on neighbor words are independent
  - Naïve Bayes Sliding Window can predict "seminar end time" before "seminar start time".
  - In BWI, searching for left and right boundaries is independent

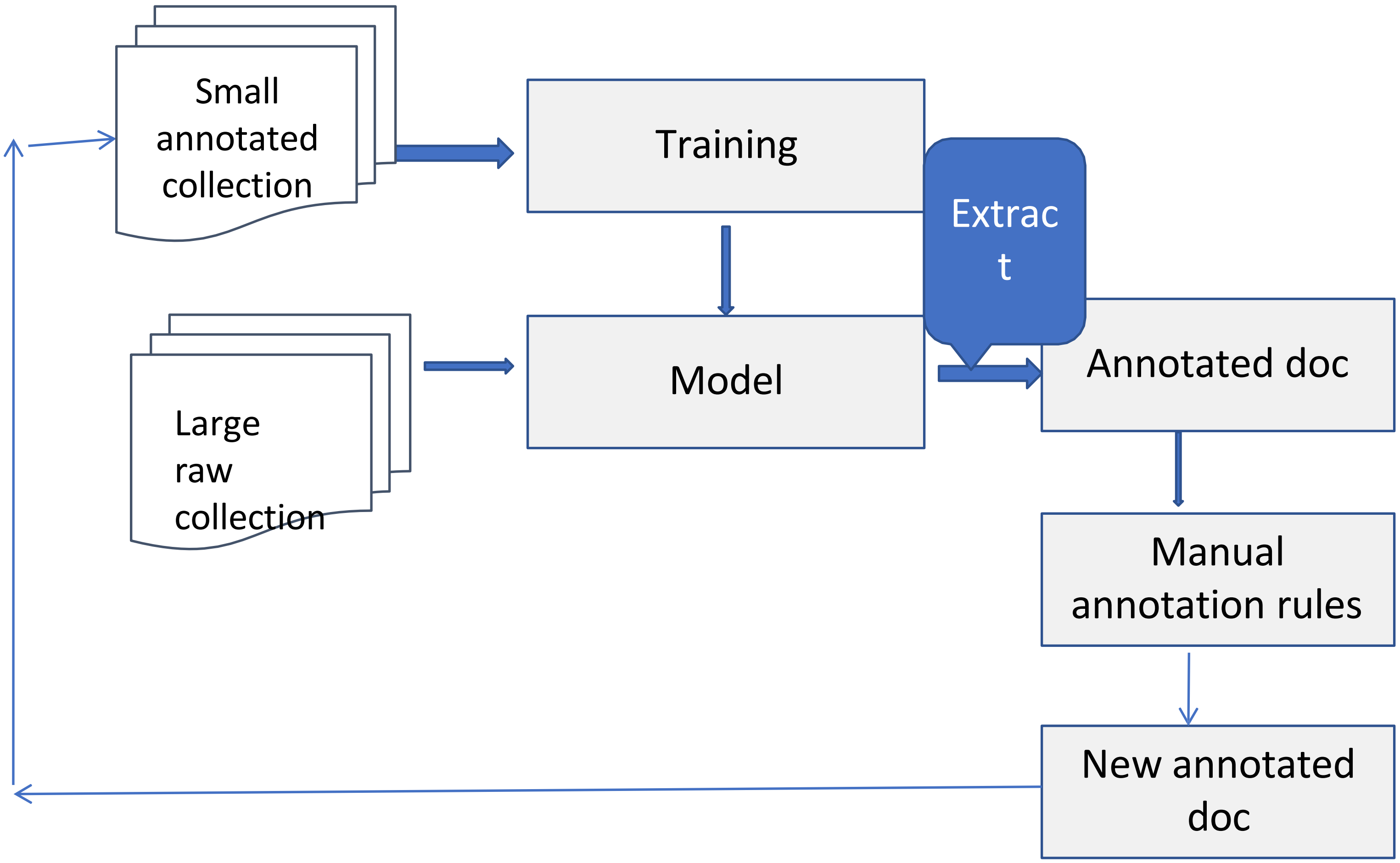
Semi-supervised learning based on coreference

[Sam Chanrathany, 2014]

Observation

- NER is able to recognize entities with contexts in the training data.
- Entites can have multiple mentions in the document in different contexts

Semi-supervised learning based on coreference



Coreference rules in Vietnamese

N_1 and N_2 are corefered if

1. Same name
2. N1 is part of N2, e.g: "*Mai Liêm Trực*" và "*Trực*".
3. Alias, e.g: "*Sài Gòn*" và "*TP Hồ Chí Minh*".
4. Abbreviation, e.g: "*IBM*" và "*International Bussiness Machines*".
5. *first k* and last m letters are the same, $k + m$ is the number of characters in N_2, e.g: "*Công ty Cổ phần Đại An*" and "*Công ty Đại An*".

