Coreference Semantics from Web Features

Mohit Bansal and Dan Klein
UC Berkeley
Motivation
Example

When *Obama* met *Jobs*, *the president* discussed the economy, technology, and education. *His election campaign is expected to* …

- World knowledge needed:
  - Obama is the president of the US
  - Presidents, not CEOs, have election campaigns
Example

When Obama met Jobs, the president discussed ...
Example

When Obama met Jobs ... His election campaign ...
This Work

- No new model

- Simple, principled features that subsumes previous work

- Features computed from Google n-grams only
Baseline System
Reconcile (Stoyanov et al., 2009)
Pairwise Supervised Coreference

\[ f(m_1, m_2) \xrightarrow{\text{score}} \text{score}(f(m_1, m_2)) \]

When \textbf{Obama} met Jobs, the \textbf{president} discussed ...
## Pairwise Features

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>POSITIONAL</td>
<td>Distance in # Sentences</td>
</tr>
<tr>
<td>LEXICAL</td>
<td>Approximate String Match</td>
</tr>
<tr>
<td>GRAMMATICAL</td>
<td>Number Agreement</td>
</tr>
<tr>
<td></td>
<td>Gender Agreement</td>
</tr>
<tr>
<td></td>
<td>Appositive Relationship</td>
</tr>
<tr>
<td>SEMANTIC</td>
<td>WordNet Synonyms</td>
</tr>
<tr>
<td></td>
<td>Alias</td>
</tr>
</tbody>
</table>

*Soon et al. 2001, Ng and Cardie 2002, Stoyanov et al. 2009*
One Important Change

- Averaged Perceptron ➔ Decision Tree

![Graph showing MUC F1 scores for AvgPerc and DecTree with values 65.9 for AvgPerc and 69.5 for DecTree.]

[Quinlan 1993, Hall et al., 2009]
World Knowledge via Web Features
Barack Obama, the 44th president of the United States, will be sworn in …
Barack Obama, the 44th president of the United States, will be sworn in …

Korea and other countries will be participating in this important event …
<table>
<thead>
<tr>
<th><strong>World Knowledge</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Barack Obama</strong>, the 44th <strong>president</strong> of the United States, will be sworn in …</td>
</tr>
<tr>
<td><strong>Korea</strong> and other <strong>countries</strong> will be participating in this important event …</td>
</tr>
<tr>
<td><strong>Asia</strong> is the largest <strong>continent</strong>, located primarily in the eastern and northern …</td>
</tr>
</tbody>
</table>
Web Features

\[ f(m_1, m_2) \]

\[ (m_1, m_2) \quad \text{score}(f(\cdot)) \]

\[ f(h_1, h_2) \text{ (Web)} \]
Feature Categories

- General Co-occurrence
- Hearst Co-occurrence [Poesio et al., 2004; Markert & Nissim, 2005; Kobdani et al., 2011]*
- Entity-Based Context
- Distributional Soft Clustering [Daume’ III & Marcu, 2005]*
- Pronoun Context [Yang et al., 2005; Bergsma & Lin, 2006]*
1. General Co-occurrence

\[ \text{count(} \text{president } * \text{ leader} \text{)} = 11383 \]

\[ \text{count(} \text{voter } * \text{ leader} \text{)} = 95 \]

When the president met a voter, the leader said …
1. General Co-occurrence

\[ C_{12} \]

where

\[ c_{12} = \text{count}("h_1 \star h_2") + \text{count}("h_1 \star \star h_2") + \text{count}("h_1 \star \star \star h_2") \]
1. General Co-occurrence

\[
\frac{C_{12}}{C_1 \cdot C_2}
\]

where

\[
c_{12} = \text{count}(“h_1 \star h_2”) + \text{count}(“h_1 \star \star h_2”) + \text{count}(“h_1 \star \star \star h_2”)
\]

\[
c_1 = \text{count}(“h_1”)
\]

\[
c_2 = \text{count}(“h_2”)
\]
1. General Co-occurrence

\[
bin \left( \log \left( \frac{c_{12}}{c_1 \cdot c_2} \right) \right)
\]

where

\[
c_{12} = \text{count}("h_1 \ast h_2") + \text{count}("h_1 \ast \ast h_2") + \text{count}("h_1 \ast \ast \ast h_2")
\]

