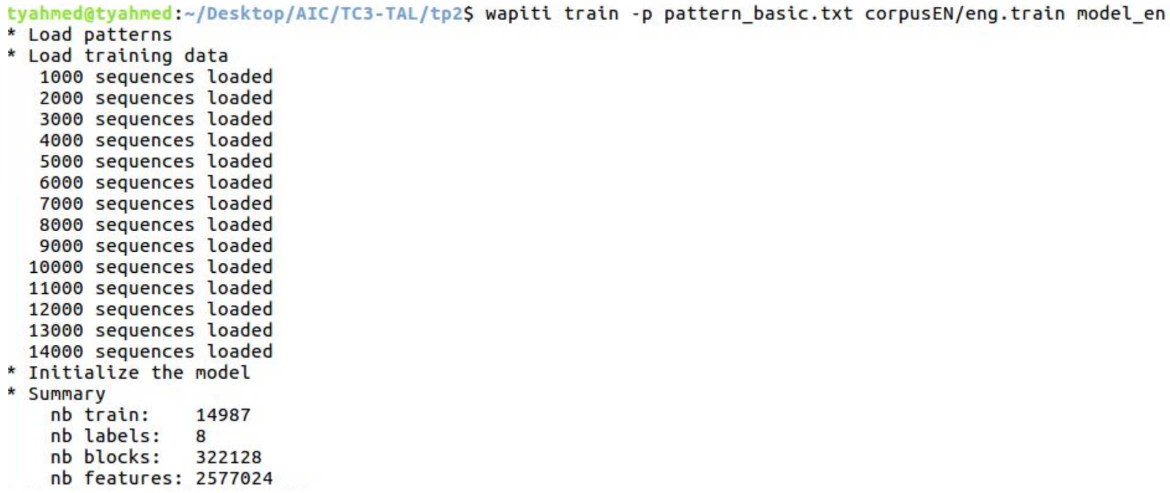
Named entity recognition with wapiti

Author : Taycir YAHMED

This document holds named entity recognition experiments with the tool wapiti.

Simple training with a basic pattern

1. Training the model

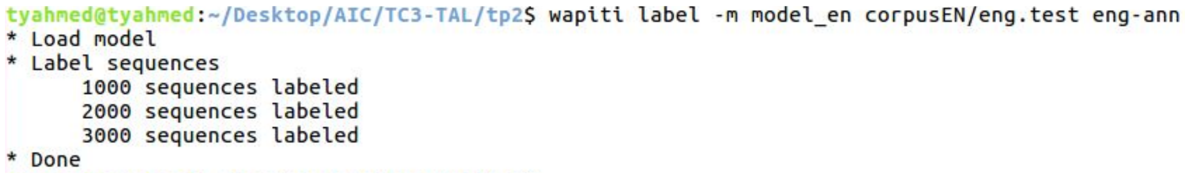


```
tyahmed@tyahmed:~/Desktop/AIC/TC3-TAL/tp2$ wapiti train -p pattern_basic.txt corpusEN/eng.train model_en
* Load patterns
* Load training data
   1000 sequences loaded
   2000 sequences loaded
   3000 sequences loaded
   4000 sequences loaded
   5000 sequences loaded
   6000 sequences loaded
   7000 sequences loaded
   8000 sequences loaded
   9000 sequences loaded
  10000 sequences loaded
  11000 sequences loaded
  12000 sequences loaded
  13000 sequences loaded
  14000 sequences loaded
* Initialize the model
* Summary
    nb train:    14987
    nb labels:   8
    nb blocks:   322128
    nb features: 2577024
```

The used pattern file contains the following content:

