Easy Victories and Uphill Battles in Coreference Resolution

Greg Durrett and Dan Klein
UC Berkeley
Aspects of Coreference
Aspects of Coreference

Binding theory (Chomsky, 1981)
Aspects of Coreference

Binding theory (Chomsky, 1981)

Pronoun agreement (Hobbs, 1977)
Aspects of Coreference

Binding theory (Chomsky, 1981)

Pronoun agreement (Hobbs, 1977)

Centering (Grosz et al., 1995)
Aspects of Coreference

Binding theory (Chomsky, 1981)

Pronoun agreement (Hobbs, 1977)

Centering (Grosz et al., 1995)

Semantic compatibility

The president₁ ... The leader₁
Aspects of Coreference

Binding theory (Chomsky, 1981)

Pronoun agreement (Hobbs, 1977)

Centering (Grosz et al., 1995)

Semantic compatibility

Definiteness

\[
S \\
\text{John}_1 \\
\text{asked} \\
\text{him}_2 \\
\ldots
\]

\[
\text{MALE} \\
\text{FEMALE} \\
\text{FEMALE}
\]

\[
\text{John}_1 \text{ talked to } \text{Jane}_2. \text{ She}_2 \text{ asked...}
\]

\[
\text{SUBJ.} \\
\text{OBJ.} \\
\text{SUBJ.}
\]

\[
\text{John}_1 \text{ talked to } \text{Bill}_2. \text{ He}_1 \text{ asked...}
\]

\[
\text{The president}_1 \ldots \text{ The leader}_1
\]

\[
\text{The president} \\
\text{A president}
\]
Definiteness: Classical

Soon et al. (2001), Ng et al. (2002), etc.
Definiteness: Classical

The president

Soon et al. (2001), Ng et al. (2002), etc.
The president

```
if (startsWith("the"))
    DEFINITE
else if (startsWith("a"))
    INDEFINITE
else
    NO_ART
```

Soon et al. (2001), Ng et al. (2002), etc.
Definiteness: Classical

The president

```java
if (startsWith("the"))
    DEFINITE
else if (startsWith("a"))
    INDEFINITE
else
    NO_ART
```

DEFINITE
INDEFINITE
NO_ART

Soon et al. (2001), Ng et al. (2002), etc.
Definiteness: Data-Driven

The president
Definiteness: Data-Driven

The president → The
Definiteness: **Data-Driven**

*The president*

- The
- that
- this
- a
- an
- U.S.
- all
- no
- some
- more
- his
- John
- their
- its
- Barack
- Israeli
- new

...
Approaches

Classical approach
Approaches

Classical approach

- Learning with heuristic features
  
  Soon et al. (2001), inter alia
Approaches

Classical approach

- Learning with heuristic features
  
  - Soon et al. (2001), inter alia
  
  - Haghighi and Klein (2009),
  
  - Lee et al. (2011)

- Rule-based
Approaches

Classical approach

- Learning with heuristic features
- Rule-based

Mixed approach

- Add data-driven features on a few axes

Soon et al. (2001), inter alia
Haghighi and Klein (2009),
Lee et al. (2011)

Bengtson and Roth (2008),
Rahman and Ng (2011),
Björkelund and Nugues (2011)
Approaches

Classical approach

- Learning with heuristic features
  - Soon et al. (2001), inter alia
  - Haghighi and Klein (2009), Lee et al. (2011)
- Rule-based

Mixed approach

- Add data-driven features on a few axes
  - Bengtson and Roth (2008), Rahman and Ng (2011), Björkelund and Nugues (2011)

Data-driven approach (this work)

- >400,000 comprehensive, uniform features
Mention-Ranking Architecture

Denis and Baldridge (2008), Durrett et al. (2013)
[Voters]$_1$ agree when [they]$_1$ are given [a chance]$_2$ to decide if [they]$_1$ ...

Denis and Baldridge (2008), Durrett et al. (2013)
Mention-Ranking Architecture

\[ A_1 \]

New

[Voters]_1 agree when [they]_1 are given [a chance]_2 to decide if [they]_1 ...

