Decentralized Entity-Level Modeling for Coreference Resolution

Greg Durrett, David Hall, and Dan Klein
UC Berkeley
Entity-Level Modeling
Entity-Level Modeling

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party. [his]
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.

[Soon et al. (2001) inter alia]
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.

[Soon et al. (2001) inter alia]
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.

[Soon et al. (2001) inter alia]
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.

[Soon et al. (2001) inter alia] Does not propagate information
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
Centralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.

[Luo et al. (2004), Rahman and Ng (2009)]
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.

Gender: **MALE**

[James Reed]
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.

[Gender: MALE [James Reed]]

[Gender: FEMALE [Rose Brooks]]

[Gender: UNKNOWN [Reed]]

[Gender: UNKNOWN [Brooks]]

[Gender: MALE [his]]

[Luo et al. (2004), Rahman and Ng (2009)]
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.

Gender: **MALE**
- James Reed
- Reed

Gender: **FEMALE**
- Rose Brooks

[Gender: **UNKNOWN**
- Reed
- Brooks

his

[Luo et al. (2004), Rahman and Ng (2009)]]
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.

[Luo et al. (2004), Rahman and Ng (2009)]
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.

[Luo et al. (2004), Rahman and Ng (2009)]
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.

Does not maintain uncertainty during inference

[Gender: MALE, Gender: FEMALE, Unknown, Unknown, MALE]

James Reed  Rose Brooks  Reed  Brooks  his

[Luo et al. (2004), Rahman and Ng (2009)]
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
Our Decentralized Approach

New York was where James Reed met Rose Brooks. Reed was introduced to Brooks at his company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company’s Christmas party.

Maintains tractability of pairwise system, incorporates entity-level information

James Reed
Rose Brooks
Reed
Brooks
his
... [James Reed]\textsubscript{1} met [Rose Brooks]\textsubscript{2}. [Reed]\textsubscript{3} was ...

[Denis and Baldridge (2008)]
BASIC Model

... [James Reed]$_1$ met [Rose Brooks]$_2$.  [Reed]$_3$ was ...  

[Denis and Baldridge (2008)]
... [James Reed]₁ met [Rose Brooks]₂. [Reed]₃ was ...

[Denis and Baldridge (2008)]
... [James Reed]_1 met [Rose Brooks]_2. [Reed]_3 was ...

[Denis and Baldridge (2008)]
... [James Reed]_1 met [Rose Brooks]_2. [Reed]_3 was ...

[Denis and Baldridge (2008)]
BASIC Model

\[ Pr(a_i | x) \propto \exp(w^T f(a_i, x)) \]

... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

[Denis and Baldridge (2008)]
\[ Pr(a_i | x) \propto \exp(w^T f(a_i, x)) \]

... [James Reed]_1 met [Rose Brooks]_2. [Reed]_3 was ...

[Denis and Baldridge (2008)]
BASIC Model

\[ Pr(a_i | x) \propto \exp(w^T f(a_i, x)) \]

- \( \text{New} \land \text{proper} \)
- \( \text{New} \land \text{two words} \)
- \( \text{Head match} \)
- \( \text{Both proper} \)

... [James Reed]_1 met [Rose Brooks]_2. [Reed]_3 was ...

[Denis and Baldridge (2008)]
BASIC Model

\[ Pr(\mathcal{a}_i | x) \propto \exp(w^T f(\mathcal{a}_i, x)) \]

... [James Reed]$_1$ met [Rose Brooks]$_2$.  

[Denis and Baldridge (2008)]
$Pr(a_i \mid x) \propto \exp(w^T f(a_i, x))$
DECENTRALIZED Model

... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...
DECENTRALIZED Model

... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...
DECENTRALIZED Model

... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...
... [James Reed]$_1$ met [Rose Brooks]$_2$.  [Reed]$_3$ was ...
DECENTRALIZED Model

... [James Reed]₁ met [Rose Brooks]₂.  [Reed]₃ was ...

