A Context Pattern Induction Method for Named Entity Extraction

Partha Pratim Talukdar

Computer & Information Science Department

University of Pennsylvania, Philadelphia

partha@cis.upenn.edu

Joint work with **Thorsten Brants** (Google), **Mark Liberman** (Penn) and **Fernando Pereira** (Penn).

Named Entity Extraction

Recognition and classification of entity names e.g. people names, organization names, place names etc.

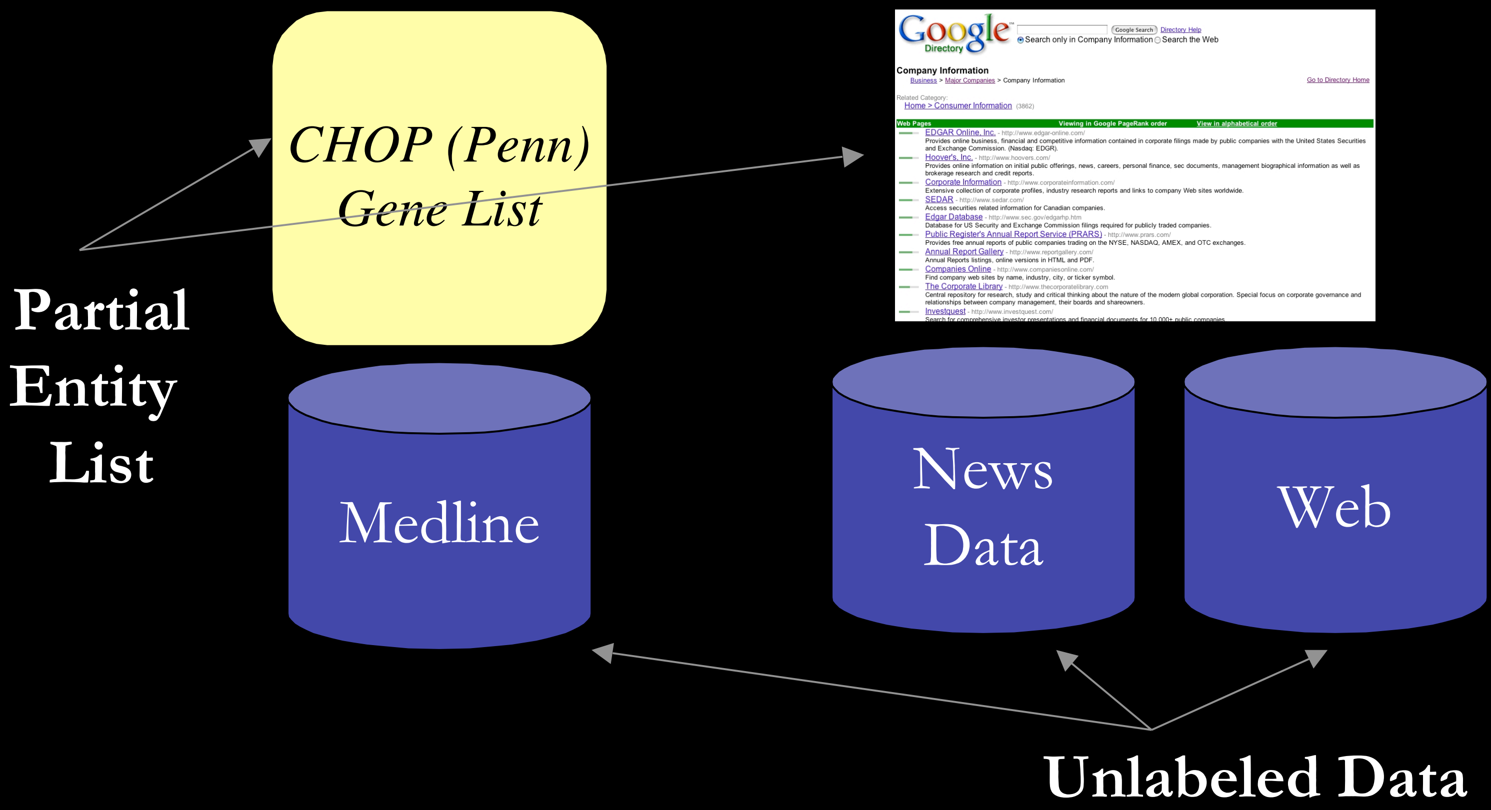
We have identified a transcriptional repressor , Nrg1, in a genetic screen designed to reveal negative factors involved in the expression of STA1.

