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
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...
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...
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**...
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

\[ \text{coref?} \quad \text{coref?} \]

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

\[ \text{not coref?} \]

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

- coordinates pairwise coreference decisions

A Machine Learning Approach

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

not coref

King George VI

Queen Elizabeth

her

husband

the King

his

Logue

a renowned speech therapist

Clustering Algorithm

not coref

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 \( (NP_1, NP_2, \ldots, NP_n) \)

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

Let \( NP_k = \text{noun phrase in } S(NP_j) \text{ with highest confidence} \)

\( c_j = c_j \cup c_k \)
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

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th>MUC-7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>Baseline</td>
<td>40.7</td>
<td>73.5</td>
</tr>
<tr>
<td>Worst MUC System</td>
<td>36</td>
<td>44</td>
</tr>
<tr>
<td>Best MUC System</td>
<td>59</td>
<td>72</td>
</tr>
</tbody>
</table>
## Baseline System: Results

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th></th>
<th>MUC-7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
</tr>
<tr>
<td>Baseline</td>
<td>40.7</td>
<td>73.5</td>
<td>52.4</td>
<td>27.2</td>
</tr>
<tr>
<td>Best MUC System</td>
<td>59</td>
<td>72</td>
<td>65</td>
<td>56.1</td>
</tr>
<tr>
<td>Worst MUC System</td>
<td>36</td>
<td>44</td>
<td>40</td>
<td>52.5</td>
</tr>
</tbody>
</table>
Problem 1

Coreference is an equivalence relation
– loss of transitivity

\[ \text{[Queen Elizabeth] set about transforming [her] [husband], ...} \]

\[ \text{not coref ?} \]
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
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...
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
Negative Example Selection

Step 1: Create all possible negative instances from $NP9$
Step 2: Locate the farthest antecedent of $NP_9$, $f(NP_9)$
Step 2: Locate the farthest antecedent of $NP9$, $f(NP9)$
Step 3: Remove all instances involving NPs that precede $f(NP9)$

farthest antecedent
Step 3: Remove all instances involving NPs that precede $f(NP9)$
Results (Negative Example Selection)

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th></th>
<th>MUC-7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
</tr>
<tr>
<td>Baseline</td>
<td>40.7</td>
<td>73.5</td>
<td><strong>52.4</strong></td>
<td>27.2</td>
</tr>
<tr>
<td>NEG-SELECT</td>
<td>46.5</td>
<td>67.8</td>
<td><strong>55.2</strong></td>
<td>37.4</td>
</tr>
</tbody>
</table>

- % 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 $BestRule = \text{best rule in } R$
        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)

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th></th>
<th>MUC-7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
</tr>
<tr>
<td>Baseline</td>
<td>40.7</td>
<td>73.5</td>
<td>52.4</td>
<td>27.2</td>
</tr>
<tr>
<td>NEG-SELECT</td>
<td>46.5</td>
<td>67.8</td>
<td>55.2</td>
<td>37.4</td>
</tr>
<tr>
<td>POS-SELECT</td>
<td>53.1</td>
<td>80.8</td>
<td>64.1</td>
<td>41.1</td>
</tr>
<tr>
<td>NEG-SELECT + POS-SELECT</td>
<td>63.4</td>
<td>76.3</td>
<td>69.3</td>
<td>59.5</td>
</tr>
</tbody>
</table>

F-measure increases by 12% using POS-SELECT
Results (Positive Example Selection)

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th></th>
<th>MUC-7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
</tr>
<tr>
<td>Baseline</td>
<td>40.7</td>
<td>73.5</td>
<td>52.4</td>
<td>27.2</td>
</tr>
<tr>
<td>NEG-SELECT</td>
<td>46.5</td>
<td>67.8</td>
<td>55.2</td>
<td>37.4</td>
</tr>
<tr>
<td>POS-SELECT</td>
<td>53.1</td>
<td>80.8</td>
<td>64.1</td>
<td>41.1</td>
</tr>
<tr>
<td>NEG-SELECT + POS-SELECT</td>
<td>63.4</td>
<td>76.3</td>
<td>69.3</td>
<td>59.5</td>
</tr>
</tbody>
</table>

F-measure increases by 16-17% using both NEG-SELECT and POS-SELECT
Results (Positive Example Selection)