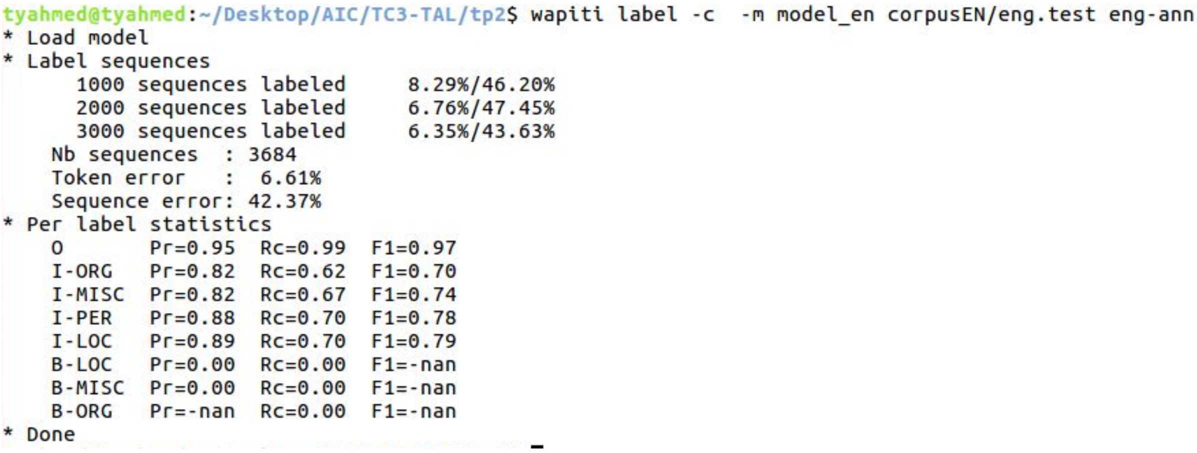
```
# Unigram
u1:%x[-2,0]
u2:%x[-1,0]
u3:%x[ 0,0]
u4:%x[ 1,0]
u5:%x[ 2,0]

# Bigram
u6:%x[-1,0]/%x[0,0]
u7:%x[ 1,0]/%x[0,0]
```

2. Labelling the test set



```
tyahmed@tyahmed:~/Desktop/AIC/TC3-TAL/tp2$ wapiti label -m model_en corpusEN/eng.test eng-ann
* Load model
* Label sequences
      1000 sequences labeled
      2000 sequences labeled
      3000 sequences labeled
* Done
```

3. Testing the model performance



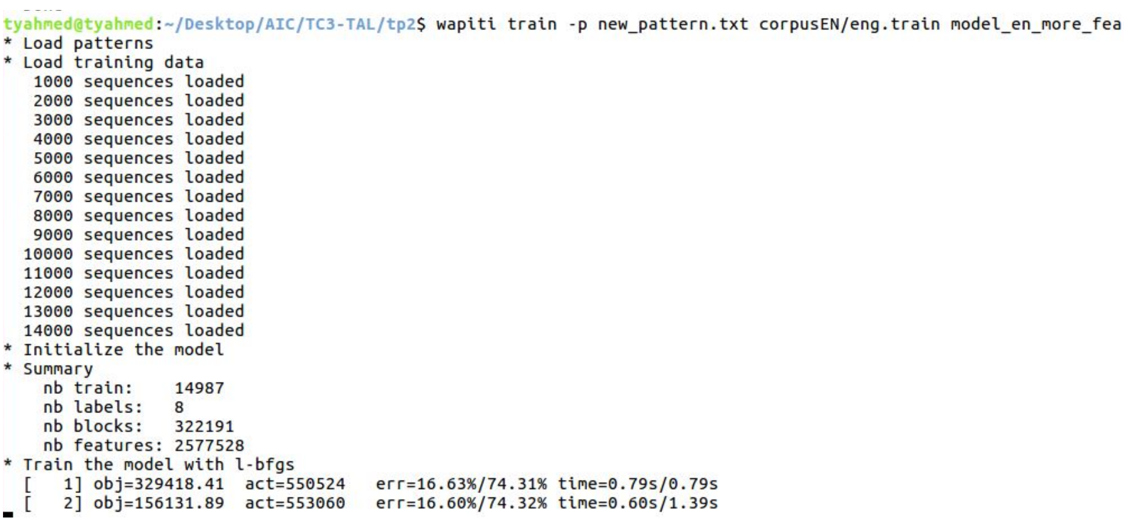
```
tyahmed@tyahmed:~/Desktop/AIC/TC3-TAL/tp2$ wapiti label -c  -m model_en corpusEN/eng.test eng-ann
* Load model
* Label sequences
      1000 sequences labeled     8.29%/46.20%
      2000 sequences labeled     6.76%/47.45%
      3000 sequences labeled     6.35%/43.63%
    Nb sequences  : 3684
    Token error   :  6.61%
    Sequence error: 42.37%
* Per label statistics
    O       Pr=0.95  Rc=0.99  F1=0.97
    I-ORG   Pr=0.82  Rc=0.62  F1=0.70
    I-MISC  Pr=0.82  Rc=0.67  F1=0.74
    I-PER   Pr=0.88  Rc=0.70  F1=0.78
    I-LOC   Pr=0.89  Rc=0.70  F1=0.79
    B-LOC   Pr=0.00  Rc=0.00  F1=-nan
    B-MISC  Pr=0.00  Rc=0.00  F1=-nan
    B-ORG   Pr=-nan  Rc=0.00  F1=-nan
* Done
```

