Supervised Models for Coreference Resolution

Altaf Rahman and Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas
What is Coreference Resolution?

– Identify all noun phrases (mentions) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affair experience as a former First Lady.
What is Coreference Resolution?

– Identify all noun phrases (mentions) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affair experience as a former First Lady.
What is Coreference Resolution?

– Identify all noun phrases (mentions) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. She had foreign affair experience.
What is Coreference Resolution?

– Identify all noun phrases (mentions) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affair experience as a former First Lady.
What is Coreference Resolution?

– Identify all noun phrases (mentions) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affair experience as a former First Lady.
Plan for the talk

• Existing learning based coreference models
  – Overview
  – Implementation details
• Our cluster ranking model
• Evaluation
• Summary
Plan for the talk

• Existing learning based coreference models
  – Overview
  – Implementation details
• Our cluster ranking model
• Evaluation
• Summary
Existing learning based coreference models

- Mention-Pair (MP) model
- Entity-Mention (EM) model
- Mention-Ranking (MR) model
Mention-Pair (MP) Model

• Soon et al. 2001 ; Ng and Cardie 2002
• Classifies whether two mentions are coreferent or not.
• Weaknesses
  – Insufficient information to make an informed coreference decision.
Mention-Pair (MP) Model

• Soon et al. 2001; Ng and Cardie 2002
• Classifies whether two mentions are coreferent or not.
• Weaknesses
  – Insufficient information to make an informed coreferenced decision.

Barack Obama ..................Hillary Rodham Clinton ........his
 ................ secretary of state ..........................He ............her
Mention-Pair (MP) Model

• Soon et al. 2001; Ng and Cardie 2002
• Classifies whether two mentions are coreferent or not.

• Weaknesses
  – Insufficient information to make an informed coreference decision.
  – Each candidate antecedents is considered independently of the others.
Mention-Pair (MP) Model

- Soon et al. 2001; Ng and Cardie 2002
- Classifies whether **two mentions** are coreferent or not.
- Weaknesses
  - Insufficient information to make an informed coreference decision.
  - Each candidate antecedents is considered independently of the others.

| Barack Obama ..........Hillary Rodham Clinton ........his ..........secretary of state ............the President........He .............her | 13 |
Entity-Mention (EM) Model

- Pasula et al. 2003; Luo et al. 2004; Yang et al. 2004
- Classifies whether a mention and a preceding, possibly partially formed cluster are coreferent or not.
- Strength
  - Improved expressiveness.
  - Allows the computation of cluster level features
- Weakness
  - Each candidate cluster is considered independently of the others.

Barack Obama .................. Hillary Rodham Clinton ...... his
............... secretary of state ......................... He .............. her
Mention-Ranking (MR) Model

• Denis & Baldridge 2007, 2008
• Imposes a **ranking** on a set of candidate antecedents

• **Strength**
  – Considers all the candidate antecedents simultaneously

• **Weakness**
  – Insufficient information to make an informed coreference decision.

Barack Obama ..................Hillary Rodham Clinton ........his

........... secretary of state .......................He ...............her
Goal

• Propose a cluster ranking (CR) model
  – ranks all the preceding clusters for a mention
  – combines the strengths of EM and MR models
    • Improve expressiveness by using cluster level features.
    • Considers all the candidate clusters simultaneously
Plan for the talk

• Existing learning based coreference models
  – Overview
  – Implementation details
• Our cluster ranking model
• Evaluation
• Summary
Mention-Pair (MP) Model

• Classifies whether two mentions are coreferent or not

• Training
  – Each instance is created between $m_j$ and $m_k$
  – 39 features
Mention-Pair (MP) Model

• Classifies whether two mentions are coreferent or not

• Training
  – Each instance is created between $m_j$ and $m_k$
  – 39 features
    § Features describing $m_j$ a candidate antecedent
    § Pronoun ? Subject ? Nested ?
Mention-Pair (MP) Model

