Higher-order Coreference Resolution with Coarse-to-fine Inference

Kenton Lee*  Luheng He  Luke Zettlemoyer

University of Washington

* Now at Google
Coreference Resolution

It’s because of what both of you are doing to have things change.

I think that’s what’s… Go ahead Linda.

Thanks goes to you and to the media to help us.

Absolutely.

Obviously we couldn’t seem loud enough to bring the attention, so our hat is off to all of you as well.

Example from Wiseman et al. (2016)
It’s because of what both of you are doing to have things change.

I think that’s what’s… Go ahead Linda.

Thanks goes to you and to the media to help us.

Absolutely.

Obviously we couldn’t seem loud enough to bring the attention, so our hat is off to all of you as well.
It’s because of what both of you are doing to have things change.

I think that’s what’s… Go ahead Linda.

Thanks goes to you and to the media to help us.

Absolutely.

Obviously we couldn’t seem loud enough to bring the attention, so our hat is off to all of you as well.

Example from Wiseman et al. (2016)
Recent Trends in Coreference Resolution

End-to-end models have achieved large improvements

Advantages
- Conceptually simple
- Minimal feature engineering

Disadvantages
- Computationally expensive
- Very little “reasoning” involved
Contributions

• Address a **modeling** challenge:
  • Enable higher-order (multi-hop) coreference

• Address a **computational** challenge:
  • Coarse-to-fine inference with a factored model
Contributions

• Address a **modeling** challenge:
  • Enable higher-order (multi-hop) coreference

• Address a **computational** challenge:
  • Coarse-to-fine inference with a factored model
Existing Approach: Span-ranking Model

Lee et al. 2017 (EMNLP):

- Consider all possible spans in the document:

\[ 1 < i < n \]

- Compute neural span representations:

\[ h(i) \]

- Estimate probability distribution over possible antecedents:

\[ P(y_i \mid h) \]
I think that’s what’s… Go ahead Linda.

Thanks goes to you and to the media to help us.

Local information not sufficient

Obviously we couldn’t seem loud enough to bring the attention, so our hat is off to all of you as well.

Example from Wiseman et al. (2016)
I think that’s what’s… Go ahead Linda.

Thanks goes to you and to the media to help us.

Global structure reveals inconsistency

Obviously we couldn’t seem loud enough to bring the attention, so our hat is off to all of you as well.

Example from Wiseman et al. (2016)
Higher-order Model

• Let span representations softly condition on previous decisions
Higher-order Model

- Let span representations softly condition on previous decisions
- For each iteration:
  - Estimation antecedent distribution
  - Attend over possible antecedents
  - Merge every span representation with its expected antecedent
I think that’s what’s… Go ahead Linda.

Thanks goes to you and to the media to help us.

Absolutely.

Obviously we couldn’t seem loud enough to bring the attention, so our hat is off to all of you as well.
I think that’s what’s… Go ahead Linda.

Thanks goes to you and to the media to help us.

Absolutely.

Obviously we couldn’t seem loud enough to bring the attention, so our hat is off to all of you as well.
I think that’s what’s… Go ahead Linda.

Absolutely.

Thanks goes to you and to the media to help us.

Obviously we couldn’t seem loud enough to bring this about.

Learn a representation of “you” w.r.t. “I”
I think that’s what’s… Go ahead Linda.

Thanks goes to you and to the media to help us.

Absolutely.

Obviously we couldn’t seem loud enough to bring the attention, so our hat is off to all of you as well.
I think that’s what’s… Go ahead Linda.

Thanks goes to you and to the media to help us.

Absolutely.

Obviously we couldn’t seem loud enough to bring the attention, so our hat is off to all of you as well.
Higher-order Model

- Let span representations softly condition on previous decisions
- Iterative inference to compute $h_n(i)$
Higher-order Model

- Let span representations softly condition on previous decisions
- Iterative inference to compute $h_n(i)$:
  - Base case: $h_0(i) = h(i)$ (from the baseline)
Higher-order Model

• Let span representations softly condition on previous decisions
• Iterative inference to compute $h_n(i)$:
  • Base case: $h_0(i) = h(i)$ (from the baseline)
  • Recursive case:
    $$a_n(i) = \sum_{y_i} P(y_i | h_{n-1}) h_{n-1}(i)$$ (attention mechanism)
Higher-order Model

- Let span representations softly condition on previous decisions
- Iterative inference to compute $h_n(i)$:
  - Base case: $h_0(i) = h(i)$ (from the baseline)
  - Recursive case:
    
    $$a_n(i) = \sum_{y_i} P(y_i| h_{n-1}) h_{n-1}(i) \quad \text{(attention mechanism)}$$
    $$f_n(i) = \sigma(W[a_n(i), h_{n-1}(i)]) \quad \text{(forget gates)}$$
Higher-order Model

• Let span representations softly condition on previous decisions

• Iterative inference to compute $h_n(i)$:
  
  • Base case: $h_0(i) = h(i)$ (from the baseline)
  
