Coreference Resolution for Biomedical Articles

Using SpanBERT for Coreference Resolution with the CRAFT Corpus

PROBLEM INTRODUCTION

- Massive amount of published biomedical articles
- COVID-19
- Coreference & Coreference resolution
- Mentions
- Real-life Entities

Computers are useless. They can only give you answers.
- Pablo Picasso

All the world's a stage, and all the men and women merely players. They have their exits and their entrances; And one man in his time plays many parts.
- William Shakespeare

SpanBERT

- BERT
- Google
- Spans & Tokens
- BERT and SpanBERT for Coreference Resolution

DATA & FORMATS

- CRAFT
- Knowtator2
- CoNLL
- OntoNotes
- Input/Output
- Jsonlines

CRAFT converted to CoNLL

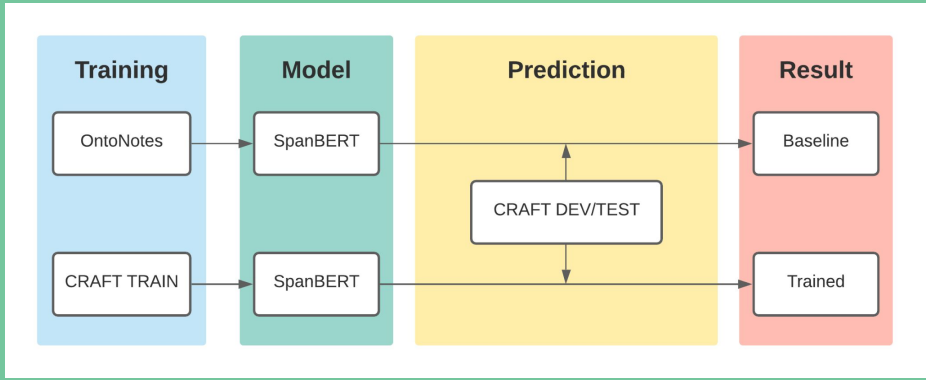
```
#begin document (1860061); part 000
1860061 0 1     Combining   VBG -   -   -   -   -   -   -   -
1860061 0 2     global  JJ  -   -   -   -   -   -   -   (1|(1b
1860061 0 3     genome  NN  -   -   -   -   -   -   -   1b)|1)
1860061 0 4     and CC  -   -   -   -   -   -   -   -
1860061 0 5     transcriptome   NN  -   -   -   -   -   -   -   (2
1860061 0 6     approaches  NNS -   -   -   -   -   -   -   (1b)|2)
1860061 0 7     to  TO  -   -   -   -   -   -   -   -
1860061 0 8     identify    VB  -   -   -   -   -   -   -   -
1860061 0 9     the DT  -   -   -   -   -   -   -   (3
1860061 0 10    candidate   NN  -   -   -   -   -   -   -   -
1860061 0 11    genes   NNS -   -   -   -   -   -   -   -
1860061 0 12    of  IN  -   -   -   -   -   -   -   -
1860061 0 13    small   JJ  -   -   -   -   -   -   -   (4|(5
1860061 0 14    -   HYPH    -   -   -   -   -   -   -   -
1860061 0 15    effect  NN  -   -   -   -   -   -   -   -
1860061 0 16    quantitative    JJ  -   -   -   -   -   -   -   -   -
1860061 0 17    trait   NN  -   -   -   -   -   -   -   -
1860061 0 18    loci    NNS -   -   -   -   -   -   -   4)
1860061 0 19    in  IN  -   -   -   -   -   -   -   -
1860061 0 20    collagen    NN  -   -   -   -   -   -   -   (6
1860061 0 21    -   HYPH    -   -   -   -   -   -   -   -
1860061 0 22    induced VBN -   -   -   -   -   -   -   -
1860061 0 23    arthritis   NN  -   -   -   -   -   -   -   6)|5)|3)
```

Input for SpanBERT: JSON Lines

```
{
  "clusters": [],
  "doc_key": "0000000",
  "sentences": [["[CLS]", "subword1", "##subword1", ".", "[SEP]"]],
  "speakers": [["[SPL]", "-", "-", "-", "[SPL]"]],
  "sentence_map": [0, 0, 0, 0, 0],
  "subtoken_map": [0, 0, 0, 1, 1]
}
```

