Information Extraction: Coreference and Relation Extraction

Lecture #20

Computational Linguistics

CMPSCI 591N, Spring 2006

University of Massachusetts Amherst

Andrew McCallum

But first...
A few slides on
Directed Graphical Models

Lecture #20

Computational Linguistics

CMPSCI 591N, Spring 2006

University of Massachusetts Amherst

Andrew McCallum

- Notation: $\mathbf{x}_A \perp \mathbf{x}_B | \mathbf{x}_C$
  Definition: two (sets of) variables $\mathbf{x}_A$ and $\mathbf{x}_B$ are conditionally independent given a third $\mathbf{x}_C$ if:

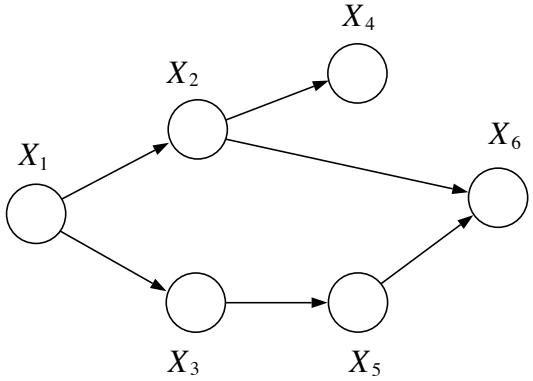
$$\mathsf{P}(\mathbf{x}_A, \mathbf{x}_B | \mathbf{x}_C) = \mathsf{P}(\mathbf{x}_A | \mathbf{x}_C) \mathsf{P}(\mathbf{x}_B | \mathbf{x}_C) \qquad \forall \mathbf{x}_C$$

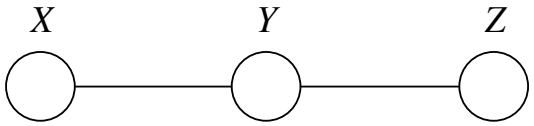
  which is equivalent to saying

$$\mathsf{P}(\mathbf{x}_A | \mathbf{x}_B, \mathbf{x}_C) = \mathsf{P}(\mathbf{x}_A | \mathbf{x}_C) \qquad \forall \mathbf{x}_C$$

- Only a subset of all distributions respect any given (nontrivial) conditional independence statement. The subset of distributions that respect all the CI assumptions we make is the *family of distributions consisitent with our assumptions*.
- Probabilistic graphical models are a powerful, elegant and simple way to specify such a family.

- Probabilistic graphical models represent large joint distributions compactly using a set of "local" relationships specified by a graph.
- Each random variable in our model corresponds to a graph node.
- There are directed/undirected *edges* between the nodes which tell us qualitatively about the *factorization* of the joint probability.
- There are *functions* stored at the nodes which tell us the quantitative details of the pieces into which the distribution factors.





- Graphical models are also known as Bayes(ian) (Belief) Net(work)s.

- Consider *directed acyclic graphs* over n variables.
- Each node has (possibly empty) set of parents π_i.
- Each node maintains a function $f_i(\mathbf{x}_i; \mathbf{x}_{\pi_i})$ such that $f_i > 0$ and $\sum_{\mathbf{x}_i} f_i(\mathbf{x}_i; \mathbf{x}_{\pi_i}) = 1 \; \forall \pi_i$.
- Define the joint probability to be:

$$\mathsf{P}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_i f_i(\mathbf{x}_i; \mathbf{x}_{\pi_i})$$

  Even with no further restriction on the the f_i, it is always true that

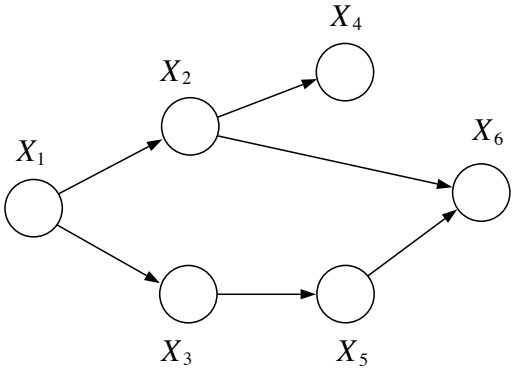
$$f_i(\mathbf{x}_i; \mathbf{x}_{\pi_i}) = \mathsf{P}(\mathbf{x}_i | \mathbf{x}_{\pi_i})$$

  so we will just write

$$\mathsf{P}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_i \mathsf{P}(\mathbf{x}_i | \mathbf{x}_{\pi_i})$$

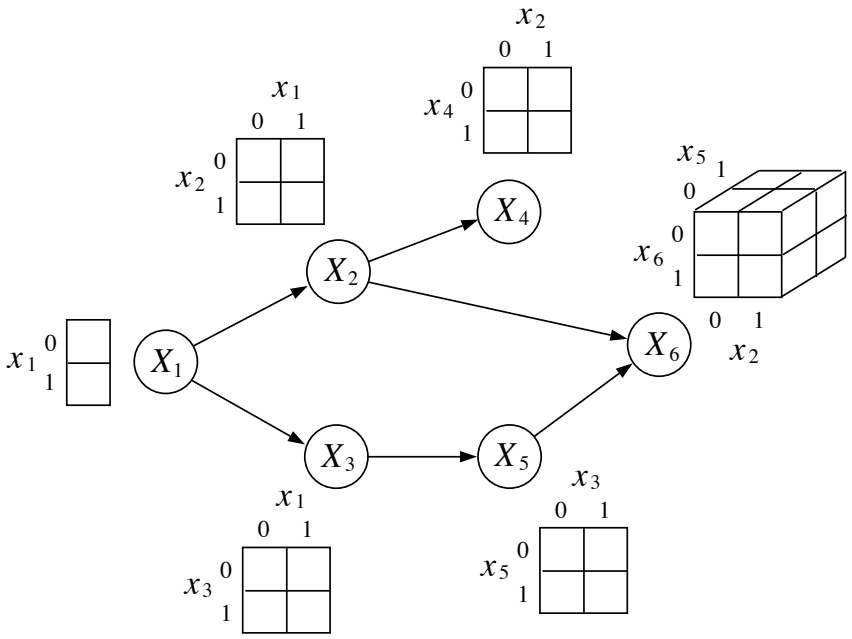
- Factorization of the joint in terms of *local conditional probabilities*. Exponential in "fan-in" of each node instead of in total variables n.

- Consider this six node network:



The joint probability is now:

$$\mathsf{P}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6) = \mathsf{P}(\mathbf{x}_1)\mathsf{P}(\mathbf{x}_2|\mathbf{x}_1)\mathsf{P}(\mathbf{x}_3|\mathbf{x}_1) \mathsf{P}(\mathbf{x}_4|\mathbf{x}_2)\mathsf{P}(\mathbf{x}_5|\mathbf{x}_3)\mathsf{P}(\mathbf{x}_6|\mathbf{x}_2, \mathbf{x}_5)$$



- Key point about directed graphical models:
  Missing edges imply conditional independence
- Remember, that by the chain rule we can always write the full joint as a product of conditionals, given an ordering:

  $$\mathsf{P}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \ldots) = \mathsf{P}(\mathbf{x}_1)\mathsf{P}(\mathbf{x}_2|\mathbf{x}_1)\mathsf{P}(\mathbf{x}_3|\mathbf{x}_1, \mathbf{x}_2)\mathsf{P}(\mathbf{x}_4|\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)\ldots$$

- If the joint is represented by a DAGM, then some of the conditioned variables on the right hand sides are missing.
  This is equivalent to enforcing conditional independence.
- Start with the "idiot's graph": each node has all previous nodes in the ordering as its parents.
- Now remove edges to get your DAG.
- Removing an edge into node i eliminates an argument from the conditional probability factor $p(\mathbf{x}_i|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{i-1})$

Information Extraction: Coreference and Relation Extraction

Lecture #20

Computational Linguistics

CMPSCI 591N, Spring 2006

University of Massachusetts Amherst

Andrew McCallum

What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

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

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

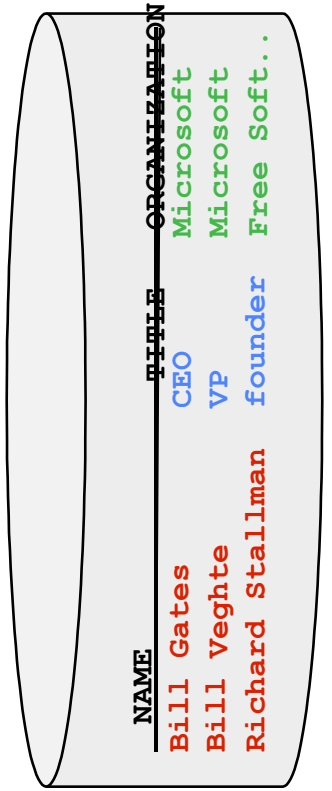
Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates

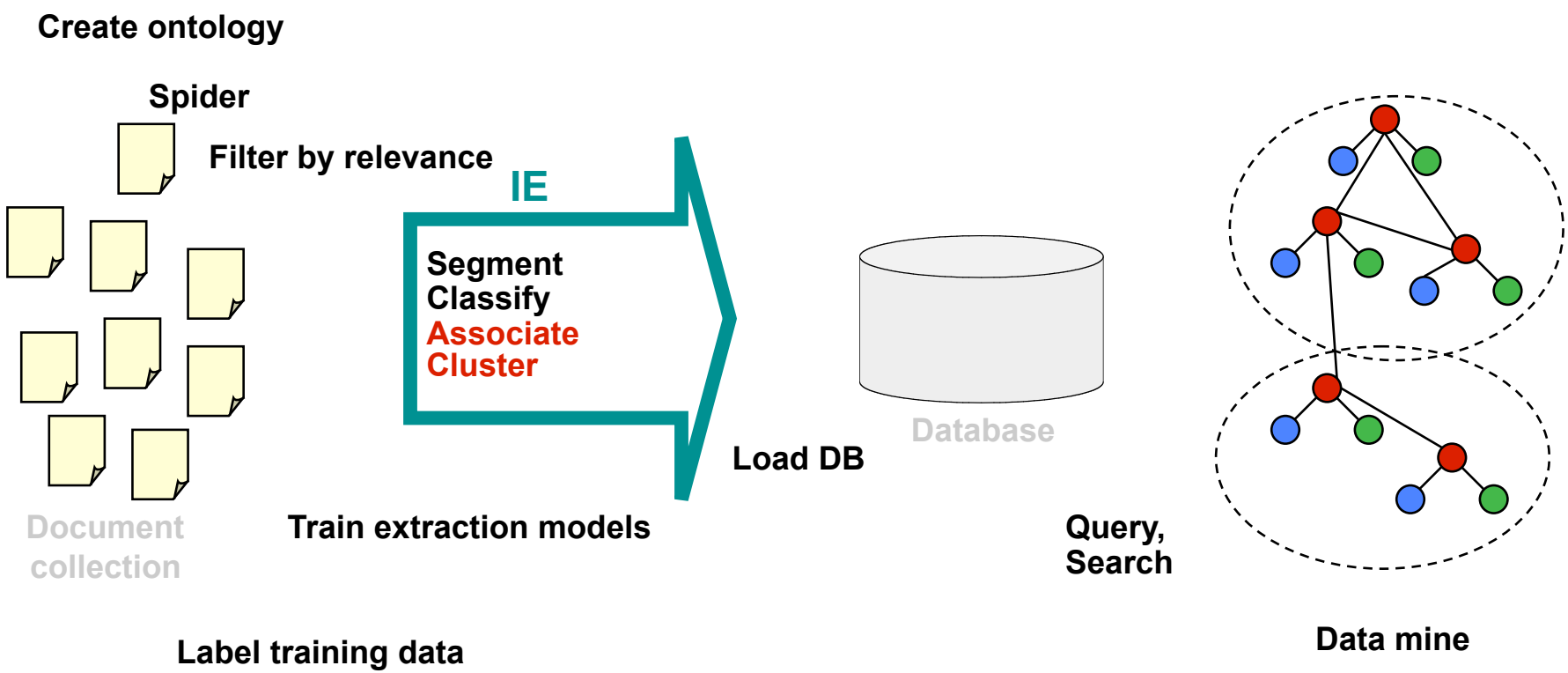
* Microsoft
Gates

* Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation



IE in Context



Main Points

Co-reference

- How to cast as classification [Cardie]
- Scaling up [McCallum et al]

Relation extraction

- With augmented grammar [Miller et al 2000]
- With joint inference [Roth & Yih]
- Semi-supervised [Brin]

Coreference Resolution

AKA "record linkage", "database record deduplication", "citation matching", "object correspondence", "identity uncertainty"

Input

News article, with named-entity "mentions" tagged

Today Secretary of State Colin Powell
met with . . . . . .
. . . . . he . . . . . .
. . . . . Condoleezza Rice . . . . .
. . . . Mr Powell . . . . . she . . . . . .
. . . . . Powell . . . . . .
. . . President Bush . . . . . .
. . . . . Rice . . . . . .
. . . . . Bush . . . . . .
. . . . . .
. . . . . .

