

Introduction to Coreference Resolution

Nathan Alan Gilbert

January 24, 2007

Natural Language Processing Group
University of Utah

Extra: <http://www.cs.utah.edu/~ngilbert/nlp/errata1.html>

Motivation

- New project in the area.
- Coreference Resolution is important.
 - Many other areas of NLP depend on Coref.
 - IE, IR, Question answering systems, Summarization, etc.
 - In other words, improving performance here could help in other areas.
- A lot of work still left to do.
 - Low hanging fruit?

What is Coreference Resolution?

- *Two textual entities that refer to the same object in “the real world.”* - Mitkov
 - **More Formal Definition:**
 - α_1 and α_2 corefer if and only if $\text{Referent}(\alpha_1) = \text{Referent}(\alpha_2)$.
 - α_n are noun phrases.
 - $\text{Referent}(\alpha)$ is 'the entity referred to by α '
 - Equivalence Relation: reflexive, symmetric and transitive.
-
-

Coreference Resolution vs Anaphora Resolution

- Anaphora \neq Coreference:
 - α_1 can take α_2 as its anaphoric antecedent if and only if α_1 depends on α_2 for interpretation.
 - Anaphora is irreflexive, nonsymmetrical, and nontransitive.
 - Examples: “President Clinton” and “Hillary Rodham-Clinton's husband.”
 - Bound Anaphor:
 - “Every TV Network reported its profits.”
 - Deemter & Kibble, “On Coreferring” (Reference 3.)
-
-

(Short) History of Research

- ...from the ML perspective.
 - Aone & Bennet: (1993) Decision Tree based trained on annotated Japanese news articles. Specifically tuned to zero-anaphora.
 - M^CCarthy & Lehnert and RESOLVE. (1995) Another decision tree approach, this time in the domain of business joint ventures.
 - Feature set: {Name, Joint Venture child, Alias, Common NP, Same sentence}
-
-

Two Newer Approaches

1) Soon et al.

- 1) First Machine Learning (ML) technique to “make good.”
- 2) Why?

2) Ng and Cardie

- 1) An improvement upon the previous algorithm.
 - 2) How and why was it better?
-
-

A Machine Learning Approach to Coreference Resolution of Noun Phrases

- Developed by Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim.
 - Corpus-based, supervised learning algorithm.
 - *Proof of concept algorithm.*
 - First ML approach to do as well or better than non-ML algorithms.
 - Defines coreference the same as the MUC-6 and MUC-7 conferences.
 - Cited by 121 (according to Google Scholar)
-
-

How MUC defines Coreference

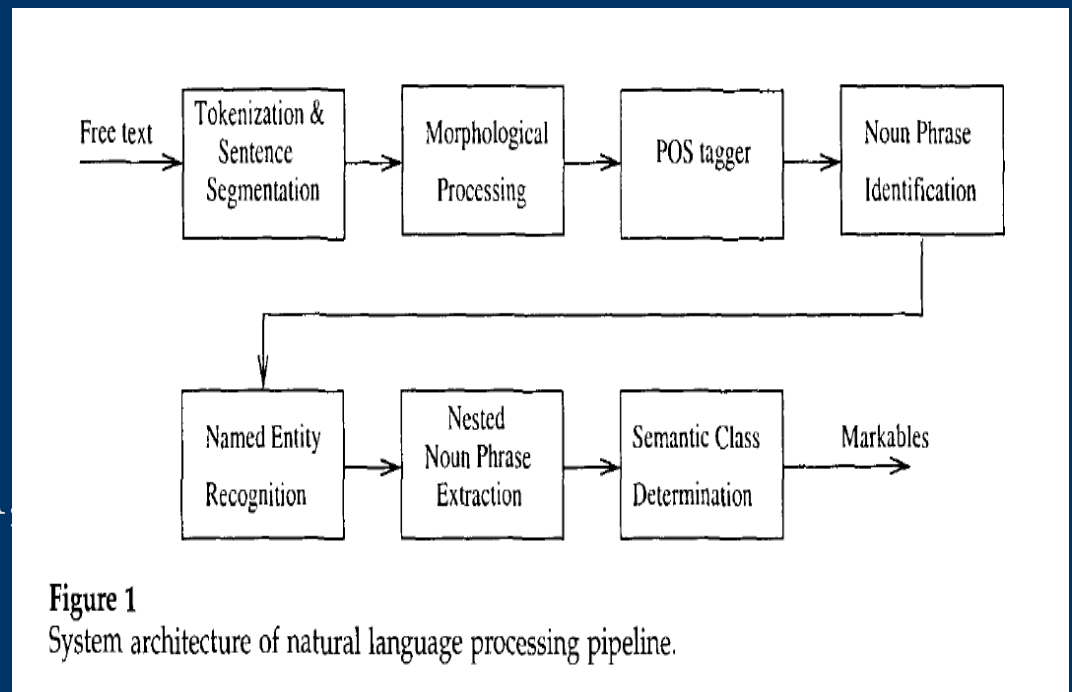
- Coreference relation:
 - An *identity-of-reference* relation between two textual elements known as *Markables*.
 - Markables can consist of definite noun phrases, demonstrative NP, proper names, and appositives.
 - Looser definition, perhaps too much?
-
-

