Information Extraction

William Wang

School of Computer Science
Carnegie Mellon University
yww@cs.cmu.edu

CIPS Summer School
07/25/2015

History of Summer School

1st MSRA Summer Workshop of **Information Extraction:**

June, 2005

IE Course Logistics

Don’t be afraid of asking questions!

Homepage:
http://www.cs.cmu.edu/~yww/ss2015.html

Prerequisite:

- No previous experience on IE is required.
- Some basic knowledge in Machine Learning.

Acknowledgement



William Cohen | Tom Mitchell | Katie Mazaitis

Some of the slides are also adapted from Andrew McCallum, Sunita Sarawagi, Luke Zettlemoyer, Rion Snow, Pedro Domingos, Ralf Grishman, Raphael Hoffmann, and many other people.

Instructor

William Wang (CMU)

Teaching experience:

CMU Machine Learning (100+ students)

CMU Machine Learning for Large Dataset (60+ students)

Affiliations:

- Yahoo! Labs NYC (2015)
- Microsoft Research Redmond (2012-2013)
- Columbia University (2009-2011)
- University of Southern California (2010)

Research Interests

- machine learning

[Machine Learning 2015] [IJCAI 2015] [ACL 2015a] [CIKM 2014] [StarAI 2014] [CIKM 2013]

- natural language processing

[NAACL 2015a] [EMNLP 2014] [ACL 2014] [EMNLP 2013a] [EMNLP 2013b] [ACL 2012] [SIGDIAL 2012] [IJCNLP 2011] [COLING 2010]

- spoken language processing

[ACL 2015b] [NAACL 2015b] [INTERSPEECH 2015] [SLT 2014] [ASRU 2013] [ICASSP 2013] [CSL 2013] [SLT 2012] [ASRU 2011] [INTERSPEECH 2011] [SIGDIAL 2011] [Book Chapter 2011]

What is Information Extraction (IE)?

And why do we care?

Named Entity Recognition

Relation Extraction

Event Extraction

Temporal IE

Multilingual Information Extraction

Tsung-Dao Lee (**T. D. Lee**, Chinese: 李政道; pinyin: *Lǐ Zhèngdào*) (born November 24, 1926) is a Chinese-born American physicist, well known for his work on parity violation, the Lee Model, particle physics, relativistic heavy ion (RHIC) physics, nontopological solitons and soliton stars. He holds the rank of University Professor Emeritus at Columbia University, where he has taught since 1953 and from which he retired in 2012.[1]

In 1957, Lee, at the age of 30, won the Nobel Prize in Physics with C. N. Yang[2] for their work on the violation of parity law in weak interaction, which Chien-Shiung Wu experimentally verified.

Information Extraction

Definition:

extracting structured knowledge from unstructured or semi-structured data (e.g. free text and tables).

In this short course: we will focus on IE from text data.

A Relation Extraction View

Input: documents.

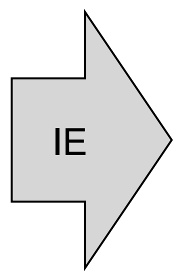
October 14, 2002, 4:00 a.m. PT

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…



Output: relation triples.

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

A Broader View of IE

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

A Broader View of IE

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

A Broader View of IE

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

A Broader View of IE

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

Complexity in IE

Closed set

U.S. states (50 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 patterns

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

Person names

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

Pawel Opalinski, Software Engineer at WhizBang Labs.

Granularity of IE Tasks

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

Single entity

Person: **Jack Welch**

Person: **Jeffrey Immelt**

Location: **Connecticut**

Binary relationship

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

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

N-ary record

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

IE Applications

Question Answering

where does td lee work

Web News Images Videos Maps More Search tools

About 80,600,000 results (0.39 seconds)

He holds the rank of University Professor Emeritus at **Columbia University**, where he has taught since 1953 and from which he retired in 2012. In 1957, Lee, at **the age** of 30, won the Nobel Prize in Physics with C. N.

Tsung-Dao Lee - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Tsung-Dao_Lee** Wikipedia

Question Answering

when did td lee win nobel prize

Web News Images Videos Shopping More ▾ Search tools

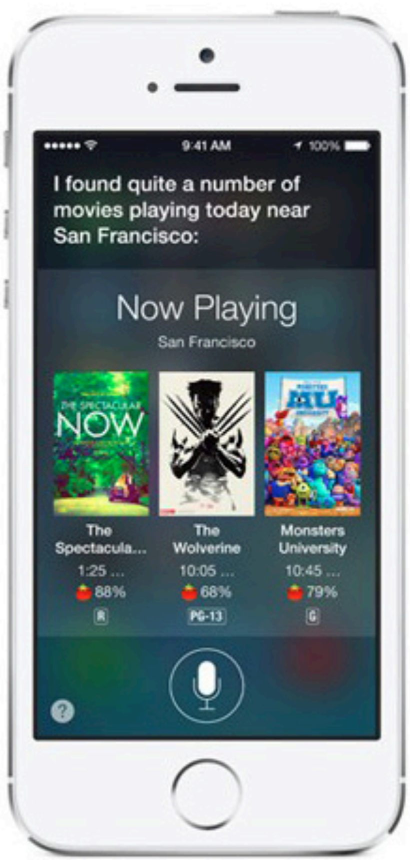
About 127,000 results (0.49 seconds)

In **1957**, Lee, at the age of 30, won the Nobel Prize in Physics with C. N. Yang for their work on the violation of parity law in weak interaction, which Chien-Shiung Wu experimentally verified. Lee was the youngest Nobel laureate after World War II until Malala Yousafzai was awarded the Nobel Peace Prize in 2014.

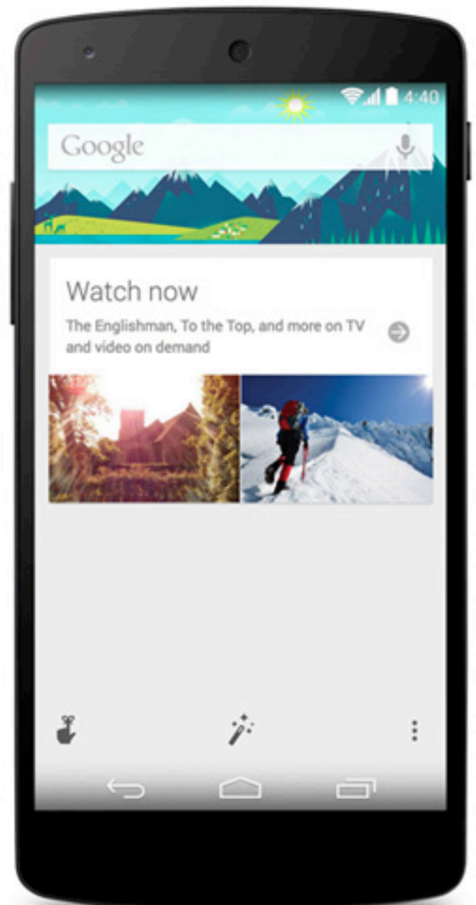
Tsung-Dao Lee - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Tsung-Dao_Lee** Wikipedia ▾

Virtual Assistant

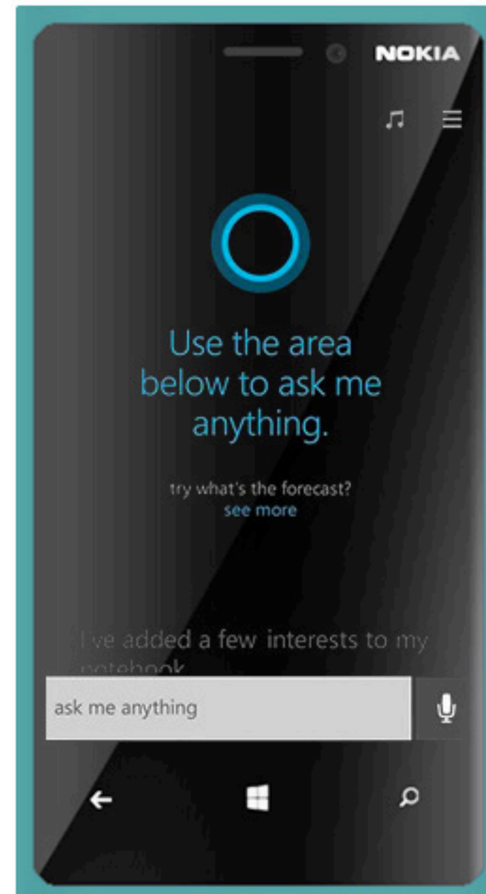
Apple Siri



Google Now



Windows Cortana



Course Outline

1. Basic theories and practices on named entity recognition: supervised, semi-supervised, unsupervised.

2. Recent advances in relation extraction:
   a. distant supervision
   b. latent variable models

3. Scalable IE and reasoning with first-order logics.

Basic Theories and Practices of NER

Named Entity Recognition

Given a sentence:

Yesterday William Wang flew to Beijing.

extract the following information:

Person name: **William Wang**
Location name: **Beijing**

What is the easiest method?

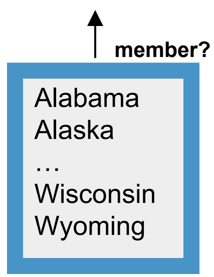
use a lexicon of person names and location names, scan the sentence and look for matches.

Why this will not work? The scalability issue.

Overview of NER Models

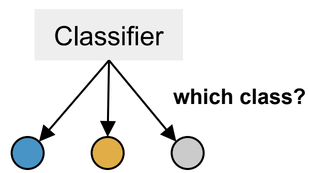
Lexicons

Abraham Lincoln was born in Kentucky.



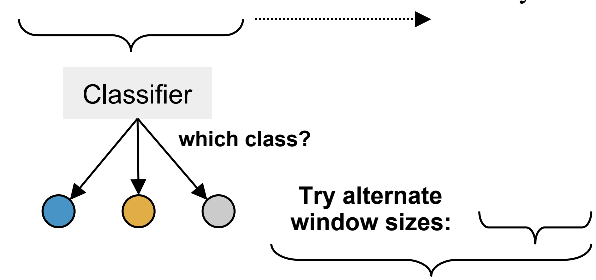
Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.

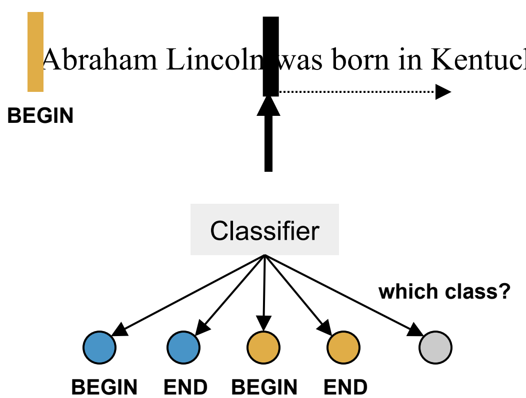


Sliding Window

Abraham Lincoln was born in Kentucky.

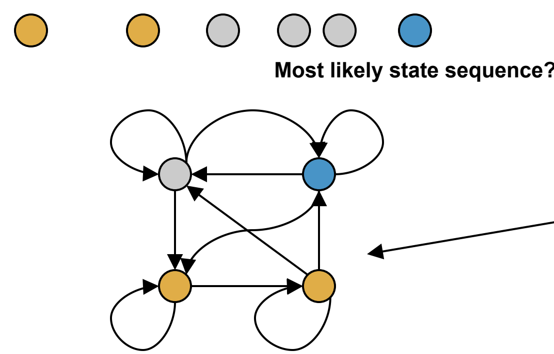


Boundary Models



Token Tagging

Abraham Lincoln was born in Kentucky.



This is often treated as a structured prediction problem…classifying tokens *sequentially*

HMMs, CRFs, ….

Sliding Window

IE by 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

IE by 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

IE by 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

IE by 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

A Naïve Bayes Sliding Window Model

[Freitag 1997]

$$\ldots\ \underbrace{\underset{w_{t-m}}{\text{00}}\ \text{: pm Place}\ \underset{w_{t-1}}{\text{:}}}_{\text{prefix}}\ \underbrace{\boxed{\underset{w_{t}}{\text{Wean}}\ \text{Hall Rm}\ \underset{w_{t+n}}{\text{5409}}}}_{\text{contents}}\ \underbrace{\underset{w_{t+n+1}}{\text{Speaker}}\ \text{: Sebastian}\ \underset{w_{t+n+m}}{\text{Thrun}}}_{\text{suffix}}\ \ldots$$

Estimate Pr(LOCATION|window) using Bayes rule

Try all "reasonable" windows (vary length, position)

Assume independence for length, prefix words, suffix words, content words

Estimate from data quantities like: Pr("Place" in prefix|LOCATION)

If $\mathrm{P}(\text{“Wean Hall Rm 5409”} = \mathrm{LOCATION})$ **is above some threshold, extract it.**

A Naïve Bayes Sliding Window Model

[Freitag 1997]

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

$$\underbrace{w_{t-m} \ \cdots \ w_{t-1}}_{\text{prefix}} \quad \underbrace{w_{t} \ \cdots \ w_{t+n}}_{\text{contents}} \quad \underbrace{w_{t+n+1} \ \cdots \ w_{t+n+m}}_{\text{suffix}}$$

1. Create dataset of examples like these:
   +(prefix00,…,prefixColon, contentWean,contentHall,….,suffixSpeaker,…)
   - (prefixColon,…,prefixWean,contentHall,….,ContentSpeaker,suffixColon,….)
   …
2. Train a NaiveBayes classifier (or YFCL), treating the examples like BOWs for text classification
3. If Pr(class=+|prefix,contents,suffix) > threshold, predict the content window is a location.
   - To think about: what if the extracted entities aren't consistent, eg if the location overlaps with the speaker?

Sliding Window Performance

[Freitag 1997]

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%

Token Tagging

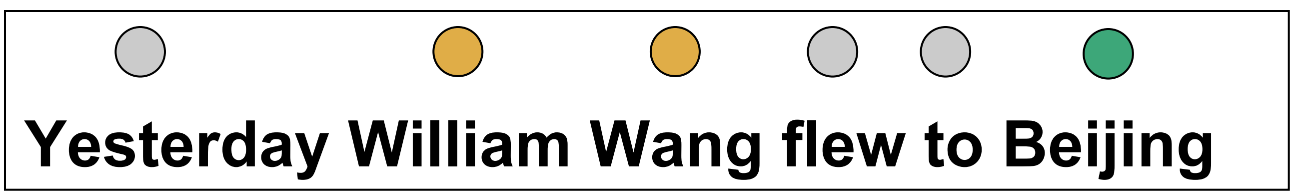
NER by Token Tagging

Given a sentence:

Yesterday William Wang flew to Beijing.

1) Break the sentence into *tokens*, and ***classify*** each token with a label indicating *what sort of entity* it's part of:

- person name
- location name
- background



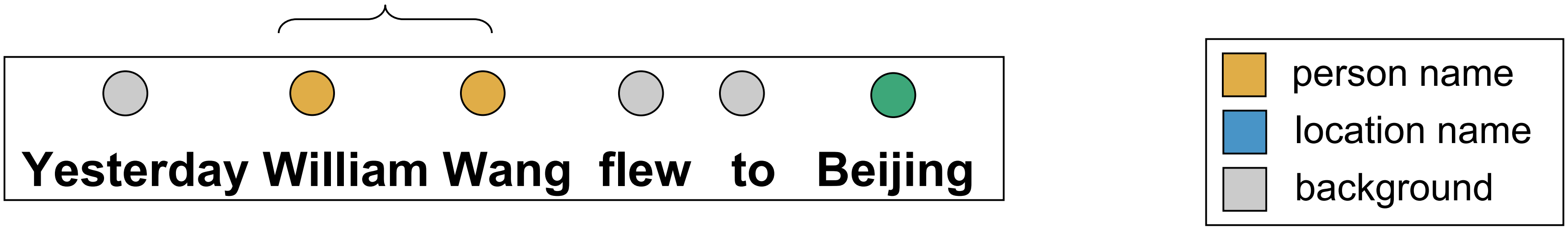
2) Identify names based on the entity labels

Person name: William Wang
Location name: Beijing

3) To learn an NER system, use YFCL.

NER by Token Tagging

Similar labels tend to *cluster together* in text



Another common labeling scheme is BIO (begin, inside, outside; e.g. beginPerson, insidePerson, beginLocation, insideLocation, outside)

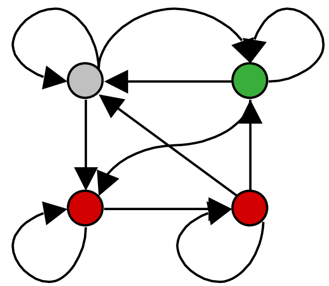
BIO also leads to *strong dependencies between nearby labels* (eg inside follows begin)

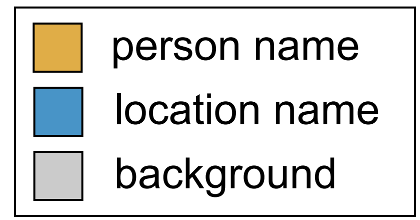
Hidden Markov Models for NER

Given a sequence of observations:

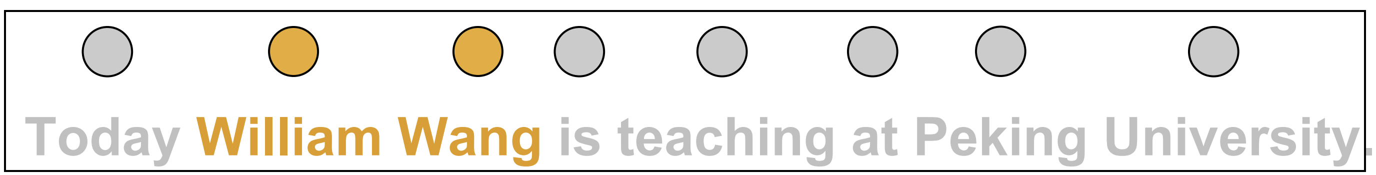
Today William Wang is teaching at Peking University.

and a trained HMM:





Find the most likely state sequence: (Viterbi) $\arg\max_{\vec{s}} P(\vec{s}, \vec{o})$



Any words said to be generated by the designated "person name" state extract as a person name:

Person name: William Wang

Review of Hidden Markov Models

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta}) = p(\mathbf{z}_1|\pi) \left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) \right] \prod_{n-1}^{N} p(\mathbf{x}_n|\mathbf{z}_n, \phi)$$

Observables:

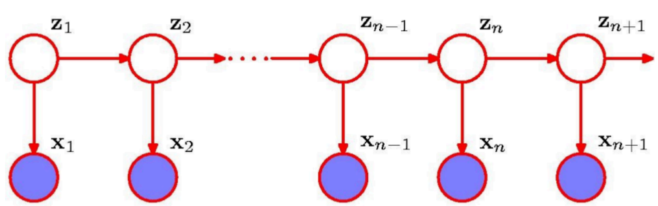
$$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$$

Latent states:

$$\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$$

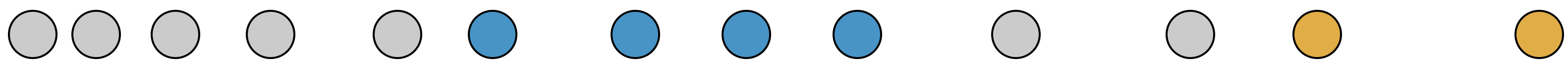
Model parameters:

$$\boldsymbol{\Theta} = \{\pi, \mathbf{A}, \phi\}$$



Hidden Markov Models for NER

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



1. The HMM consists of two probability tables
   - *Pr(currentState=s|previousState=t)* for s=background, location, speaker,
   - *Pr(currentWord=w|currentState=s)* for s=background, location, …
2. Estimate these tables with a (smoothed) CPT
   - Prob(location|location) = #(loc->loc)/#(loc->*) transitions
3. Given a new sentence, find the most likely sequence of hidden states using Viterbi method:

MaxProb(curr=s|position k)=

$\mathrm{Max}_{\text{state } t}$ MaxProb(curr=t|position=k-1) * Prob(word=w_{k-1}|t)*Prob(curr=s|prev=t)

Performance: Sliding Window vs HMMs

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
Speaker:	30%
Location:	61%
Start Time:	98%

Field	F1
Speaker:	77%
Location:	79%
Start Time:	98%

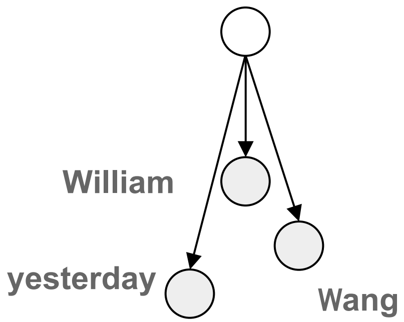
Improving the HMMs

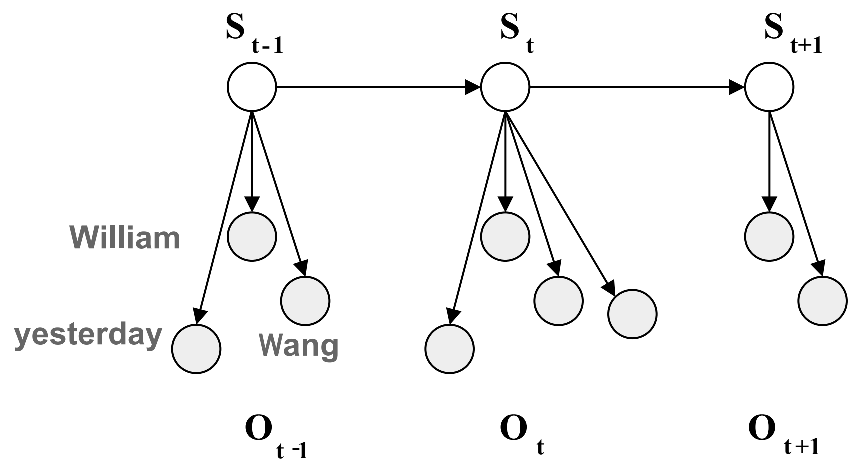
• we need richer representation for the observations e.g., overlapping features.

• we would like to consider modeling the discriminative/ conditional probability model of P(Z|X), rather than the joint/generative probability model of P(Z,X).