Output

Number of entities, *N* = *3*

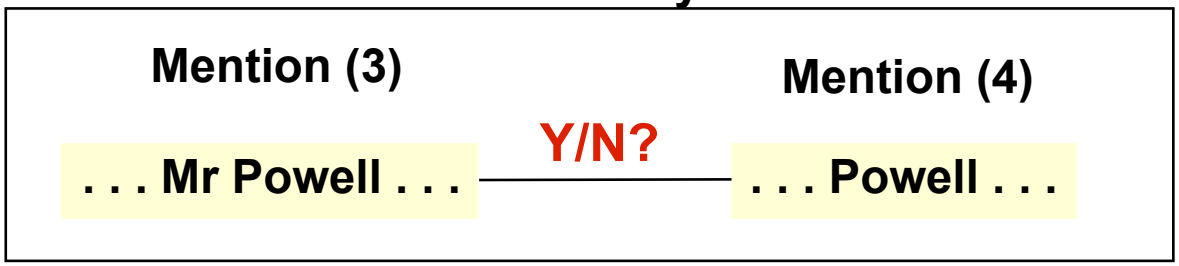
#1
- Secretary of State Colin Powell
- he
- Mr. Powell
- Powell

#2
- Condoleezza Rice
- she
- Rice

#3
- President Bush
- Bush

Inside the Traditional Solution

Pair-wise Affinity Metric



N	Two words in common	29
Y	One word in common	13
Y	"Normalized" mentions are string identical	39
Y	Capitalized word in common	17
Y	> 50% character tri-gram overlap	19
N	< 25% character tri-gram overlap	-34
Y	In same sentence	9
Y	Within two sentences	8
N	Further than 3 sentences apart	-1
Y	"Hobbs Distance" < 3	11
N	Number of entities in between two mentions = 0	12
N	Number of entities in between two mentions > 4	-3
Y	Font matches	1
Y	Default	-19
	OVERALL SCORE =	98 > threshold=0

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...

IE Example: Coreference

SAN SALVADOR, 15 JAN 90 (ACAN-EFE) -- [TEXT] ARMANDO CALDERON SOL, PRESIDENT OF THE NATIONALIST REPUBLICAN ALLIANCE (ARENA), THE RULING SALVADORAN PARTY, TODAY CALLED FOR AN INVESTIGATION INTO ANY POSSIBLE CONNECTION BETWEEN THE MILITARY PERSONNEL IMPLICATED IN THE ASSASSINATION OF JESUIT PRIESTS.

"IT IS SOMETHING SO HORRENDOUS, SO MONSTROUS, THAT WE MUST INVESTIGATE THE POSSIBILITY THAT THE FMLN (FARABUNDO MARTI NATIONAL LIBERATION FRONT) STAGED THESE MURDERS TO DISCREDIT THE GOVERNMENT," CALDERON SOL SAID.

SALVADORAN PRESIDENT ALFREDO CRISTIANI IMPLICATED FOUR OFFICERS, INCLUDING ONE COLONEL, AND FIVE MEMBERS OF THE ARMED FORCES IN THE ASSASSINATION OF SIX JESUIT PRIESTS AND TWO WOMEN ON 16 NOVEMBER AT THE CENTRAL AMERICAN UNIVERSITY.

Why It's Hard

Many sources of information play a role

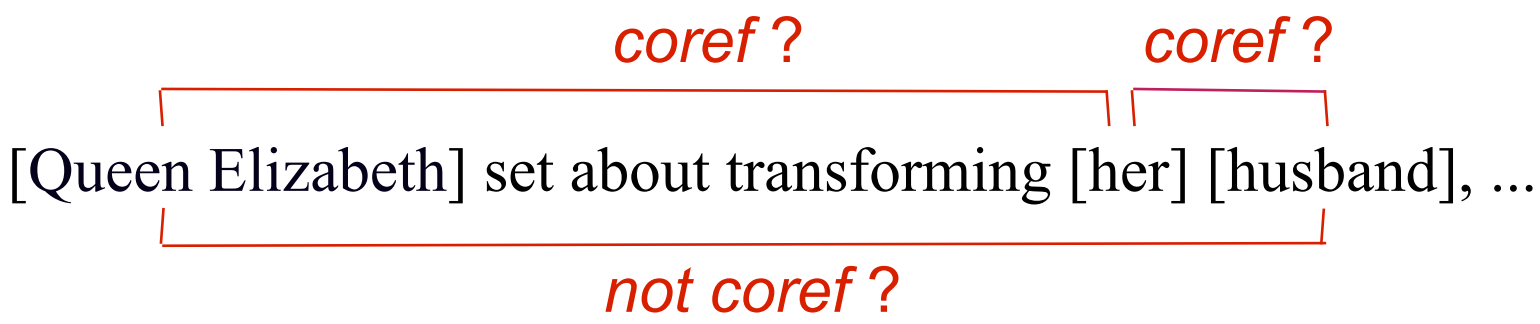
- head noun matches
  - IBM *executives* = the *executives*
- syntactic constraints
  - John helped himself to...
  - John helped him to…
- number and gender agreement
- discourse focus, recency, syntactic parallelism, semantic class, world knowledge, …

Why It's Hard

- No single source is a completely reliable indicator
  - number agreement
    - the assassination = these murders
- Identifying each of these features automatically, accurately, and in context, is hard
- Coreference resolution subsumes the problem of pronoun resolution…

A Machine Learning Approach

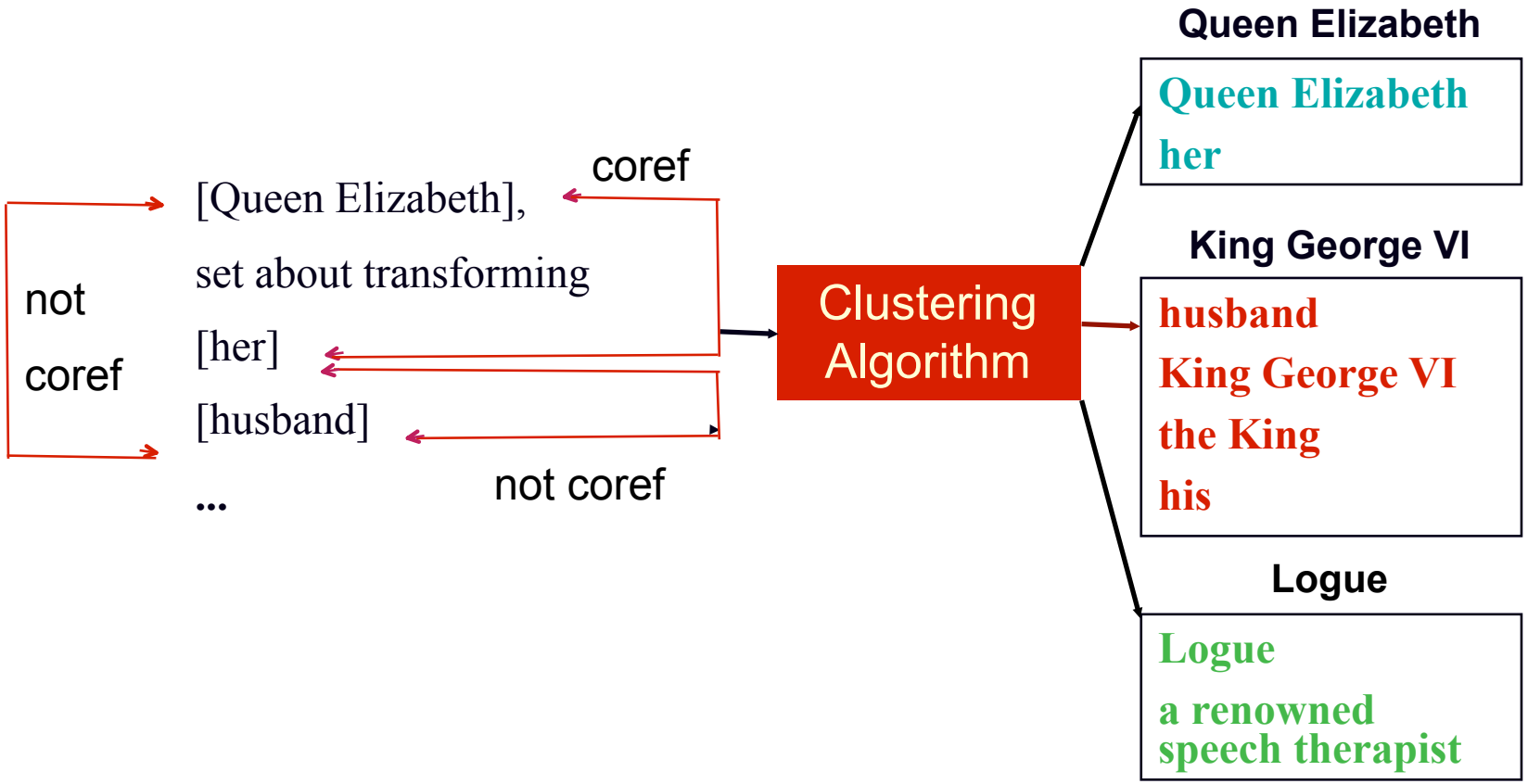
- Classification
  - given a description of two noun phrases, NP_i and NP_j, classify the pair as *coreferent* or *not coreferent*



Aone & Bennett [1995]; Connolly et al. [1994]; McCarthy & Lehnert [1995]; Soon et al. [2001]; Ng & Cardie [2002]; …

A Machine Learning Approach

- Clustering
  - coordinates pairwise coreference decisions



Machine Learning Issues

- Training data creation
- Instance representation
- Learning algorithm
- Clustering algorithm

Training Data Creation

- Creating training instances
  - texts annotated with coreference information
  - one instance $inst(NP_i, NP_j)$ for each pair of NPs
    - assumption: NP_i precedes NP_j
    - feature vector: describes the two NPs and context
    - class value:

coref	pairs on the same coreference chain
not coref	otherwise

Instance Representation

- 25 features per instance
  - lexical (3)
    - string matching for pronouns, proper names, common nouns
  - grammatical (18)
    - pronoun, demonstrative (the, this), indefinite (it is raining), …
    - number, gender, animacy
    - appositive (george, the king), predicate nominative (a horse is a mammal)
    - binding constraints, simple contra-indexing constraints, …
    - span, maximalnp, …
  - semantic (2)
    - same WordNet class
    - alias
  - positional (1)
    - distance between the NPs in terms of # of sentences
  - knowledge-based (1)
    - naïve pronoun resolution algorithm

Learning Algorithm

- RIPPER (Cohen, 1995)
  C4.5 (Quinlan, 1994)
  - rule learners
    - input: set of training instances
    - output: coreference classifier

- Learned classifier
  - input: test instance (represents pair of NPs)
  - output: classification
    confidence of classification

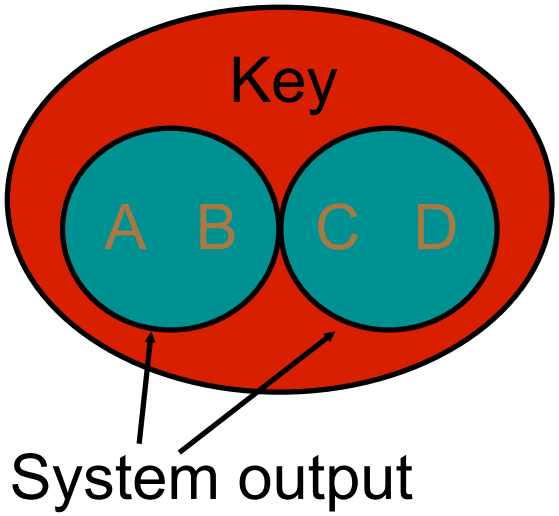
Clustering Algorithm

- Best-first single-link clustering
  - Mark each NP_j as belonging to its own class: $NP_j \in c_j$
  - Proceed through the NPs in left-to-right order.
    - For each NP, NP_j, create test instances, $inst(NP_i, NP_j)$, for all of its preceding NPs, NP_i.
    - Select as the antecedent for NP_j the highest-confidence coreferent NP, NP_i, according to the coreference classifier (or none if all have below .5 confidence);

      Merge c_j and c_j .

