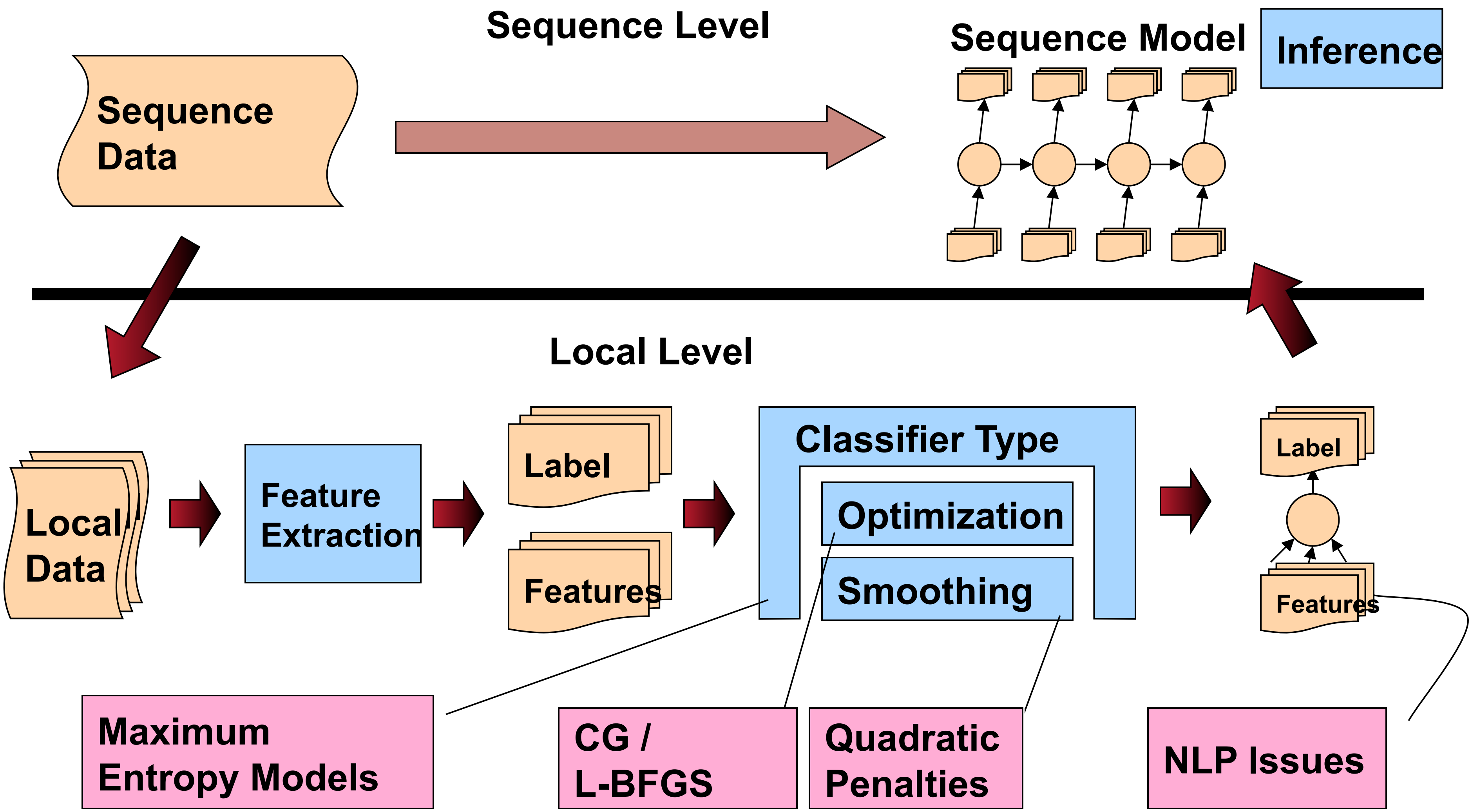
Information Extraction:
Sequence Models, Information Extraction Tasks and Information Integration

CS 224N

2009

Sequence Inference



MEMM inference in systems

- For a Conditional Markov Model (CMM) a.k.a. a Maximum Entropy Markov Model (MEMM), the classifier makes a single decision at a time, conditioned on evidence from observations and previous decisions.
- A larger space of sequences is explored via search

Decision Point

Local Context

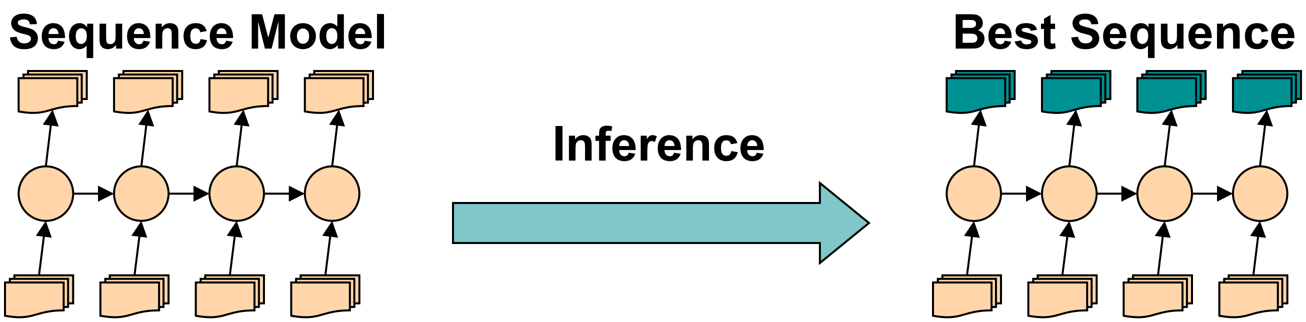
-3	-2	-1	0	+1
DT	NNP	VBD	???	???
The	Dow	fell	22.6	%

Features

W_0	22.6
W_{+1}	%
W_{-1}	fell
T_{-1}	VBD
T_{-1}-T_{-2}	NNP-VBD
hasDigit?	true
…	…

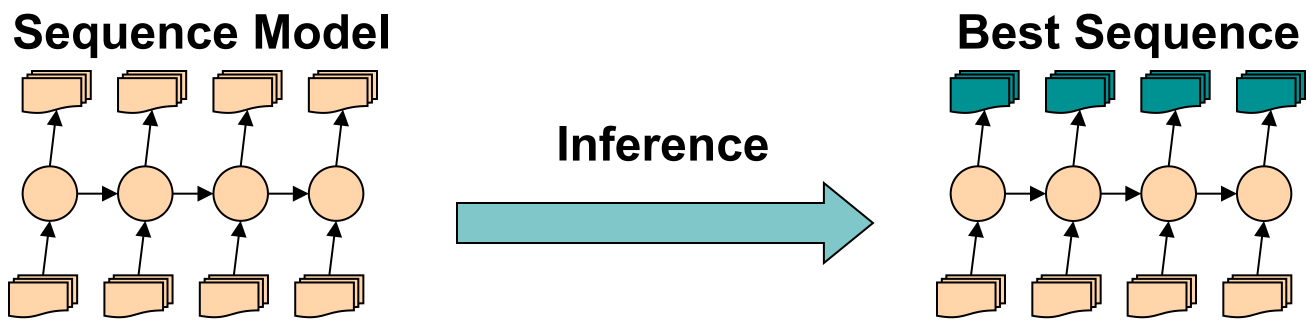
(Ratnaparkhi 1996; Toutanova et al. 2003, etc.)

Two ways to Search: Beam Inference



- Beam inference:
  - At each position keep the top k complete sequences.
  - Extend each sequence in each local way.
  - The extensions compete for the k slots at the next position.
- Advantages:
  - Fast; and beam sizes of 3–5 are as good or almost as good as exact inference in many cases.
  - Easy to implement (no dynamic programming required).
- Disadvantage:
  - Inexact: the globally best sequence can fall off the beam.

Two ways to Search: Viterbi Inference



- Viterbi inference:
  - Dynamic programming or memoization.
  - Requires small window of state influence (e.g., past two states are relevant).
- Advantage:
  - Exact: the global best sequence is returned.
- Disadvantage:
  - Harder to implement long-distance state-state interactions (but beam inference tends not to successfully capture long-distance resurrection of sequences anyway).

Viterbi Inference: J&M Ch. 6

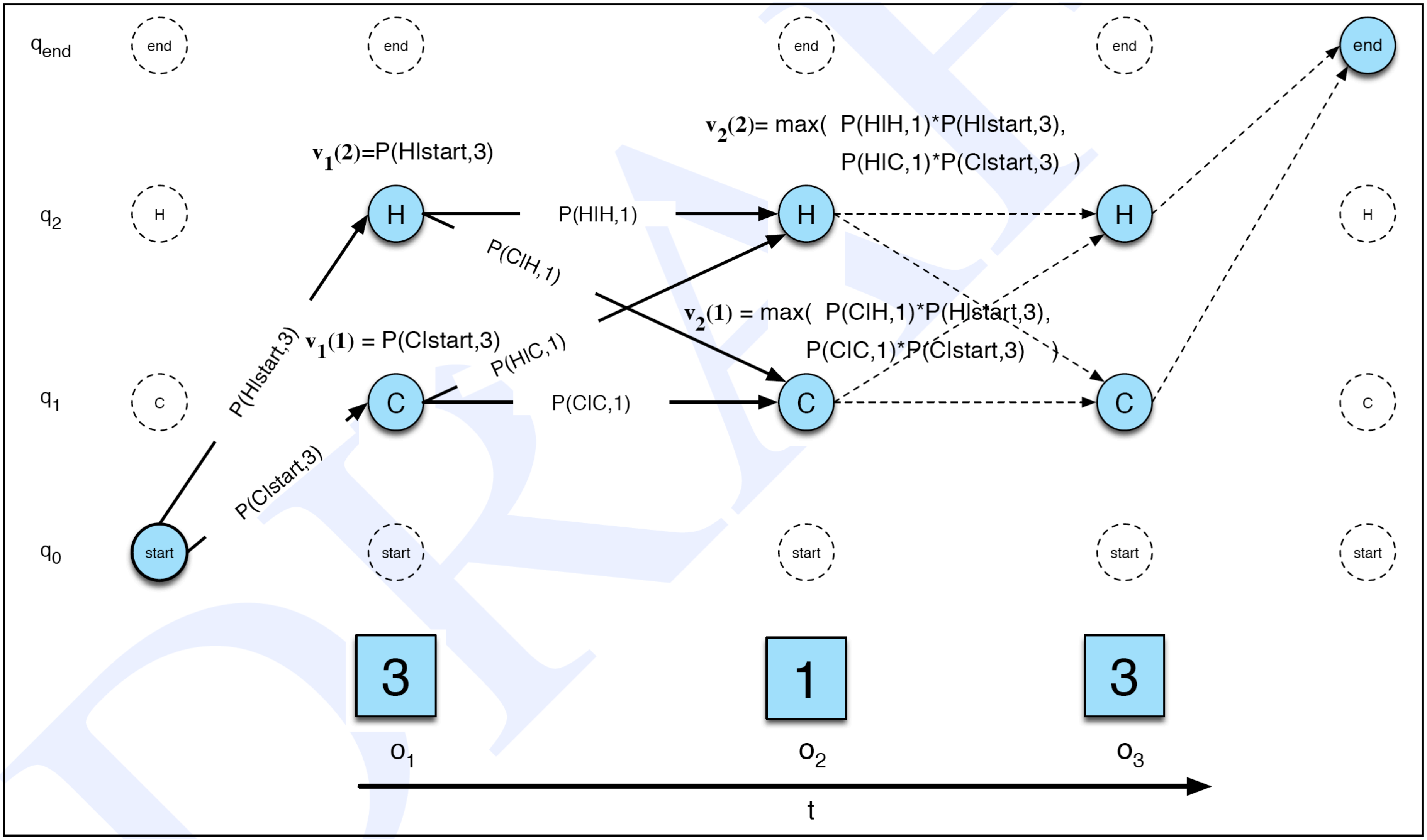
- I'm basically punting on this … read Ch. 6.
  - I'll do dynamic programming for parsing
- It's a small change from HMM Viterbi
  - From:

$$v_t(j) \;=\; \max_{i=1}^{N}\; v_{t-1}(i)\, P(s_j|s_i)\, P(o_t|s_j) \quad 1 \leq j \leq N, 1 < t \leq T$$

