Named entity recognition report

Author: Alassane Watt -- alassane.watt@student.ecp.fr

Plan:

1) Experimentations

2) Performance comparison

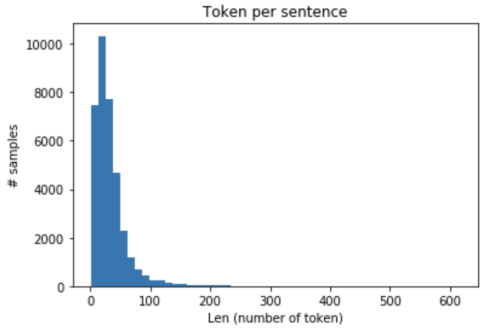
Experimentations

We evaluated 10 configurations of the NeuroNLP2 NER model on the test set of the dataset the ner model was trained on:

- w2v embeddings trained on small press corpus + ner trained on press fra4_ID
- w2v embeddings trained on med corpus + ner trained on emea
- w2v embeddings trained on med corpus + ner trained on medline
- w2v embeddings trained on full press corpus + ner trained on fra4
- fasttext embeddings trained on small press corpus + ner trained on fra4
- fasttext embeddings trained on full press corpus + ner trained on fra4
- fasttext embeddings trained on med corpus + ner trained on emea
- fasttext embeddings trained on med corpus + ner trained on medline
- fasttext embeddings trained on med corpus + ner trained on fra4 press
- fasttext embeddings trained on small press corpus + ner trained on emea

w2v embeddings trained on small press corpus + ner trained on press fra4_ID

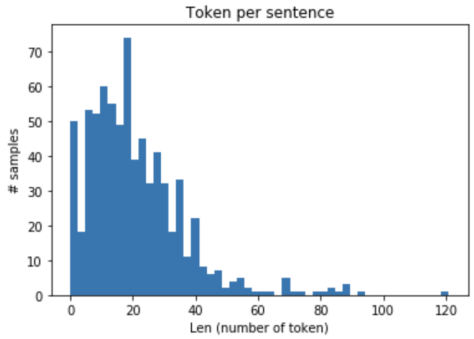
Number of sentences: 35723
Number of words: 1156339
Number of unique words: 37863
Number of tags: 11
Different tags: {'i-func', 'b-org', 'b-pers', 'O', 'b-loc', 'b-func', 'i-loc', 'i-prod', 'i-pers', 'b-prod', 'i-org'}



test acc: 97.29%, precision: 69.52%, recall: 64.44%, F1: 66.88%

w2v embeddings trained on med corpus + ner trained on emea

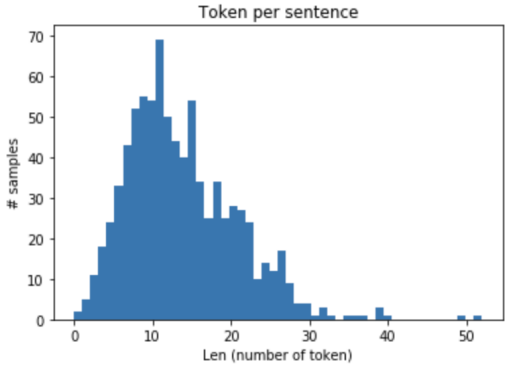
Number of sentences: 738
Number of words: 15339
Number of unique words: 2599
Number of tags: 21
Different tags: {'B-DISO', 'O', 'I-ANAT', 'B-PHEN', 'I-PROC', 'B-CHEM', 'I-GEOG', 'I-DEVI', 'B-GEOG', 'I-LIVB', 'I-PHYS', 'B-PROC', 'I-OBJC', 'B-LIVB', 'B-PHYS', 'B-DEVI', 'B-ANAT', 'I-PHEN', 'B-OBJC', 'I-DISO', 'I-CHEM'}



