



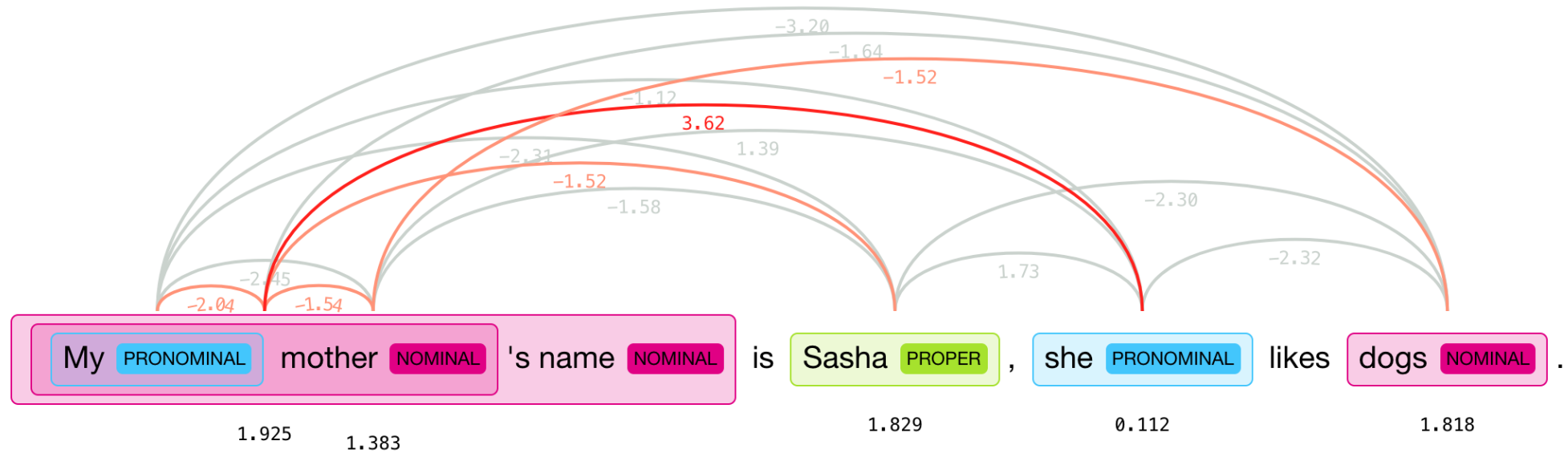
LUND
UNIVERSITY

Coreference Resolution on Biomedical Texts

By Nicolás Jaua



What's Coreference?



Tools

- Hugging Face's neuralcoref
- CRAFT
- Google Colab
- Alvis Cluster

```
194730 0 1 The DT - - - - - - - (2
194730 0 2 protein NN - - - - - - - 2)
194730 0 3 belongs VBZ - - - - - - - -
194730 0 4 to IN - - - - - - - -
194730 0 5 a DT - - - - - - - -
194730 0 6 family NN - - - - - - - -
194730 0 7 of IN - - - - - - - -
194730 0 8 evolutionarily RB - - - - - - -
194730 0 9 conserved VBN - - - - - - - -
194730 0 10 proteins NN - - - - - - - -
194730 0 11 of IN - - - - - - - -
194730 0 12 a DT - - - - - - - -
194730 0 13 bipartite JJ - - - - - - - -
194730 0 14 structure NN - - - - - - - -
194730 0 15 with IN - - - - - - - -
194730 0 16 a DT - - - - - - - (32a
194730 0 17 variable JJ - - - - - - - -
194730 0 18 N NN - - - - - - - -
194730 0 19 - HYPH - - - - - - - -
194730 0 20 terminal JJ - - - - - - - 32a)
194730 0 21 and CC - - - - - - - -
194730 0 22 a DT - - - - - - - (33
194730 0 23 conserved VBN - - - - - - - -
194730 0 24 C NN - - - - - - - -
194730 0 25 - HYPH - - - - - - - -
194730 0 26 terminal JJ - - - - - - - -
194730 0 27 domain NN - - - - - - - (32a)|33)
194730 0 28 . . - - - - - - - -
```