Coreference rules in Vietnamese

6. Except prefix, all letters in N_2 is in N_1 and prefix of N_2 is either the same as N_1 or abrre of N_1, e.g: "*Công ty TNHH Apave Việt Nam*", "*Cty Apave Việt Nam*", "*Công ty Apave*"
7. A name is the last part of the other, e.g: "*Trịnh Chân Trâu*" và "*Chân Trâu*".
8. The last part of a name is abbreviation of last part of the other, the rests are the same, e.g, với "*Bộ Giáo dục và Đạo tạo*" and "*Bộ GD & ĐT*"

Coreference rules in Vietnamese

9. *k* last letters are the same, first part of N_2 is abbre of first part of N_1, N_2 has $k + 1$ letter, e.g: "*Công ty HP VN*" and "*Cty HP VN*".

10. Abbre in N_2 for phrases in N_1 and the rest in N_2 is in N_1, e.g: "*Công ty TNHH Hewlett Packard Việt Nam*", "*Cty HP VN*", "*HP VN*", "*HP Việt Nam*" and "*Công ty HP Việt Nam*"

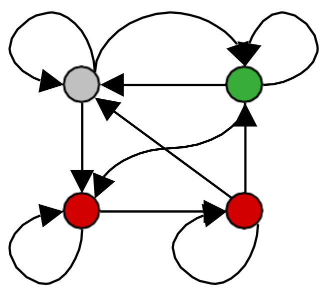
11. $N_1(N_2)$, N_2 has one syllable and is an organization. e.g: "*Phòng Thương mại và Công nghiệp Việt Nam (VCCI)*", or "*Liên đoàn Bóng đá Việt Nam (VFF)*", or "*Tổng công ty Cao su VN (Geruco)*".

Finite State Machines

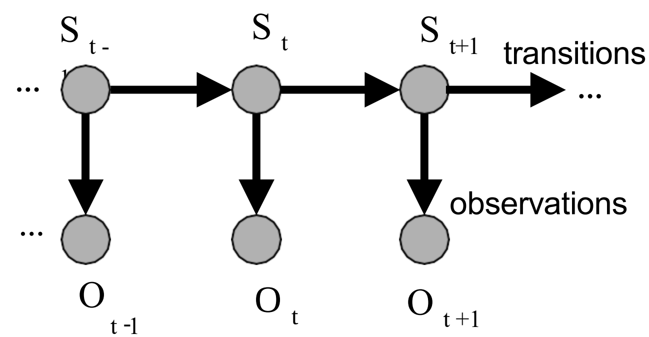
Hidden Markov Models

HMMs is a sequence model used in speech processing, NLP, audio processing...

Finite state model

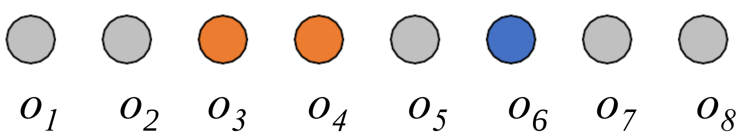


Graphical model



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:

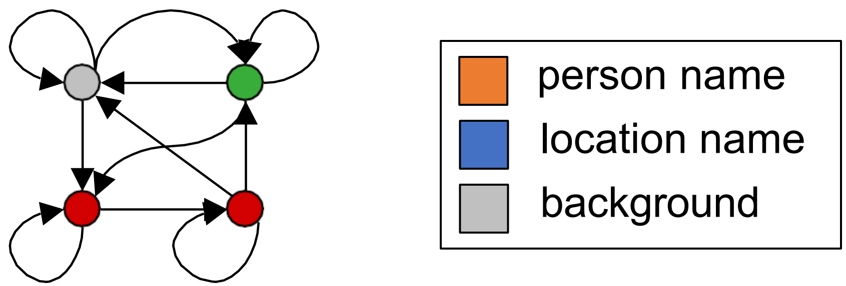
Maximize probability of training observations (w/ prior)