↓

*We have identified a transcriptional repressor , **Nrg1**, in a genetic screen designed to reveal negative factors involved in the expression of **STA1**.*

Penn
UNIVERSITY of PENNSYLVANIA

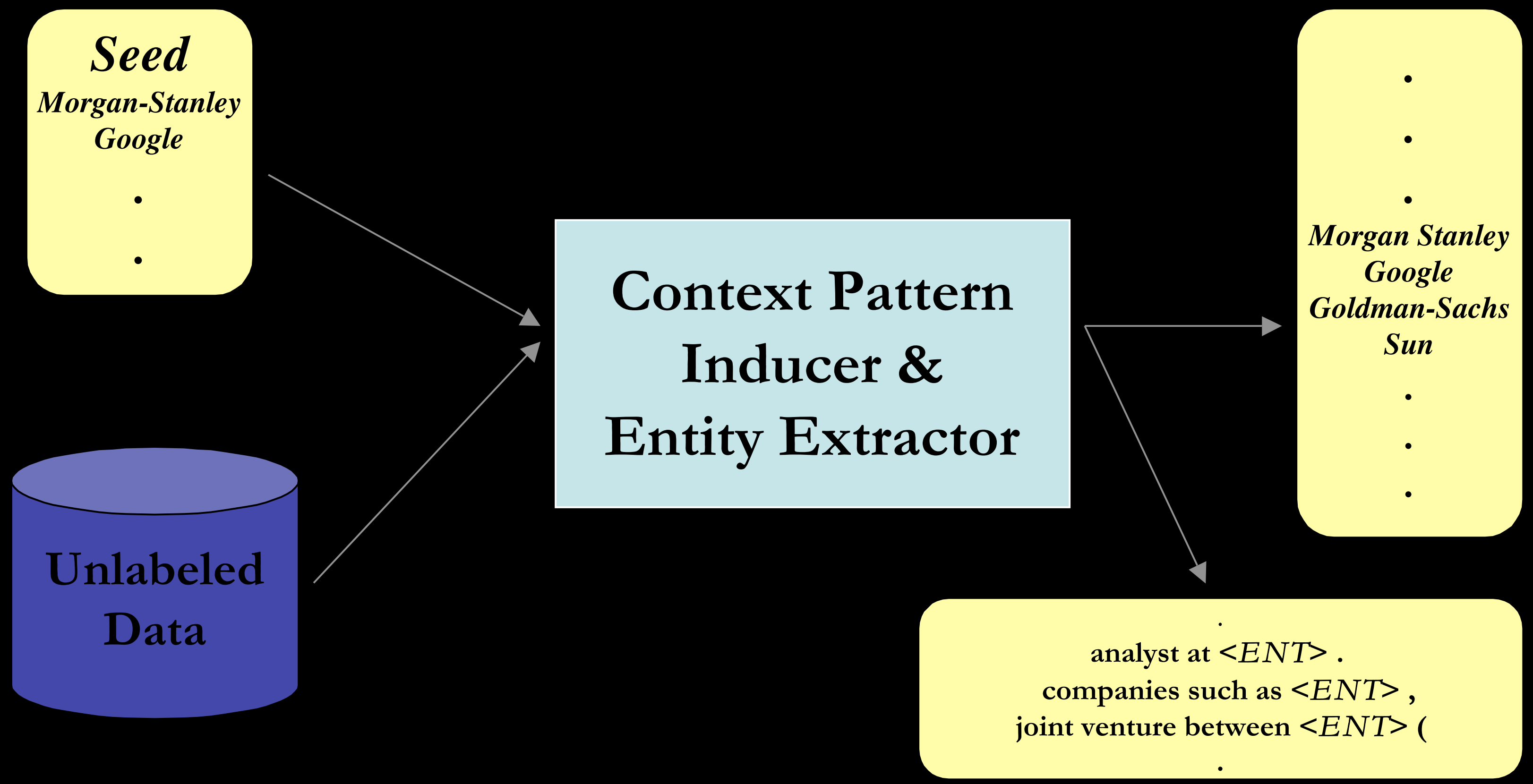
Motivation



Can anything be done by combining unlabeled data with partial entity lists ?

