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

Corbèn Poot, Andreas van Cranenburgh
@andreasvc

university of groningen

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

Definition

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

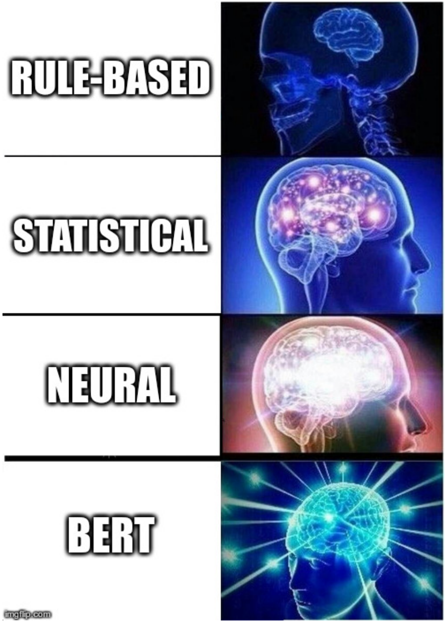
Definition

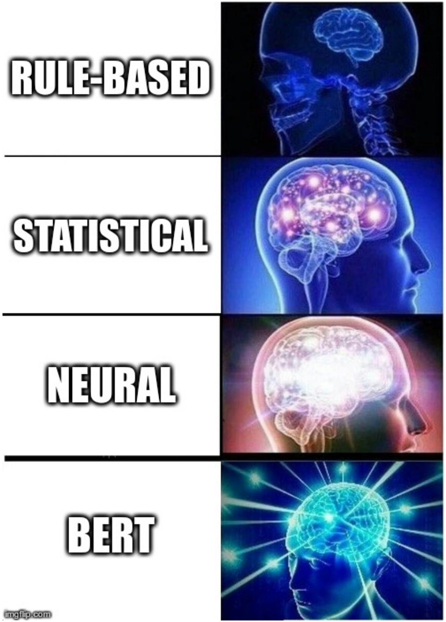
Coreference resolution is the task of clustering mentions in text that refer to the same persons or objects.

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

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

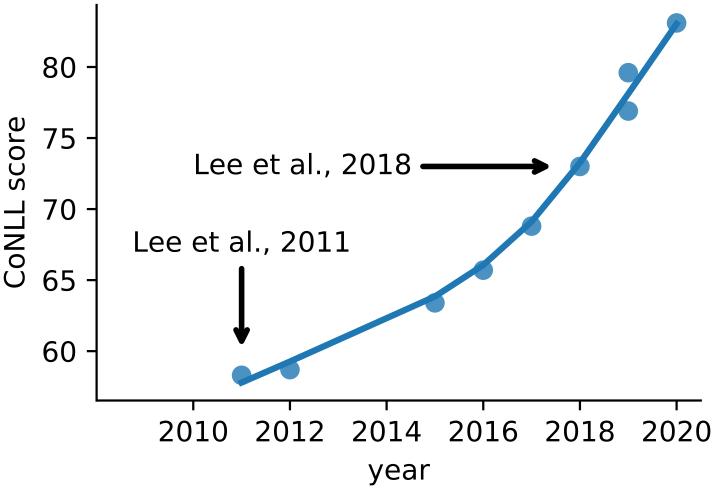
http://nlpprogress.com/english/coreference_resolution.html


RULE-BASED
STATISTICAL
NEURAL
BERT
imgflip.com



- 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

	SoNaR-1	RiddleCoref
Domain	news, wiki, etc	novels
Docs	861	33
Tokens	1M	160k
Tokens/doc	≈ 1166	≈ 4900
Pron/Nom/Name %	11/71/18	40/47/13

