A Dutch coreference resolution system with an evaluation on literary fiction

Andreas van Cranenburgh

University of Groningen, The Netherlands

November 7, 2019, Düsseldorf, CL colloquium talk
WHAT DO WE WANT?

NATURAL LANGUAGE PROCESSING

WHEN DO WE WANT IT?

WHEN DO WE WANT WHAT?

https://twitter.com/JenMsft/status/1132306345787568128
Plan for today

1. Background
2. Annotating Dutch novels
3. The coreference system
4. Evaluation
5. Future work
1. Background
2. Annotating Dutch novels
3. The coreference system
4. Evaluation
5. Future work
Definition

Coreference resolution is the task of clustering mentions in text that refer to the same underlying real world entities.

http://nlpprogress.com/english/coreference_resolution.html
Definition

Coreference resolution is the task of clustering mentions in text that refer to the same underlying real world entities.

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

http://nlpprogress.com/english/coreference_resolution.html
Definition

Coreference resolution is the task of clustering mentions in text that refer to the same underlying real world entities.

"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
Mentions

Definition

**Mention** or referring expression: span of text that refers to a person or object in the real or mention world.

NB: contrast with *markable*, a *potentially* referring expression.
Definition

Mention or referring expression: span of text that refers to a person or object in the real or mention world.

NB: contrast with markable, a potentially referring expression.

Pronouns  I, he, my, his, that, [each other], himself, …
Names  [John], [John Smith], [Mr. Smith], …
Nominals  [the man], [the flowers on [the table]], …

But not:
- Events, actions, times
Coreference: mentions of the same entity

- [Bond]₁, [James Bond]₁.
- [I]₁ took [[my]₁ bike]₂.
Winograd schemes

The [city councilmen]$_1$ refused [the demonstrators]$_2$ a permit because . . .

1. . . .[they]$_1$ feared violence.
2. . . .[they]$_2$ advocated violence.

“Al-complete” problem

History of coreference resolution

Various datasets, languages:

1996  MUC-6 shared task, English
2004  ACE shared task, English/Chinese/Arabic
2010  SemEval shared task, multilingual including Dutch
2011  CoNLL shared task, English
2012  CoNLL shared task, English/Chinese/Arabic
State of the art: from rules to a neural arms race . . .

OntoNotes (English), CoNLL scores:

- **CoNLL 2011** shared task, winner: Lee et al., rule-based \(58.3\%\)
- **CoNLL 2012** shared task, winner: Fernandes et al., perceptron \(58.7\%\)
- **EMNLP 2017** end-to-end coref. resolution, deep learning \(67.2\%\)
- **NAACL 2018** e2e + ELMO + c2f, deeper learning \(73.0\%\)
- **EMNLP 2019** e2e + BERT Large, even deeper learning \(76.9\%\)
Evaluation metrics

Coreference evaluation is a mess!

Fatally flawed metrics:
- 1996 MUC
- 1998 B³
- 2005 CEAFm, CEAFE
- 2011 CoNLL score (= avg of MUC, B³, CEAFE)
- 2011 BLANC

No known issues (yet!):  
- 2016 Link-based Entity-Aware metric (LEA)

Moosavi & Strube (ACL 2016) Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric
1. Background
2. Annotating Dutch novels
3. The coreference system
4. Evaluation
5. Future work
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/
I made a new annotation scheme...

How standards proliferate:
(see: a/c chargers, character encodings, instant messaging, etc.)

**Situation:** There are 14 competing standards.

**Soon:**

14?! Ridiculous! We need to develop one universal standard that covers everyone's use cases. Yeah!

**Situation:** There are 15 competing standards.

https://xkcd.com/927/
Annotation scheme

Simplified annotation scheme:

- Annotate mentions: include singletons, exclude non-referring expressions.
- Avoid difficult mention boundaries: no discontinuity, relative clauses
- Only annotate entity clusters, not directed anaphor-antecedent relations

https://github.com/andreasvc/dutchcoref/
Annotation workflow

1. Tokenize, parse with Alpino
2. Run coreference system
3. Manually correct output with CorefAnnotator
4. Optional: correction by second annotator

http://www.let.rug.nl/vannoord/alp/Alpino/
https://github.com/nilsreiter/CorefAnnotator/
Annotation workflow

1. Tokenize, parse with Alpino
2. Run coreference system
3. Manually correct output with CorefAnnotator
4. Optional: correction by second annotator

Result: tabular CoNLL 2012 file

```
#begin document (example); part 000
example 1  John  (0)
example 2  sees  -
example 3  Mary  (1)
#end document
```

http://www.let.rug.nl/vannoord/alp/Alpino/
https://github.com/nilsreiter/CoRefAnnotator/
## Annotated texts