- Classifies whether two mentions are coreferent or not
- Training
  - Each instance is created between $m_j$ and $m_k$
  - 39 features
    - Features describing $m_j$, a candidate antecedent
      - Pronoun ? Subject ? Nested ?
    - Features describing $m_k$, the mention to be resolved
      - Number ? Gender ? Pronoun2 ? Semantic Class ? Animacy ?
Mention-Pair (MP) Model

• Classifies whether two mentions are coreferent or not

• Training
  – Each instance is created between $m_j$ and $m_k$
  – 39 features
    ￭ Features describing $m_j$, a candidate antecedent
      ￭ Pronoun ? Subject ? Nested ?
    ￭ Features describing $m_k$, the mention to be resolved
      ￭ Number ? Gender ? Pronoun2 ? Semantic Class ? Animacy ?
    ￭ Features describing the relation between $m_j$, a candidate antecedent $m_k$, the mention to be resolved
Mention-Pair (MP) Model

• Training instance creation (Soon et al.)
  – create
    § positive instance for each anaphoric mention, \( m_j \) and its closest preceding antecedent mention, \( m_i \)
    § negative instance for \( m_j \) and each intervening mention, \( m_{i+1}, m_{i+2}, \ldots, m_{j-1} \)
    § No instance for non-anaphors.
Mention-Pair (MP) Model

• Training instance creation (Soon et al.)
  – create
    § positive instance for each anaphoric mention, \( m_j \), and its closest preceding antecedent mention, \( m_i \)
    § negative instance for \( m_j \) and each intervening mention, \( m_{i+1}, m_{i+2}, \ldots, m_{j-1} \)
    § No instance for non-anaphors.

• Testing (Soon et al.)
  – For each \( m_j \)
    § Select as the antecedent of \( m_j \) the closest preceding mention that is classified as the coreferent with \( m_j \)
    § if no such mention exist \( m_j \) is considered non-anaphoric
Entity-Mention (EM) Model

• Classifies whether a mention and a preceding cluster are coreferent or not.

• Training
  – Each instance is between a mention and a preceding partially formed cluster.
  – Cluster level features
• Training instance creation:
  - Positive Instance
    • For each anaphoric mention $m_k$ and preceding cluster $c_j$
      to which it belongs
Entity-Mention (EM) Model

• Training instance creation:
  – Positive Instance
    • For each anaphoric mention $m_k$ and preceding cluster $c_j$ to which it belongs

  – No instance for non-anaphors.
• Training instance creation:
  – Positive Instance
    • For each anaphoric mention $m_k$ and preceding cluster $c_j$ to which it belongs
  – No instance for non-anaphors.
  – Negative Instance
    • For each anaphoric mention $m_k$ and partial cluster whose last mention appears between $m_k$ and its closest antecedent in $c_j$ to which it belongs
For mention $m_6$

- Positive instance:
  - features between $m_6$ and the cluster \{m_2, m_4\}
For mention $m_6$

- **Positive instance:**
  - features between $m_6$ and the cluster $\{m_2, m_4\}$

- **Negative instance:**
  - features between $m_6$ and the cluster $\{m_1, m_5\}$
For mention $m_6$

- **Positive instance**: features between $m_6$ and the cluster $\{m_2, m_4\}$
- **Negative instance**: features between $m_6$ and the cluster $\{m_1, m_5\}$
- No negative instance created between $m_6$ and $\{m_3\}$
Entity-Mention (EM) Model

• Testing
  – Like MP model except now we resolve the mention to the closest *preceding cluster* that is classified as coreferent.
Entity-Mention(EM) Model continued

- For each relational feature used by MP model we create a set of **cluster level** features
  - Example: Gender = **Compatible** ? **Incompatible** ? **Not Applicable**?

