LOT Winter School 2005

scourse Anaphora: Theories, Data and Applications

Bonnie Webber
School of Informatics
University of Edinburgh
January 2005

4: Corpus-based Approaches to Anaphor Resolution

Day 4: Corpus-based Approaches to Anaphor Resolution

- Statistical & Machine-learning for Coreference
- Machine-learning for Resolving Pronouns in Dialogue
- Machine-learning for Comparative Anaphora
- Summary

Background

the early 90s, *Information Extraction* was a major stimulus to developments nguage Technology (MUC challenge). It involves automatically identifying xtracting from text, information relevant to filling in a structured, labelled ate.

nple:

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs o be shipped to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with roduction of 20,000 iron and "metal wood" clubs a month.

Suppose the template to be filled in is:

TIE-UP-1:

Relationship:	Tie-Up
Entities:	
Joint Venture Company:	
Activity:	
Amount:	

Recognising that what should be extracted for the "entities" being "tied up" are: "Bridgestone Sports Co", "a local concern" and "a Japanese trading house" requires

- knowing that it should be the arguments to "set up a joint venture";
- resolving "it".

N.B. Since this is *extraction*, no attempt was made to associate "local" with "Taiwan" – an interesting but separate problem.

TIE-UP-1:

Relationship:	Tie-Up
Entities:	"Bridgestone Sports Co", "a local concern", "a Japanese trading house"
Joint Venture Company:	
Activity:	
Amount:	

gnising that what should be extracted for the Joint Venture Company is "Bridgestone s Taiwan Co." and filling in the rest of the template, requires resolving the anaphoric te NP "The joint venture" with its antecedent "a joint venture".

TIE-UP-1:

Relationship:	Tie-Up
Entities:	"Bridgestone Sports Co", "a local concern", "a Japanese trading house"
Joint Venture Company:	"Bridgestone Sports Taiwan Co."
Activity:	ACTIVITY-1
Amount:	NT$20000000

ACTIVITY-1:

Company:	"Bridgestone Sports Taiwan Co."
Product:	"iron and 'metal wood' clubs"
Start date:	DURING: January 1990

e needs of IE led to an equivalent *coreference resolution* challenge aimed at mining whether two NL expressions *markables* corefer – e.g.,

Bridgestone Sports Co" – "it"
a joint venture" – "the joint venture" – "Bridgestone Sports Taiwan Co."

eeds of IE meant that *markables* were defined to include: definite NPs, nstrative NPs, proper names, appositives, predicate NPs, "sub-NP modifiers" ronouns – e.g.,

a. Higgins, *formerly sales director of Sudsy Soaps*, became *the president of Dreamy Detergents*.

b. The price of *aluminum* siding has steadily increased, as the market for *aluminum* reacts to the strike in Chile.

we saw from [van Deemter & Kibble 2001], this blurs the distinction een *anaphora resolution* and *coreference resolution*, but it has fueled the bulk rpus-based approaches to coreference resolution.

Topics to be covered

Both *anaphora resolution* and *coreference resolution* use methods based on:

- statistics
- classifiers induced through Machine Learning

So we will look at:

1. A statistical approach to resolving 3^{rd} person pronouns: [Ge, Hale & Charniak]
2. Two ML approaches to coreference resolution: [Soon, Lim & Ng], [Ng & Cardie]
3. A ML approach to resolving pronouns in dialogue [Strube & Müller]
4. A ML approach to resolving *comparative anaphora* [Modjeska, Markert & Nissim 03]

Statistical Resolution of 3^{rd} person Pronouns

Hale & Charniak 98] base their approach on statistics characterising relations
' person pronouns (*he*, *she*, *it*, *they* and their variants) and their antecedents in
en text.

acterised in terms of:

Distance between pronoun and potential antecedent (recency preference);

Syntactic context of the pronoun;

Words in the potential antecedent (for deriving likely gender, number and
nimacy, and supporting a preference for agreement w.r.t. these features);

Number of times the referent of the antecedent has been mentioned,
upporting a preference for repeatedly mentioned NPs.

