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

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

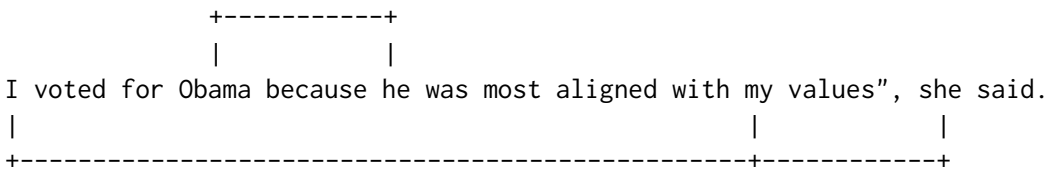
Definition

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

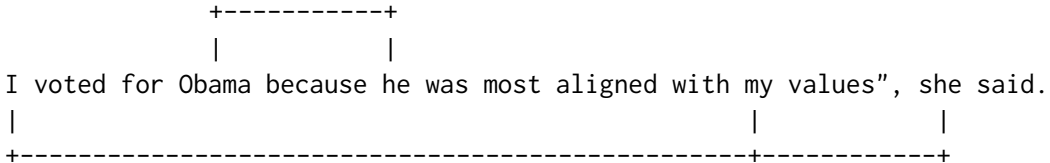
Coreference resolution is the task of clustering mentions in text that refer to the same underlying real world entities.



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.



- 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.

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.

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: equivalent mentions

- $[\text{John}]_1$ sees $[\text{Mary}]_2$. $[\text{He}]_1$ waves at $[\text{her}]_2$.
- $[\text{Bond}]_1$, $[\text{James Bond}]_1$.
- $[\text{I}]_1$ took $[[\text{my}]_1$ 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.

"AI-complete" problem

Kehler (2002). Coherence, reference, and the theory of grammar, chapter 1.
http://web.stanford.edu/group/cslipublications/cslipublications/pdf/1575862166.pdf

History

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 B^3

2005 CEAFm, CEAFe

2011 CoNLL score (= avg of MUC, B^3, 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

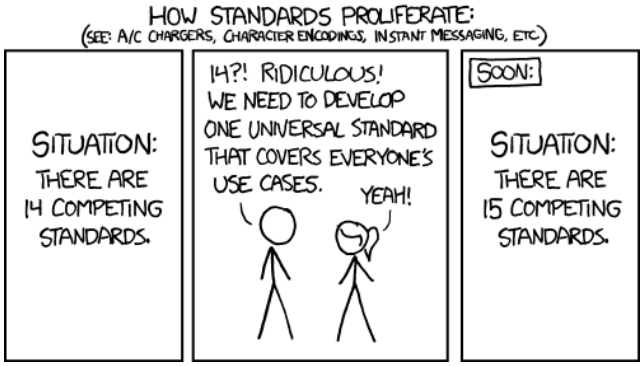
By the way . . .

#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 . . .



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: correct 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: correct 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

	CLIN26 dev set	SemEval 2010 dev	Novels, dev set	Novels, test set
documents	30	23	10	11
tokens	4018	9164	19,051	88,092
sents per doc	7	21.4	100	491.5
avg sent len	19.3	18.4	19.0	16.3

Annotated texts

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

108k tokens of annotated literary text!

Mention detection

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

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

Bergsma & Lin (COLING-ACL 2006). Bootstrapping path-based pronoun resolution.
`http://aclweb.org/anthology/P06-1005`

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

	Animacy	Gender	Number
Pronouns	parse	parse	parse
Nominals	wordnet	wordnet	parse
Names	NER, web text	web text	web text

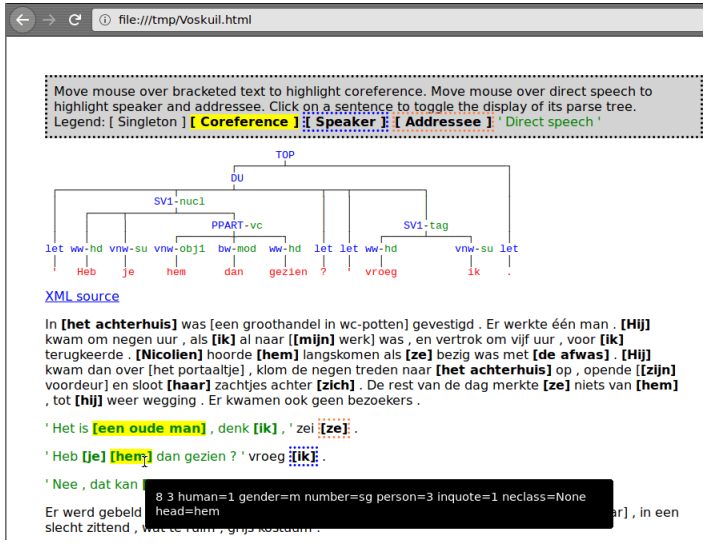
Bergsma & Lin (COLING-ACL 2006). Bootstrapping path-based pronoun resolution.
`http://aclweb.org/anthology/P06-1005`

Sieves: link mentions with deterministic rules

Quote attribution
String match
Precise constructs
Head match
Proper head noun match
Pronoun resolution

Muzny et al. (EACL 2017) A two-stage sieve approach for quote attribution
Heeyoung Lee et al. (CL 2013) Deterministic coreference resolution [...]

Demo



https://andreasvc.github.io/voskuil.html

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

Evaluation: shared tasks

CLIN26 shared task	Mentions	BLANC
GroRef, Boeing test set	59.34	30.96
This Work, Boeing test set	**59.49**	**31.48**
GroRef, GM test set	**60.40**	**31.31**
This Work, GM test set	59.26	31.07
GroRef, Stock test set	53.70	25.40
This Work, Stock test set	**54.68**	**26.09**

SemEval 2010, Dutch, test set	Mentions	BLANC	MUC	B^3	CEAFm
SemEval 2010: Sucre	42.3	**46.9**	29.7	11.7	15.9
SemEval 2010: UBIU	34.7	32.3	8.3	17.0	17.0
This Work	**64.27**	41.48	**51.95**	**45.85**	**51.20**

Evaluation: novels

	mention F1	recall	precision	LEA F1
SemEval 2010 (test set)	64.27	36.00	39.96	37.88
CLIN26 shared task (Boeing test set)	59.49	29.83	33.95	31.76
Literary texts (dev set)	87.05	57.13	61.71	59.33
Literary texts (test set)	87.10	49.27	57.45	53.05

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
- Pronoun resolution
- Quote attribution

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
- Pronoun resolution
- Quote attribution

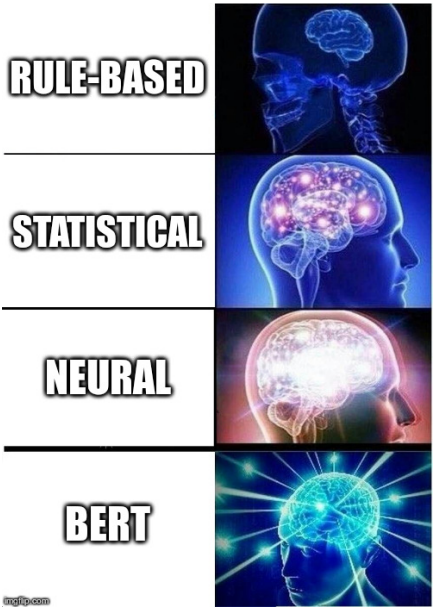
Procedure:

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!

Dilbert cartoon, syndicated by Bruno Publications B.V.