Improving Machine Learning Approaches to Coreference Resolution

Vincent Ng and Claire Cardie
Department of Computer Science
Cornell University

Plan for the Talk

- noun phrase coreference resolution
- standard machine learning approach
  - an instantiation (W. M. Soon, H. T. Ng, and D. Lim [2001])
- two extensions
  - extra-linguistic modifications to the learning framework
  - large-scale expansion in linguistic knowledge sources

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 <u>a viable monarch</u>. Logue,
a renowned speech therapist, was summoned to help
the King overcome his <u>speech impediment</u>...

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...

Plan for the Talk

- noun phrase coreference resolution
- standard machine learning approach
  - an instantiation (Soon *et al.* [2001])
- two extensions

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*
  - create one training instance for each pair of noun phrases from texts annotated with coreference information
    - feature vector: describes the two NPs and context

coref ? *coref ?*

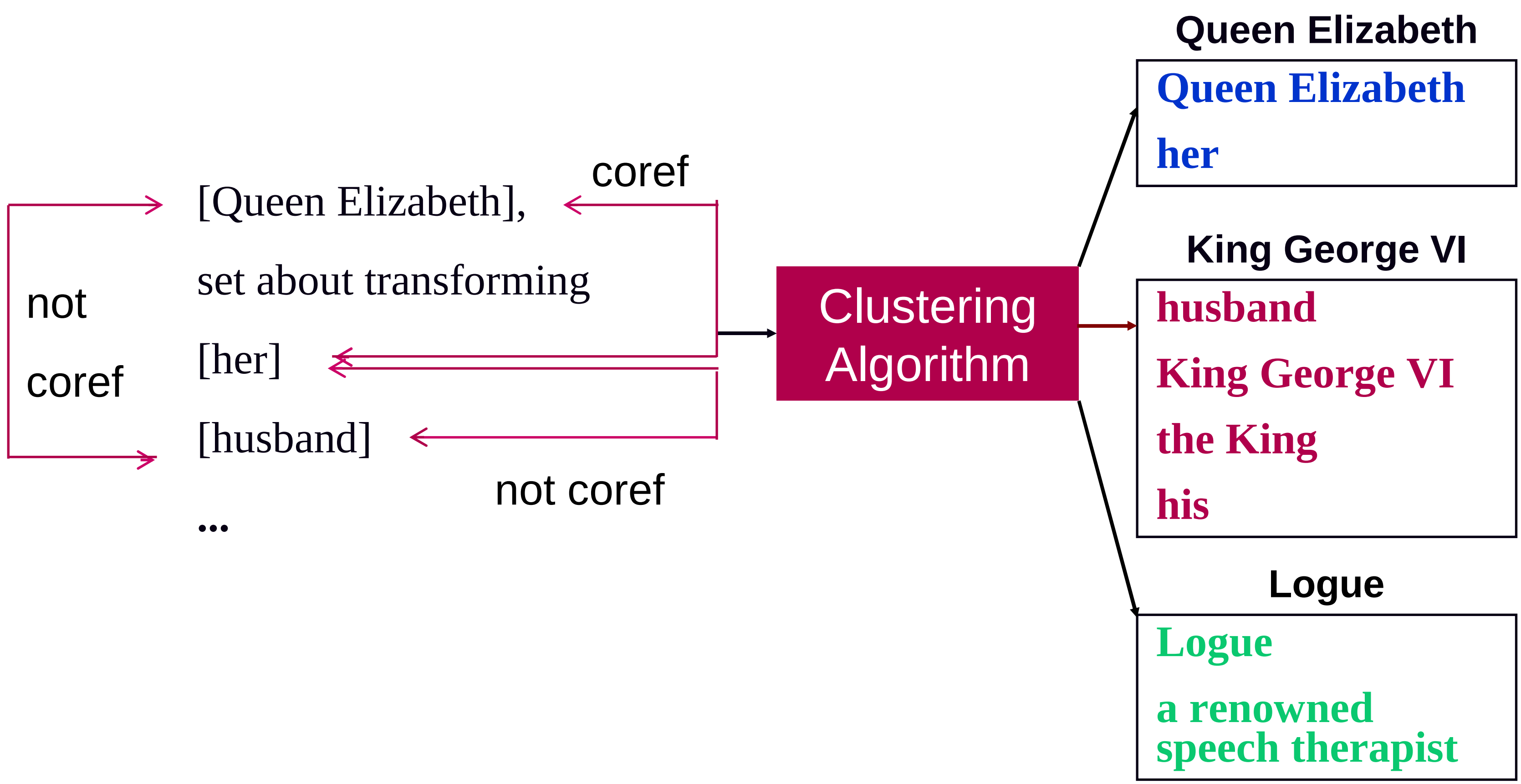
[Queen Elizabeth] set about transforming [her] [husband], ...