  - To:

$$v_t(j) \;=\; \max_{i=1}^{N}\; v_{t-1}(i)\, P(s_j|s_i, o_t) \quad 1 \leq j \leq N, 1 < t \leq T$$

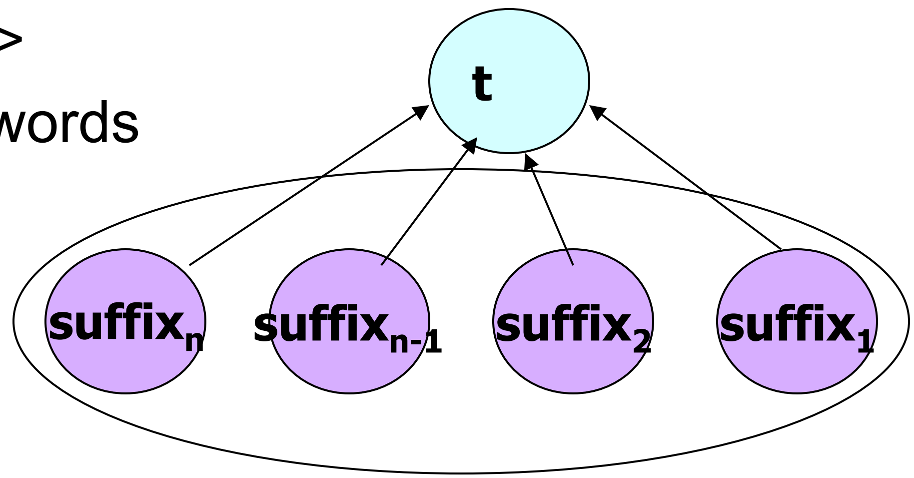
Viterbi Inference: J&M Ch. 6



Part-of-speech tagging: HMM Tagging Models of Brants 2000

- Highly competitive with other state-of-the art models
- Trigram HMM with smoothed transition probabilities
- Capitalization feature becomes part of the state – each tag state is split into two e.g.

  NN → <NN,cap>,<NN,not cap>
- Suffix features for unknown words



$$P(w \mid tag) = P(suffix \mid tag)(w \mid suffix)$$

$$\approx \hat{P}(suffix)\tilde{P}(tag \mid suffix) / \hat{P}(tag)$$

$$\tilde{P}(tag \mid suffix_n) = \lambda_1 \hat{P}(tag \mid suffix_n) + \lambda_2 \hat{P}(tag \mid suffix_{n-1}) + \ldots + \lambda_n \hat{P}(tag)$$

MEMM Tagging Models -II

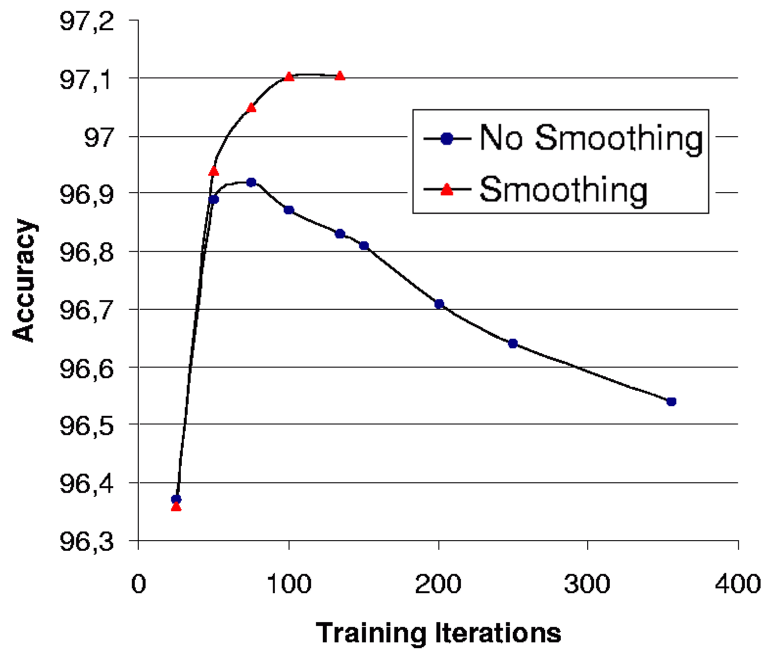
- Ratnaparkhi (1996): local distributions are estimated using maximum entropy models
  - Previous two tags, current word, previous two words, next two words, suffix, prefix, hyphenation, and capitalization features for unknown words
- Toutanova et al. (2003)
  - Richer features, bidirectional inference, better smoothing, better unknown word handling

Model	Overall Accuracy	Unknown Words
MEMM (Ratn. 1996)	96.63	85.56
HMM (Brants 2000)	96.7	85.5
MEMM (T. et al 2003)	97.24	89.04

Smoothing: POS Tagging

- From (Toutanova et al., 2003):

	Overall Accuracy	Unknown Word Acc
Without Smoothing	96.54	85.20
With Smoothing	97.10	88.20



- Smoothing helps:
  - Softens distributions.
  - Pushes weight onto more explanatory features.
  - Allows many features to be dumped safely into the mix.
  - Speeds up convergence (if both are allowed to converge)!

Summary of POS Tagging

For POS tagging, the change from generative to discriminative model **does not by itself** result in great improvement

One profits from discriminative models for specifying dependence on **overlapping features of the observation** such as spelling, suffix analysis, etc.

This additional power (of the CMM ,CRF, Perceptron models) has been shown to result in improvements in accuracy

A CMM allows integration of rich features of the observations, but can suffer from assuming independence from following observations; this effect can be relieved by moving to a CRF, but also by adding dependence on following words

The **higher accuracy** of discriminative models comes at the price of **much slower training**

CRFs [Lafferty, Pereira, and McCallum 2001]

- Another sequence model: Conditional Random Fields (CRFs)
- A whole-sequence conditional model rather than a chaining of local models.

$$P(c \mid d, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

- The space of c's is now the space of sequences
  - But if the features f_i remain local, the conditional sequence likelihood can still be calculated exactly using dynamic programming
- Training is slower, but CRFs avoid causal-competition biases
- These (or a variant using a max margin criterion) are seen as the state-of-the-art these days, and fairly standardly used

NER Results:
CoNLL (2003) Named Entity Recognition task

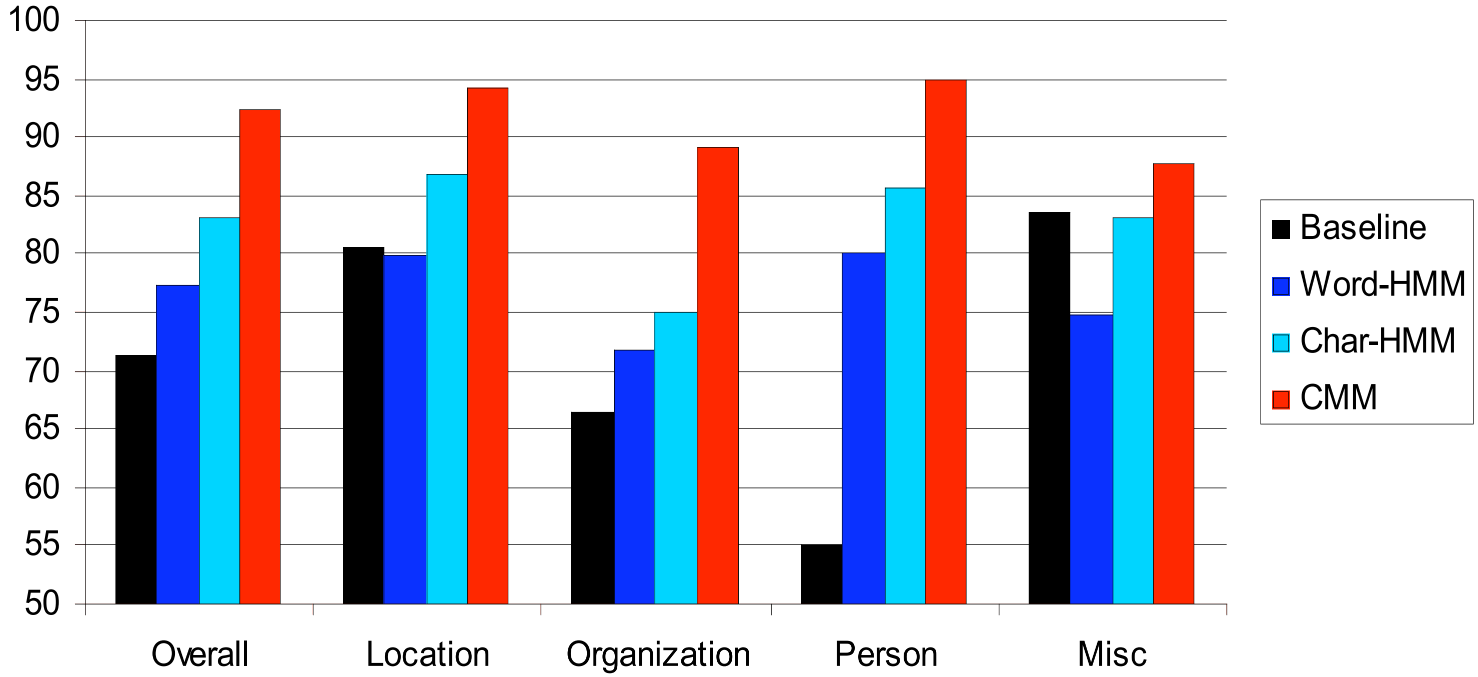
Task: Predict semantic label of each word in text

Foreign	NNP	I-NP	ORG
Ministry	NNP	I-NP	ORG
spokesman	NN	I-NP	O
Shen	NNP	I-NP	PER
Guofang	NNP	I-NP	PER
told	VBD	I-VP	O
Reuters	NNP	I-NP	ORG
:	:	:	:

Standard evaluation is per entity, *not* per token

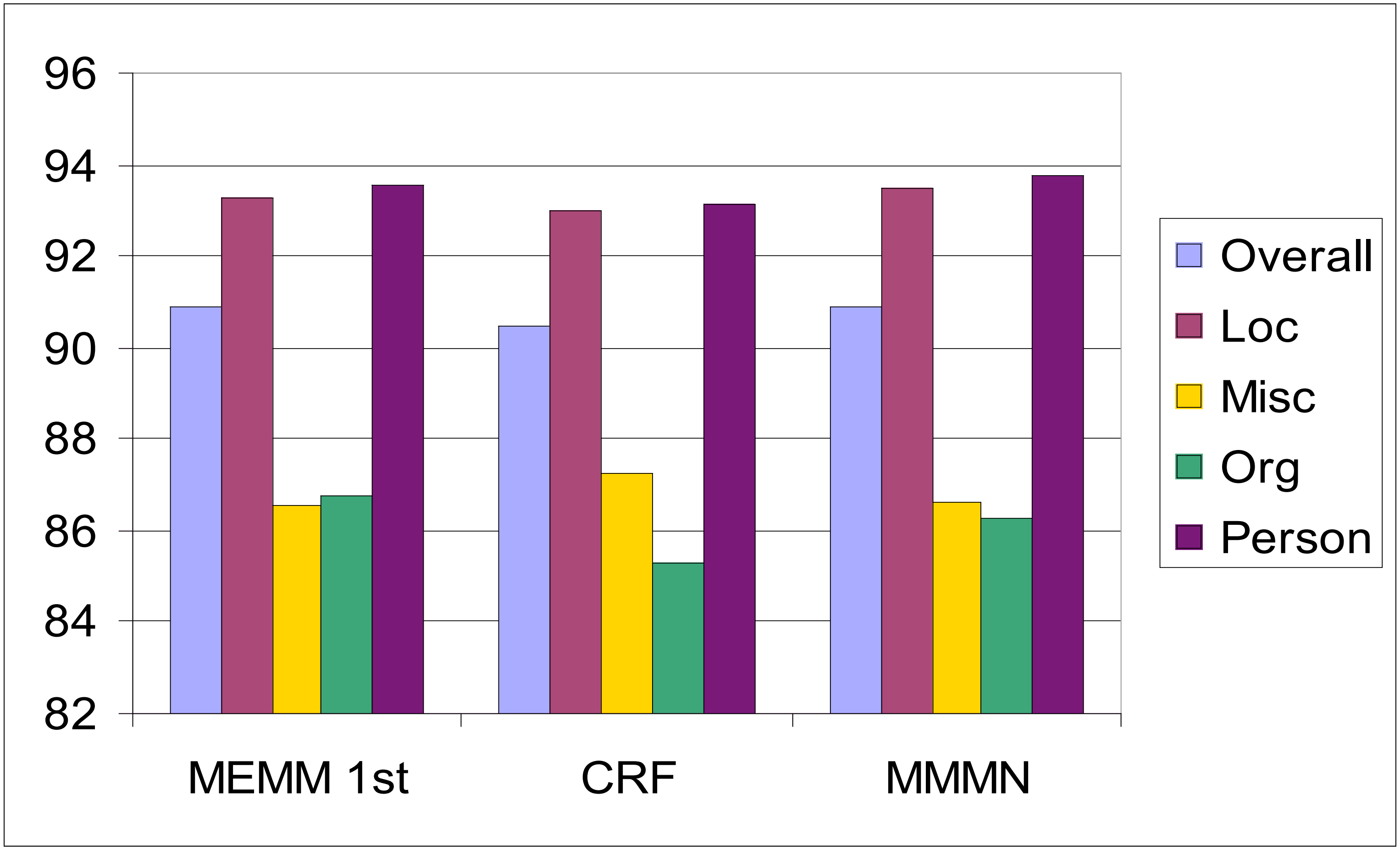
NER Results: Discriminative Model

- Increases from better features, a better classification model.

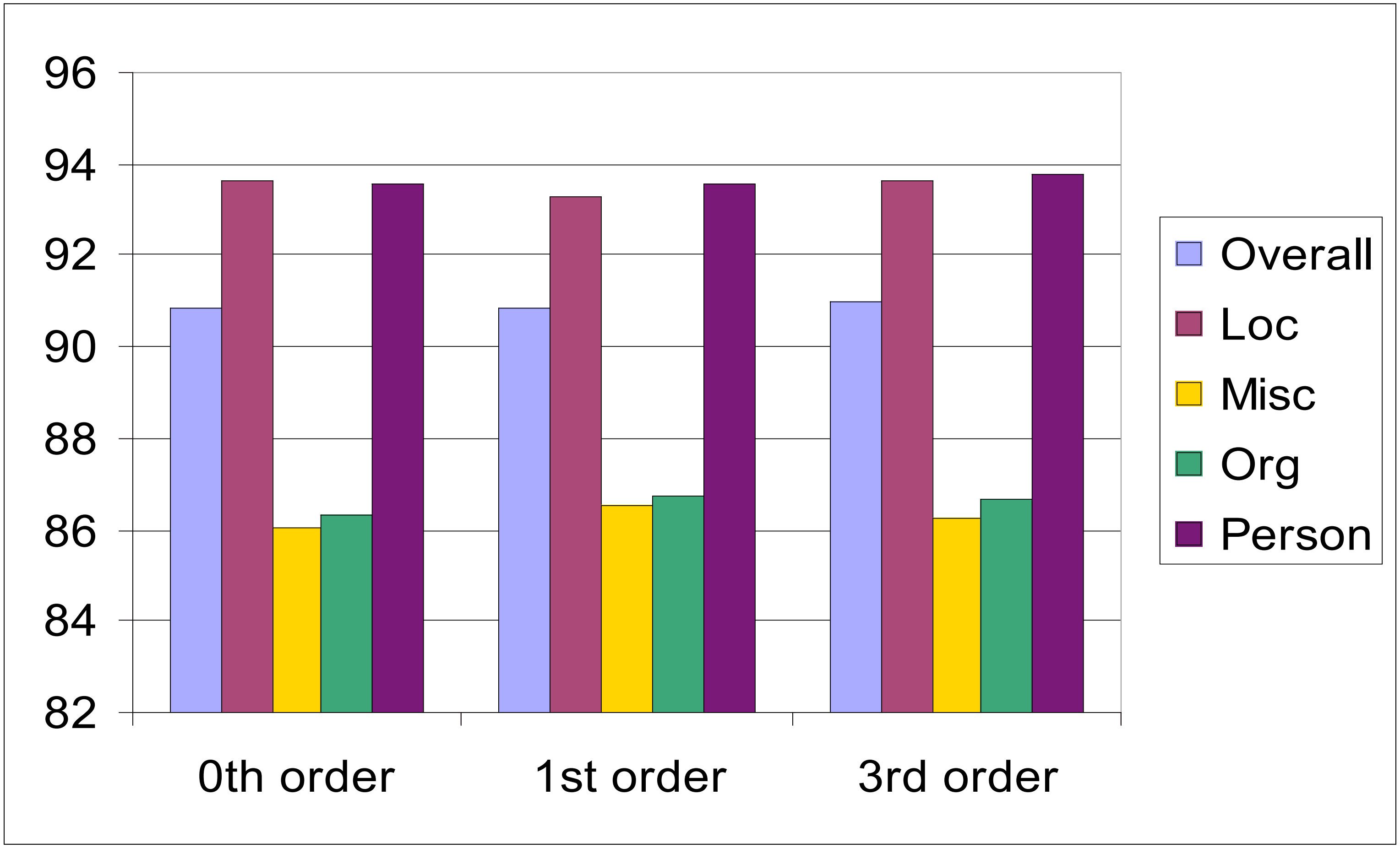


CoNLL 2003 Shared Task: English NER; entity precision/recall F1

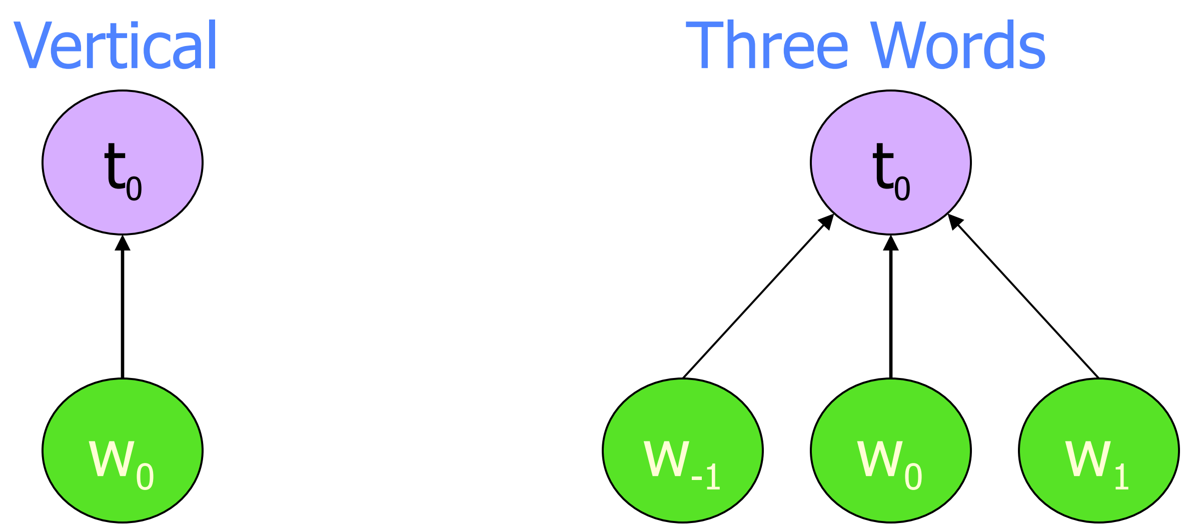
Sequence models? CoNLL 2003 NER shared task Results on English Devset