#C    = 0 Normalized_C = 0/3   =   0.00 #I      = 2 Normalized_I = 2/3   =   0.66
#NA  = 1 Normalized_NA= 1/3   =   0.33

0.0 = NONE
<0.5 = MOST-FALSE
>0.5 = MOST-TRUE
1.0 = ALL

GenderC= NONE
GenderI= MOST-TRUE
GenderNA= MOST-FALSE
Entity-Mention (EM) Model continued

• For each relational feature used by MP model we create a set of **cluster level** features
  
  – Example: Gender = **Compatible**? **Incompatible**? **Not Applicable**?

<table>
<thead>
<tr>
<th>Hillary Clinton</th>
<th>whose</th>
<th>she</th>
<th>him</th>
</tr>
</thead>
<tbody>
<tr>
<td>(female)</td>
<td>(neutral)</td>
<td>(female)</td>
<td>(male)</td>
</tr>
</tbody>
</table>

For the mention “him”

- \#C = 0 \hspace{1cm} \text{Normalized}_C = 0/3 = 0.00
- \#I = 2 \hspace{1cm} \text{Normalized}_I = 2/3 = 0.66
- \#NA = 1 \hspace{1cm} \text{Normalized}_{NA} = 1/3 = 0.33
For each relational feature used by MP model we create a set of cluster level features.

- Example: Gender = Compatible ? Incompatible ? Not Applicable?

For the mention “him”

- #C = 0  Normalized_C = 0/3 = 0.00
- #I = 2  Normalized_I = 2/3 = 0.66
- #NA = 1  Normalized_NA = 1/3 = 0.33

0.0 = NONE
<0.5 = MOST-FALSE
>0.5 = MOST-TRUE
1.0 = ALL

Hillary Clinton.... whose.............she.............him....... (female) (neutral) (female) (male)
Entity-Mention(EM) Model continued

• For each relational feature used by MP model we create a set of **cluster level** features
  — Example: Gender = **Compatible** ? **Incompatible** ? **Not Applicable**?

<table>
<thead>
<tr>
<th>Hillary Clinton... whose... she... him...</th>
<th>(female)</th>
<th>(neutral)</th>
<th>(female)</th>
<th>(male)</th>
</tr>
</thead>
</table>

For the mention “him”

— #C = 0 Normalized_C = 0/3 = 0.00
— #I = 2 Normalized_I = 2/3 = 0.66
— #NA = 1 Normalized_NA = 1/3 = 0.33

0.0 = **NONE**  
<0.5 = **MOST-FALSE**  
>0.5 = **MOST-TRUE**  
1.0 = **ALL**

— GenderC=**NONE** GenderI=**MOST-TRUE** GenderNA=**MOST-FALSE**
Mention-Ranking (MR) Model

• Ranks a set of candidate antecedents for each mention

• Training
  – Each instance represents 2 mentions \((m_j, m_k)\)
  – Same 39 features as in Mention-pair (MP) model
  – Used the SVM ranker learning algorithm (Joachims 2002).
• Training instance creation
  – According to Soon et al’s method
    § Rank value is 2 if positive
    § Otherwise rank 1
Mention-Ranking (MR) Model continued

• Training instance creation
  – According to Soon et al’s method
    § Rank value is 2 if positive
    § Otherwise rank 1

• Testing
  – First check anaphoricity of \( m_j \) using separate anaphoricity classifier.
    § If \( m_j \) is non-anaphoric then create a new cluster.
    § Otherwise, resolve \( m_j \) to the highest ranked \( m_k \) among ALL the candidate antecedents.
Plan for the talk

• Existing learning based coreference models
  – Overview
  – Implementation details
• Our cluster ranking model
• Evaluation
• Summary
Cluster Ranking (CR) model

• Combines the strength of MP and EM model
• Ranks all the preceding clusters for each mention
• Training
  – Each instance is comprised of features between a mention \( m_k \) and its preceding cluster \( c_j \)
  – Instances are created like the EM model.
  – Rank values are assigned like the MR model.
Cluster Ranking (CR) Model

• Since no instances were created from the non-anaphors, we need to rely on a separate classifier to determine whether a mention is anaphoric

