Understanding Mention Detector-Linker Interaction in Neural Coreference Resolution

Zhaofeng Wu and Matt Gardner
Neural Coreference Resolution
Neural Coreference Resolution

He and I laughed at my dog who bit himself.
Neural Coreference Resolution

He and I laughed at my dog who bit himself.

SpanBERT
Neural Coreference Resolution

He and I laughed at my dog who bit himself.
He and I laughed at my dog who bit himself.
Neural Coreference Resolution

He and I laughed at my dog who bit himself.
Neural Coreference Resolution

He and I laughed at my dog who bit himself.
He and I laughed at my dog who bit himself.
Neural Coreference Resolution

He and I laughed at my dog who bit himself.
Neural Coreference Resolution

He and I laughed at my dog who bit himself.

“IT is important to note that the fi rst system stage, i.e., the mention detection fi e, favors recall heavily ...”
— Lee et al. (2011)
Is Detector Recall Always More Important Than Precision?
Oracle Detectors
Oracle Detectors

He who bit I my dog laughed

Original
Oracle Detectors

Original

He
I
my dog
laughed
who bit

Perfect Recall

He
I
my
dog
myself
laughed
who bit

Perfect Recall
Oracle Detectors

Original
- He
- I
- my dog
- laughed
- who bit

Perfect Precision
- He
- I
- my dog

Perfect Recall
- He
- I
- my dog
- laughed
- my
- himself
- who bit
Datasets & Model Settings
Datasets & Model Settings

• Ontonotes (English)
Datasets & Model Settings

- Ontonotes (English)
- PreCo
Datasets & Model Settings

- Ontonotes (English)
- PreCo
- Embedder: SpanBERT-large
Datasets & Model Settings

• Ontonotes (English)
• PreCo
• Embedder: SpanBERT-large
• Experimental configuration
  • [https://github.com/allenai/allennlp-models/blob/main/training_config/coref/coref_spanbert_large.jsonnet](https://github.com/allenai/allennlp-models/blob/main/training_config/coref/coref_spanbert_large.jsonnet)
Importance of Precision
Importance of Precision

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>Recall &amp; Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td>84.6</td>
</tr>
<tr>
<td>Anaphoric Mentions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Mentions</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Importance of Precision

Baseline: 84.6
Anaphoric Mentions: 85.2
All Mentions: 85.2
Importance of Precision

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>Recall &amp; Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>84.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anaphoric Mentions</td>
<td>85.2</td>
<td>95.1</td>
<td></td>
</tr>
<tr>
<td>All Mentions</td>
<td>85.2</td>
<td>88.4</td>
<td></td>
</tr>
</tbody>
</table>

The chart above illustrates the F1 scores for different scenarios. The highest score is 95.1 for Anaphoric Mentions.
Importance of Precision

![Graph showing F1 scores for Baseline, Anaphoric Mentions, and All Mentions.]

- Baseline: Recall 84.6, Precision 85.2, Recall & Precision 85.2
- Anaphoric Mentions: Recall 95.1, Precision 96.5, Recall & Precision 88.4
- All Mentions: Recall 88.4, Precision 88.4, Recall & Precision 89.1
Importance of Recall
Importance of Recall

Per Operation Effect

- Anaphoric: Recall 0.71, Precision 0.13
- All: Recall 0.14, Precision 0.12
Precision-Recall Trade-Off

F1 (Ontonotes)

Max Length  Spans considered per word

- 30 • 0.4
- 32 • 0.45
- 34 • 0.5
- 36 • 0.55
- 38 • 0.6
- 40 • 0.65

Graph showing the relationship between Max Length and Spans considered per word and their impact on F1 (Ontonotes) scores.
Previously …

![Bar chart showing recall, precision, and F1 scores for Baseline, Anaphoric Mentions, and All Mentions.

- Baseline: Recall 84.6, Precision 95.1, F1 89.1
- Anaphoric Mentions: Recall 85.2, Precision 96.5, F1 95.1
- All Mentions: Recall 85.2, Precision 88.4, F1 89.1]
Previously …

![Graph showing F1 scores for Baseline, Anaphoric Mentions, and All Mentions with Recall, Precision, and Recall & Precision categories.](image)
How Well Can the Detector Make Anaphoricity Decisions?
Anaphoricity Decisions
Anaphoricity Decisions

• Detector-like span classifiers that recognize:
Anaphoricity Decisions

• Detector-like span classifiers that recognize:
  • All mentions: 79.9 classification F1
Anaphoricity Decisions

- Detector-like span classifiers that recognize:
  - All mentions: 79.9 classification F1
  - Anaphoric mentions: 54.3 classification F1
Anaphoricity Decisions

• Detector-like span classifiers that recognize:
  • All mentions: 79.9 classification F1
  • Anaphoric mentions: 54.3 classification F1
• Confusion Index (CI): (singleton recall) / (anaphoric mention recall)
Anaphoricity Decisions

- Detector-like span classifiers that recognize:
  - All mentions: 79.9 classification F1
  - Anaphoric mentions: 54.3 classification F1
- Confusion Index (CI): (singleton recall) / (anaphoric mention recall)
  - A perfect detector has 0 CI; a random one has 1 CI
Anaphoricity Decisions

• Detector-like span classifiers that recognize:
  • All mentions: 79.9 classification F1
  • Anaphoric mentions: 54.3 classification F1
  • Confusion Index (CI): (singleton recall) / (anaphoric mention recall)
    • A perfect detector has 0 CI; a random one has 1 CI
    • has 0.81 CI
Anaphoricity Decisions

• Detector-like span classifiers that recognize:
  • All mentions: 79.9 classification F1
  • Anaphoric mentions: 54.3 classification F1
• Confusion Index (CI): (singleton recall) / (anaphoric mention recall)
  • A perfect detector has 0 CI; a random one has 1 CI
  • has 0.81 CI
  • Degrades to 0.997 CI with text appearing as both mention types
What About the Linker?
## Linker Errors

<table>
<thead>
<tr>
<th>Category</th>
<th>Sentence A</th>
<th>Sentence B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronoun (109)</td>
<td>Tom lives in <strong>Seattle</strong>.</td>
<td><strong>It</strong> was successfully launched.</td>
</tr>
<tr>
<td>Exact Match (6)</td>
<td><strong>Disney</strong> is a global brand.</td>
<td>The subway to <strong>Disney</strong> has been constructed.</td>
</tr>
<tr>
<td>Head Match (11)</td>
<td>Those landmark buildings are tall.</td>
<td>He has not seen <strong>these small buildings</strong>.</td>
</tr>
<tr>
<td>Other Match (7)</td>
<td>Dr. <strong>Henry</strong> notices something else.</td>
<td>Dr. <strong>Mann</strong> is a successful researcher.</td>
</tr>
<tr>
<td>Semantic Proximity (12)</td>
<td>Hong Kong cinema has nurtured many directors.</td>
<td>It memorializes Hong Kong’s film history.</td>
</tr>
<tr>
<td>Others (5)</td>
<td><strong>Paul Kelly</strong> and <strong>Steve Sodbury</strong> have no idea.</td>
<td></td>
</tr>
</tbody>
</table>
Summary
Summary

- Detector precision should not be overlooked
Summary

- Detector precision should not be overlooked
- Anaphoricity decisions are important, though they are difficult to make
Summary

• Detector precision should not be overlooked
• Anaphoricity decisions are important, though they are difficult to make
• Most linker errors require enhanced contextual reasoning
Summary

• Detector precision should not be overlooked

• Anaphoricity decisions are important, though they are difficult to make

• Most linker errors require enhanced contextual reasoning