Given a pronoun ρ, choose as its referent

$$\arg\max_a P(A(\rho)=a \mid \rho, h, \vec{W}, t, l, s_\rho, \vec{d}, \vec{M})$$

$A(\rho)$ random variable denoting the referent of ρ

a potential antecedent

h head constituent above ρ

$\vec{W}$ full list of potential antecedents

t type of phrase of the potential antecedent (always NP here)

l type of head constituent

$\mathbf{s}_\rho$ syntactic structure in which ρ appears

$\vec{d}$ distance of each potential antecedent from ρ

$\vec{M}$ number of times the referent of each potential antecedent is mentioned.

mposing $P(A(\rho)=a \mid \rho, h, \vec{W}, t, l, s_\rho, \vec{d}, \vec{M})$, using Bayes theorem and some
endence assumptions yields

$$P(a \mid \vec{M}) * P(s_\rho, \vec{d} \mid a) * P(\vec{W} \mid h,t,l,a) * P(\rho \mid w_a)$$

e w_a is the a^{th} candidate in $\vec{W}$.

er assumptions and rewriting yields

$$P(A(\rho)=w_a) \propto P(s_\rho, d_a \mid a) * P(\rho \mid w_a) * \frac{P(w_a \mid h,t,l)}{P(w_a \mid t)} * P(a \mid m_a)$$

e m_a is the number of times the referent of a is mentioned.

or 1 relates to both the syntactic position of the anaphor and its distance from
otential antecedent. Ge et al. treat it as a *Hobbs distance*, d_H as follows:

Use the Hobbs' algorithm to successively number the NPs reached on its
earch for an antecedent. Then

$$P(d_H=i \mid a) = \frac{|correct\ antecedents\ at\ Hobbs\ distance\ i|}{|correct\ antecedents|}$$

Factor 2 requires getting counts on gender/animacy.

$$P(\rho \mid w_a) = \frac{|w_a\ in\ the\ antecedent\ for\ \rho|}{|w_a|}$$

Factor 3 reflects selectional restrictions, and use frequencies collected when building a statistical parser for the Penn TreeBank.

Factor 4 is a *mention count* statistic, for which they also consider where in the paragraph the mention is made.

$$P(a \mid m_a, j) = \frac{|a\ is\ antecedent, m_a, j|}{|m_a, j|}$$

Where to get the other frequencies needed to derive these probabilities?

⇒ Manually annotated Gold Standard

How to ensure that one has all the probabilities one needs?

⇒ Smoothing

Evaluation: Statistical Pronoun Resolution

:

371 instances of *he*, *she* and *it* and their variants in 3975 sentences (about 00 news stories);

Manually excluding all instances of *it* that are purely syntactic.

ach pronoun, they collected candidate antecedents at *Hobbs distances* 1-15, omputed the probability of each pair. That candidate was chosen that mises the computed probability.

Evaluation used 10-fold cross validation.

Results: 82.9% correct. (This isn't a binary test, so no *recall* or *precision*.)

Found gender/animacy to be a big factor: If this decision were perfect, results would increase to 89.3%. (Rest of paper records work on probability model or gender/animacy. See [Zaenen et al 2004] for corpus-based work on determining animacy.)

ML Approaches to Coreference Resolution

Training Phase: Learn a classifier from a generated set of positive and negative examples, each example being a *pair* consisting of:

- a pronoun or other potential coferential "markable";
- a potential antecedent.

Evaluation Phase: For each pronoun or other "markable",

- identify potential antecedents and create a set of pairs;
- apply classifier to each;
- decide what to do if classifier accepts 0, 1 or >1 pairs for that markable.

Application Phase

w do ML approaches to Coreference Resolution Differ?

Space of positive/negative examples used in training

– What markables found and considered for coreference;

– What potential antecedents are considered.