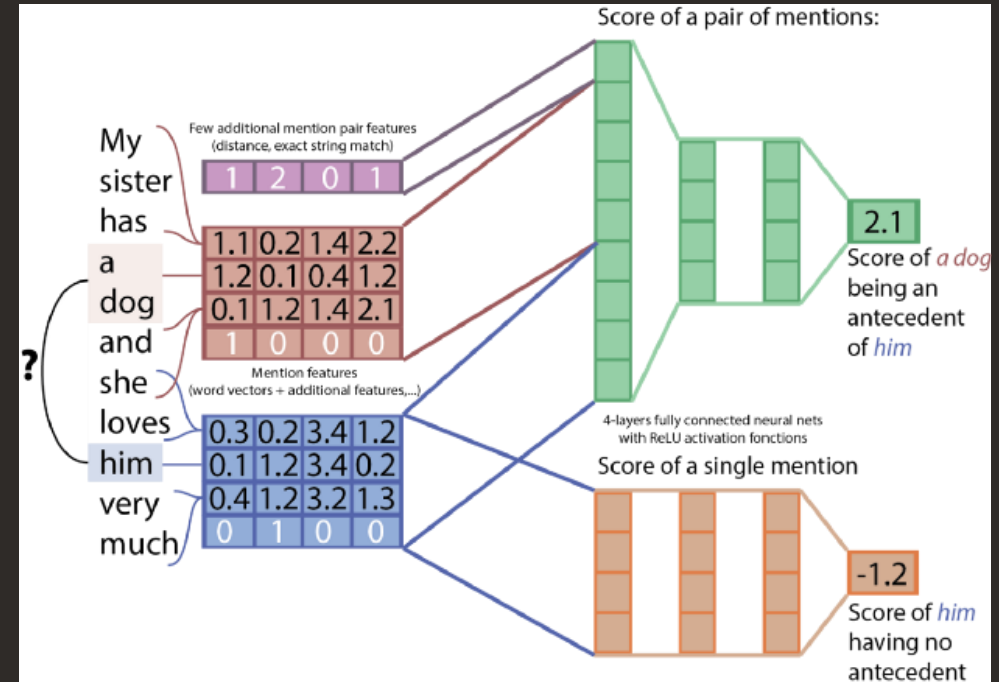
Method

- Prepare the data for training.
 - Get Corpus in expected format.
 - Parse the files into numpy arrays.
- Train a new model.
- Evaluate and compare results.





```
1 #begin document (bc/cctv/00/cctv_0005); part 003
2
3 bc/cctv/00/cctv_0005 -3- 0- Yes UH (TOP(S(INTJ*)) - Wang_shilin * (ARGM-DIS*)
4 bc/cctv/00/cctv_0005 -3- 1- , , - Wang_shilin * *
5 bc/cctv/00/cctv_0005 -3- 2- I PRP (NP*) - Wang_shilin * (ARG0*) (12)
6 bc/cctv/00/cctv_0005 -3- 3- noticed VBD (VP* notice 01 1 Wang_shilin * (V*)
7 bc/cctv/00/cctv_0005 -3- 4- that IN (SBAR* - Wang_shilin * (ARG1*)
8 bc/cctv/00/cctv_0005 -3- 5- many JJ (S(NP(NP*) - Wang_shilin * (ARG0*)
9 bc/cctv/00/cctv_0005 -3- 6- friends NNS *) - Wang_shilin * *
10 bc/cctv/00/cctv_0005 -3- 7- , , * - Wang_shilin * *
11 bc/cctv/00/cctv_0005 -3- 8- around IN (PP* - Wang_shilin * *
12 bc/cctv/00/cctv_0005 -3- 9- me PRP (NP*)) - Wang_shilin * * (12)
13 bc/cctv/00/cctv_0005 -3- 10- received VBD (VP* receive 01 1 Wang_shilin * (V*)
14 bc/cctv/00/cctv_0005 -3- 11- it PRP (NP*)) - Wang_shilin * (ARG1*) (119)
15 bc/cctv/00/cctv_0005 -3- 12- . . *) - Wang_shilin * *
16
17 bc/cctv/00/cctv_0005 -3- 0- It PRP (TOP(S(NP*)) - Wang_shilin * *
18 bc/cctv/00/cctv_0005 -3- 1- seems VBZ (VP* seem 01 1 Wang_shilin * (V*)
19 bc/cctv/00/cctv_0005 -3- 2- that IN (SBAR* - Wang_shilin * (ARG1*)
20 bc/cctv/00/cctv_0005 -3- 3- almost RB (S(NP* - Wang_shilin * (ARG0*)
21 bc/cctv/00/cctv_0005 -3- 4- everyone NN *) - Wang_shilin * *
22 bc/cctv/00/cctv_0005 -3- 5- received VBD (VP* receive 01 1 Wang_shilin * (V*)
23 bc/cctv/00/cctv_0005 -3- 6- this DT (NP*) - Wang_shilin * (ARG1*) (119)
24 bc/cctv/00/cctv_0005 -3- 7- SMS NN *) - Wang_shilin * * (119)
25 bc/cctv/00/cctv_0005 -3- 8- . . *) - Wang_shilin * *
26
27 #end document
```


Method

- Prepare the data for training.
 - Get Corpus in expected format.
 - Parse the files into numpy arrays.
- Train a new model.
- Evaluate and compare results.