<table>
<thead>
<tr>
<th></th>
<th>CLIN26 dev set</th>
<th>SemEval 2010 dev</th>
<th>Novels, dev set</th>
<th>Novels, test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>documents</td>
<td>30</td>
<td>23</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>tokens</td>
<td>4018</td>
<td>9164</td>
<td>19,051</td>
<td>88,092</td>
</tr>
<tr>
<td>sents per doc</td>
<td>7</td>
<td>21.4</td>
<td>100</td>
<td>491.5</td>
</tr>
<tr>
<td>avg sent len</td>
<td>19.3</td>
<td>18.4</td>
<td>19.0</td>
<td>16.3</td>
</tr>
</tbody>
</table>
### Annotated texts

<table>
<thead>
<tr>
<th></th>
<th>CLIN26 dev set</th>
<th>SemEval 2010 dev</th>
<th>Novels, dev set</th>
<th>Novels, test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>documents</td>
<td>30</td>
<td>23</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>tokens</td>
<td>4018</td>
<td>9164</td>
<td>19,051</td>
<td>88,092</td>
</tr>
<tr>
<td>sents per doc</td>
<td>7</td>
<td>21.4</td>
<td>100</td>
<td>491.5</td>
</tr>
<tr>
<td>avg sent len</td>
<td>19.3</td>
<td>18.4</td>
<td>19.0</td>
<td>16.3</td>
</tr>
<tr>
<td>entities</td>
<td>273</td>
<td>424</td>
<td>1798</td>
<td>8337</td>
</tr>
<tr>
<td>mentions</td>
<td>663</td>
<td>1010</td>
<td>4243</td>
<td>20,873</td>
</tr>
<tr>
<td>% pronouns</td>
<td>7.69</td>
<td>14.45</td>
<td>43.3</td>
<td>36.5</td>
</tr>
<tr>
<td>% nominal</td>
<td>52.34</td>
<td>54.35</td>
<td>46.2</td>
<td>52.2</td>
</tr>
<tr>
<td>% names</td>
<td>39.97</td>
<td>31.20</td>
<td>10.5</td>
<td>11.2</td>
</tr>
</tbody>
</table>

107k tokens of annotated literary text!
1. Background
2. Annotating Dutch novels
3. The coreference system
4. Evaluation
5. Future work
Mention detection

1. Extract candidate constituents
2. Adjust spans
3. Filter with patterns
4. Detect features

- [John], [who], [the book on [the table]], ...
- [Mr Smith], [San Jose, California], ...
- Drop: yesterday, about two meters, ...
- gender, animacy, number
Mention feature detection

- parse features in the Alpino parse tree (HPSG-inspired)
- NER part of Alpino; person/org/loc/misc
- wordnet animacy & gender of head nouns (hand-corrected)
- web text names extracted w/heuristic patterns from 30GB English text

Mention feature detection

- **parse** features in the Alpino parse tree (HPSG-inspired)
- **NER** part of Alpino; person/org/loc/misc
- **wordnet** animacy & gender of head nouns (hand-corrected)
- **web text** names extracted w/heuristic patterns from 30GB English text

<table>
<thead>
<tr>
<th></th>
<th>Pronouns</th>
<th>Nominals</th>
<th>Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>parse</td>
<td>parse</td>
<td>web text</td>
</tr>
<tr>
<td>Gender</td>
<td>parse</td>
<td>wordnet</td>
<td>web text</td>
</tr>
<tr>
<td>Animacy</td>
<td>parse</td>
<td>wordnet</td>
<td>NER, web text</td>
</tr>
</tbody>
</table>

Bergsma & Lin (COLING-ACL 2006). Bootstrapping path-based pronoun resolution.
http://aclweb.org/anthology/P06-1005
Sieves: link mentions with deterministic rules

Quote attribution  find speaker & addressee of direct speech
String match  [The boy]₁ ...[The boy]₁ ... 
Precise constructs  [the boy]₁ [who]₁ ... 
Head match  [The clever boy]₁ ...[the boy]₁ 
Proper head noun match  [Bond]₂, [James Bond]₂ 
Pronoun resolution  [He]₁ ...[his]₁ ... 

Muzny et al. (EACL 2017) A two-stage sieve approach for quote attribution
Heeyoung Lee et al. (CL 2013) Deterministic coreference resolution [...]
Move mouse over bracketed text to highlight coreference. Move mouse over direct speech to highlight speaker and addressee. Click on a sentence to toggle the display of its parse tree.