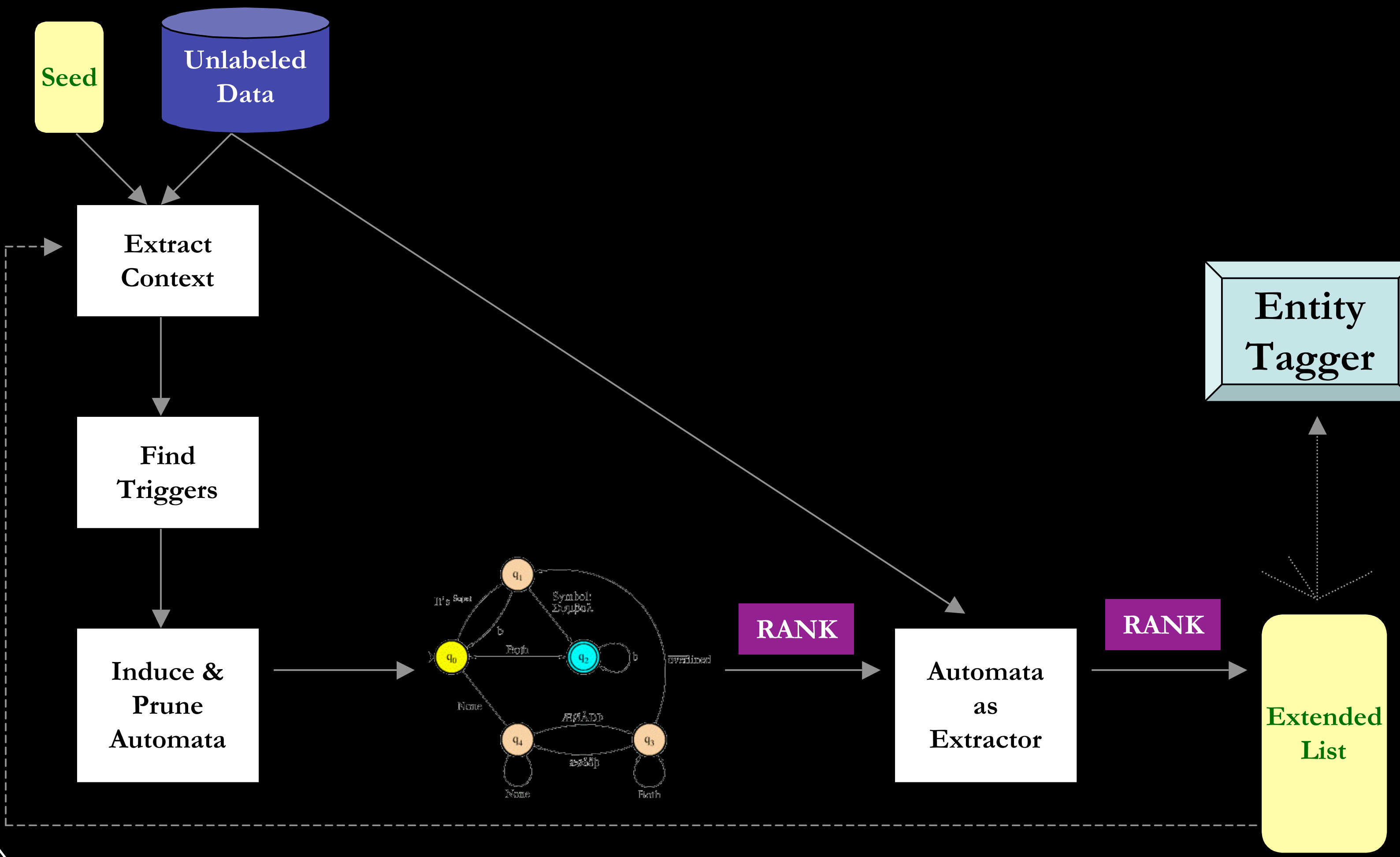
Objective

To Capture Redundancy in Expression.