Maximum Entropy Markov Model (MEMM)

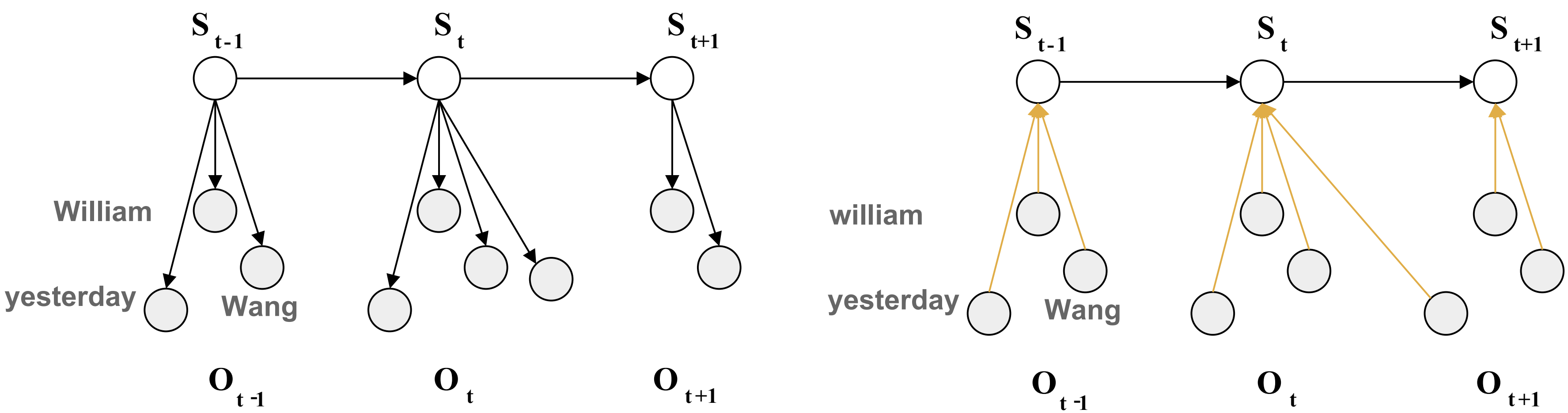
Naïve Bayes vs HMM





HMM = sequential Naïve Bayes

From HMM to MEMM



Replace generative model in HMM with
a MaxEnt/Logistic Regression model

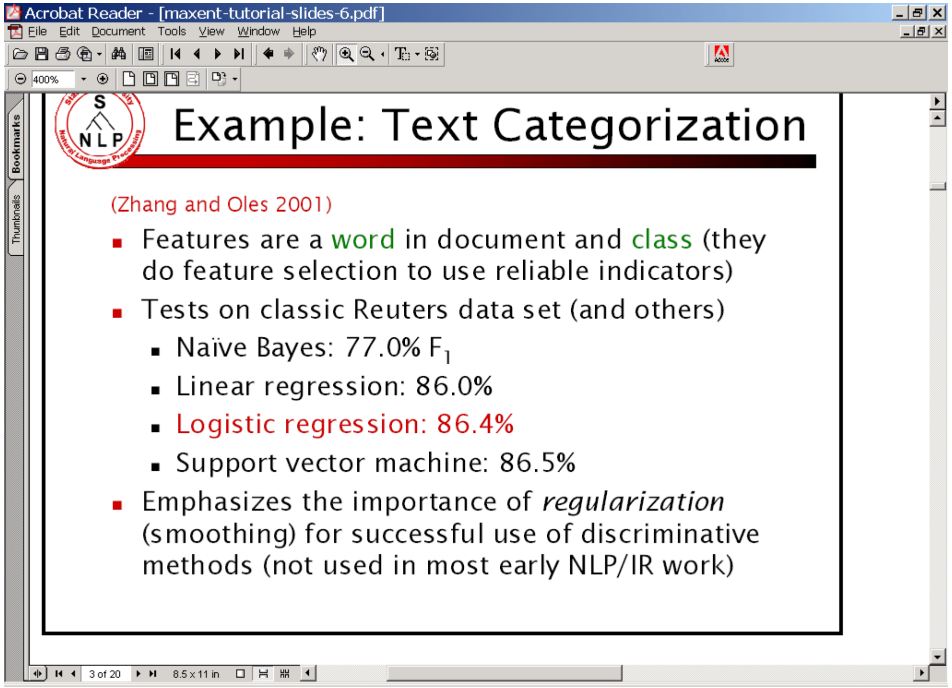
Why MaxEnt Model?

- **Performance:**

Good MaxEnt methods are competitive with linear SVMs and other state of are classifiers in accuracy.

- **Embedding in a larger system:**

MaxEnt optimizes Pr(y|x), not error rate.



From Naïve Bayes to MaxEnt

$$\Pr(y \mid x) = \frac{1}{Z}\Pr(y)\prod_j \Pr(w_k \mid y) \quad = \alpha_0 \prod_i \alpha_\iota^{f_i(x)}$$

where w_k is word j in x

$$\exp\left(\sum_i \lambda_i f_i(x)\right)$$

$$f_{j,k}(doc) = [\text{word k appears at position j of doc?} 1 : 0]$$

$$f_i(doc) = \text{i-th j,k combination}$$

$$\alpha_\iota = \Pr(w_k \mid y)$$

$$\alpha_0 = \Pr(y)/Z$$

MEMMs

- Basic difference from ME tagging:

1. ME tagging: previous state is feature of MaxEnt classifier
2. MEMM: build a **separate** MaxEnt classifier for each state.

Can build any HMM architecture you want; eg parallel nested HMM’s, etc.

- MEMM does allow possibility of “hidden” states and Baum-Welsh like training
- Viterbi is the most natural inference scheme

MEMM task: FAQ parsing

```
    <head>X-NNTP-Poster: NewsHound v1.33
    <head>
    <head>Archive-name: acorn/faq/part2
    <head>Frequency: monthly
    <head>
<question>2.6) What configuration of serial cable should I use
  <answer>
  <answer>   Here follows a diagram of the necessary connections
  <answer>programs to work properly. They are as far as I know t
  <answer>agreed upon by commercial comms software developers fo
  <answer>
  <answer>   Pins 1, 4, and 8 must be connected together inside
  <answer>is to avoid the well known serial port chip bugs. The
```

MEMM features

begins-with-number
begins-with-ordinal
begins-with-punctuation
begins-with-question-word
begins-with-subject
blank
contains-alphanum
contains-bracketed-number
contains-http
contains-non-space
contains-number
contains-pipe
contains-question-mark
contains-question-word
ends-with-question-mark
first-alpha-is-capitalized
indented
indented-1-to-4
indented-5-to-10
more-than-one-third-space
only-punctuation
prev-is-blank
prev-begins-with-ordinal
shorter-than-30

MEMM Performance

Table 4. Co-occurrence agreement probability (COAP), segmentation precision (SegPrec) and segmentation recall (SegRecall) of four learners on the FAQ dataset. All these averages have 95% confidence intervals of 0.01 or less.

Learner	*COAP*	*SegPrec*	*SegRecall*
ME-Stateless	0.520	0.038	0.362
TokenHMM	0.865	0.276	0.140
FeatureHMM	0.941	0.413	0.529
MEMM	0.965	0.867	0.681

Conditional Random Fields

Label Bias Problem of MEMM

- Consider a simple MEMM for person and location names

  all names are two tokens states:

  other

  b-person and e-person for person names

  b-locn and e-locn for location names

Label Bias Problem of MEMM

corpus:

Harvey Ford

(person 9 times, location 1 time)

Harvey Park

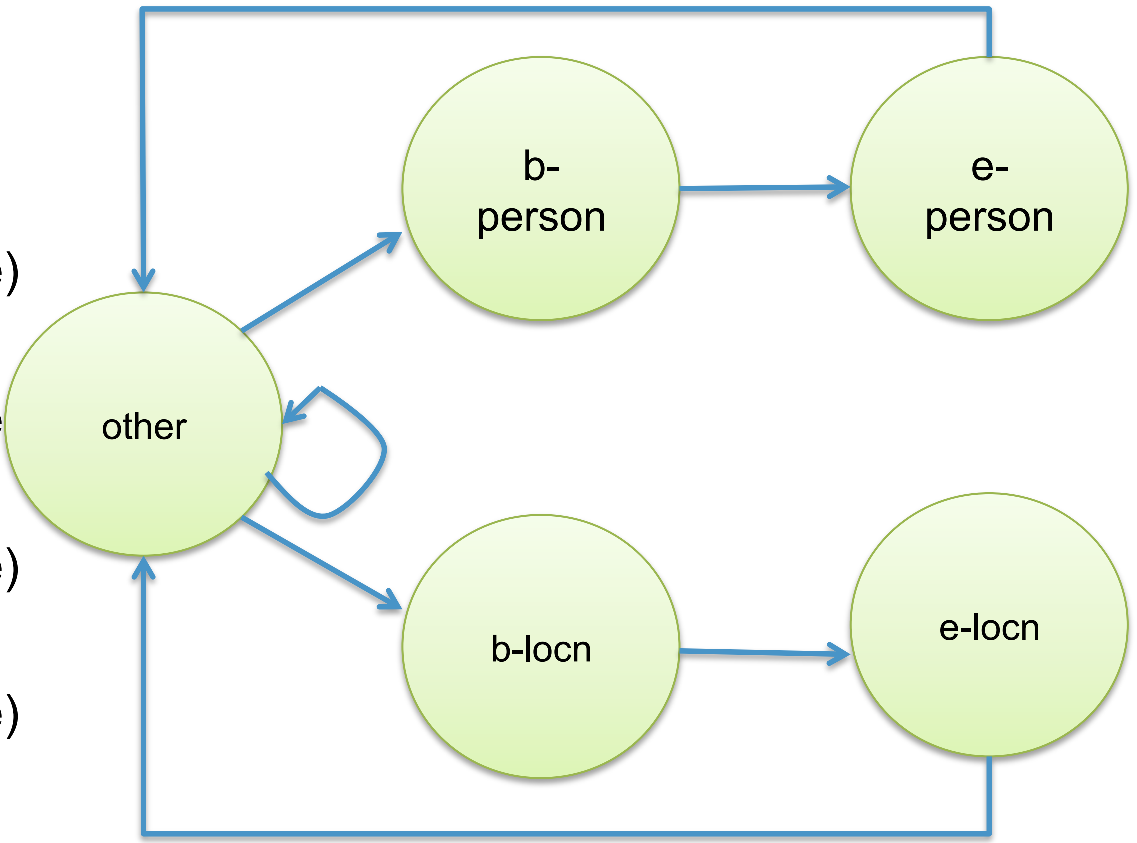
(location 9 times, person 1 time

Myrtle Ford

(person 9 times, location 1 time)

Myrtle Park

(location 9 times, person 1 time)



second token a good indicator of person vs. location

Label Bias Problem of MEMM

Conditional probabilities:

p(b-person | other, w = Harvey) = 0.5

p(b-locn | other, w = Harvey) = 0.5

p(b-person | other, w = Myrtle) = 0.5

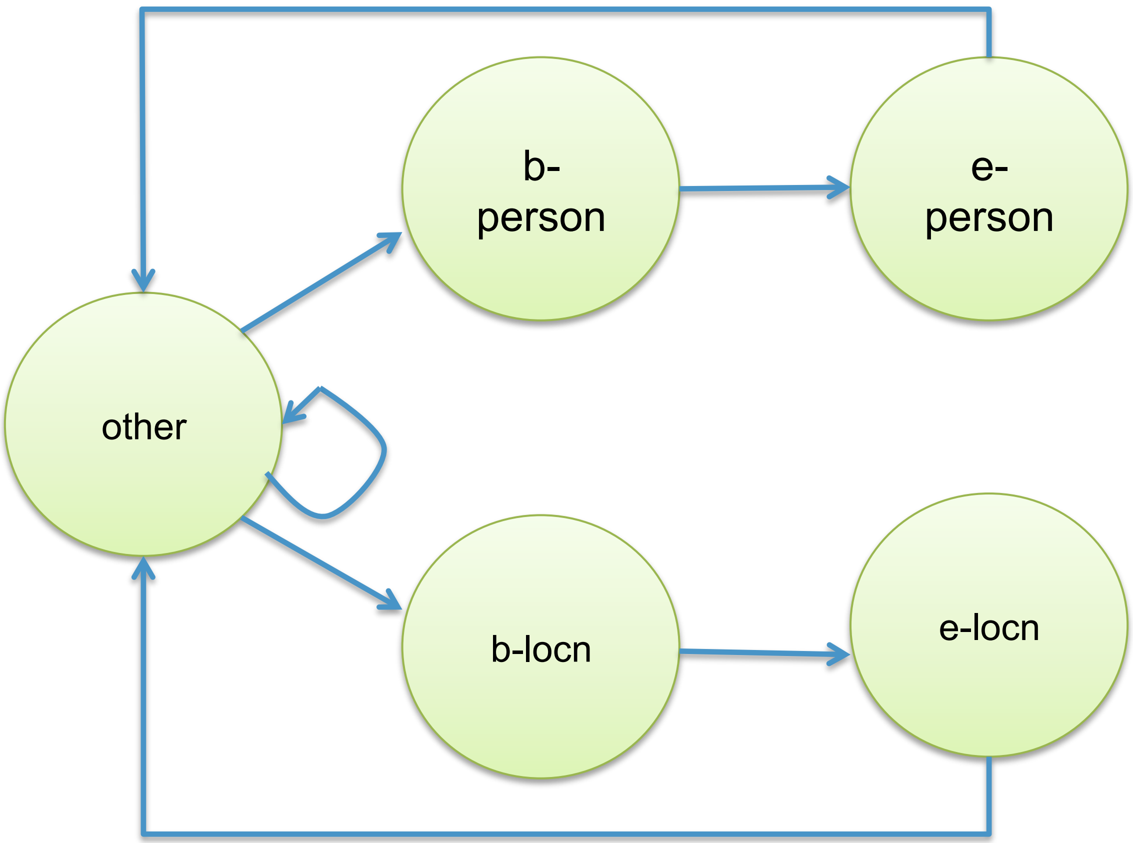
p(b-locn | other, w = Myrtle) = 0.5

p(e-person | b-person, w = Ford) = 1

p(e-person | b-person, w = Park) = 1

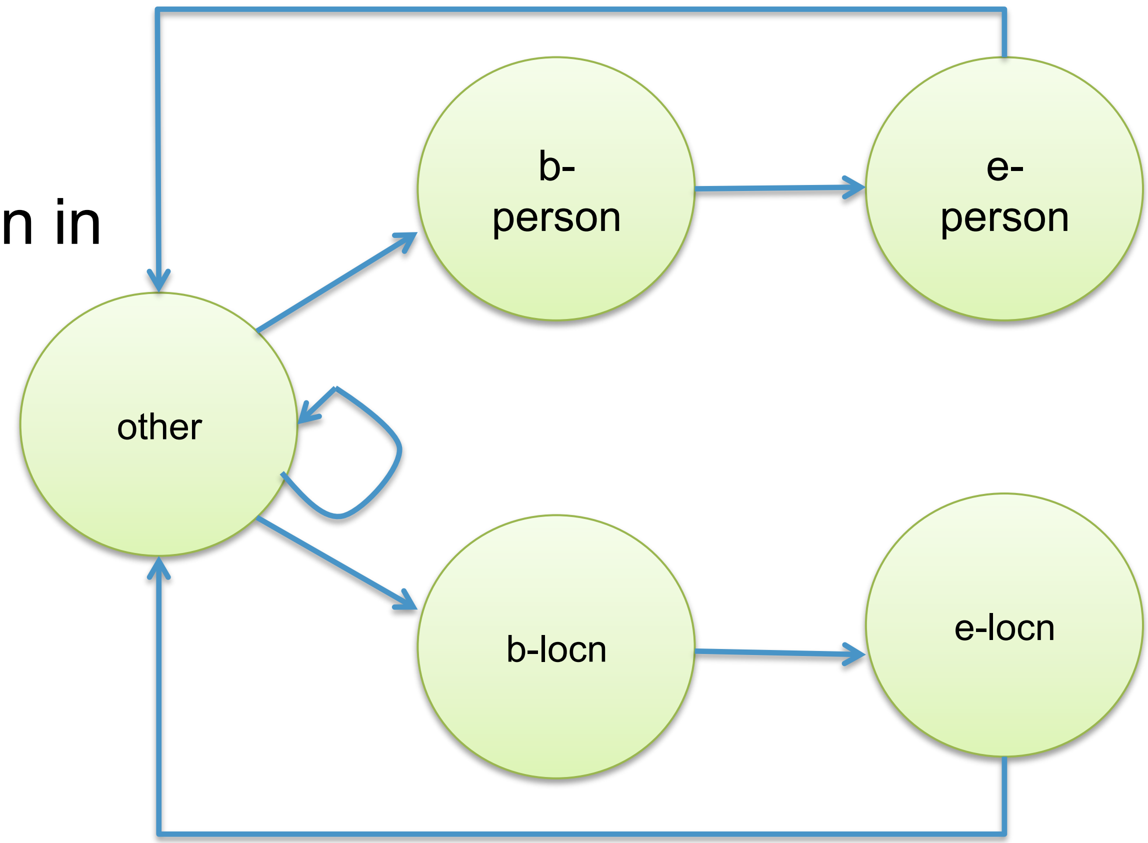
p(e-locn | b-locn, w = Ford) = 1

p(e-locn | b-locn, w = Park) = 1



Label Bias Problem of MEMM

Role of second token in distinguishing person vs. location completely lost

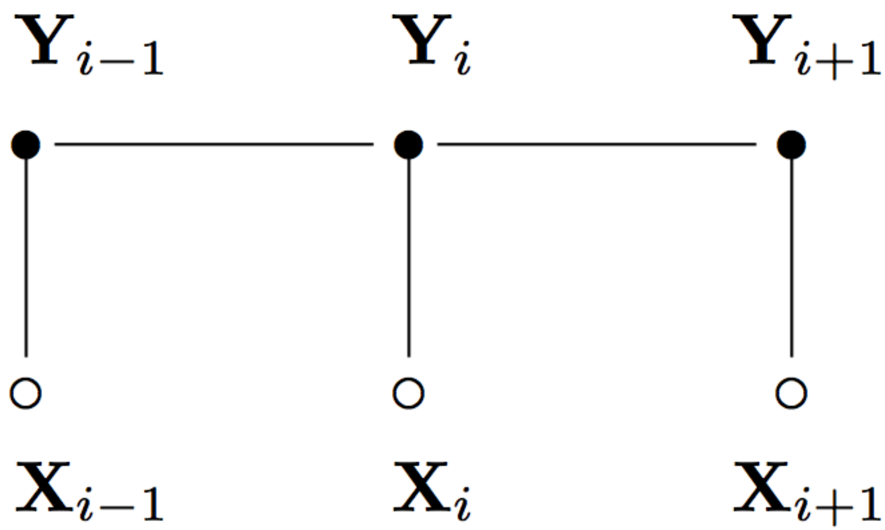


Label Bias Problem of MEMM

- Problem:

Probabilities of outgoing arcs normalized separately for each state.

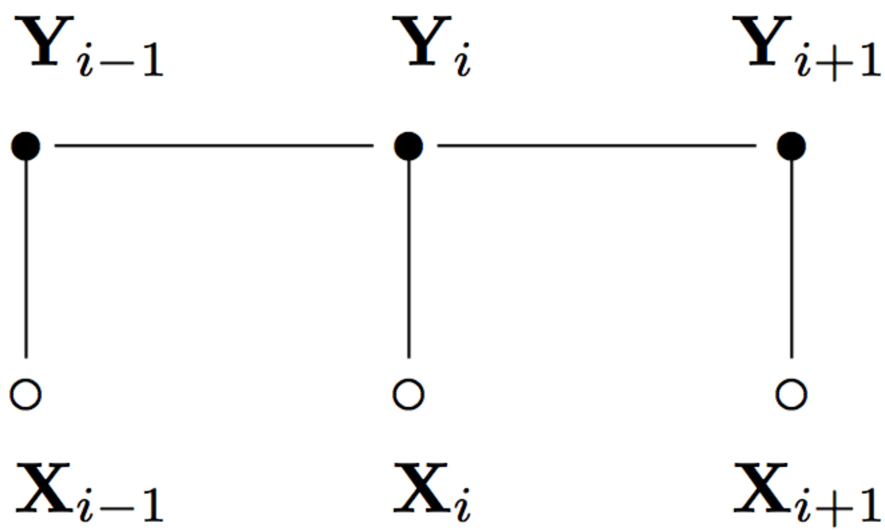
Conditional Random Fields



CRFs' advantages

- over HMM: the independence assumption is relaxed, allowing overlapping features.
- over MEMM: undirected graphical model, a single exponential model for the joint probability of the entire label sequence.

Linear Chain CRFs



$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \sum_t \left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}) + \sum_{h=1}^{h} \mu_h f_h(x_t, y_t) \right)$$

Sha & Pereira results

$q(y_{i-1}, y_i)$	$p(\boldsymbol{x}, i)$
$y_i = y$ $y_i = y,\ y_{i-1} = y'$ $c(y_i) = c$	true
$y_i = y$ or $c(y_i) = c$	$w_i = w$ $w_{i-1} = w$ $w_{i+1} = w$ $w_{i-2} = w$ $w_{i+2} = w$ $w_{i-1} = w',\ w_i = w$ $w_{i+1} = w',\ w_i = w$ $t_i = t$ $t_{i-1} = t$ $t_{i+1} = t$ $t_{i-2} = t$ $t_{i+2} = t$ $t_{i-1} = t',\ t_i = t$ $t_{i-2} = t',\ t_{i-1} = t$ $t_i = t',\ t_{i+1} = t$ $t_{i+1} = t',\ t_{i+2} = t$ $t_{i-2} = t'',\ t_{i-1} = t',\ t_i = t$ $t_{i-1} = t'',\ t_i = t',\ t_{i+1} = t$ $t_i = t'',\ t_{i+1} = t',\ t_{i+2} = t$

Table 1: Shallow parsing features

Model	F score
SVM combination (Kudo and Matsumoto, 2001)	94.39%
CRF	94.38%
Generalized winnow (Zhang et al., 2002)	93.89%
Voted perceptron	94.09%
MEMM	93.70%

Table 2: NP chunking F scores

CRF beats MEMM (McNemar's test); MEMM *probably* beats voted perceptron

Sha & Pereira results

training method	time	F score	$\mathcal{L}'_\lambda$
Precond. CG	130	94.19%	-2968
Mixed CG	540	94.20%	-2990
Plain CG	648	94.04%	-2967
L-BFGS	84	94.19%	-2948
GIS	3700	93.55%	-5668

Table 3: Runtime for various training methods

in minutes, 375k examples

Sequential Models for IE: Practical Advice

Implementing an HMM

- Follow Larry Rabiner's classic HMM tutorial:

A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition

LAWRENCE R. RABINER, FELLOW, IEEE

- Debugging an HMM:

Training (forward-backward): check your transition probability matrix.

Decoding (Viterbi): check the output state sequence.

Understanding CRFs

• actually Lafferty's paper is pretty hard to understand. Instead, try to read Hanna Wallach's CRF introduction.

Conditional Random Fields: An Introduction*

Hanna M. Wallach

February 24, 2004

CRF Tools

- CRF++： probably most widely used. Fast, multithreaded L-BFGS training. Support CoNLL format only.
- CRFsuite： flexible data input format. No parallelization.
- Wapiti (recommended)： Support CoNLL and customized data format. Fast, multithreaded L-BFGS training.
- Stochastic Gradient CRFs： using SGD training instead of L-BFGS.
- Mallet： CRFs in Java.

CRF Demo: Wapiti
https://wapiti.limsi.fr

Training sentence:
Yesterday William Wang flew to Beijing.

Testing sentence:
Yesterday William Cohen flew to Buenos Aires.

Semi-supervised IE

Semi-supervised IE

- Basic idea:
    - Find where a known fact occurs in text, by matching/alignment/…
    - Use this as training data for a conventional IE learning system.
- Once you've learned an extractor from that data
    - Run the extractor on some (maybe additional) text
    - Take the (possibly noisy) new facts and start over

- This is called: "Self-training" or "bootstrapping"

Macro-reading *c.* 1992

Automatic Acquisition of Hyponyms from Large Text Corpora

Marti A. Hearst
Computer Science Division, 571 Evans Hall
University of California, Berkeley
Berkeley, CA 94720
and
Xerox Palo Alto Research Center
marti@cs.berkeley.edu

[Coling 1992]

Idea: write some *specific patterns* that indicate A is a kind of B:

1. **… such NP as NP ("at such schools as CMU, students rarely need extensions")**
2. **NP, NP, or other NP ("William, Carlos or other machine learning professors")**
3. **NP including NP ("struggling teams including the Pirates")**
4. **NP, especially NP (prestigious conferences, especially NIPS)**

Results: 8.6M words of Grolier's encyclopedia → 7067 pattern instances → 152 relations

Many were not in WordNet.

