A Benchmark of Rule-Based and Neural Coreference Resolution in Dutch Novels and News

Corbèn Poot, Andreas van Cranenburgh
@andreascvc

CRAC 2020 @ COLING, December 12
This talk:

Introduction
Setup and Results
Analysis

https://twitter.com/JenMsft/status/1132306345787568128
Definition

Coreference resolution is the task of clustering mentions in text that refer to the same persons or objects.

http://nlpprogress.com/english/coreference_resolution.html
Coreference resolution is the task of clustering mentions in text that refer to the same persons or objects.

"I voted for Obama because he was most aligned with my values", she said.

http://nlpprogress.com/english/coreference_resolution.html
Coreference resolution is the task of clustering mentions in text that refer to the same persons or objects.

"I voted for Obama because he was most aligned with my values", she said.

- Entity 1 = \{Obama, he\}
- Entity 2 = \{I, my, she\}

http://nlpprogress.com/english/coreference_resolution.html
Rule-based: deterministic, hand-written rules

Statistical: traditional (non-neural) machine learning

Neural: embeddings, CNN, recurrent nets etc.

BERT: contextual-word embeddings
- Rule-based: deterministic, hand-written rules
- Statistical: traditional (non-neural) machine learning
- Neural: embeddings, CNN, recurrent nets etc.
- BERT: contextual-word embeddings
State of the art: from rules to a neural arms race ...

OntoNotes (English)
By the way ...

#BenderRule:

The rest of this talk is about Dutch!

https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/
Research agenda/background

- Project The Riddle of Literary Quality (2012–2020)
- Next goal: Analyze plot, characters, dialogue of novels
- Domain-adaptation of NLP for literature

https://literaryquality.huygens.knaw.nl/
Datasets

<table>
<thead>
<tr>
<th></th>
<th>SoNaR-1</th>
<th>RiddleCoref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>news, wiki, etc</td>
<td>novels</td>
</tr>
<tr>
<td>Docs</td>
<td>861</td>
<td>33</td>
</tr>
<tr>
<td>Tokens</td>
<td>1M</td>
<td>160k</td>
</tr>
<tr>
<td>Tokens/doc</td>
<td>$\approx 1166$</td>
<td>$\approx 4900$</td>
</tr>
<tr>
<td>Pron/Nom/Name %</td>
<td>11/71/18</td>
<td>40/47/13</td>
</tr>
</tbody>
</table>

- SoNaR-1: automatically extracted markables
- RiddleCoref: manually annotated mentions

Van Cranenburgh (CLIN journal 2019). A Dutch coref. res. system w/evaluation on literary fiction.
System

<table>
<thead>
<tr>
<th></th>
<th>dutchcoref</th>
<th>e2e-Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>rule-based</td>
<td>neural</td>
</tr>
<tr>
<td></td>
<td>entity-based</td>
<td>mention-ranking</td>
</tr>
<tr>
<td></td>
<td>knowledge-driven</td>
<td>data-driven</td>
</tr>
<tr>
<td>Features</td>
<td>Parse trees, NER, Gazetteer etc.</td>
<td>embeddings</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(fastText, BERT)</td>
</tr>
<tr>
<td>Based on</td>
<td>Stanford sieves</td>
<td>e2e, higher-order, c2f</td>
</tr>
<tr>
<td></td>
<td>Lee et al 2013</td>
<td>Lee et al 2018</td>
</tr>
</tbody>
</table>

https://github.com/andreasvc/dutchcoref/
https://github.com/Filter-Bubble/e2e-Dutch
Rule-based system: precision-ranked sieves

End-to-end neural system

Figure adapted from Lee et al (EMNLP 2017). End-to-end neural coreference resolution. We use Lee et al (NAACL 2018). Higher-order coref. resolution w/coarse- to-fine inf.
## Results

<table>
<thead>
<tr>
<th></th>
<th>CoNLL score</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RiddleCoref</td>
</tr>
<tr>
<td><strong>dutchcoref</strong></td>
<td>69.9</td>
<td>55.9</td>
</tr>
<tr>
<td><strong>e2e-Dutch</strong></td>
<td>63.6</td>
<td><strong>68.5</strong></td>
</tr>
</tbody>
</table>