IE with HMM

Given a text

Yesterday Pedro Domingos spoke this example sentence.

and an HMM



Find the best tag sequence $\arg\max_{\vec{s}} P(\vec{s}, \vec{o})$

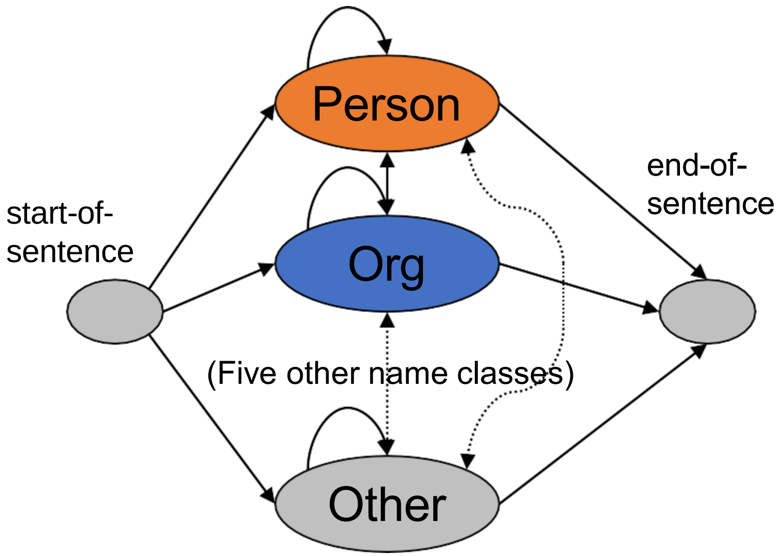
Yesterday Pedro Domingos spoke this example sentence.

Person name: Pedro Domingos

Example HMM : "Nymble"

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

Task: Named Entity Extraction



Transition probabilities	Observation probabilities
$P(s_t \mid s_{t-1}, o_{t-1})$	$P(o_t \mid s_t, s_{t-1})$ or $P(o_t \mid s_t, o_{t-1})$
Back-off to:	Back-off to:
$P(s_t \mid s_{t-1})$	$P(o_t \mid s_t)$
$P(s_t)$	$P(o_t)$

Trained on ~500k words from news

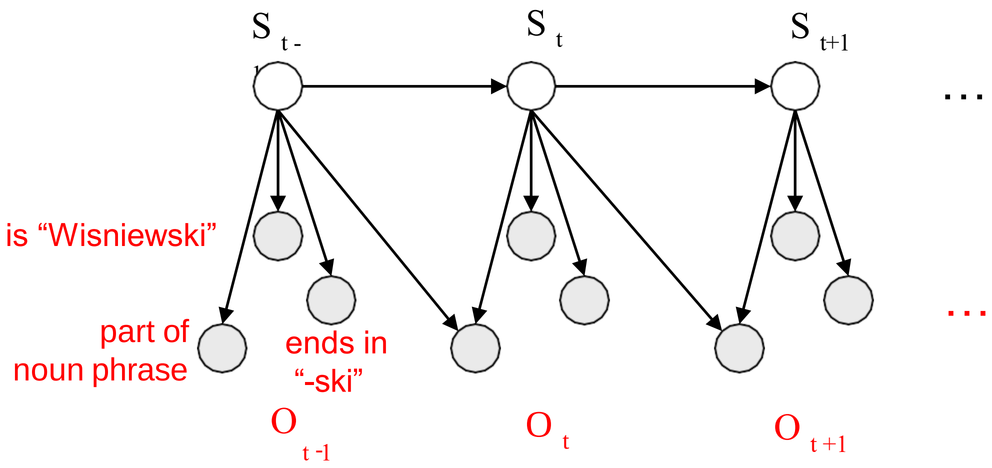
Result:

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

More complicated models

Overlapped features

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”