Penn
University of Pennsylvania

Approach



** One automaton induced for each trigger word.

Preparing for Grammar Induction

an increased expression of ## adenosine deaminase ## in vad mic e
expression of a murine ## adenosine deaminase ## gene in rhesus monkey
contrast the expression of ## apolipoprotein e ## mrna was greater than

- Type of grammar: regular or context free ?
- Where do we start: *ideally patterns should be variable length.*
- What about starting from a token which is specific to the context of entities: *Trigger words.*

Penn
UNIVERSITY of PENNSYLVANIA

Trigger Words

Objective:

Automatically find out tokens which are specific to extracted entity contexts and which can indicate occurrence of entities in its neighbourhood.

- What about frequent tokens in entire corpus ?
- What about frequent tokens in extracted context ?

  - These tokens can be common everywhere.
- What about those with high term weights ?

  - Noise and very specific words can fill top slots.

Penn
University of Pennsylvania

Trigger Words: Dominating Words

- Assign term weight W_t to each token in context.
- From each context segment C_j, find *dominating word* (DW_j), the token with highest term weight:

$$DW_j = argmax_t W_t, \forall t \in C_j$$

- Exactly one dominating word is selected from each context. Compute frequency (multiplicity) of these dominating words .
- Consider top n as trigger words.

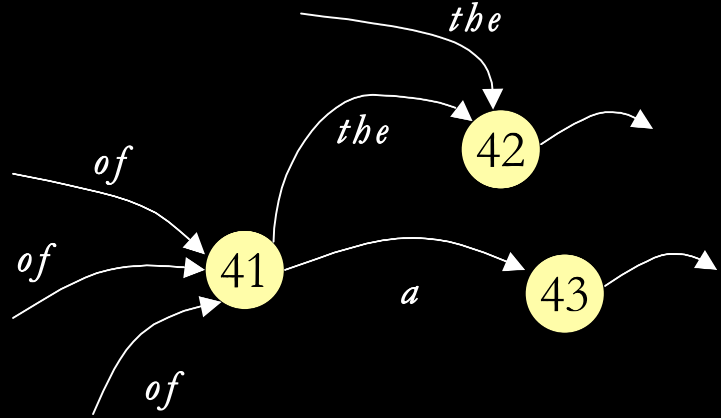
Trigger Words: Example

showed an increased expression of <ENT> in vad mice colon
vivo expression of a murine <ENT> gene in rhesus monkey hematopoietic
plasmodium falciparum expression of the <ENT> gene in mouse l cells
in contrast the expression of <ENT> mrna was greater than that

Token	Dominating Frequency	
expression	2	n = 1
murine	1	
falciparum	1	

Automata Induction

- One automaton induced for each trigger word.
- Given a token, we can uniquely identify the single state it points to: ***1-reversible***.



- Captures bi-gram statistics and helps combine evidence.
- Cycles are allowed.
- Induced automaton is to be used as an acceptor and not as generator.

Automaton Pruning

expression of -<ENT>- …

expression of ~~a~~ ~~murine~~ -<ENT>- …

expression of ~~the~~ -<ENT>- …

expression of -<ENT>- …

- Posterior score of each transition is computed using forward-backward algorithm.

- A transition is pruned if its posterior score is significantly lower than the best outgoing transition.

Penn
UNIVERSITY of PENNSYLVANIA

Automaton as Extractor

- Induced automata are used as extractors.
- Tokens that fit patterns' slots are ***candidate entities.***
- But can we directly consider candidate entity tokens as part of valid entity names ?

  - No. But simple heuristics work very well.