A₁  A₂  A₃
DECENTRALIZED Model

... [James Reed]₁ met [Rose Brooks]₂. [Reed]₃ was ...
... [James Reed]₁ met [Rose Brooks]₂. [Reed]₃ was ...
DECENTRALIZED Model

... [James Reed]_1 met [Rose Brooks]_2.  [Reed]_3 was ...
... [James Reed]_1 met [Rose Brooks]_2. [Reed]_3 was ...
DECENTRALIZED Model

... [James Reed]$_1$ met [Rose Brooks]$_2$.  [Reed]$_3$ was ...

\[A_1\] \[A_2\] \[A_3\] \[P_1\] \[P_2\] \[P_3\]
... [James Reed]_1 met [Rose Brooks]_2. [Reed]_3 was ...
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

New
DECENTRALIZED Model

... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

New
... [James Reed]_1 met [Rose Brooks]_2. [Reed]_3 was ...
A DECENTRALIZED Model

... [James Reed]₁ met [Rose Brooks]₂. [Reed]₃ was ...

\[ A₁ = M \quad F \quad P₁ \]
\[ A₂ = P₂ \]
\[ A₃ = M \quad F \quad P₃ \]

New

1 2
... [James Reed]₁ met [Rose Brooks]₂. [Reed]₃ was ...

New
... [James Reed]₁ met [Rose Brooks]₂. [Reed]₃ was ...

DECENTRALIZED Model
... [James Reed]₁ met [Rose Brooks]₂. [Reed]₃ was ...

**DECENTRALIZED Model**

\[ A_1 \] \[ P_1 \] \[ M \] \[ F \] \[ P_2 \] \[ P_3 \] \[ A_2 \] \[ A_3 \]

1 2 New
DECENTRALIZED Model

... [James Reed]_1 met [Rose Brooks]_2. [Reed]_3 was ...
... [James Reed]₁ met [Rose Brooks]₂. [Reed]₃ was ...
DECENTRALIZED Model

\[ P_1 \rightarrow P_2 \rightarrow P_3 = A_1 \rightarrow A_2 \rightarrow A_3 \]
DECENTRALIZED Model
DECENTRALIZED Model

\[ P_1 \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad P_2 \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad P_3 \]

\[ A_1 \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad A_2 \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad A_3 \]
DECENTRALIZED Model

\[ P_1 = P_2 = P_3 \]

\{ \]

Pairwise model
DECENTRALIZED Model

\[ P_1 \rightarrow P_2 \rightarrow P_3 \]

\[ A_1 \rightarrow A_2 \rightarrow A_3 \]

Property model

Pairwise model
Decentralized Model

- $P_1$
- $P_2$
- $P_3$
- $A_1$
- $A_2$
- $A_3$

Property model
Equality factors
Pairwise model
Inference
Inference

Need to compute expected feature counts:

\[ \mathbb{E}_{\text{gold}} f(P_1 P_2 P_3 A_1 A_2 A_3) - \mathbb{E}_{\text{all}} f(P_1 P_2 P_3 A_1 A_2 A_3) \]
Inference

Need to compute expected feature counts:

\[ \mathbb{E}_{\text{gold}} f(A_1, A_2, A_3) - \mathbb{E}_{\text{all}} f(A_1, A_2, A_3) \]

Use belief propagation to compute marginals over variables.
Inference

- Need to compute expected feature counts:

\[ \mathbb{E}_{\text{gold}} f(A) - \mathbb{E}_{\text{all}} f(A) \]

- Use belief propagation to compute marginals over variables

- Decoding: max over each \( A_i \) marginal
Learning
Learning

- Optimize conditional log likelihood of training data
Optimize conditional log likelihood of training data

\[ \sum_{i} \log \left( Pr(a^i_g | x^i) \right) \]
Optimize conditional log likelihood of training data

\[ \sum_i \log (P_r(a^i_g | x^i)) \]
Learning

Optimize conditional log likelihood of training data

\[ \sum_i \log \left( Pr \left( a_g^i | x^i \right) \right) \]

Gold antecedent vector

Training examples
Optimize conditional log likelihood of training data

\[ \sum_i \log \left( Pr(a_g^i | x^i) \right) \]

Gold antecedent vector

Training examples

Observed document properties
Optimize conditional log likelihood of training data

$$\sum_i \log \left( \sum_{a_g^i \in A(C)} Pr(a_g^i | x^i) \right)$$
Learning

Optimize conditional log likelihood of training data

\[
\sum_i \log \left( \sum_{a_g^i \in A(C)} Pr(a_g^i | x^i) \right)
\]

Antecedent choices consistent with gold standard
Learning
Learning

Want to optimize for MUC, $B^3$, CEAF, etc.
Learning

- Want to optimize for MUC, B³, CEAF, etc.
- Use a decomposable metric as a proxy
Want to optimize for MUC, B³, CEAF, etc.