Method

- Prepare the data for training. 
- Get Corpus in expected format. 
- Parse the files into numpy arrays. 
- Train a new model. 
- Evaluate and compare results.

Issues

- Memory issues.
- Dependency issues when training on clusters.
- Issues with scoring wrapper.

Method

- Prepare the data for training. ✓
 - Get Corpus in expected format. ✓
 - Parse the files into numpy arrays. ✓
- Train a new model. ✗
- Evaluate and compare results.

Issues

- Memory issues. ✓
- Dependency issues when training on clusters. ✓
- Issues with scoring wrapper. ✗

Method

- Prepare the data for training.
 - Get Corpus in expected format.
 - Parse the files into numpy arrays.
- Train a new model.
- Evaluate and compare results.



17194222

muc

===== TOTALS =====

Identification of Mentions: Recall: (8 / 1671) 0.47% Precision: (8 / 245) 3.26% F1: 0.83%

Coreference: Recall: (1 / 1268) 0.07% Precision: (1 / 167) 0.59% F1: 0.13%

bcub

===== TOTALS =====

Identification of Mentions: Recall: (8 / 1671) 0.47% Precision: (8 / 245) 3.26% F1: 0.83%

Coreference: Recall: (1.16066578458937 / 1671) 0.06% Precision: (3.27310924369748 / 245) 1.33% F1: 0.13%

ceafe

===== TOTALS =====

Identification of Mentions: Recall: (8 / 1671) 0.47% Precision: (8 / 245) 3.26% F1: 0.83%

Coreference: Recall: (0.912842712842713 / 403) 0.22% Precision: (0.912842712842713 / 78) 1.17% F1: 0.37%

Evaluation

- Identification of mentions:

- Recall:

$$\frac{|\text{mentions in common}|}{|\text{mentions in gold standard}|}$$

- Precision:

$$\frac{|\text{mentions in common}|}{|\text{mentions in output}|}$$

- F1: $\frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$

- MUC: $\frac{|\text{links in common}|}{|\text{links in file}|}$

17194222

muc

===== TOTALS =====

Identification of Mentions: Recall: (8 / 1671) 0.47% Precision: (8 / 245)
3.26% F1: 0.83%

Coreference: Recall: (1 / 1268) 0.07% Precision: (1 / 167) 0.59% F1: 0.13%

bcub

===== TOTALS =====

Identification of Mentions: Recall: (8 / 1671) 0.47% Precision: (8 / 245)
3.26% F1: 0.83%

Coreference: Recall: (1.16066578458937 / 1671) 0.06% Precision:
(3.27310924369748 / 245) 1.33% F1: 0.13%

ceafe

===== TOTALS =====

Identification of Mentions: Recall: (8 / 1671) 0.47% Precision: (8 / 245)
3.26% F1: 0.83%

Coreference: Recall: (0.912842712842713 / 403) 0.22% Precision:
(0.912842712842713 / 78) 1.17% F1: 0.37%

Evaluation

- B-Cubed:

- $Recall(m_i) = \frac{|R_{m_i} \cap K_{m_i}|}{|K_{m_i}|}$
- $Precision(m_i) = \frac{|R_{m_i} \cap K_{m_i}|}{|R_{m_i}|}$

- CEAF:

- $\phi(K_i, R_i) = \frac{2|R_i \cap K_i|}{|R_i| + |K_i|}$
- $\frac{\Phi(g^*)}{|entities\ in\ file|}$

17194222

muc

===== TOTALS =====

Identification of Mentions: Recall: (8 / 1671) 0.47% Precision: (8 / 245) 3.26% F1: 0.83%

Coreference: Recall: (1 / 1268) 0.07% Precision: (1 / 167) 0.59% F1: 0.13%

bcub

===== TOTALS =====

Identification of Mentions: Recall: (8 / 1671) 0.47% Precision: (8 / 245) 3.26% F1: 0.83%

Coreference: Recall: (1.16066578458937 / 1671) 0.06% Precision: (3.27310924369748 / 245) 1.33% F1: 0.13%

ceafe

===== TOTALS =====

Identification of Mentions: Recall: (8 / 1671) 0.47% Precision: (8 / 245) 3.26% F1: 0.83%

Coreference: Recall: (0.912842712842713 / 403) 0.22% Precision: (0.912842712842713 / 78) 1.17% F1: 0.37%