Optimizing performance by adding more features

1. Adding features: here for instance I added the Pos tagging column and the chunk column. Hence, the pattern file contains the following:

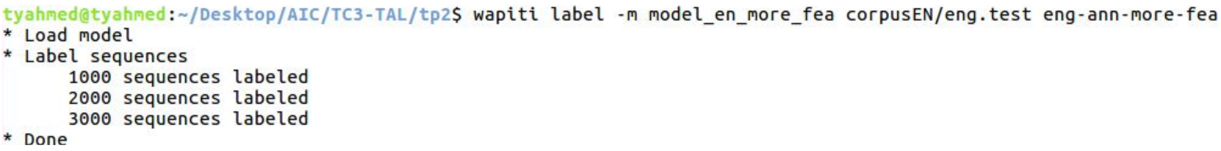
```
# Unigram
u1:%x[-2,0]
u2:%x[-1,0]
u3:%x[ 0,0]
u4:%x[ 1,0]
u5:%x[ 2,0]
u6:%x[ 0,1]
u7:%x[ 0,2]

# Bigram
u6:%x[-1,0]/%x[0,0]
u7:%x[ 1,0]/%x[0,0]
```

2. Training the model



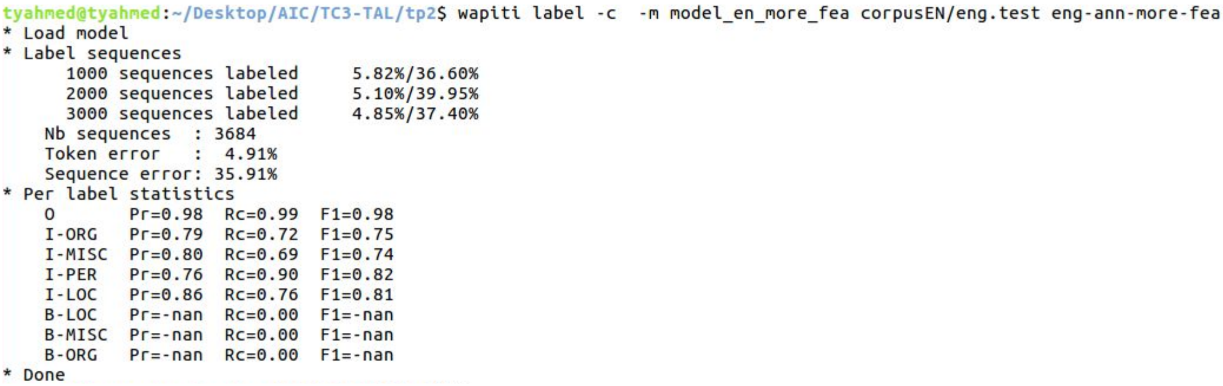
```
tyahmed@tyahmed:~/Desktop/AIC/TC3-TAL/tp2$ wapiti train -p new_pattern.txt corpusEN/eng.train model_en_more_fea
* Load patterns
* Load training data
   1000 sequences loaded
   2000 sequences loaded
   3000 sequences loaded
   4000 sequences loaded
   5000 sequences loaded
   6000 sequences loaded
   7000 sequences loaded
   8000 sequences loaded
   9000 sequences loaded
  10000 sequences loaded
  11000 sequences loaded
  12000 sequences loaded
  13000 sequences loaded
  14000 sequences loaded
* Initialize the model
* Summary
    nb train:    14987
    nb labels:   8
    nb blocks:   322191
    nb features: 2577528
* Train the model with l-bfgs
  [   1] obj=329418.41  act=550524   err=16.63%/74.31% time=0.79s/0.79s
  [   2] obj=156131.89  act=553060   err=16.60%/74.32% time=0.60s/1.39s
```

3. Labelling the test set



```
tyahmed@tyahmed:~/Desktop/AIC/TC3-TAL/tp2$ wapiti label -m model_en_more_fea corpusEN/eng.test eng-ann-more-fea
* Load model
* Label sequences
      1000 sequences labeled
      2000 sequences labeled
      3000 sequences labeled
* Done
```