Use a decomposable metric as a proxy

... [James Reed]₁ met [Rose Brooks]₂. [Reed]₃ was ...
Learning

- Want to optimize for MUC, $B^3$, CEAF, etc.
- Use a decomposable metric as a proxy

... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

False Anaphor
Learning

- Want to optimize for MUC, B³, CEAF, etc.
- Use a decomposable metric as a proxy

... [James Reed]₁ met [Rose Brooks]₂. [Reed]₃ was ...
Learning

- Want to optimize for MUC, B³, CEAF, etc.
- Use a decomposable metric as a proxy

... [James Reed]₁ met [Rose Brooks]₂. [Reed]₃ was ...

False Anaphor  Wrong Link  False New

New

New

New

New
Want to optimize for MUC, B³, CEAF, etc.

Use a decomposable metric as a proxy

\[ k_1(\text{False Anaphors}) + k_2(\text{False News}) + k_3(\text{Wrong Links}) \]

... [James Reed]₁ met [Rose Brooks]₂. [Reed]₃ was...
\[
\sum_{i} \log \left( \sum_{a^i_g \in A(C)} Pr(a^i_g | x^i) \right)
\]
Incorporate this loss with *softmax-margin* by adding it as a feature to the pairwise model.

\[
\sum_i \log \left( \sum_{a^i_g \in A(C')} Pr(a^i_g | x^i) \right)
\]

[Gimpel and Smith (2010)]
Incorporate this loss with softmax-margin by adding it as a feature to the pairwise model.

\[
\sum_i \log \left( \sum_{a^i_g \in A(C')} \frac{Pr'}{Pr(a^i_g|x^i)} \right)
\]

[Gimpel and Smith (2010)]
Incorporate this loss with *softmax-margin* by adding it as a feature to the pairwise model:

\[
\sum_i \log \left( \sum_{a_g^i \in A(C')} \frac{Pr'}{Pr(a_g^i | x_i)} \right) + \lambda \| w \|_1
\]

[Gimpel and Smith (2010)]
Experiments
Experiments

- CoNLL 2011 dataset, system mentions from Lee et al. (2011)
Experiments

- CoNLL 2011 dataset, system mentions from Lee et al. (2011)

- Baselines:
  - Pairwise system
  - Centralized entity-level system following Rahman and Ng (2009)
Experiments

- CoNLL 2011 dataset, system mentions from Lee et al. (2011)

- Baselines:
  - Pairwise system
  - Centralized entity-level system following Rahman and Ng (2009)

- Two settings:
  - Synthetic features to contrast architectures
  - Standard entity features
Synthetic Properties
For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$
Synthetic Properties

- For each gold cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

[James Reed]
[Reed]
[his]
For each gold cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$
For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

- **James Reed** $\rightarrow$ 5
- **Reed** $\rightarrow$ 5
- **his** $\rightarrow$ 5
- **Rose Brooks** $\rightarrow$ 3
- **Brooks** $\rightarrow$ 3
For each gold cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

- [James Reed] [Reed] [his] $\rightarrow 5$
- [Rose Brooks] [Brooks] $\rightarrow 3$
- [New York] $\rightarrow 5$
For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

- [James Reed] [Reed] [his] $\rightarrow 5$
- [Rose Brooks] [Brooks] $\rightarrow 3$
- [New York] $\rightarrow 5$

For each mention, sample from Dirichlet peaked on $\ell$
Synthetic Properties

- For each gold cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

- For each mention, sample from Dirichlet peaked on $\ell$

- $[\text{James Reed}]$ $\rightarrow$ 5
- $[\text{Reed}]$ $\rightarrow$ 5
- $[\text{his}]$ $\rightarrow$ 5
- $[\text{Rose Brooks}]$ $\rightarrow$ 3
- $[\text{Brooks}]$ $\rightarrow$ 3
- $[\text{New York}]$ $\rightarrow$ 5
- $[\text{James Reed}]$ $\rightarrow$ 5
- $[\text{Reed}]$ $\rightarrow$ 5
- $[\text{his}]$ $\rightarrow$ 5
- $[\text{Rose Brooks}]$ $\rightarrow$ 3
- $[\text{Brooks}]$ $\rightarrow$ 3
- $[\text{New York}]$ $\rightarrow$ 5
For each gold cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

- For each mention, sample from Dirichlet peaked on $\ell$

- [James Reed] $\rightarrow$ 5
- [Rose Brooks] $\rightarrow$ 3
- [New York] $\rightarrow$ 5

- James Reed $\rightarrow$ 1, 2, 3, 4, 5
- Reed $\rightarrow$ 1, 2, 3, 4, 5
For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