test acc: 84.05%, precision: 47.17%, recall: 35.24%, F1: 40.34%

w2v embeddings trained on med corpus + ner trained on medline

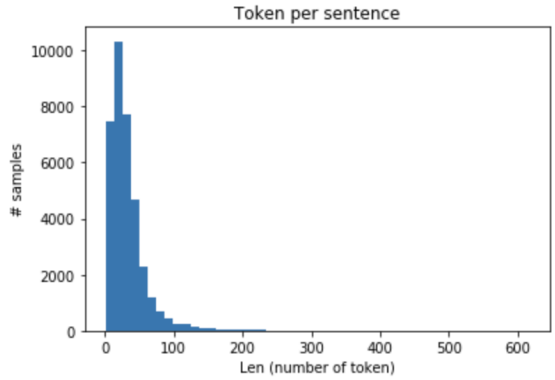
Number of sentences: 835
Number of words: 11525
Number of unique words: 3862
Number of tags: 21
Different tags: {'B-DISO', 'O', 'I-ANAT', 'B-PHEN', 'I-PROC', 'B-CHEM', 'I-GEOG', 'I-DEVI', 'B-GEOG', 'I-LIVB', 'I-PHYS', 'B-PROC', 'I-OBJC', 'B-LIVB', 'B-PHYS', 'B-DEVI', 'B-ANAT', 'I-PHEN', 'B-OBJC', 'I-DISO', 'I-CHEM'}



test acc: 72.69%, precision: 27.78%, recall: 23.28%, F1: 25.33%

w2v embeddings trained on full press corpus + ner trained on fra4

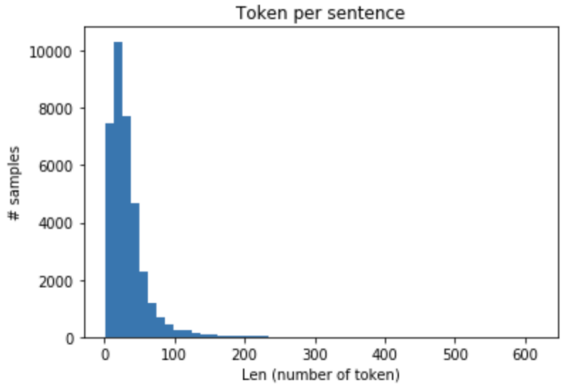
Number of sentences: 35723
Number of words: 1156339
Number of unique words: 37863
Number of tags: 11
Different tags: {'i-func', 'b-org', 'b-pers', 'O', 'b-loc', 'b-func', 'i-loc', 'i-prod', 'i-pers', 'b-prod', 'i-org'}



test acc: 97.50%, precision: 73.81%, recall: 66.02%, F1: 69.70%

fasttext embeddings trained on small press corpus + ner trained on fra4

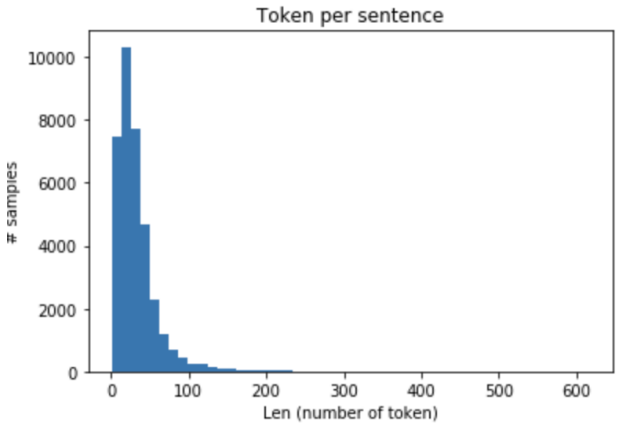
Number of sentences: 35723
Number of words: 1156339
Number of unique words: 37863
Number of tags: 11
Different tags: {'i-func', 'b-org', 'b-pers', 'O', 'b-loc', 'b-func', 'i-loc', 'i-prod', 'i-pers', 'b-prod', 'i-org'}



test acc: 97.43%, precision: 72.46%, recall: 66.62%, F1: 69.41%

fasttext embeddings trained on full press corpus + ner trained on fra4

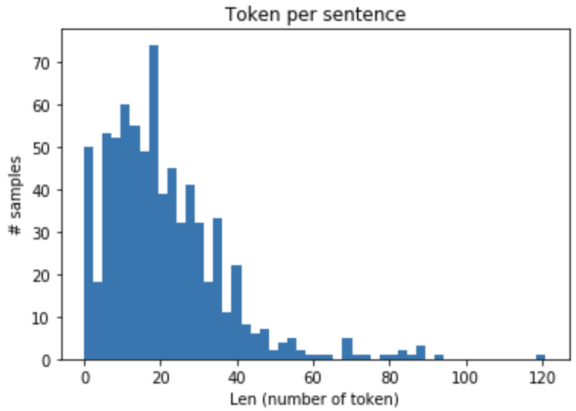
Number of sentences: 35723
Number of words: 1156339
Number of unique words: 37863
Number of tags: 11
Different tags: {'i-func', 'b-org', 'b-pers', 'O', 'b-loc', 'b-func', 'i-loc', 'i-prod', 'i-pers', 'b-prod', 'i-org'}