CoNLL NER Results: CMM Order



Sequence Tagging Without Sequence Information: POS tagging

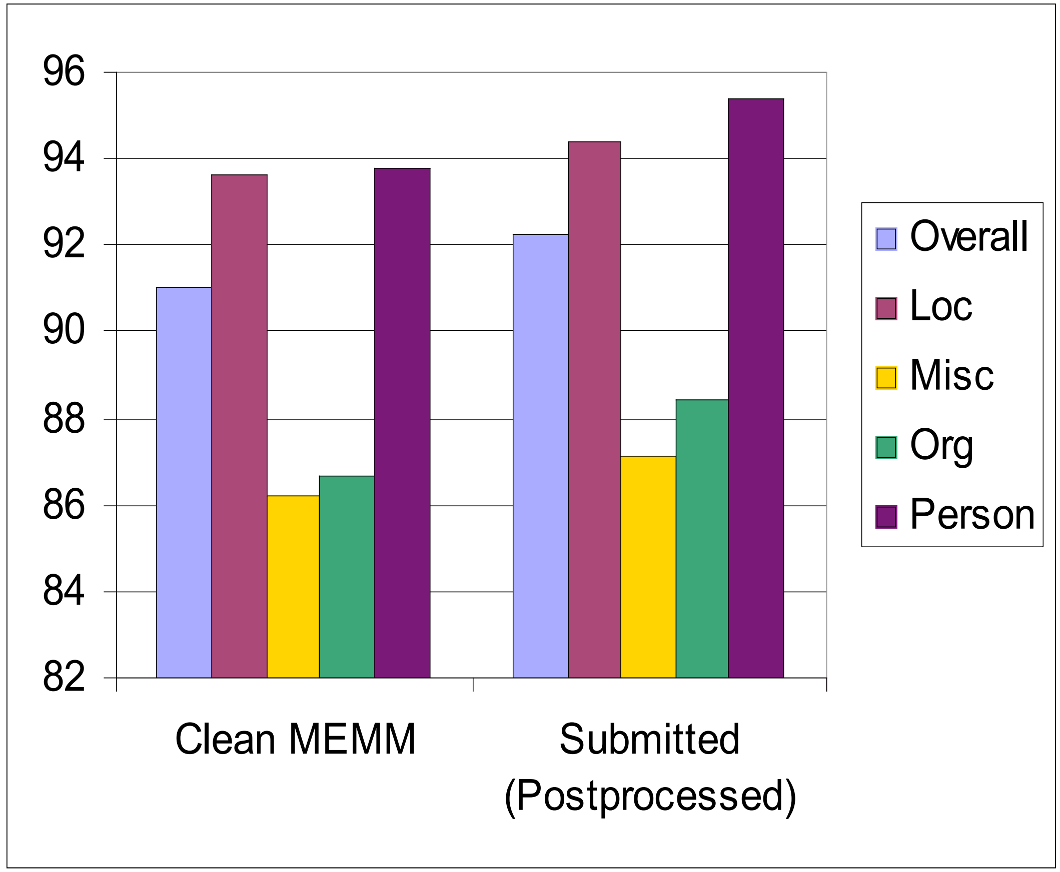


Model	Features	Token	Unknown	Sentence
Vertical	56,805	**93.69%**	82.61%	26.74%
3Words	239,767	**96.57%**	86.78%	48.27%

Using 3 words works significantly better than using only the current word and the previous two or three tags instead! (Toutanova et al. 2003)

CoNLL NER: A real difference

- A difference of about 0.7% gives significance among good CoNLL results
- Here we get one!
- It was done with some Perl regular expressions



Biomedical NER Motivation

- The biomedical world has a huge body of information, which is growing rapidly.
  - MEDLINE, the primary research database serving the biomedical community, currently contains over 12 million abstracts, with 60,000 new abstracts appearing each month.
  - There is also an impressive number of biological databases containing information on genes, proteins, nucleotide and amino acid sequences, including *GenBank*, *Swiss-Prot*, and *Fly-Base*; each contains entries numbering from the thousands to the millions and are multiplying rapidly.

Motivation

- Currently, all of these resources are curated by hand by expert annotators at enormous expense.
- The information overload from the massive growth in the scientific literature has shown the necessity to automatically locate, organize and manage facts relating to experimental results
- Natural Language Processing can aid researchers and curators of biomedical databases by automating these tasks.

Named Entity Recognition

- General NER vs. Biomedical NER

<PER> Christopher Manning </PER> is a professor at <ORG> Stanford University </ORG>, in <LOC> Palo Alto </LOC>.

<RNA> TAR </RNA> independent transactivation by <PROTEIN> Tat </PROTEIN> in cells derived from the <CELL> CNS </CELL> - a novel mechanism of <DNA> HIV-1 gene </DNA> regulation.

Why is this difficult?

- The list of biomedical entities is growing.
  - New genes and proteins are constantly being discovered, so explicitly enumerating and searching against a list of known entities is not scalable.
  - Part of the difficulty lies in identifying previously unseen entities based on contextual, orthographic, and other clues.
- Biomedical entities don’t have strict naming conventions.
  - Common English words such as *period*, *curved*, and *for* are used for gene names.
  - Entity names can be ambiguous. For example, in FlyBase, “clk” is the gene symbol for the “Clock” gene but it also is used as a synonym of the “period” gene.
- Biomedical entity names are ambiguous
  - Experts only agree on whether a word is even a gene or protein 69% of the time. (Krauthammer *et al*., 2000)
  - Often systematic polysemies between gene, RNA, DNA, etc.

Interesting Features

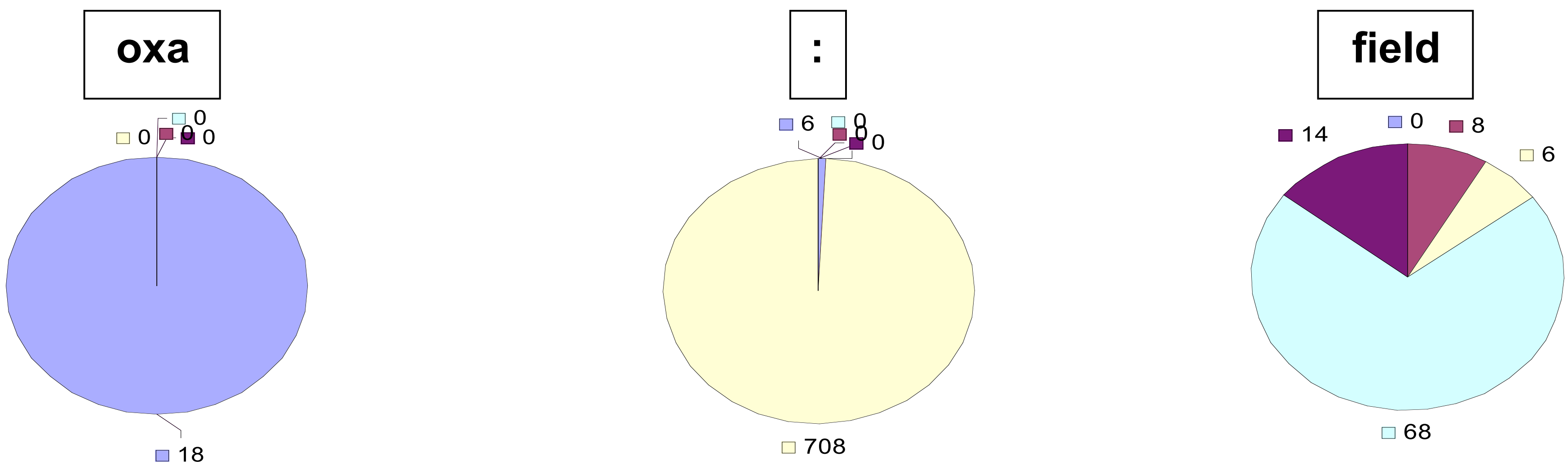
- Word, and surrounding context
- Word Shapes
  - Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd

- Character substrings

#Dpp# → **#Dpp#, #Dpp, Dpp#, #Dp, Dpp, pp#, #D, Dp, pp, p#, D, p**

Features: What's in a Name?



Cotrimoxazole

Wethersfield

Alien Fury: Countdown to Invasion

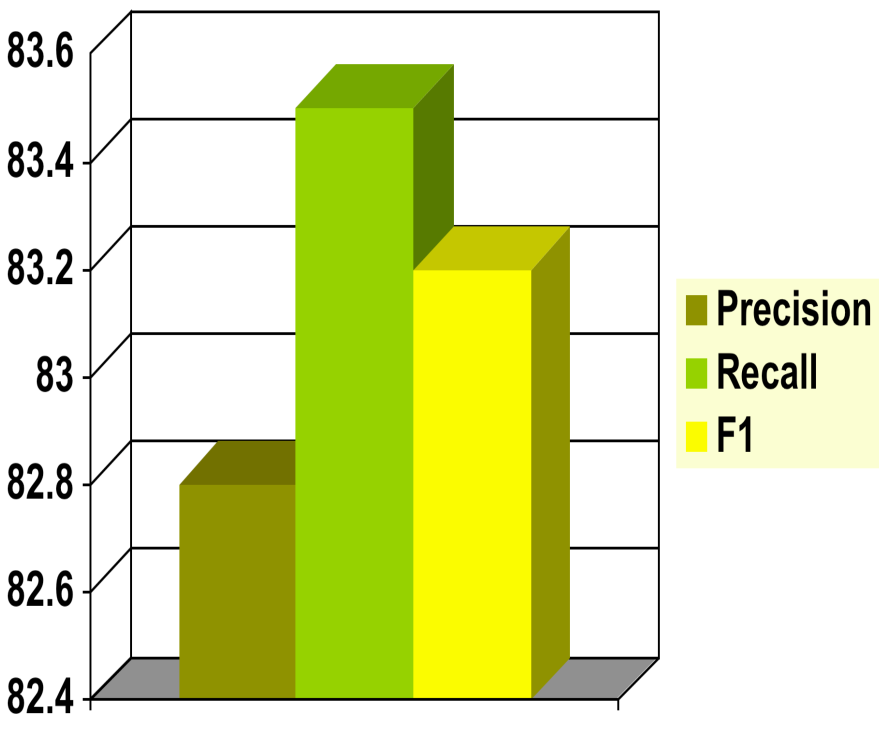
Interesting Features

- Part−of−Speech tags
- Parsing information
- Searching the web for the word in a given context
  - *X gene*, *X mutation*, *X antagonist*
- Gazetteer
  - list words whose classification is known
- Abbreviation extraction (Schwartz and Hearst, 2003)
  - Identify short and long forms when occurring together in text

… Zn finger homeodomain 2 (Zfh 2) …

Finkel et al. (2004) Results

- BioCreative task – Identify genes and proteins



Precision	Recall	F1
81.3%	86.1%	83.6%

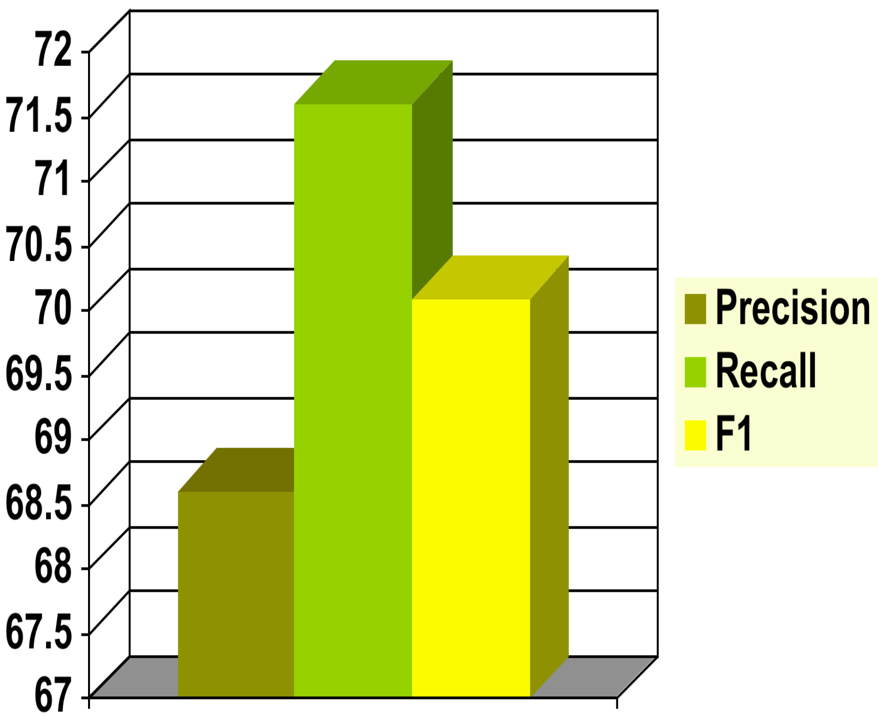
precision = tp / (tp + fp)

recall = tp / (tp + fn)

F1 = 2(precision)(recall) / (precision + recall)

Finkel et al. (2004) Results

- BioNLP task − Identify genes, proteins, DNA, RNA, and cell types



Precision	Recall	F1
68.6%	71.6%	70.1%

precision = tp / (tp + fp)

recall = tp / (tp + fn)

F1 = 2(precision)(recall) / (precision + recall)

Quiz question!

