Combining Sample Selection and Error-Driven Pruning for Machine Learning of Coreference Rules

Vincent Ng and Claire Cardie
Department of Computer Science
Cornell University

Plan for the talk

- Noun phrase coreference resolution
- Baseline coreference resolution system
  - standard machine learning approach
- Problems and potential solutions

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...

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...

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*

coref ? *coref* ?

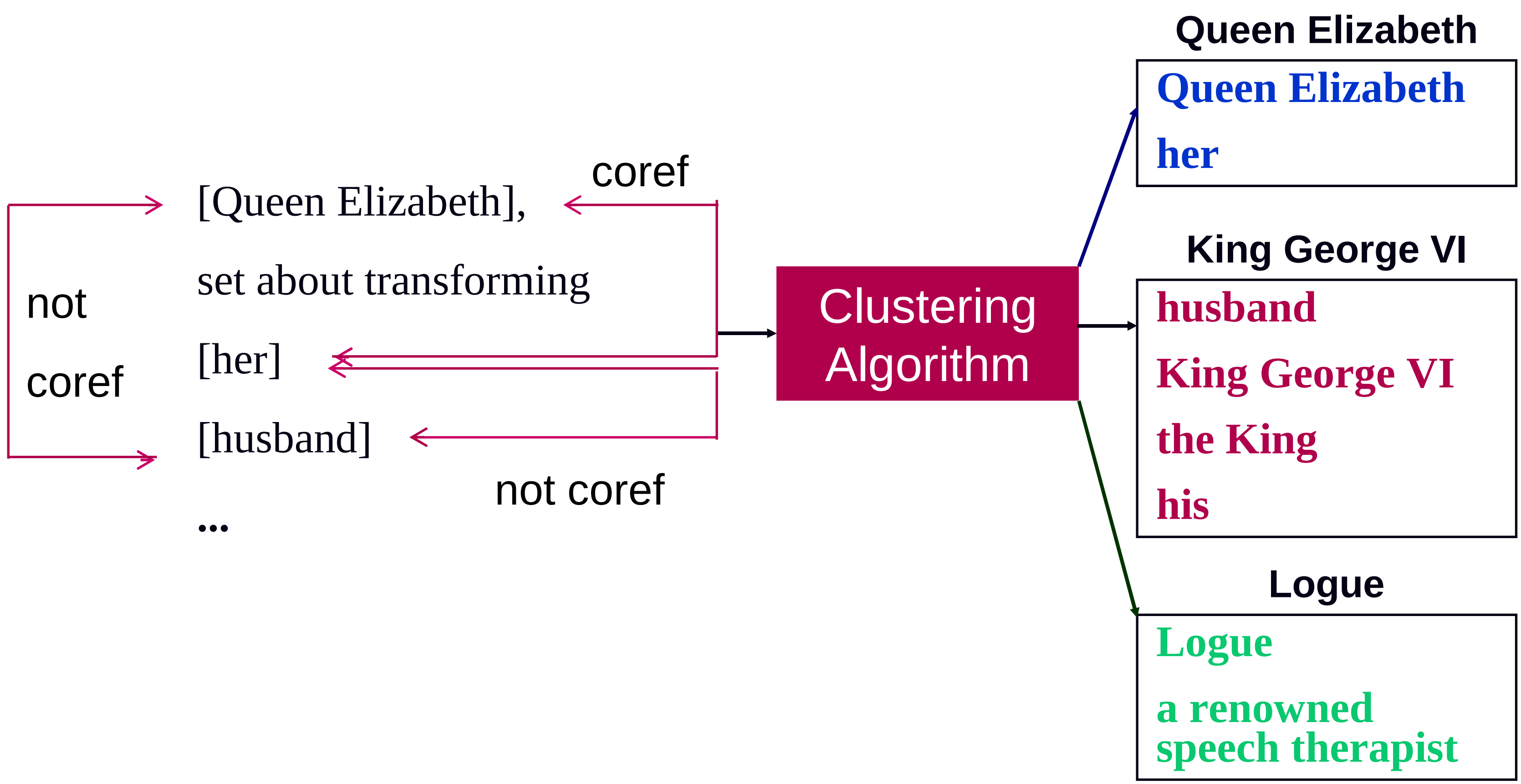
[Queen Elizabeth] set about transforming [her] [husband], ...

not coref ?

