Structured Linear Models

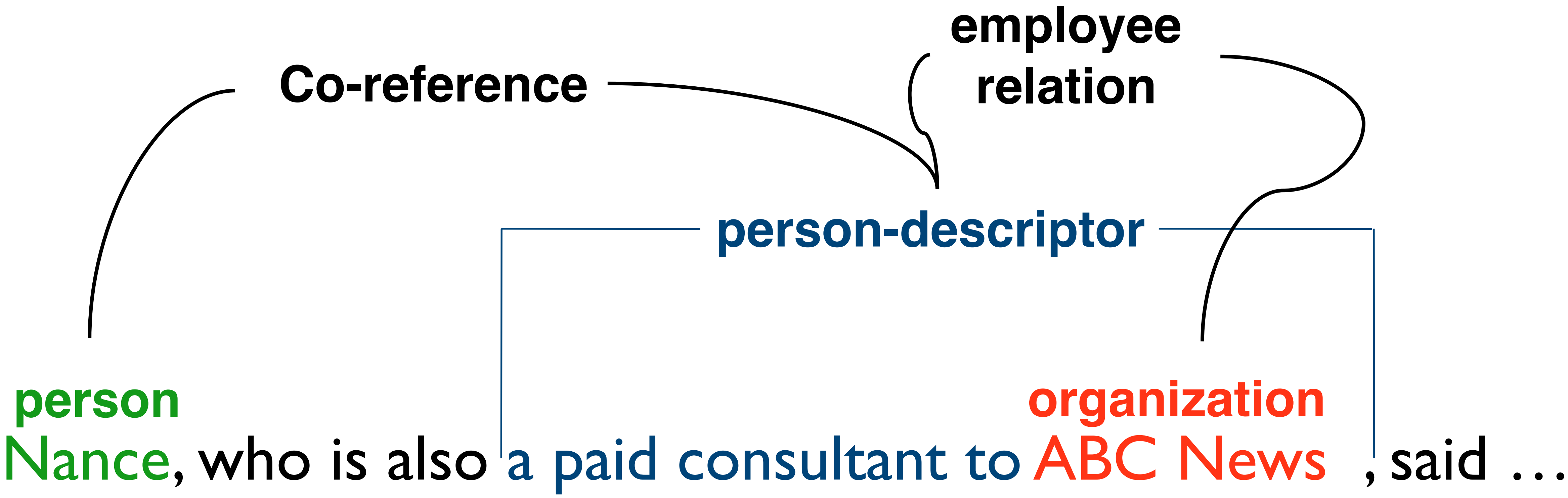
Fernando Pereira
with
Koby Crammer, Ryan McDonald, Fei Sha,
ParthaTalukdar
Department of Computer and Information Science
University of Pennsylvania

Thanks: Thorsten Brants, John Lafferty, Mark Liberman,Andrew McCallum, Pete White,
NSF EIA0205456, EIA 0205448, IIS 0428193, and DARPA/SRI contract NBCHD030010

Goals

- *What*: link document and structured databases
- *How*: *information extraction*:
  - *Tag* entity and relation mentions in text
  - Map (“normalize”) the mentions to database entities and relations
- *Example*: biomedical databases

Information Extraction



Co-reference

employee relation

person-descriptor

person

organization

Nance, who is also a paid consultant to ABC News , said …

Biomedical Examples

- Gene/protein mentions:

  `In the absence of` **`MHC class II`**`,` **`purified soluble D10 TCR`** `bound to` **`Staphylococcus aureus enterotoxin C2`** `with an association rate of 1.`

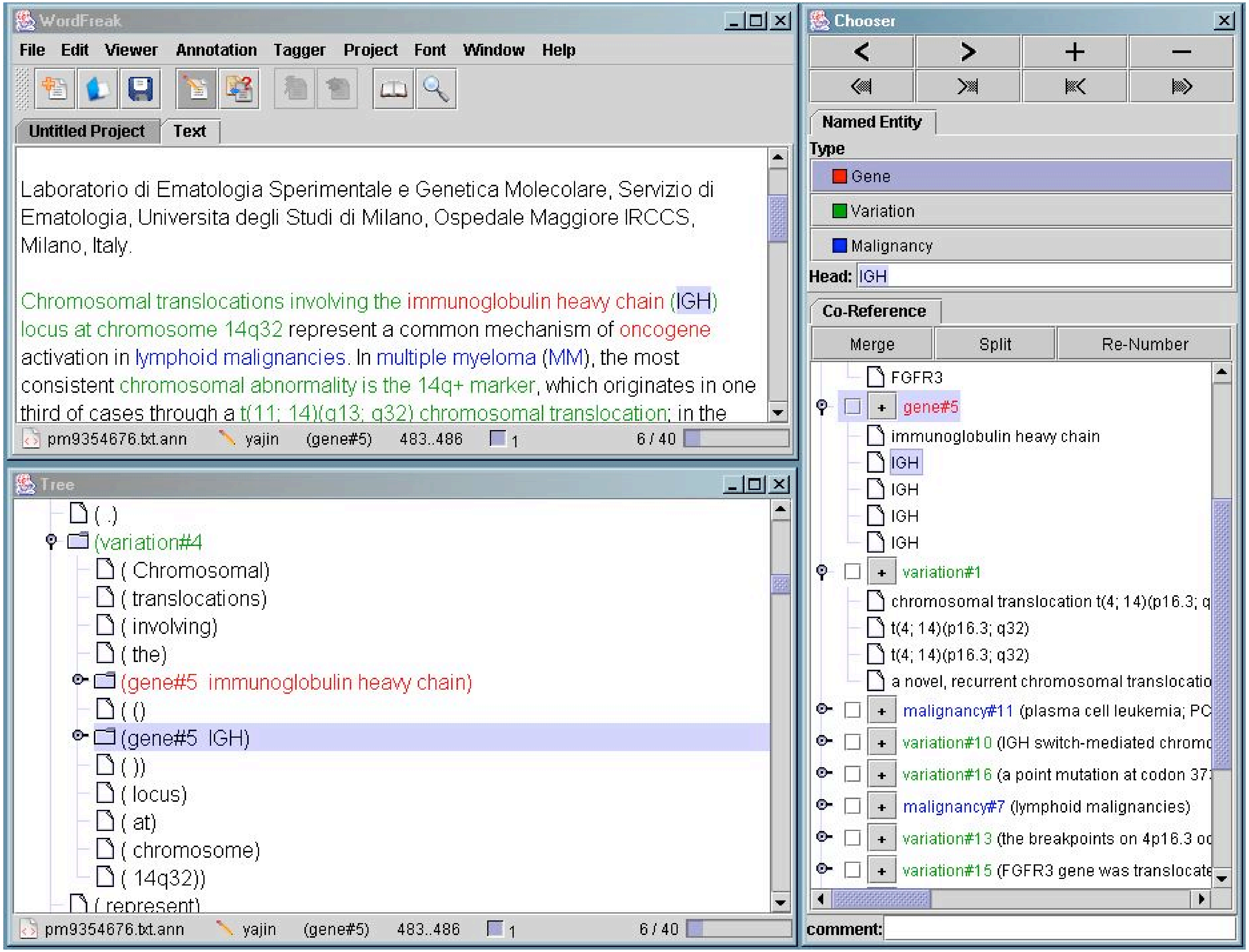
- Variation events: type, location, and state change

  `One ER showed a` **`G`** `to` **`T`** **`point mutation`** `in the` **`second position of codon 12`**

Approach

- Develop text annotation guidelines
- Annotate initial training documents
- Train machine learning algorithms for extraction
- Automatically label more documents and correct (active annotation)

Annotation Tool



Laboratorio di Ematologia Sperimentale e Genetica Molecolare, Servizio di Ematologia, Universita degli Studi di Milano, Ospedale Maggiore IRCCS, Milano, Italy.