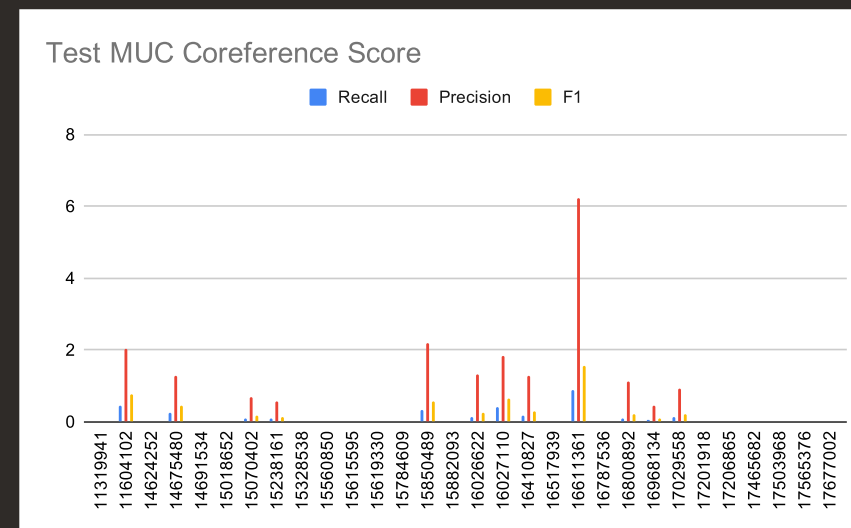
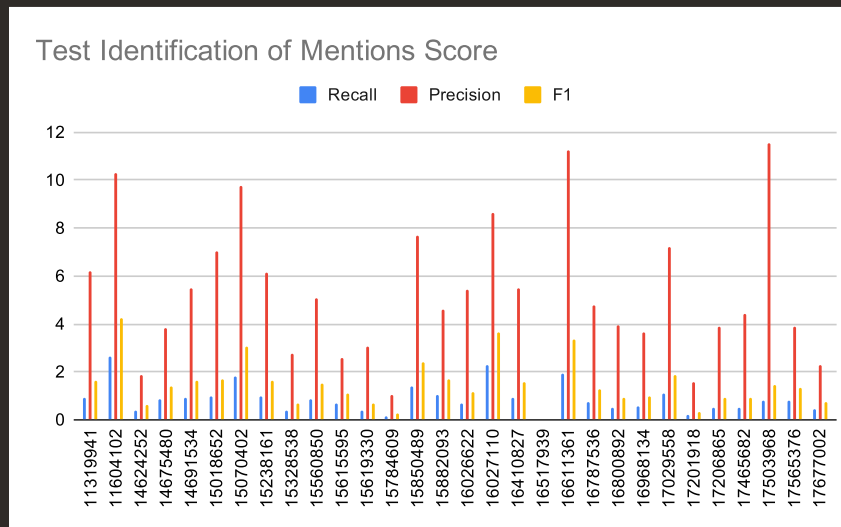
Results

```
1858683 0 1 The DT - - - - - (5
1858683 0 2 pygopus NN - - - - - -
1858683 0 3 gene NN - - - - - -
1858683 0 4 of IN - - - - - -
1858683 0 5 Drosophila NNP - - - - - (6)|5)
1858683 0 6 encodes VBZ - - - - - -
1858683 0 7 an DT - - - - - -
1858683 0 8 essential JJ - - - - - -
1858683 0 9 component NN - - - - - -
1858683 0 10 of IN - - - - - -
1858683 0 11 the DT - - - - - (7
1858683 0 12 Armadillo NN - - - - - -
1858683 0 13 ( -LRB- - - - - -
1858683 0 14  $\beta$  NN - - - - - -
1858683 0 15 - HYPH - - - - - -
1858683 0 16 catenin NN - - - - - -
1858683 0 17 ) -RRB- - - - - -
1858683 0 18 transcription NN - - - - - -
1858683 0 19 factor NN - - - - - -
1858683 0 20 complex NN - - - - - -
1858683 0 21 of IN - - - - - -
1858683 0 22 canonical JJ - - - - - (8
1858683 0 23 Wnt NN - - - - - -
1858683 0 24 signaling NN - - - - - 8)|7)
1858683 0 25 . . - - - - -
```

```
1713256 0 1 The - - - - - -
1713256 0 2 pygopus - - - - - -
1713256 0 3 gene - - - - - -
1713256 0 4 of - - - - - -
1713256 0 5 Drosophila - - - - - -
1713256 0 6 encodes - - - - - -
1713256 0 7 an - - - - - -
1713256 0 8 essential - - - - - -
1713256 0 9 component - - - - - -
1713256 0 10 of - - - - - -
1713256 0 11 the - - - - - -
1713256 0 12 Armadillo - - - - - -
1713256 0 13 ( - - - - - -
1713256 0 14  $\beta$  - - - - - -
1713256 0 15 - - - - - -
1713256 0 16 catenin - - - - - -
1713256 0 17 ) - - - - - -
1713256 0 18 transcription - - - - - -
1713256 0 19 factor - - - - - -
1713256 0 20 complex - - - - - -
1713256 0 21 of - - - - - -
1713256 0 22 canonical - - - - - (1
1713256 0 23 Wnt - - - - - 1)
1713256 0 24 signaling - - - - - -
1713256 0 25 . - - - - -
```