- Answer in one sentence:

In discriminative sequence labeling tasks like NER, why do sequence models (that condition on other labels) often offer little value over using straight classifiers (which don't condition on other labels)?

Information Extraction and Integration

Following slides from:

William Cohen

Andrew McCallum

Eugene Agichtein

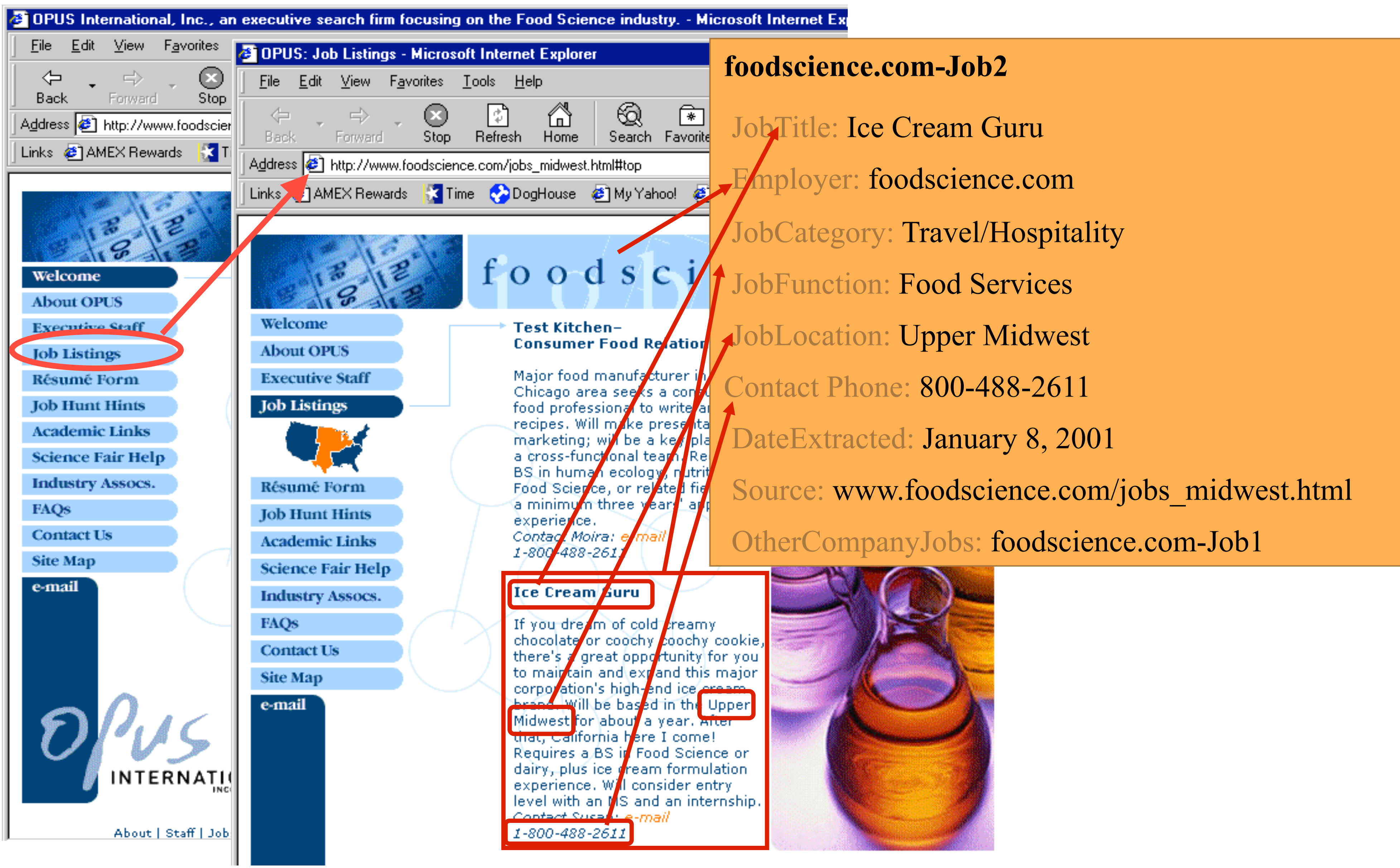
Sunita Sarawagi

The Value of Text Data

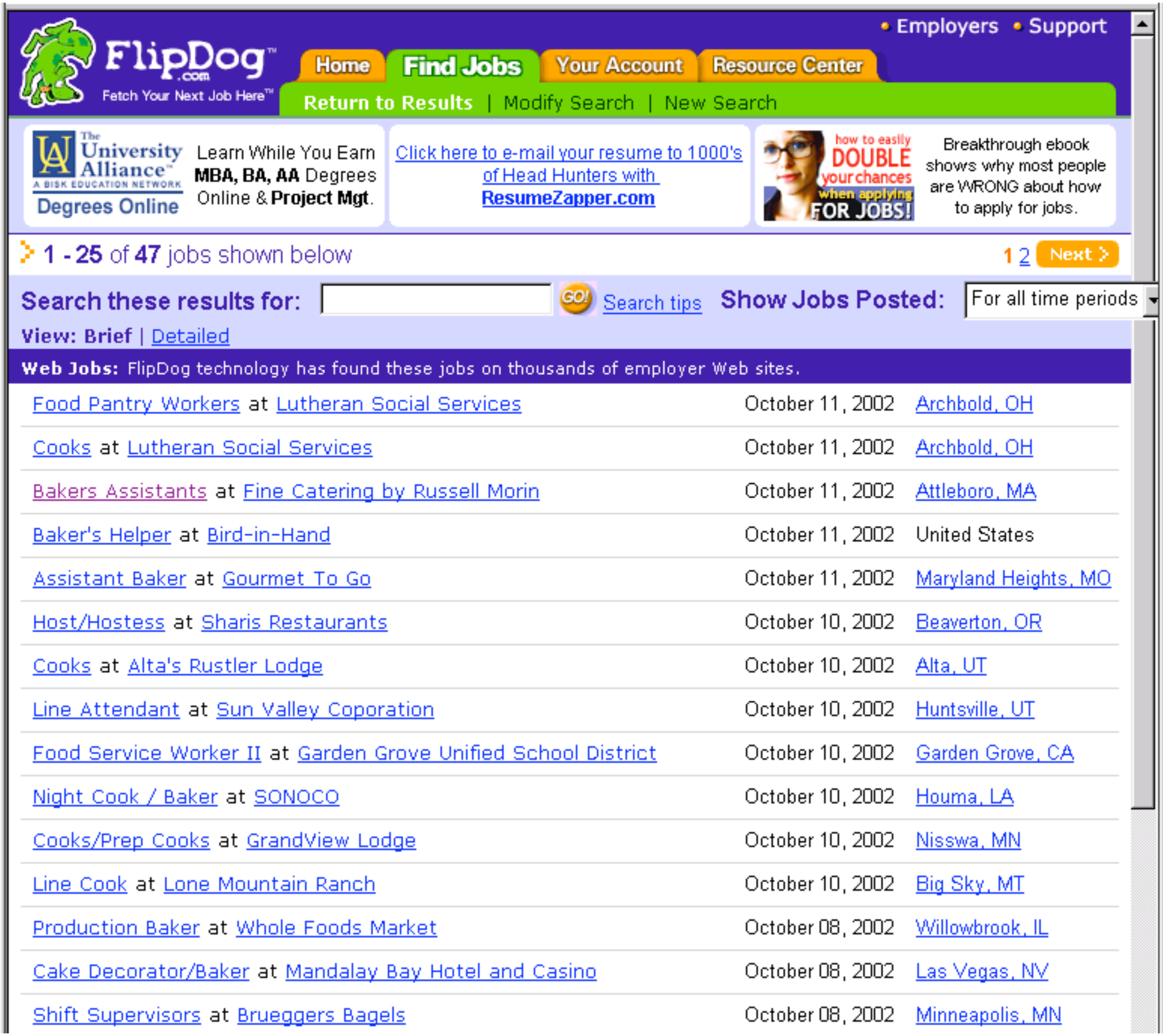
- "Unstructured" text data is the primary source of human-generated information
  – Citeseer, comparison shopping, PIM systems, web search, data warehousing, scientific literature
- Managing and utilizing text: information extraction and integration
- Scalability: a bottleneck for deployment
- Relevance to data mining community

Example: A Solution

Extracting Job Openings from the Web



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



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

What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

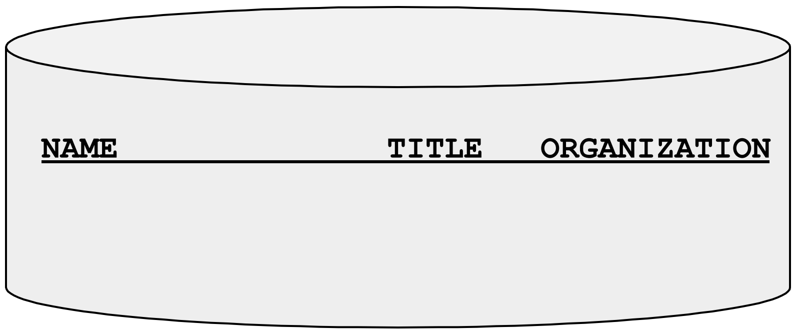
October 14, 2002, 4:00 a.m. PT

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…



What is “Information Extraction”

As a task: **Filling slots in a database from sub-segments of text.**

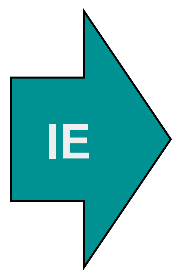
October 14, 2002, 4:00 a.m. PT

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…



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

What is "Information Extraction"

As a family of techniques:

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

What is “Information Extraction”

As a family of techniques:

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

These two steps aka “named entity recognition”

What is "Information Extraction"

As a family of techniques:

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”

As a family of techniques:

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…

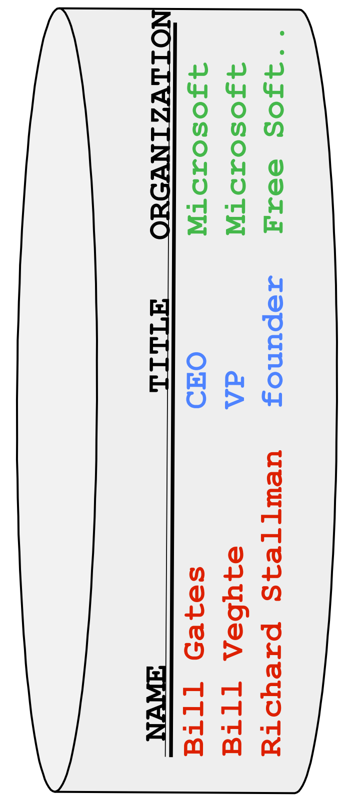
* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates

* Microsoft

Bill Veghte
* Microsoft
VP

Richard Stallman
founder
Free Software Foundation



IE is different in different domains!