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

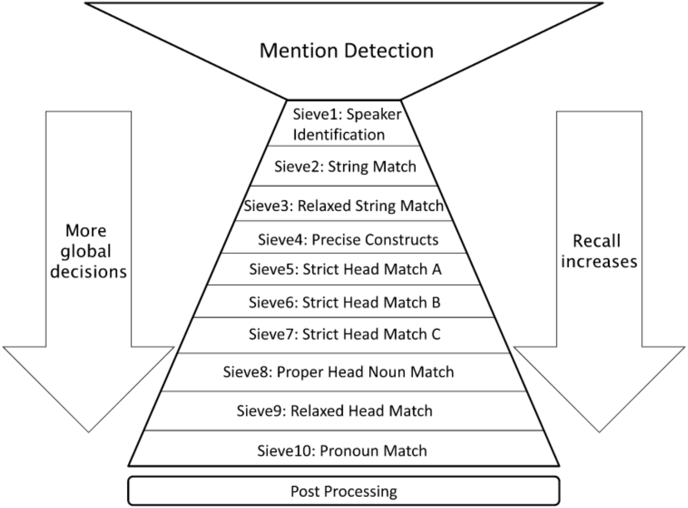
Schuurman et al (LREC 2010). [...] SoNaR, a reference corpus of contemporary written Dutch.
Van Cranenburgh (CLIN journal 2019). A Dutch coref. res. system w/evaluation on literary fiction.

Systems

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

https://github.com/andreasvc/dutchcoref/
https://github.com/Filter-Bubble/e2e-Dutch

Rule-based system: precision-ranked sieves



Lee et al (CL 2013). Deterministic coref. res. based on entity-centric, precision-ranked rules.

End-to-end neural system

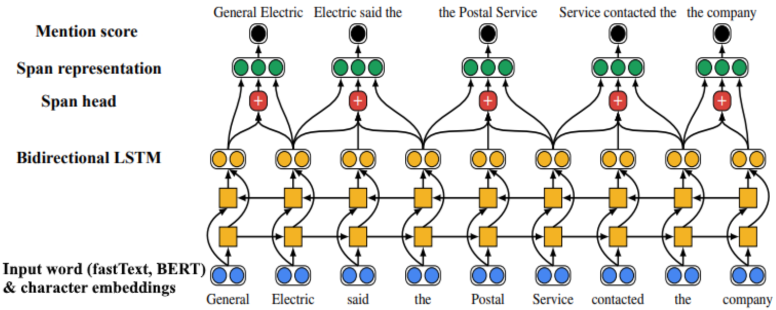


Figure adapted from Lee et al (EMNLP 2017). End-to-end neural coreference resolution.
We use Lee et al (NAACL 2018). Higher-order coref. resolution w/coarse- to-fine inf.

Results

	CoNLL score	
	RiddleCoref	SoNaR-1
dutchcoref	**69.9**	55.9
e2e-Dutch	63.6	**68.5**

- Large coref. performance differences

Results

	CoNLL score	
	RiddleCoref	SoNaR-1
dutchcoref	**69.9**	55.9
e2e-Dutch	63.6	**68.5**

- Large coref. performance differences

	Mention F1	
	RiddleCoref	SoNaR-1
dutchcoref	**89.2**	74.2
e2e-Dutch	85.3	**87.9**

- dutchcoref is limited
  by mention performance?

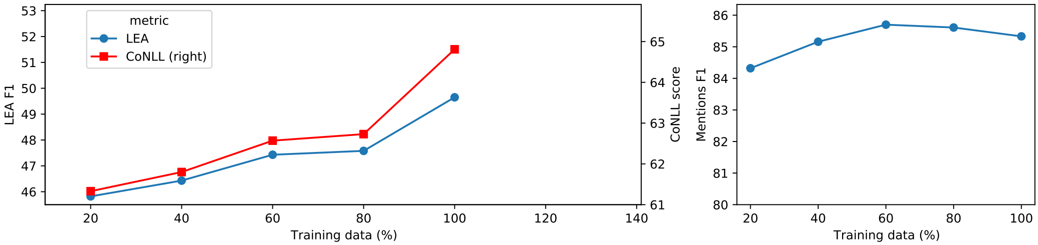
Detailed results (test set, predicted mentions, incl/singletons