Aone & Bennett [1995]; Connolly et al. [1994]; McCarthy & Lehnert [1995]; Soon, Ng & Lim [2001]

A Machine Learning Approach

- Clustering
  - coordinates pairwise coreference decisions



Machine Learning Issues

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

Baseline System: 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

Baseline System: Instance Representation

- 25 features per instance
  - lexical (3)
  - grammatical (18)
  - semantic (2)
  - positional (1)
  - knowledge-based (1)

Baseline System: Learning Algorithm

- RIPPER (Cohen, 1995): positive rule learner
  - input: set of training instances
  - output: coreference classifier

- Classifier outputs
  - classification
  - confidence of classification

Baseline System: Clustering Algorithm

- Best-first single-link clustering

CREATE-COREF-CHAINS (NP_1, NP_2, ..., NP_n)

Mark each NP_j as belonging to its own class: $NP_j \in c_j$

For each NP_j do

Form an instance from NP_j with each preceding NP

Let $S(NP_j) = \{NP_i \mid NP_i$ is classified as coreferent with $NP_j\}$

Let NP_k = noun phrase in $S(NPj)$ with highest confidence

$c_j = c_j \cup c_k$

Baseline System: Evaluation

- MUC-6 and MUC-7 coreference data sets
- documents annotated w.r.t. coreference
- MUC-6: 30 training texts + 30 test texts
- MUC-7: 30 training texts + 20 test texts
- MUC scoring program
  - recall, precision, F-measure

Baseline System: 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

Baseline System: Results

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

Problem 1

- Coreference is an equivalence relation
  - loss of transitivity

coref ? *coref ?*

[Queen Elizabeth] set about transforming [her] [husband], ...

not coref ?

Problem 2

- Coreference is a rare relation
  - skewed class distributions
  - MUC-6 and MUC-7 dry run data sets each contains only 2% positive instances

Problem 3

- Coreference is a discourse-level problem
  - different solutions for different types of NPs
    - pronouns: locality constraints
    - proper names: string matching and aliasing

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

  - inclusion of “hard” positive training instances

Problem 3

- Coreference is a discourse-level problem
  - different solutions for different types of NPs
    - pronouns: locality constraints
    - proper names: string matching and aliasing

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

  - inclusion of “hard” positive training instances

Problem 3

- Coreference is a discourse-level problem
  - different solutions for different types of NPs
    - pronouns: locality constraints
    - proper names: string matching and aliasing

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

  - inclusion of “hard” positive training instances

Problem 3

- Coreference is a discourse-level problem
  - different solutions for different types of NPs
    - pronouns: locality constraints
    - proper names: string matching and aliasing

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

  - inclusion of “hard” positive training instances

Problem 3

- Coreference is a discourse-level problem
  - different solutions for different types of NPs
    - pronouns: locality constraints
    - proper names: string matching and aliasing

Queen Elizabeth set about transforming her husband,

King George VI, into a viable monarch. Logue,

the renowned speech therapist, was summoned to help

the King overcome his speech impediment...