Locating Markables

- A good system finds them all.
 - Pipelined approach
 - Tokenizer, sentence splitter, morphological/lexical processing, POS tagger, NP location, NER, nested-NP ID, and semantic classifier.
 - All this, to determine the boundaries of markables.
 - Also generates information for features in the training examples.
-
-

NLP Tools

- Mostly HMM that learn from corpora.
 - POS Tagger (Church 1988)
 - Noun Phrase Identifier Module (Soon et al.)
 - NER System (Bikel, Schwartz, and Weischedel 1999)



Nested-Noun Phrase Module

- Noun Phrase ID and NER output merged, then delivered to NNP extraction module.
 - Accepts Noun Phrase, returns all nested noun phrases.
 - Two flavors of NNP:
 - Possessive NP [“his long-term strategy”], the NNP is “his.”
 - Prenominals [“wage increases”], the NNP would be “wage” in this case.
-
-

Markables for Resolution

- The markables that are used for coreference resolution:

The union of the NER system output, nested-NP module and the NP detection module.



Feature Vector Creation

- Considerations:
 - Strong enough to determine if two markables corefer.
 - Weak enough to be stretched across multiple domains.
- What considerations for different kinds of noun phrases?
 - Ended up using 5 features to catch 5 types of NP.
[Definite, Demonstrative, pronouns, and proper names]



Feature Set

- **Distance** – The distance, in sentences, between anaphor and antecedent.
 - **Pronoun** – Two features, one for the antecedent and anaphor apiece
 - **String Match** – articles and demonstratives are removed, thus: “this car” = “that car”, etc.
 - **Definite NP** – Checked the anaphor only.
 - **Demonstrative NP** – Checked the demonstrative only.
-
-

Feature Set {cont.}

- **Number Agreement** – Single or plural? The morphological root is used to determine this.
 - **Semantic Class Agreement** – Uses WordNet to determine semantic class.
 - **Gender Agreement** – Uses name lists, and titles to determine gender of a markable.
 - **Both-Propor-Names** – Determined by capitalization...
 - **Alias** – If both markables are named entities of the same type, other constraints.
 - **Appositive** – At least one markable must be a proper name, also checks for proper punctuation and the existence of verbs.
-
-

Feature Vector

Table 1

Feature vector of the markable pair ($i = \textit{Frank Newman}$, $j = \textit{vice chairman}$).