Evaluation

- MUC-6 and MUC-7 coreference data sets
- documents annotated w.r.t. coreference
- 30 + 30 training texts (dry run)
- 30 + 20 test texts (formal evaluation)
- scoring program
  - recall
  - precision
  - F-measure: 2PR/(P+R)

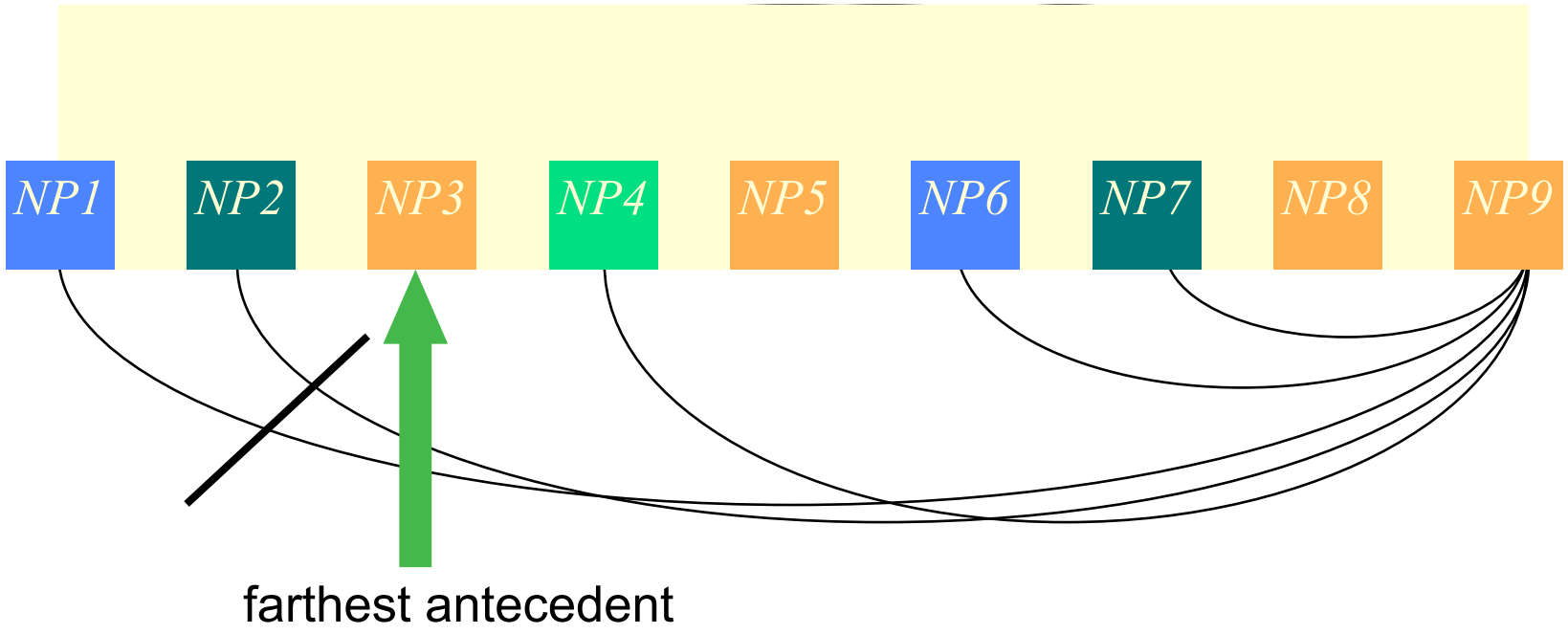


Baseline Results

	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	40.7	73.5	**52.4**	27.2	86.3	**41.3**
Worst MUC System	36	44	40	52.5	21.4	30.4
Best MUC System	59	72	65	56.1	68.8	61.8

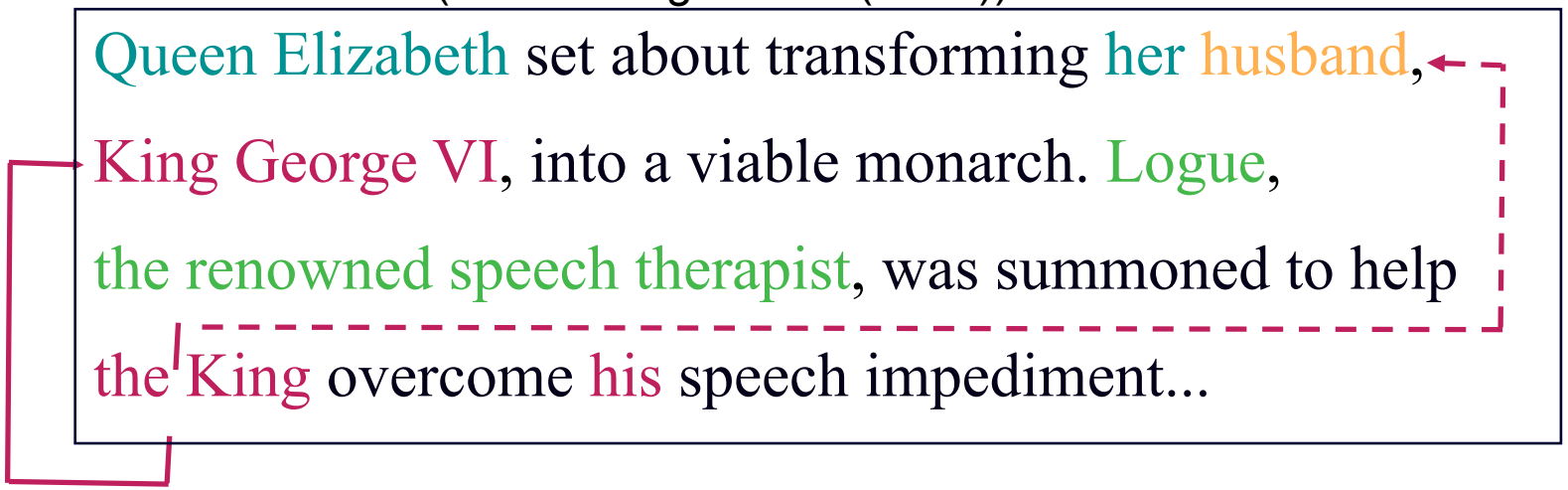
Problem 1

- Coreference is a rare relation
  - skewed class distributions (2% positive instances)
  - *remove some negative instances*



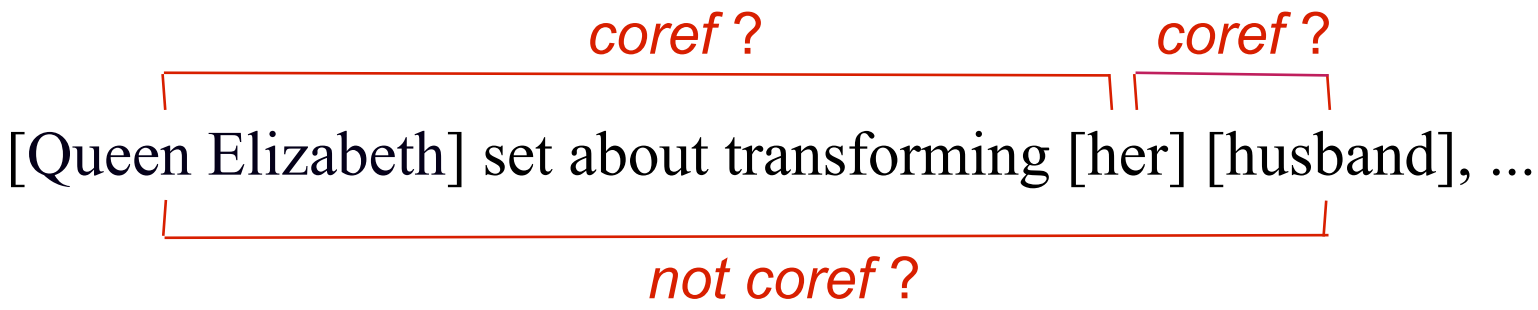
Problem 2

- Coreference is a discourse-level problem
  - different solutions for different types of NPs
    - proper names: string matching and aliasing
  - inclusion of “hard” positive training instances
  - *positive example selection*: selects easy positive training instances (cf. Harabagiu *et al.* (2001))



Problem 3

- Coreference is an equivalence relation
  - loss of transitivity
  - need to tighten the connection between classification and clustering
  - *prune learned rules w.r.t. the clustering-level coreference scoring function*



Results

	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	40.7	73.5	52.4	27.2	86.3	41.3
NEG-SELECT	46.5	67.8	55.2	37.4	59.7	46.0
POS-SELECT	53.1	80.8	64.1	41.1	78.0	53.8
NEG-SELECT + POS-SELECT	63.4	76.3	69.3	59.5	55.1	57.2
NEG-SELECT + POS-SELECT + RULE-SELECT	63.3	76.9	**69.5**	54.2	76.3	**63.4**

- Ultimately: large increase in F-measure, due to gains in recall

Comparison with Best MUC Systems

	MUC-6			MUC-7		
	R	P	F	R	P	F
NEG-SELECT + POS-SELECT + RULE-SELECT	63.3	76.9	**69.5**	54.2	76.3	**63.4**
Best MUC System	59	72	**65**	56.1	68.8	**61.8**

Main Points

Co-reference

- How to cast as classification [Cardie]
- **Scaling up [McCallum et al]**

Relation extraction

- With augmented grammar [Miller et al 2000]
- With joint inference [Roth]
- Semi-supervised [Brin]

Reference Matching

- Fahlman, Scott & Lebiere, Christian (1989). The cascade-correlation learning architecture. In Touretzky, D., editor, Advances in Neural Information Processing Systems (volume 2), (pp. 524-532), San Mateo, CA. Morgan Kaufmann.

- Fahlman, S.E. and Lebiere, C., "The Cascade Correlation Learning Architecture," NIPS, Vol. 2, pp. 524-532, Morgan Kaufmann, 1990.

- Fahlman, S. E. (1991) The recurrent cascade-correlation learning architecture. In Lippman, R.P. Moody, J.E., and Touretzky, D.S., editors, NIPS 3, 190-205.

The Citation Clustering Data

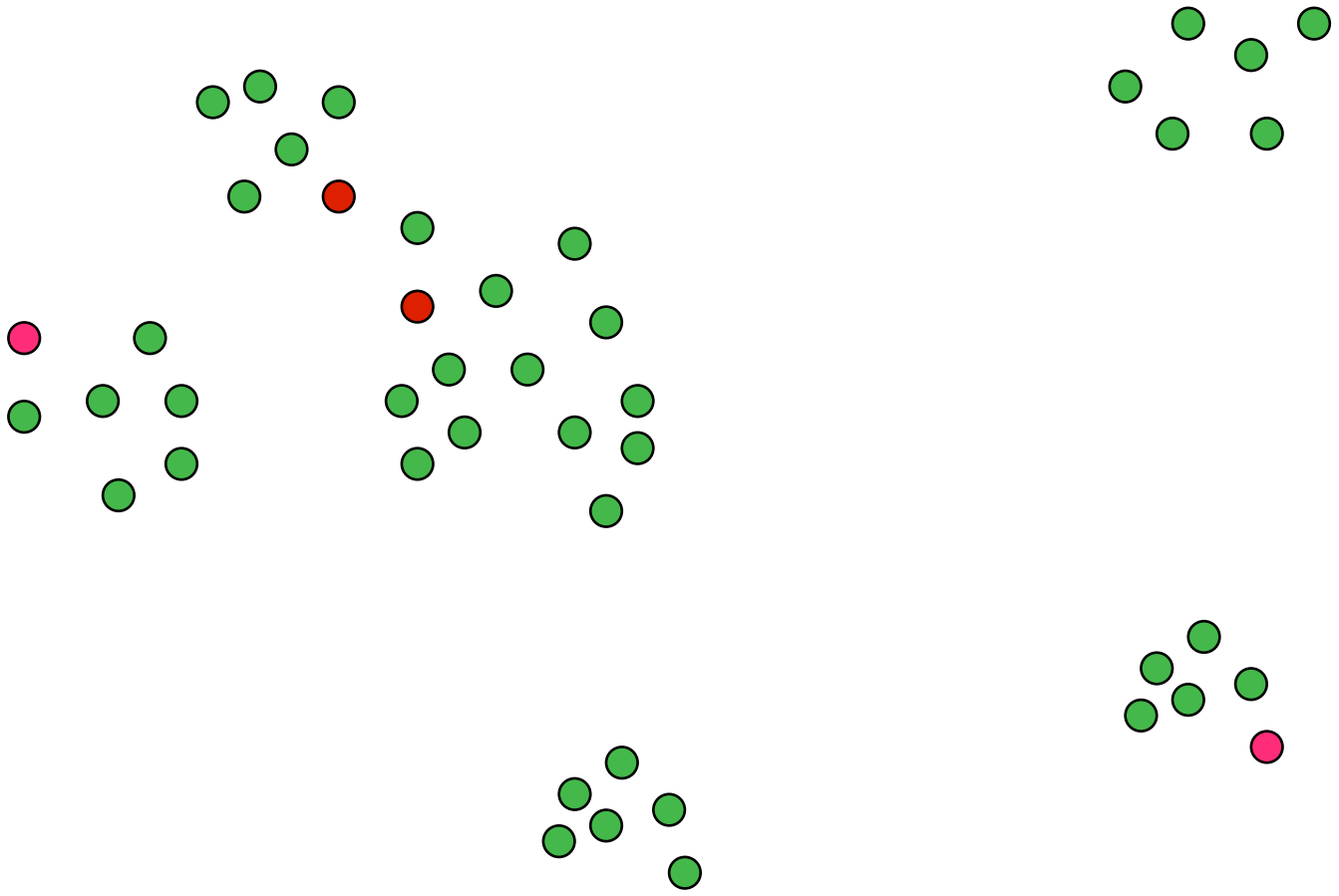
- Over 1,000,000 citations
- About 100,000 unique papers
- About 100,000 unique vocabulary words