System	dataset	Mentions			LEA			CoNLL
		R	P	F1	R	P	F1	
dutchcoref	RiddleCoref	87.7	90.8	89.2	50.8	64.8	**57.0**	**69.9**
e2e-Dutch	RiddleCoref	82.0	89.0	85.3	44.8	50.5	47.5	63.6
dutchcoref	SoNaR-1	65.3	85.9	74.2	37.9	52.6	44.0	55.9
e2e-Dutch	SoNaR-1	89.0	86.8	87.9	60.8	62.5	**61.6**	**68.5**

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

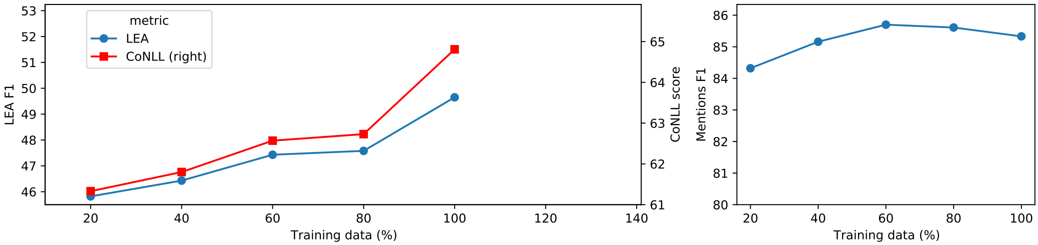
Moosavi & Strube (ACL 2016). Which coreference evaluation metric do you trust?.
https://github.com/ns-moosavi/coval/

Learning curve (% training data)



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

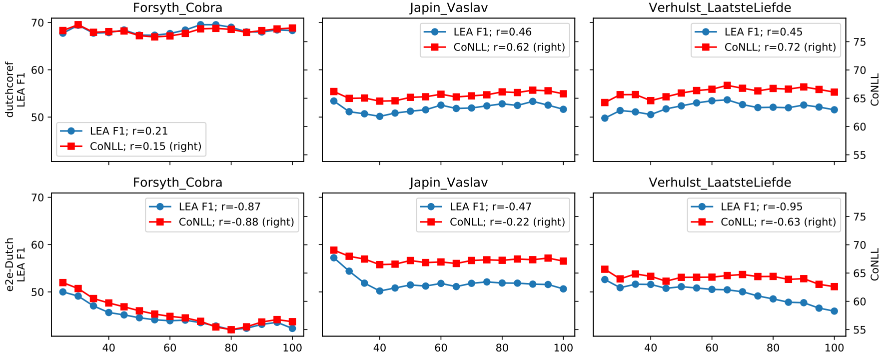
Learning curve (% training data)



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

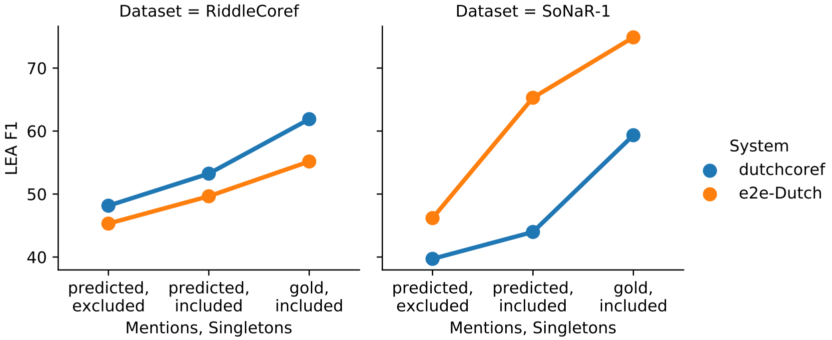
- need more training data to beat dutchcoref
- mention performance does reach plateau

Document length



- 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)



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).

Lee et al (NLE 2017). A scaffolding approach to coref. res. integrating statistical and rule-based models.
Joshi et al (EMNLP 2019). BERT for coreference resolution: Baselines and analysis.

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

Paper: https://arxiv.org/abs/2011.01615

Thanks to my BSc thesis students for helping with annotation!

Dilbert cartoon, syndicated by Bruno Publications B.V.