Marti's system was iterative

Another iterative, high-precision system

Extracting Patterns and Relations from the World Wide Web

Sergey Brin

Computer Science Department
Stanford University
sergey@cs.stanford.edu

[some workshop, 1998]

Unlike Hearst, Brin learned the patterns; and learned very *high-precision, easy-to-match* patterns using regular expressions.

Result: 24M web pages + 5 books → 199 occurrences → 3 patterns → 4047 occurrences + 5M pages → 3947 occurrences → 105 patterns → … 15,257 books *with some manual tweaks

Idea: exploit "pattern/relation duality":

1. **Start with some *seed* instances of (*author,title)* pairs ("Isaac Asimov", "The Robots of Dawn")**
2. **Look for *occurrences* of these pairs on the web.**
3. **Generate *patterns* that match the seeds.**

   - URLprefix, prefix, middle, suffix
4. **Extract new (*author, title*) pairs that match the patterns.**
5. **Go to 2.**

Key Ideas: So Far

- High-precision low-coverage extractors and large redundant corpora (macro-reading)
- Self-training/bootstrapping

1) Advantage: train on a small corpus, test on a larger one
   - You can use more-or-less off-the-shelf learning methods
   - You can work with very large corpora
2) But, data gets noisier and noisier as you iterate
3) Need either
   - *really* high-precision extractors, or
   - some way to cope with the noise

A variant of bootstrapping: co-training

Redundantly Sufficient Features:

- **features x can be separated into two types x_1,x_2**
- **either x_1 or x_2 is sufficient for classification – i.e.**

  there exists functions f_1 and f_2 such that

  $f(x) = f_1(x_1)=f_2(x_2)$ has low error

person

.., says Mr. Cooper, a vice president of ..

spelling feature | **context feature**

e.g. Capitalization=X+.X+
Prefix=Mr.

e.g., based on words nearby
in dependency parse

Another kind of self-training

Combining Labeled and Unlabeled Data with Co-Training[‡]

Avrim Blum
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3891
avrim+@cs.cmu.edu

Tom Mitchell
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3891
mitchell+@cs.cmu.edu

[COLT 98]

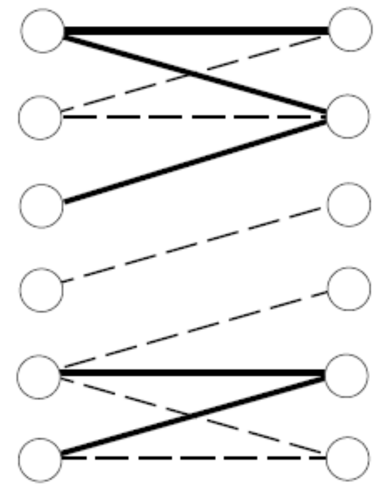


Figure 1: Graphs $G_{\mathcal{D}}$ and G_S. Edges represent examples with non-zero probability under $\mathcal{D}$. Solid edges represent examples observed in some finite sample S. Notice that given our assumptions, even without seeing any labels the learning algorithm can deduce that any two examples belonging to the same connected component in G_S must have the same classification.

Given:

- a set L of labeled training examples
- a set U of unlabeled examples

Create a pool U' of examples by choosing u examples at random from U
Loop for k iterations:

Use L to train a classifier h_1 that considers only the x_1 portion of x
Use L to train a classifier h_2 that considers only the x_2 portion of x
Allow h_1 to label p positive and n negative examples from U'
Allow h_2 to label p positive and n negative examples from U'
Add these self-labeled examples to L
Randomly choose $2p + 2n$ examples from U to replenish U'

A co-training algorithm

Given:

- a set L of labeled training examples
- a set U of unlabeled examples

Create a pool U' of examples by choosing u examples at random from U

Loop for k iterations:

> Use L to train a classifier h_1 that considers only the x_1 portion of x
>
> Use L to train a classifier h_2 that considers only the x_2 portion of x
>
> Allow h_1 to label p positive and n negative examples from U'
>
> Allow h_2 to label p positive and n negative examples from U'
>
> Add these self-labeled examples to L
>
> Randomly choose $2p + 2n$ examples from U to replenish U'

Unsupervised Models for Named Entity Classification

Michael Collins and Yoram Singer [EMNLP 99]

Redundantly Sufficient Features:

- **features** x **can be separated into two types** x_1, x_2
- **either** x_1 **or** x_2 **is sufficient for classification – i.e. there exists functions** f_1 **and** f_2 **such that**

$f(x) = f_1(x_1) = f_2(x_2)$ has low error

Candidate entities x segmented using a POS pattern

person

..., says Mr. Cooper, a vice president of ..

spelling feature

context feature

e.g., Capitalization=X+.X+
Prefix=Mr.

Based on dependency parse

Evaluation for Collins and Singer

Learning Algorithm	Accuracy (Clean)	Accuracy (Noise)
Baseline	45.8%	41.8%
EM	83.1%	75.8%
(Yarowsky 95)	81.3%	74.1%
Yarowsky-cautious	91.2%	83.2%
DL-CoTrain	91.3%	83.3%
CoBoost	91.1%	83.1%

Table 2: Accuracy for different learning methods. The baseline method tags all entities as the most frequent class type (organization).

88,962 examples (spelling,context) pairs

7 seed rules are used

1000 examples are chosen as test data (85 noise)

We label the examples as (location, person, organization, noise)

$$\text{Accuracy : Noise} = \frac{N_c}{962}$$

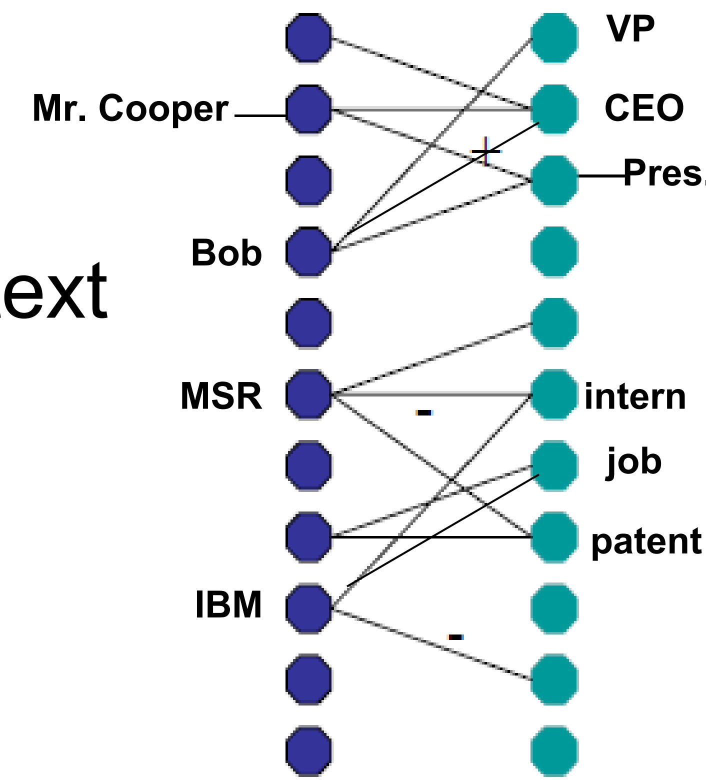
$$\text{Accuracy : Clean} = \frac{N_c}{962 - 85}$$

Key Ideas: So Far

- High-precision low-coverage extractors and large redundant corpora (macro-reading)
- Self-training/bootstrapping
- Co-training
- Clustering phrases by context

Don't propagate labels;

Instead do without them entirely



Induction of Semantic Classes from Natural Language Text

Dekang Lin and Patrick Pantel
University of Alberta
Department of Computing Science
Edmonton, Alberta T6H 2E1 Canada
{lindek, ppantel}@cs.ualberta.ca

[KDD 2002]

Basic idea: parse a big corpus, then cluster NPs by their contexts

CONCEPT	MEMBERS
Nq178	Toyota, Honda, Volkswagen, Mazda, Oldsmobile, BMW Audi, Mercedes-Benz, Cadillac, Volvo, Subaru, Chevrolet, Mercedes, Buick, Porsche, Nissan, VW, Mitsubishi, Renault, Hyundai, Isuzu, Jaguar, Suzuki, Dodge, Rolls-Royce, Pontiac, Fiat, Chevy, Saturn, Yugo, Ferrari, "Mercedes Benz", Plymouth, mustang, Beretta, Panasonic, Corvette, Nintendo, Camaro
Nq352	heroin, cocaine, marijuana, narcotic, alcohol, steroid, crack, opium
Nq356	Saskatchewan, Alberta, Manitoba, "British Columbia", Ontario, "New Brunswick", Newfoundland, Quebec, Guangdong, "Prince Edward Island", "Nova Scotia", "Papua New Guinea", "Northwest Territories", Luzon

Table 1. Excerpts of entries in the collocation database for *duty* and *responsibility* [12].

DUTY		RESPONSIBILITY	
modified-by adjectives	fiduciary 319, active 251, other 82, official 76, additional 47, administrative 44, military 44, constitutional 41, reserve 24, high 23, moral 21, double 16, day-to-day 15, normal 15, specific 15, assigned 14, extra 13, operating 13, temporary 13, corporate 12, peacekeeping 12, possible 12, regular 12, retaliatory 12, heavy 11, routine 11, sacred 11, stiff 11, congressional 10, fundamental 10, hazardous 10, main 10, patriotic 10, punitive 10, special 10, …	modified-by adjectives	more 107, full 92, fiduciary 89, primary 88, personal 79, great 69, financial 64, fiscal 59, social 59, moral 48, additional 46, ultimate 39, day-to-day 37, special 37, individual 36, legal 35, other 35, corporate 30, direct 30, constitutional 29, given 29, overall 29, added 28, sole 25, operating 23, broad 22, political 22, heavy 20, main 18, shared 18, professional 17, current 15, federal 14, joint 14, enormous 13, executive 13, operational 13, similar 13, administrative 10, fundamental 10, specific 10, …
object-of verbs	have 253, assume 190, perform 153, do 131, impose 118, breach 112, carry out 79, violate 54, return to 50, fulfill 44, handle 42, resume 41, take over 35, pay 26, see 26, avoid 19, neglect 18, shirk 18, include 17,	object-of verbs	have 747, claim 741, take 643, assume 390, accept 220, bear 187, share 103, deny 86, fulfill 53, meet 48, feel 47, retain 47, shift 47, carry out 45, take over 41, shoulder 29, escape 28, transfer 28, delegate 26, give 25, admit 23, do 21, acknowledge 20, exercise 20,

Key Ideas: So Far

- High-precision low-coverage extractors and large redundant corpora (macro-reading)
- Self-training/bootstrapping or co-training
- Other semi-supervised methods:

1) Expectation-maximization: like self-training but you "soft-label" the unlabeled examples with the *expectation* over the labels in each iteration.
2) Works for almost any generative model (e.g., HMMs)
3) Learns directly from all the data

   Maybe better; Maybe slower

   Extreme cases:

   supervised learning …. **clustering** + cluster-labeling

Key Ideas: So Far

- High-precision low-coverage extractors and large redundant corpora (macro-reading)
- Self-training/bootstrapping or co-training
- Other semi-supervised methods:

  Expectation-maximization

  Transductive margin-based methods (e.g., transductive SVM)

  Graph-based methods

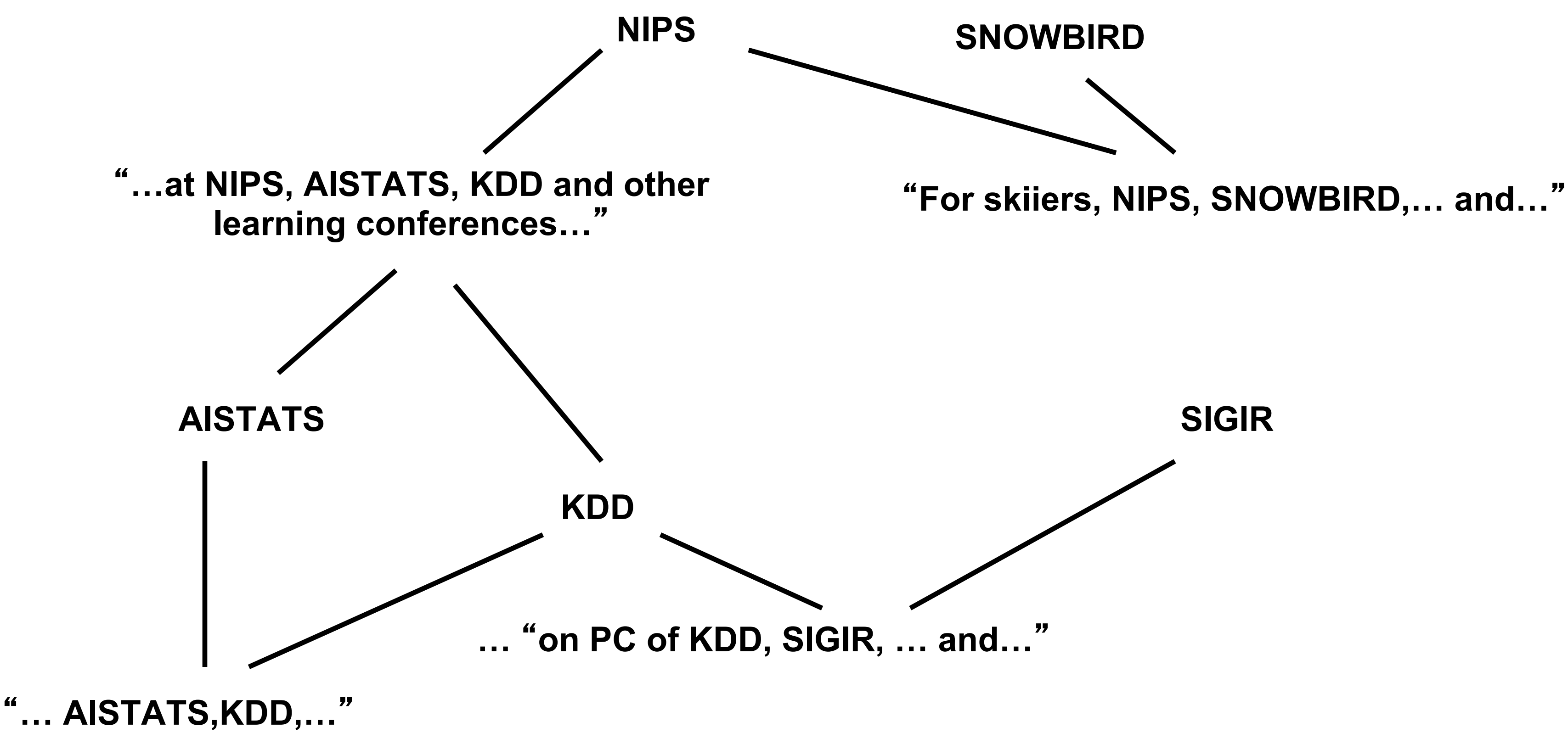
History: Open-domain IE by pattern-matching (Hearst, 92)

- Start with seeds: “NIPS”, “ICML”
- Look thru a corpus for certain patterns:
    - … “at NIPS, AISTATS, KDD and other learning conferences…”
- Expand from seeds to new instances

Repeat….until ___

“on PC of KDD, SIGIR, … and…”

Bootstrapping as graph proximity



shorter paths ~ earlier iterations

many paths ~ additional evidence

Similarity of Nodes in Graphs: Personal PageRank/RandomWalk with Restart

- Similarity defined by PageRank
- Similarity between nodes *x* and *y*:

"Random surfer model": from a node *z,*

with probability α, stop and "output" *z*

pick an edge label (rel) *r* using Pr(*r* | *z) ...* e.g. uniform

pick a *y* given *x, r:* e.g. uniform from { *y'* : *z* → *y* with label *r* }

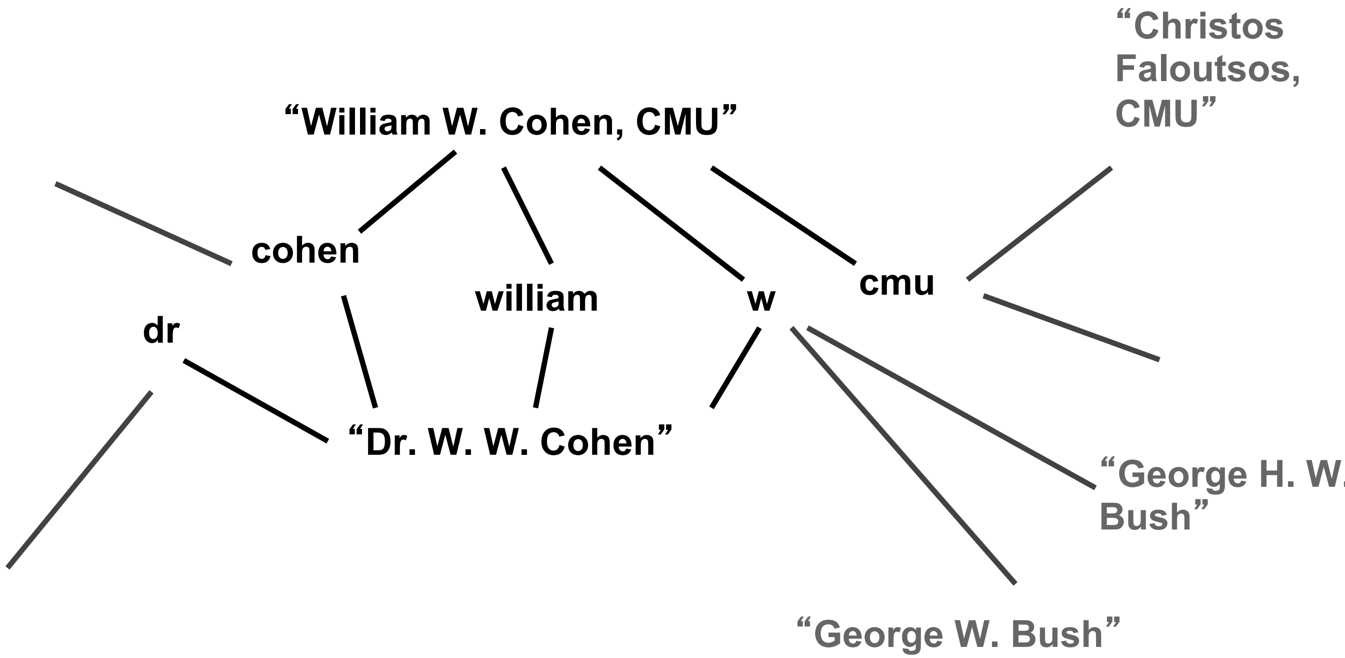
repeat from node *y* ....

Similarity *x~y* = Pr("output" y | start at *x)*

Bootstrapping: propagate from labeled data to "similar" unlabeled data.

Intuitively, *x~y* is summation of weight of all paths from *x* to *y,* where *weight* of path decreases exponentially with length.

PPR/RWR on a Graph



A little math exercise…

Let *x* be less than 1 and larger than 0. Then

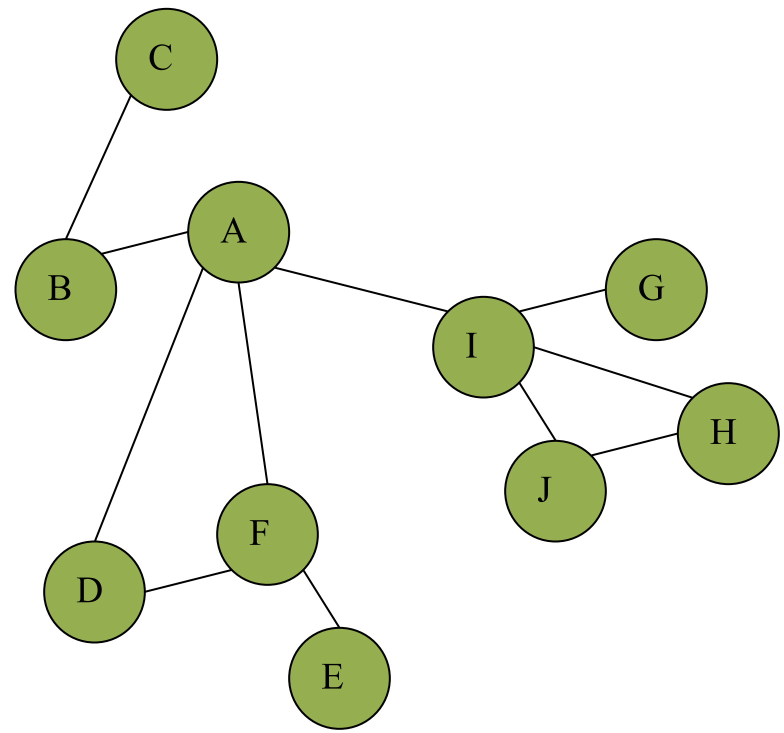
$$y = 1 + x + x^2 + x^3 + \ldots + x^n$$

$$y \approx (1 - x)^{-1}$$

Example: x=0.1, and 1+0.1+0.01+0.001+…. = 1.11111 = 10/9.