- Over 1 trillion distance calculations

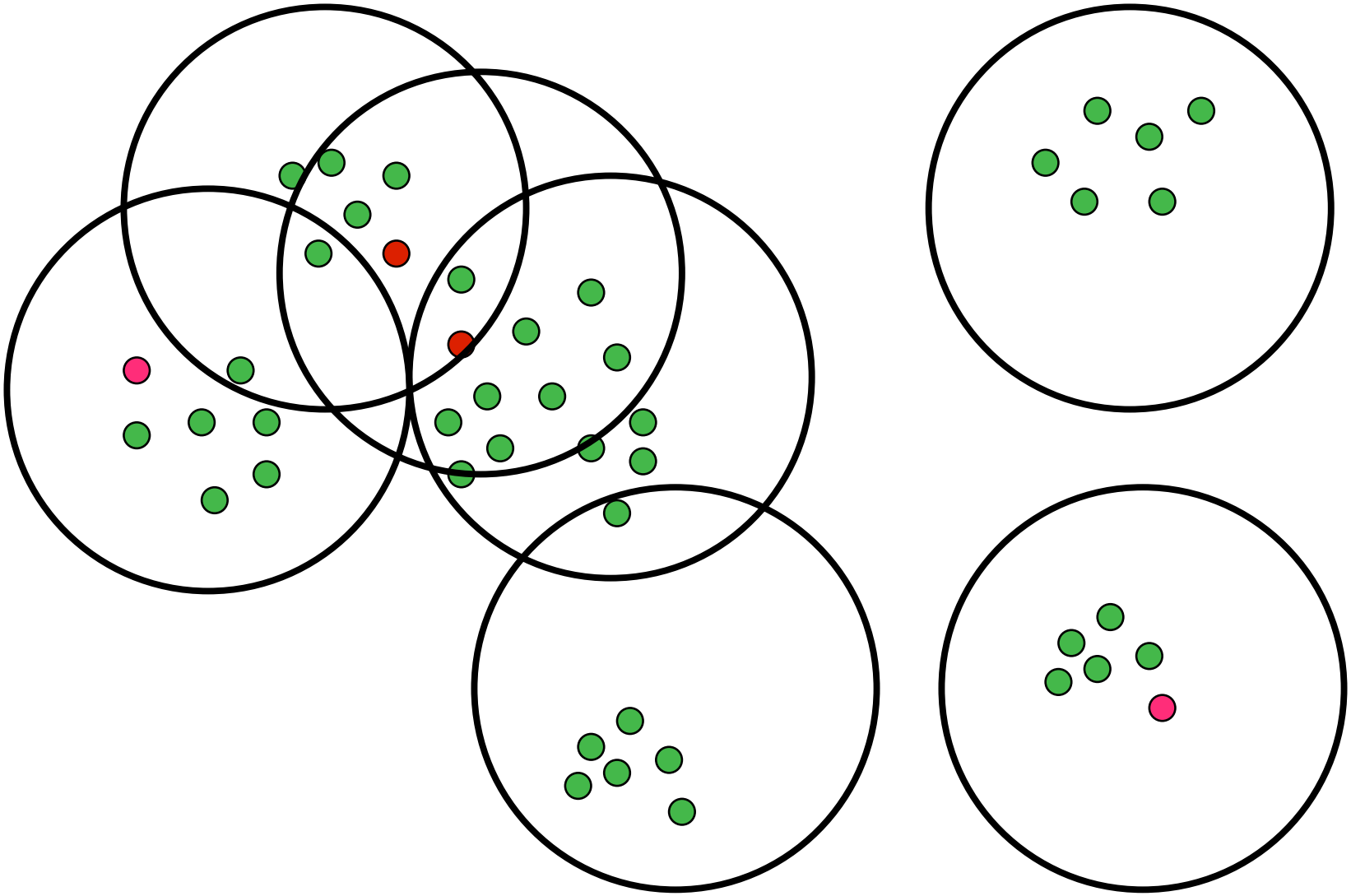
The Canopies Approach

- Two distance metrics: cheap & expensive
- First Pass
  - very inexpensive distance metric
  - create overlapping canopies
- Second Pass
  - expensive, accurate distance metric
  - canopies determine which distances calculated

Illustrating Canopies

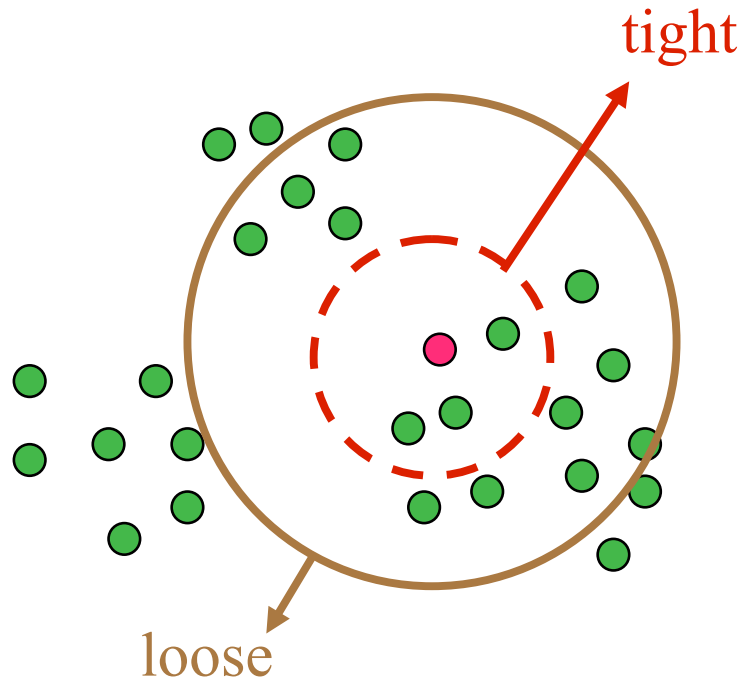


Overlapping Canopies



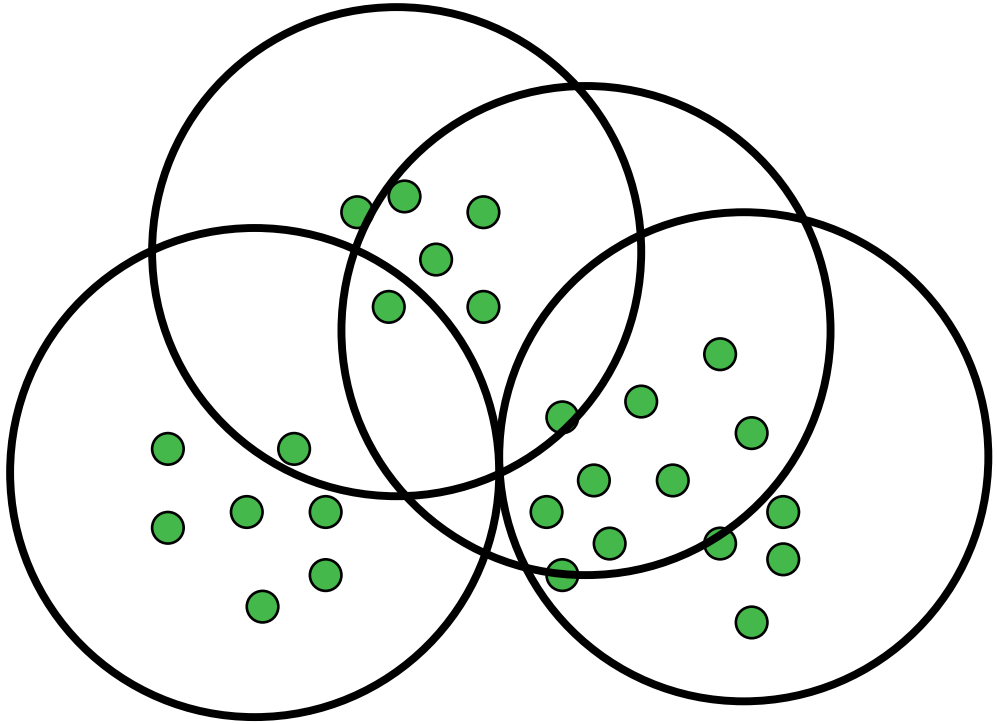
Creating canopies with two thresholds

- Put all points in D
- Loop:
  - Pick a point X from D
  - Put points within K_{loose} of X in canopy
  - Remove points within K_{tight} of X from D



Using canopies with Greedy Agglomerative Clustering

- Calculate expensive distances between points in the same canopy
- All other distances default to infinity
- Sort finite distances and iteratively merge closest



Computational Savings

- inexpensive metric << expensive metric
- # canopies per data point: f (small, but > 1)
- number of canopies: c (large)

- complexity reduction:

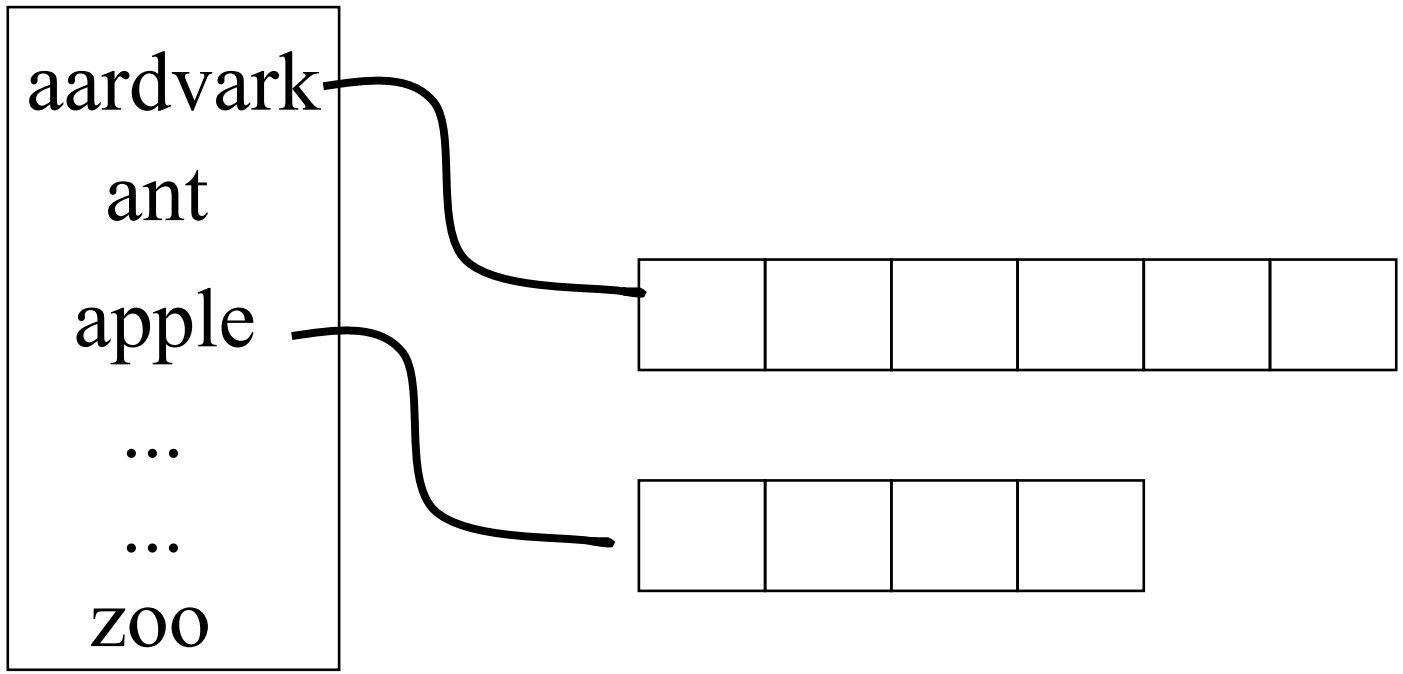
$$O\left(\frac{f^2}{c}\right)$$

The Experimental Dataset

- All citations for authors:
  - Michael Kearns
  - Robert Schapire
  - Yoav Freund
- 1916 citations
- 121 unique papers
- Similar dataset used for parameter tuning

Inexpensive Distance Metric for Text

- Word-level matching (TFIDF)
- Inexpensive using an inverted index



Expensive Distance Metric for Text

- String edit distance
- Compute with Dynamic Programming
- Costs for character:
  - insertion
  - deletion
  - substitution
  - ...