Feature	Value	Comments
DIST	0	i and j are in the same sentence
LPRONOUN	-	i is not a pronoun
JPRONOUN	-	j is not a pronoun
STR.MATCH	-	i and j do not match
DEF.NP	-	j is not a definite noun phrase
DEM.NP	-	j is not a demonstrative noun phrase
NUMBER	+	i and j are both singular
SEMCLASS	1	i and j are both persons (This feature has three values: false(0), true(1), unknown(2).)
GENDER	1	i and j are both males (This feature has three values: false(0), true(1), unknown(2).)
PROPER.NAME	-	Only i is a proper name
ALIAS	-	j is not an alias of i
APPOSITIVE	+	j is in apposition to i

Where i is the antecedent and j is the anaphor.

Collecting Training Data

- Given a coreferent chain, $A1 \rightarrow A2 \rightarrow A3 \rightarrow A4$, from an annotated training document.
 - Pairs of noun phrases that are immediately adjacent are used to generate positive training data.
 - Therefore, $A1 - A2$, $A2 - A3$, $A3 - A4$ are used as positive examples.
-
-

Negative Training Examples

- Other markables that are not found in any chains are paired with anaphor to form a negative example.
 - If a,b are markables that appear between A1 – A2, then negative examples would be:
 - a – A2, b – A2, etc...
 - For both positive and negative examples, annotated noun phrases must be recognized as noun phrases in the NLP pipeline.
-
-

Examples

Annotated Text:

((Union)_{a1} representatives that could be reached)_{b1} said (they)_{b2} hadn't decided whether (they)_{b3} would respond. By proposing (a meeting time)_{c1}, (Eastern)_{d1} moved one step closer toward reopening (high-cost contract agreements)_{e1} with ((its)_{d2} unions)_{a2}.

Results from NLP pipeline must extract these noun phrases in order to generate training data!

Generating Coreference Chains in Test Documents

- “Every markable” is a possible anaphor, every markable before a given anaphor is a possible antecedent.
 - However, antecedents do not share this characteristic.
 - Example: *“Mr. Smith's daughter ... and his daughter's eyes.”*
 - The above rule restricts the possible antecedents for “his”.
-
-

The Coreference Algorithm

For every markable j , select as possible anaphor:

For every i , where $i < j$:

generateFeatureVector(i,j) \rightarrow DecisionTree:

If(DecisionTree == TRUE): *antecedent found*

Else: continue.

An Example

(Mr. Washington)₇₃'s candidacy is being championed by (several powerful lawmakers)₇₄ including ((her)₇₆ boss)₇₅ (Chairman John Dingell)₇₇ (D., Mich)₇₈ of (the House Energy and Commerce Committee)₇₉. (She)₈₀ currently is a counsel to (the committee)₈₂. (Mrs. Washington)₈₃ and (Mr. Dingell)₈₄ have been considered (allies)₈₅ of (the (securities)₈₇ exchanges)₈₆, while (banks)₈₈ and ((futures)₉₀ exchanges)₈₉ have often fought with (them)₉₁.

Example Output

Antecedent	Anaphor	Feature Vector	Corefers?
(several powerful lawmakers) ₇₄	(her) ₇₆	0,1,-2,-,-,+,-,-,-,-,-	No
(Ms. Washington) ₇₃	(her) ₇₆	0,1,+1,-,-,+,-,-,-,-,-	Yes
(the House Energy and Commerce Committee) ₇₉	(She) ₈₀	1,0,+0,-,-,+,-,-,-,-,-	No
(Mich.) ₇₈	(She) ₈₀	2,0,+0,-,-,+,-,-,-,-,-	No
(Chairman John Dingell) ₇₇	(She) ₈₀	3,1,+0,-,-,+,-,-,-,-,-	No
(her) ₇₆	(She) ₈₀	3,1,+1,-,-,+,-,-,-,-,+	Yes
(the committee) ₈₂	(Ms. Washington) ₈₃	1,0,+0,-,-,-,-,-,-,-,-	No
(a counsel) ₈₁	(Ms. Washington) ₈₃	1,1,+2,-,-,-,-,-,-,-,-	No
(She) ₈₀	(Ms. Washington) ₈₃	1,1,+1,-,-,-,-,-,-,-,+	No
(the House Energy and Commerce Committee) ₇₉	(Ms. Washington) ₈₃	2,0,+0,+,-,-,-,-,-,-,-	No
(Mich.) ₇₈	(Ms. Washington) ₈₃	3,0,+0,+,-,-,-,-,-,-,-	No
(Chairman John Dingell) ₇₇	(Ms. Washington) ₈₃	4,1,+0,+,-,-,-,-,-,-,-	No
(her) ₇₆	(Ms. Washington) ₈₃	4,1,+1,-,-,-,-,-,-,-,+	No
(her boss) ₇₅	(Ms. Washington) ₈₃	4,1,-0,-,-,-,-,-,-,-,-	No
(several powerful lawmakers) ₇₄	(Ms. Washington) ₈₃	4,1,-2,-,-,-,-,-,-,-,-	No
(Ms. Washington) ₇₃	(Ms. Washington) ₈₃	4,1,+1,+,-,-,-,-,+,-,-	Yes