- Large coref. performance differences
Results

<table>
<thead>
<tr>
<th></th>
<th>CoNLL score</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RiddleCoref</td>
<td>SoNaR-1</td>
<td></td>
</tr>
<tr>
<td>dutchcoref</td>
<td>69.9</td>
<td>55.9</td>
<td></td>
</tr>
<tr>
<td>e2e-Dutch</td>
<td>63.6</td>
<td>68.5</td>
<td></td>
</tr>
</tbody>
</table>

▶ Large coref. performance differences

<table>
<thead>
<tr>
<th></th>
<th>Mention F1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RiddleCoref</td>
<td>SoNaR-1</td>
<td></td>
</tr>
<tr>
<td>dutchcoref</td>
<td>89.2</td>
<td>74.2</td>
<td></td>
</tr>
<tr>
<td>e2e-Dutch</td>
<td>85.3</td>
<td>87.9</td>
<td></td>
</tr>
</tbody>
</table>

▶ dutchcoref is limited by mention performance?
Detailed results (test set, predicted mentions, incl/singletons

<table>
<thead>
<tr>
<th>System</th>
<th>dataset</th>
<th>Mentions</th>
<th></th>
<th>LEA</th>
<th></th>
<th>CoNLL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R</td>
<td>P</td>
<td>F1</td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>dutchcoref</td>
<td>RiddleCoref</td>
<td>87.7</td>
<td>90.8</td>
<td>89.2</td>
<td>50.8</td>
<td>64.8</td>
</tr>
<tr>
<td>e2e-Dutch</td>
<td>RiddleCoref</td>
<td>82.0</td>
<td>89.0</td>
<td>85.3</td>
<td>44.8</td>
<td>50.5</td>
</tr>
<tr>
<td>dutchcoref</td>
<td>SoNaR-1</td>
<td>65.3</td>
<td>85.9</td>
<td>74.2</td>
<td>37.9</td>
<td>52.6</td>
</tr>
<tr>
<td>e2e-Dutch</td>
<td>SoNaR-1</td>
<td>89.0</td>
<td>86.8</td>
<td>87.9</td>
<td>60.8</td>
<td>62.5</td>
</tr>
</tbody>
</table>

- RiddleCoref: Large LEA precision difference
- SoNaR-1: Large mention/LEA recall differences

Learning curve (% training data)

e2e-Dutch performance on RiddleCoref dev set, as function of training data (initial segments of novels).
e2e-Dutch performance on RiddleCoref dev set, as function of training data (initial segments of novels).

- need more training data to beat dutchcoref
- mention performance does reach plateau
Coreference scores as a function of document length being evaluated.

- Gold and system output are truncated at different lengths (% of words);
- $r$ is correlation coefficient.
Singletons and gold mentions (dev set)

Dataset = RiddleCoref

Dataset = SoNaR-1

System
dutchcoref
e2e-Dutch
SoNaR-1 annotation issues

From a cursory inspection:
- Missing links for string matches: 5x “Amsterdam” etc.
- Missing anaphoric links
- Mention boundaries not corrected
SoNaR-1 annotation issues

From a cursory inspection:
- Missing links for string matches: 5x “Amsterdam” etc.
- Missing anaphoric links
- Mention boundaries not corrected

Remarks:
- Neural system adapts to all annotation conventions/Issues
- Rule-based system is penalized for annotation differences
Conclusions

- Neural system struggles with long documents but needs more training data to reach full potential
- Singletons inflate the scores, esp. with e2e-Dutch on SoNaR-1
- Rule-based system is affected by annotation differences/issues
- Next steps: add classifiers to rule-based system (Lee et al 2017); BERT finetuning for neural system (Joshi et al 2019).

Recommendations:

- Evaluate on long(er) documents
- Exclude singletons for evaluation
- Use semi-automatic annotation
Recommendations:
▶ Evaluate on long(er) documents
▶ Exclude singletons for evaluation
▶ Use semi-automatic annotation

Open questions:
▶ Exclude singletons during training?
▶ Why is performance gap between datasets and systems so big?
▶ What has best return on investment:
  ▶ Rule-based system (add classifiers, harmonize annotation)
  ▶ Neural system (annotate more novel data, throw more compute at it)
THE END

Models: https://github.com/andreasvc/crac2020
Thanks to my BSc thesis students for helping with annotation!

Dilbert cartoon, syndicated by Bruno Publications B.V.