		S	e	c	a	t
	0.0	0.7	1.4	2.1	2.8	3.5
S	0.7	0.0	0.7	1.1	1.4	1.8
c	1.4	0.7	1.0	0.7	1.4	1.8
o	2.1	1.1	1.7	1.4	1.7	2.4
t	2.8	1.4	2.1	1.8	2.4	1.7
t	3.5	1.8	2.4	2.1	2.8	2.4

do Fahlman vs Falman

Experimental Results

	F1	Minutes
Canopies GAC	0.838	7.65
Complete GAC	0.835	134.09
Old Cora	0.784	0.03
Author/Year	0.697	0.03

Add precision, recall along side F1

Main Points

Co-reference

- How to cast as classification [Cardie]
- Measures of string similarity [Cohen]
- Scaling up [McCallum et al]

Relation extraction

- **With augmented grammar [Miller et al 2000]**
- With joint inference [Roth & Yih]
- Semi-supervised [Brin]

Information Extraction

Named Entity Recognition

INPUT: Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT: Profits soared at [Company Boeing Co.], easily topping forecasts on [Location Wall Street], as their CEO [Person Alan Mulally] announced first quarter results.

Relationships between Entities

INPUT: Boeing is located in Seattle. Alan Mulally is the CEO.

OUTPUT:

{Relationship = Company-Location
Company = Boeing
Location = Seattle}

{Relationship = Employer-Employee
Employer = Boeing Co.
Employee = Alan Mulally}

Extraction From Entire Documents

Hi [PERSON Ted] and [PERSON Hill],
Just a reminder that the game move will need to be entered [TIME tonight]. We will need data on operations, rawmaterials ordering, and details of the bond to be sold.
[PERSON Hill]: I will be in the [LOCATION lobby] after the class at [TIME 9 pm]. how about we meet in the [LOCATION lobby] around that time (i.e when both our classes are over).
[PERSON Ted]: Let me know how you are going to provide the bond related input information. We can either meet in the [LOCATION lobby] around [TIME 5.30 pm] or you can e-mail me the info.
Thanks, [PERSON Ajay]

TIME	9 pm, 18th September	TIME	5.30 pm, 18th September
LOCATION	Lobby, Building NE43	LOCATION	Lobby, Building NE43
PERSON	David Hill, Ajay Sinclair	PERSON	Ted Jones, Ajay Sinclair
TOPIC	data on operations. . .	TOPIC	bond related input information

10TH DEGREE is a full service advertising agency specializing in direct and interactive marketing. Located in Irvine CA, 10TH DEGREE is looking for an Assistant Interactive Account Manager to help manage and coordinate interactive marketing initiatives for a marquee automotive account. Experience in online marketing, automotive and/or the advertising agency field is a plus.
Assistant Account Manager Responsibilities
Ensures smooth implementation of programs and initiatives Helps manage the delivery of projects and key client deliverables ...
Compensation: $50,000 – $80,000 Hiring Organization: 10TH DEGREE
Principals only. Recruiters, please don't contact this job poster. Please, no phone calls about this job! Please do not contact job poster about other services, products or commercial interests. Reposting this message elsewhere is NOT OK. this is in or around Orange County - Irvine

INDUSTRY	Advertising
POSITION	Assistant Account Manager
LOCATION	Irvine, CA
COMPANY	10th Degree
SALARY	$50,000 – $80,000

Relationship Extraction

[Miller et. al, 2000]

An example:

Donald M. Goldstein, a historian at the University of Pittsburgh . . .

- Entity information to be extracted:
  - Named entity boundaries:
    Organizations, people, and locations
  - Person descriptors: "a historian at the University of Pittsburgh" refers to "Donald M. Goldstein"
- Entity relationships to be extracted:
  - Employer/Employee relations
    (e.g., *Goldstein* is employed at *University of Pittsburgh*)
  - Company/product relations
  - Organization/headquarters-location relation

Relationship Extraction: Annotation

Another example:

Nance, who is a paid consultant to ABC News, said . . .

- The following information was annotated:
  - *Nance* as a person; *ABC News* as an organization; *a paid consultant to ABC News* as a descriptor
  - A *coreference* link between *Nance* and *a paid consultant to ABC News*
  - An *employer-relation* link from *a paid consultant to ABC News* to *ABC News*

Next question: how can we build a model which recovers this information?

The Basic Approach

- Build a statistical parsing model which simultaneously recovers syntactic relation and the information extraction information

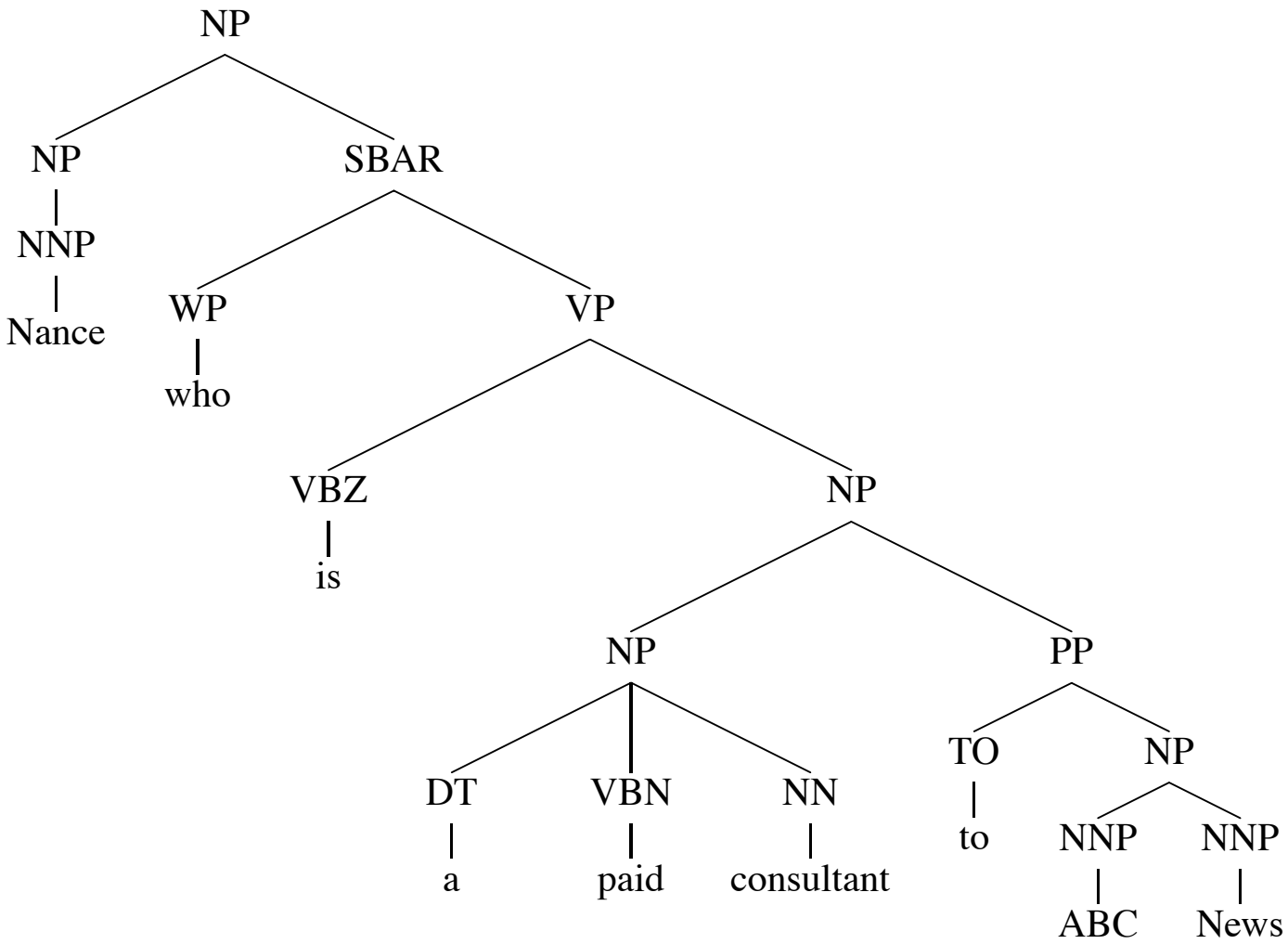
  To do this:

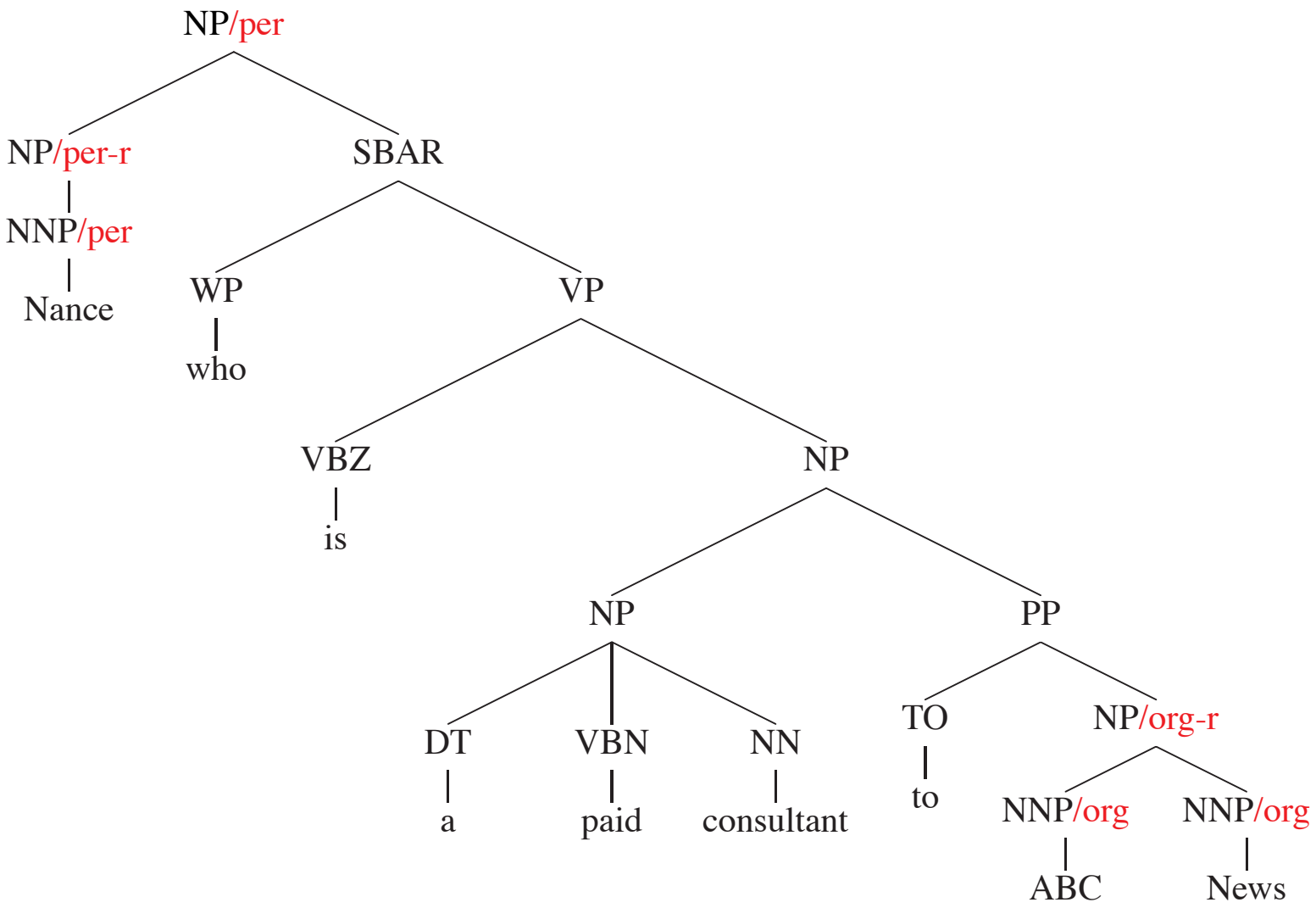
- Step 1: annotate training sentences for entities, descriptors, coreference links, and relation links