Choosing the Right Classifier

- Developed a C5 classifier, which is decision tree based.
 - Improve C4.5 (which was an improvement upon ID3) (Quinlan)
 - Binary Classifier, either i and j corefer, or not.
 - For more information see References.
- Decision Trees: A structure where non-terminal nodes represent tests on one or more attributes and leaves reflect decision outcomes.



Decision Tree Output

```
STR_MATCH = +: +
STR_MATCH = -:
: ... J_PRONOUN = -:
    : ... APPOSITIVE = +: +
    :     APPOSITIVE = -:
    :     : ... ALIAS = +: +
    :     :     ALIAS = -: -
J_PRONOUN = +:
: ... GENDER = 0: -
    GENDER = 2: -
    GENDER = 1:
        : ... I_PRONOUN = +: +
        I_PRONOUN = -:
            : ... DIST > 0: -
            DIST <= 0:
                : ... NUMBER = +: +
                NUMBER = -: -
```

- Only 8 of 12 features were learned from MUC-6 texts.
- Only 2 of 5 “noun-type” features learned.

Evaluation (Training)

- MUC-6
 - 30 documents used as training. (12,400 words)
 - 20,910 Training examples. (Coreferent pairs.)
 - 6.5% positive.
- MUC-7
 - 30 documents (19,000 words)
 - 48,872 Training examples.
 - 4.4% positive.