Chromosomal translocations involving the immunoglobulin heavy chain (IGH) locus at chromosome 14q32 represent a common mechanism of oncogene activation in lymphoid malignancies. In multiple myeloma (MM), the most consistent chromosomal abnormality is the 14q+ marker, which originates in one third of cases through a t(11; 14)(q13; q32) chromosomal translocation; in the

pm9354676.txt.ann yajin (gene#5) 483..486 1 6 / 40

(.)
(variation#4
(Chromosomal)
(translocations)
(involving)
(the)
(gene#5 immunoglobulin heavy chain)
(()
(gene#5 IGH)
( ))
(locus)
(at)
(chromosome)
(14q32))
(represent)

pm9354676.txt.ann yajin (gene#5) 483..486 1 6 / 40

Type:
- 🟥 Gene
- 🟩 Variation
- 🟦 Malignancy

Head: IGH

- FGFR3
- □ + gene#5
  - immunoglobulin heavy chain
  - IGH
  - IGH
  - IGH
  - IGH
- □ + variation#1
  - chromosomal translocation t(4; 14)(p16.3; q
  - t(4; 14)(p16.3; q32)
  - t(4; 14)(p16.3; q32)
  - a novel, recurrent chromosomal translocatio
- □ + malignancy#11 (plasma cell leukemia; PC
- □ + variation#10 (IGH switch-mediated chromo
- □ + variation#16 (a point mutation at codon 37
- □ + malignancy#7 (lymphoid malignancies)
- □ + variation#13 (the breakpoints on 4p16.3 oc
- □ + variation#15 (FGFR3 gene was translocate

comment:

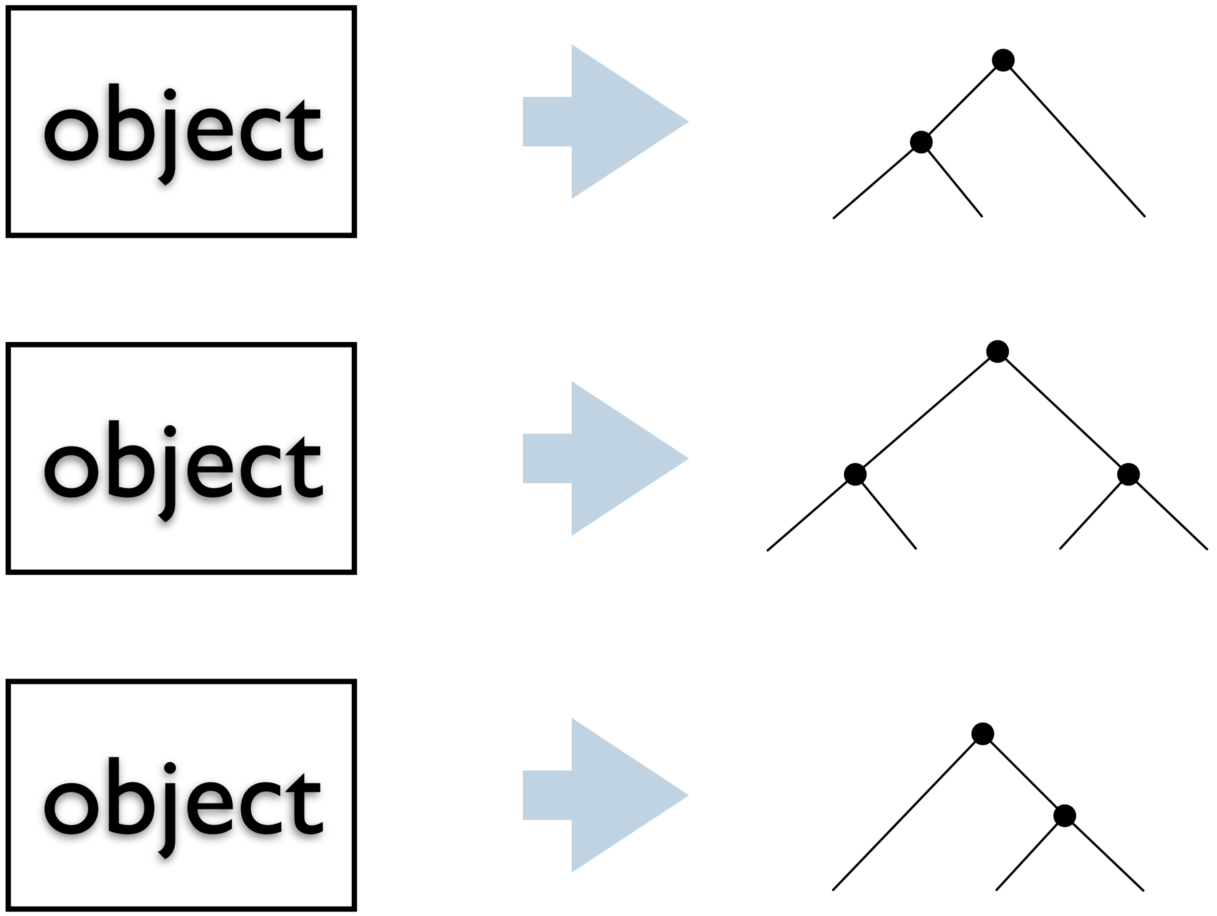
Penn
University of Pennsylvania

Analyzing Text

- Segmentation
  - units (paragraphs, sentences)
  - layout (lists, FAQs,...)
- Tagging
  - part of speech
  - sense
- *Information extraction*
- Parsing

Structured Classification

- Learn mapping from objects (documents, sentences,...) to structures



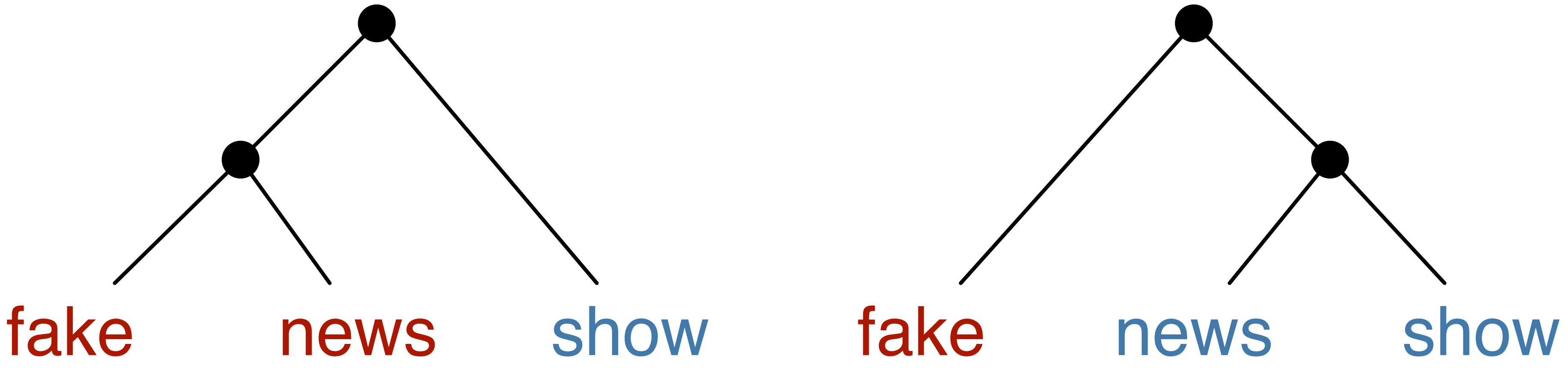
object

object

object

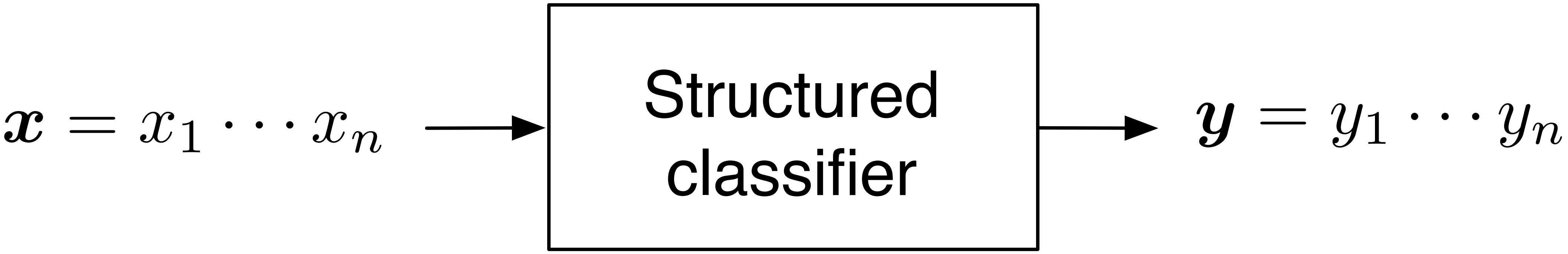
Challenges

- Interacting decisions



- Many types of sequence features
- Computing an answer is relatively costly

Analysis by Tagging



$\boldsymbol{x} = x_1 \cdots x_n$

Structured classifier

$\boldsymbol{y} = y_1 \cdots y_n$

- Labels give the role of corresponding inputs
    - *Information extraction*
    - Part-of-speech tagging
    - Shallow parsing
    - Other segmentation/labeling tasks (speech, genomic sequences,...)