Features associated with anaphor/antecedent pairs, and the tools used to derive those features;

Type of classifier learned (e.g., decision tree, Naive Bayes classifiers, maximum entropy classifiers, support vector machine classifiers);

Character of the decision process in which the classifier(s) is/are embedded;

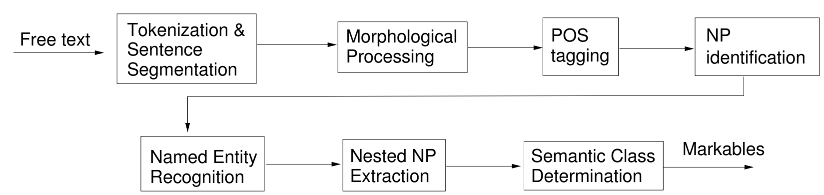
Space of test data;

Evaluation procedure.

Automatic Identification of Markables [Soon et al]

While *markables* are hand-annotated in training data, they must be identified automatically in test data (**evaluation**) and in any subsequent **application**.

This is a *boundary detection* task, also requiring identification of markables *embedded* in other markables.



[Soon et al. 2001] use HMMS for POS tagging, (preliminary) NP identification and NER, adjusting boundaries of NPs (when necessary) to conform to recognised named entities.

Nested NP Extraction [Soon et al]

ed **possessive NPs**: Probably detected by presence of possessive pronoun
"his", "its") or possessive marker somewhere in a *markable*, but Soon et al.
say how boundary of the *nested* NP is detected.

with possessive pronouns or possessive markers on Named Entities – e.g.

their long-range strategy) → ((their) long-range strategy)
Eastern ' s parent) → ((Eastern) ' s parent)

mbiguity in other cases – e.g.

A country ' s rule book) should express (its people ' s preferences).
→ ?? ((A country) ' s rule book) should express (((its) people) ' s
references).
→ ?? (A (country) ' s rule book) should express ((its) ((people) ' s
references)).

Positive and Negative Training Examples [Soon et al]

Positive Examples:

Given a *coreference chain* A1 ← A2 ← A3 ← A4, create a positive example from each adjacent pair of markables.

A1 – A2   A2 – A3   A3 – A4

Call the first member of the pair the *antecedent*, the second, the *anaphor*.

Negative Examples:

For each markable b, c, d **in between** an adjacent pair $A_{j-1} - A_j$ on a *coreference chain*, create a negative example consisting of the in-between markable and the *anaphor*.

$b - A_j \quad c - A_j \quad d - A_j$

Q? Do you see any potential problem with this choice of negative examples, for when the classifier would later be used in a real application?

eatures associated with +/- Examples [Soon et al]

eatures hold of either the *anaphor* (**anaph**), the potential *antecedent* (**ante**) or
air (**pair**).

Distance feature: 0 (same S), 1 (previous S), 2, 3, ... (**pair**)

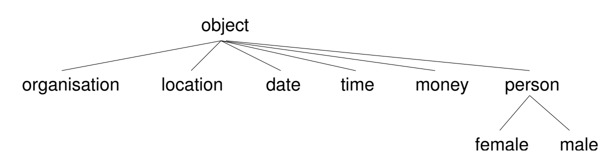
-Pronoun: true if *antecedent* is a pronoun, false otherwise. (**ante**)

-Pronoun: true if *anaphor* is a pronoun, false otherwise. (**anaph**)

tring match: true if *antecedent* and *anaphor* are the same word sequence
ignoring determiners), false otherwise. (**pair**)

Definite NP: true if *anaphor* is a definite NP, false otherwise. (**anaph**)

6. Demonstrative NP: true if *anaphor* is an NP that starts with a demonstrative (this/that/these/those), false otherwise. (**anaph**)
7. Number agreement: true if *antecedent* and *anaphor* agree in number – both singular or both plural. (**pair**)
8. Semantic class agreement: true if the most frequent sense of the *antecedent* and of the *anaphor* agree with respect to a small set of semantic classes organised into a simple ISA hierarchy. Unknown if the semantic class of one of them is unknown. Otherwise false. (**pair**)



Gender agreement: true if semantic class of both *antecedent* and *anaphor* are not person, or if both are person and male or person and female, or one is person and the other is male or female. Unknown if the semantic class of either is unknown. Otherwise false. (**pair**)

Both proper names: true if both *antecedent* and *anaphor* are proper names. Otherwise false. (**pair**)

Alias: true if both *antecedent* and *anaphor* are Named Entities and one is an alias of the other – e.g., different format for dates or names, acronyms for companies. False otherwise (**pair**)

Appositive: true if the *anaphor* is an appositive with respect to the *antecedent*. False otherwise. (**pair**)

of these features are defined such that they can be determined automatically 100% accuracy.