  - inclusion of “hard” positive training instances

Problem 3

- Coreference is a discourse-level problem
  - different solutions for different types of NPs
    - » pronouns: locality constraints
    - » proper names: string matching and aliasing

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

  - inclusion of “hard” positive training instances

Classification-based Single-link Clustering

- Problems
  - skewed class distributions
  - inclusion of hard positive training instances
  - loss of transitivity

Skewed Class Distributions

- negative example selection
- variant of the Soon *et al.* (2001) algorithm
- NEG-SELECT retains only negative instances for non-coreferent NPs that lie between an anaphoric NP and its farthest preceding antecedent

Negative Example Selection

- An example
  - create negative instances from *NP9*

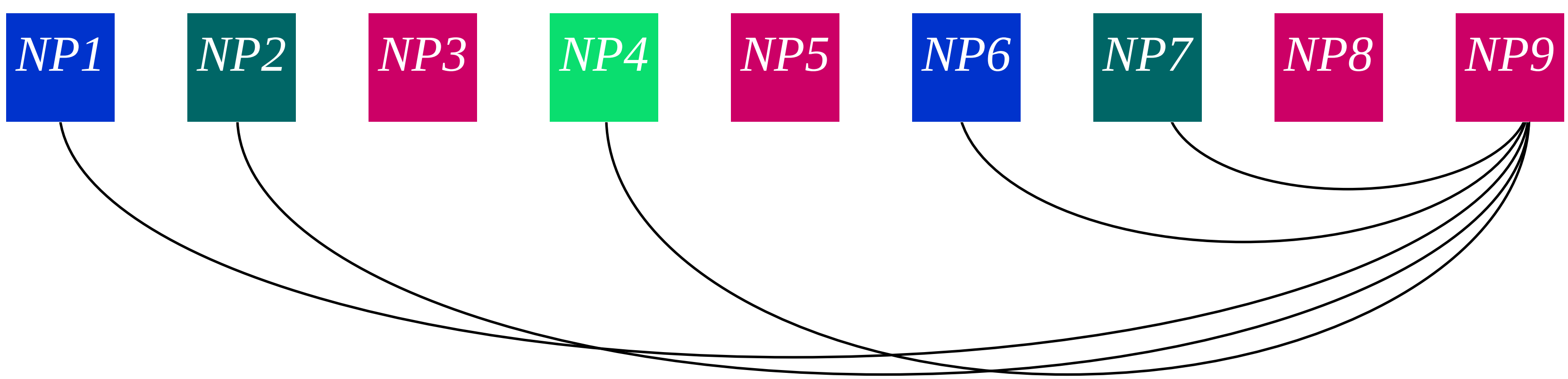
Negative Example Selection

Step 1: Create all possible negative instances from *NP9*

NP1 *NP2* *NP3* *NP4* *NP5* *NP6* *NP7* *NP8* *NP9*

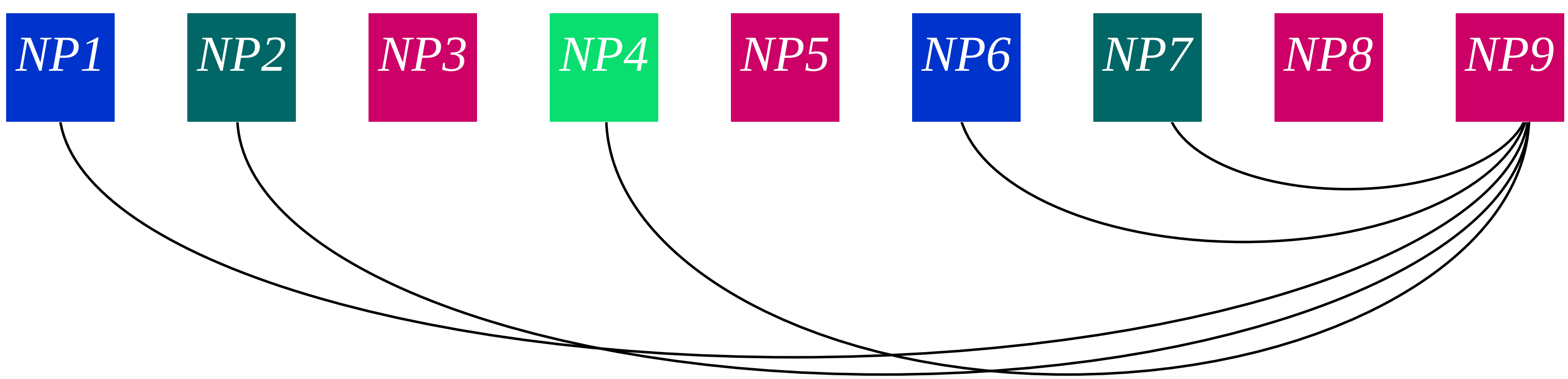
Negative Example Selection

Step 1: Create all possible negative instances from *NP9*



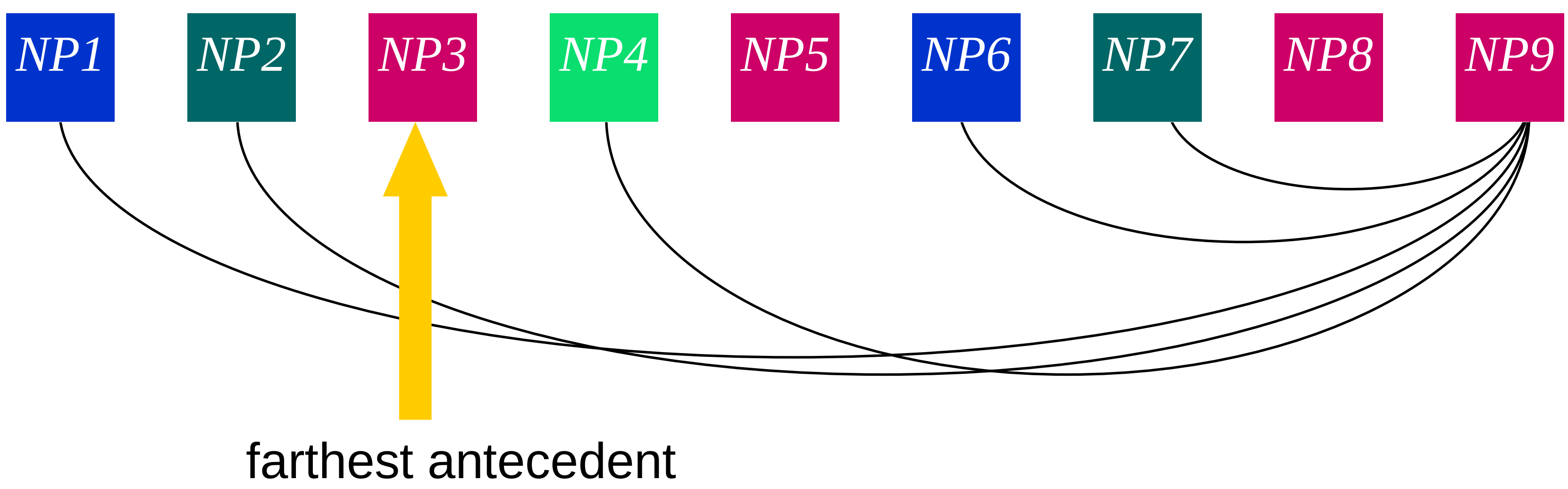
Negative Example Selection