• Problem
  • Errors in anaphoricity determination can be propagated to coreference resolution

• Hypothesis
  • A model for jointly determining anaphoricity and coreference resolution can overcome this problem with the pipeline approach
Joint anaphoricity determination and coreference resolution

• Idea
  • Create additional training instances from non-anaphors
    – If $m_k$ is non-anaphoric, assign rank value
      § 1 to each instance formed between $m_k$ and each preceding cluster
      § 2 to all instances formed between $m_k$ and a (hypothetical) null cluster
      § Use only features describing $m_k$
  • Same idea can be applied to create joint version of MR model
Plan for the talk

• Existing learning based coreference models
  – Overview
  – Implementation details
• Our cluster ranking model
• Evaluation
• Summary
Evaluation

• Experimental setup
  § ACE2005 corpus
  § 599 documents of 7 sources - BC, BN, CTS, NW, UN, WL
  § 80% for Training and 20% for Testing.
  § True mentions
  § System mentions (extracted by a learned mention extractor that is trained on train docs)
  § Scoring programs
    • MUC (Vilain et al. 1995)
    • CEAF (Lu et al. 2005)
    • $B^3$ (Bagga & Baldwin 1998)
  § Recall, precision and f-measure
System Mention Results

• Baseline Systems
  – MP model
  – EM model
  – MR model

All models are trained using SVM-light. All learning parameters are set to default values.
## System Mention Results (MP baseline)

<table>
<thead>
<tr>
<th>Coreference Models</th>
<th>CEAF</th>
<th>B³</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>MP model</td>
<td>56.1</td>
<td>51.0</td>
</tr>
</tbody>
</table>

- CEAF F score is 53.4
- B³ F score is 54.1
### System Mention Results (EM baseline)

<table>
<thead>
<tr>
<th>Coreference Models</th>
<th>CEAF</th>
<th>(B^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>MP model</td>
<td>56.1</td>
<td>51.0</td>
</tr>
<tr>
<td>EM model</td>
<td>56.3</td>
<td>50.2</td>
</tr>
</tbody>
</table>

- F score change is insignificant despite the improved expressiveness of EM model.
- Similar trends have been reported by Luo et al 2004.
2 architectures for using anaphoricity information
  – Pipeline
  – Joint
System Mention Results (MR baseline)

<table>
<thead>
<tr>
<th>Coreference Models</th>
<th>CEAF</th>
<th>B³</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>MP model</td>
<td>56.1</td>
<td>51.0</td>
</tr>
<tr>
<td>EM model</td>
<td>56.3</td>
<td>50.2</td>
</tr>
<tr>
<td>MR model (pipeline)</td>
<td>51.6</td>
<td>56.7</td>
</tr>
<tr>
<td>MR model (joint)</td>
<td>53.0</td>
<td>58.5</td>
</tr>
</tbody>
</table>

- 2 architectures for using anaphoricity information
  - Pipeline
  - Joint
- Both show significant improvements over MP baseline
### System Mention Results (MR baseline)

<table>
<thead>
<tr>
<th>Coreference Models</th>
<th>CEAF</th>
<th>B³</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>MP model</td>
<td>56.1</td>
<td>51.0</td>
</tr>
<tr>
<td>EM model</td>
<td>56.3</td>
<td>50.2</td>
</tr>
<tr>
<td>MR model (pipeline)</td>
<td>51.6</td>
<td>56.7</td>
</tr>
<tr>
<td>MR model (joint)</td>
<td>53.0</td>
<td>58.5</td>
</tr>
</tbody>
</table>

- 2 architectures for using anaphoricity information
  - Pipeline
  - Joint
- Both show significant improvements over MP baseline
- Joint architecture outperforms pipeline architecture.
# System Mention Results (CR model)