Graph = Matrix

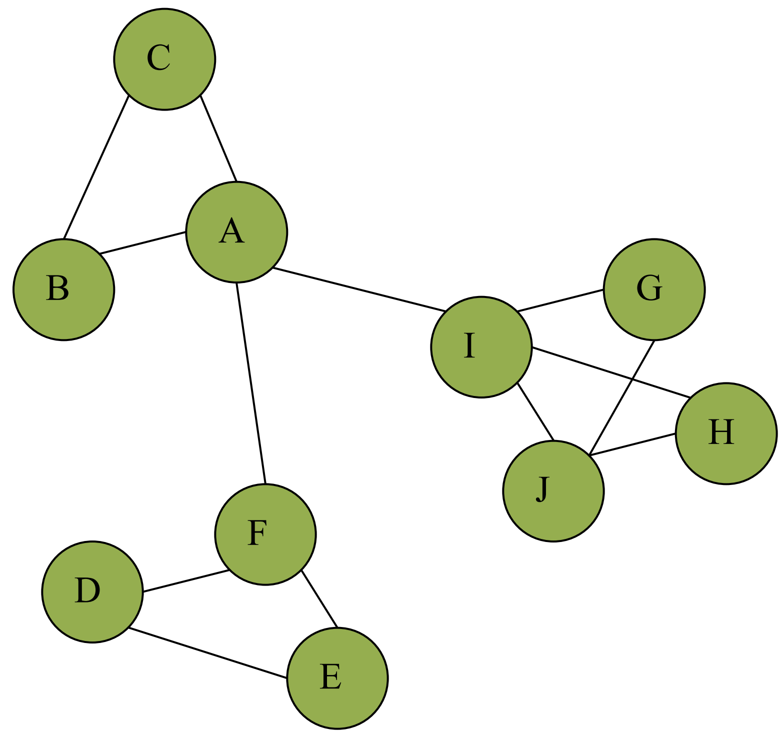
	A	B	C	D	E	F	G	H	I	J
A		1		1		1			1	
B	1		1							
C		1								
D	1					1				
E						1				
F	1			1	1					
G									1	
H							1		1	1
I	1						1	1		1
J								1	1	



Graph = Matrix

Transitively Closed Components = "Blocks"

	A	B	C	D	E	F	G	H	I	J
A	_	1	1			1			1	
B	1	_	1							
C	1	1	_							
D				_	1	1				
E				1	_	1				
F	1			1	1	_				
G							_		1	1
H								_	1	1
I	1						1	1	_	1
J							1	1	1	_



Of course we can't see the "blocks" unless the nodes are sorted by cluster…

Graph = Matrix

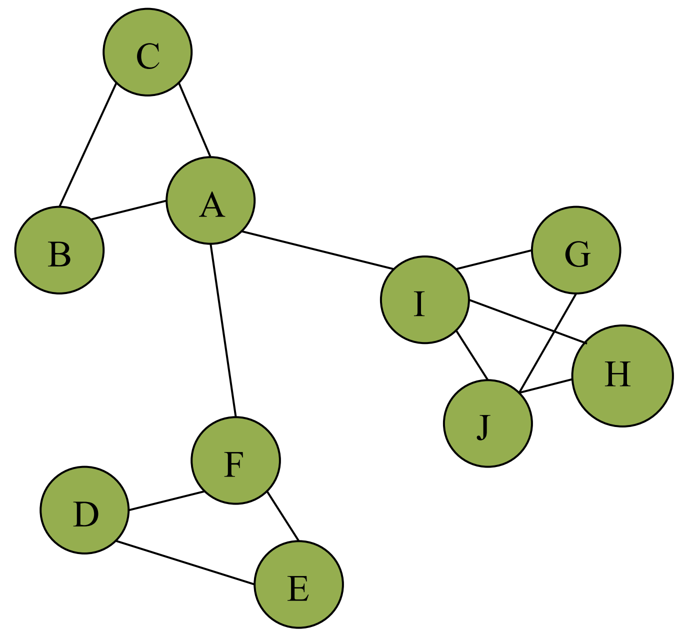
Vector = Node → Weight

M

	A	B	C	D	E	F	G	H	I	J
A	_	1	1			1			1	
B	1	_	1							
C	1	1	_							
D				_	1	1				
E				1	_	1				
F	1			1	1	_				
G							_		1	1
H								_	1	1
I	1						1	1	_	1
J							1	1	1	_

v

	A
A	4
B	2
C	3
D	
E	
F	
G	
H	
I	
J	



M

Graph = Matrix

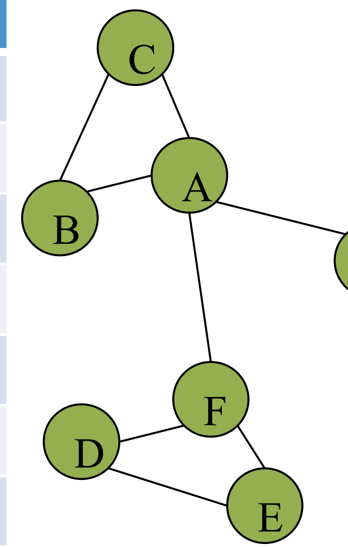
$M*v_1 = v_2$ "propagates weights from neighbors"

$M \quad * \quad v_1 \quad = v_2$

	A	B	C	D	E	F	G	H	I	J
A	_	1	1			1				
B	1	_	1							
C	1	1	_							
D				_	1	1				
E				1	_	1				
F				1	1	_				
G							_		1	1
H								_	1	1
I							1	1	_	1
J							1	1	1	_

A	4
B	2
C	3
D	
E	
F	
G	
H	
I	
J	

A	2*1+3*1+0*1
B	4*1+3*1
C	4*1+2*1
D	
E	
F	
G	
H	
I	
J	



M

A little math…

Let W[*i,j*] **be Pr(walk to** *j* **from** *i***)and let** *α* **be less than 1. Then:**

$$\mathbf{Y} = \mathbf{I} + \alpha\mathbf{W} + (\alpha\mathbf{W})^2 + (\alpha\mathbf{W})^3 + \ldots(\alpha\mathbf{W})^n$$

$$\mathbf{Y}(\mathbf{I} - \alpha\mathbf{W}) = (\mathbf{I} + \alpha\mathbf{W} + (\alpha\mathbf{W})^2 + (\alpha\mathbf{W})^3 + \ldots)(\mathbf{I} - \mathbf{W})$$

$$\mathbf{Y}(\mathbf{I} - \alpha\mathbf{W}) = (\mathbf{I} - \alpha\mathbf{W}) + (\alpha\mathbf{W} - (\alpha\mathbf{W})^2 + \ldots)(\mathbf{I} - \mathbf{W})$$

$$\mathbf{Y}(\mathbf{I} - \alpha\mathbf{W}) = \mathbf{I} - (\alpha\mathbf{W})^{n+1}$$

$$\mathbf{Y} \approx (\mathbf{I} - \alpha\mathbf{W})^{-1} \qquad \mathbf{Y}[i, j] = \frac{1}{Z}\Pr(j \mid i)$$

The matrix **(I-** *α***W)** is the *Laplacian* of *α***W.**

Generally the Laplacian is **(D - A)** where **D**[*i,i*] is the degree of *i* in the adjacency matrix **A.**

A little math…

Let W[*i,j*] be Pr(walk to *j* from *i*)and let *α* be less than 1. Then:

component i

$$\mathbf{v}^0 = \langle 0,0,0,\ldots,0,1,0,\ldots,0 \rangle$$

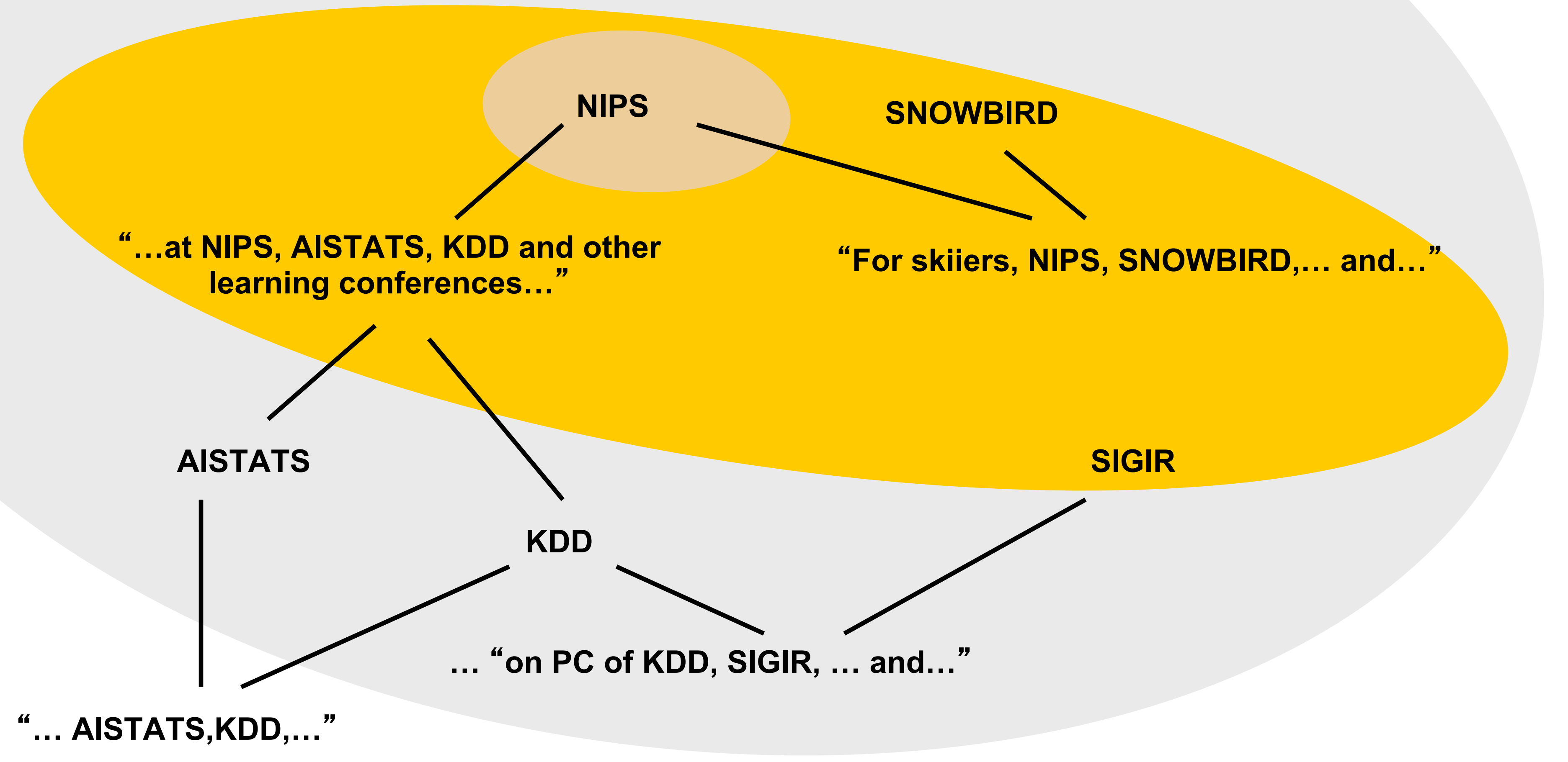
$$\mathbf{v}^{t+1} = (1-\alpha)\mathbf{v}^0 + \alpha \mathbf{W}\mathbf{v}^{t-1}$$

$$\mathbf{v}^n \rightarrow \mathbf{Y}\mathbf{v}^0 \text{ so } \mathbf{v}^n[j] \approx \Pr(j \mid i)$$

The matrix **(I-** *α***W)** is the *Laplacian* of *α***W.**

Generally the Laplacian is **(D- A)** where **D**[*i,i*] is the degree of *i* in the adjacency matrix **A.**

Bootstrapping via PPR/RWR on graph of patterns and nodes



Examples: Cohen & Minkov EMNLP 2008; Komachi et al EMLNLP 2008; Talukdar et al, EMNLP 2008, ACL 2010

Key Ideas: So Far

- High-precision low-coverage extractors and large redundant corpora (macro-reading)
- Self-training/bootstrapping or co-training
- Other semi-supervised methods:

  Expectation-maximization

  Transductive margin-based methods (e.g., transductive SVM)

  Graph-based methods

  Label propogation via random walk with reset

Bootstrapping

Clustering by distributional similarity…

Lin & Pantel '02

Hearst '92

Deeper linguistic features, free text…

BlumMitchell '98

Learning, semi-supervised learning, dual feature spaces…

Brin'98

Scalability, surface patterns, use of web crawlers…

Bootstrapping

Clustering by distributional similarity…

Lin & Pantel '02

Hearst '92

Deeper linguistic features, free text…

Collins & Singer '99

Boosting-based co-train method using content & context features; context based on Collins' parser; learn to *classify* three types of NE

BM'98

Learning, semi-supervised learning, dual feature spaces…

Brin'98

Scalability, surface patterns, use of web crawlers…

Bootstrapping

Clustering by distributional similarity…

Lin & Pantel '02

Hearst '92

Deeper linguistic features, free text…

Riloff & Jones '99

Hearst-like patterns, Brin-like bootstrapping (+ "meta-level" bootstrapping) on MUC data

Collins & Singer '99

BM' 98

Learning, semi-supervised learning, dual feature spaces…

Brin' 98

Scalability, surface patterns, use of web crawlers…

Bootstrapping

Clustering by distributional similarity…

Lin & Pantel '02

Hearst '92

Deeper linguistic features, free text…

Riloff & Jones '99

Collins & Singer '99

BM' 98

Learning, semi-supervised learning, dual feature spaces…

Cucerzan & Yarowsky '99

EM like co-train method with context & content both defined by character-level **tries**

Brin' 98

Scalability, surface patterns, use of web crawlers…

Bootstrapping

Clustering by distributional similarity…

Lin & Pantel '02

Hearst '92

Deeper linguistic features, free text…

Riloff & Jones '99

Collins & Singer '99

…

BM' 98

Learning, semi-supervised learning, dual feature spaces…

Etzioni et al 2005

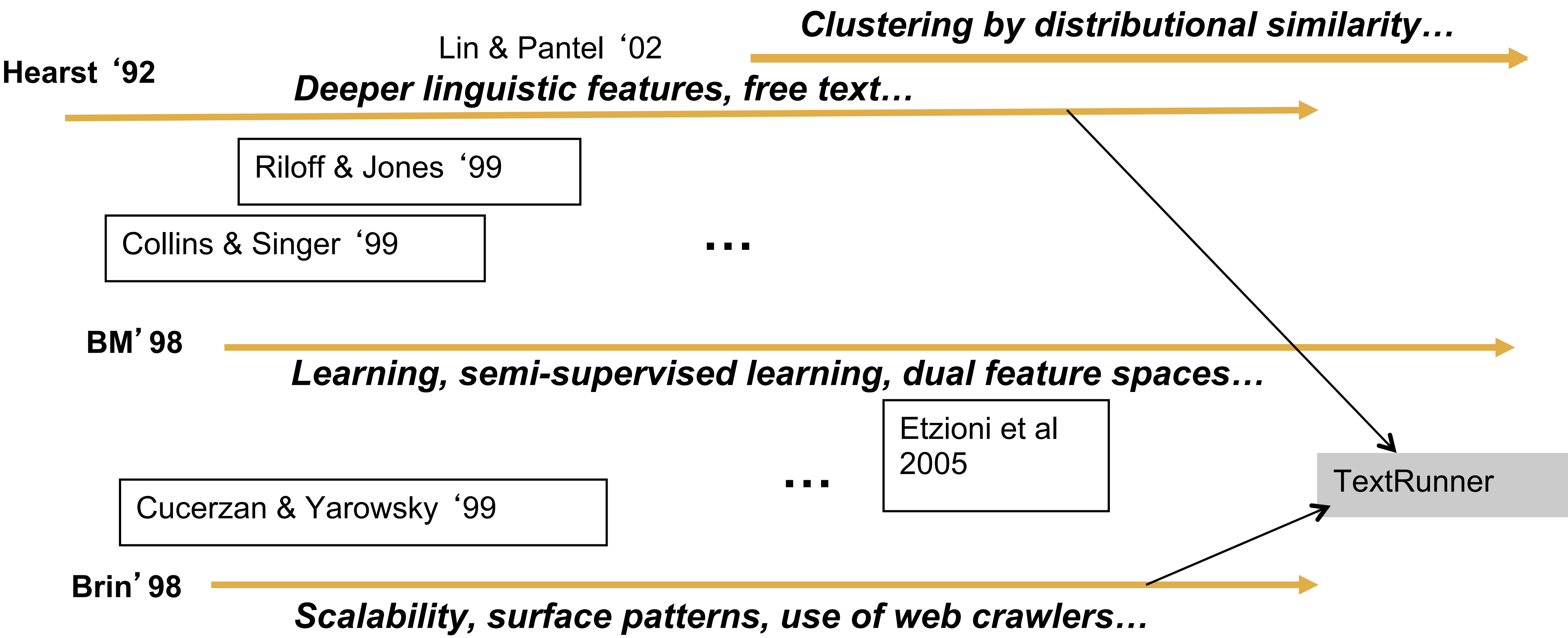
…

Cucerzan & Yarowsky '99 (morphology)

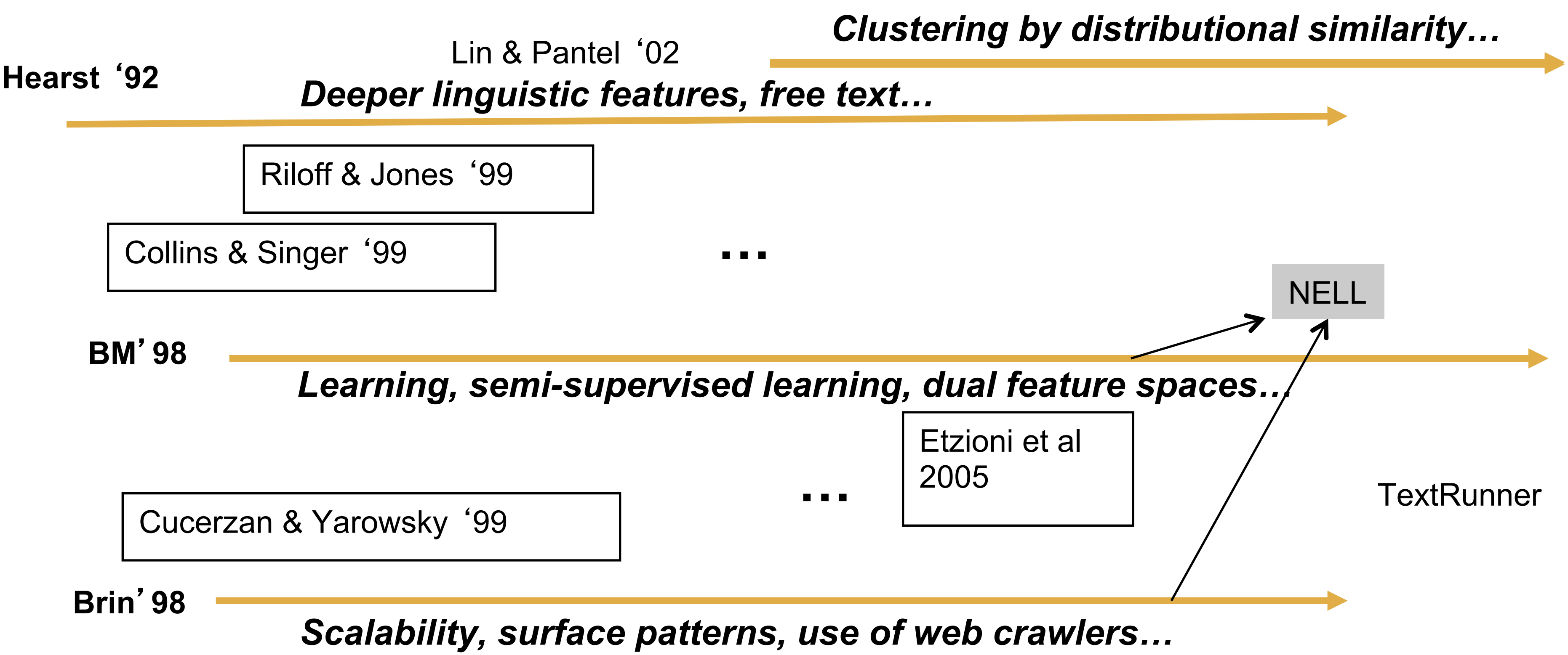
Brin' 98

Scalability, surface patterns, use of web crawlers…

Bootstrapping



Bootstrapping



OpenIE Demo

http://knowitall.github.io/openie/

Never Ending Language Learning

PI: Tom M. Mitchell

Machine Learning Department
Carnegie Mellon University

NELL Theses

1. we'll never understand learning until we build never-ending machine learners

2. background knowledge is key to deep semantic analysis

   NELL KB, plus

   large scale corpus statistics

NELL today

Running 24x7, since January, 12, 2010

Today:

- knowledge base with ~100 million confidence-weighted beliefs
- learning to read
- gradually improving reading accuracy
- learning to reason

gradually improving KB size,

> 100,000 learned rules, scalable probabilistic inference

- extending ontology

new relations: clustering typed pairs

new categories: (developing) clustering + reading subsets

- beginning to include image analysis (via NEIL)

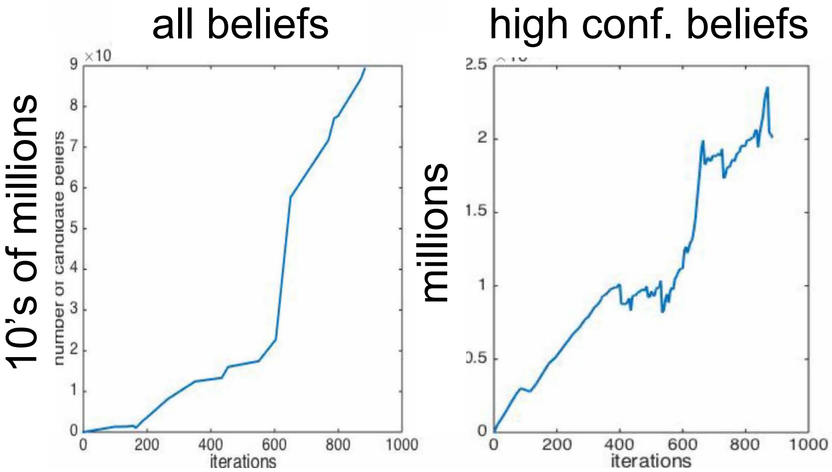
NELL Web Interface

Recently-Learned Facts

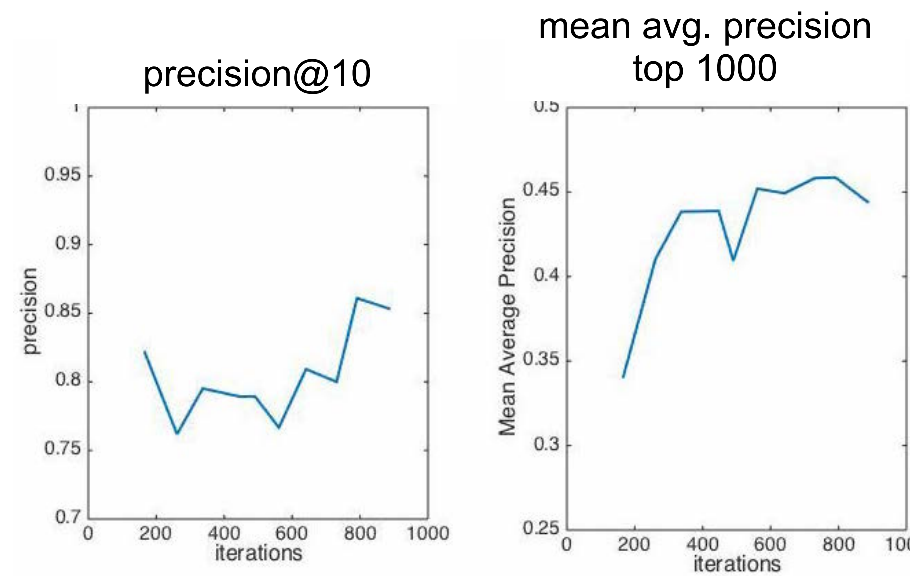
Refresh

instance	iteration	date learned	confidence
african americans at siege of petersburg 1 is a military conflict	938	10-jul-2015	90.6
david koch is a professor	934	25-jun-2015	100.0
california sacramento farm is a farm	934	25-jun-2015	99.0
estate referal services is a profession	934	25-jun-2015	94.4
japanese chicken wings is a type of meat	937	07-jul-2015	99.4
banc of america securities is a company in the economic sector of investment	934	25-jun-2015	99.6
fcc is headquartered in the city washington d c	939	16-jul-2015	96.9
patrick vieira plays for the team france	939	16-jul-2015	93.8
tom anderson is a top member of myspace	939	16-jul-2015	93.8
office is a synonym for united states department	934	25-jun-2015	100.0

NELL Is Improving Over Time (Jan 2010 to Nov 2014)

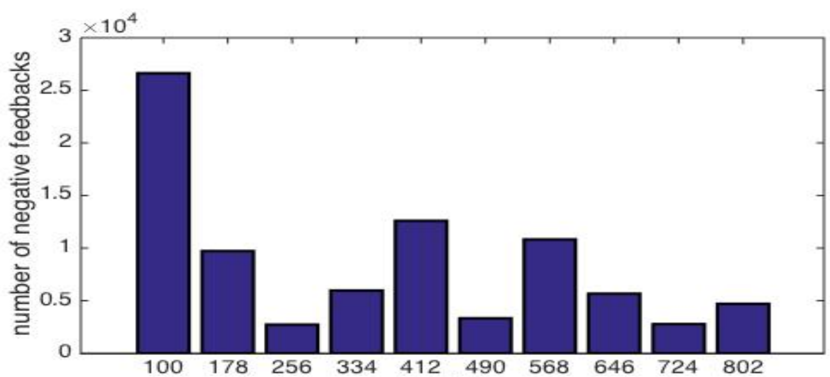


number of NELL beliefs vs. time



reading accuracy vs. time
(average over 31 predicates)

