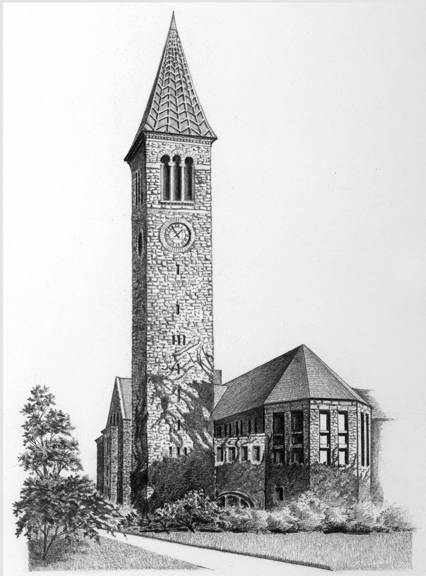
Machine Learning for Noun Phrase Coreference



Claire Cardie
Department of Computer Science
Cornell University

Last Class

- noun phrase coreference resolution
  - what it is
  - why it's important
  - why it's hard
- a (supervised) machine learning approach
- weakly supervised approaches

1. Illustrate how much you've learned
2. Realities of doing research in NLP+ML
3. Introduce some cool weakly supervised learning methods

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...

Why It's Hard

Many sources of information play a role

- string matching
- syntactic constraints
- number agreement
- gender agreement
- discourse focus
- recency
- syntactic parallelism
- semantic class
- world knowledge…

Why It's Hard

- No single source is a completely reliable indicator
- Identifying each of these features automatically, accurately, and in context, is hard

Last Class

- noun phrase coreference resolution
- a (supervised) machine learning approach
  - evaluation
  - problems…some solutions
- weakly supervised approaches

Knowledge-based approaches are still common. E.g.

- Lappin & Leass [1994]
- CogNIAC [Baldwin, 1996]

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* ?

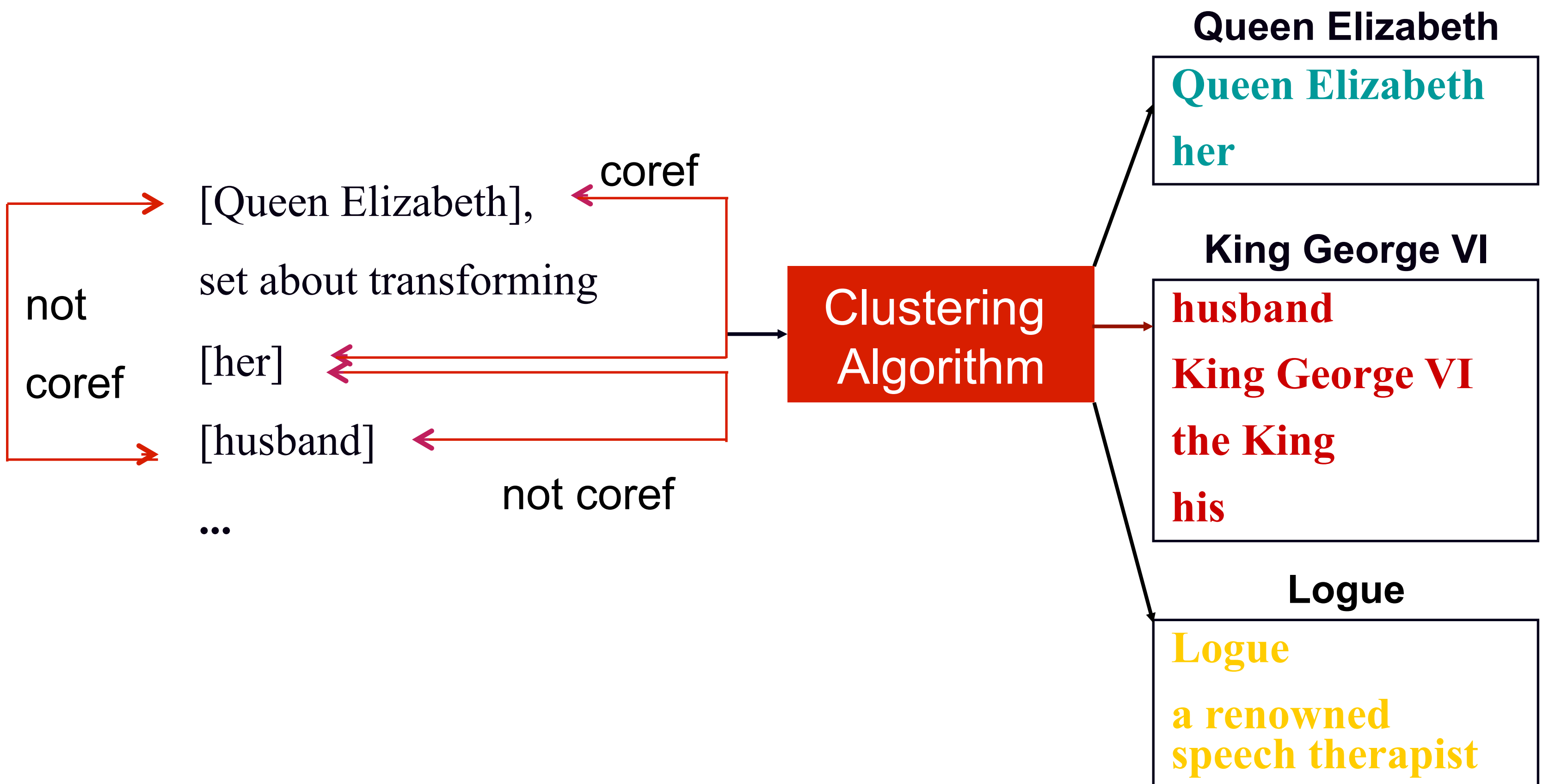
[Queen Elizabeth] set about transforming [her] [husband], ...