Data & Machine Learning Method [Soon et al]

30 MUC-6 documents and 30 MUC-7 documents, both annotated with coreference, yielded

- 20,910 training examples, of which 6.5% were positive (MUC-6)
- 48,872 training examples, of which 4.4% were positive (MUC-7)

ML Method: C5.0 Decision Tree Classifier

Classifier parameters set using 10-fold cross-validation:

- 20% pruning confidence, and at least 5 instances per leaf node (MUC-6)
- 60% pruning confidence, and at least 2 instances per leaf node (MUC-7)

When applied to the test data, classifer makes the decision **+coref** or **-coref**.

Evaluation Results

MUC-6: recall 58.6%, precision 67.3%, F-measure 62.6%

MUC-7: recall 56.1%, precision 65.5%, F-measure 60.4%

They don't say how the **test set** examples were generated, since one should know what the correct antecedent is when creating the test set.

If test set examples have a **different distribution** than those in the training set, could effect the quality of the classifier w.r.t. the test set.

importantly, while [Soon et al] carry out a detailed error analysis, they give results per anaphor, distinguishing

No *antecedent* is found for an *anaphor*

N>1 *antecedents* are found for an *anaphor*

One *antecedent* found but the wrong one 1 false positive)

Cases that are not distinguished

Case 1

- Anaphor 1: 1 TP, 2 FP, 2 TN
- Anaphor 2: 1 TP, 2 FP, 2 TN

	Label	
Actual	A	$\bar{A}$
A	2	0
$\bar{A}$	4	4

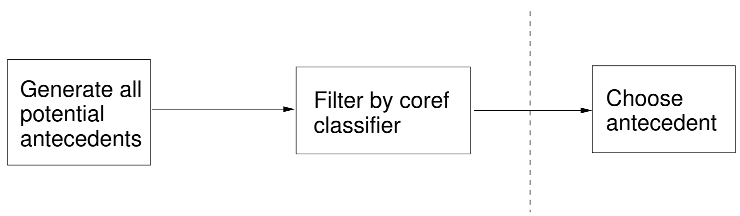
precision: 33% *recall*: 100%

Case 2

- Anaphor 1: 1 TP, 0 FP, 2 TN
- Anaphor 2: 1 TP, 4 FP, 2 TN

While the confusion matrix is the same, in Case 2, we know the antecedent of Anaphor 1.

Schematic Diagram [Soon et al 01]



[Ng & Cardie 02] extends [Soon et al 01]

[Ng & Cardie] follow similar procedure as [Soon et al] for creating +/- training examples, with one exception:

> For *non-pronominal* NPs in coreference chains, choose as + example, the closest non-pronoun preceding antecedent (*most confident antecedent*).

Their training and test sets are derived from the same MUC-6 and MUC-7 texts.

A decision tree (DT) is induced using C4.5.

However, decisions using this DT are not made *directly*, based on the +/- label on the leaf. Rather, a leaf is labelled with ratio

$$\frac{positive\ instances\ at\ leaf + 1}{all\ instances\ at\ leaf + 2}$$

and only assigns **+coref** if ratio > 0.5

ɛ Cardie] make two other extra-linguistic modifications to the ML framework:

ielect "most likely" antecedent from among those above 0.5 threshhold.