- Step 2: train a parser on the Penn treebank, and apply it to the new training sentences. **Force the parser to produce parses that are consistent with the entity/descriptor etc. boundaries**

- Step 3: enhance the parse trees to include the information extraction information (we'll come to this soon)

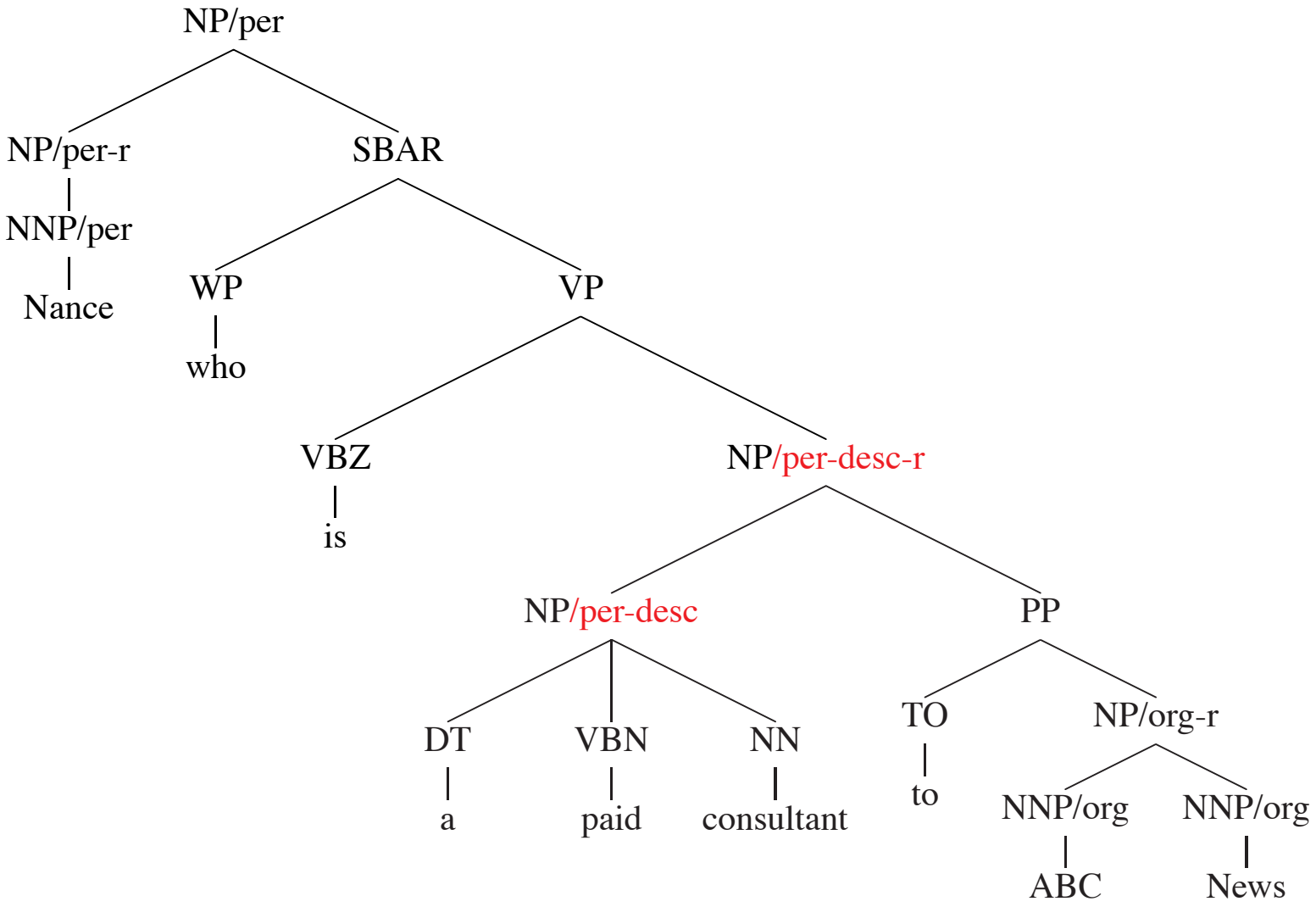
- Step 4: **re-train** the parser on the new training data, and with the new annotations


NP
NP
SBAR
NNP
WP
VP
Nance
who
VBZ
NP
is
NP
PP
DT
VBN
NN
TO
NP
a
paid
consultant
to
NNP
NNP
ABC
News



Add semantic tags showing named entities

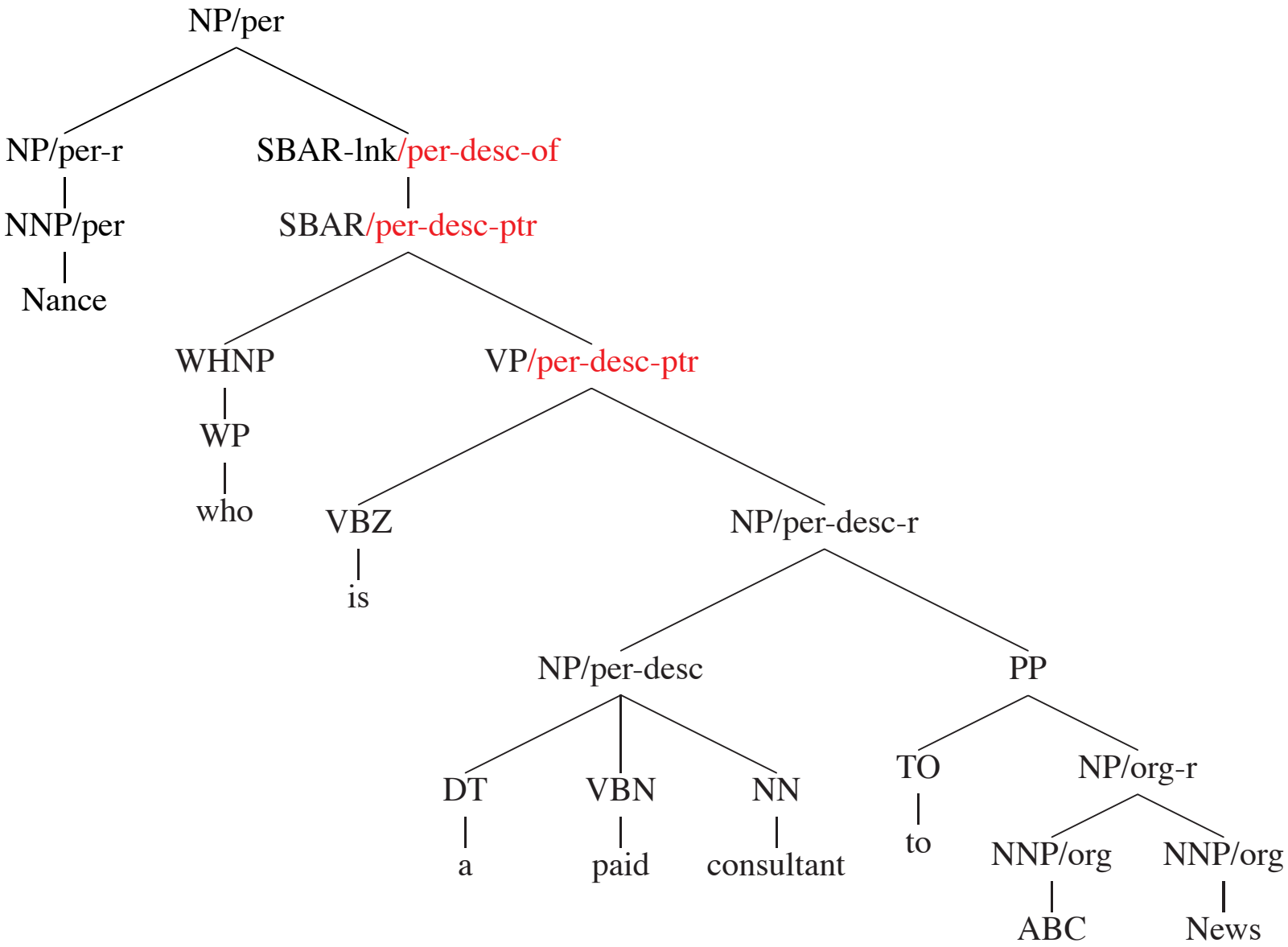
org = organization, per = person, org-r = organization "reportable" (complete), per-r = person "reportable" (complete)



Add semantic tags showing descriptors

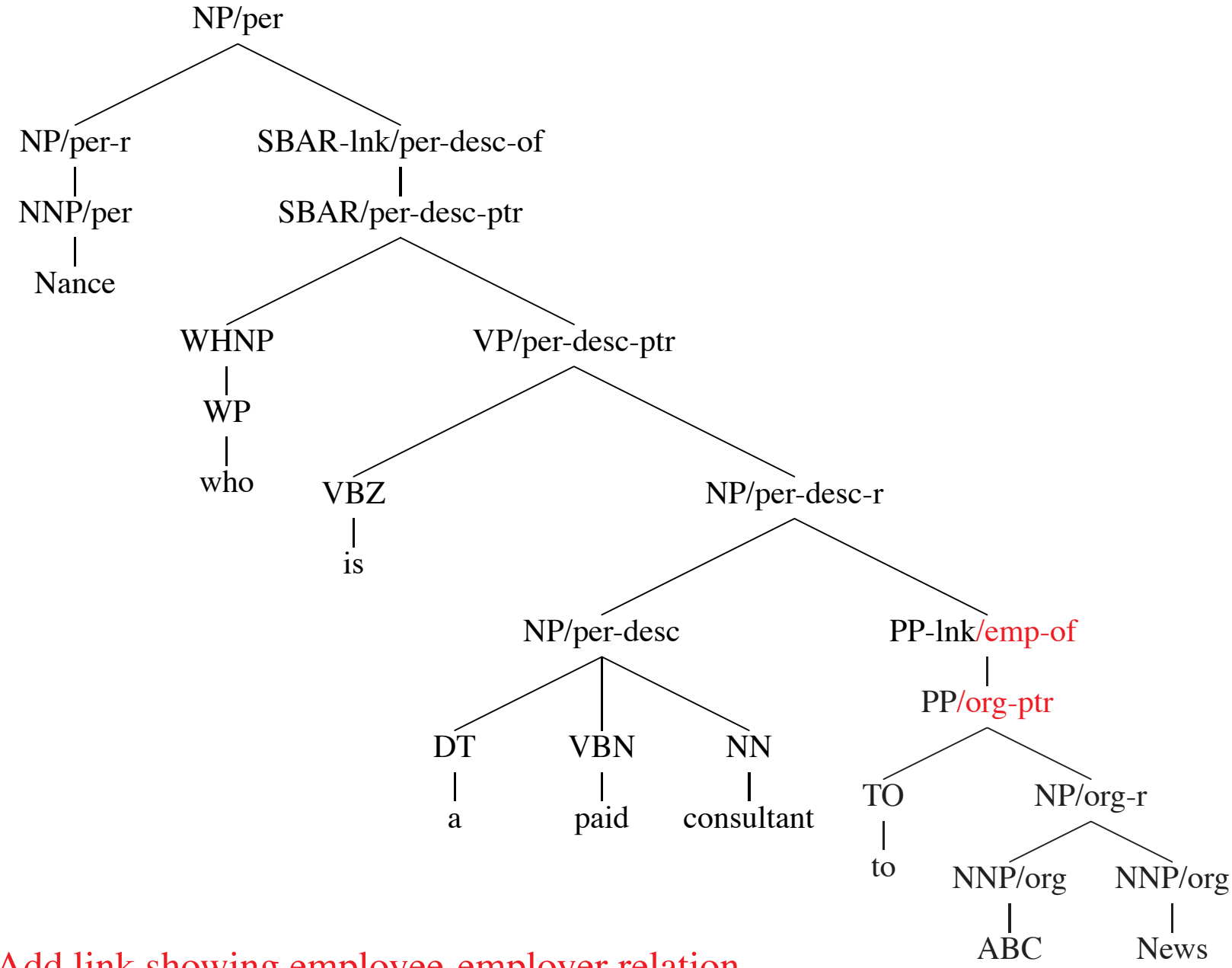
per-desc = person descriptor,

per-desc-r = person descriptor "reportable" (complete)