Step 2: Locate the farthest antecedent of *NP9* , *f(NP9)*



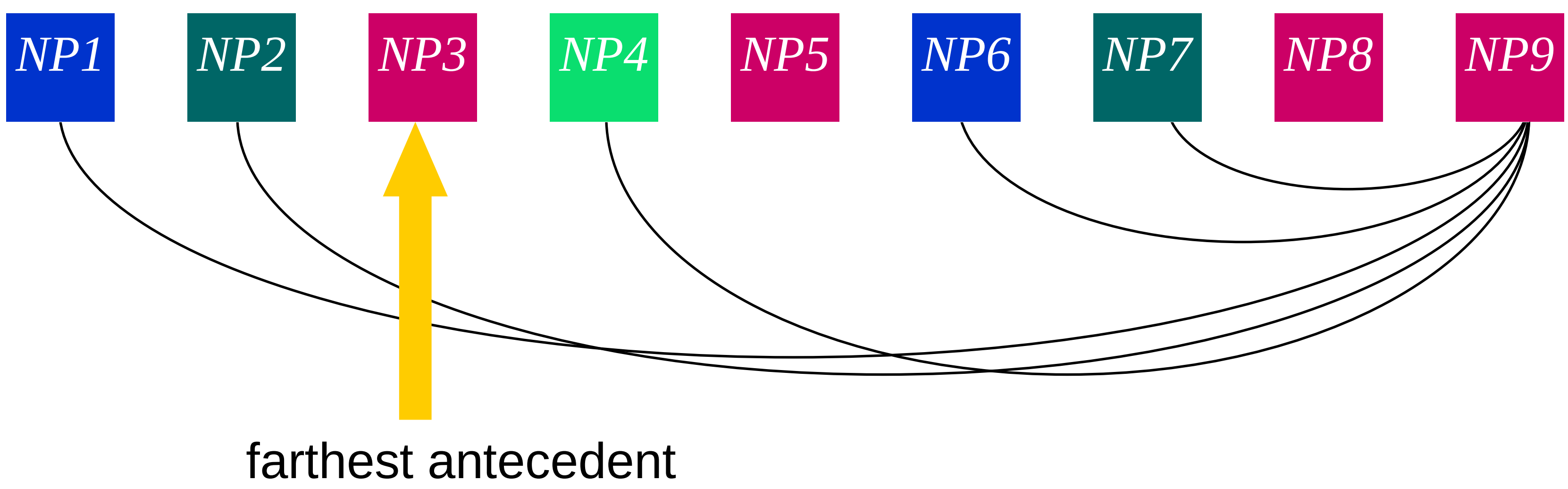
Negative Example Selection

Step 2: Locate the farthest antecedent of *NP9*, *f(NP9)*



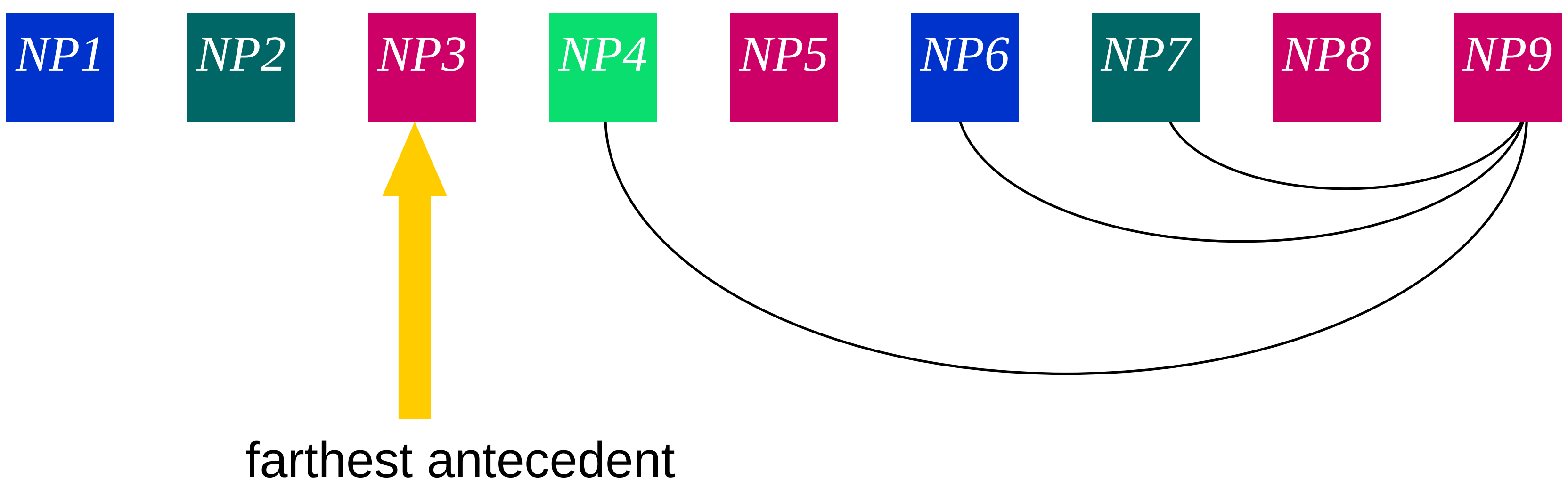
Negative Example Selection

Step 3: Remove all instances involving NPs that precede *f(NP9)*



Negative Example Selection

Step 3: Remove all instances involving NPs that precede *f(NP9)*



Results (Negative Example Selection)

	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**

- % of positive instances: 8% (MUC-6) and 7% (MUC-7)
- gain in recall but larger loss in precision
- overall performance (F-measure) increases

Inclusion of Hard Training Instances

- positive example selection
- selects easy positive training instances
- automatic variant of the Harabagiu *et al.* (2001) algorithm

POS-SELECT(*L:* positive rule learner, *T:* set of training instances*)*

repeat

Induce a ranked set of positive rules *R* on *T* using *L*

Let *BestRule* = best rule in R

Add *BestRule* to *FinalRuleSet*

For each $inst(NP_i, NP_j) \in T$ correctly covered by *BestRule,*

remove all instances of the form $inst(*, NP_j)$ from *T*.