Segmentation as Tagging

Rockwell International Corp. ’s Tulsa unit said

B I I B I I O

it signed a tentative agreement extending its contract

B O B I I O B I

with Boeing Co. to provide structural parts

O B I O O B I

for Boeing ’s 747 jetliners

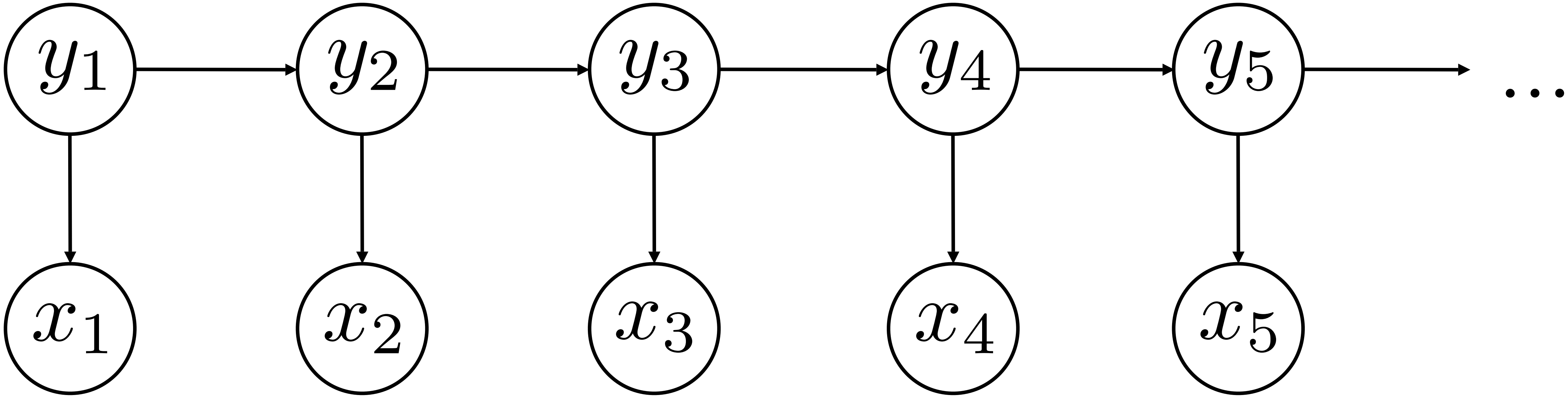
O B B I I

Traditional Approaches

- *Generative modeling*: probabilistic generators of sequence-structure pairs
  - HMMs, probabilistic CFGs
  - Hard to model non-independent features
- *Sequential classification*: decompose structure assignment into a sequence of structural decisions
  - Cannot trade-off decisions at different locations: *label-bias* problem

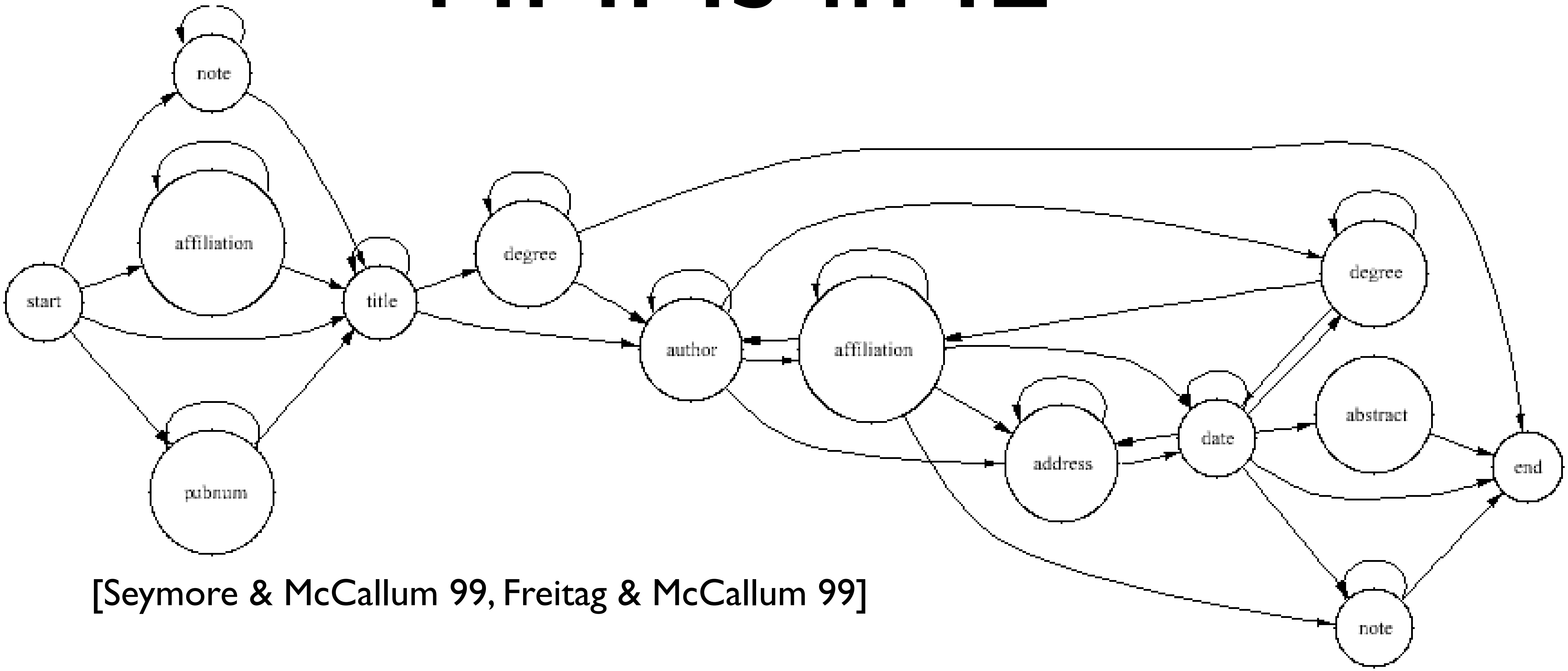
Hidden Markov Model

- Instances: symbol sequences
- Labels: state sequences



$$p(\boldsymbol{x}, \boldsymbol{y}) = p(y_1)p(x_1|y_1)\prod_{i=2}^{n} p(y_i|y_{i-1})p(x_i|y_i)$$

HMMs in IE



[Seymore & McCallum 99, Freitag & McCallum 99]

- *Inputs* x: words
- *States* y: fields to extract

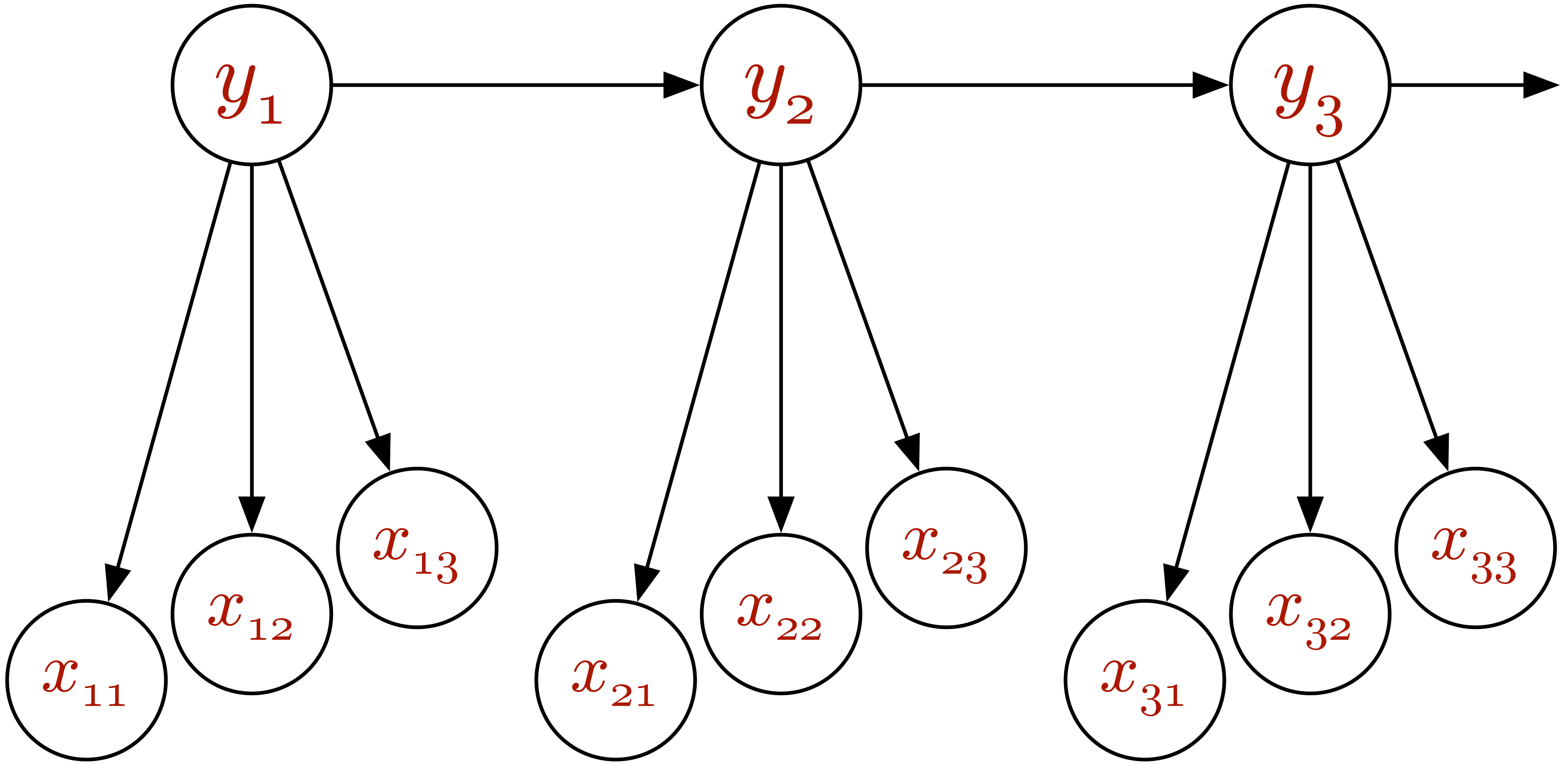
$$p(\boldsymbol{x}, \boldsymbol{y}) = \prod_i p(y_i | y_{i-1}) p(x_i | y_i)$$