Problems

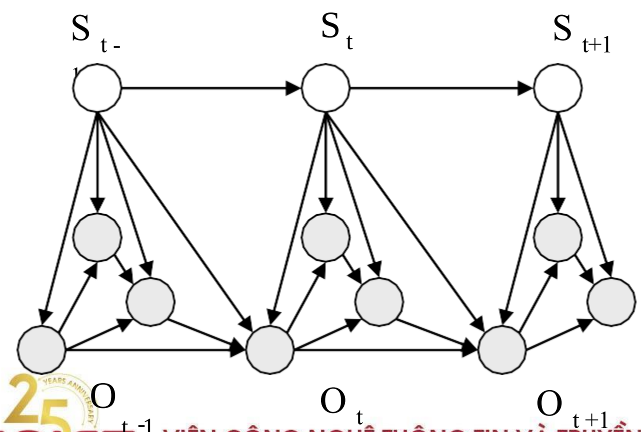
Dependent features

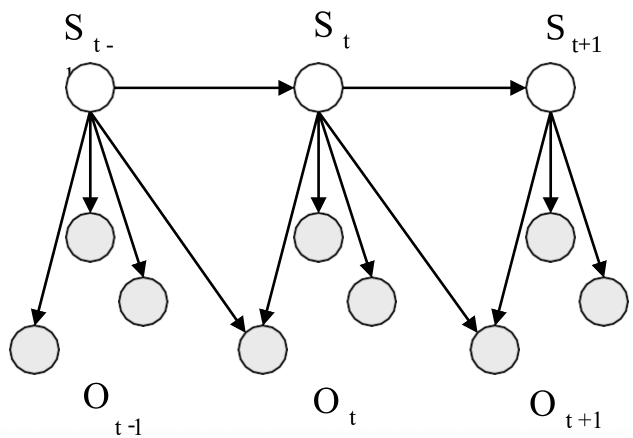
- Multiple unit levels: char, word, segment
- Multiple feature types: word, word shape, pattern

Two choices:

Modeling dependence
Each state has a Bayesian network. But lack of training data

Ignore dependence
Repeatly count events (naïve Bayes). Big problem when combining events





Conditional Sequence Models

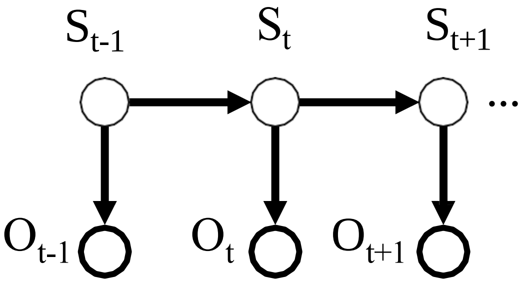
- Maximize conditional prob instead of joint prob
  P($\overline{s}$|$\overline{o}$) instead of P($\overline{s}$,$\overline{o}$):

  - Could check features, but do not generate them
  - Could not explicitly model feature dependence

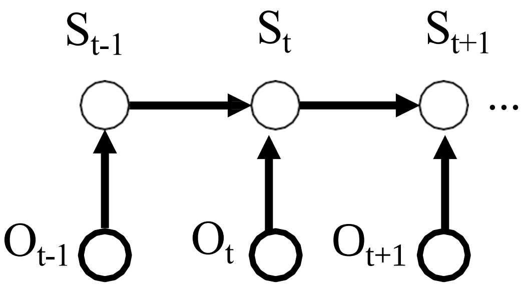
Conditional Markov Models (CMMs) vs HMMS

HMM

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



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



Several ways to infer Pr(y | x)

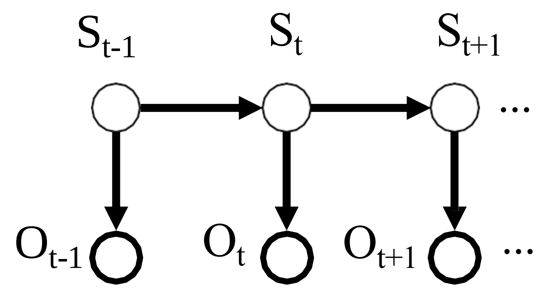
Conditional Finite State Sequence Models

[McCallum, Freitag & Pereira, 2000]

[Lafferty, McCallum, Pereira 2001]

From HMMs to CRFs

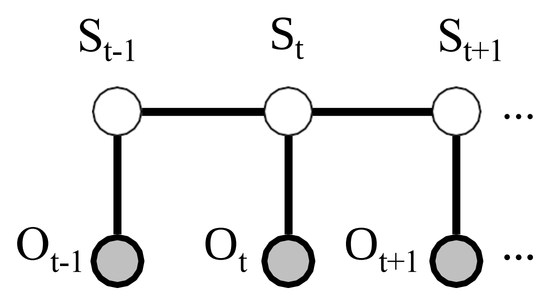
$$\vec{s} = s_1, s_2, \ldots s_n \qquad \vec{o} = o_1, o_2, \ldots o_n$$



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



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

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

(a special case of Conditional Random Fields.)