until *L* cannot induce any rule for the positive instances

return *FinalRuleSet*

Results (Positive Example Selection)

	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

- F-measure increases by 12% using POS-SELECT

Results (Positive Example Selection)

	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**

- F-measure increases by 16-17% using both NEG-SELECT and POS-SELECT

Results (Positive Example Selection)

	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**

- using both NEG-SELECT and POS-SELECT leads to better performance than using POS-SELECT alone

Loss of Transitivity

- rule pruning
- tightens connection between classification and clustering

RULE-SELECT(*R*: ruleset, *P*: pruning corpus; *S*: scoring function)

Let *BestScore* = score of the coref system using *R* on *P* w.r.t. *S*
repeat
Let *r* = the rule in *R* whose removal yields a ruleset with which coref system achieves the best score *b* on *P* w.r.t. *S*
If *b* > *BestScore*
then set *BestScore* to *b* and remove *r* from *R*
otherwise **return** *R*
while *true*

- optimizes w.r.t. the clustering-level coref scoring function

Results (Rule Selection)

	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
NEG-SELECT + POS-SELECT (more data)	64.8	70.6	67.6	60.0	55.7	57.8

- pruning corpus
  - MUC-6: MUC-7 formal
  - MUC-7: MUC-6 formal

Results (Rule Selection)

	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**
NEG-SELECT + POS-SELECT (more data)	64.8	70.6	67.6	60.0	55.7	57.8

- gains in precision; increase in F-measure
- effective at improving precision

Results (Rule Selection)

	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**
NEG-SELECT + POS-SELECT (more data)	64.8	70.6	**67.6**	60.0	55.7	**57.8**

- RULE-SELECT has made a more effective use of the additional data provided by the pruning corpus

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**

- performs better than the best MUC coreference systems

Summary

- Examined three problems with recasting noun phrase coreference resolution as a classification task
- Showed how the problems can be handled via example selection and error-driven pruning of classification rules

Properties of Coreference	Problems	Solutions
Coref is a rare relation	Skewed distributions	Negative example selection
Coref is a discourse-level problem	Inclusion of hard training instances	Positive example selection
Coref is an equivalence relation	Loss of transitivity	Rule pruning