Problems with HMMs

- Applications need richer input representation

Word features	*Formatting features*
word identity	centered
capitalization	indentation
ends in "-tion"	white space ratio
word in word list	begins with number
word font	ends with "?"

Generating Multiple Features



- Relax conditional independence of features on labels ⇒ *intractability*

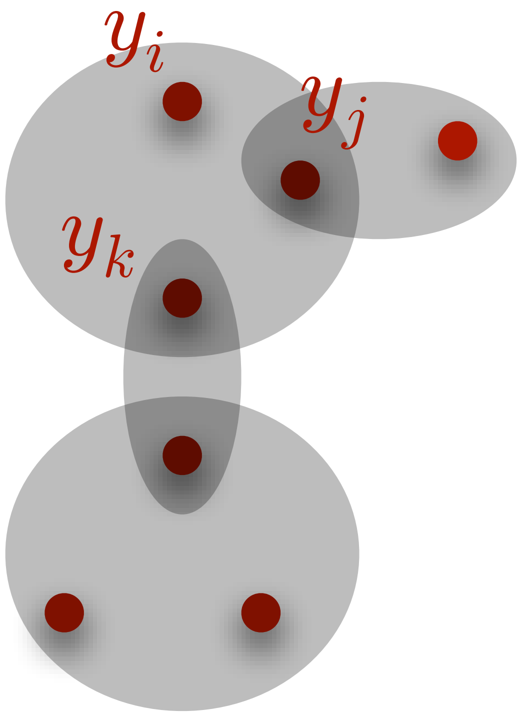
Structured Linear Models

- Generalize linear classification

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} \boldsymbol{w} \cdot \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y})$$

- Features based on local domains

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) &= \textstyle\sum_{C \in \mathcal{C}(\boldsymbol{x})} \boldsymbol{f}_C(\boldsymbol{x}, \boldsymbol{y}) \\ \boldsymbol{f}_C(\boldsymbol{x}, \boldsymbol{y}) &= \boldsymbol{f}_C(\boldsymbol{x}, \boldsymbol{y}_C) \end{aligned}$$



- Efficient Viterbi decoding for tree-structured interactions

Learning

- Prior knowledge
  - local domains $\mathcal{C}(x)$
  - local feature functions f_C
- Adjust $\boldsymbol{w}$ to optimize objective function on some training data

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \underbrace{\lambda \left\|\boldsymbol{w}\right\|^2}_{\text{regularizer}} + \sum_i \underbrace{L(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w})}_{\text{loss}}$$

Margin

- Score advantage between correct and candidate classifications

$$m(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{y}'; \boldsymbol{w}) = \boldsymbol{w} \cdot F(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{w} \cdot F(\boldsymbol{x}, \boldsymbol{y}')$$

Losses

- Log loss ⇒ maximize probability of correct output

$$L(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = \log \sum_{\boldsymbol{y}'} e^{-m(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{y}'; \boldsymbol{w})}$$

- Hamming loss ⇒ minimize distance-adjusted misclassification

$$L(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = \max_{\boldsymbol{y}'} \left[d(\boldsymbol{y}, \boldsymbol{y}') - m(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{y}'; \boldsymbol{w}) \right]_+$$

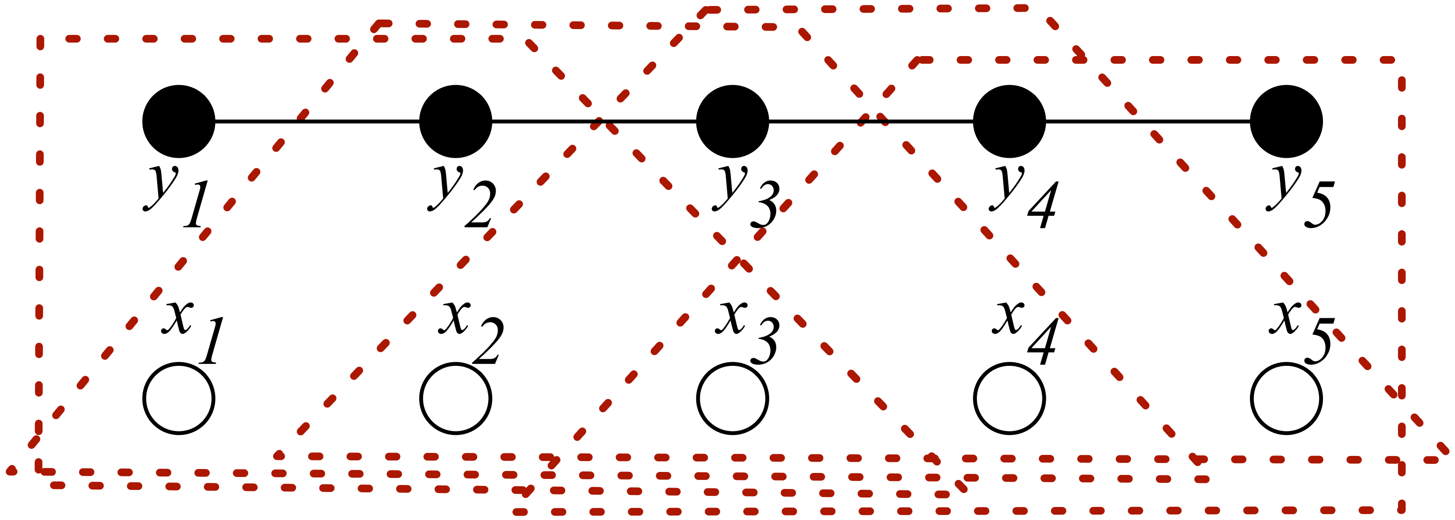
- Search over $\boldsymbol{y}'$: dynamic programming on "good" graphs

Why?