In which $\Phi_o(t) = \exp\left(\sum_k \boxed{\lambda_k f_k(s_t, o_t)} \right)$

Random features of s,o, and t

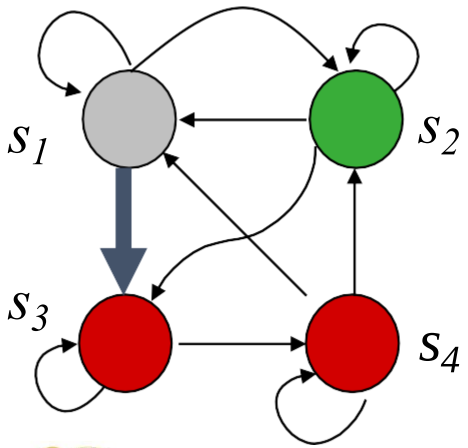
Feature functions

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

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

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

o_1 o_2 o_3 o_4 o_5 o_6 o_7



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

Parameter learning

Given training data D, maximize log-likelihood with $\Lambda = \{\lambda_k\}$

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

Log-likelihood gradient:

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

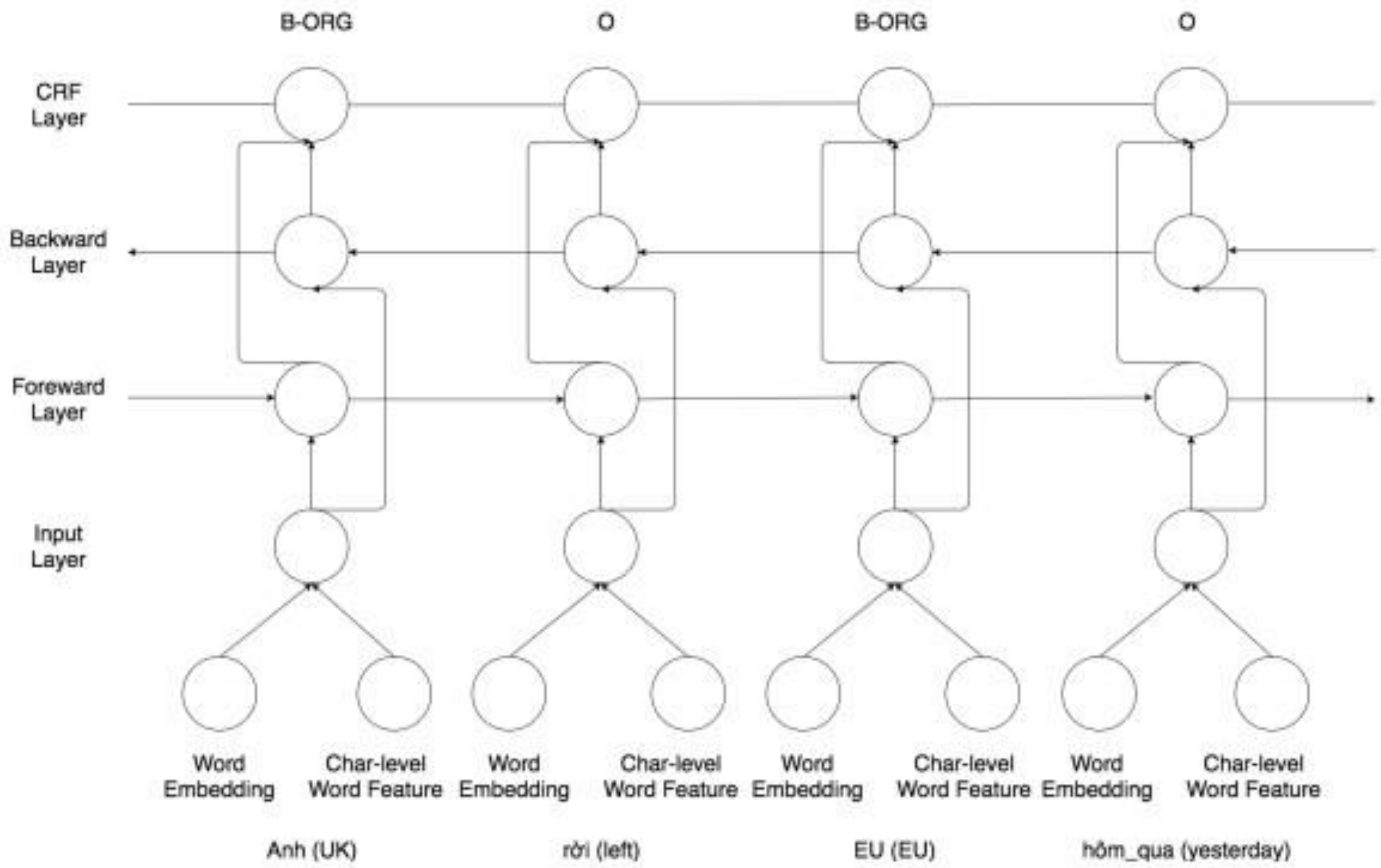
where $\#_k(\vec{s}, \vec{o}) = \sum_t f_k(s_{t-1}, s_t, \vec{o}, t)$