- Only candidates who together satisfy **K [D K]* K** are retained *e.g.*:

physicist at the University of Pennsylvania and

D K D K

Pattern: *physicist at <ENT> and*

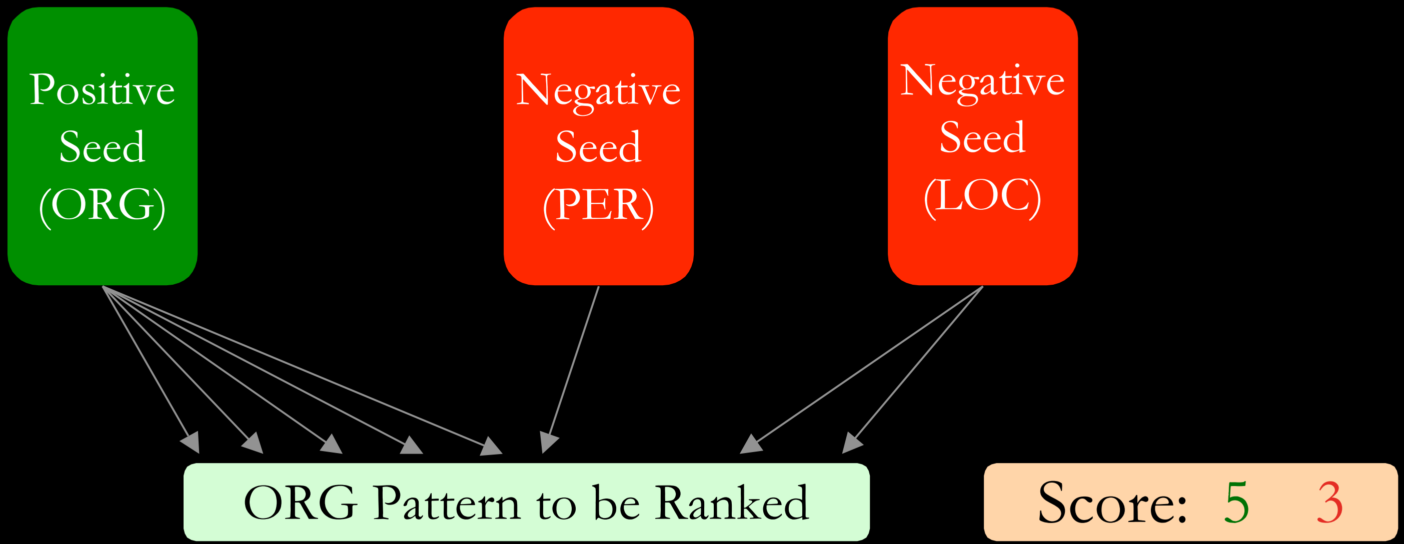
Extracted Entity: *University of Pennsylvania*

Penn

UNIVERSITY *of* PENNSYLVANIA

Pattern Ranking

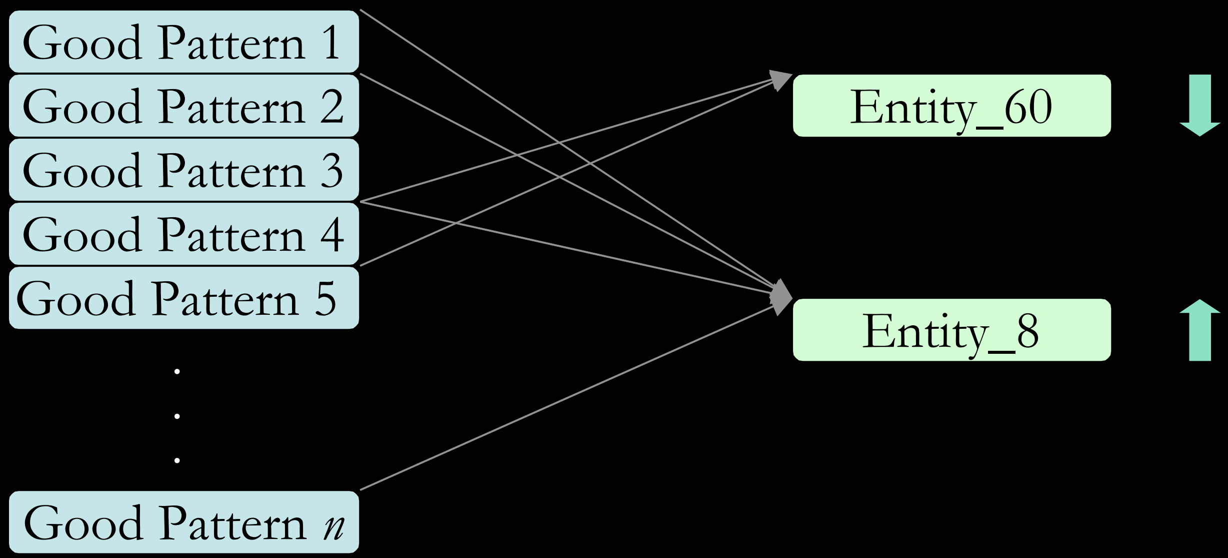
- All induced patterns are not equally good.