- Combine the best of generative and classification models:
  - Trade off labeling decisions at different positions
  - Allow overlapping features
- Modular
  - factored scoring
  - loss function

Probabilistic Version

- Sequence *conditional random fields (CRFs)*

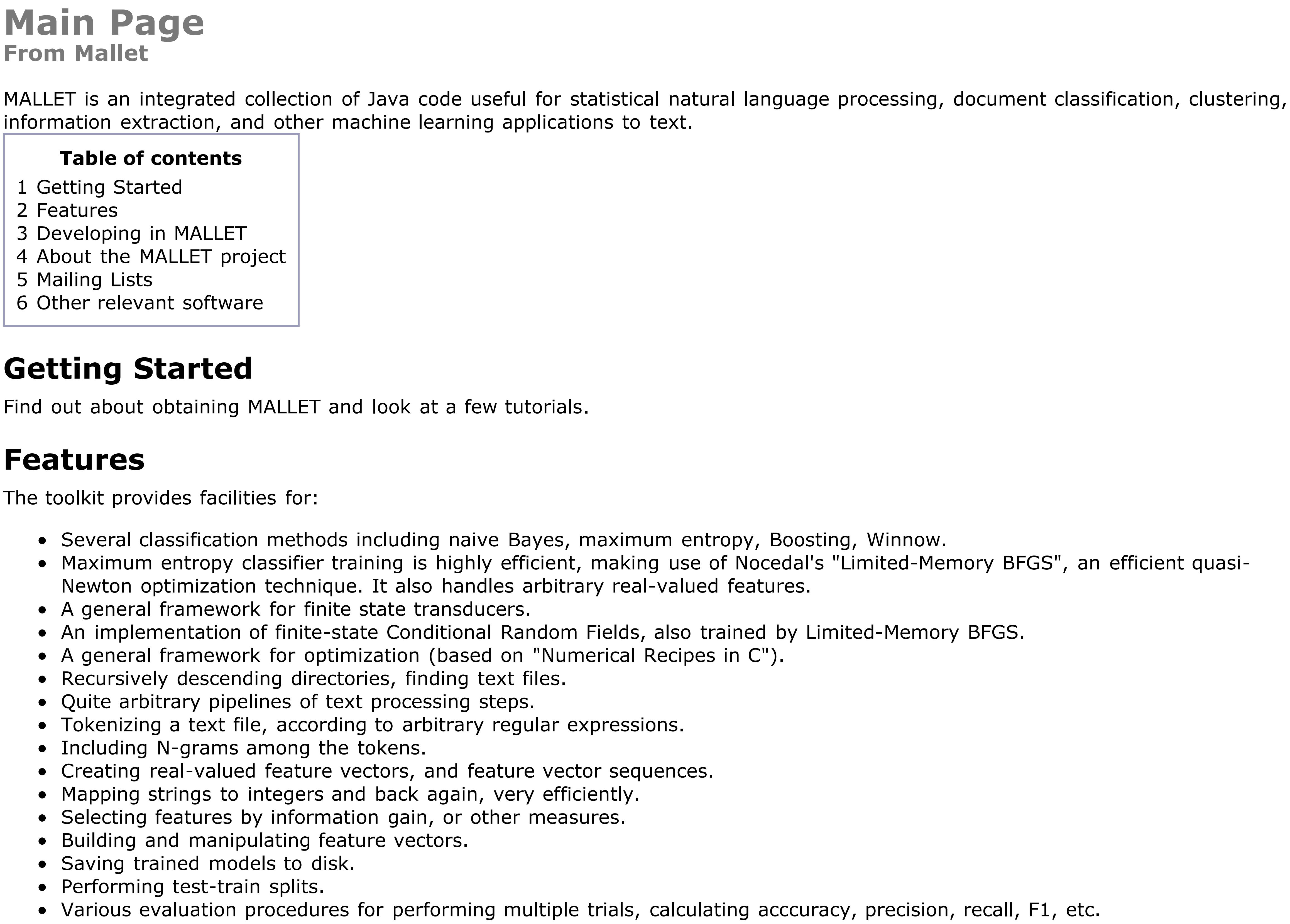


$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{w}) &= \frac{\exp \boldsymbol{w}\cdot\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})}{Z(\boldsymbol{x};\boldsymbol{w})} \\
Z(\boldsymbol{x};\boldsymbol{w}) &= \textstyle\sum_{\boldsymbol{y}} \exp \boldsymbol{w}\cdot\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) \\
\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) &= \textstyle\sum_i \boldsymbol{f}_i(\boldsymbol{x},\boldsymbol{y}) \\
\boldsymbol{f}_i(\boldsymbol{x},\boldsymbol{y}) &= \boldsymbol{f}_i(y_{i-1},y_i,\boldsymbol{x})
\end{aligned}
$$

- Training criterion: log loss

Features

- Conjunctions of
  - Label configuration
  - Input properties
    - Term identity
    - Membership in term list
    - Orthographic patterns
    - Conjunctions of the these for current and surrounding words
      - *Feature induction*: generate only those conjunctions that help prediction



Main Page

From Mallet

MALLET is an integrated collection of Java code useful for statistical natural language processing, document classification, clustering, information extraction, and other machine learning applications to text.

Table of contents

Getting Started

Find out about obtaining MALLET and look at a few tutorials.

Features

The toolkit provides facilities for:

- Several classification methods including naive Bayes, maximum entropy, Boosting, Winnow.
- Maximum entropy classifier training is highly efficient, making use of Nocedal's "Limited-Memory BFGS", an efficient quasi-Newton optimization technique. It also handles arbitrary real-valued features.
- A general framework for finite state transducers.
- An implementation of finite-state Conditional Random Fields, also trained by Limited-Memory BFGS.
- A general framework for optimization (based on "Numerical Recipes in C").
- Recursively descending directories, finding text files.
- Quite arbitrary pipelines of text processing steps.
- Tokenizing a text file, according to arbitrary regular expressions.
- Including N-grams among the tokens.
- Creating real-valued feature vectors, and feature vector sequences.
- Mapping strings to integers and back again, very efficiently.
- Selecting features by information gain, or other measures.
- Building and manipulating feature vectors.
- Saving trained models to disk.
- Performing test-train splits.
- Various evaluation procedures for performing multiple trials, calculating acccuracy, precision, recall, F1, etc.

http://mallet.cs.umass.edu/index.php/Main_Page

Evaluation

- *Precision* P: what proportion of predicted entities are correct
- *Recall* R: what proportion of correct entities are predicted
- F_1 *measure*:

$$\frac{2PR}{P+R}$$

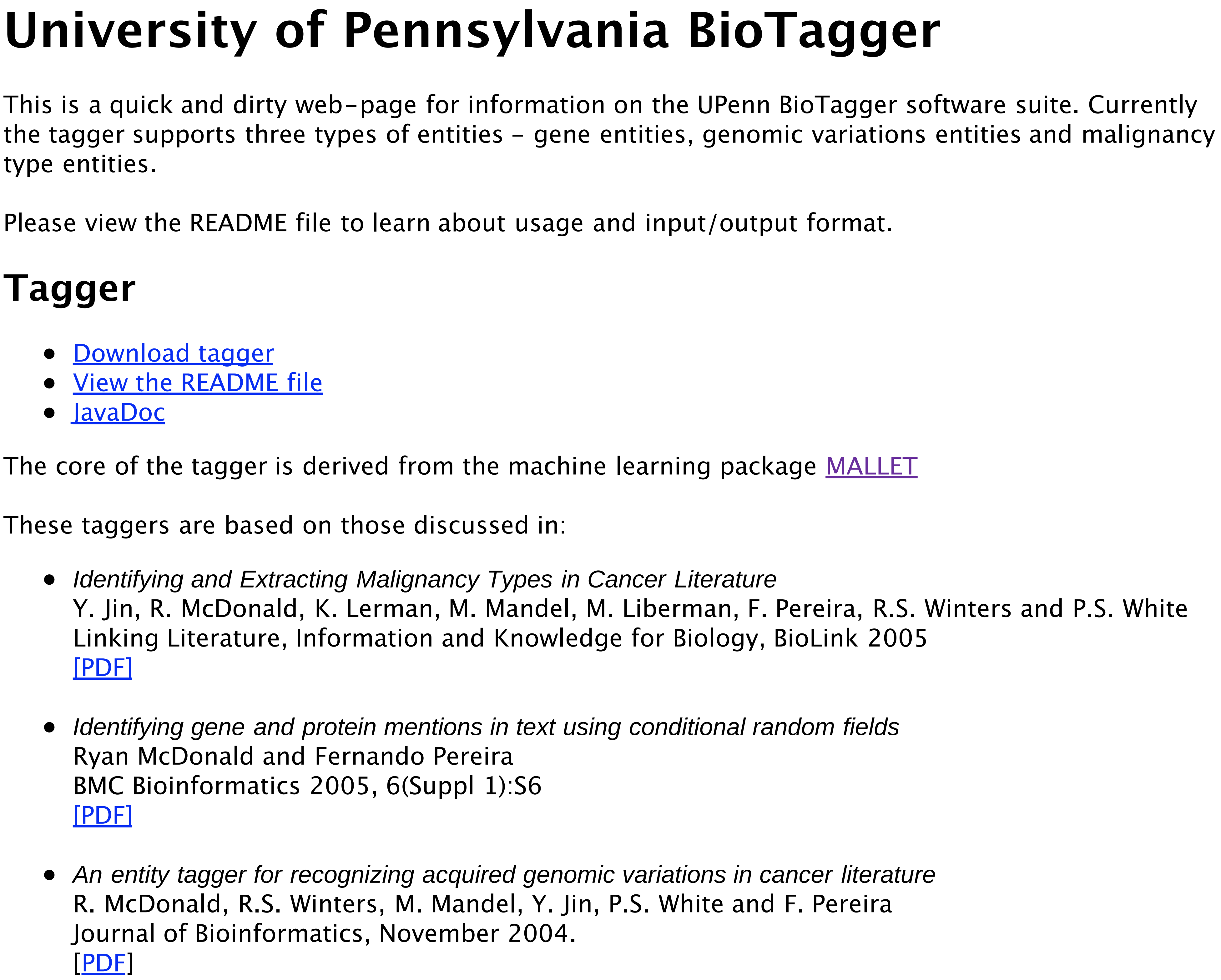
Gene/protein results

		Precision	Recall	F_1
AbGene		63	65	64
CRF	words + spelling	83	77.3	80.1
	(non-)gene tokens + rare trigrams	86.4	78.7	82.4

- Exact match
- AbGene: Brill-style POS and gene tagger, post-processor

Variation Results

	Precision	Recall	F
Type	0.80	0.72	0.76
Location	0.85	0.73	0.79
State	0.90	0.80	0.85



University of Pennsylvania BioTagger

This is a quick and dirty web-page for information on the UPenn BioTagger software suite. Currently the tagger supports three types of entities - gene entities, genomic variations entities and malignancy type entities.