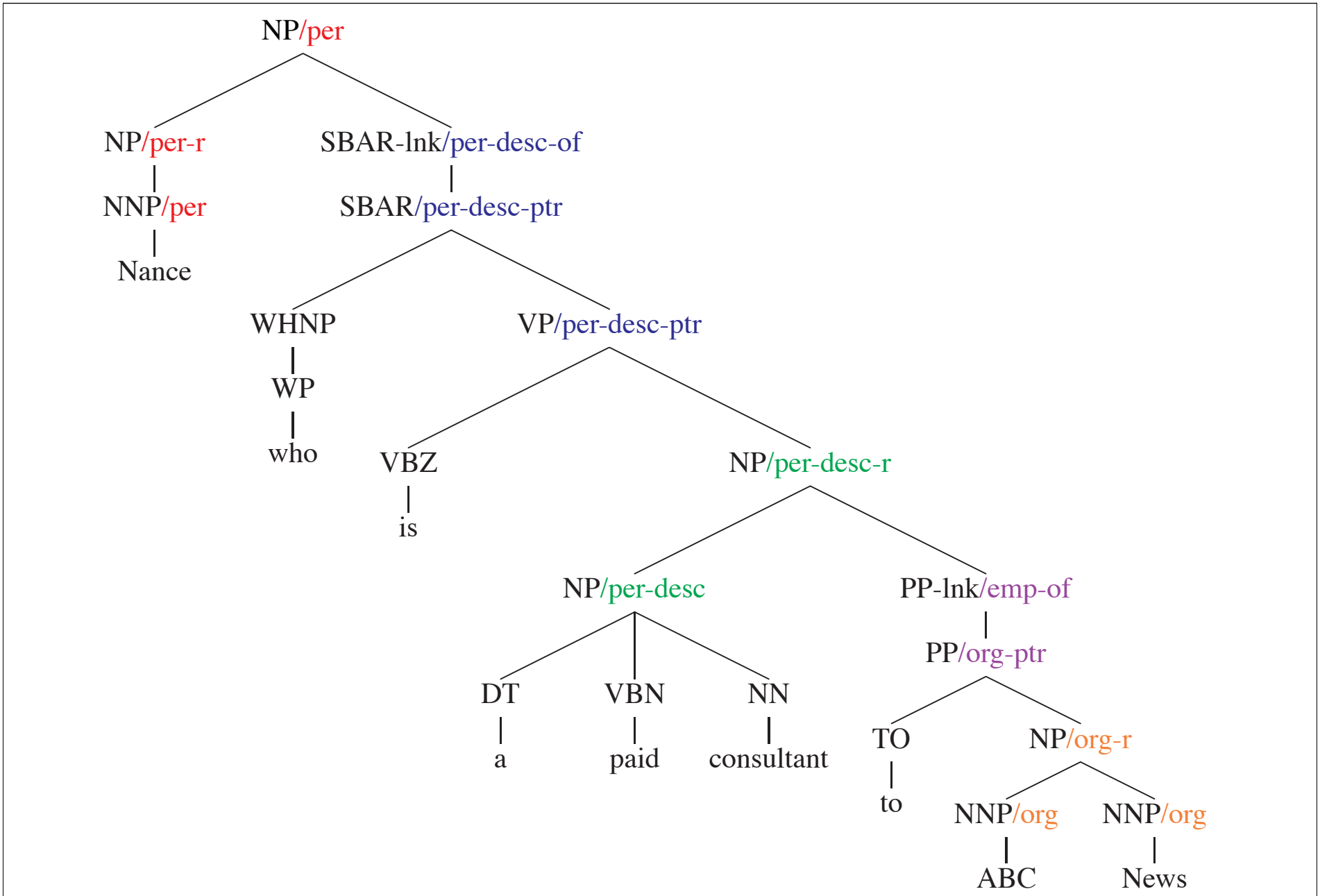
Add semantic tags showing link between "Nancy" and the descriptor

per-desc-of = person/descriptor link, per-desc-ptr = person/descriptor pointer



Add link showing employee-employer relation

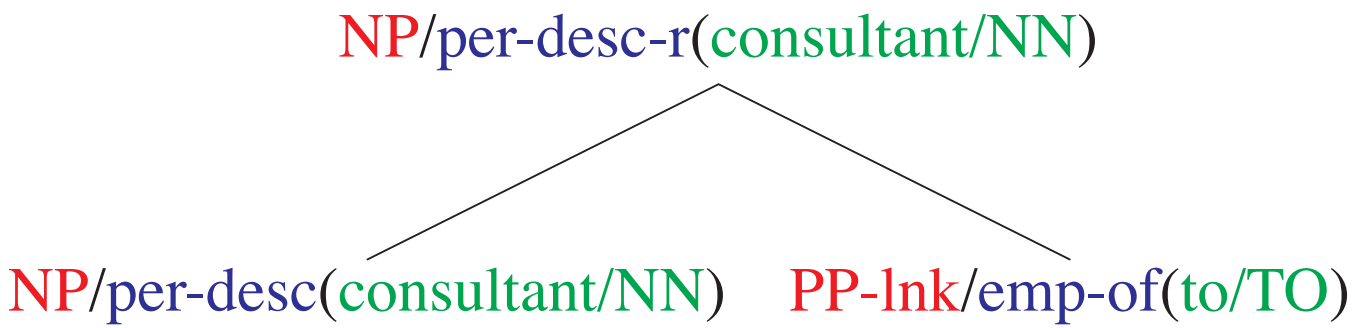
emp-of = employee-of link, emp-ptr = employee-of pointer



PERSON entity, PERSON descriptor link, DESCRIPTOR, EMPLOYER-OF relation, ORG entity

Building a Parser

- We now have context-free rules where each non-terminal in the grammar has
  - A syntactic category
  - A semantic label
  - A head-word/head-tag



- It's possible to modify syntactic parsers to estimate rule probabilities in this case

Summary

- Goal: build a parser that recovers syntactic structure, named entities, descriptors, and relations
- Annotation: mark entity boundaries, descriptor boundaries, links between entities and descriptors
- Enriched parse trees: given annotation, and a parse tree, form a new **enriched** parse tree
- The statistical model: non-terminals now include syntactic category, semantic label, head word, head tag. Rule probabilities are estimated using similar methods to syntactic parsers
- Results: precision = 81%, recall = 64% in recovering relations (employer/employee, company/product, company/headquarters-location)

Main Points

Co-reference

- How to cast as classification [Cardie]
- Measures of string similarity [Cohen]
- Scaling up [McCallum et al]

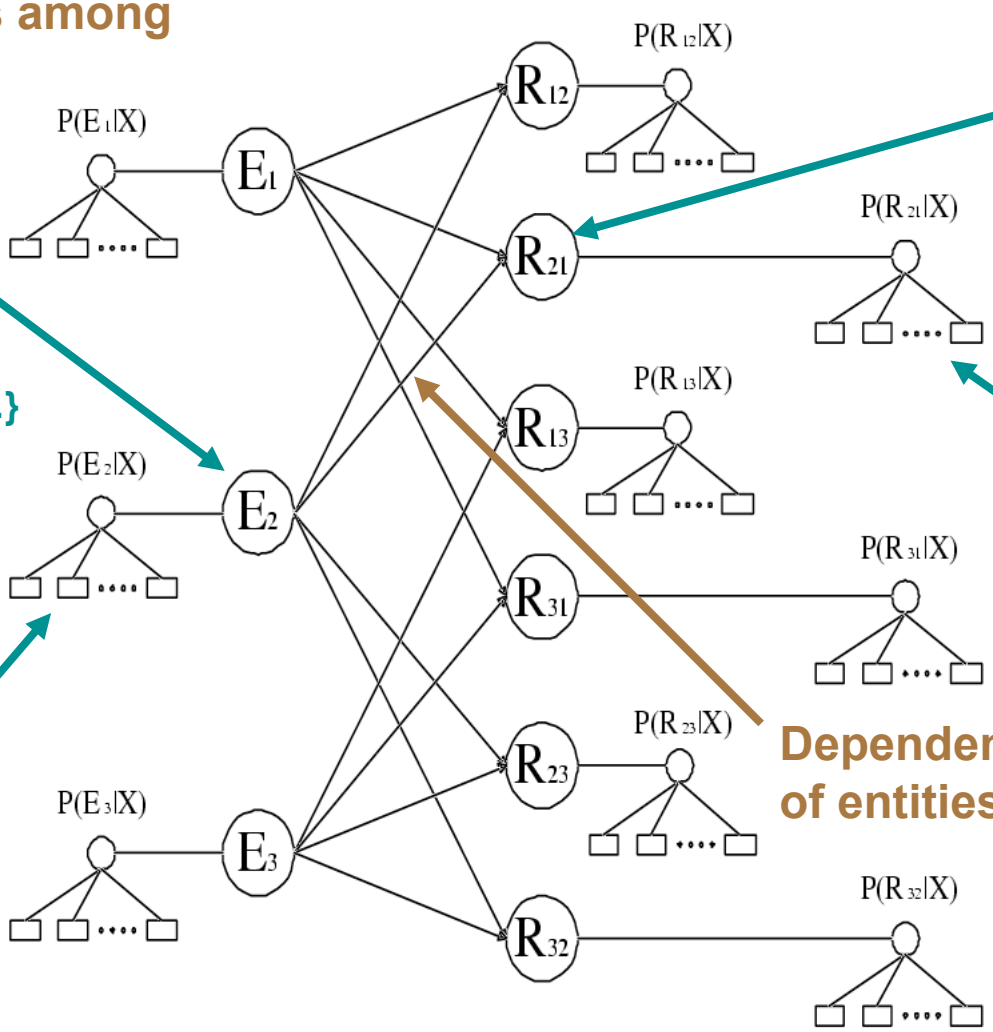
Relation extraction

- With augmented grammar [Miller et al 2000]
- **With joint inference [Roth & Yih]**
- Semi-supervised [Brin]

(1) Association with Graphical Models

[Roth & Yih 2002]

Capture arbitrary-distance dependencies among predictions.



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

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

Local language models contribute evidence to relation classification.

Local language models contribute evidence to entity classification.

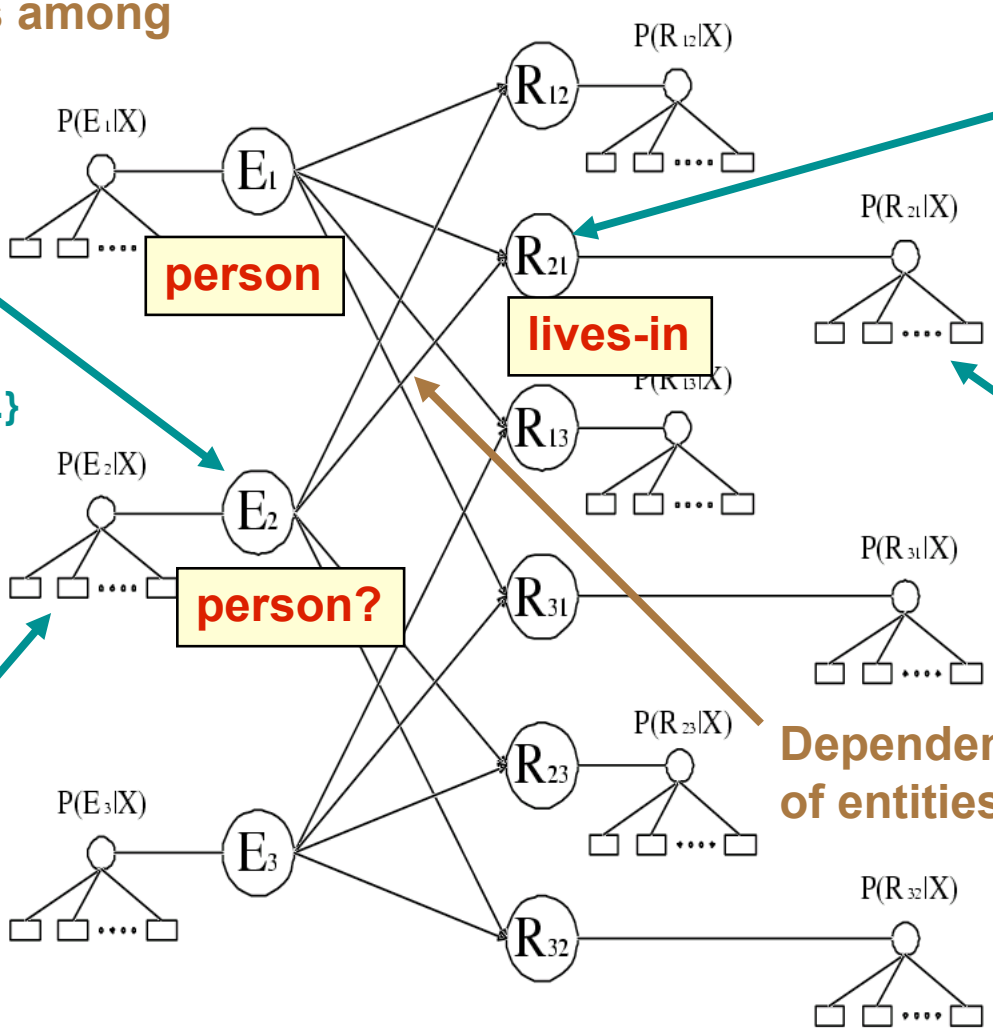
Dependencies between classes of entities and relations!