- Easier when working with multiple ambiguous classes at the same time.
- Finally select top ranking *n* patterns.

Extracted Entity Ranking

- An extracted entity gets a higher score if more number of *good patterns* (ranked as shown previously) extract it.



Experimental Results

Experiment with Watch Brand Names

- gold *-ENT-* watch
- diamond *-ENT-* watch
- fake *-ENT-* watches
- bought *-ENT-* watch
- encrusted *-ENT-* watch
- stole *-ENT-* watch
- Richemont AG , *-ENT-* watches
- Rolex and *-ENT-* watches
- buy *-ENT-* watches
- Cartier and *-ENT-* watches

…

Rolex
Cartier
Swiss
Movado
Seiko
Gucci
Patek
Piaget
Omega
Citizen
…

English Organization Name Experiment

- analyst at -*ENT*- .
- companies such as -*ENT*- .
- analyst with -*ENT*- in
- series against the -*ENT*-tonight
- Today 's Schaeffer 's Option Activity Watch features -*ENT*- (
- Cardinals and -*ENT*- ,
- sweep of the -*ENT*- with
- joint venture with -*ENT*- (
- rivals -*ENT*- Inc.
- Friday night 's game against -*ENT*- .

Boston Red Sox
St. Louis Cardinals
Chicago Cubs
Florida Marlins
Montreal Expos
San Francisco Giants
Red Sox
Cleveland Indians
Chicago White Sox
Atlanta Braves
…

Penn
University *of* Pennsylvania

English Person Name Experiment

- 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
…

- *More examples in the paper.*

Penn
University of Pennsylvania

Entity List Extension Results

Category	Seed Size	Extended Size	Precision
LOC	379	3001	70%
ORG	1597	33369	85%
PER	3616	86265	88%

- Precision is based on random evaluation of 100 entities.
- The method also works for very small seed list: watch brand name experiment with seed set size of 17.
- It is the **quality of the seed entities** (their unambiguous nature) that is more important than their number.

Penn
University of Pennsylvania

Influence on Supervised CRF Tagger

PER, LOC, ORG

Training Data (Tokens)	Test-a			Test-b		
	No List	Seed List	Unsup. List	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			Test-b		
	No List	Seed List	Unsup. List	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

Related Work

- Most of the previous methods *([Riloff & Jones '99]*, generic extractor in *[Etzioni et.al. '05*]) are language dependent (*e.g.* need chunking information) but current method is completely language independent.

- Successfully used features derived from unlabeled data (token membership in extended lists) to improve a high-performing CRF tagger.

- We report effectiveness of the algorithm on relatively large dataset of 18 billion tokens.