Please view the README file to learn about usage and input/output format.

Tagger

- Download tagger
- View the README file
- JavaDoc

The core of the tagger is derived from the machine learning package MALLET

These taggers are based on those discussed in:

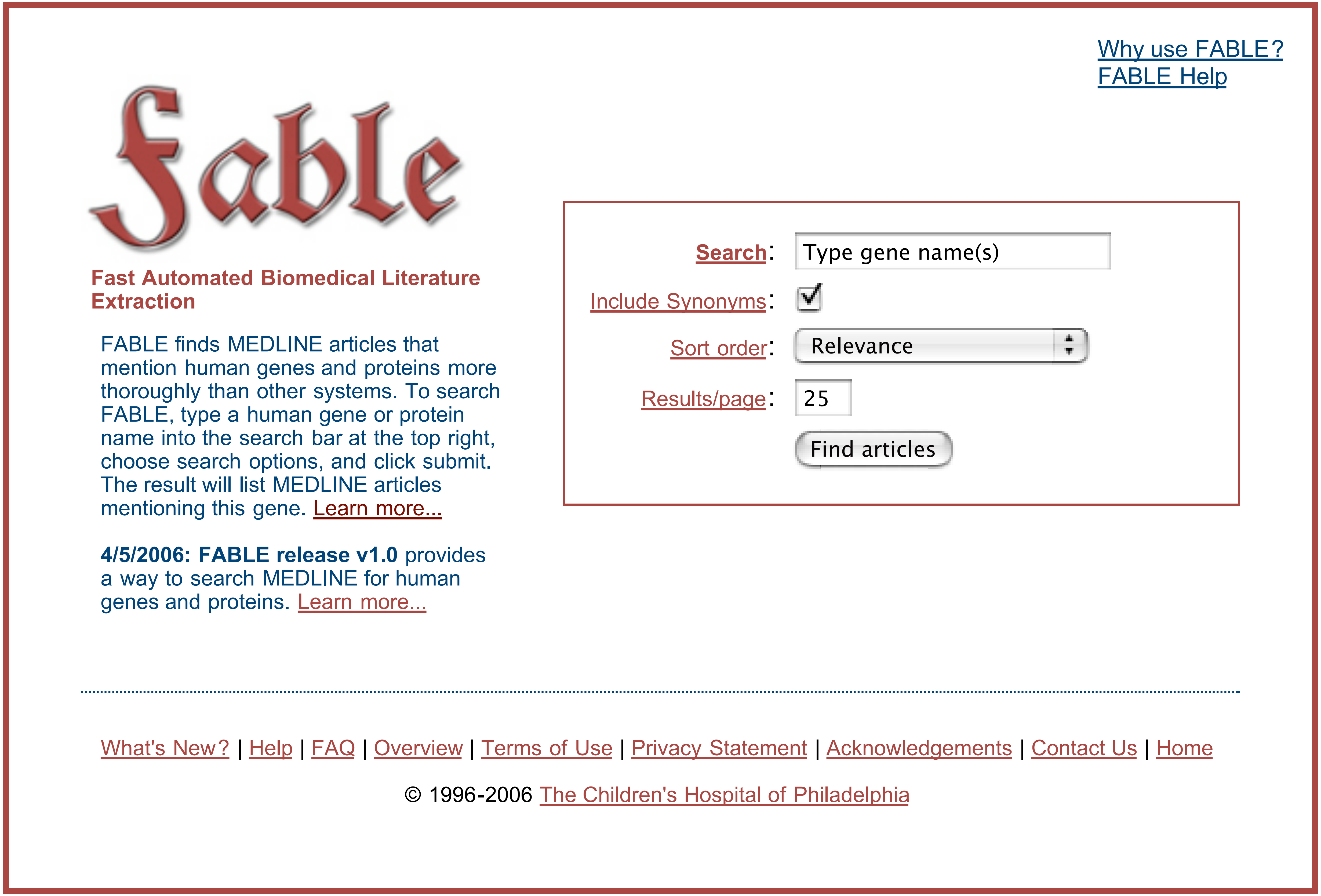
- *Identifying and Extracting Malignancy Types in Cancer Literature*
  Y. Jin, R. McDonald, K. Lerman, M. Mandel, M. Liberman, F. Pereira, R.S. Winters and P.S. White
  Linking Literature, Information and Knowledge for Biology, BioLink 2005
  [PDF]

- *Identifying gene and protein mentions in text using conditional random fields*
  Ryan McDonald and Fernando Pereira
  BMC Bioinformatics 2005, 6(Suppl 1):S6
  [PDF]

- *An entity tagger for recognizing acquired genomic variations in cancer literature*
  R. McDonald, R.S. Winters, M. Mandel, Y. Jin, P.S. White and F. Pereira
  Journal of Bioinformatics, November 2004.
  [PDF]

Programming Credits: Kevin Lerman, Yang Jin, Eric Pancoast and Ryan McDonald.
Questions: ryantm at cis dot upenn dot edu

http://www.cis.upenn.edu/~ryantm/software/BioTagger/



Fast Automated Biomedical Literature Extraction

FABLE finds MEDLINE articles that mention human genes and proteins more thoroughly than other systems. To search FABLE, type a human gene or protein name into the search bar at the top right, choose search options, and click submit. The result will list MEDLINE articles mentioning this gene. Learn more...

4/5/2006: FABLE release v1.0 provides a way to search MEDLINE for human genes and proteins. Learn more...

Search: Type gene name(s)

Include Synonyms: ☑

Sort order: Relevance

Results/page: 25

What's New? | Help | FAQ | Overview | Terms of Use | Privacy Statement | Acknowledgements | Contact Us | Home

© 1996-2006 The Children's Hospital of Philadelphia

http://fable.chop.edu/index.jsp

Penn
University of Pennsylvania

Technical challenges

- Very large number of features:
  - 820,000 at least once on training set
  - 3,800,000 input tests true at least once
  - most features are term-based
- Slow training
  - *online methods*
  - stochastic gradient
- Overfitting
  - *improve term lists*
  - *large margin methods*

Alternative: online training

- Process one training instance at a time
- Very simple
- Predictable runtime, small memory
- Adaptable to different loss functions
- Basic idea:

```
$\boldsymbol{w} = \mathbf{0}$
for $t = 1, \ldots, T$ :
    for $i = 1, \ldots, N$ :
        classify $\boldsymbol{x}_i$ incurring loss $l$
        update $\boldsymbol{w}$ to reduce $l$
```

Online maximum margin

(MIRA)