ịplit single "string match" feature into ones specific for pronouns, proper
ıames and non-pro NPs.

also consider 41 additional features involving:

nore complex string matching

rammatical role

greement and binding constraints

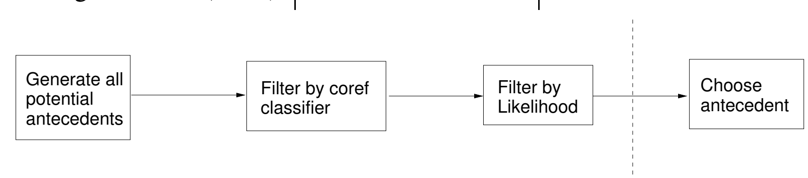
dditional preferences

votes" from two naive pronoun resolution algorithms

Results [Ng & Cardie 02]

Changing the ML framework by using "most confident" antecedent in training data, choosing "most likely" antecedent in classification, and splitting string matching improved results over Soon et al. (statistically significant).

	MUC-6			MUC-7		
	R	P	F	R	P	F
Soon et al.	58.6	67.3	62.6	56.1	65.5	60.4
Ng & Cardie (basic)	62.4	73.5	67.5	56.3	71.5	63.0



n Ng & Cardie used **all** their additional features, R **increased** (across the
) and P, for names. More drastically, P **decreased** for pronouns and non-pro
despite DT algorithm being able to choose what features to use.

	MUC-6			MUC-7		
	R	P	F	R	P	F
Ng & Cardie (basic)	62.4	73.5	67.5	56.3	71.5	63.0
all features	70.3	58.3	63.8	65.5	58.2	61.6
pronouns only	-	66.3	-	-	62.1	-
names only	-	84.2	-	-	77.7	-
common nouns only	-	40.1	-	-	45.2	-

e F = $\frac{2*R*P}{R+P}$ (the *harmonic mean*).

They also tried discarding features manually detected as being low-P, producing better results, suggesting this be automated.

	MUC-6			MUC-7		
	R	P	F	R	P	F
ML framework	62.4	73.5	67.5	56.3	71.5	63.0
all features	70.3	58.3	63.8	65.5	58.2	61.6
pronouns only	-	66.3	-	-	62.1	-
names only	-	84.2	-	-	77.7	-
common nouns only	-	40.1	-	-	45.2	-
hand-selected features	64.1	74.9	69.1	57.4	70.8	63.4
pronouns only	-	67.4	-	-	54.4	-
names only	-	93.3	-	-	86.6	-
common nouns only	-	63.0	-	-	64.8	-

ML Approach to Resolving Pronouns in Dialogue

l Eckert & Strube's observation that about 33% of pronouns in the
chBoard (dialogue) corpus lacked an antecedent and about 22% of them had
edents from units other than pronouns.

be & Müller 2003] attempt to deal with this in an approach based on machine
ng.

lem: Identifying antecedents for both 3^{rd} person and demonstrative
uns, where a pronoun may have

o antecedent

vague antecedent

n NP antecedent (coreference)

clausal antecedent (event/AO reference)

Data:

- 20 randomly chosen Switchboard dialogues: 3275 sentences / 1771 turns, 16601 markables
- *NP markables*: referring expressions (NPs, pronouns, proper names)
- *VP markables*: verb phrases
- *S markables*: sentences
- *Disfluency markables*: NPs or pronouns in unfinished or abandoned utterances
- Among features associated with each markable is ID for its *coreference class*, or NULL if it's an *isolated mention*.

Form of NP-markables:

	defNP	**indefNP**	**NNP**	**prp**	**prp$**	**dtpro**
Total	1080	1899	217	1075	70	392
In coreference rel	219	163	94	786	56	309
Singletons	861	1736	123	289	14	83

ber/gender features on pronoun markables:

	3m	3f	3n	3p	Total
prp	67 63	49 47	541 318	418 358	1075 786
prp$	18 15	14 11	3 3	35 27	70 56
dtpro	0 0	0 0	380 298	12 11	392 309
Total	85 78	63 58	924 619	465 396	

Training and Test Examples (Pairs)

1. Markables for 1^{st} and 2^{nd} person pronouns eliminated.

2. For each NP-markable not an indefinite NP, generate a set of pairs of potentially coreferring expressions by combining the markable (considered a potential anaphor) with each *compatible* markable preceding it. The example was labelled *positive* if the two belonged to the same coreference class, *negative* otherwise.