METHOD & STRATEGY

- Tokenization
- Minimize
- Chunks
- Baseline
- Training
- Evaluation



Workflow of the project

EVALUATION

- CoNLL-2012
- MUC, BCUB & CEAFE
- Built-in Scorer
- F1, Recall & Precision
- Baseline & Trained Model

```
Official result for muc
version: 8.01 /cephyr/NOBACKUP/groups/snic2021-23-312/Oskar/coref/conll-2012/scorer/v8.01/lib/CorScorer.pm

====== TOTALS =======
Identification of Mentions: Recall: (516619 / 935320) 55.23%    Precision: (516619 / 728274) 70.93% F1: 62.1%
--------------------------------------------------------------------------------
Coreference: Recall: (371877 / 912826) 40.73%   Precision: (371877 / 718293) 51.77% F1: 45.59%
--------------------------------------------------------------------------------

Official result for bcub
version: 8.01 /cephyr/NOBACKUP/groups/snic2021-23-312/Oskar/coref/conll-2012/scorer/v8.01/lib/CorScorer.pm

====== TOTALS =======
Identification of Mentions: Recall: (516619 / 935320) 55.23%    Precision: (516619 / 728274) 70.93% F1: 62.1%
--------------------------------------------------------------------------------
Coreference: Recall: (65434.3082716004 / 935320) 6.99%  Precision: (42831.8092850174 / 728274) 5.88%    F1: 6.39%
--------------------------------------------------------------------------------

Official result for ceafe
version: 8.01 /cephyr/NOBACKUP/groups/snic2021-23-312/Oskar/coref/conll-2012/scorer/v8.01/lib/CorScorer.pm

====== TOTALS =======
Identification of Mentions: Recall: (516619 / 935320) 55.23%    Precision: (516619 / 728274) 70.93% F1: 62.1%
--------------------------------------------------------------------------------
Coreference: Recall: (2334.19333245107 / 22494) 10.37%  Precision: (2334.19333245107 / 9981) 23.38% F1: 14.37%
--------------------------------------------------------------------------------

Average F1 (conll): 22.12%
Average F1 (py): 50.11% on 60 docs
Average precision (py): 60.02%
Average recall (py): 43.18%
```

BASELINE - DEV

Average F1 (CoNLL): 19.40% | Average F1 (py): 25.88%

Evaluation	F1 - Score (%)	Precision (%)	Recall (%)
Mentions	42	72	29
muc	33	57	23
bcub	11	19	8
ceafe	14	31	9

BASELINE - TEST

Average F1 (CoNLL): 14.32% | Average F1 (py): 23.56%

Evaluation	F1 - Score (%)	Precision (%)	Recall (%)
Mentions	37	77	24
muc	26	54	17
bcub	4	10	2
ceafe	13	26	9

TRAINED - DEV

Average F1 (CoNLL): 27.16% | Average F1 (py): 52.71%

Evaluation	F1 - Score (%)	Precision (%)	Recall (%)
Mentions	66	71	62
muc	51	53	49
bcub	15	13	17
ceafe	16	30	11

TRAINED-TEST

Average F1 (CoNLL): 22.12% | Average F1 (py): 50.11%

Evaluation	F1 - Score (%)	Precision (%)	Recall (%)
Mentions	62	71	55
muc	46	52	41
bcub	6	6	7
ceafe	14	23	10

CONCLUSIONS

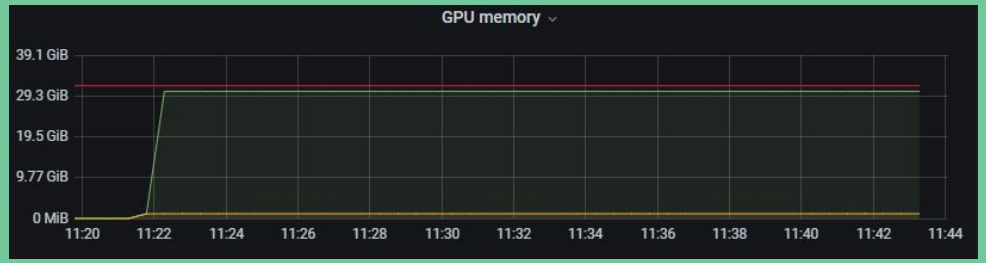
- Trained model performs better than baseline
- Built in scorer & CoNLL-2012
- F1 increase of 7.8% & 26.55%
- Coreference Resolution is hard
- Compared to OntoNotes performance

Scorer	Baseline	Trained
CoNLL	14.32%	22.12%
Built-in	23.56%	50.11%

DIFFICULTIES

- File format & structure
- Different annotations in same format
- Memory Problems
- Hardware

Memory usage for a prediction on the cluster



FUTURE WORK

- Training
- Hyperparameters
- SpanBERT Large

Default Configuration for SpanBERT Base

```
spanbert_base = ${best}{
  num_docs = 2802
  bert_learning_rate = 2e-05
  task_learning_rate = 0.0001
  max_segment_len = 384
  ffnn_size = 3000
  train_path = ${data_dir}/train.english.384.jsonlines
  eval_path = ${data_dir}/dev.english.384.jsonlines
  conll_eval_path = ${data_dir}/dev.english.v4_gold_conll
  max_training_sentences = 3
  bert_config_file = ${best.log_root}/spanbert_base/bert_config.json
  vocab_file = ${best.log_root}/spanbert_base/vocab.txt
  tf_checkpoint = ${best.log_root}/spanbert_base/model.max.ckpt
  init_checkpoint = ${best.log_root}/spanbert_base/model.max.ckpt
}
```

THANKS!

Oskar Jönsson
os2947jo-s@student.lu.se

SPECIAL THANKS TO
Sonja Aits & others involved in the project.

HAVE ANY QUESTIONS?