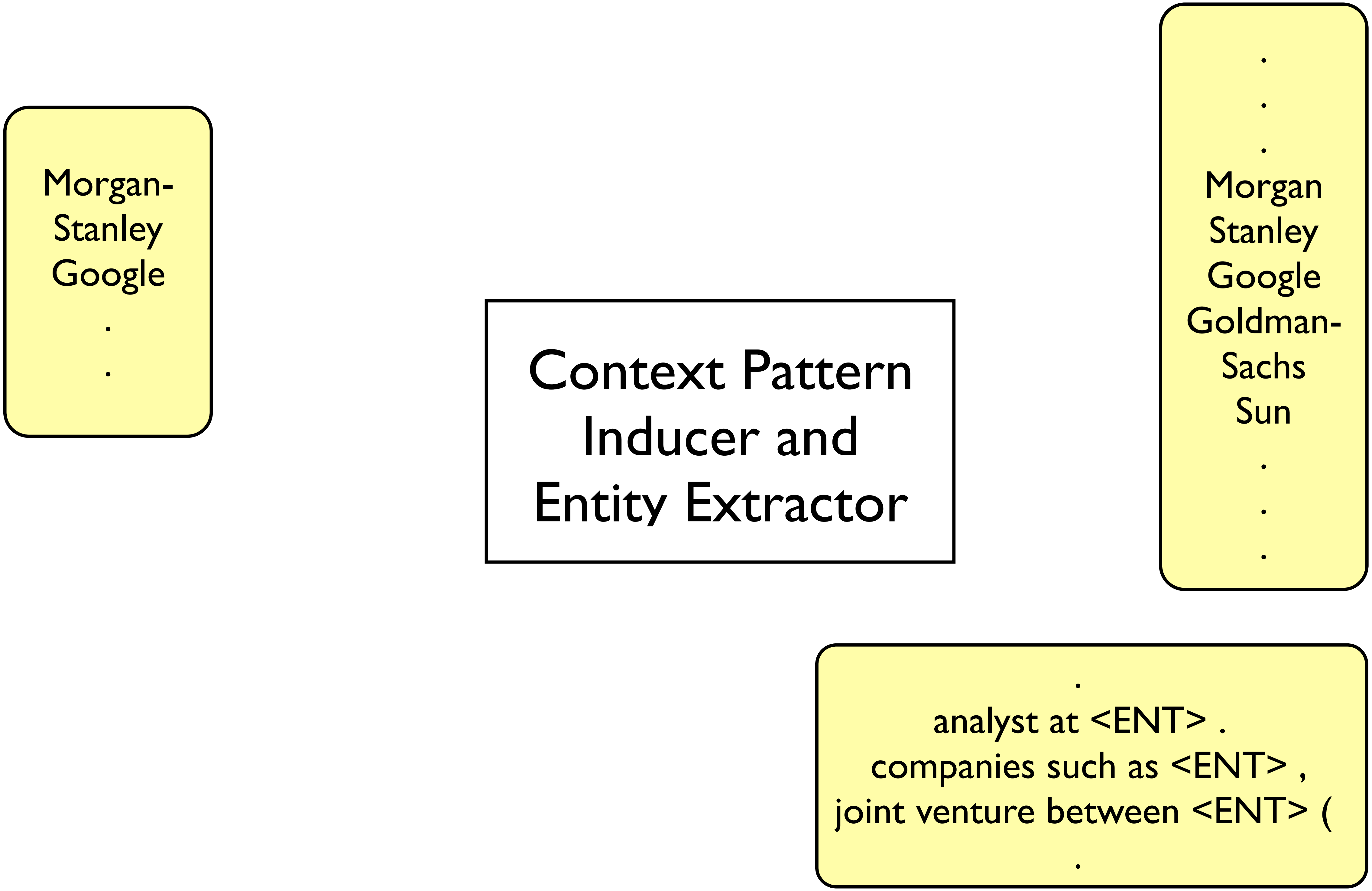
- Project onto subspace where the correct structure scores "far enough" above all incorrect ones

```
$\boldsymbol{w} = \mathbf{0}$
for $t = 1, \ldots, T$ :
   for $i = 1, \ldots, N$ :
      $\boldsymbol{w} \leftarrow \arg\min_{\boldsymbol{w}'} \frac{1}{2} \|\boldsymbol{w}' - \boldsymbol{w}\|^2$
         s.t. $\forall \boldsymbol{y} : \boldsymbol{w}' \cdot \boldsymbol{F}(\boldsymbol{x}_i, \boldsymbol{y}_i) - \boldsymbol{w}' \cdot \boldsymbol{F}(\boldsymbol{x}_i, \boldsymbol{y}) \geq d(\boldsymbol{y}_i, \boldsymbol{y})$
```

- Exponentially many ys: select best k instead

- Related to Hamming loss

Lists and Unlabeled Text



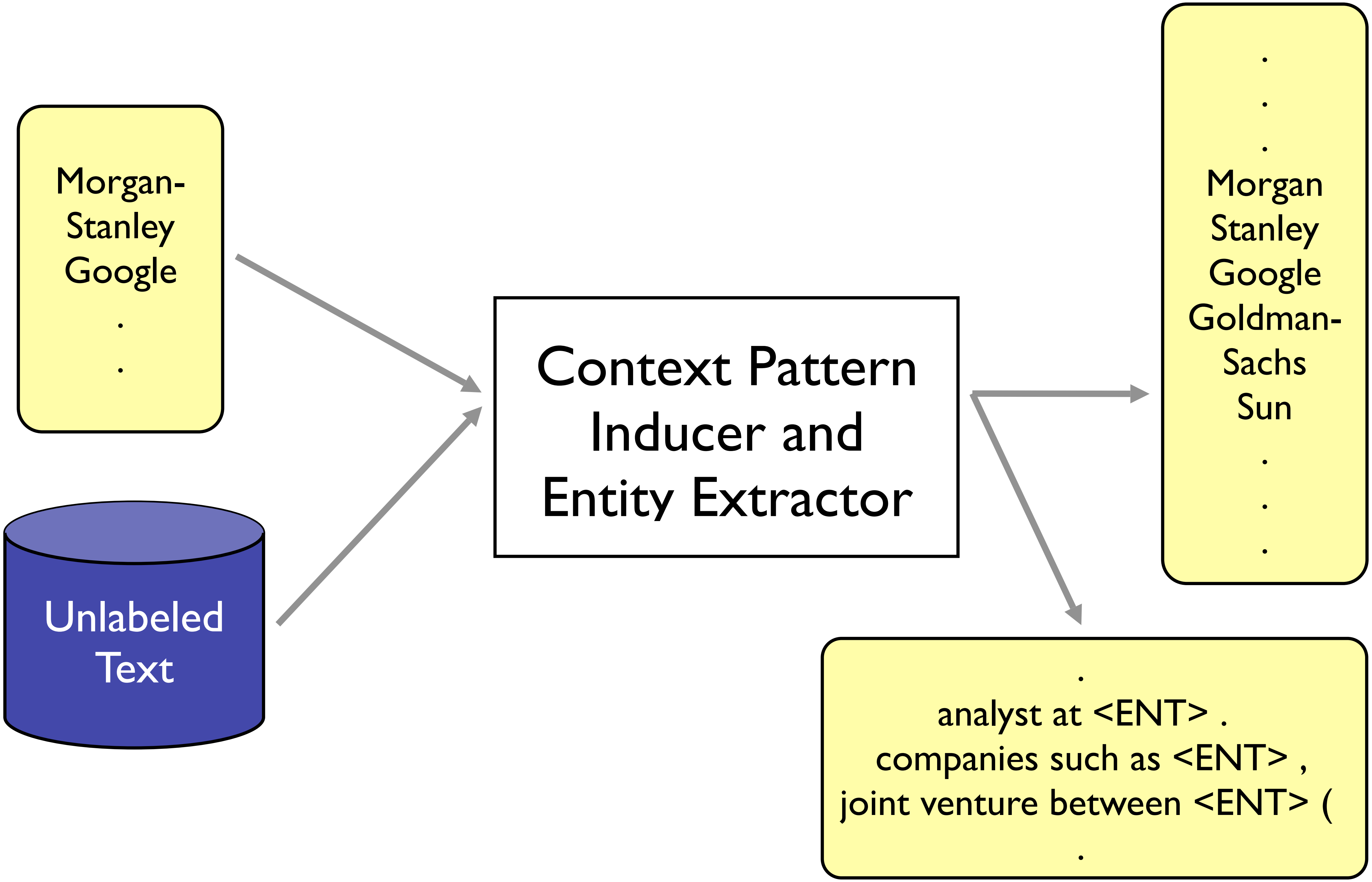
Morgan-
Stanley
Google
.
.

Context Pattern
Inducer and
Entity Extractor

.
.
.
Morgan
Stanley
Google
Goldman-
Sachs
Sun
.
.
.

.
analyst at <ENT> .
companies such as <ENT> ,
joint venture between <ENT> (
.

Lists and Unlabeled Text



Morgan-Stanley
Google
.
.