[Mitchell et al., 2015]



human feedback vs. time
(average 2.4 feedbacks per predicate per month)

Portuguese NELL

[Estevam Hruschka]

Recently-Lea

instance
adriane_galisteu is
basf_e_faber_caste
manaus_cavaliers i
jacutinga_campina
fim_da_guerra is a
bamerindus is a ba
nissan is a compar
susana_vieira is a p
campeonato_brasil
toyota_mitsubishi_

- mes
- ano
- dataliteral
- evento
  - eventoesportista
    - olimpiadas
    - grandepremio
    - corrida
    - jogoesportivo
  - convencao
  - fenomenometeo
  - tipodeeventomil
    - conflitomilitar
  - conferencia
    - conferenciade
  - eleicao
  - festivaldemusica
  - festivaldefilmes
  - resultadodeeven
  - crimeouacusaca
- contapolitica
- coordernadas
- metricadeam
- emocao

conflitomilitar

(category)

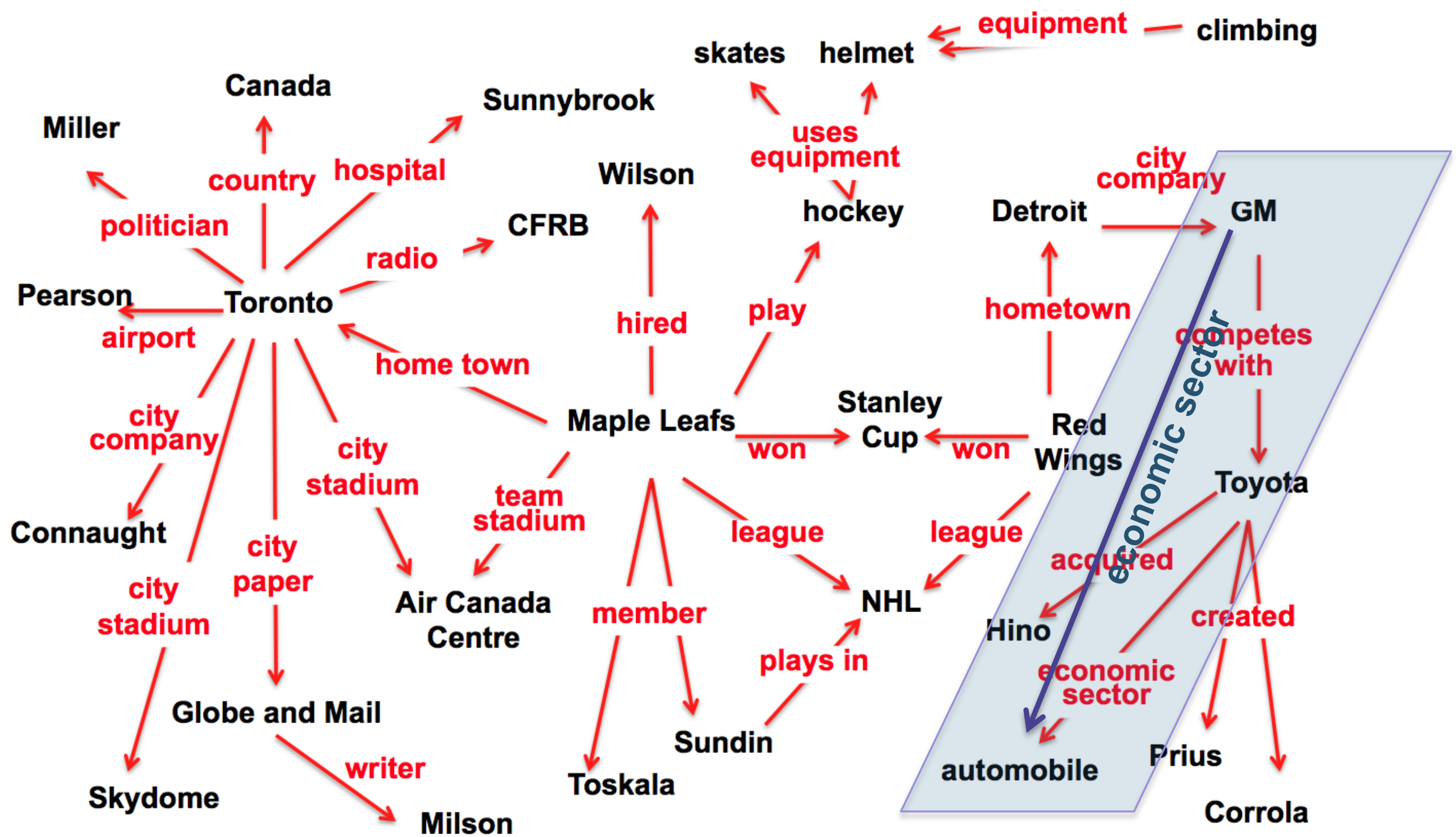
See learned instances of conflitomilitar as a list or on a

Metadata

- **allLearnedPatterns**
  - "a armada durante _" "a causa diplom�tica _" "a armamentista durante _" "a declara��o de capi data _" "a disputa teconol�gica _" "a fronteira i imin�ncia _" "a guarni��o francesa durante _" _" "a P.Y.S.B.E. na _" "a P.Y.S.B.E. _" "a ponte res promover� _" "acabaram a produ��o no _" "a "agudiza��o no _" "antecederam os conflitos d "arquiinimigos na _" "As d�cadas do cabar� Ap _" "As origens do conflito A _" "as raz�es te�ri iraquianas durante _" "avi�es de luta e _" "bacil "batalha da propaganda durante _" "bimotor na _" "cidades do leste durante _" "combates de avi _" "conflito militar apelidado de _" "conflito milit militar chamado de _" "conflito militar como _" " _" "conflito militar tal como _" "conflitos militare

Infer New Beliefs

[Lao, Mitchell, Cohen, *EMNLP* 2011]

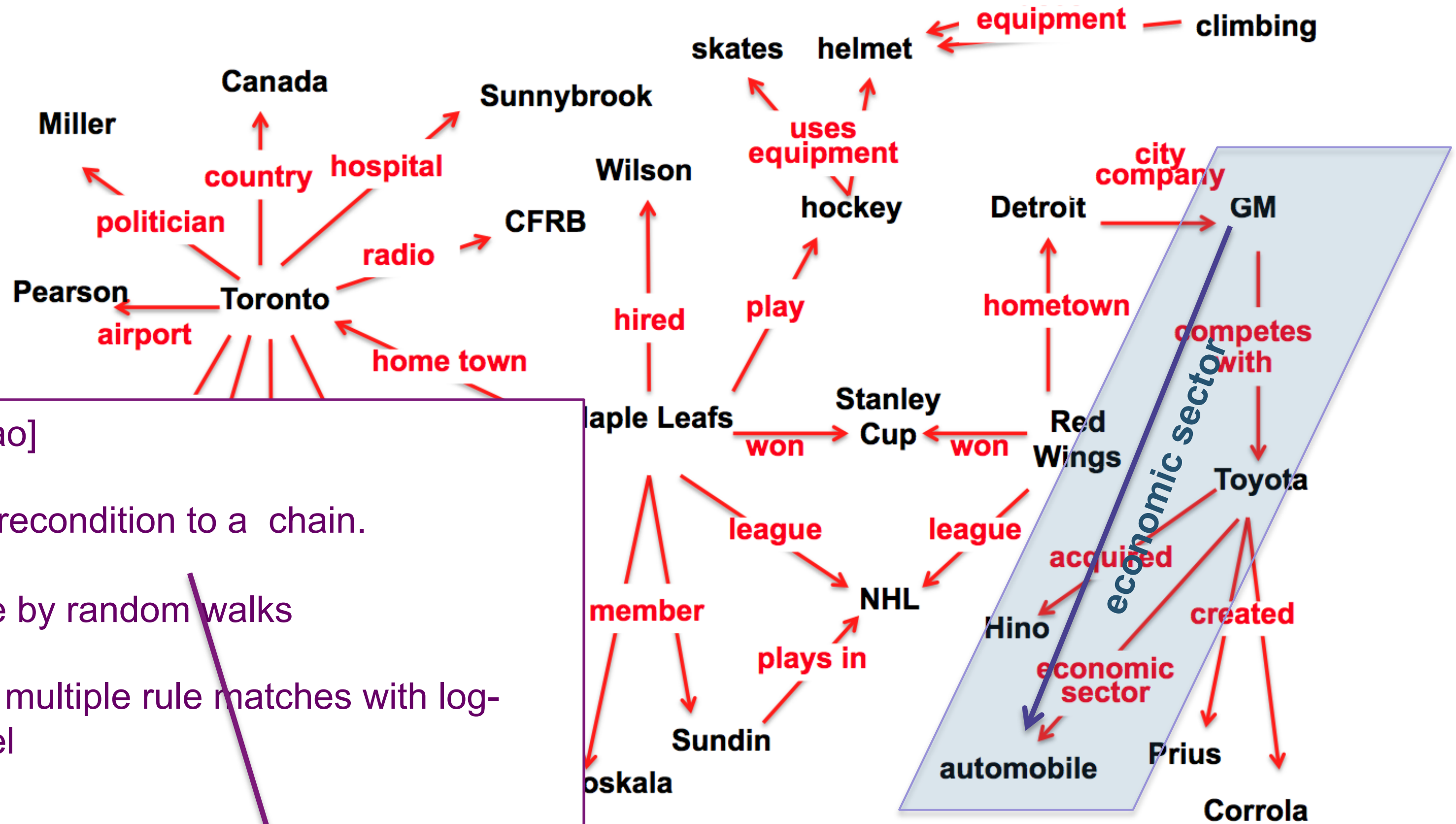


If: x1 — competes with (x1,x2) → x2 — economic sector (x2, x3) → x3

Then: economic sector (x1, x3)

Inference by Random Walks

PRA: [Lao, Mitchell, Cohen, *EMNLP* 2011]



PRA: [Ni Lao]

1. restrict precondition to a chain.

2. inference by random walks

3. combine multiple rule matches with log-linear model

If: x1 — competes with (x1,x2) → x2 — economic sector (x2, x3) → x3

Then: economic sector (x1, x3)

Course Outline

1. Basic theories and practices on named entity recognition.

2. Recent advances in relation extraction:
   a. distant supervision
   b. latent variable models

3. Scalable IE and reasoning with first-order logics.

Recent Advances in IE: Distant Supervision

Relation Extraction

Predict relations between entities based on mentions (Cullota and Sorenson, 2004)

Example: learn the *mascot(object, org)* relation.

Training data:

*“A **Scottish Terrier** has clearly won the hearts of the campus community and will become **Carnegie Mellon**'s new official mascot”*

Challenge

It is very expensive to obtain labeled training data.

Distant Supervision

Idea: if we know the relation between two entities, then any sentence that includes these two entities is likely to express the same relation.

Distant Supervision

Mintz, Bills, Snow, Jurafsky. 2009. Distant supervision for relation extraction without labeled data. ACL-2009.

Use a knowledge base of known relations to collect a lot of noisy training data.

Distant Supervision

Example: *mascot(Stanford_tree,Stanford_Band).*

High quality examples:
*“The **Stanford Tree** is the **Stanford Band**'s mascot.”*

*“Called — appropriately — the **Stanford Tree**, it is the official mascot of the **band**.”*

Noisy examples:
*“The **Stanford band** invites you to be **Tree** for a day.”*

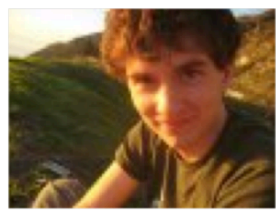
Distant Supervision: Pros

- **Has the advantages of supervised learning**
  - leverage rich, reliable hand-created knowledge
  - can use rich features (e.g. syntactic features)

- **Has the advantages of unsupervised learning**
  - leverage unlimited amounts of text data
  - allows for very large number of weak features
  - not sensitive to training corpus: genre independent

Mintz et al., (2009) ACL

Mintz, Bills, Snow, Jurafsky (2009).
Distant supervision for relation extraction without labeled data.

Training set	**Corpus**
Freebase	WIKIPEDIA
102 relations 940,000 entities 1.8 million instances	1.8 million articles 25.7 million sentences

Frequent Freebase Relations

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Collecting Training Data

Corpus text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Training data

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Collecting Training Data

Corpus text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Training data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Collecting Training Data

Corpus text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Training data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Processing Testing Data

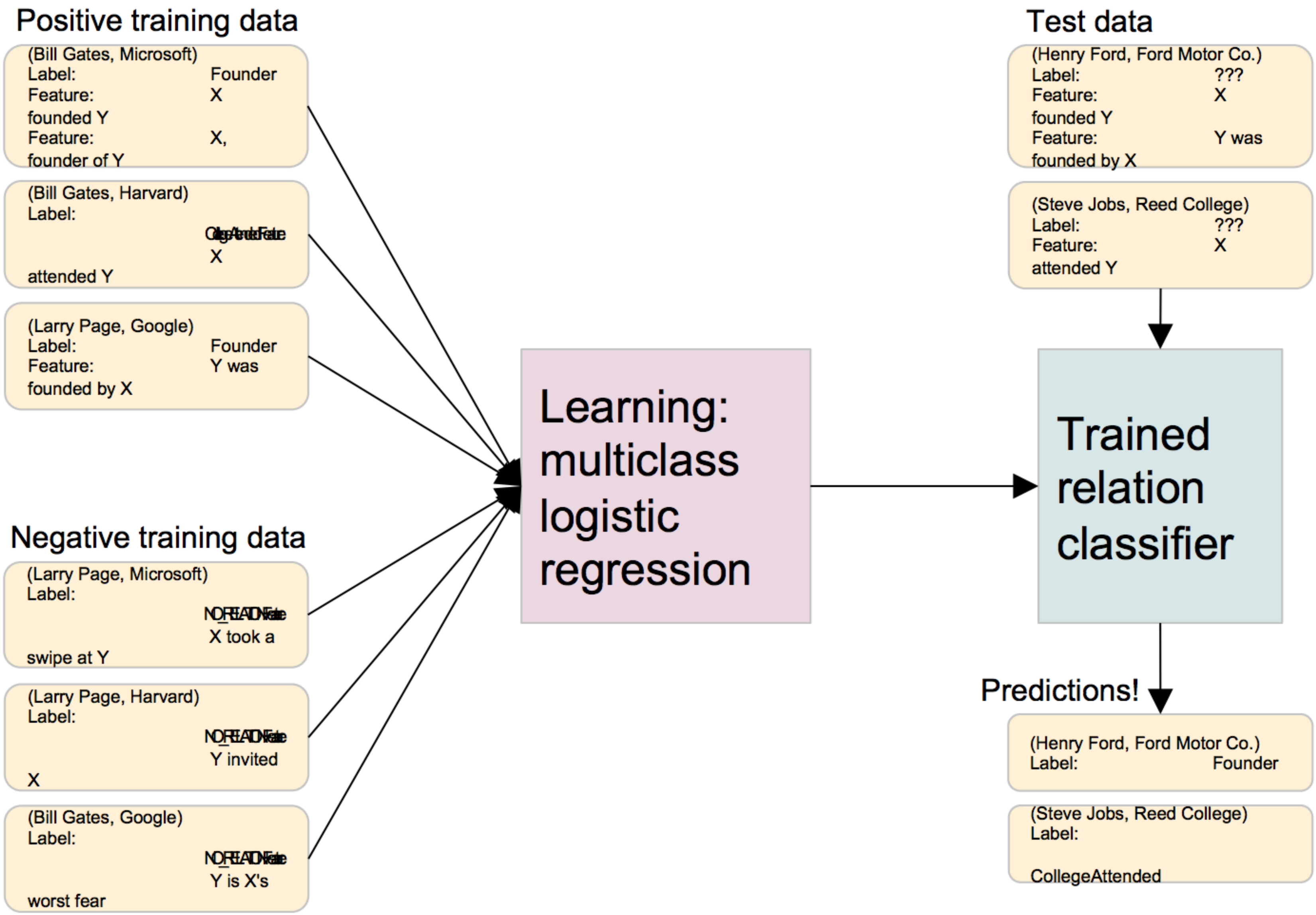
Test data

(Henry Ford, Ford Motor Co.)
Label: ???
Feature: X founded Y
Feature: Y was founded by X

Corpus text

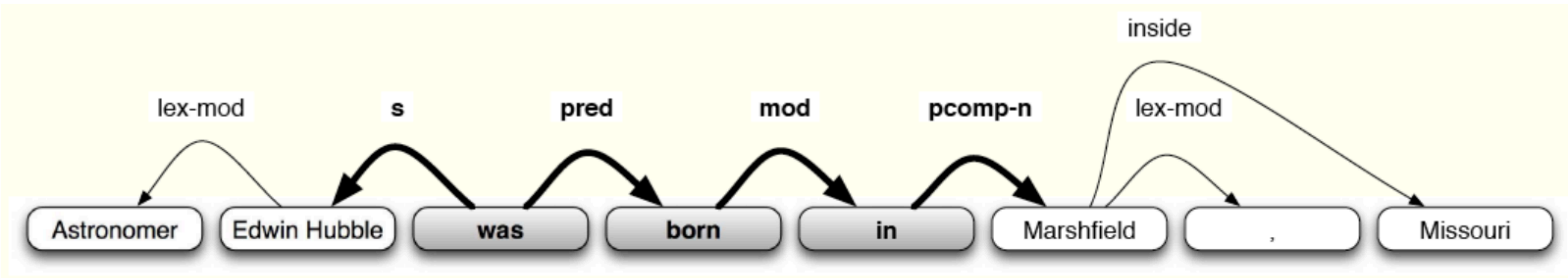
Henry Ford founded Ford Motor Co. in…
Ford Motor Co. was founded by Henry Ford…
Steve Jobs attended Reed College from…

The Experiment



Lexical and Dependency Path Features

Astronomer Edwin Hubble was born in Marshfield, Missouri.



Feature type	Left window	NE1	Middle	NE2	Right window
Lexical	[]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[]
Lexical	[Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[,]
Lexical	[#PAD#, Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[, Missouri]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]

Experimental Settings

- **1.8 million relation instances used for training**
- **800,000 Wikipedia articles used for training, 400,000 different articles used for testing**
- **Only extract relation instances not already in Freebase**

Learned Relational Facts

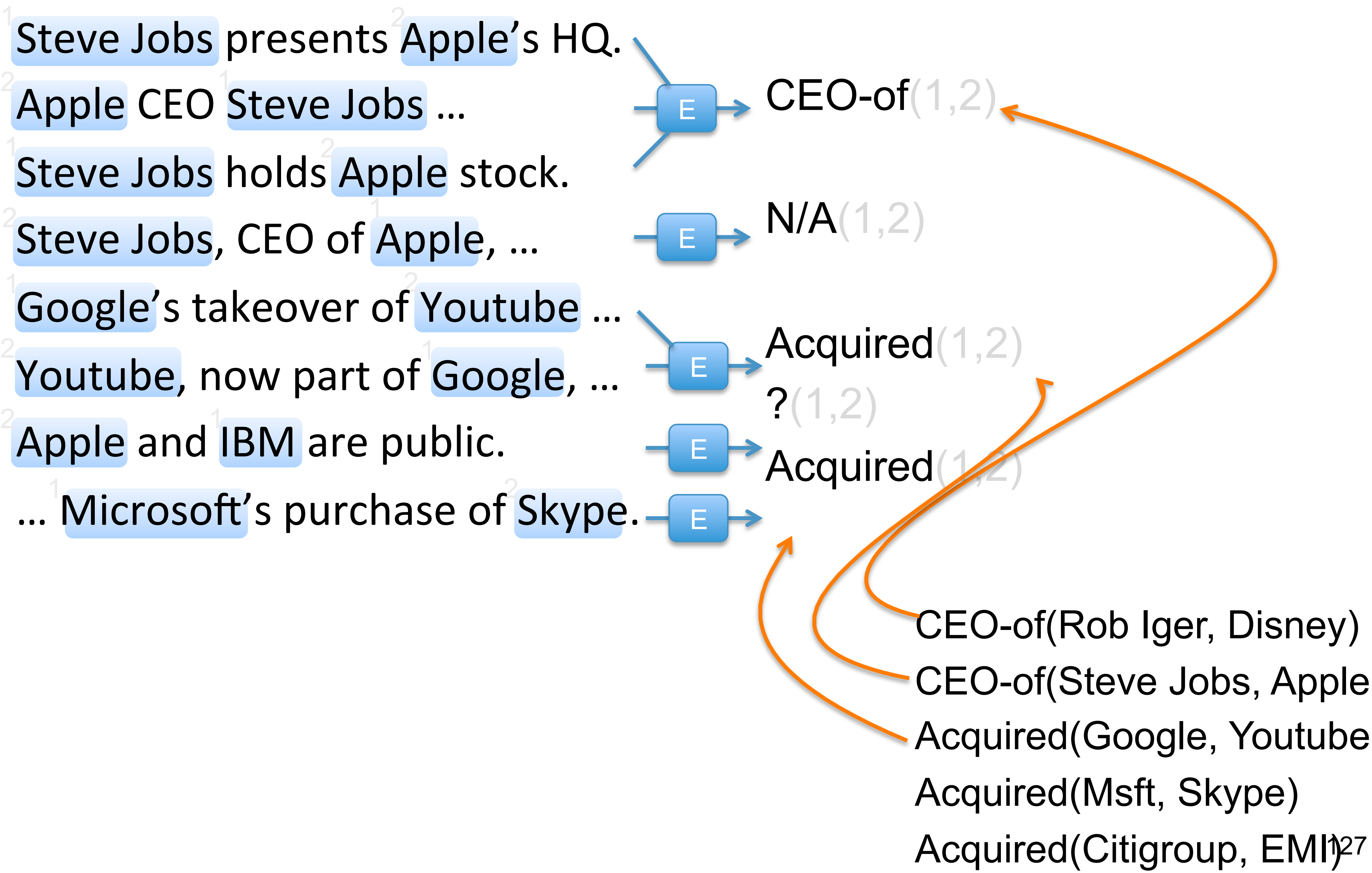
Relation name	New instance
/location/location/contains	Paris, Montmartre
/location/location/contains	Ontario, Fort Erie
/music/artist/origin	Mighty Wagon, Cincinnati
/people/deceased_person/place_of_death	Fyodor Kamensky, Clearwater
/people/person/nationality	Marianne Yvonne Heemskerk, Netherlands
/people/person/place_of_birth	Wavell Wayne Hinds, Kingston
/book/author/works_written	Upton Sinclair, Lanny Budd
/business/company/founders	WWE, Vince McMahon
/people/person/profession	Thomas Mellon, judge

Human Evaluation

Precision, using Mechanical Turk labelers:

Relation name	100 instances			1000 instances		
	Syn	Lex	Both	Syn	Lex	Both
/film/director/film	**0.49**	0.43	0.44	**0.49**	0.41	0.46
/film/writer/film	**0.70**	0.60	0.65	**0.71**	0.61	0.69
/geography/river/basin_countries	0.65	0.64	**0.67**	**0.73**	0.71	0.64
/location/country/administrative_divisions	0.68	0.59	**0.70**	**0.72**	0.68	**0.72**
/location/location/contains	0.81	**0.89**	0.84	**0.85**	0.83	0.84
/location/us_county/county_seat	0.51	0.51	**0.53**	0.47	**0.57**	0.42
/music/artist/origin	0.64	0.66	**0.71**	0.61	**0.63**	0.60
/people/deceased_person/place_of_death	0.80	0.79	**0.81**	0.80	**0.81**	0.78
/people/person/nationality	0.61	0.70	**0.72**	0.56	0.61	**0.63**
/people/person/place_of_birth	**0.78**	0.77	**0.78**	0.88	0.85	**0.91**
Average	0.67	0.66	**0.69**	**0.68**	0.67	0.67

Mintz et al. : Aggregate Extraction



Mintz et al. (2009)

Issues?