Example: on web there is less grammar, but more formatting & linking

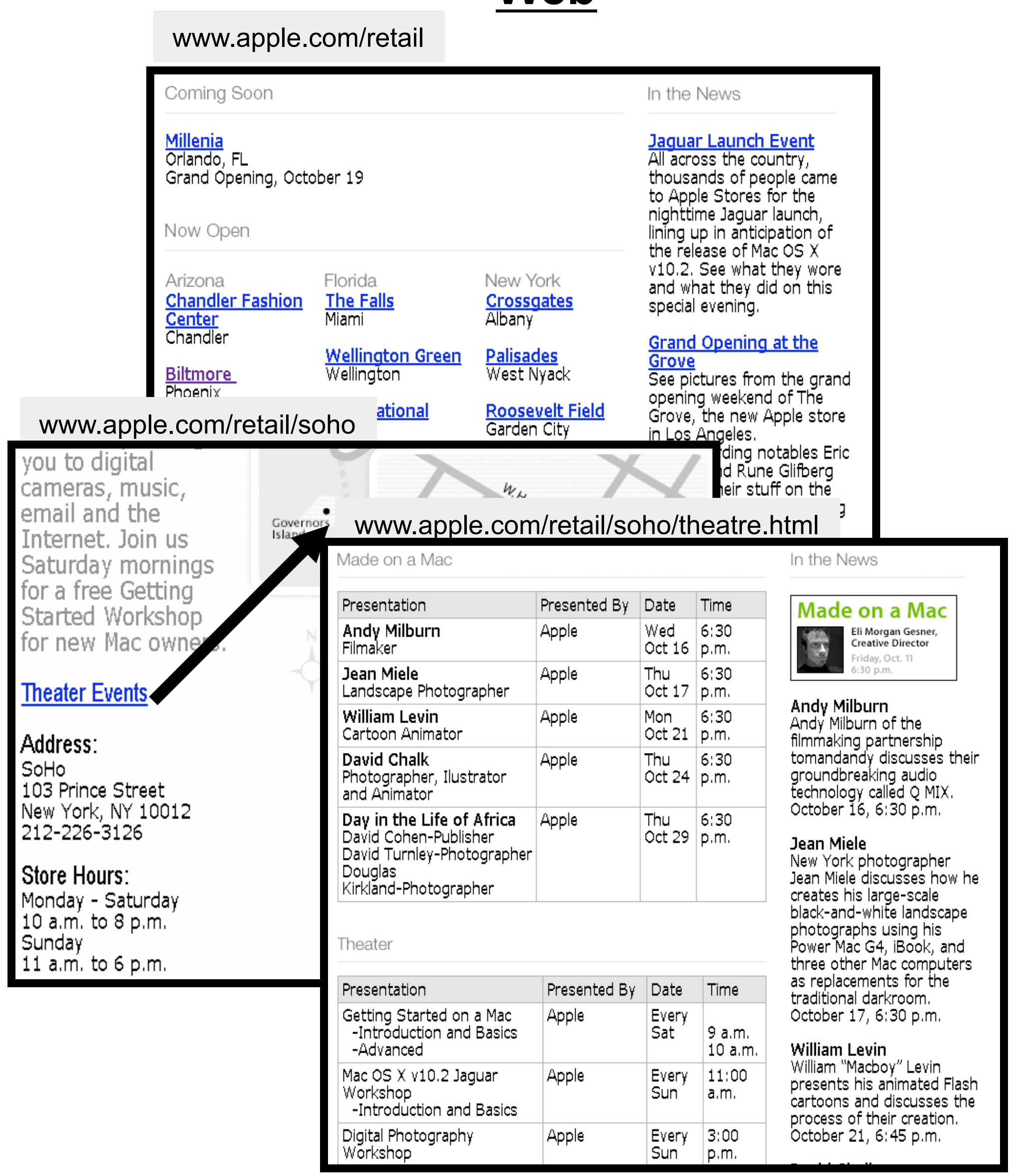
Newswire

Apple to Open Its First Retail Store in New York City

MACWORLD EXPO, NEW YORK--July 17, 2002--Apple's first retail store in New York City will open in Manhattan's SoHo district on Thursday, July 18 at 8:00 a.m. EDT. The SoHo store will be Apple's largest retail store to date and is a stunning example of Apple's commitment to offering customers the world's best computer shopping experience.

"Fourteen months after opening our first retail store, our 31 stores are attracting over 100,000 visitors each week," said Steve Jobs, Apple's CEO. "We hope our SoHo store will surprise and delight both Mac and PC users who want to see everything the Mac can do to enhance their digital lifestyles."

Web



The directory structure, link structure, formatting & layout of the Web is its own new grammar.

Landscape of IE Tasks (1/4): Degree of Formatting

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.

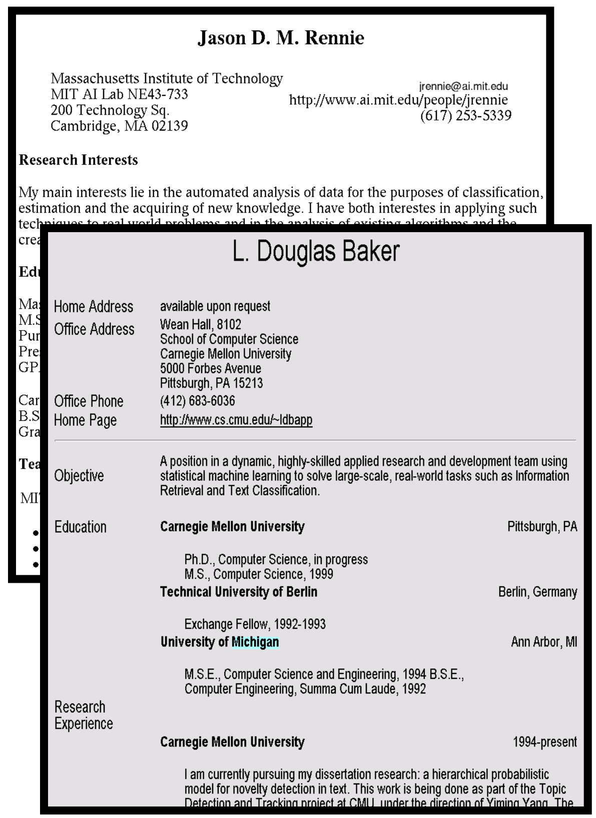
Cohen, Paul R. (413) 545-3638 cohen@cs.umass.edu CS278
Professor.
Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.

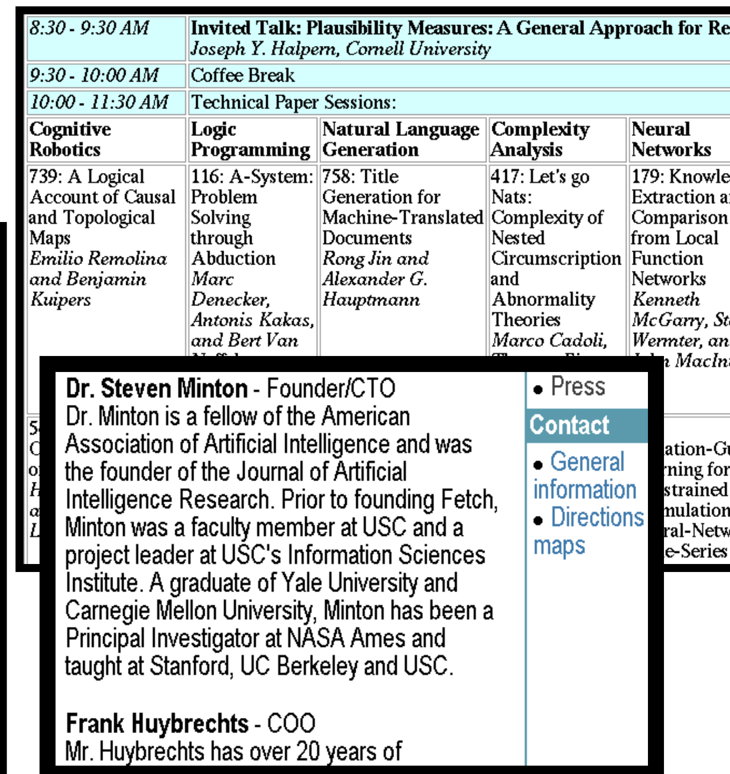
Tables

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

Landscape of IE Tasks (2/4):
Intended Breadth of Coverage

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





Landscape of IE Tasks (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.

Landscape of IE Tasks (4/4):
Single Field/Record

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**

"Named entity" extraction

Binary relationship

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

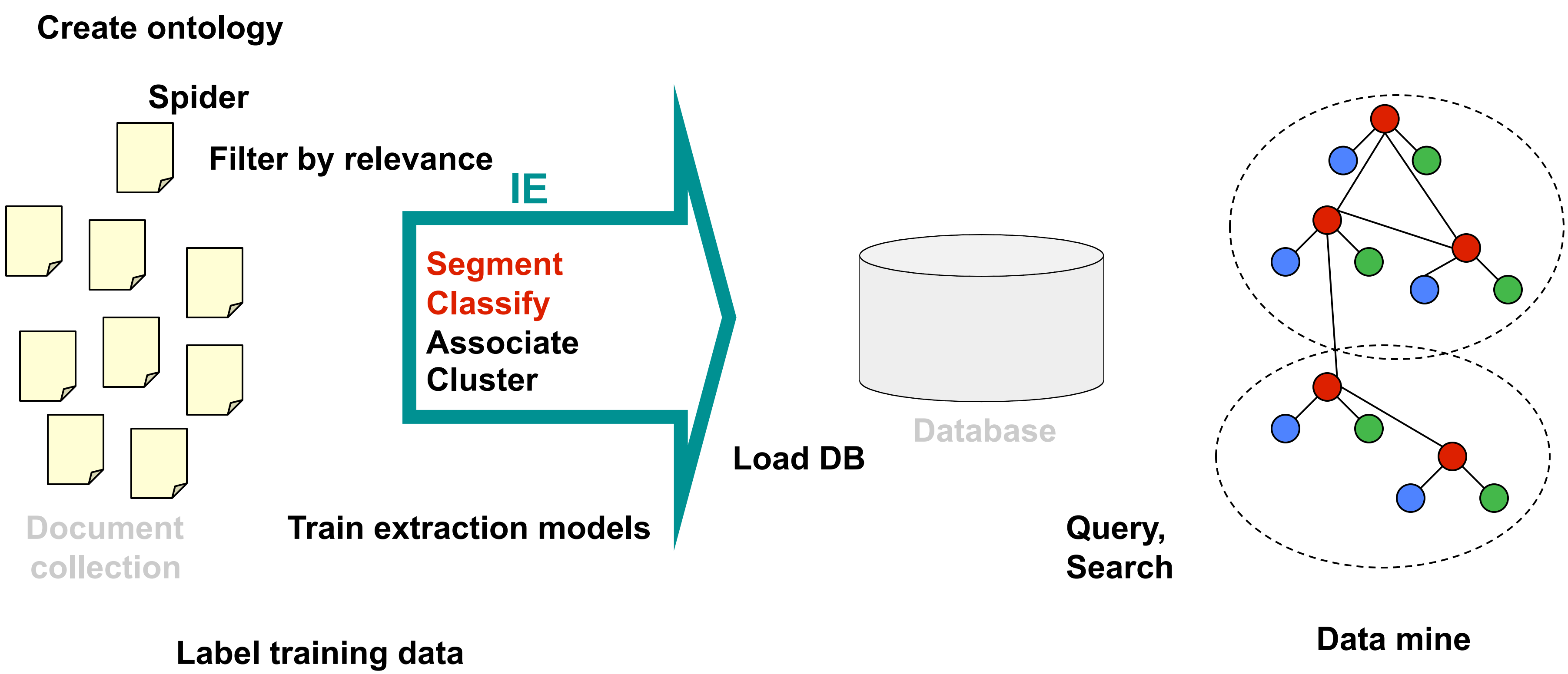
Relation: **Company-Location**
Company: **General Electric**
Location: **Connecticut**