Denis and Baldridge (2008), Durrett et al. (2013)
[Voters]$_1$ agree when [they]$_1$ are given [a chance]$_2$ to decide if [they]$_1$ ...

Denis and Baldridge (2008), Durrett et al. (2013)
[Voters]$_1$ agree when [they]$_1$ are given [a chance]$_2$ to decide if [they]$_1$ ...

Denis and Baldridge (2008), Durrett et al. (2013)
[Voters]_1 agree when [they]_1 are given [a chance]_2 to decide if [they]_1 ...

Denis and Baldridge (2008), Durrett et al. (2013)
Mention-Ranking Architecture

\[ Pr(A_i = a \mid x) \propto \exp(w^\top f(i, a, x)) \]

[Voters]_1 agree when [they]_1 are given [a chance]_2 to decide if [they]_1 ... 

Denis and Baldridge (2008), Durrett et al. (2013)
Mention-Ranking Architecture

\[ Pr(A_i = a | x) \propto \exp(w^\top f(i, a, x)) \]

[Voters]-[they]

[NOM-PRONOUN] ... 

\[ A_1 \]
\[ \text{New} \]

\[ 1 \]

[Voters] \_1 agree when [they] \_1 are given [a chance] \_2 to decide if [they] \_1 ...

\[ A_2 \]
\[ \text{New} \]

\[ 1 \]

[Dennis and Baldridge (2008), Durrett et al. (2013)]
Mention-Ranking Architecture

\[ Pr(A_i = a | x) \propto \exp(w^\top f(i, a, x)) \]

[Voters] agree when [they] are given [a chance] to decide if [they] ... 

Denis and Baldridge (2008), Durrett et al. (2013)
BASIC Features
BASIC Features

New

[Voters]_1 agree when [they]_1 ...
BASIC Features

[ Voters]₁ agree when [they]₁ ...
BASIC Features

Types
Sentence distance (+ types)
Mention distance (+ types)
Head match

[Voters]₁ agree when [they]₁...
BASIC Features

(Dev set CoNLL-F1, predicted mentions)
Definiteness

Soon et al. (2001), Ng et al. (2002), etc.
New

The president

Soon et al. (2001), Ng et al. (2002), etc.
Definiteness

[DEFINITE]

New

The president

Soon et al. (2001), Ng et al. (2002), etc.
Definiteness

[1stWord=The]

New

The president
Definiteness

Dev set CoNLL-F1, predicted mentions

Basic: 44.6
Definiteness

(Dev set CoNLL-F1, predicted mentions)
Definiteness

Dev set CoNLL-F1, predicted mentions

BASIC
FIRST WORD

49.7
44.6
Definiteness
Definiteness

Definiteness

First Word: “the, a, an” 47.0

Definiteness 47.0
Definiteness

**Definiteness**

- **FIRST WORD:** “the, a, an”  
  - 47.0

+ “some, all, no” + 9 more
  - 47.7
Definiteness

**Definiteness**

**First Word:** “the, a, an”

- **Definiteness** 47.0
- “some, all, no” + 9 more 47.7
- “this, that, these, those” 48.1
Definiteness

Definiteness: 47.0

First Word: “the, a, an” 47.0

+ “some, all, no” + 9 more 47.7

+ “this, that, these, those” 48.1

+ “U.S., new” + 9 more 48.5
Definiteness

Definiteness

**FIRST WORD: “the, a, an”**

+ “some, all, no” + 9 more

+ “this, that, these, those”

+ “U.S., new” + 9 more

+ rest

---

<table>
<thead>
<tr>
<th>Definiteness</th>
<th>47.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>“the, a, an”</td>
<td>47.0</td>
</tr>
<tr>
<td>“some, all, no”</td>
<td>47.7</td>
</tr>
<tr>
<td>“this, that, these, those”</td>
<td>48.1</td>
</tr>
<tr>
<td>“U.S., new”</td>
<td>48.5</td>
</tr>
<tr>
<td>rest</td>
<td>49.7</td>
</tr>
</tbody>
</table>
Centering

Grosz al. (1995)
[Barack Obama]$_1$ met with [David Cameron]$_2$. [He]$_1$ said ...
[Barack Obama]₁ met with [David Cameron]₂. [He]₁ said ...