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



Plan for the Talk

- noun phrase coreference resolution
- standard machine learning approach
  - an instantiation (Soon *et al.* [2001])
- two extensions

Soon *et al.* Training Instance Selection

- Creating training instances
  - Naïve method: one instance for each pair of noun phrases
  - Problem: all NP pairs produces highly skewed data set
  - Soon *et al.*'s solution: create
    - » *positive instance* for each anaphoric noun phrase, NP_j, and its closest preceding antecedent, NP_i
    - » *negative instance* for NP_j and each intervening noun phrase, NP_{i+1}, NP_{i+2} , … , NP_{j-1}

Soon Instance Representation

	Soon_str	Pronoun_1 Pronoun_2 Definite_2 Demonst_2	Number	Gender	Both_ proper_ nouns	Appos- itive	Semantic class	Alias	Sent_ dist	Class
NP1/ NP2	✗	No; Yes; No; No	✓	✓	✗	✗	✓	✗	0	Coref
NP2/ NP3	✗	Yes; No; No; No	✓	✗	✗	✗	✓	✗	0	Not coref

NP1 NP2 NP3

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

Soon *et al.* Clustering Algorithm

CREATE-COREF-CHAINS ($NP_1, NP_2, \ldots, NP_n; doc$)

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_j encountered, consider each preceding NP_i:

Let c_i = class of NP_i and c_j = class of NP_j

Let *coref-likelihood* =

dtree (feat_vec (NP_i, NP_j, *doc*))

If *coref-likelihood* > *0.5* then $c_j = c_j \cup c_i$; break

- selects as antecedent the closest preceding coreferent NP

Soon *et al.* 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

Results (Soon *et al.* system)

	MUC-6			MUC-7		
	R	P	F	R	P	F
Soon	58.6	67.3	**62.6**	56.1	65.5	**60.4**
Best MUC System	59	72	**65**	56.1	68.8	**61.8**
Worst MUC System	36	44	**40**	52.5	21.4	**30.4**

- Soon's system is the first learning-based system that achieves performance comparable to the best MUC systems

Results (Baseline System)

	MUC-6			MUC-7		
	R	P	F	R	P	F
Original Soon	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**

- improvements over Soon:
  - MUC-6: better identification of NPs (93.8 vs. 89.9 max recall)
  - MUC-7: more accurate feature value computation

Results (Baseline System)

	MUC-6			MUC-7		
	R	P	F	R	P	F
Duplicated Soon Baseline	62.4	70.7	**66.3**	55.2	68.5	**61.2**
Best MUC System	59	72	**65**	56.1	68.8	**61.8**

Plan for the Talk

- noun phrase coreference resolution
- standard machine learning approach
  - an instantiation (Soon *et al.* [2001])
- two extensions
  - three extra-linguistic modifications to the learning framework
  - large-scale expansion in linguistic knowledge sources

Modification 1: Best-first Clustering

- Focus on improving precision

Modify create-coref-chain to select as the antecedent of NP_j the noun phrase with the

highest coref-likelihood score

from among preceding noun phrases with coref-likelihood scores > 0.5.

Assume non-anaphoric otherwise.

Modification 2: Training Instance Selection

- Focus on improving precision

Generate a positive training example for the

most confident antecedent

of each anaphoric NP_j. Negative examples are generated as in the Baseline system.

non-pronominal NP: closest preceding non-pronominal antecedent

pronouns: closest preceding antecedent

Modification 2: Training Instance Selection

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. The Queen was said to have acted as a buffer for the King's fierce temper during the process.

Modification 2: Training Instance Selection

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. The Queen was said to have acted as a buffer for the King's fierce temper during the process.

Modification 2: Training Instance Selection

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. The Queen was said to have acted as a buffer for the King's fierce temper during the process.

Modification 2: Training Instance Selection

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. The Queen was said to have acted as a buffer for the King's fierce temper during the process.

Modification 2: Training Instance Selection

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. The Queen was said to have acted as a buffer for the King's fierce temper during the process.

Modification 2: Training Instance Selection

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. The Queen was said to have acted as a buffer for the King's fierce temper during the process.

Modification 2: Training Instance Selection

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. The Queen was said to have acted as a buffer for the King's fierce temper during the process.

Modification 3: String Match Feature

- Focus on improving precision

Split the SOON_STR feature into several primitive features:

PRONOUN_STR

PROPER_NAME_STR

WORDS_STR

These restrict string matching to pronouns, proper names, and non-pronominal noun phrases, respectively.