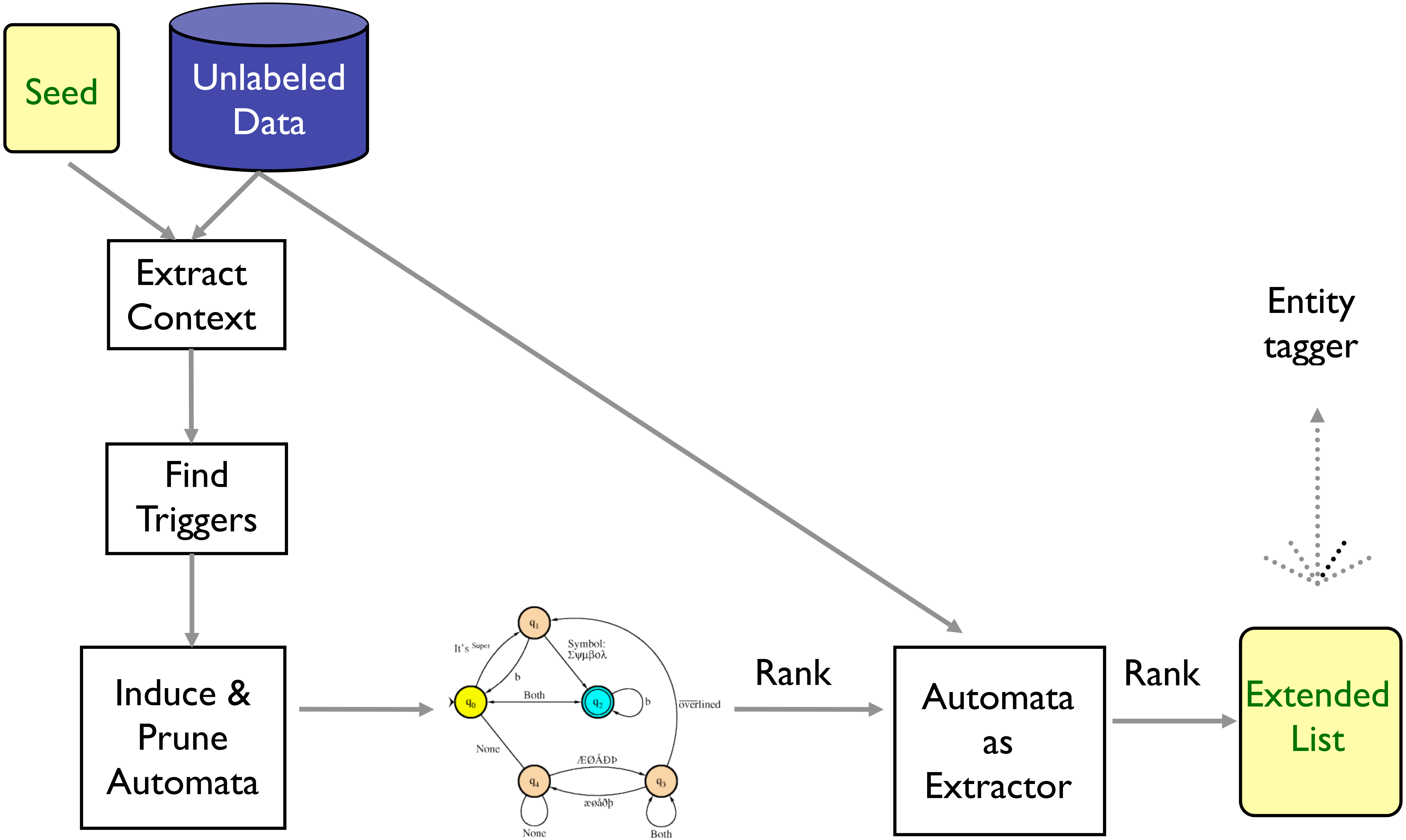
Unlabeled Text

Context Pattern Inducer and Entity Extractor

.
.
.
Morgan Stanley
Google
Goldman-Sachs
Sun
.
.
.

.
analyst at <ENT> .
companies such as <ENT> ,
joint venture between <ENT> (
.

Pattern Induction



Seed

Unlabeled Data

Extract Context

Find Triggers

Induce & Prune Automata

Rank

Automata as Extractor

Rank

Extended List

Entity tagger

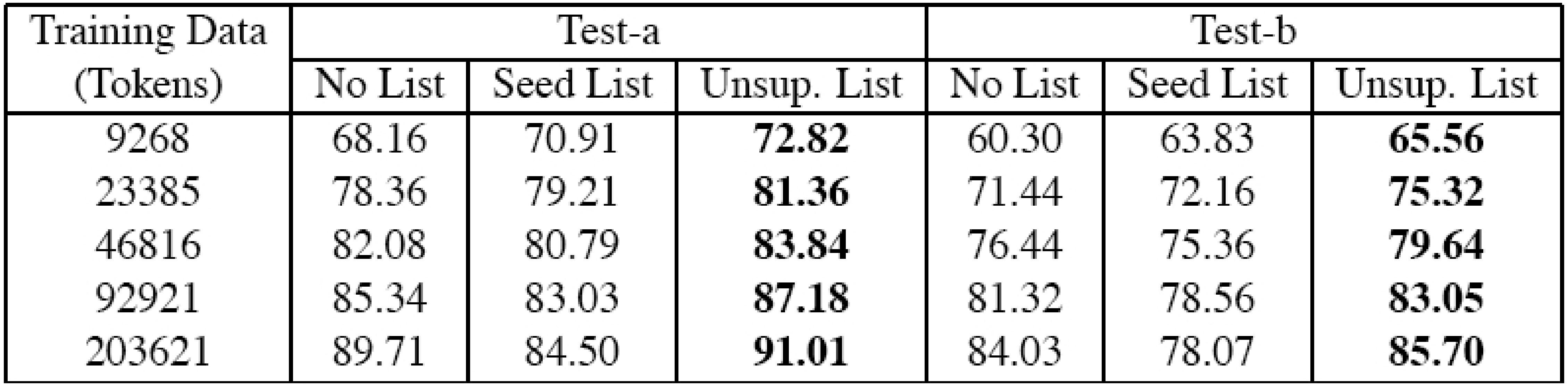
Person Names

compatriot -*ENT-* .
compatriot -*ENT-* in
Rep. -*ENT-* ,
Actor -*ENT-* is
Sir -*ENT-* ,
Actor -*ENT-* ,
Tiger Woods , -*ENT-* and
movie starring -*ENT-* .
compatriot -*ENT-* and
movie starring -*ENT-* and

Tiger Woods
Andre Agassi
Lleyton Hewitt
Ernie Els
Serena Williams
Andy Roddick
Retief Goosen
Vijay Singh
Jennifer Capriati
Roger Federer
…

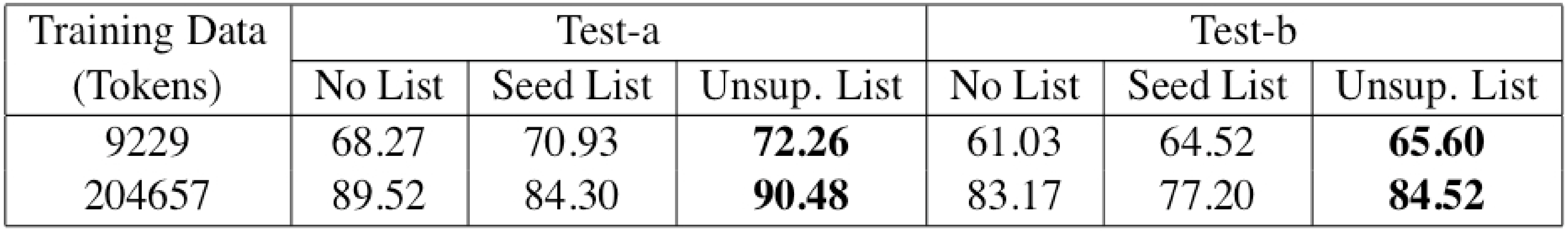
Improving CRF Tagger

PER, LOC, ORG



Training Data (Tokens)	Test-a No List	 Seed List	 Unsup. List	Test-b No List	 Seed List	 Unsup. List
9268	68.16	70.91	**72.82**	60.30	63.83	**65.56**
23385	78.36	79.21	**81.36**	71.44	72.16	**75.32**
46816	82.08	80.79	**83.84**	76.44	75.36	**79.64**
92921	85.34	83.03	**87.18**	81.32	78.56	**83.05**
203621	89.71	84.50	**91.01**	84.03	78.07	**85.70**

PER, LOC, ORG, MISC



Training Data (Tokens)	Test-a No List	 Seed List	 Unsup. List	Test-b No List	 Seed List	 Unsup. List
9229	68.27	70.93	**72.26**	61.03	64.52	**65.60**
204657	89.52	84.30	**90.48**	83.17	77.20	**84.52**

Test Data Sizes: Test-a 51362 tokens, Test-b 46435 tokens

Extensions

- Reducing training data requirements
  - *Pattern induction*
  - Unsupervised domain adaptation for linear models: *structural correspondence learning*
- Deeper analysis
  - Syntactic features
    - Structured linear models for *dependency parsing*
- Joint entity-relation extraction
  - Computational challenges in inference and learning