- No multi-instance learning

- No multi-relation learning

Multi-Instance Learning

[1]Steve Jobs presents [2]Apple's HQ. → E → ?(1,2)=N/A(1,2)

[2]Apple CEO [1]Steve Jobs … → E → ?(1,2)=CEO-of(1,2)

[1]Steve Jobs holds [2]Apple stock. → E → ?(1,2)=N/A(1,2)

[2]Steve Jobs, CEO of [1]Apple, … → E → ?(1,2)

[1]Google's takeover of [2]Youtube … → E → ?(1,2)

[2]Youtube, now part of [1]Google, … → E → ?(1,2)

[2]Apple and [1]IBM are public. → E → ?(1,2)

… [1]Microsoft's purchase of [2]Skype. → E → ?(1,2)

∨

CEO-of(Rob Iger, Disney)

CEO-of(Steve Jobs, Apple)

Acquired(Google, Youtube)

Acquired(Msft, Skype)

Acquired(Citigroup, EMI)

Cf. [Bunescu, Mooney 07], [Riedel, Yao, McCallum 10])

Overlapping Relations

[1]Steve Jobs presents [2]Apple's HQ. → E → ?(1,2)=N/A(1,2)

[2]Apple CEO [1]Steve Jobs … → E → ?(1,2)=CEO-of(1,2)

[1]Steve Jobs holds [2]Apple stock. → E → ?(1,2)=SH-of(1,2)

[2]Steve Jobs, CEO of [1]Apple, … → E → ?(1,2)

[1]Google's takeover of [2]Youtube … → E → ?(1,2)

[2]Youtube, now part of [1]Google, … → E → ?(1,2)

[2]Apple and [1]IBM are public. → E → ?(1,2)

… [1]Microsoft's purchase of [2]Skype. → E → ?(1,2)

V

SH-of(Steve Jobs, Apple)

CEO-of(Rob Iger, Disney)

CEO-of(Steve Jobs, Apple)

Acquired(Google, Youtube)

Acquired(Msft, Skype)

Acquired(Citigroup, EMI)

Hoffman et al. (2011)

Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, Daniel S. Weld
Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
{raphaelh,clzhang,xiaoling,lsz,weld}@cs.washington.edu

Sentence-Level Learning

Steve Jobs presents Apple's HQ. → E → ?(1,2)

Apple CEO Steve Jobs … → E → ?(1,2)

Steve Jobs holds Apple stock. → E → ?(1,2)

Steve Jobs, CEO of Apple, … → E → ?(1,2)

Google's takeover of Youtube … → E → ?(1,2)

Youtube, now part of Google, … → E → ?(1,2)

Apple and IBM are public. → E → ?(1,2)

… Microsoft's purchase of Skype. → E → ?(1,2)

$\vee$

Train so that extracted facts match facts in DB

CEO-of(Rob Iger, Disney)

CEO-of(Steve Jobs, Apple)

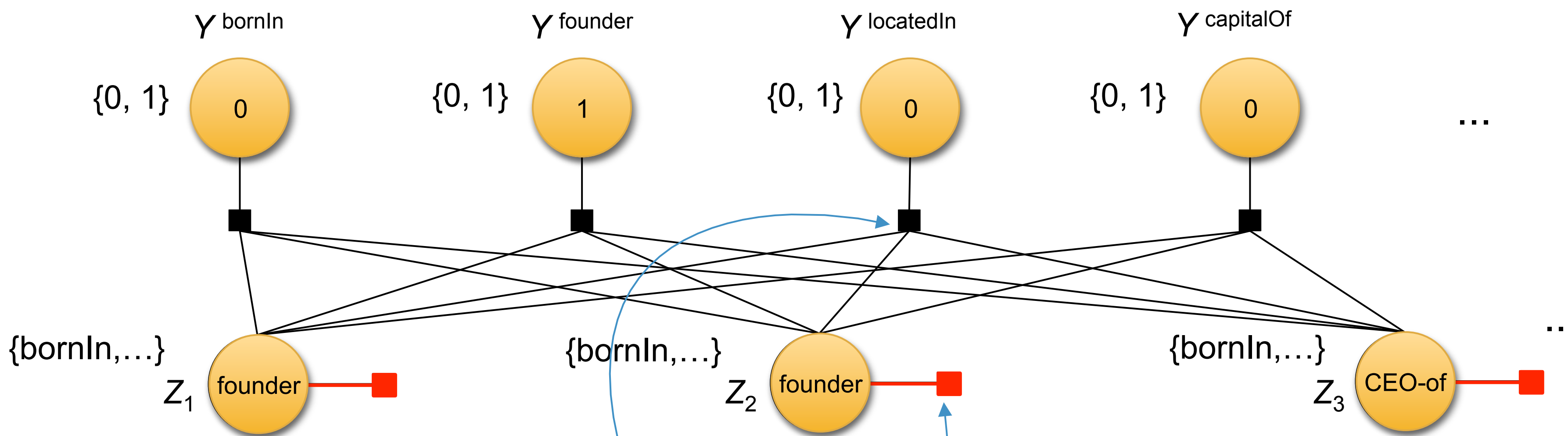
Acquired(Google, Youtube)

Acquired(Msft, Skype)

Acquired(Citigroup, EMI)

Steve Jobs, Apple:

Model



Steve Jobs was founder of **Apple**.

Steve Jobs, Steve Wozniak and Ronald Wayne founded **Apple**.

Steve Jobs is CEO of **Apple**.

...

$$p(\mathbf{Y}=\mathbf{y},\mathbf{Z}=\mathbf{z}|\mathbf{x};\theta) \stackrel{\text{def}}{=} \frac{1}{Z_x}\prod_r \Phi^{\text{join}}(y^r,\mathbf{z})\prod_i \Phi^{\text{extract}}(z_i,x_i)$$

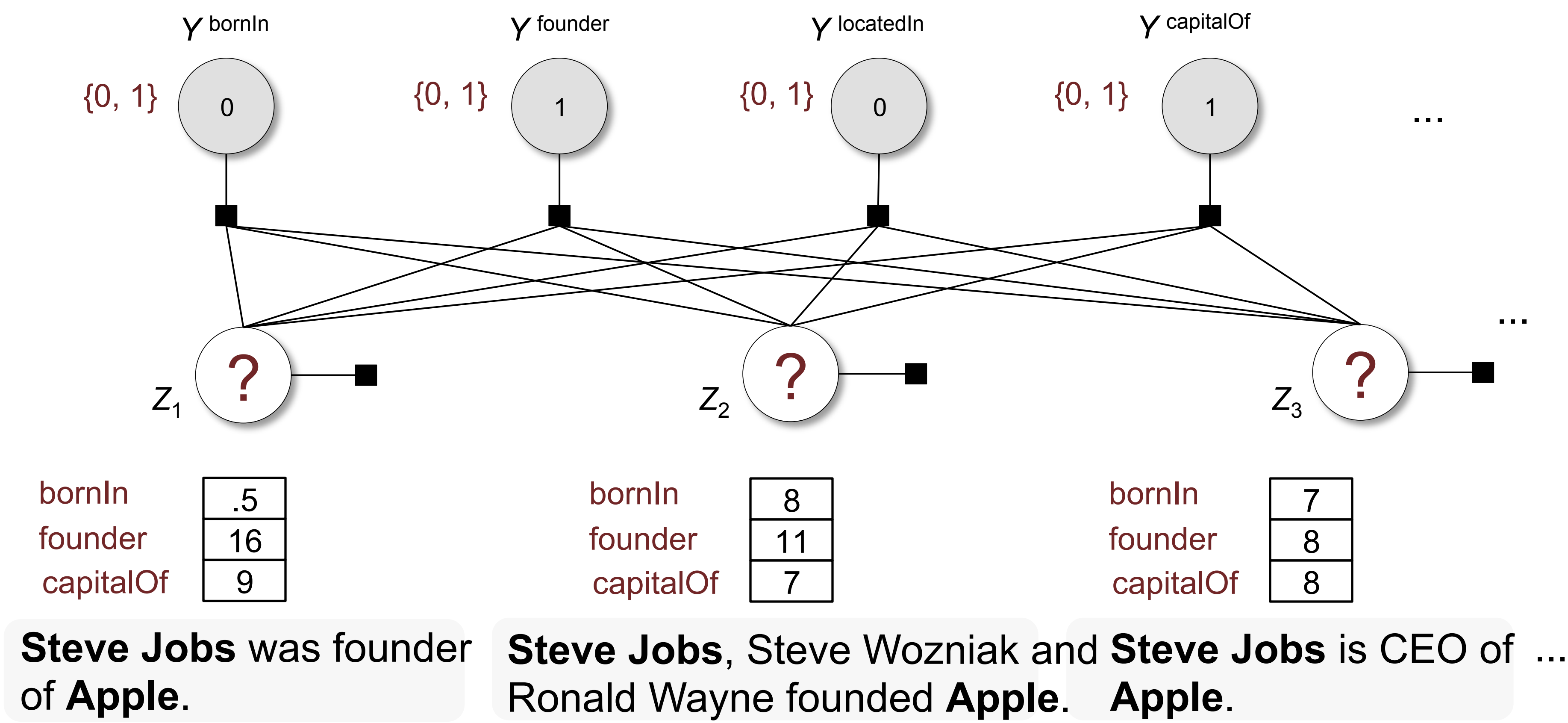
$$\Phi^{\text{join}}(y^r,\mathbf{z}) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } y^r = true \wedge \exists i : z_i = r \\ 0 & \text{otherwise} \end{cases}$$

All features at sentence-level

(join factors are deterministic ORs)

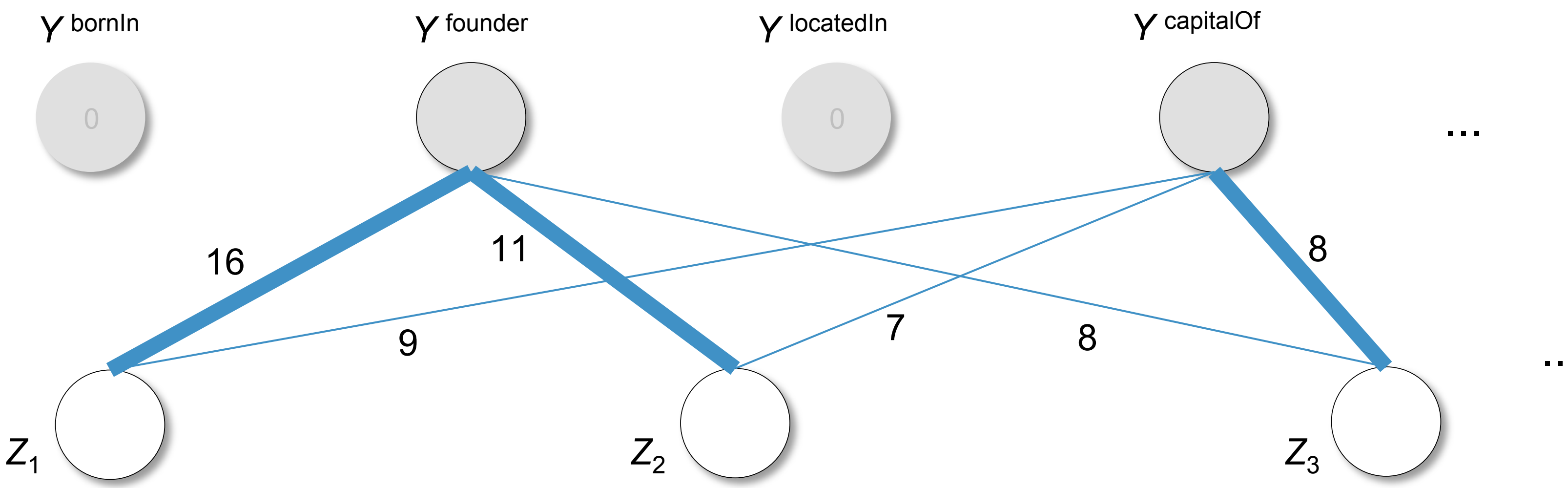
Inference

Computing $\arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}, \mathbf{y}; \theta)$:



Inference

Variant of the weighted, edge-cover problem:



bornIn	.5
founder	16
capitalOf	9

bornIn	8
founder	11
capitalOf	7

bornIn	7
founder	8
capitalOf	8

Steve Jobs was founder of **Apple**.

Steve Jobs, Steve Wozniak and Ronald Wayne founded **Apple**.

Steve Jobs is CEO of **Apple**.

...

Learning

Training set $\{(\mathbf{x}_i, \mathbf{y}_i) | i = 1 \ldots n\}$, where

i corresponds to a particular entity pair

$\mathbf{x}_i$ contains all sentences with mentions of pair

$\mathbf{y}_i$ bit vector of facts about pair from database

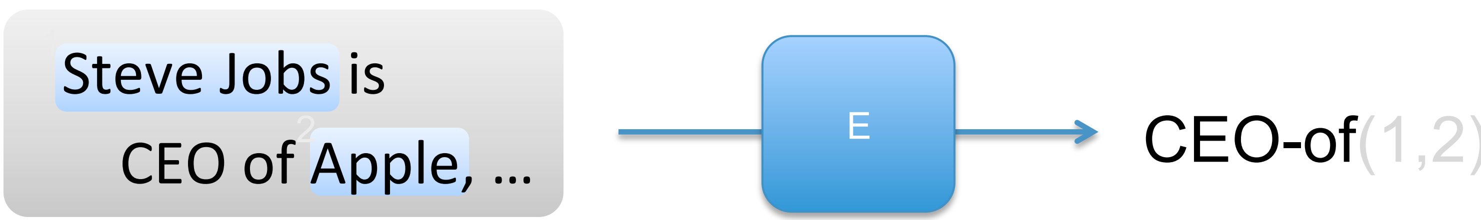
Maximize Likelihood

$$O(\theta) = \prod_i p(\mathbf{y}_i | \mathbf{x}_i; \theta) = \prod_i \sum_{\mathbf{z}} p(\mathbf{y}_i, \mathbf{z} | \mathbf{x}_i; \theta)$$

Sentential vs. Aggregate Extraction

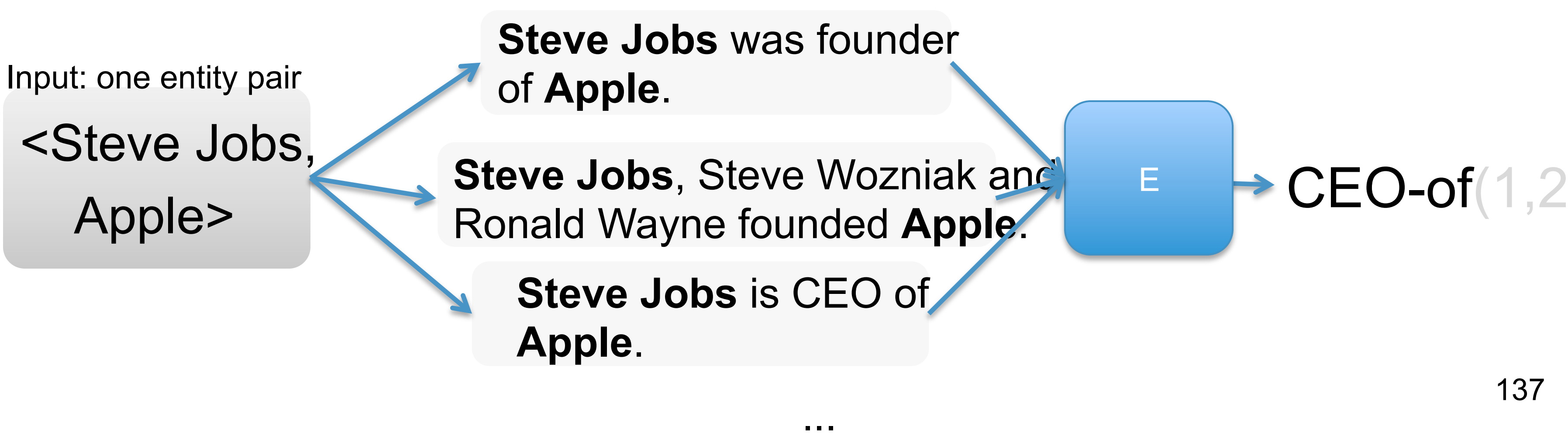
Sentential

Input: one sentence



Aggregate

Input: one entity pair



Distant Supervision: Related Work

- Mintz, Bills, Snow, Jurafsky 09:
  - Extraction at aggregate level
  - Features: conjunctions of lexical, syntactic, and entity type info along dependency path
- Riedel, Yao, McCallum 10:
  - Extraction at aggregate level
  - Latent variable on sentence
- Bunescu, Mooney 07:
  - Multi-instance learning for relation extraction
  - Kernel-based approach

Experimental Setup

- Data as in Riedel et al. 10:

LDC NYT corpus, 2005-06 (training), 2007 (testing)

Data first tagged with Stanford NER system

Entities matched to Freebase, ~ top 50 relations

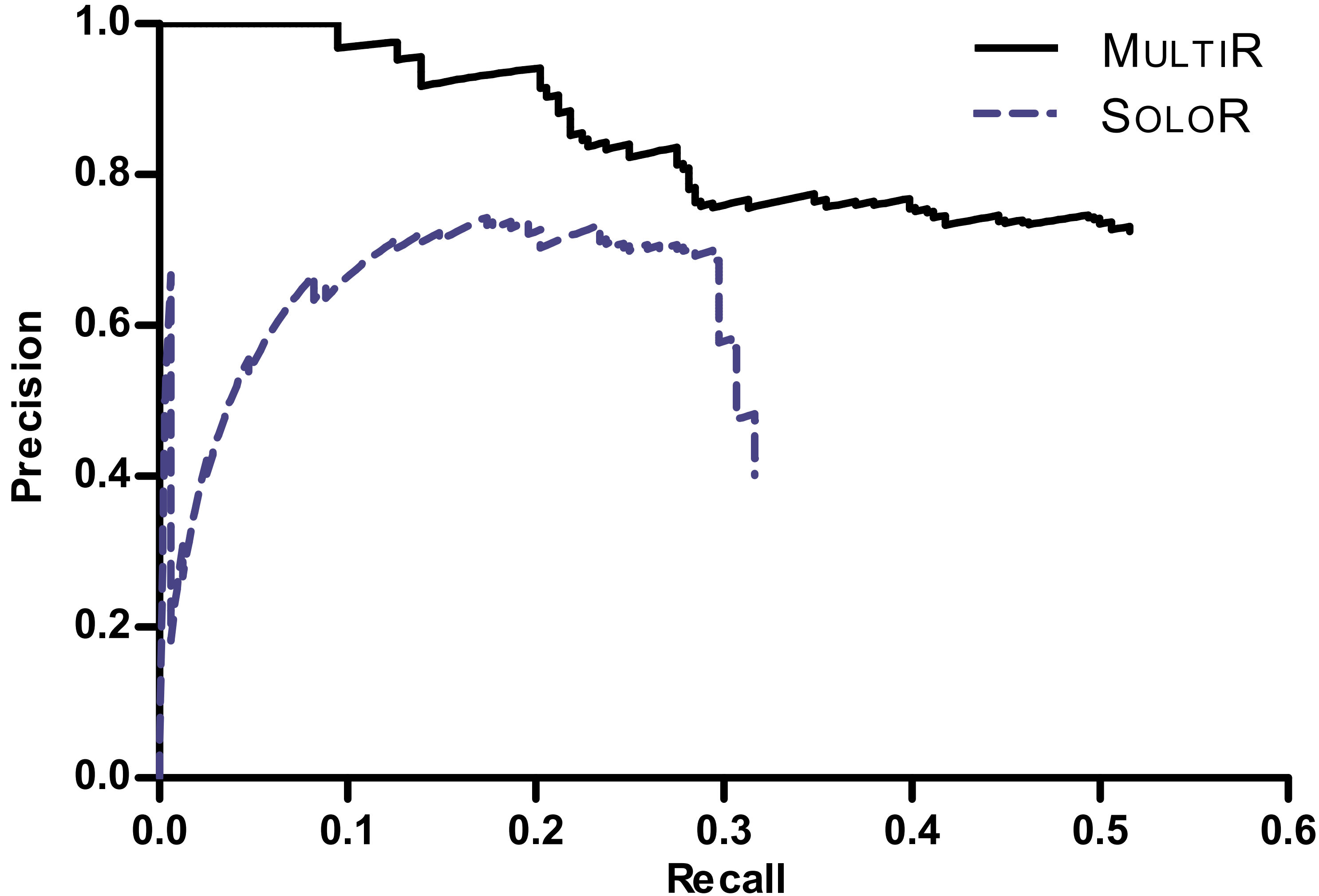
Mention-level features as in Mintz et al. 09

- Systems:

MultiR: proposed approach

SoloR: re-implementation of Riedel et al. 2010

Sentential Extraction



Distant Supervision: Conclusion

- Widely used in the IE community nowadays.
- A much cheaper way of obtaining training data
- Still, there‘s room for improvement:
- what about entities that are not in Freebase?
- what if entities are in Freebase, but no relation is recorded?