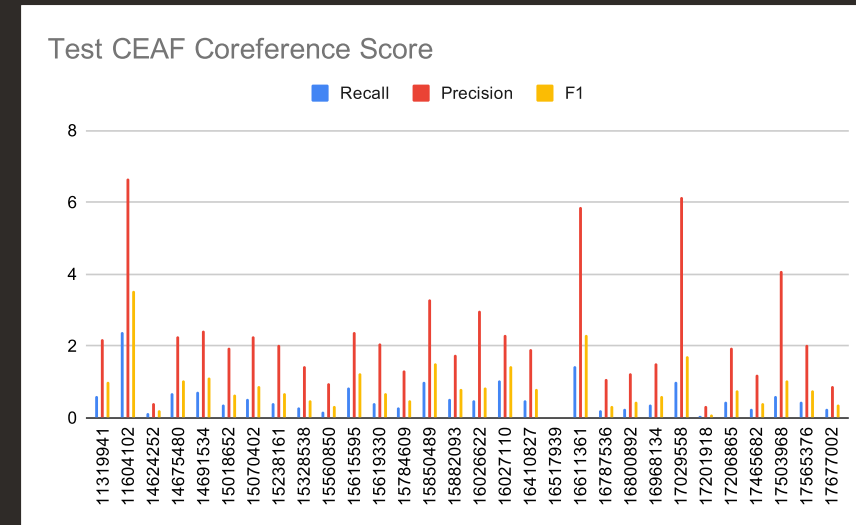
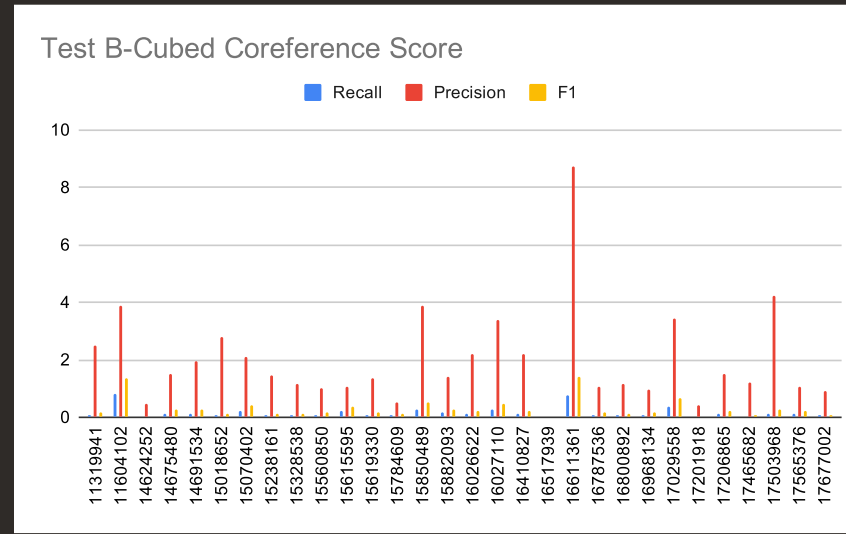
Results

- Identification:
 - Recall: 0.878%
 - Precision: 5.173%
 - F1: 1.487%
- Coreference (MUC):
 - Recall: 0.105%
 - Precision: 0.666%
 - F1: 0.181%



Results

- Coreference (B-Cubed):
 - Recall: 0.160%
 - Precision: 1.981%
 - F1: 0.293%
- Coreference (CEAF):
 - Recall: 0.563%
 - Precision: 2.238%
 - F1: 0.889%



Conclusion and Future Work

- Neuralcoref does not work on biomedical text without training.
- Fix issues with scorer and use parsed files and fixed scripts to train a neuralcoref model using CRAFT.
- Find another tool for coreference resolution training.
- Find other pretrained models that might give better results.

Thank you for listening!

Any questions?



LUND
UNIVERSITY