  • Recursive case:

    $$a_n(i) = \sum_{y_i} P(y_i|h_{n-1}(i))h_{n-1}(i)$$ (attention mechanism)
    $$f_n(i) = \sigma(W [a_n(i), h_{n-1}(i)])$$ (forget gates)
    $$h_n(i) = f_n(i) \circ a_n(i) + (1 - f_n(i)) \circ h_{n-1}(i)$$
Higher-order Model

- Let span representations softly condition on previous decisions
- Iterative inference to compute $h_n(i)$:
  - Base case: $h_0(i) = h(i)$ (from the baseline)
  - Recursive case:
    
    $$a_n(i) = \sum_{y_i} P(y_i|h_{n-1})h_{n-1}(i)$$
    
    (attention mechanism)

    $$f_n(i) = \sigma(W[a_n(i), h_{n-1}(i)])$$
    
    (forget gates)

    $$h_n(i) = f_n(i) \circ a_n(i) + (1 - f_n(i)) \circ h_{n-1}(i)$$

  - Final result:
    
    $$P(y_i|h_n)$$
Higher-order Model

- Let span representations softly condition on previous decisions
- Iterative inference to compute $h_n(i)$:
  - Base case: $h_0(i) = h(i)$ (from the baseline)
  - Recursive case:
    
    $a_n(i) = \sum_{y_i} P(y_i|h_{n-1})h_{n-1}(i)$
    
    (attention mechanism)
    
    $f_n(i) = \sigma(W[a_n(i), h_{n-1}(i)])$
    
    $h_n(i) = f_n(i)$
  - Final result:
    
    $P(y_i|h_n)$

Final coreference decision conditions on clusters of size $n + 2$
Recent Trends in Coreference Resolution

End-to-end models have achieved large improvements

Advantages

• Conceptually simple
• Minimal feature engineering

Disadvantages

• Computationally expensive
• Very little “reasoning” involved
Recent Trends in Coreference Resolution

End-to-end models have achieved large improvements

**Advantages**
- Conceptually simple
- Minimal feature engineering

**Disadvantages**
- Computationally expensive
- Very little "reasoning" involved

2nd order model already runs out of memory
Contributions

• Address a **modeling** challenge:
  • Enable higher-order (multi-hop) coreference

• Address a **computational** challenge:
  • Coarse-to-fine inference with a factored model
Computational Challenge

- It's because of what both of you are doing to have things change.

- Mention candidates just for exposition
It’s because of what both of you are doing to have things change.

• Mention candidates just for exposition
• $O(n^2)$ spans to consider in practice
Computational Challenge

It's because of what both of you are doing to have things change.

• Mention candidates just for exposition
• $O(n^2)$ spans to consider in practice
• $O(n^4)$ coreference links to consider
Coarse-to-fine Inference

\[ P(y_i|h) = \text{softmax}(s(i, y_i, h)) \]
Coarse-to-fine Inference

\[ P(y_i|h) = \text{softmax}(s(i, y_i, h)) \]

Existing scoring function:

\[ s(i, j, h) = \text{FFNN}(h(i)) + \text{FFNN}(h(j)) + \text{FFNN}(h(i), h(j), h(i) \circ h(j)) \]

Mention scores

Antecedent scores
Coarse-to-fine Inference

\[ P(y_i|h) = \text{softmax}(s(i, y_i, h)) \]

Coarse-to-fine scoring function:

\[ s(i, j, h) = \text{FFNN}(h(i)) + \text{FFNN}(h(j)) + h(i)^\top W_c h(j) + \text{FFNN}(h(i), h(j), h(i) \circ h(j)) \]

Mention scores

Cheap/inaccurate antecedent scores

Antecedent scores
Coarse-to-fine Inference

\[ P(y_i|h) = \text{softmax}(s(i, y_i, h)) \]

Coarse-to-fine scoring function:

\[ s(i, j, h) = \text{FFNN}(h(i)) + \text{FFNN}(h(j)) \]

\[ + h(i)\top W_c h(j) \]

\[ + \text{FFNN}(h(i), h(j), h(i) \circ h(j)) \]

Only compute expensive scores for the top K span pairs

Mention scores

Cheap/inaccurate antecedent scores

Antecedent scores
Experimental Setup

**Dataset**: English OntoNotes (CoNLL-2012)

**Baseline**: Lee et al. 2017 with:

1. Better hyperparameters (deeper LSTMs, longer spans, etc.)
2. ELMo (Peters et al. 2018) embeddings
Coreference Results

Lee et al. (2017) + ELMo + hyperparameter tuning

Test Avg. F1 (%)

72.3
Coreference Results

<table>
<thead>
<tr>
<th>Test Avg. F1 (%)</th>
<th>Lee et al. (2017) + ELMo + hyperparameter tuning</th>
<th>+ coarse-to-fine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>72.3</td>
<td>72.6</td>
</tr>
</tbody>
</table>

Lee et al. (2017) + ELMo + hyperparameter tuning has a test average F1 score of 72.3%, while adding coarse-to-fine improves it to 72.6%.
Coreference Results

![Bar chart showing test average F1 scores.]

- Lee et al. (2017)
  - + ELMo
  - + hyperparameter tuning: 72.3
- + coarse-to-fine: 72.6
- + second-order inference: 73.1
Summary

• Improve structural consistency via multi-hop coreference

• Enable more complex inference via coarse-to-fine beam search