Recent Advances in IE: Latent Variable Modeling

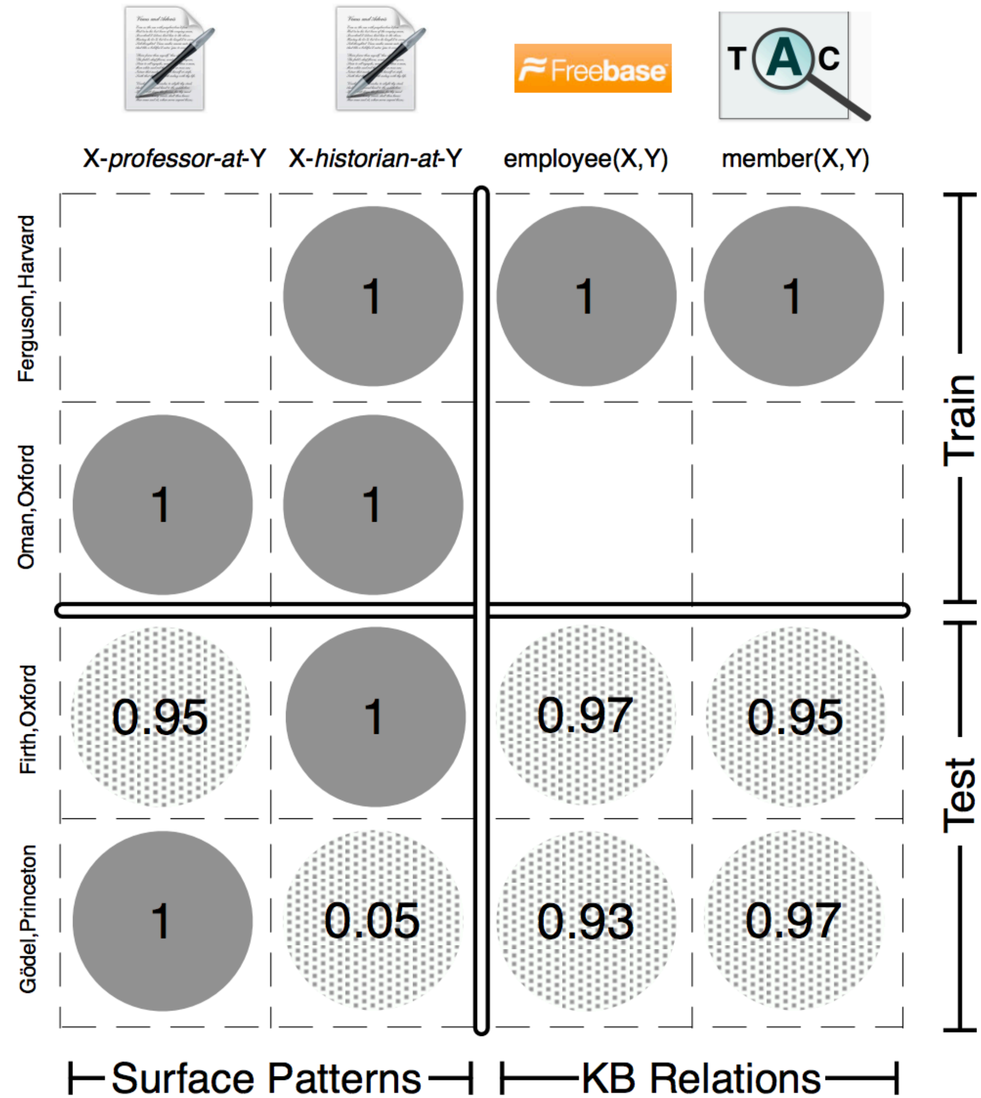
Universal Schema

- Riedel et al., NAACL 2013. Relation Extraction with Matrix Factorization and Universal Schemas.
- Motivation: use **matrix representation** for relation extraction.
- Idea: put all training and testing data into a matrix, and fill in the missing values.
- Jointly learn latent factor representation for surface patterns and multiple relations.

Universal Schema

- Rows: pair of entities. e.g., (William, CMU)

- Columns: surface patterns and relations. e.g., X-is_a_professor_at-Y teaches (X, Y)

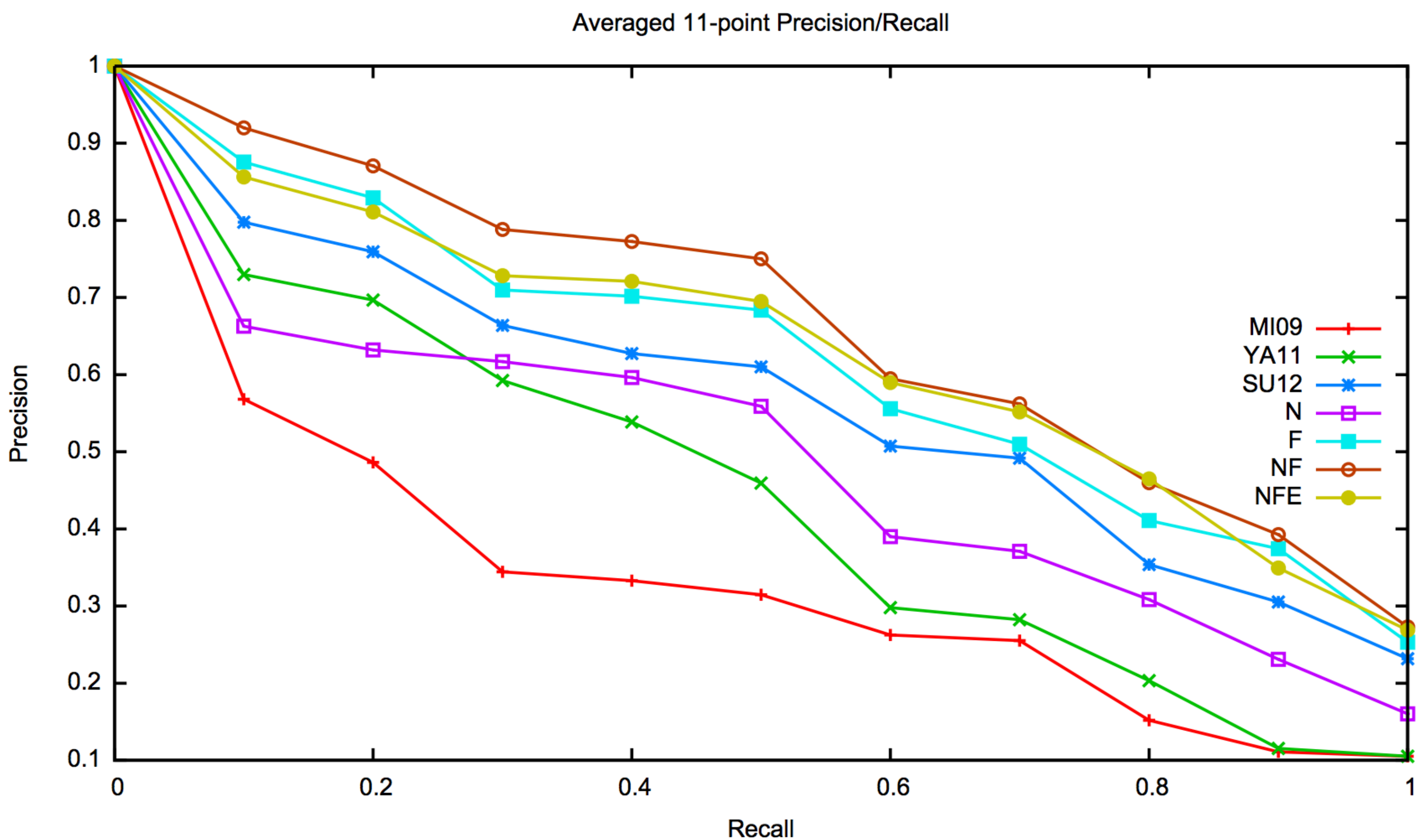


Matrix Factorization

- Approach: Bayesian Personalized Ranking
  (Rendle et al., 2009)
- Requires: negative training data.
- How to collect negative data: both entities of the entity pair occur in Freebase, however, Freebase does not say there is a relation between them.

Performance

- Dataset: Freebase + NewYorkTimes.



Universal Schema

- Pros:

1) language, schema independent
2) joint learning of surface patterns and relations
3) scalability

- Cons:

1) explainability
2) requires negative examples

Course Outline

1. Basic theories and practices on named entity recognition: supervised and semi-supervised.

2. Recent advances in relation extraction:
    a. distant supervision
    b. latent variable models

3. Scalable IE and reasoning with first-order logics.

Joint IE and Reasoning

A Motivating Example…

An elementary school student was sent to detention by his Math teacher after school. When he got home, his father said: “Ma Yun, what happen to you at school today?” Ma: “Sorry dad, I was playing with a magnet, but it attracted Mrs. Smith’s golden ring. Then, Mrs. Smith went out to cry, and slapped the P.E. teacher in the face.”

Query:
Who is most likely the husband of Mrs. Smith?

This example was adapted from Weibo.

Reasoning

An elementary school student was sent to detention by his Math teacher after school. When he got home, his father said: “Ma Yun, what happen to you at school today?” : “Sorry dad, I was playing with a magnet, but it attracted Mrs. Smith’s golden ring. Then, Mrs. Smith went out to cry, and slapped the P.E. teacher in the face.”

Magnet

From Wikipedia, the free encyclopedia

This article is about objects and devices that produce magnetic fields. For a description of magnetic materials Magnet (disambiguation).

This article **needs additional citations for verification**. Relevant discussion may be page. Please help improve this article by adding citations to reliable sources. Unsourc challenged and removed. *(July 2011)*

A **magnet** (from Greek μαγνήτις λίθος *magnḗtis líthos*, "Magnesian stone") is a material or object that produces a magnetic field. This magnetic field is invisible but is responsible for the most notable property of a magnet: a force that pulls on other ferromagnetic materials, such as iron, and attracts or repels other magnets.

attract (magnet, golden_ring)

conflict (iron, golden_ring)

attract (magnet, iron)

slap (Mrs. Smith, P.E. Teacher)

husband (Mrs. Smith, P.E. Teacher)

This example was adapted from Weibo.

Issues with Modern IE Systems

- No relational KB inference is performed at extraction time (or no inference at all).
- Classification is not the panacea.
- Big pipeline: error cascades.

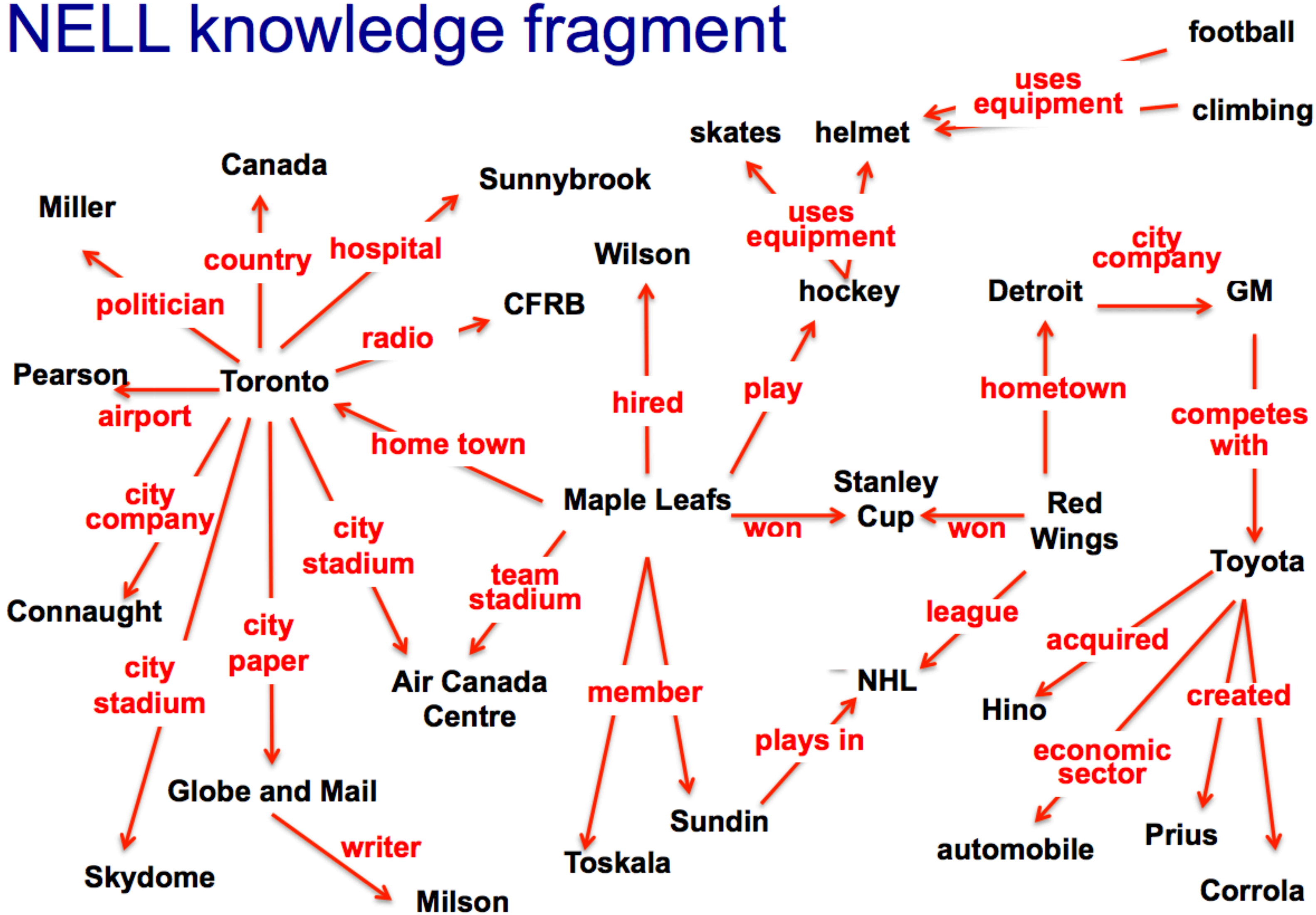
Motivations

- To deal with complexity, we need first-order logics to perform reasoning.

- To deal with uncertainty, we need statistical/ probabilistic approaches, at the same time.

Knowledge Base Inference

NELL knowledge fragment



Issues with KB Reasoning Systems

- Often done using relational triples (e.g., wife(barack,michelle)) after IE, and key contextual information is lost.

E.g., Path-Ranking Algorithm (Ni et al., 2010)

PRA Paths for inferring **athletePlaysSport**:

athletePlaysSport(A,S):- factAthletePlaysForTeam(A,T),factTeamPlaysSport(T,S).

PRA Paths for inferring **teamPlaysSport**:

teamPlaysSport(T,S):-
    factMemberOfConference(T,C),factConferenceHasMember(C,T'),factTeamPlaysSport(T',S).

teamPlaysSport(T,S):-
    factTeamHasAthlete(T,A),factAthletePlaysSport(A,S).

Our Approach

- presents a joint IE and reasoning model in a statistical relational learning setting;
- incorporates latent contexts into probabilistic first-order logics.

Agenda

- Motivation
- Background: ProPPR
- Datasets
- Joint IE and Structure Learning
- Experiments
- Conclusion

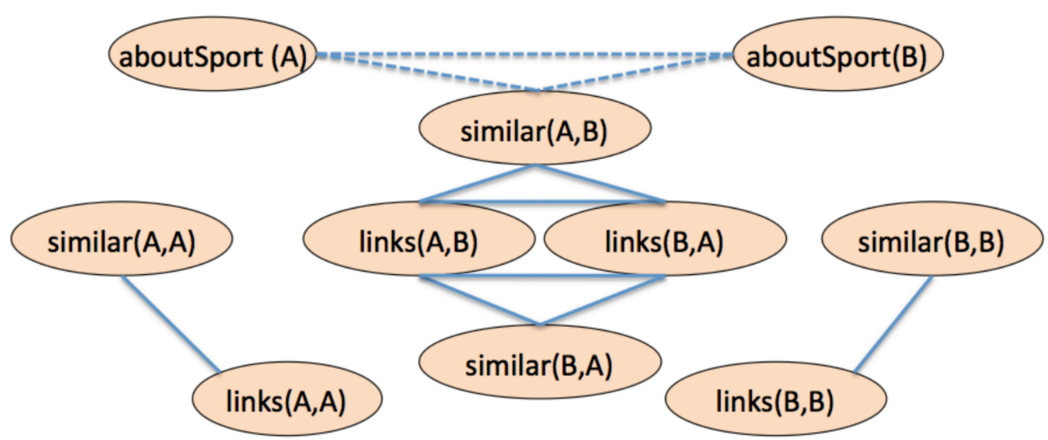
Wait, Why Not Markov Logic Network?

network size is $O(n^a)$, where a = #arity.

e.g., holdStock(person,company)

R1 2.0 $\forall X,Y \; links(X,Y) \vee links(Y,X) \Rightarrow similar(X,Y)$

R2 1.5 $\forall X,Y \; similar(X,Y) \Rightarrow \left(aboutSports(X) \Leftrightarrow aboutSports(Y)\right)$

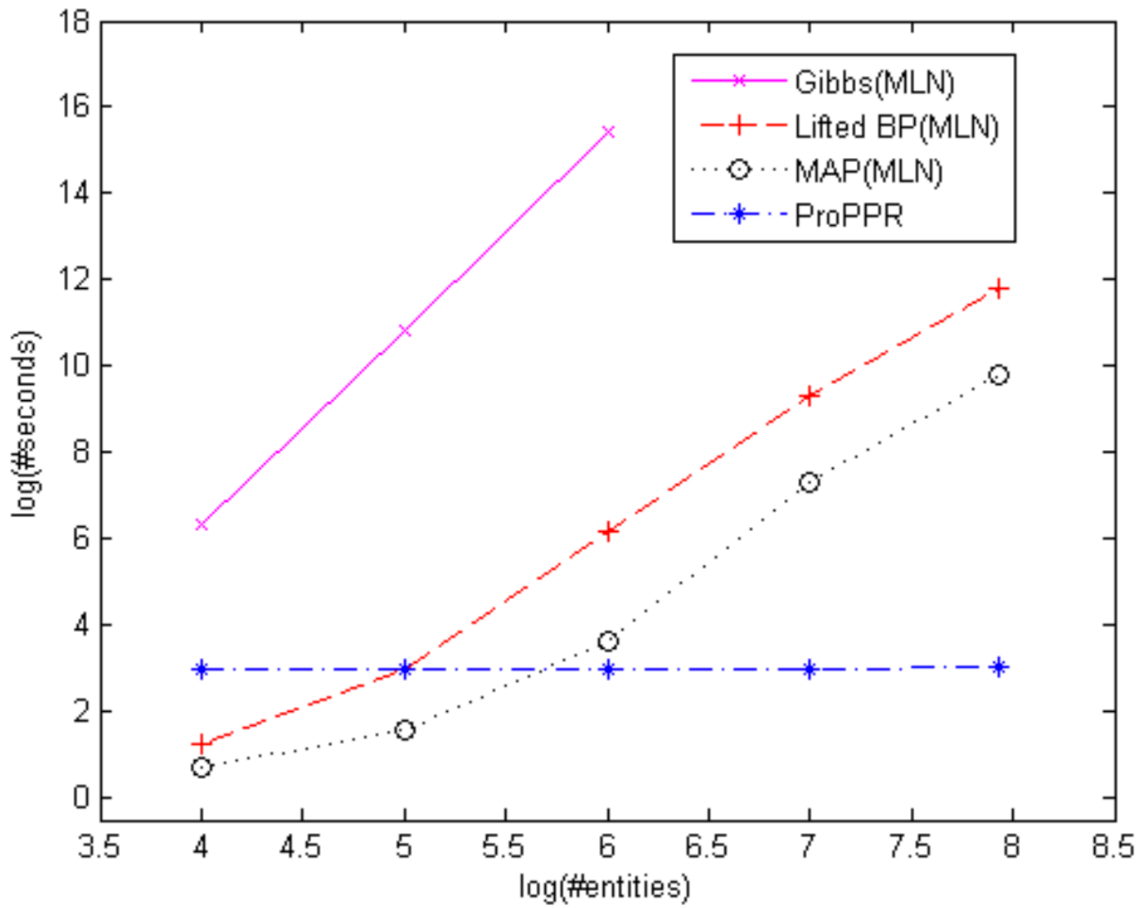


Inference time often depends on graph size.

Programming with Personalized PageRank (ProPPR)

- CIKM 2013 best paper honorable mention
- is a probabilistic first-order logic
- can be used in:
- entity resolution, classification (Wang et al., 2013)
- dependency parsing (Wang et al., 2014 EMNLP)
- large-scale KB inference (Wang et al., 2015 MLJ)
- logic programming (Wang et al., 2015 IJCAI)

Inference Time Comparison



ProPPR's inference time is independent of the size of the graph (Wang et al., 2013).

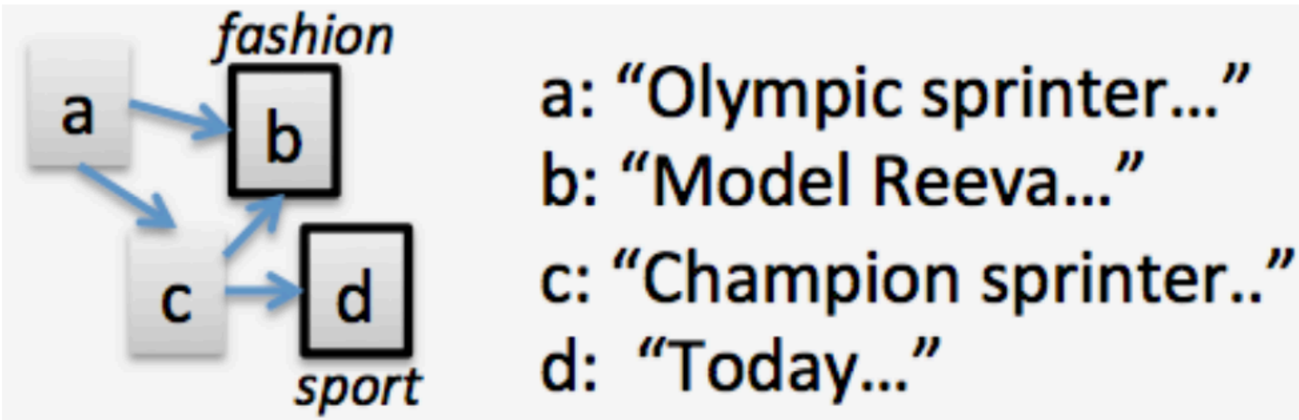
Accuracy: Citation Matching

	Cites	Authors	Venues	Titles
MLN Our rules	0.513	0.532	0.602	0.544
ProPPR(**w**=1)	0.680	0.836	0.860	**0.908**
ProPPR	**0.800**	**0.840**	**0.869**	0.900

AUC scores: 0.0=low, 1.0=hi
w=1 is before learning
(i.e., heuristic matching rules, weighted with PPR)

ProPPR Example

Input:



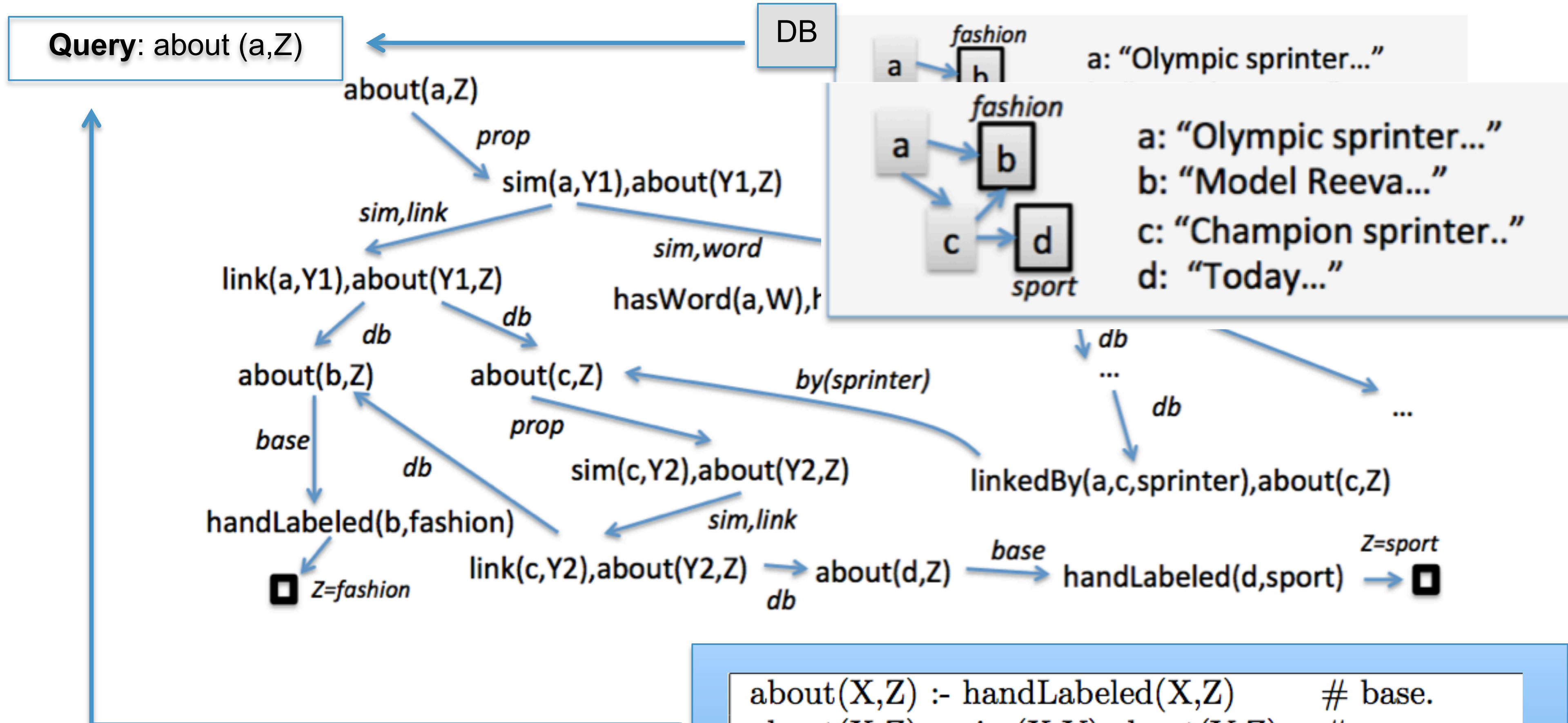
Query: *about(a,?)*

An Example ProPPR Program

```
about(X,Z) :- handLabeled(X,Z)              # base.
about(X,Z) :- sim(X,Y),about(Y,Z)           # prop.
sim(X,Y) :- links(X,Y)                      # sim,link.
sim(X,Y) :-
    hasWord(X,W),hasWord(Y,W),
    linkedBy(X,Y,W)                         # sim,word.
linkedBy(X,Y,W) :- true                     # by(W).
```