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th></th>
<th>MUC-7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
</tr>
<tr>
<td>Baseline</td>
<td>40.7</td>
<td>73.5</td>
<td>52.4</td>
<td>27.2</td>
</tr>
<tr>
<td>NEG-SELECT</td>
<td>46.5</td>
<td>67.8</td>
<td>55.2</td>
<td>37.4</td>
</tr>
<tr>
<td>POS-SELECT</td>
<td>53.1</td>
<td>80.8</td>
<td>64.1</td>
<td>41.1</td>
</tr>
<tr>
<td>NEG-SELECT + POS-SELECT</td>
<td>63.4</td>
<td>76.3</td>
<td>69.3</td>
<td>59.5</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th></th>
<th>MUC-7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
</tr>
<tr>
<td><strong>Baseline</strong></td>
<td>40.7</td>
<td>73.5</td>
<td>52.4</td>
<td>27.2</td>
</tr>
<tr>
<td><strong>NEG-SELECT</strong></td>
<td>46.5</td>
<td>67.8</td>
<td>55.2</td>
<td>37.4</td>
</tr>
<tr>
<td><strong>POS-SELECT</strong></td>
<td>53.1</td>
<td>80.8</td>
<td>64.1</td>
<td>41.1</td>
</tr>
<tr>
<td><strong>NEG-SELECT + POS-SELECT</strong></td>
<td>63.4</td>
<td>76.3</td>
<td>69.3</td>
<td>59.5</td>
</tr>
<tr>
<td><strong>NEG-SELECT + POS-SELECT + RULE-SELECT</strong></td>
<td>63.3</td>
<td>76.9</td>
<td>69.5</td>
<td>54.2</td>
</tr>
<tr>
<td><strong>NEG-SELECT + POS-SELECT (more data)</strong></td>
<td>64.8</td>
<td>70.6</td>
<td>67.6</td>
<td>60.0</td>
</tr>
</tbody>
</table>

*pruning corpus*
- MUC-6: MUC-7 formal
- MUC-7: MUC-6 formal
## Results (Rule Selection)

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th></th>
<th>MUC-7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
</tr>
<tr>
<td>Baseline</td>
<td>40.7</td>
<td>73.5</td>
<td>52.4</td>
<td>27.2</td>
</tr>
<tr>
<td>NEG-SELECT</td>
<td>46.5</td>
<td>67.8</td>
<td>55.2</td>
<td>37.4</td>
</tr>
<tr>
<td>POS-SELECT</td>
<td>53.1</td>
<td>80.8</td>
<td>64.1</td>
<td>41.1</td>
</tr>
<tr>
<td>NEG-SELECT + POS-SELECT</td>
<td>63.4</td>
<td>76.3</td>
<td><strong>69.3</strong></td>
<td>59.5</td>
</tr>
<tr>
<td>NEG-SELECT + POS-SELECT + RULE-SELECT</td>
<td>63.3</td>
<td>76.9</td>
<td><strong>69.5</strong></td>
<td>54.2</td>
</tr>
<tr>
<td>NEG-SELECT + POS-SELECT (more data)</td>
<td>64.8</td>
<td>70.6</td>
<td>67.6</td>
<td>60.0</td>
</tr>
</tbody>
</table>

- gains in precision; increase in F-measure
- effective at improving precision
### Results (Rule Selection)

<table>
<thead>
<tr>
<th>Rule Selection</th>
<th>MUC-6</th>
<th>MUC-7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>Baseline</td>
<td>40.7</td>
<td>73.5</td>
</tr>
<tr>
<td>NEG-SELECT</td>
<td>46.5</td>
<td>67.8</td>
</tr>
<tr>
<td>POS-SELECT</td>
<td>53.1</td>
<td>80.8</td>
</tr>
<tr>
<td>NEG-SELECT + POS-SELECT</td>
<td>63.4</td>
<td>76.3</td>
</tr>
<tr>
<td>NEG-SELECT + POS-SELECT + RULE-SELECT</td>
<td>63.3</td>
<td>76.9</td>
</tr>
<tr>
<td>NEG-SELECT + POS-SELECT (more data)</td>
<td>64.8</td>
<td>70.6</td>
</tr>
</tbody>
</table>

 RULE-SELECT has made a more effective use of the additional data provided by the pruning corpus
Comparison with Best MUC Systems

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th>MUC-7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>NEG-SELECT + POS-SELECT + RULE-SELECT</td>
<td>63.3</td>
<td>76.9</td>
</tr>
<tr>
<td>Best MUC System</td>
<td>59</td>
<td>72</td>
</tr>
</tbody>
</table>

S 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.

<table>
<thead>
<tr>
<th>Properties of Coreference</th>
<th>Problems</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coref is a rare relation</td>
<td>Skewed distributions</td>
<td>Negative example selection</td>
</tr>
<tr>
<td>Coref is a discourse-level problem</td>
<td>Inclusion of hard training instances</td>
<td>Positive example selection</td>
</tr>
<tr>
<td>Coref is an equivalence relation</td>
<td>Loss of transitivity</td>
<td>Rule pruning</td>
</tr>
</tbody>
</table>