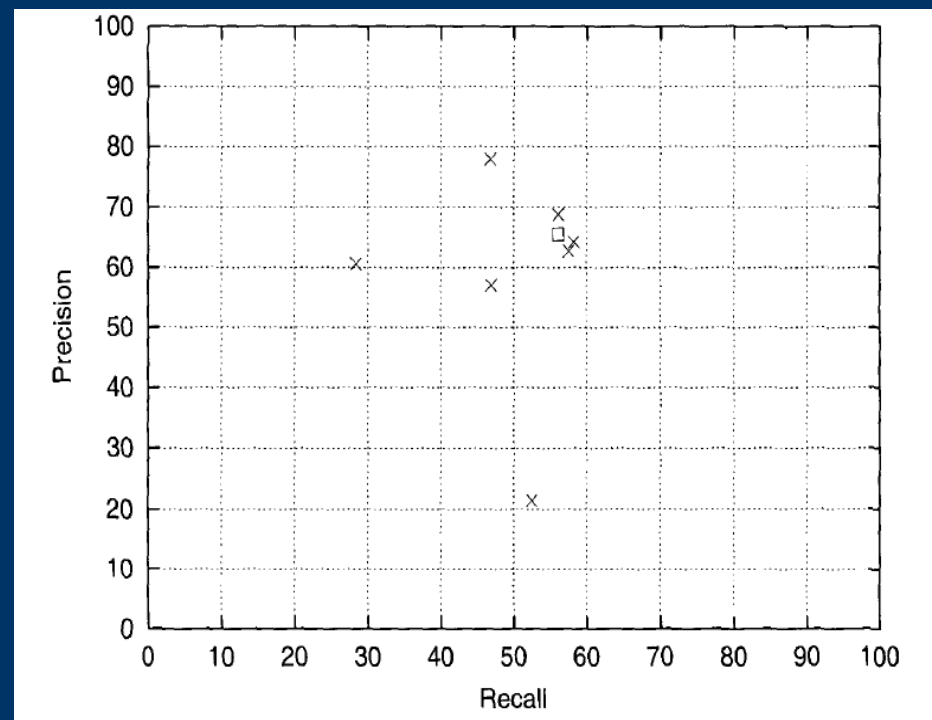
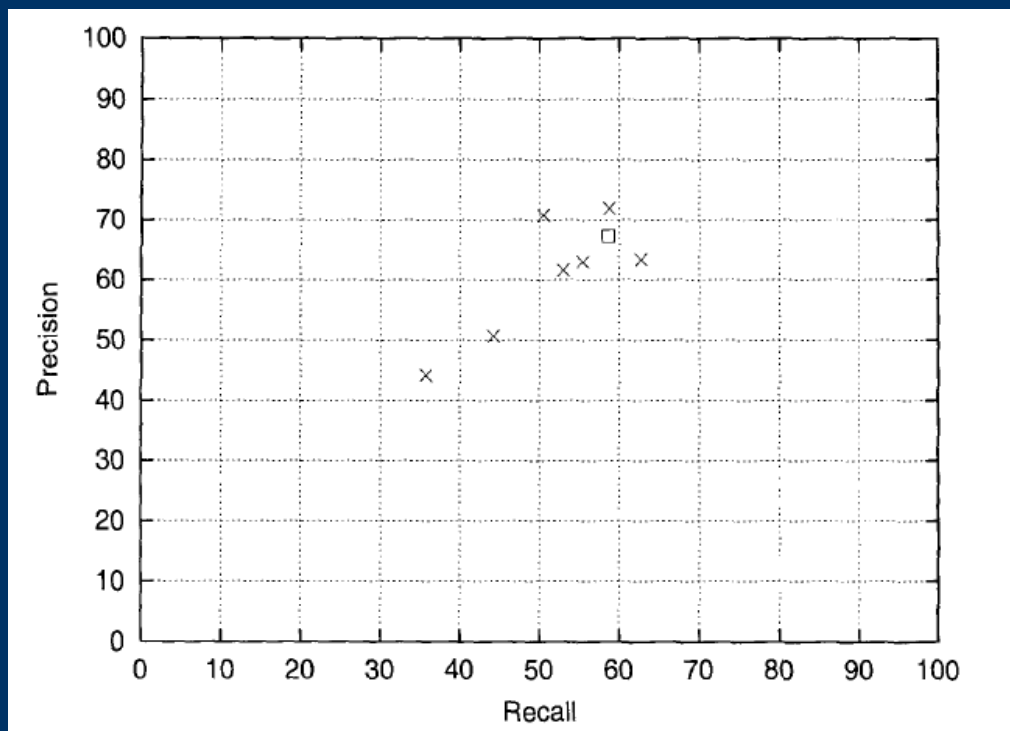
Evaluation (C5 Classifier)

- MUC-6
 - Minimum instances = 5
 - Pruning confidence = 20%
 - MUC-7
 - Minimum instances = 2
 - Pruning confidence = 60%
 - Lower **PC** means more drastic pruning of Decision Tree.
 - **MI** indicates how many occurrences must be present in the training data to create a new leaf. Lower mean more specialized tree.
-
-