\[
c_1 = \text{count}("h_1")
\]

\[
c_2 = \text{count}("h_2")
\]
China and Japan are geographically separated only by a relatively narrow stretch of ocean. China has strongly influenced Japan with its writing system, architecture, culture, religion, philosophy, and law.
Asia is the largest continent of the world and has the highest population. It is located primarily in the eastern and northern hemispheres. It covers 8.7% of the Earth’s total surface area and 30% of its land area.
ACL and other scientific societies are for people working on problems involving natural language and computation. An annual meeting is held each summer in locations where significant computational linguistics …
2. Hearst Co-occurrence

\[
\text{count}(\text{president } \ast \text{ leader}) \quad \text{count}(\text{voter } \ast \text{ leader}) \\
= 752 \quad = 0
\]

When the president met a voter, the leader said …
2. Hearst Co-occurrence

- Hypernymy patterns:
  \[ h_1 \{is \mid are \mid was \mid were\} \{a \mid an \mid the\}? h_2 \]
2. Hearst Co-occurrence

Hypernymy patterns:

- $h_1 \{is | are | was | were\} \{a | an | the\}? h_2$
- $h_1 \{and | or\} \{other | the other | another\} h_2$
2. Hearst Co-occurrence

Hypernymy patterns:

- \( h_1 \{ \text{is} \mid \text{are} \mid \text{was} \mid \text{were} \} \{ \text{a} \mid \text{an} \mid \text{the} \}? h_2 \)
- \( h_1 \{ \text{and} \mid \text{or} \} \{ \text{other} \mid \text{the other} \mid \text{another} \} \ h_2 \)
- \( h_1 \text{ other than } \{ \text{a} \mid \text{an} \mid \text{the} \}? h_2 \)
- \( h_1 \text{ such as } \{ \text{a} \mid \text{an} \mid \text{the} \}? h_2 \)
2. Hearst Co-occurrence

Hypernymy patterns:

- $h_1 \{is \mid are \mid was \mid were\} \{a \mid an \mid the\} \ L h_2$
- $h_1 \{and \mid or\} \{other \mid the \text{other} \mid another\} \ R h_2$
- $h_1 \text{other than} \{a \mid an \mid the\} \ L h_2$
- $h_1 \text{such as} \{a \mid an \mid the\} \ R h_2$
- $h_1 \text{, including} \{a \mid an \mid the\} \ R h_2$
- $h_1 \text{, especially} \{a \mid an \mid the\} \ L h_2$

[Hearst 1992]
2. Hearst Co-occurrence

Hyponymy patterns:

- $h_1 \{is \mid are \mid was \mid were\} \{a \mid an \mid the\} \? h_2$
- $h_1 \{and \mid or\} \{other \mid the other \mid another\} \ h_2$
- $h_1 \text{other than} \{a \mid an \mid the\} \? h_2$
- $h_1 \text{such as} \{a \mid an \mid the\} \? h_2$
- $h_1 \text{, including} \{a \mid an \mid the\} \? h_2$
- $h_1 \text{, especially} \{a \mid an \mid the\} \? h_2$
- $h_1 \text{ of} \{the \mid all\} \? h_2$
3. Entity-Based Context

\[ h_1 = \text{president} \]

- president is elected
- president is authorized
- president is responsible
- president is the chief
- president is above
- president is the head

\[ h_2 = \text{leader} \]

- leader is responsible
- leader is expected
- leader is able
- leader is elected
- leader is chosen
- leader is best

\[ \cap \]
3. Entity-Based Context

\[ h_1 = \textit{president} \quad \cap \quad h_2 = \textit{leader} \]

\textit{president} is \textit{elected}
\textit{president} is \textit{authorized}
\textit{president} is \textit{responsible}
\textit{president} is \textit{the chief}
\textit{president} is \textit{above}
\textit{president} is \textit{the head}
\ldots

\textit{leader} is \textit{responsible}
\textit{leader} is \textit{expected}
\textit{leader} is \textit{able}
\textit{leader} is \textit{elected}
\textit{leader} is \textit{chosen}
\textit{leader} is \textit{best}
\ldots
3. Entity-Based Context

\[ h_1 = \textit{president} \]
\[ h_2 = \textit{leader} \]

president is \textit{elected}
president is \textit{authorized}
president is \textit{responsible}
president is the chief
president is above
president is the head

\[ \cap \]

leader is \textit{responsible}
leader is \textit{expected}
leader is \textit{able}
leader is elected
leader is chosen
leader is best