N.B. This allows there to be >1 antecedent for an anaphor.

3. For each NP-markable for *it* and *that*, generate additional pairs for event reference, consisting of the potential anaphor and each S- and VP-markable from the last two preceding valid sentences. Label the pairs *positive* and *negative* as above.

Features:

- NP-level features for each member of the pair
- Coreference features for the pair

features tailored to spoken dialogue, some to identified VP argument
rences for NP vs. non-NP fillers (but possibly specific to the 553
hBoard dialogues).

Results

Masculine and feminine pronouns (3mf)

	correct found	total found	total correct	P	R	F
baseline: basic features	120	150	1250	80%	9.6%	17.14%
+ mdist_3mf3p	121	153	1250	79.08%	9.68%	17.25%

Neuter pronouns (3n)

	correct found	total found	total correct	P	R	F
baseline: basic features	109	235	1250	46.38	8.72	14.68
+ none	97	232	1250	41.81	7.76	13.09
+ ante_exp_typ	137	359	1250	38.16	10.96	17.03
+ wdist_ic	154	389	1250	39.59	12.32	18.79
+ ante_tfidf	158	391	1250	40.41	12.64	19.26

bined Results for all pronouns

	correct found	total found	total correct	P	R	F
line: basic features	456	739	1250	61.71	36.48	45.85
bined	509	897	1250	56.74	40.72	47.42

ML Approach to 'Other'-NPs [Modjeska et al 03]

We'll look at resolving a non-pronominal anaphor in order to illustrate:

- Similar +/- example and feature selection applied to a different problem;
- Use of another web frequency feature.

Other anaphors are referential NPs modified by "other" or "another", having *non-structural antecedents*:

(2) An exhibition of American design and architecture opened in September in *Moscow* and will travel to **eight other Soviet cities**.

(3) You either believe Seymour can do it again or you don't. Beside *the designer's age*, **other risk factors for Mr. Cray's company** include Cray-3's [. . .] chip technology.

ally excluded from this set were features that can probably be excluded
atically if applied to parsed text:

other than' constructions, where a *structural* element is excluded:

4) Few clients other than the state . . . .

other'-NPs in *list contexts* since 99% of time, the antecedent is the preceding
ist element (though not necessarily):

5) Research shows AZT can relieve *dementia* and **other symptoms** . . . .

6) *One aardvark* stayed by the water hole. The giraffes and **the other ardvaarks** scattered.

diomatic phrases such as "on the other hand";

eciprocals ("each other", "one another");

the other(s)", "the other one(s)", "another one", which **probably** behave like
naphoric pronouns (though no **definitive** study yet of *one anaphora*).

other'-NPs with non-NP antecedents. (Can't be easily excluded
utomatically, though again, no **definitive** study.)

Training/Test Data [Modjeska et al 03]

500 'other'-NPs were extracted from Penn TreeBank *Wall Street Journal* (PTB WSJ) corpus, along with a 5-sentence preceding context.

All *base NPs* in context were tagged, and NPs with possessive NPs were split:

(7) [An exhibition] of [American design] and [architecture] opened in [September] in [Moscow] and will travel to **eight other Soviet cities**.

(8) You either believe [Seymour] can do it again or you don't. Beside [the designer] ' s [age], **other risk factors for Mr. Cray's company** include Cray-3's chip technology.

- +examples: anaphor and closest preceding antecedent (500 pairs, 16%)
- -examples: anaphor and each base NP between anaphor and antecedent (2584 pairs, 84%)

Features [Modjeska et al 03]

NP_FORM (NP Surface Form): def, indef, dem, pro, name,?
Feature of each NP, derived from POS-tags and use of heuristics.

RESTR_SUBSTR (Antecedent contains anaphor?): yes,no
Feature of pair (each lemmatised), derived with Perl script.