N-ary record

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

Broader View

Up to now we have been focused on ML methods for segmentation and classification



Steps 1 & 2: Hand Coded Rule Example: Conference Name

```
# These are subordinate patterns
$wordOrdinals="(?:first|second|third|fourth|fifth|sixth|seventh|eighth|ninth|tenth|eleventh|twelfth|thirteenth|
fourteenth|fifteenth)";
my $numberOrdinals="(?:\\d?(?:1st|2nd|3rd|1th|2th|3th|4th|5th|6th|7th|8th|9th|0th))";
my $ordinals="(?:$wordOrdinals|$numberOrdinals)";
my $confTypes="(?:Conference|Workshop|Symposium)";
my $words="(?:[A-Z]\\w+\\s*)"; # A word starting with a capital letter and ending with 0 or more spaces
my $confDescriptors="(?:international\\s+|[A-Z]+\\s+)"; # .e.g "International Conference ...' or the conference
name for workshops (e.g. "VLDB Workshop ...")
my $connectors="(?:on|of)";
my $abbreviations="(?:\\([A-Z]\\w\\w+[\\W\\s]*?(?:\\d\\d+)?\\))"; # Conference abbreviations like "(SIGMOD'06)"
# The actual pattern we search for.  A typical conference name this pattern will find is
# "3rd International Conference on Blah Blah Blah (ICBBB-05)"
my $fullNamePattern="((?:$ordinals\\s+$words*|$confDescriptors)?$confTypes(?:\\s+$connectors\\s+.*?|\\s+)?
$abbreviations?)(?:\\n|\\r|\\.|<)";
############################### #################################
# Given a <dbworldMessage>, look for the conference pattern
#################################################################
lookForPattern($dbworldMessage, $fullNamePattern);
###########################################################
# In a given <file>, look for occurrences of <pattern>
# <pattern> is a regular expression
###########################################################
sub lookForPattern {
   my ($file,$pattern) = @_;
```

Machine Learning Methods

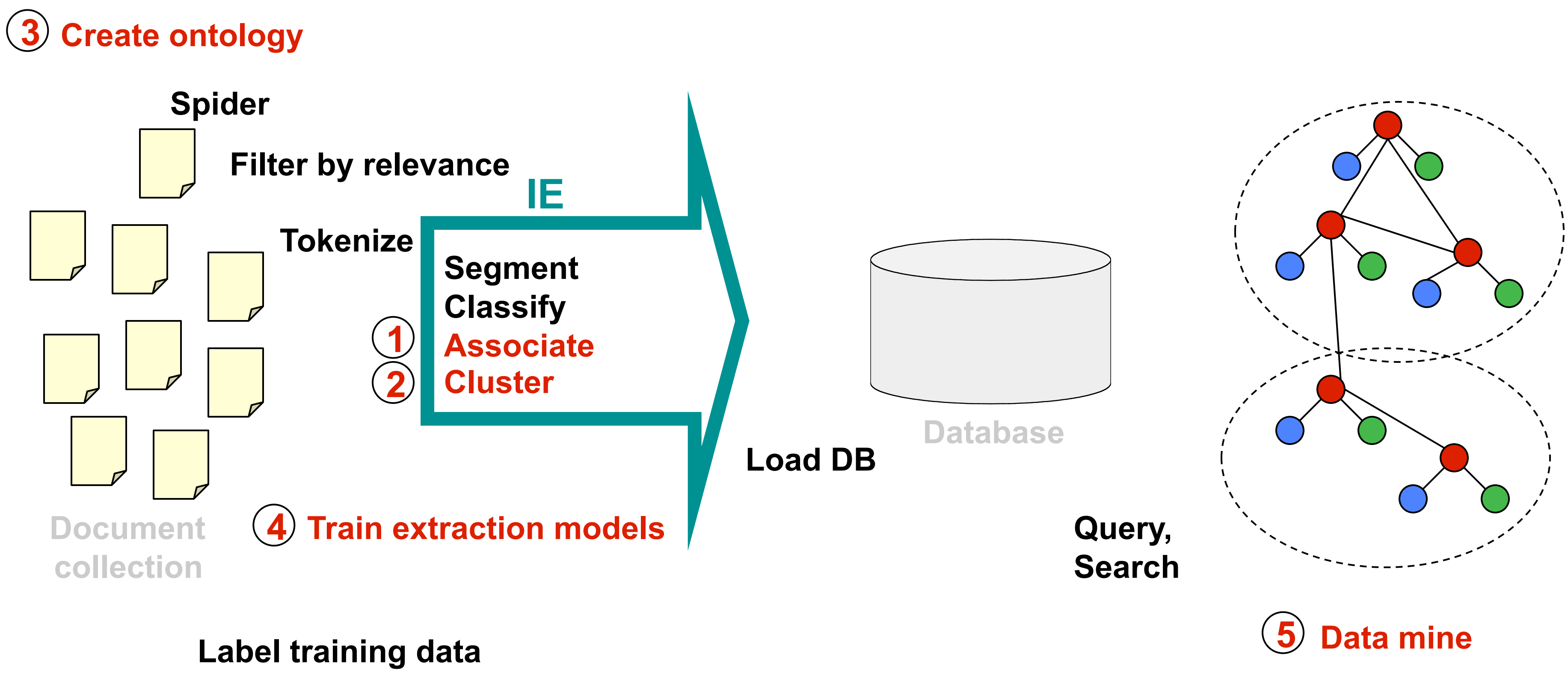
- Sequence models: HMMs, CMMs/MEMMs, CRFs
- Can work well when training data is easy to construct and is plentiful
- Can capture complex patterns that are hard to encode with hand-crafted rules
  - e.g., determine whether a review is positive or negative
  - extract long complex gene names

The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300."

- Can be labor intensive to construct training data
  - Question: how much training data is sufficient?

Broader View

Now touch on some other issues



Relation Extraction: Disease Outbreaks

- **E**xtract structured relations from text

May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly **Ebola** epidemic in **Zaire** , is finding itself hard pressed to cope with the crisis…

Information Extraction System (e.g., NYU's Proteus)

Disease Outbreaks in ***The New York Times***

Date	***Disease Name***	***Location***
Jan. 1995	Malaria	Ethiopia
July 1995	Mad Cow Disease	U.K.
Feb. 1995	Pneumonia	U.S.

Example: Protein Interactions

„We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex."

$\text{CBF-A} \xleftrightarrow[\text{complex}]{\text{interact}} \text{CBF-C}$

$\text{CBF-B} \xrightarrow{\text{associates}} \text{CBF-A-CBF-C complex}$

Relation Extraction

- Typically requires Entity Tagging as preprocessing

- Knowledge Engineering
  - Rules defined over lexical items
    - “<company> located in <location>”
  - Rules defined over parsed text
    - “((Obj <company>) (Verb located) (*) (Subj <location>))”
  - Proteus, GATE, …

- Machine Learning-based
  - Supervised: Learn rules/patterns from examples
    Roth 2005, Cardie 2006, Mooney 2005, Bunescu 2007, …
  - Partially-supervised: bootstrap from “seed” examples
    Agichtein & Gravano 2000, Etzioni et al., 2004, …

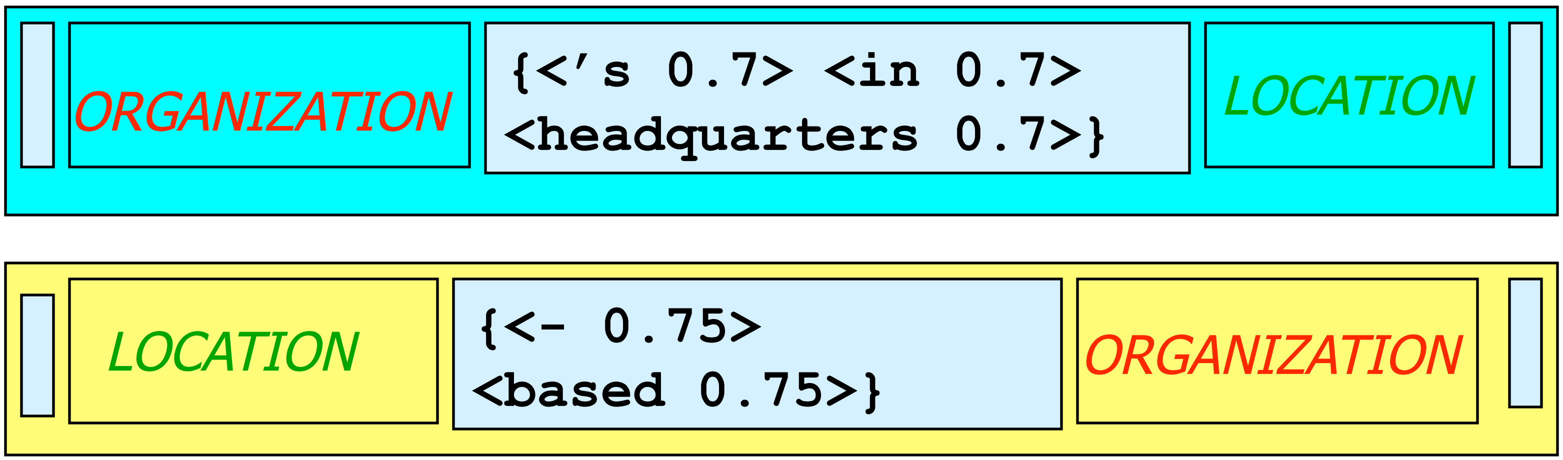
Example Extraction Rule [NYU Proteus]

```
;;; For <company> appoints <person> <position>

(defpattern appoint
   "np-sem(C-company)?  rn?  sa?  vg(C-appoint) np-sem(C-person) ´,´?
    to-be?  np(C-position) to-succeed?:
    company-at=1.attributes, sa=3.span, lv=4.span, person-at=5.attributes
    position-at=8.attributes |
...
```

```
(defun when-appoint (phrase-type)
    (let ((person-at (binding ´person-at))
        (company-entity (entity-bound ´company-at))
        (person-entity (essential-entity-bound ´person-at ´C-person))
        (position-entity (entity-bound ´position-at))
        (predecessor-entity (entity-bound ´predecessor-at))
        new-event)
     (not-an-antecedent position-entity)
     ;; if no company is specified for position, use agent
...
```

Example of Learned Extraction Patterns: Snowball [AG2000]



(1) Association as Binary Classification

Christos Faloutsos conferred with Ted Senator, the KDD 2003 General Chair.