Method:

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

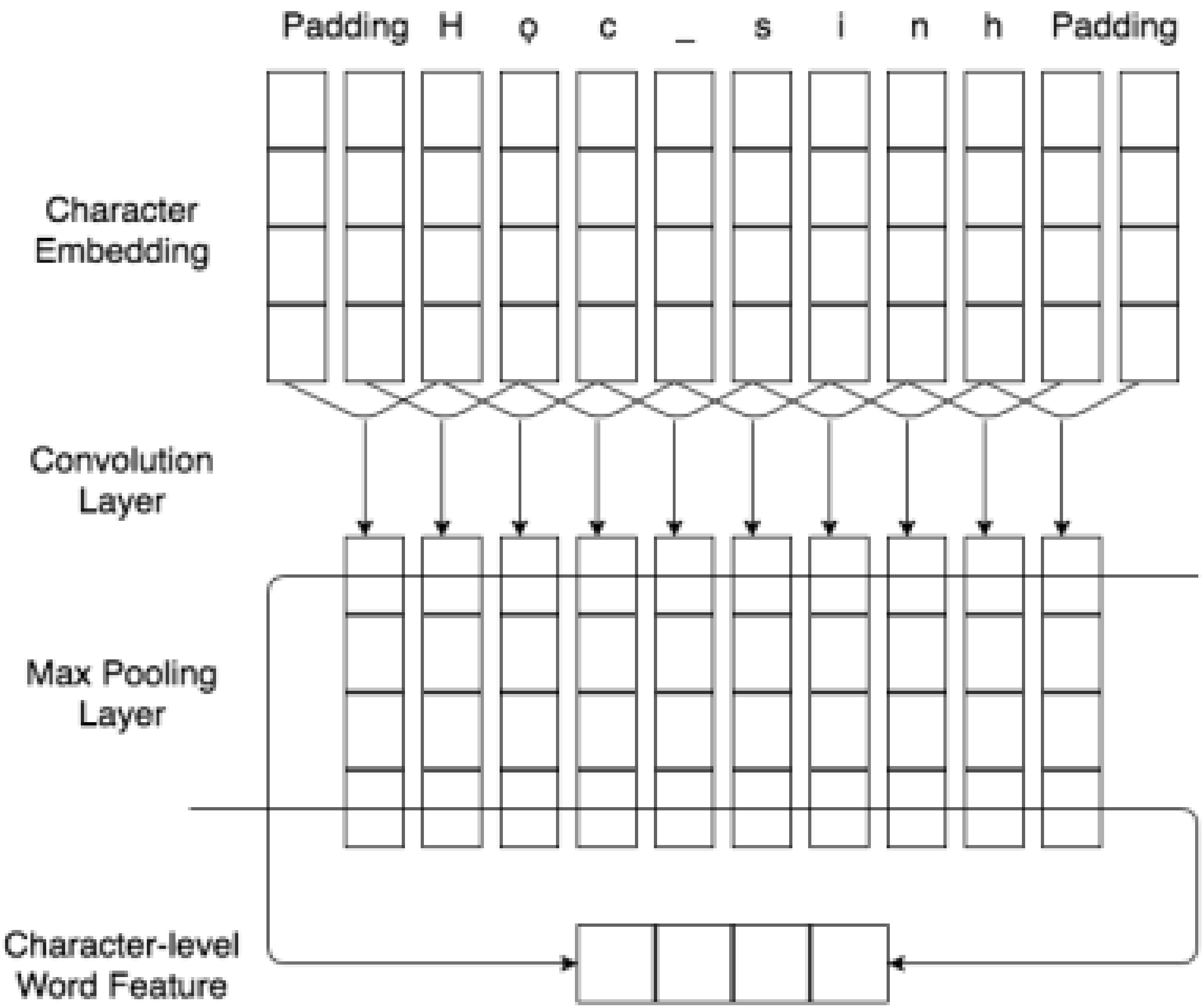
[Sha & Pereira 2002] & [Malouf 2002]