Inference with loopy belief propagation.

(1) Association with Graphical Models

[Roth & Yih 2002]

Also capture long-distance dependencies among predictions.



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

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

Local language models contribute evidence to relation classification.

Local language models contribute evidence to entity classification.

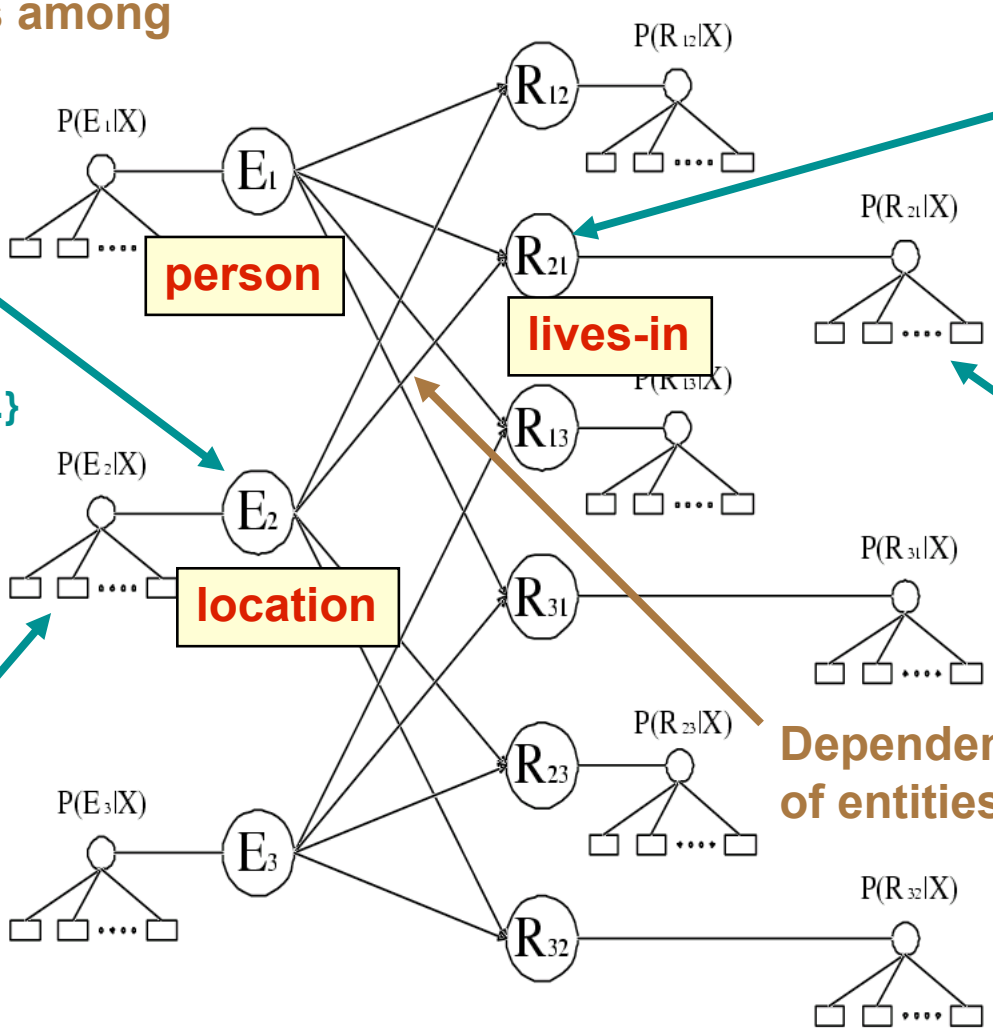
Dependencies between classes of entities and relations!

Inference with loopy belief propagation.

(1) Association with Graphical Models

[Roth & Yih 2002]

Also capture long-distance dependencies among predictions.



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

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

Local language models contribute evidence to relation classification.

Local language models contribute evidence to entity classification.

Dependencies between classes of entities and relations!

Inference with loopy belief propagation.

Main Points

Co-reference

- How to cast as classification [Cardie]
- Measures of string similarity [Cohen]
- Scaling up [McCallum et al]

Relation extraction

- With augmented grammar [Miller et al 2000]
- With joint inference [Roth & Yih]
- **Semi-supervised [Brin]**

Partially Supervised Approaches to Relation Extraction

- Last lecture: introduced a partially supervised method for named entity classification
- Basic observation: "redundancy" in that either spelling or context of an entity is often sufficient to determine its type
- Lead to *cotraining* approaches, where two classifiers bootstrap each other from a small number of seed rules
- **Can we apply these kind of methods to relation extraction?**

From [Brin, 1998]

The World Wide Web provides a vast source of information of almost all types, ranging from DNA databases to resumes to lists of favorite restaurants. However, this information is often scattered among many web servers and hosts using many different formats. If these chunks of information could be extracted from the World Wide Web and integrated into a structure form, they would form an unprecedented source of information. It would include the largest international directory of people, the largest and most diverse databases of products, the greatest bibliography of academic works, and many other useful resources.

From [Brin, 1998]

> For data we used a repository of 24 million web pages totalling 147 gigabytes. This data is part of the Stanford WebBase and is used for the Google search engine [BP], and other research projects. As a part of the search engine, we have built an inverted index of the entire repository.
>
> The repository spans many disks and several machines. It takes a considerable amount of time to make just one pass over the data even without doing any substantial processing. Therefore, in these [sic] we only made passes over subsets of the repository on any given iteration.

[BP] Sergey Brin and Larry Page. Google search engine.
`http://google.stanford.edu`

- From [Brin, 1998]:
  authors/book-titles, data = web data, seeds are

Isaac Asimov	The Robots of Dawn
David Brin	Startide Rising
James Gleik	Chaos: Making a New Science
Charles Dickens	Great Expectations
William Shakespeare	The Comedy of Errors

DIPRE [Brin, 1998]

A pattern is a 5 tuple:

- *Order*: author preceding title, or visa versa
- *URL-prefix*: a prefix of the URL of the page of the pattern
- *prefix*: up to 10 characters preceding the author/title pair
- *middle*: the characters between the author and title
- *suffix*: up to 10 characters following the author/title pair

DIPRE: Inducing Patterns from Data

- Find all instances of seeds on web pages.
  Basic question: how do we induce patterns from these examples?

- Answer = Following procedure:

  1. Group all occurrences together which have the same values for *order*, *middle*
  2. For any group: Set *url-prefix* to be longest common prefix of the group's URLs, *prefix* to be the longest common prefix of the group, *suffix* to be the longest common suffix
  3. For each group's pattern, calculate its specificity as

     $$spec(p) = n|middle||url\text{-}prefix||prefix||suffix|$$

     where n is the number of examples in the group, $|x|$ is the length of x in characters
  4. **If** specificity exceeds some threshold, include the pattern
  5. **Else If** all patterns occur on same webpage, reject the pattern
  6. **Else** create new sub-groups grouped by characters in the urls which is one past *url-prefix*, and repeat the procedure in step 2 for these new sub-groups.

The Overall Algorithm

1. Use the seed examples to label some data

2. Induce patterns from the labeled examples, using method described on the previous slide

3. Apply the patterns to data, to get a new set of author/title pairs

4. Return to step 2, and iterate

DIPRE: Inducing Patterns from Data

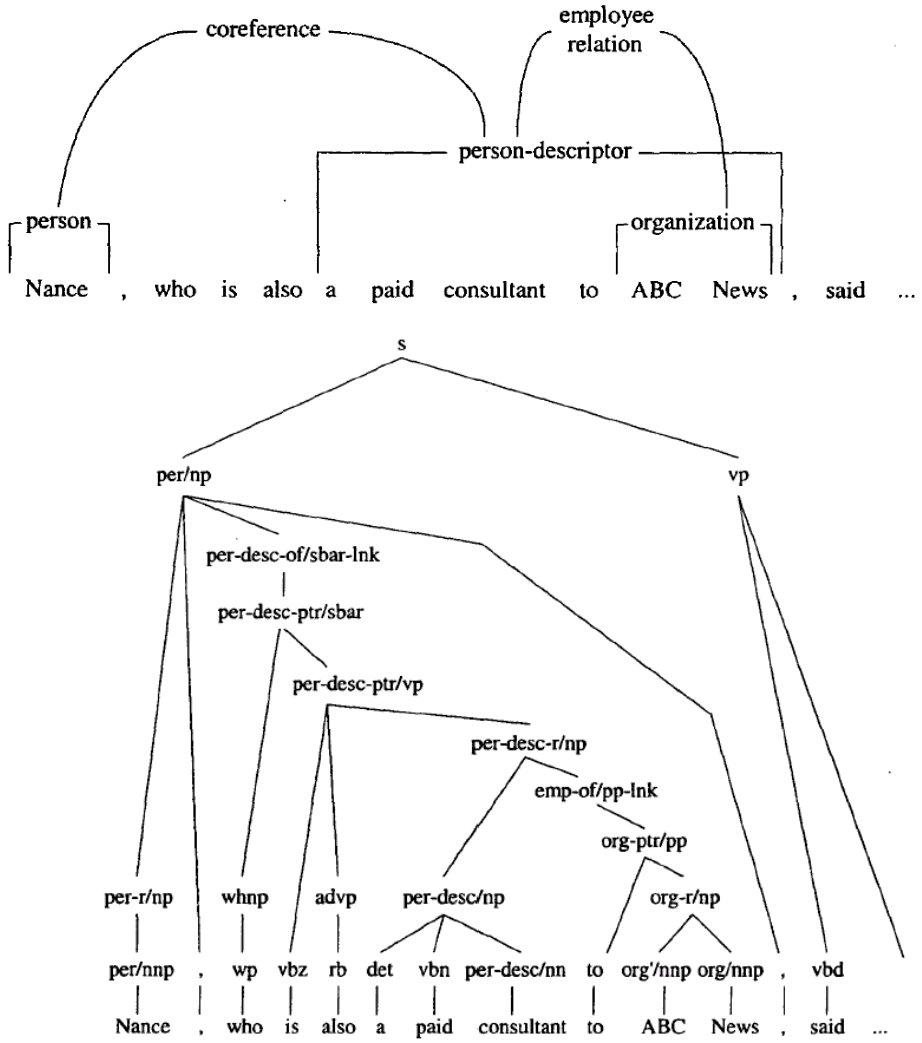
The patterns found in the first iteration:

`www.sff.net/locus/c.*`	<LI><B>*title*</B> by *author* (
`dns.city-net.com/lmann/awards/hugos/1984.html`	*title* by *author* (
`dolphin-upenn.edu/dcummins/texts/sf-award.htm`	*author* || *title* || (

- The 5 seeds produced 199 labeled instances, giving the 3 patterns above
- Applying the three patterns gave 4047 new book instances
- Searching 5 million web pages gave 3972 occurrences of these books
- This gave 105 patterns, 24 applied to more than one URL
- Applied to 2 million URLS produced 9369 unique (author,title) pairs
- Manual intervention: removed 242 "bogus" items where the author was "Conclusion"
- Final iteration: ran over 156,000 documents which contained the word "books"; induced 346 patterns, 15,257 (author,title) pairs

(1) Association using Parse Tree

Simultaneously POS tag, parse, extract & associate! *[Miller et al 2000]*



Increase space of parse constitutes to include entity and relation tags

Notation	Description
c_h	head constituent category
c_m	modifier constituent category
X_p	X of parent node
t	POS tag
w	word

Parameters	e.g.
$P(c_h \mid c_p)$	P(vp|s)
$P(c_m \mid c_p, c_{hp}, c_{m-1}, w_p)$	P(per/np|s,vp,null,said)
$P(t_m \mid c_m, t_h, w_h)$	P(per/nnp|per/np,vbd,
$P(w_m \mid c_m, t_m, t_h, w_h)$	P(nance|per/np,per/nnp,vbd,said)

(This is also a great example of extraction using a tree model.)