test acc: 97.50%, precision: 73.43%, recall: 66.14%, F1: 69.59%

fasttext embeddings trained on med corpus + ner trained on emea

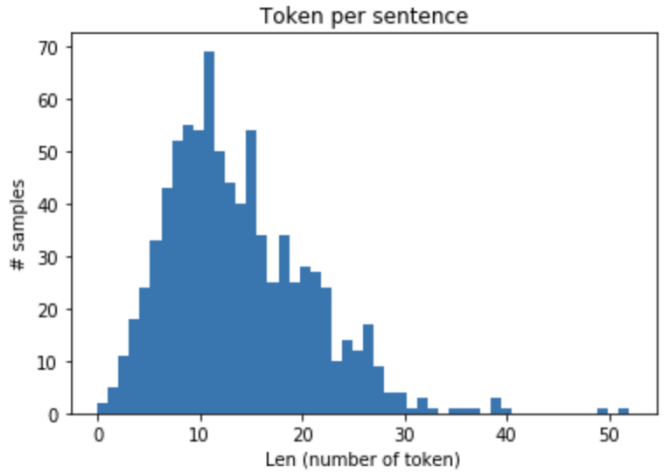
Number of sentences: 738
Number of words: 15339
Number of unique words: 2599
Number of tags: 21
Different tags: {'B-DISO', 'O', 'I-ANAT', 'B-PHEN', 'I-PROC', 'B-CHEM', 'I-GEOG', 'I-DEVI', 'B-GEOG', 'I-LIVB', 'I-PHYS', 'B-PROC', 'I-OBJC', 'B-LIVB', 'B-PHYS', 'B-DEVI', 'B-ANAT', 'I-PHEN', 'B-OBJC', 'I-DISO', 'I-CHEM'}



test acc: 84.95%, precision: 61.07%, recall: 30.57%, F1: 40.75%

fasttext embeddings trained on med corpus + ner trained on medline

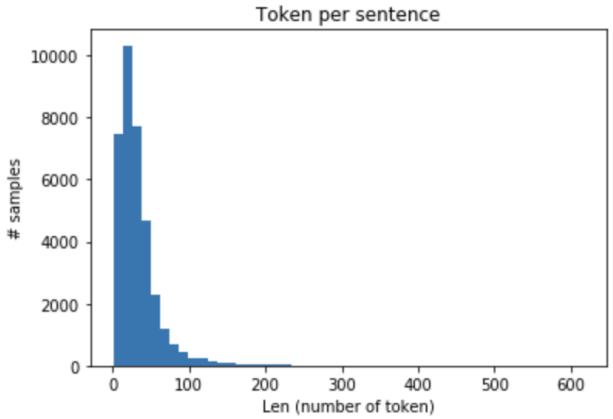
Number of sentences: 835
Number of words: 11525
Number of unique words: 3862
Number of tags: 21
Different tags: {'B-DISO', 'O', 'I-ANAT', 'B-PHEN', 'I-PROC', 'B-CHEM', 'I-GEOG', 'I-DEVI', 'B-GEOG', 'I-LIVB', 'I-PHYS', 'B-PROC', 'I-OBJC', 'B-LIVB', 'B-PHYS', 'B-DEVI', 'B-ANAT', 'I-PHEN', 'B-OBJC', 'I-DISO', 'I-CHEM'}



test acc: 75.77%, precision: 42.71%, recall: 26.02%, F1: 32.34%

fasttext embeddings trained on med corpus + ner trained on fra4 press

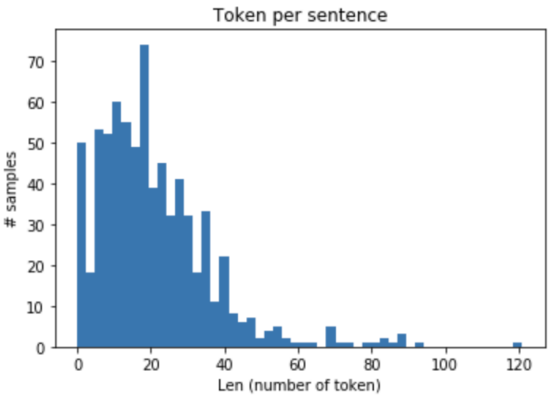
Number of sentences: 35723
Number of words: 1156339
Number of unique words: 37863
Number of tags: 11
Different tags: {'i-func', 'b-org', 'b-pers', 'O', 'b-loc', 'b-func', 'i-loc', 'i-prod', 'i-pers', 'b-prod', 'i-org'}