<table>
<thead>
<tr>
<th>Coreference Models</th>
<th>CEAF</th>
<th></th>
<th>B³</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
<td>F</td>
<td>Precision</td>
</tr>
<tr>
<td>MP model</td>
<td>56.1</td>
<td>51.0</td>
<td>53.4</td>
<td>50.8</td>
</tr>
<tr>
<td>EM model</td>
<td>56.3</td>
<td>50.2</td>
<td>53.1</td>
<td>51.2</td>
</tr>
<tr>
<td>MR model (pipeline)</td>
<td>51.6</td>
<td>56.7</td>
<td>54.1</td>
<td>52.3</td>
</tr>
<tr>
<td>MR model (joint)</td>
<td>53.0</td>
<td>58.5</td>
<td>55.6</td>
<td>50.4</td>
</tr>
<tr>
<td>CR model (pipeline)</td>
<td>54.1</td>
<td>59.3</td>
<td><strong>56.6</strong></td>
<td>55.3</td>
</tr>
<tr>
<td>CR model (joint)</td>
<td>56.7</td>
<td>62.6</td>
<td><strong>59.5</strong></td>
<td>54.4</td>
</tr>
</tbody>
</table>
### System Mention Results (CR model)

<table>
<thead>
<tr>
<th>Coreference Models</th>
<th>CEAF</th>
<th>B³</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>MP model</td>
<td>56.1</td>
<td>51.0</td>
</tr>
<tr>
<td>EM model</td>
<td>56.3</td>
<td>50.2</td>
</tr>
<tr>
<td>MR model (pipeline)</td>
<td>51.6</td>
<td>56.7</td>
</tr>
<tr>
<td>MR model (joint)</td>
<td>53.0</td>
<td>58.5</td>
</tr>
<tr>
<td>CR model (pipeline)</td>
<td>54.1</td>
<td>59.3</td>
</tr>
<tr>
<td>CR model (joint)</td>
<td>56.7</td>
<td>62.6</td>
</tr>
</tbody>
</table>

- Cluster-ranking model outperforms Mention-ranking model
- Due to simultaneous gains in recall and precision
System Mention Results (CR model)

<table>
<thead>
<tr>
<th>Coreference Models</th>
<th>CEAF</th>
<th>B³</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>MP model</td>
<td>56.1</td>
<td>51.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EM model</td>
<td>56.3</td>
<td>50.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MR model (pipeline)</td>
<td>51.6</td>
<td>56.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MR model (joint)</td>
<td>53.0</td>
<td>58.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CR model (pipeline)</td>
<td>54.1</td>
<td>59.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CR model (joint)</td>
<td>56.7</td>
<td>62.6</td>
</tr>
</tbody>
</table>

- Cluster-ranking model outperforms Mention-ranking model
  - Due to simultaneous gains in recall and precision
System Mention Results (CR model)

<table>
<thead>
<tr>
<th>Coreference Models</th>
<th>CEAF</th>
<th>B³</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>MP model</td>
<td>56.1</td>
<td>51.0</td>
</tr>
<tr>
<td>EM model</td>
<td>56.3</td>
<td>50.2</td>
</tr>
<tr>
<td>MR model (pipeline)</td>
<td>51.6</td>
<td>56.7</td>
</tr>
<tr>
<td>MR model (joint)</td>
<td>53.0</td>
<td>58.5</td>
</tr>
<tr>
<td>CR model (pipeline)</td>
<td>54.1</td>
<td>59.3</td>
</tr>
<tr>
<td>CR model (joint)</td>
<td>56.7</td>
<td>62.6</td>
</tr>
</tbody>
</table>

- Joint architecture outperforms pipeline architecture.
  - Due to simultaneous gains in recall and precision
Plan for the talk

• Existing learning based coreference models
  – Overview
  – Implementation details
• Our cluster ranking model
• Evaluation
• Summary
Summary

• Proposed a cluster ranking approach
  – Combines the strengths of EM and MR models.
  – Jointly learns coreference resolution and anaphoricity determination
  – Significantly outperforms three commonly-used learning-based coreference models