NER using biLSTM + CRF

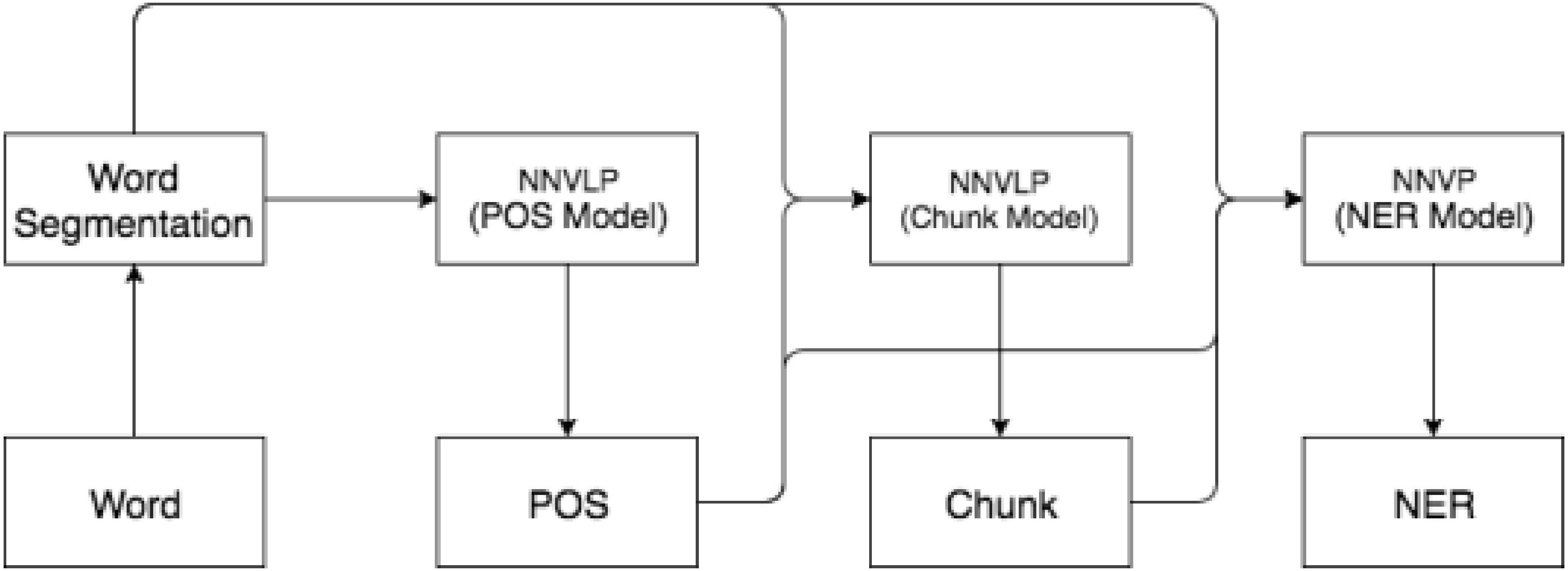


Thai-Hoang Pham, Phuong Le-Hong, "End-to-end Recurrent Neural Network Models for Vietnamese Named Entity Recognition: Word-level vs. Character-level" (PACLING 2017)

NER using biLSTM + CRF



NER using biLSTM + CRF



- Experiments on VLSP, performance F1=91.92%

Working with IE data

- Some characteristics of IE:
  - Based on extraction from documents
  - Noise (lack of events, unnormalized entities)
  - Need data cleansing

- Applications
  - Data mining
  - Fuzzy query *[Cohen 1998]*
  - Use as learning features *[Cohen 2000]*