GRAM_FUNC (grammatical role): subj, predNP, IndObj, Obj, Oblique, ?
Feature of each NP, derived from PTB.

SYN_PAR (Anaphor and antecedent have same gram fn?): yes,no
Feature of pair, derived from PTB.

SDIST (anaphor-antecedent distance in Ss): 1, 2, 3, 4, 5
Feature of pair, derived with Perl script.

6. SEMCLASS (semantic class): person, org, loc, date, money, number, thing, abst, ?
   Feature of each NP, derived using GATE2 and WordNet.
7. SEMCLASS_AGR (Anaphor and antecedent agree on SEMCLASS?): yes,no,?
   Feature of pair, derived from SEMCLASS.
8. GENDER_AGR (Anaphor and antecedent agree on gender?): same, compat, incompat, ?
   Feature of pair, derived using lexical resources.
9. RELATION (anaph-ante relation): same-pred, hypernym, meronym, compat, incompat, ?
   Feature of pair, derived using Perl scripts and WordNet

Performance using Naive Bayes Classifier

baseline to be closest possible antecedent.

	P	R	F
baseline	27.8	27.8	27.8
F1	51.7	40.6	45.5

st class of errors due to insufficient semantic knowledge (e.g. gaps in Net, general world knowledge, bridging)

While Mr. Dallara and Japanese say the question of investors' access to the US and Japanese markets may get a disproportionate share of the public's attention, a number of **other important issues** will be on the table at next week's talks.

The Justice Department's view is share by **other lawyers**.

Web Frequency of Patterns indicating Semantic Relationships

Basic Idea: The web implicitly contains alot of information on semantic relations in the form of *lexical patterns*.

Even though these relations may be *under-specified*, but they can still be used to *discriminate* cases where a relation holds from cases where one doesn't.

E.g. *X(s) and other Ys*

age and other risk factors ??designer and other risk factors

E.g. *Ys such as X(s)*

??issues such as attention issues such as access

E.g. *XY* (*PP-attachment*: "as cases in centrifuges for rotors")

rotor cases ??rotor centrifuges

ach anaphor (N2) and head N of each possible antecedent (N1), Google is
he pattern:

```
N1{sg} OR N1{pl}) and other N2{pl}
```

ing a frequency count: M_{ant}

Named Entity possible antecedents are treated specially, given *sparse data*
em.)

le frequency counts also gotten for:

N1{sg} OR N1{pl} (call this M_{N1})

N2{pl} (call this M_{N2})

al Information (MI) is then computed: $\log \frac{M_{ant} * GP}{M_{N1} * M_{N2}}$
e *GP* is number of Google pages.

n the set of possible antecedents for an anaphor, the one with the highest MI
igned value +*first* for web feature. Other antecedents were assigned value
.

Re-train Naive Bayes classifier using all features, including web.

	P	R	F
baseline	27.8	27.8	27.8
F1	51.7	40.6	45.5
F1+web	**60.8**	**53.4**	**56.9**

Clear Improvement.

Remaining Issues

em 1: The large number of Named Entity antecedents (39.6%). Errors in NE
ing can lead to mis-judgments in classification.

em 2: Method fails when only correct antecedent is a pronoun, since it lacks
aphor resolution component for pronouns.

em 3: Only using *head* of NP, not all of it, to avoid *data sparsity* problem.
hat may miss the relevant semantically predictive element.

ly, these results are just for the classifier. It still needs to be incorporated in
hor resolution procedure for *other*-NPs, since classification may still result in
r more possible antecedents for each anaphor.

Summary

Statistical and ML approaches are being developed for resolving anaphoric pronouns and establishing coreference. They are also being extended to other types of anaphors.

In general, they all work by re-interpreting the problem as one of classifying pairs of anaphor and possible antecedent into two classes: +/-coref.

Thus they act as *filters*, yielding 0, 1 or more pairs for each anaphor.

Unless possible antecedents may themselves be +coref, a *resolution process* is still needed, either to choose among remaining candidates or to suggest an alternative if no candidates pass the filter.

R&D question: Where are the big performance gains going to come from?