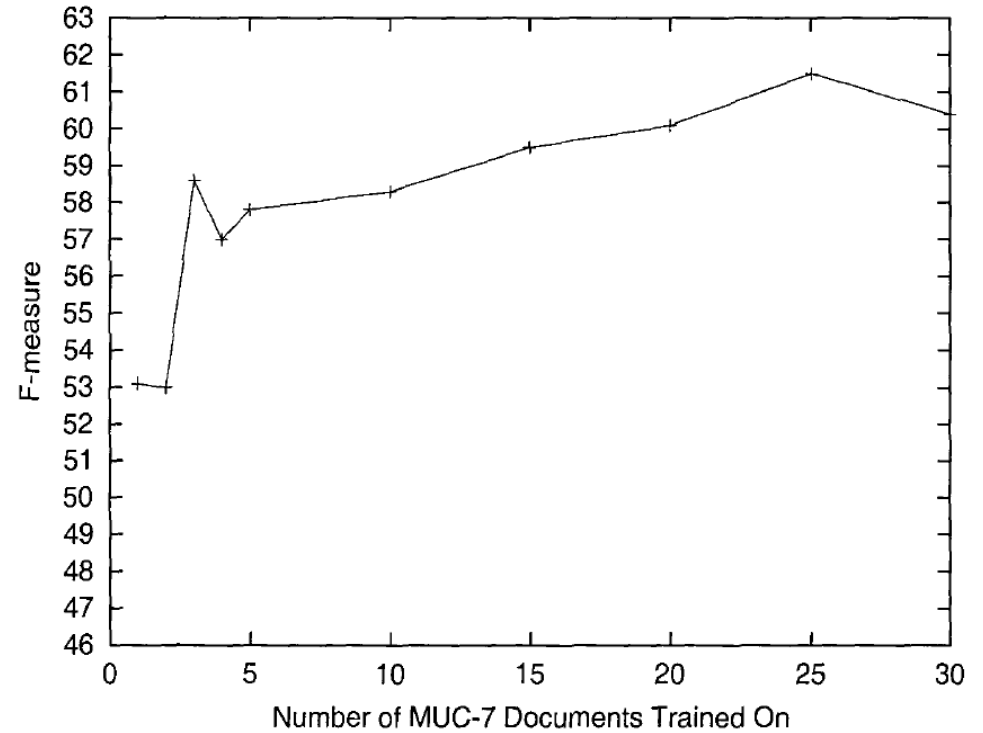
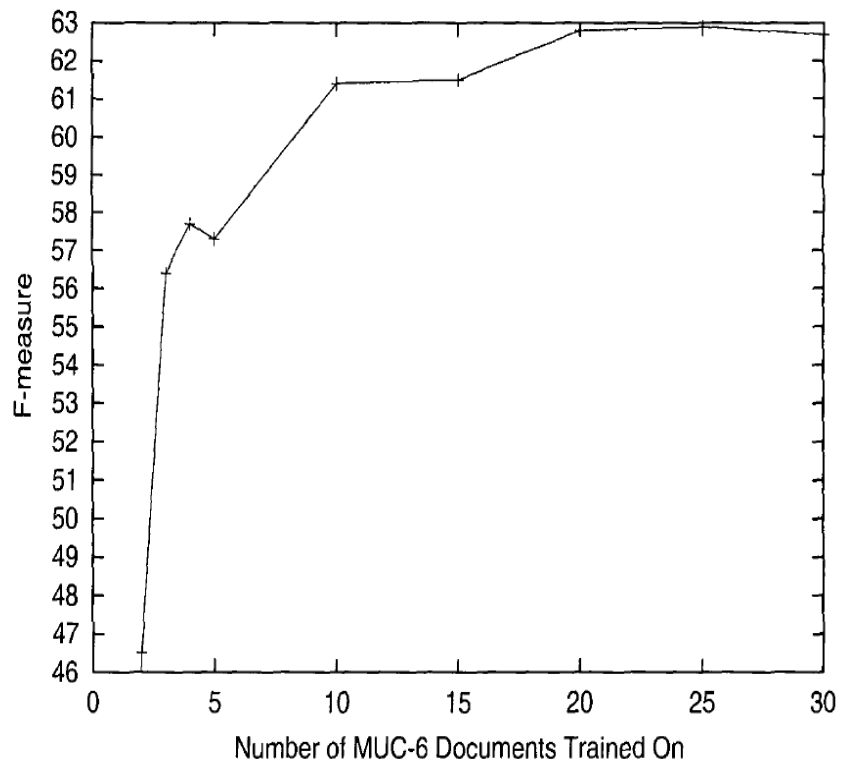
Results 1