Results (Learning Framework Modifications)

	MUC-6			MUC-7		
	R	P	F	R	P	F
Duplicated Soon Baseline	62.4	70.7	**66.3**	55.2	68.5	**61.2**
Learning Framework Modifications	62.4	73.5	**67.5**	56.3	71.5	**63.0**
clustering	60.4	74.4	66.7	54.3	72.1	62.0
instance selection	61.9	70.3	65.8	55.2	68.3	61.1
string match	62.4	70.8	66.3	56.5	69.6	62.3

§ gain in precision without loss in recall

Results (Learning Framework Modifications)

	MUC-6			MUC-7		
	R	P	F	R	P	F
Duplicated Soon Baseline	62.4	70.7	66.3	55.2	68.5	61.2
Learning Framework Modifications	62.4	73.5	**67.5**	56.3	71.5	**63.0**
clustering	60.4	74.4	**66.7**	54.3	72.1	**62.0**
instance selection	61.9	70.3	**65.8**	55.2	68.3	**61.1**
string match	62.4	70.8	**66.3**	56.5	69.6	**62.3**

- F-measure increases in most cases with weaker effects

Can we stop here?

	MUC-6			MUC-7		
	R	P	F	R	P	F
Duplicated Soon Baseline	62.4	70.7	66.3	55.2	68.5	61.2
Learning Framework Modifications	62.4	73.5	**67.5**	56.3	71.5	**63.0**
Best MUC System	59	72	**65**	56.1	68.8	**61.8**

NP Coreference Using Many Features

Will incorporating additional knowledge further improve performance?

- propose 41 additional features
- similar features have been explored in the past [Lappin & Leass, 1994; Grishman, 1995; Lin, 1995]
  - » but not in a machine learning framework
  - » previous work generally treats linguistic constraints as *broadly and unconditionally applicable hard constraints*

Additional Features

Lexical (8)	More complex string matching
Semantic (4)	Finer-grained semantic compatibility tests
Positional (1)	Distance in terms of number of paragraphs
Knowledge-based (2)	Naïve pronoun resolution, rule-based coref resolution
Grammatical (26)	NP type Grammatical role Linguistic constraints Linguistic preferences Heuristics

Results (Expanded Feature Set)

	MUC-6			MUC-7		
	R	P	F	R	P	F
Duplicated Soon Baseline	62.4	70.7	66.3	55.2	68.5	61.2
Learning Framework Modifications	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	-	66.3	-	-	62.1	-
proper nouns	-	84.2	-	-	77.7	-
common nouns	-	40.1	-	-	45.2	-

- with the expanded feature set performance drops!
- gain in recall but larger loss in precision

Results (Expanded Feature Set)

	MUC-6			MUC-7		
	R	P	F	R	P	F
Duplicated Soon Baseline	62.4	70.7	66.3	55.2	68.5	61.2
Learning Framework Modifications	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	-	66.3	-	-	62.1	-
proper nouns	-	84.2	-	-	77.7	-
common nouns	-	**40.1**	-	-	**45.2**	-

- poor precision on common nouns

Fixing Low Precision Rules

- example:
  - NP_i and NP_j ($i < j$) are *coreferent* if
    - » NP_i is a proper name
    - » NP_j is a definite noun phrase in the subject position
    - » NP_i and NP_j have the same semantic class
    - » NP_i and NP_j are at most one sentence apart
  - covers 38 examples with 18 exceptions
- manually discard features used primarily to induce low-precision rules for common noun resolution
- re-train coreference classifier using the reduced feature set consisting of 22-26 features

Results (Manual Feature Selection)

	MUC-6			MUC-7		
	R	P	F	R	P	F
Duplicated Soon Baseline	62.4	70.7	66.3	55.2	68.5	61.2
Learning Framework Modifications	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**
Hand-selected Features	64.1	74.9	**69.1**	57.4	70.8	**63.4**
pronouns	-	67.4	-	-	54.4	-
proper nouns	-	93.3	-	-	86.6	-
common nouns	-	63.0	-	-	64.8	-

- gain in precision and smaller loss in recall

Results (Manual Feature Selection)

	MUC-6			MUC-7		
	R	P	F	R	P	F
Duplicated Soon Baseline	62.4	70.7	66.3	55.2	68.5	61.2
Learning Framework Modifications	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
Hand-selected Features	64.1	74.9	69.1	57.4	70.8	63.4
pronouns	-	67.4	-	-	54.4	-
proper nouns	-	93.3	-	-	86.6	-
common nouns	-	**63.0**	-	-	**64.8**	-

- precision on common nouns rises from 40s to 60s.

Summary

- Investigated two methods to improve the best existing machine learning approach to noun phrase coreference resolution
  - Baseline: better NP identification -> gain in recall
  - Changes to ML framework: gain in precision (without loss in recall)
  - Expanded feature set: gain in recall but larger loss in precision
  - Manual feature selection: small loss in recall and larger gain in precision