4. Testing the model

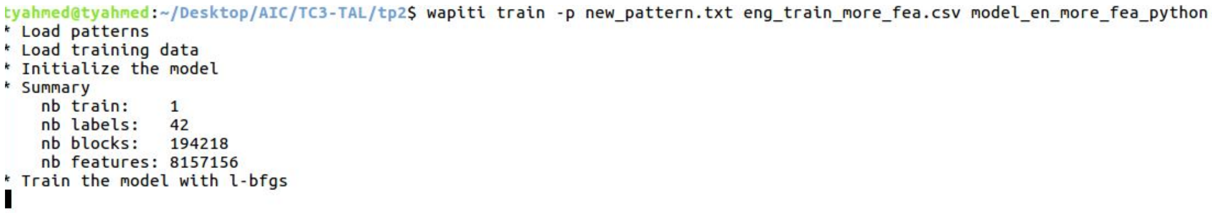


```
tyahmed@tyahmed:~/Desktop/AIC/TC3-TAL/tp2$ wapiti label -c  -m model_en_more_fea corpusEN/eng.test eng-ann-more-fea
* Load model
* Label sequences
      1000 sequences labeled     5.82%/36.60%
      2000 sequences labeled     5.10%/39.95%
      3000 sequences labeled     4.85%/37.40%
    Nb sequences  : 3684
    Token error   :  4.91%
    Sequence error: 35.91%
* Per label statistics
    O       Pr=0.98  Rc=0.99  F1=0.98
    I-ORG   Pr=0.79  Rc=0.72  F1=0.75
    I-MISC  Pr=0.80  Rc=0.69  F1=0.74
    I-PER   Pr=0.76  Rc=0.90  F1=0.82
    I-LOC   Pr=0.86  Rc=0.76  F1=0.81
    B-LOC   Pr=-nan  Rc=0.00  F1=-nan
    B-MISC  Pr=-nan  Rc=0.00  F1=-nan
    B-ORG   Pr=-nan  Rc=0.00  F1=-nan
* Done
```

Adding more features to the corpus using python

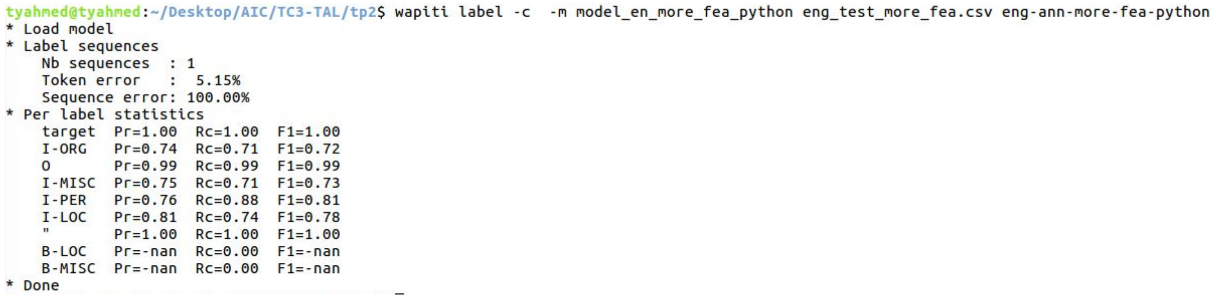
The code to perform this task is joined to this report.

1. Training with the new features



```
tyahmed@tyahmed:~/Desktop/AIC/TC3-TAL/tp2$ wapiti train -p new_pattern.txt eng_train_more_fea.csv model_en_more_fea_python
* Load patterns
* Load training data
* Initialize the model
* Summary
    nb train:    1
    nb labels:   42
    nb blocks:   194218
    nb features: 8157156
* Train the model with l-bfgs
```

2. Testing the model



```
tyahmed@tyahmed:~/Desktop/AIC/TC3-TAL/tp2$ wapiti label -c  -m model_en_more_fea_python eng_test_more_fea.csv eng-ann-more-fea-python
* Load model
* Label sequences
    Nb sequences  : 1
    Token error   :  5.15%
    Sequence error: 100.00%
* Per label statistics
    target  Pr=1.00  Rc=1.00  F1=1.00
    I-ORG   Pr=0.74  Rc=0.71  F1=0.72
    O       Pr=0.99  Rc=0.99  F1=0.99
    I-MISC  Pr=0.75  Rc=0.71  F1=0.73
    I-PER   Pr=0.76  Rc=0.88  F1=0.81
    I-LOC   Pr=0.81  Rc=0.74  F1=0.78
    "       Pr=1.00  Rc=1.00  F1=1.00
    B-LOC   Pr=-nan  Rc=0.00  F1=-nan
    B-MISC  Pr=-nan  Rc=0.00  F1=-nan
* Done
```