Grosz al. (1995)
[Barack Obama]$_1$ met with [David Cameron]$_2$. [He]$_1$ said ...

Grosz al. (1995)
[Barack Obama]_1 met with [David Cameron]_2. [He]_1 said ...

Grosz al. (1995)
[Barack Obama]_1 met with [David Cameron]_2. [He]_1 said ...
[Barack Obama]₁ met with [David Cameron]₂. [He]₁ said ...

Haghighi and Klein (2010)
Centering

[Barack Obama]₁ met with [David Cameron]₂. [He]₁ said ...

Subject  Object  Subject

Haghighi and Klein (2010)
[Barack Obama]₁ met with [David Cameron]₂. [He]₁ said ...
[Barack Obama]$_1$ met with [David Cameron]$_2$. [He]$_1$ said ... with [X].
[Barack Obama]$_1$ met with [David Cameron]$_2$. [He]$_1$ said ...

\textit{with [X]. . . [X] said}
Centering

\[ \text{[with X – . Y]} \]
\[ \text{[with X – Y said]} \]

... [Barack Obama]_1 met with [David Cameron]_2. [He]_1 said ...

\[ \text{with [X].} \ \ \ \ \ \ [X] \text{ said} \]
Centering

(Dev set CoNLL-F1, predicted mentions)
Centering

![Bar graph showing comparison of So Far and Position with CoNLL-F1 scores]

- **So Far**: 50.9
- **Position**: 51.9

*(Dev set CoNLL-F1, predicted mentions)*
Centering

(Dev set CoNLL-F1, predicted mentions)
Pronoun Agreement

Bergsma and Lin (2006)
Pronoun Agreement

[Barack Obama]₁ ... [He]₁ ...

Bergsma and Lin (2006)
Pronoun Agreement

[Bergsma and Lin (2006)]

[MALÉ–MÁLE] 

[Barack Obama]₁ ... [He]₁ ...
Pronoun Agreement

[MALE–MALE]

[Barack Obama]₁ ... [He]₁ ...

Obama  MALE=3603  FEMALE=14

MALE

Bergsma and Lin (2006)
Pronoun Agreement

\[ \text{[MALE-MALE]} \]

\[ (\text{Barack Obama})_1 \ldots (\text{He})_1 \ldots \]

\[ \text{Obama MALE}=3603 \]
\[ \text{FEMALE}=14 \]

Bergsma and Lin (2006)
Pronoun Agreement

\[ [\text{Barack Obama}]_1 \ldots [\text{He}]_1 \ldots \]
Pronoun Agreement

[Obama–He]

[Barack Obama]₁ ... [He]₁ ...
Pronoun Agreement

(Dev set CoNLL-F1, predicted mentions)
Pronoun Agreement

(Dev set CoNLL-F1, predicted mentions)
Pronoun Agreement

SO FAR

INDICATORS

58.2

53.4

(Dev set CoNLL-F1, predicted mentions)
SURFACE Information

Features conjoin surface-level mention attributes
SURFACE Information

Features conjoin surface-level mention attributes

<s>[President Barack Obama] signed the bill ... Afterwards [he] said ...
SURFACE Information

- Features conjoin surface-level mention attributes

<s>[President Barack Obama] signed the bill ... Afterwards [he] said ...
SURFACE Information

Features conjoin surface-level mention attributes

<s>[
President Barack Obama
] signed the bill ... Afterwards [ he ] said ...
SURFACE Information

Features conjoin surface-level mention attributes

<s>[President Barack Obama] signed the bill ... Afterwards [he] said ...