...
3. Entity-Based Context

\[ h \{ \text{is} \mid \text{are} \mid \text{was} \mid \text{were} \} \{ \text{a} \mid \text{an} \mid \text{the} \}? y \]

\[ Y_h = \{ y \} \]
3. Entity-Based Context

$h_1 \quad \downarrow \quad Y_{h_1} \quad \Rightarrow \quad Y_{h_1} \cap Y_{h_2} \neq \emptyset \quad \downarrow \quad Y_{h_2} \quad h_2
4. Distributional Soft Clustering

- Distributional hypothesis of Harris (1954)

Words that occur in similar contexts tend to have similar linguistic behavior.

- Applied to Web-scale clusters (Lin et al., 2010)

- Soft clustering assigns up to 20 clusters / word
4. Distributional Soft Clustering

president

- C734
- C284
- C450
- C976
- C447
- ...

leader

- C926
- C985
- C734
- C974
- C450
- ...

Ranked

[Lin et al., 2010]
4. Distributional Soft Clustering

\[ \text{bin}(i + j) \]
5. Pronoun Context

\( h_2 + \text{context} = \text{his election} \)

\( h_1 = \text{Obama} \)

\[
\frac{\text{count}(\text{"Obama 's election"})}{\text{count}(\text{" 's election"}) \cdot \text{count}(\text{"Obama"})}
\]
5. Pronoun Context

\[
\frac{\text{count}(\text{"h}_1 \ 's \ \star \ r")}{\text{count}(\text{" \ \star \ 's \ \star \ r") } \text{count}(\text{"h}_1")}
\]
Web Data
Datasets

Google \( n \)-grams corpus (Brants & Franz, 2006)

- president is the law \( \rightarrow \) 60
- president is the leader \( \rightarrow \) 245
- president is the least \( \rightarrow \) 58
- president is the legal \( \rightarrow \) 50
- president is the main \( \rightarrow \) 79

Word-clusters from Lin et al. (2010)

- presidency” \( \rightarrow \) C229 C127 C114 C129 C611 ...
- president \( \rightarrow \) C734 C284 C450 C976 C447 ...
- president! \( \rightarrow \) C548 C368 C645 C842 C583 ...
- president & CEO \( \rightarrow \) C560 C293 C358 C944 C284 ...
- president’s ability \( \rightarrow \) C876 C754 C770 C212 C215 ...
Coreference Experiments
Datasets and Metrics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#docs</th>
<th>#mentions</th>
<th>#chains</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE04</td>
<td>128</td>
<td>3037</td>
<td>1332</td>
</tr>
<tr>
<td>ACE05</td>
<td>81</td>
<td>1991</td>
<td>775</td>
</tr>
<tr>
<td>ACE05-ALL</td>
<td>599</td>
<td>9217</td>
<td>3050</td>
</tr>
</tbody>
</table>

- 2 popular and complementary metrics
  - MUC
  - B³

[Vilain et al., 1995; Bagga & Baldwin, 1998]
ACE04-DEV Incremental Results

MUC F1

Baseline: 69.5
+Coocc: 69.8
+Hearst: 70.0
+Entity: 70.4
+Cluster: 70.7
+Pronoun: 71.3

B3 F1

Baseline: 77.9
+Coocc: 78.1
+Hearst: 78.5
+Entity: 78.6
+Cluster: 78.6
+Pronoun: 79.0
Analysis
Error Correction Analysis

Coreferent pairs corrected by Web features:

Barry Bonds
athletic director
Democrat Al Gore
Iran
the EPA
Vojislav Kostunica

the best baseball player
Mulcahy
the vice president
the country
the agency
the pro-democracy leader
Decision Tree Analysis

- ~30% of the decision nodes are Web features
- Avg. classification error at Web leaves < 3%
- Strongly discriminative nodes:
  - Hearst feature for its zero-count value
  - Cluster feature for its no-match value
Conclusion

- Simple Web features help significantly
- World knowledge via co-occurrence, context, hypernymy, and compatibility
- State-of-the-art results
Thank you!

Berkeley

Questions?