test acc: 97.33%, precision: 73.10%, recall: 62.88%, F1: 67.60%

fasttext embeddings trained on small press corpus + ner trained on emea

Number of sentences: 738
Number of words: 15339
Number of unique words: 2599
Number of tags: 21
Different tags: {'B-DISO', 'O', 'I-ANAT', 'B-PHEN', 'I-PROC', 'B-CHEM', 'I-GEOG', 'I-DEVI', 'B-GEOG', 'I-LIVB', 'I-PHYS', 'B-PROC', 'I-OBJC', 'B-LIVB', 'B-PHYS', 'B-DEVI', 'B-ANAT', 'I-PHEN', 'B-OBJC', 'I-DISO', 'I-CHEM'}

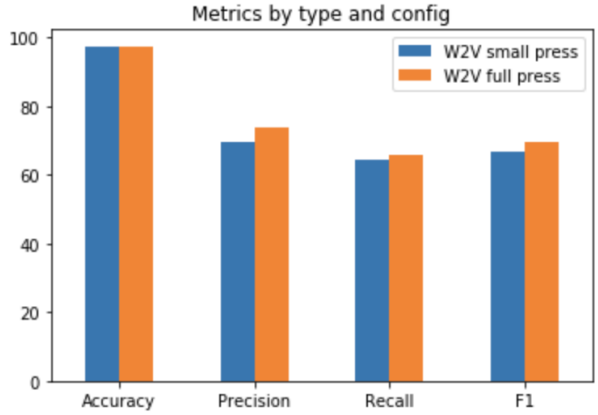


test acc: 84.19%, precision: 54.85%, recall: 31.34%, F1: 39.89%

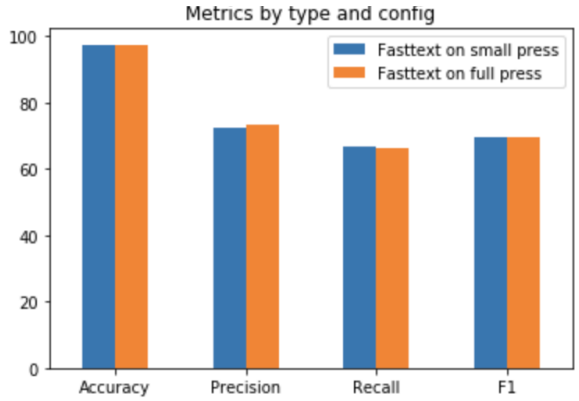
Performance comparison

The size of the corpus used to train word embedding

Comparing w2v trained on small press vs w2v trained on full press



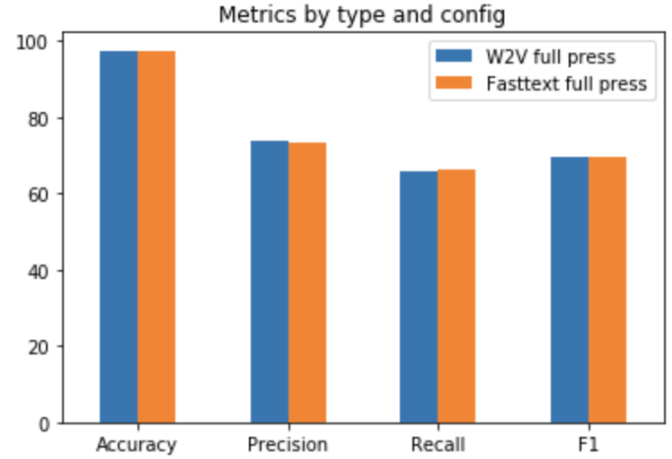
Comparing Fasttext trained on small press vs Fasttext trained on full press



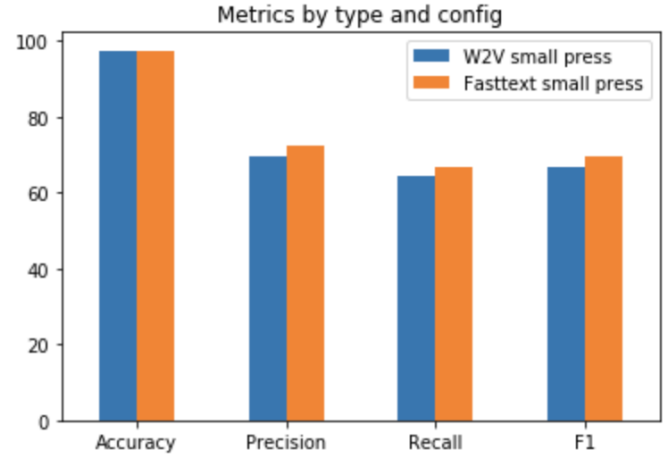
Mini-conclusion: The two previous graphs of this section show that embeddings trained on a bigger corpus improve the ner model performance.