LENGTH = 3

LENGTH = 1
Features conjoin surface-level mention attributes

<s>[
  President Barack Obama
]<signed>

LENGTH = 3
PROPER

Afterwards[
  he
]<said>

LENGTH = 1
PRONOUN
SURFACE Information

Features conjoin surface-level mention attributes

<s> [President Barack Obama] signed the bill ... Afterwards [he] said ... 

LENGTH = 3

PROPER

LENGTH = 1

PRONOUN
SURFACE Information

Features conjoin surface-level mention attributes

Mention distance
Sentence distance

<s>[President Barack Obama] signed the bill ... Afterwards [he] said ...

LENGTH = 3
PROPER

LENGTH = 1
PRONOUN
SURFACE Information

Features conjoin surface-level mention attributes

Mention distance
Sentence distance
Head match
Exact match

<s>[President Barack Obama] signed the bill ... Afterwards [he] said ...

LENGTH = 3
PROPER

LENGTH = 1
PRONOUN
**SURFACE Accuracy**

(Dev set CoNLL-F1, predicted mentions)
SURFACE Accuracy

![Bar chart showing comparison between SO FAR and SURFACE.]

58.2

60.1

(Dev set CoNLL-F1, predicted mentions)
Error Analysis
Error Analysis

Anaphoric pronouns

Obama ← he
Anaphoric pronouns

Obama ← he

72.0%
Error Analysis

Anaphoric pronouns

*Obama* ← *he*

Referring: head match

*the U.S. president* ← *president*

72.0%
Error Analysis

Anaphoric pronouns

Obama ← he

72.0%

Referring: head match

the U.S. president ← president

82.7%
Error Analysis

Anaphoric pronouns

Obama ← he

72.0%

Referring: head match

the U.S. president ← president

82.7%

Referring: no head match

David Cameron ← prime minister
Error Analysis

Anaphoric pronouns

- Obama ← he

Referring: head match

- the U.S. president ← president

Referring: no head match

- David Cameron ← prime minister

- Obama: 72.0%
- Referring: head match: 82.7%
- Referring: no head match: 6.2%
Compatibility
Compatibility

[David Cameron]₁ ... [The prime minister]₁ ...
Compatibility

[David Cameron]_1 \ldots [The prime minister]_1 \ldots

\{ \text{Soon et al. (2001)} \}

\textit{inter alia}
Compatibility

[David Cameron]₁ ... [The prime minister]₁ ... 

Number, gender}

Soon et al. (2001)
inter alia
Compatibility

[David Cameron]$_1$ ... [The prime minister]$_1$ ...

- Number, gender
- Named entity type

\{ Soon et al. (2001) inter alia \}
Compatibility

[David Cameron]₁ ... [The prime minister]₁ ...

- Number, gender
- Named entity type
- Unsupervised clustering labels

Soon et al. (2001)
inter alia
Compatibility

[David Cameron]_1 ... [The prime minister]_1 ...

- Number, gender
- Named entity type
- Unsupervised clustering labels
- WordNet hypernymy / synonymy

{ Soon et al. (2001)  
inter alia }
SURFACE Accuracy

(Dev set CoNLL-F1, predicted mentions)
SURFACE Accuracy

\[ \begin{array}{c|c|c}
\text{SURFACE} & 60.1 & 60.4 \\
\text{HEURISTICS} & 60.1 & 60.4 \\
\end{array} \]

(Dev set CoNLL-F1, predicted mentions)
What works?
What works?

- Importing external information with sophisticated heuristics
  
  Ponzetto and Strube (2006)
  Rahman and Ng (2011)
  Bansal and Klein (2012)
What works?