Legend: [Singleton] [Coreference] [Speaker] [Addressee] [Direct speech]

XML source


'Het is [een oude man], denk [ik]. ' Zei [ze].

'Heb [je] [hem] dan gezien? 'vroeg [ik].

'Nee, dat kan ik niet beantwoorden. Er werd gebeld, in een slecht zittend, twee keer, hij is weg. Hij kwam niet terug, in een slecht zittend....'

https://andreasvc.github.io/voskuil.html
1. Background
2. Annotating Dutch novels
3. The coreference system
4. Evaluation
5. Future work
## Evaluation: shared tasks

<table>
<thead>
<tr>
<th>CLIN26 shared task</th>
<th>Mentions</th>
<th>BLANC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GroRef, Boeing test set</td>
<td>59.34</td>
<td>30.96</td>
</tr>
<tr>
<td>This Work, Boeing test set</td>
<td><strong>59.49</strong></td>
<td><strong>31.48</strong></td>
</tr>
<tr>
<td>GroRef, GM test set</td>
<td>60.40</td>
<td>31.31</td>
</tr>
<tr>
<td>This Work, GM test set</td>
<td>59.26</td>
<td>31.07</td>
</tr>
<tr>
<td>GroRef, Stock test set</td>
<td>53.70</td>
<td>25.40</td>
</tr>
<tr>
<td>This Work, Stock test set</td>
<td><strong>54.68</strong></td>
<td><strong>26.09</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SemEval 2010, Dutch, test set</th>
<th>Mentions</th>
<th>BLANC</th>
<th>MUC</th>
<th>B³</th>
<th>CEAFm</th>
</tr>
</thead>
<tbody>
<tr>
<td>SemEval 2010: Sucre</td>
<td>42.3</td>
<td><strong>46.9</strong></td>
<td>29.7</td>
<td>11.7</td>
<td>15.9</td>
</tr>
<tr>
<td>SemEval 2010: UBIU</td>
<td>34.7</td>
<td>32.3</td>
<td>8.3</td>
<td>17.0</td>
<td>17.0</td>
</tr>
<tr>
<td>This Work</td>
<td><strong>64.27</strong></td>
<td>41.48</td>
<td><strong>51.95</strong></td>
<td><strong>45.85</strong></td>
<td><strong>51.20</strong></td>
</tr>
<tr>
<td>Dataset</td>
<td>mention F1</td>
<td>recall</td>
<td>precision</td>
<td>LEA F1</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>------------</td>
<td>--------</td>
<td>-----------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>SemEval 2010 (test set)</td>
<td>64.27</td>
<td>36.00</td>
<td>39.96</td>
<td>37.88</td>
<td></td>
</tr>
<tr>
<td>CLIN26 shared task (Boeing test set)</td>
<td>59.49</td>
<td>29.83</td>
<td>33.95</td>
<td>31.76</td>
<td></td>
</tr>
<tr>
<td>Literary texts (dev set)</td>
<td>87.05</td>
<td>57.13</td>
<td>61.71</td>
<td>59.33</td>
<td></td>
</tr>
<tr>
<td>Literary texts (test set)</td>
<td>87.10</td>
<td>49.27</td>
<td>57.45</td>
<td>53.05</td>
<td></td>
</tr>
</tbody>
</table>
Discussion

- High variance among novels; matter of style?
- Better performance on novels than news! Surprising?
  - More dialogue and pronouns in novels (some long chains)
  - Novels are longer documents
    (including our annotated fragments)
  - Not all errors are created equal . . .
1. Background
2. Annotating Dutch novels
3. The coreference system
4. Evaluation
5. Future work
Improve components

Components:
- Mention detection/spans
- Pleonastic pronoun detection
- Gender/animacy
- Quote attribution
- Pronoun resolution

Heeyoung Lee et al. (NLE 2017) A scaffolding approach to coreference resolution integrating statistical and rule-based models. https://doi.org/10.1017/S1351324917000109
Improve components

Components:
▶ Mention detection/spans
▶ Pleonastic pronoun detection
▶ Gender/animacy
▶ Quote attribution
▶ Pronoun resolution

Procedure:
1. Acquire/annotate more data
2. Train supervised classifier
3. ???
4. Profit!

Heeyoung Lee et al. (NLE 2017) A scaffolding approach to coreference resolution integrating statistical and rule-based models. https://doi.org/10.1017/S1351324917000109
Neural coreference

Train
End-to-end coreference system with BERT for Dutch ...

Kenton Lee et al. (EMNLP 2017) End-to-end neural coreference resolution
THE END

Code: https://github.com/andreasvc/dutchcoref

Paper: Coming soon™

Thanks to my BSc thesis students for helping with annotation!