coref ?

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

Supervised Inductive Learning

Examples of NP pairs (features + class)

↓

ML Algorithm

↓

(novel) pair of NPs (features) → Concept description (program) → class

Training Data Creation

- Creating training instances
  - texts annotated with coreference information

candidate antecedent anaphor

  - one instance $inst(NP_i, NP_j)$ for each *ordered* pair of NPs
    - » 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_1, pronoun_2, demonstrative_2, indefinite_2, …
    - number, gender, animacy
    - appositive, predicate nominative
    - 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

Lie #1: 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);

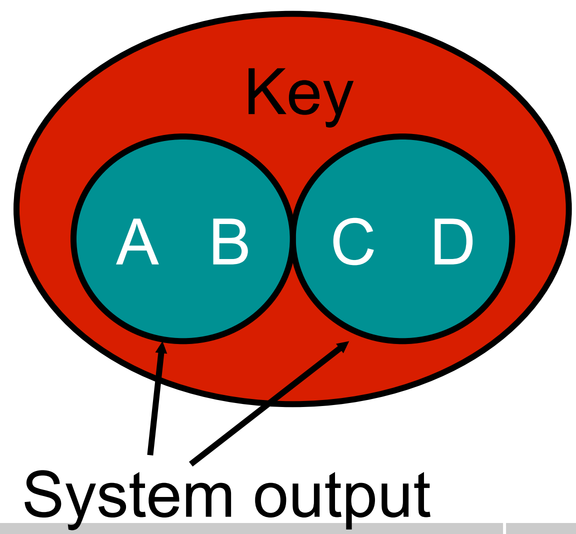
      Merge c_j and c_j .

Plan for the Talk

- noun phrase coreference resolution
- a (supervised) machine learning approach
  - evaluation
  - problems…some solutions
- weakly supervised approaches

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

Classifier for MUC-6 Data Set

```
ALIAS = C: +
ALIAS = I:
| SOON_STR_NONPRO = C:
| | ANIMACY = NA: -
| | ANIMACY = I: -
| | ANIMACY = C: +
| SOON_STR_NONPRO = I:
| | PRO_STR = C: +
| | PRO_STR = I:
| | | PRO_RESOLVE = C:
| | | | EMBEDDED_1 = Y: -
| | | | EMBEDDED_1 = N:
| | | | | PRONOUN_1 = Y:
| | | | | | ANIMACY = NA: -
| | | | | | ANIMACY = I: -
| | | | | | ANIMACY = C: +
| | | | | PRONOUN_1 = N:
| | | | | | MAXIMALNP = C: +
| | | | | | MAXIMALNP = I:
| | | | | | | WNCLASS = NA: -
| | | | | | | WNCLASS = I: +
| | | | | | | WNCLASS = C: +
| | | PRO_RESOLVE = I:
| | | | APPOSITIVE = I: -
| | | | APPOSITIVE = C:
| | | | | GENDER = NA: +
| | | | | GENDER = I: +
| | | | | GENDER = C: -
```