- Importing external information with sophisticated heuristics
  - Ponzetto and Strube (2006)
  - Rahman and Ng (2011)
  - Bansal and Klein (2012)

- We can support additional heuristic features, including number and gender information (Bergsma and Lin, 2006)
Test Set Accuracy

STANFORD (Lee et al., 2011)

56.7

(Test set CoNLL-F1, predicted mentions)
Test Set Accuracy

- **STANFORD** (Lee et al., 2011) - 56.7
- **IMS** (Björkelund and Farkas, 2012) - 58.3

(Test set CoNLL-F1, predicted mentions)
Test Set Accuracy

- **STANFORD** (Lee et al., 2011) - 56.7%
- **IMS** (Björkelund and Farkas, 2012) - 58.3%
- **SURFACE** (This work) - 59.0%

(Test set CoNLL-F1, predicted mentions)
Test Set Accuracy

- **STANFORD** (Lee et al., 2011)
- **IMS** (Björkelund and Farkas, 2012)
- **SURFACE** (This work)
- **BEST** (This work)

*(Test set CoNLL-F1, predicted mentions)*
Conclusion
Conclusion

Surface lexical features capture a wide range of linguistic phenomena in a unified way.
Conclusion

- Surface lexical features capture a wide range of linguistic phenomena in a unified way

- Semantic errors require heavy-duty information from other knowledge sources
Conclusion

- Surface lexical features capture a wide range of linguistic phenomena in a unified way.

- Semantic errors require heavy-duty information from other knowledge sources.

- Extensible system that achieves state-of-the-art performance.
Conclusion

- The Berkeley Coreference Resolution System: http://nlp.cs.berkeley.edu
- Full end-to-end system (accepts raw text as input)
Conclusion

- The Berkeley Coreference Resolution System: http://nlp.cs.berkeley.edu
- Full end-to-end system (accepts raw text as input)

Thank you!
Stacking

<table>
<thead>
<tr>
<th>Surface</th>
<th>+Definiteness</th>
<th>+Position</th>
<th>+Gender/Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>60.1</td>
<td>60.1</td>
<td>60.2</td>
<td>60.4</td>
</tr>
</tbody>
</table>
SURFACE Accuracy

SURFACE
HEURISTICS
INDICATORS

60.1
60.4
60.2

Predicted mentions
SURFACE Accuracy

SURFACE

HEURISTICS

INDICATORS

Predicted mentions

Gold mentions

60.1
60.4
60.2
75.1
76.7
76.0
SURFACE Accuracy

Predicted mentions

- SURFACE: 60.1
- Heuristics: 60.4
- Indicators: 60.2

Δ = +0.3

Gold mentions

- SURFACE: 75.1
- Heuristics: 76.7
- Indicators: 76.0

Δ = +1.6
SURFACE Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Predicted mentions</th>
<th>Gold mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>SURFACE</td>
<td>60.1</td>
<td>75.1</td>
</tr>
<tr>
<td>HEURISTICS</td>
<td>60.4</td>
<td>76.7</td>
</tr>
<tr>
<td>INDICATORS</td>
<td>60.2</td>
<td>76.0</td>
</tr>
</tbody>
</table>

Δ = +1.6
Δ = +0.9
New

[the president]
Gold vs. System Mentions

New
[the president]
PERSON
Gold vs. System Mentions

[Barack Obama]
[the GOP]
[the Capitol]

1

New

[the president]

PERSON

2

3
Gold vs. System Mentions

[Barack Obama] \textbf{PERSON}

[the GOP] \textbf{ORG}

[the Capitol] \textbf{LOC}

[the president] \textbf{PERSON}

\textcolor{red}{New}
Gold vs. System Mentions

[Barack Obama] \[PERSON\]
[the GOP] \[ORG\]
[the Capitol] \[LOC\]

1
2
3

New

[the president] \[PERSON\]

3x more mentions when singletons are included
Gold vs. System Mentions

[Barack Obama]  [the GOP]  [the Capitol]  [the advisor]  [the Chief of Staff]  [somebody]  [the party]

PERSON   ORG   LOC   PERSON   PERSON   PERSON   ORG

1  2  3  4  5  6  7

New

[the president]  PERSON
Gold vs. System Mentions

[Barack Obama] [the GOP] [the Capitol] [the advisor] [the Chief of Staff] [somebody] [the party]

PERSON ORG LOC PERSON PERSON PERSON PERSON

1 2 3 4 5 6 7

UPHILL BATTLE

New

[the president]