Feature Vector

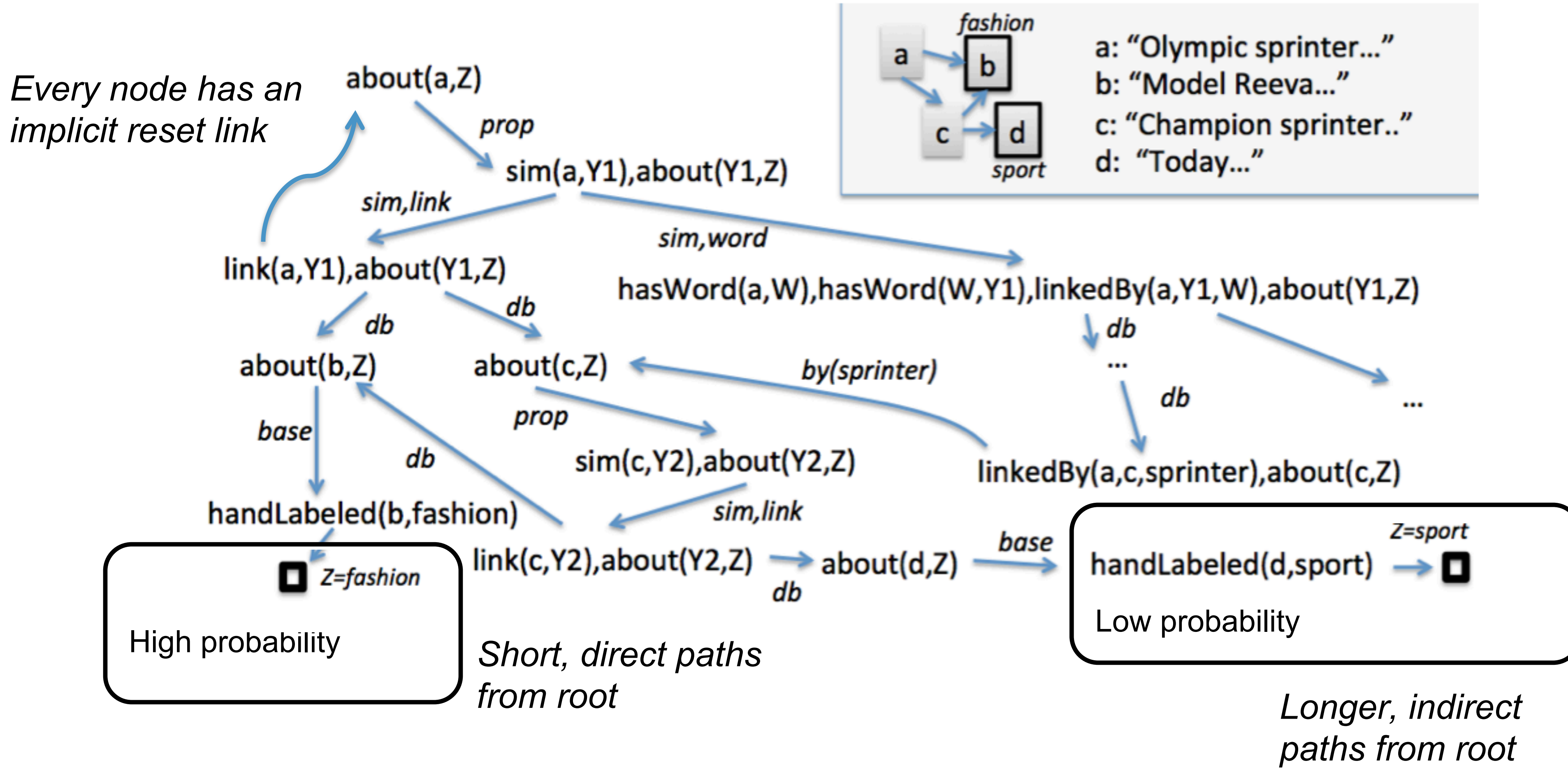
Feature Template



Program + DB + Query define a *proof graph*, where nodes are *conjunctions of goals* and edges are labeled with sets of *features.*

```
about(X,Z) :- handLabeled(X,Z)          # base.
about(X,Z) :- sim(X,Y),about(Y,Z)       # prop.
sim(X,Y) :- links(X,Y)                  # sim,link.
sim(X,Y) :-
    hasWord(X,W),hasWord(Y,W),
    linkedBy(X,Y,W)                     # sim,word.
linkedBy(X,Y,W) :- true                 # by(W).
```

Program (label propagation) | LHS ➔ features



Transition probabilities, Pr(child|parent), plus Personalized PageRank (aka Random-Walk-With-Reset) define a *distribution over nodes.*

Very fast *approximate* methods for PPR

Transition probabilities, Pr(child|parent), are defined by **weighted sum of edge features**, followed by normalization.

Learning via pSGD

Approximate Inference in ProPPR

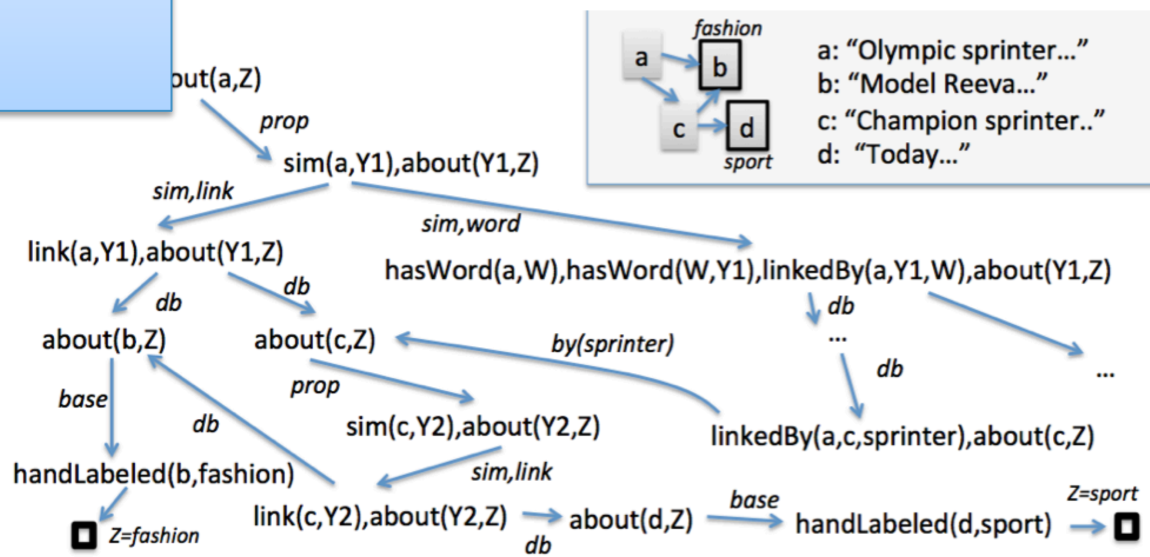
- Score for a query soln (e.g., "Z=sport" for "about(a,Z)") depends on *probability* of reaching a ☐ node*

"Grounding" (proof tree) size is O(1/αε) … ie *independent* of DB size ➔ fast approx incremental inference (Reid,Lang,Chung, 08)

α is reset probability

*as in Stochastic Logic Programs [Cussens, 2001]

Basic idea: **incrementally expand the tree from the query node** until all nodes *v* accessed have weight below *ε/degree(v)*



Parameter Learning in ProPPR

PPR probabilities are a stationary distribution of a Markov chain

reset

$$\mathbf{p}^{t+1} \equiv \alpha\mathbf{s} + (1-\alpha)\mathbf{M}\mathbf{p}^{t}$$

Transition probabilities u→v are derived by **linearly** combining features of an edge, applying a **squashing** function *f*, and normalizing

$$s_{uv} \equiv \vec{\phi}_{uv} \cdot \mathbf{w}$$

$$t_u \equiv \sum_{v'} f(s_{uv'})$$

$$\mathbf{M}_{u,v} \equiv \frac{f(s_{uv})}{t_u}$$

f is exp, truncated *tanh,* ReLU…

Parameter Learning in ProPPR

PPR probabilities are a stationary distribution of a Markov chain

$$\mathbf{p}^{t+1} \equiv \alpha\mathbf{s} + (1-\alpha)\mathbf{M}\mathbf{p}^t$$

Learning uses gradient descent: derivative $\mathbf{d}^t$ of $\mathbf{p}^t$ is :

$$\mathbf{d}^{t+1} = \frac{\partial}{\partial \mathbf{w}}\mathbf{p}^{t+1} = (1-\alpha)\left((\frac{\partial}{\partial \mathbf{w}}\mathbf{M})\mathbf{p}^t + \mathbf{M}\mathbf{d}^t \right)$$

Overall algorithm not unlike backprop…we use parallel SGD

Where Does the Program Come From?

- Traditionally by hand.
- We use structure learning to automatically learn first-order logic clauses from data.
- Idea (CIKM 2014):

    build a second-order abductive logic

    whose parameters correspond to 1st-order theory

    reduce the structure learning to parameter learning.

Logic program is an *interpreter* for a program containing *all possible rules* from a sublanguage

Query$_0$: sibling(malia,Z)

DB$_0$: sister(malia,sasha), mother(malia,michelle), …

Query: interp(sibling,malia,Z)

DB: rel(sister,malia,sasha), rel(mother,malia,michelle), …

Interpreter for all clauses of the form P(X,Y) :- Q(X,Y):

```
interp(P,X,Y) :- rel(P,X,Y).
interp(P,X,Y) :- interp(Q,X,Y), assumeRule(P,Q).
assumeRule(P,Q) :- true   # f(P,Q).   // P(X,Y):-Q(X,Y)
```

Features correspond to *specific* rules

interp(sibling,malia,Z)

rel(Q,malia,Z),
assumeRule(sibling,Q),…

assumeRule(sibling,sister),…

f(sibling,sister)

…

Z=sasha

assumeRule(sibling,mother),…

f(sibling,mother)

…

Z=michelle

Logic program is an *interpreter* for a program containing *all possible rules* from a sublanguage

Features ~ rules. For example:
f(sibling,sister) ~ sibling(X,Y):- sister(X,Y)

DB: rel(sister

p(sibling,malia,Z)

Gradient of *parameters* (feature weights) informs you about what *rules* could be added to the theory…

Interpreter for all clauses of the form P(X,Y) :- Q(X,Y):

interp(P,X,Y) :- rel(P,X,Y).
interp(P,X,Y) :- interp(Q,X,Y), assumeRule(P,Q).
assumeRule(P,Q) :- true # f(P,Q). *// P(X,Y):-Q(X,Y)*

Added rule:
Interp(sibling,X,Y) :- interp(sister,X,Y).

interp(sibling,malia,Z)

rel(Q,malia,Z),
assumeRule(sibling,Q),…

assumeRule(sibling,sister),…

assumeRule(sibling,mother),…

f(sibling,sister)

…

f(sibling,mother)

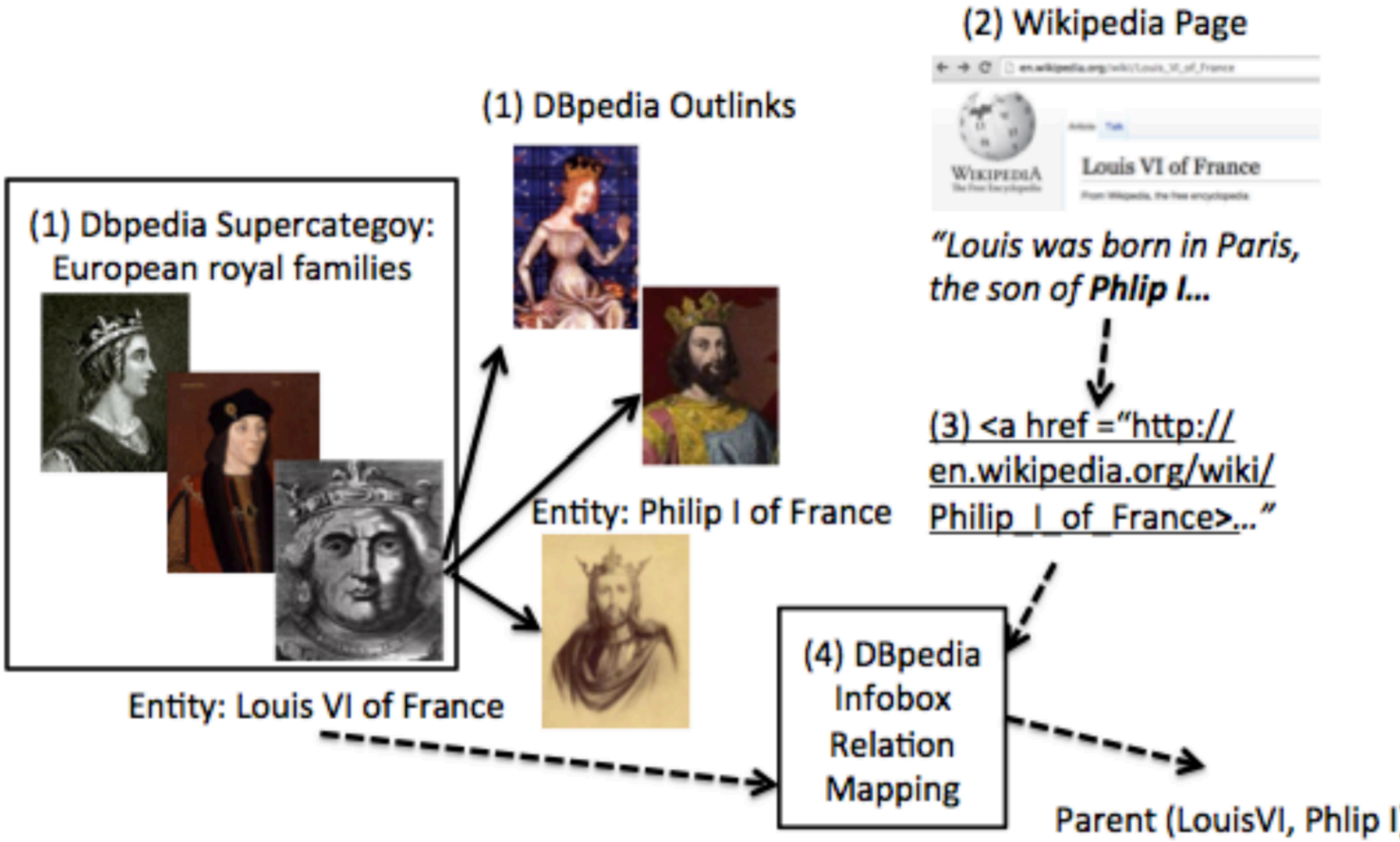
…

Z=sasha

Z=michelle

Joint IE and Structure learning

Data Collection



Joint IE+SL Theory

	Rule template	ProPPR clause
Structure learning		
(a)	P(X,Y) :- R(X,Y)	interp(P,X,Y) :- interp0(R,X,Y),abduce_if(P,R). abduce_if(P,R) :- true # f_if(P,R).
(b)	P(X,Y) :- R(Y,X)	interp(P,X,Y) :- interp0(R,Y,X),abduce_ifInv(P,R). abduce_ifInv(P,R) :- true # f_ifInv(P,R).
(c)	P(X,Y) :- R1(X,Z),R2(Z,Y)	interp(P,X,Y) :- interp0(R1,X,Z),interp0(R2,Z,Y), abduce_chain(P,R1,R2). abduce_chain(P,R1,R2) :- true # f_chain(P,R1,R2).
	base case for SL interpreter	interp0(P,X,Y) :- rel(R,X,Y).
	insertion point for learned rules	interp0(P,X,Y) :- *any rules learned by SL.*
Information extraction		
(d)	R(X,Y) :- link(X,Y,W), indicates(W,R).	interp(R,X,Y) :- link(X,Y,W),abduce_indicates(W,R). abduce_indicates(W,R) :- true #f_ind1(W,R).
(e)	R(X,Y) :- link(X,Y,W1), link(X,Y,W2), indicates(W1,W2,R).	interp(R,X,Y) :- link(X,Y,W1),link(X,Y,W2), abduce_indicates(W1,W2,R). abduce_indicates(W1,W2,R) :- true #f_ind2(W1,W2,R).

Experiments

- Task: KB Completion.
- Three Wikipedia Datasets:

  royal, geo, american.

  67K, 12K, and 43K links respectively.

	10% deleted	50% deleted
ProPPR/SL	79.5	61.9
ProPPR/IE	**81.1**	**70.6**

Results on Royal, similar results on two other InfoBox datasets.

Joint Relation Learning IE in ProPPR

- Experiment

    Combine IE and SL rules

	10% deleted	50% deleted
ProPPR/SL	79.5	61.9
ProPPR/IE	81.1	70.6
ProPPR/Joint IE,SL	**82.8**	**78.6**

Similar results on two other InfoBox datasets

Joint IE and Relation Learning

- Baselines: MLNs (Richardson and Domingos, 2006), Universal Schema (Riedel et al., 2013), IE- and structure-learning-only models.

	Royal				
% missing	10%	20%	30%	40%	50%
Baselines					
MLN	60.8	43.7	44.9	38.8	38.8
Universal Schema	48.2	53.0	52.9	47.3	41.2
SL	79.5	77.2	74.8	65.5	61.9
IE only					
IE (U)	81.3	78.5	76.4	75.7	70.6
IE (U+B)	81.1	78.1	76.2	75.5	70.3
Joint					
SL+IE (U)	82.8	80.9	79.1	77.9	78.6
SL+IE (U+B)	83.4	82.0	80.7	79.7	80.3

Latent Context Invention

Making the classifier more powerful: introduce latent classes (analogous to invented predicates) which can be combined with the context words in the features used by the classifier.

Latent context invention

(f)	R(X,Y) :- latent(L), link(X,Y,W), indicate(W,L,R)	interp(R,X,Y) :- latent(L),link(X,Y,W),abduce_latent(W,L,R). abduce_latent(W,L,R) :- true #f_latent1(W,L,R).
(g)	R(X,Y) :- latent(L1),latent(L2) link(X,Y,W), indicate(W,L1,L2,R)	interp(R,X,Y) :- latent(L1),latent(L2),link(X,Y,W), abduce_latent(W,L1,L2,R). abduce_latent(W,L1,L2,R) :- true #f_latent2(W,L1,L2,R).

Joint IE and Relation Learning

- Task: Knowledge Base Completion.
- Baselines: MLNs (Richardson and Domingos, 2006), Universal Schema (Riedel et al., 2013), IE- and structure-learning-only models.

	Royal					Geo					American				
% missing	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
Baselines															
MLN	60.8	43.7	44.9	38.8	38.8	80.4	79.2	68.1	66.0	68.0	54.0	56.0	51.2	41.0	13.8
Universal Schema	48.2	53.0	52.9	47.3	41.2	82.0	84.0	75.7	77.0	65.2	56.7	51.4	55.9	54.7	51.3
SL	79.5	77.2	74.8	65.5	61.9	83.8	80.4	77.1	72.8	67.2	73.1	70.0	71.3	67.1	61.7
IE only															
IE (U)	81.3	78.5	76.4	75.7	70.6	83.9	79.4	73.1	71.6	65.2	63.4	61.0	60.2	61.4	54.4
IE (U+B)	81.1	78.1	76.2	75.5	70.3	84.0	79.5	73.3	71.6	65.3	64.3	61.2	61.1	62.1	55.7
Joint															
SL+IE (U)	82.8	80.9	79.1	77.9	78.6	89.5	89.4	89.3	88.1	87.6	74.0	73.3	73.7	70.5	68.0
SL+IE (U+B)	83.4	82.0	80.7	79.7	80.3	89.6	89.6	89.5	88.4	87.7	**74.6**	73.5	74.2	70.9	68.4
Joint + Latent															
Joint + Clustering	**83.5**	82.3	81.2	80.2	80.7	89.8	89.6	89.5	88.8	88.4	**74.6**	73.9	74.4	71.5	69.7
Joint + LCI	**83.5**	**82.5**	81.5	80.6	81.1	**89.9**	**89.8**	**89.7**	89.1	89.0	**74.6**	74.1	74.5	72.3	70.3
Joint + LCI + hLCI	**83.5**	**82.5**	**81.7**	**81.0**	**81.3**	**89.9**	89.7	**89.7**	**89.6**	**89.5**	**74.6**	**74.4**	**74.6**	**73.6**	**72.1**

Explaining the Parameters

indicates("mother",parent)
indicates("king",parent)
indicates("spouse",spouse)
indicates("married",spouse)
indicates("succeeded",successor)
indicates("son",successor)

parent(X,Y) :- successor(Y,X)
successor(X,Y) :- parent(Y,X)
spouse(X,Y) :- spouse(Y,X)
parent(X,Y) :- predecessor(X,Y)
successor(Y,X) :- spouse(X,Y)
predecessor(X,Y) :- parent(X,Y)

Discussions

- Comparing to latent variable models, our method is explainable.
- This is multi-instance multi-relation distant supervision with logic.
- This framework allows us to recursively learn relations, and jointly reason with IE clauses.
- Our structure learning method is efficient: according to Kok & Domingos's (2010, ICML), LSM sometimes takes 28 days to learn on a moderate-small dataset, where as our method needs a few minutes on a similar-sized dataset.

Conclusion

- We introduce a probabilistic logic programming method for joint IE and reasoning.
- We briefly show how to incorporate latent classes in first-order logic.
- Our system outperforms state-of-the-art IE systems.

ProPPR Demo

Course Conclusion

1. Basic theories and practices on named entity recognition: supervised, semi-supervised, and unsupervsed.

2. Recent advances in relation extraction:
   a. distant supervision
   b. latent variable models

3. Scalable IE and reasoning with first-order logics.

Acknowledgement

- CIPS Executives
- Peking University
- General Chair: Prof. Le Sun
- PC Chair: Prof. Heng Ji
- Org. Chairs: Profs. Wang, Zhao, and Sui.
- Volunteers
- Participants

Ask Me Anything!

yww@cs.cmu.edu