The word embedding model

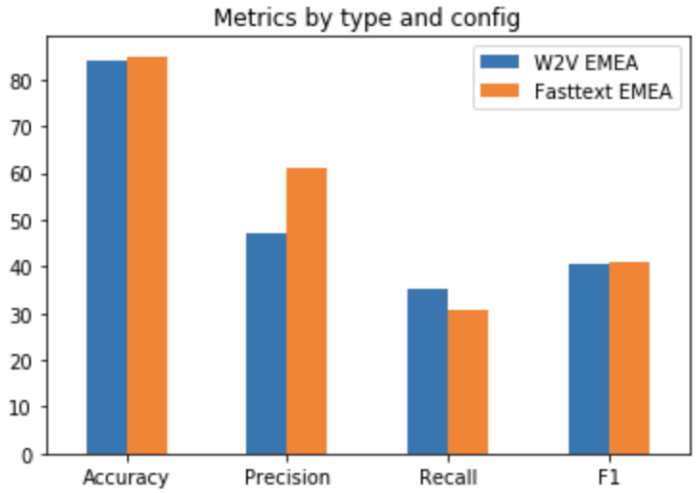
Comparing w2v trained on full press vs fasttext trained on full press



Comparing w2v trained on small press vs fasttext trained on small press



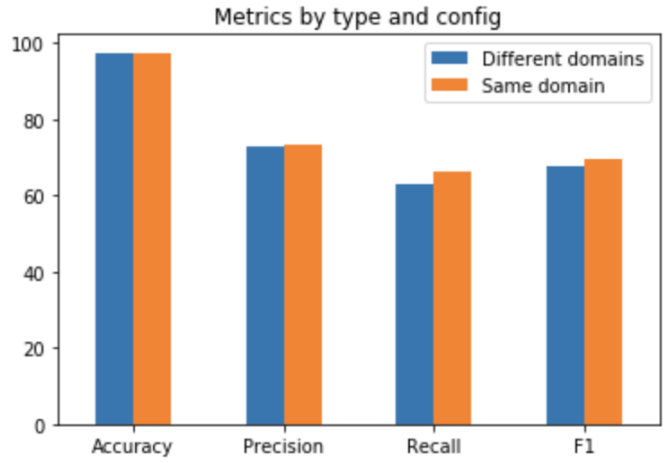
Comparing w2v trained on medical corpus vs fasttext trained on medical corpus and ner on emea



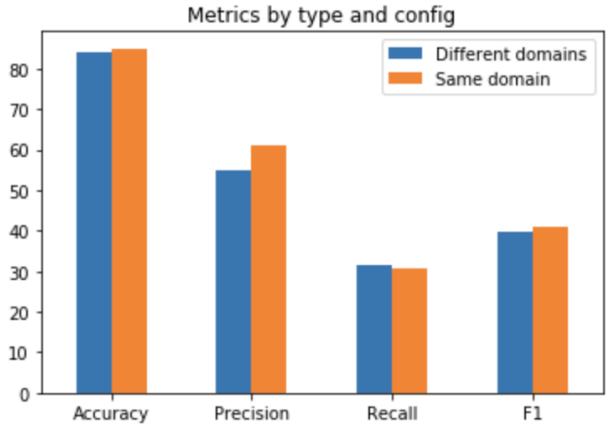
Mini-conclusion: Fasttext seems to perform better than word2vec overall.

The fit of the domain of the corpus used for named entity recognition and that of the word embeddings.

Comparing BETWEEN ner trained on press corpus with fasttext embeddings trained on medical corpus AND ner trained on press with fasttext embeddings trained on press



Comparing BETWEEN ner trained on medical emea corpus with fasttext embeddings trained on small press corpus AND ner trained on medical emea corpus with fasttext embeddings trained on medical corpus



Mini_conclusion: As expected, having the word embeddings and the ner model trained on the same domain achieves higher performance than having them trained on different corpora.