	Recall	Precision	F-Measure
MUC-6	58.6%	67.3%	62.6%
MUC-7	56.1%	65.5%	60.4%

Results 2



Learning Curve



Contribution of Features

Table 3
MUC-6 results of complete and baseline systems to study the contribution of the features.

System ID	Recall	Prec	F	Remarks
Complete systems				
DSO	58.6	67.3	62.6	Our system
DSO_TRG	52.6	67.6	59.2	Our system using RESOLVE's method of generating positive and negative examples
RESOLVE	44.2	50.7	47.2	The RESOLVE coreference system at the University of Massachusetts
Baseline systems using just one feature				
DIST	0.0	0.0	0.0	Only "distance" feature is used
SEMCLASS	0.0	0.0	0.0	Only "semantic class agreement"
NUMBER	0.0	0.0	0.0	Only "number agreement"
GENDER	0.0	0.0	0.0	Only "gender agreement"
PROPER_NAME	0.0	0.0	0.0	Only "both proper names"
ALIAS	24.5	88.7	38.4	Only "alias"
J_PRONOUN	0.0	0.0	0.0	Only "j-pronoun"
DEF_NP	0.0	0.0	0.0	Only "definite noun phrase"
DEM_NP	0.0	0.0	0.0	Only "demonstrative noun phrase"
STR_MATCH	45.7	65.6	53.9	Only "string match"
APPPOSITIVE	3.9	57.7	7.3	Only "appositive"
I_PRONOUN	0.0	0.0	0.0	Only "i-pronoun"
Other baseline systems				
ALIAS_STR	51.5	66.4	58.0	Only the "alias" and "string match" features are used
ALIAS_STR_APPPOS	55.2	66.4	60.3	Only the "alias," "string match," and "appositive" features are used
ONE_CHAIN	89.9	31.8	47.0	All markables form one chain
ONE_WRD	55.4	36.6	44.1	Markables corefer if there is at least one common word
HD_WRD	56.4	50.4	53.2	Markables corefer if their head words are the same

Error Analysis

Table 6

The types and frequencies of errors that affect recall.

Types of Errors Causing Missing Links	Frequency	%
Inadequacy of current surface features	38	63.3%
Errors in noun phrase identification	7	11.7%
Errors in semantic class determination	7	11.7%
Errors in part-of-speech assignment	5	8.3%
Errors in apposition determination	2	3.3%
Errors in tokenization	1	1.7%

- Randomly chose 5 documents to determine reason for erroneous coreference links.

Improving Machine Learning Approaches to Coreference Resolution

- Vincent Ng and Claire Cardie, Cornell University
- Paper cited by 101.



Using Soon for a Baseline...

- Ng & Cardie pick up where Soon and gang left off.
 - Reimplemented their works as a baseline, with a few changes:
 - Use C4.5 instead of C5.
 - Class values are assigned for coreferent pairs, values of 0.5 or greater result in match. Only one match sought.
 - Trained bases system on same corpora.
-
-

New Features

- Best-first clustering
 - Searched for *highly likely antecedent*, instead of first matched.
 - Antecedent was chosen by highest class value of all previous possibilities (with class value > 0.5)
 - Training Set Creation
 - Generated positive training data for most confident antecedent.
 - Several heuristics used to determine most confident antecedent.
 - Negative examples created just as in Soon.
-
-

More New Features

- Improved String Match
 - Split Soon String match into several primitive features.
- Increased feature set from 12 to 53!
- Employed several different variants of system
 - with different classifying algorithm, RIPPER.
 - with hand crafted feature sets.