Penn
UNIVERSITY of PENNSYLVANIA

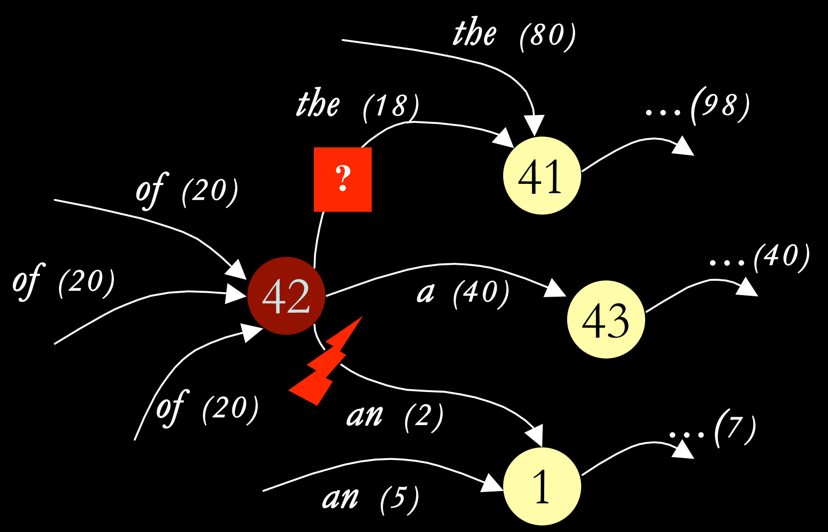
Future Work

- Empirical comparison with other methods.
- Better pattern and entity ranking.
- Compare to see whether features derived in this paper can complement other recent methods that also generate features from unlabeled data.
- Experiment with other languages and domains.

Thanks

Automaton Pruning (contd.)

- Which transitions to prune (remove) ?
- How about taking pruning decision locally ?



- There is possibility of transition (42, 41) getting pruned in some threshold based scheme when decision is taken locally.

Pruning

- For numerical stability, log probabilities are used which are processed as per following log-semiring definition:

  Set: [-inf, inf]

  Plus: log(exp(x) + exp(y))

  Zero: -inf

  Times: +

  One: 0

- After pruning, automata are trimmed.
- Automata are stored in AT&T FSM format.

Penn
UNIVERSITY *of* PENNSYLVANIA

German ORG & PER Experiment

Organization Patterns

- Tageszeitung " -<**ENT**>- "
- Zeitung -<**ENT**>- Â»
- Aktie von -<**ENT**>- mit
- laut " -<**ENT**>- "
- Laut " -<**ENT**>- "
- Heimspiel gegen -<**ENT**>- .
- empfehlen die Aktie von -<**ENT**>- (
- vwd) - Die -<**ENT**>- Inc
- Bei -<**ENT**>- geht
- Bericht der -<**ENT**>- Â»
- Wie die -<**ENT**>- Â»
- Airlines , -<**ENT**>- ,
- berichtete die -<**ENT**>- Â»
- berichtet die -<**ENT**>- Â»
- Analysten von -<**ENT**>- .
- Laut -<**ENT**>- Â»
- Analysten von -<**ENT**>- stufen
- Analysten von -<**ENT**>- die
- MarktfÃ¼hrer -<**ENT**>- .
- Klubs -<**ENT**>- und
- ......

Person Patterns

- s. -<**ENT**>- (
- Landsmann -<**ENT**>- .
- Nachfolger -<**ENT**>- ,
- Wer -<**ENT**>- ?
- Landsmann -<**ENT**>- (
- Seite von -<**ENT**>- in
- Seite von -<**ENT**>- und
- Superstars -<**ENT**>- und
- 7:5 , -<**ENT**>- (
- Kollege -<**ENT**>- .
- Prominente wie -<**ENT**>- ,
- Hollywoodstar -<**ENT**>- (
- Schauspielerin -<**ENT**>- ,
- Weltstars wie -<**ENT**>- ,
- Schauspieler -<**ENT**>- und
- Nationalspieler -<**ENT**>- (
- 6:1 , -<**ENT**>- (
- Angeles (dpa) - -<**ENT**>- (
- verletzten -<**ENT**>- und
- Schauspieler -<**ENT**>- (
- ......

Influence on Supervised Tagger

- Conditional Random Field (CRF) based tagger trained on CoNLL-2003 English data for LOC, ORG and PER names.
- Tested with and without automatically generated entity lists as additional features.
- Tested with varying amount of training data to test the hypothesis that the tagger benefits most from using unsupervised generated list when there is less training data.

Automata Induction

- All entity names are replaced by token *"<ENT>"*
- Only one token to the right of *"<ENT>"* considered.
- Cycles are allowed.
- Induced automaton is to be used as an acceptor and not as generator.
- Each transition is initially scored as follows:

$$Score(a_i, a_j) = \frac{TransCount(a_i, a_j)}{\sum_k TransCount(a_i, a_k)}$$