- [James Reed] $\rightarrow 5$
- [Reed] $\rightarrow 5$
- [his] $\rightarrow 5$
- [Rose Brooks] $\rightarrow 3$
- [Brooks] $\rightarrow 3$
- [New York] $\rightarrow 5$

For each mention, sample from Dirichlet peaked on $\ell$

- James Reed $\rightarrow$
- Reed $\rightarrow$
- his $\rightarrow$
Synthetic Properties

BASIC  65

60.0

55

(CoNLL scores, 10-fold cross-validation on train set)
Synthetic Properties

* uses gold information

(CoNLL scores, 10-fold cross-validation on train set)
Synthetic Properties

- **BASIC**
- **PAIRWISE***
- **CENTRALIZED***

* uses gold information

(CoNLL scores, 10-fold cross-validation on train set)
Synthetic Properties

- **BASIC**
- **PAIRWISE***
- **CENTRALIZED***
- **DECENTRALIZED***

* uses gold information

(CoNLL scores, 10-fold cross-validation on train set)
φ-feature Properties

[Rahman and Ng (2009), Lee et al. (2011), inter alia]
Properties based on linguistic $\phi$-features:

- Number
- Gender
- Animacy
- NE type

[Rahman and Ng (2009), Lee et al. (2011), inter alia]
(CoNLL scores, 10-fold cross-validation on train set)
φ-feature Properties

(CoNLL scores, 10-fold cross-validation on train set)
(CoNLL scores, 10-fold cross-validation on train set)
φ-features do not capture fine-grained semantic distinctions between entities
Semantic Properties

- φ-features do not capture fine-grained semantic distinctions between entities
- Use properties derived from unsupervised clustering of headwords and their governors
Semantic Properties

BASIC

(CoNLL scores, 10-fold cross-validation on train set)
Semantic Properties

(CoNLL scores, 10-fold cross-validation on train set)
Semantic Properties

(Basic, Pairwise, Centralized)

(CoNLL scores, 10-fold cross-validation on train set)
Semantic Properties

CoNLL scores, 10-fold cross-validation on train set
Overall Results

<table>
<thead>
<tr>
<th></th>
<th>Pairwise</th>
<th>Centralized</th>
<th>Decentralized</th>
</tr>
</thead>
<tbody>
<tr>
<td>φ-features</td>
<td>59.9</td>
<td>59.8</td>
<td>59.9</td>
</tr>
<tr>
<td>Semantic</td>
<td>60.4</td>
<td>59.9</td>
<td>60.4</td>
</tr>
</tbody>
</table>
## Overall Results

<table>
<thead>
<tr>
<th></th>
<th>Pairwise</th>
<th>Centralized</th>
<th>Decentralized</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>φ-features</strong></td>
<td>59.9</td>
<td>59.8</td>
<td>59.9</td>
</tr>
<tr>
<td><strong>Semantic</strong></td>
<td>60.4</td>
<td>59.9</td>
<td>60.4</td>
</tr>
<tr>
<td><strong>Synthetic</strong></td>
<td>62.7</td>
<td>63.0</td>
<td>63.7</td>
</tr>
</tbody>
</table>
Conclusion

- Our model effectively integrates entity-level features in an end-to-end way
Conclusion

- Our model effectively integrates entity-level features in an end-to-end way.

- Good entity-level features are hard to find: simple $\phi$-feature and semantic type propagation give little benefit.
Conclusion

- Our model effectively integrates entity-level features in an end-to-end way

- Good entity-level features are hard to find: simple $\phi$-feature and semantic type propagation give little benefit

Thank you!
Projected Properties

raw input

$P_1$
Projected Properties

raw input

\[ R_1 \]

\[
\begin{array}{c c}
M & F \\
\end{array}
\]
Projected Properties

raw input

R₁

“projection” factor

P₁

M  F
Projected Properties

“How willing is the model to switch from M to F”

raw input

“projection” factor

\[ R_1 \]

\[ P_1 \]

\[
\begin{array}{cc}
\theta_{M-M} & \theta_{M-F} \\
\theta_{F-M} & \theta_{F-F}
\end{array}
\]
Projected Properties

“How willing is the model to switch from M to F”

raw input

“projection” factor

θ_{M-M}  θ_{M-F}
θ_{F-M}  θ_{F-F}

θ_{M-M}  θ_{M-F}
θ_{F-M}  θ_{F-F}

“projection” factor

θ_{M-M}  θ_{M-F}
θ_{F-M}  θ_{F-F}

θ_{M-M}  θ_{M-F}
θ_{F-M}  θ_{F-F}
Final Results

(CoNLL scores, blind test set)