Results

System Variation	C4.5						RIPPER					
	MUC-6			MUC-7			MUC-6			MUC-7		
	R	P	F	R	P	F	R	P	F	R	P	F
Original Soon et al.	58.6	67.3	62.6	56.1	65.5	60.4	-	-	-	-	-	-
Duplicated Soon Baseline	62.4	70.7	66.3	55.2	68.5	61.2	60.8	68.4	64.3	54.0	69.5	60.8
Learning Framework	62.4	73.5	67.5	56.3	71.5	63.0	60.8	75.3	67.2	55.3	73.8	63.2
String Match	60.4	74.4	66.7	54.3	72.1	62.0	58.5	74.9	65.7	48.9	73.2	58.6
Training Instance Selection	61.9	70.3	65.8	55.2	68.3	61.1	61.3	70.4	65.5	54.2	68.8	60.6
Clustering	62.4	70.8	66.3	56.5	69.6	62.3	60.5	68.4	64.2	55.6	70.7	62.2
All Features	70.3	58.3	63.8	65.5	58.2	61.6	67.0	62.2	64.5	61.9	60.6	61.2
Pronouns only	-	66.3	-	-	62.1	-	-	71.3	-	-	62.0	-
Proper Nouns only	-	84.2	-	-	77.7	-	-	85.5	-	-	75.9	-
Common Nouns only	-	40.1	-	-	45.2	-	-	43.7	-	-	48.0	-
Hand-selected Features	64.1	74.9	69.1	57.4	70.8	63.4	64.2	78.0	70.4	55.7	72.8	63.1
Pronouns only	-	67.4	-	-	54.4	-	-	77.0	-	-	60.8	-
Proper Nouns only	-	93.3	-	-	86.6	-	-	95.2	-	-	88.7	-
Common Nouns only	-	63.0	-	-	64.8	-	-	62.8	-	-	63.5	-

Table 2: Results for the MUC-6 and MUC-7 data sets using C4.5 and RIPPER. Recall, Precision, and F-measure are provided. Results in boldface indicate the best results obtained for a particular data set and classifier combination.

Final Thoughts on Ng and Cardie

- Increased feature set size without increasing training size.
- In the end, the hand crafted feature produced some of the strongest results.
- All 53 features when used in unison produced poor results.

```
ALIAS = C: + (347.0/23.8)
ALIAS = I:
|
|   SOON_STR_NONPRO = C:
|   | ANIMACY = NA: - (4.0/2.2)
|   | ANIMACY = I: + (0.0)
|   | ANIMACY = C: + (259.0/45.8)
|   SOON_STR_NONPRO = I:
|   | PRO_STR = C: + (39.0/2.6)
|   | PRO_STR = I:
|   | | PRO_RESOLVE = C:
|   | | | EMBEDDED_1 = Y: - (7.0/3.4)
|   | | | EMBEDDED_1 = N:
|   | | | | PRONOUN_1 = Y:
|   | | | | | ANIMACY = NA: - (6.0/2.3)
|   | | | | | ANIMACY = I: - (1.0/0.8)
|   | | | | | ANIMACY = C: + (10.0/3.5)
|   | | | | PRONOUN_1 = N:
|   | | | | | MAXIMALNP = C: + (108.0/18.2)
|   | | | | | MAXIMALNP = I:
|   | | | | | | WNCLASS = NA: - (5.0/1.2)
|   | | | | | | WNCLASS = I: + (0.0)
|   | | | | | | WNCLASS = C: + (12.0/3.6)
|   | | PRO_RESOLVE = I:
|   | | | APPOSITIVE = I: - (26806.0/713.8)
|   | | | APPOSITIVE = C:
|   | | | | GENDER = NA: + (28.0/2.6)
|   | | | | GENDER = I: + (5.0/3.2)
|   | | | | GENDER = C: - (17.0/3.7)
```

Conclusions

- Machine Learning is becoming more viable for coreference resolution.
- Next step, better unsupervised algorithms?



Questions or Comments?

Thank You!