Person Person Role

Person-Role (Christos Faloutsos, KDD 2003 General Chair) → NO

Person-Role (Ted Senator, KDD 2003 General Chair) → YES

Do this with SVMs and tree kernels over parse trees.

[Zelenko et al, 2002]

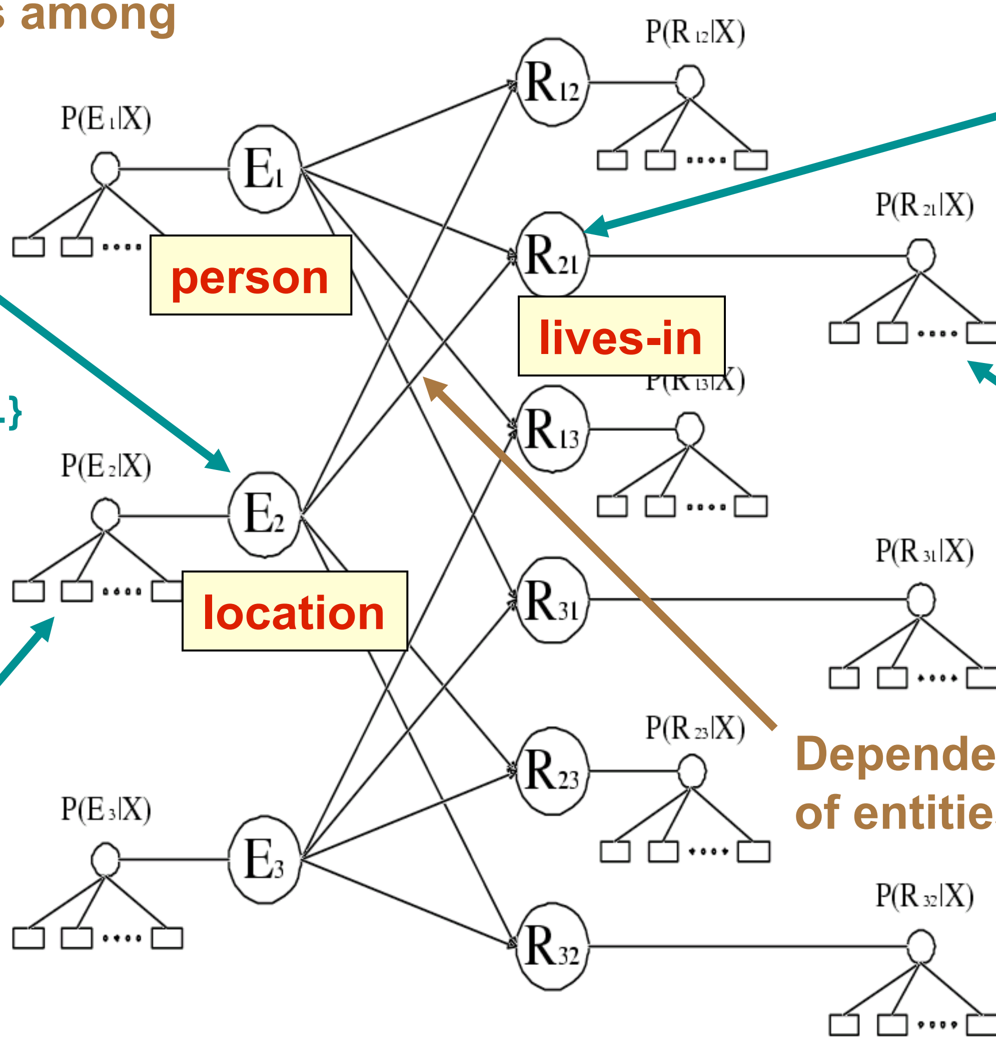
(2) Association with Graphical Models

[Roth & Yih 2002]

Capture arbitrary-distance dependencies among predictions.

Random variable over the class of entity #1, e.g. over {person, location,…}

Local language models contribute evidence to entity classification.



Random variable over the class of relation between entity #2 and #1, e.g. over {lives-in, is-boss-of,…}

Local language models contribute evidence to relation classification.

Dependencies between classes of entities and relations!

Inference with loopy belief propagation.

Accuracy of Information Extraction

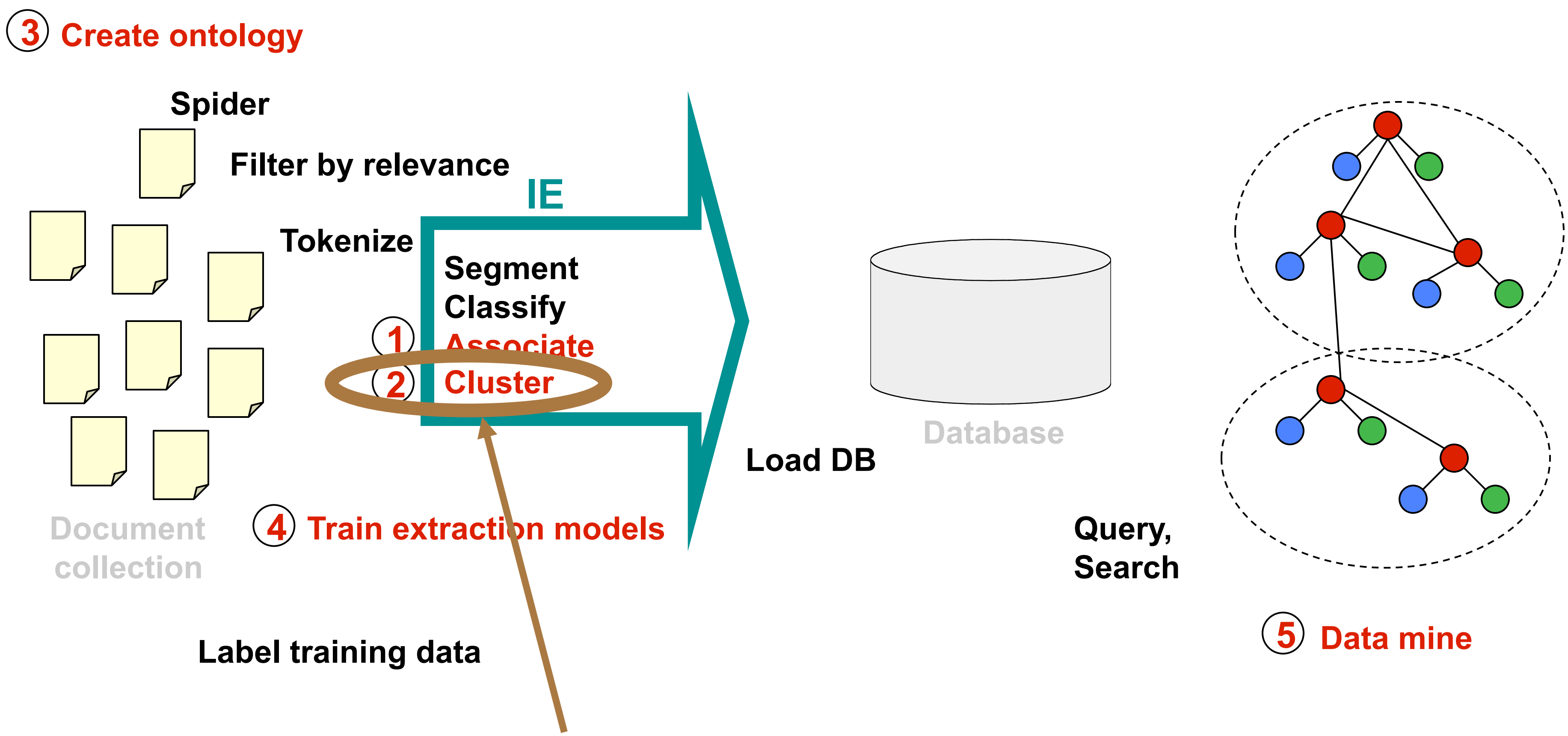
Information Type	Accuracy
Entities	90-98%
Attributes	80%
Facts	60-70%
Events	50-60%

[Feldman, ICML 2006 tutorial]

- Errors cascade (error in entity tag → error in relation extraction)
- This estimate is optimistic:
  - Holds for well-established tasks
  - Many specific/novel IE tasks exhibit much lower accuracy

Broader View

Now touch on some other issues



When do two extracted strings refer to the same object?

Extracted Entities: Resolving Duplicates

Document 1: *The Justice Department has officially ended its inquiry into the assassinations of* **John F. Kennedy** *and Martin Luther King Jr., finding ``no persuasive evidence'' to support conspiracy theories, according to department documents. The House Assassinations Committee concluded in 1978 that* **Kennedy** *was ``probably'' assassinated as the result of a conspiracy involving a second gunman, a finding that broke from the* ***Warren Commission*** *'s belief that Lee Harvey Oswald acted alone in* ***Dallas*** *on Nov. 22, 1963.*

Document 2: *In 1953, Massachusetts* **Sen. John F. Kennedy** *married Jacqueline Lee Bouvier in Newport, R.I. In 1960, Democratic presidential candidate* **John F. Kennedy** *confronted the issue of his Roman Catholic faith by telling a Protestant group in Houston, ``I do not speak for my church on public matters, and the church does not speak for me.''*

Document 3: David Kennedy *was born in Leicester, England in 1959. …***Kennedy** *co-edited The New Poetry (Bloodaxe Books 1993), and is the author of New Relations: The Refashioning Of British Poetry 1980-1994 (Seren 1996).*

[From Li, Morie, & Roth, AI Magazine, 2005]

Important Problem

- Appears in numerous real-world contexts
- Plagues many applications
  - Citeseer, DBLife, AliBaba, Rexa, etc.

(2) Information Integration

[Minton, Knoblock, et al 2001], [Doan, Domingos, Halevy 2001], [Richardson & Domingos 2003]

Goal might be to merge results of two IE systems:

Name:	Introduction to Computer Science
Number:	CS 101
Teacher:	M. A. Kludge
Time:	9-11am
Name:	Data Structures in Java
Room:	5032 Wean Hall

Title:	Intro. to Comp. Sci.
Num:	101
Dept:	Computer Science
Teacher:	Dr. Klüdge
TA:	John Smith
Topic:	Java Programming
Start time:	9:10 AM