

Task 5. CRF implementation and Pat template test of Wapiti sequence annotation

github link: https://github.com/Allen-ZKW/NLP_HZAU/tree/Task5

Author: Kewei Zhao

Date: 2021-4-23

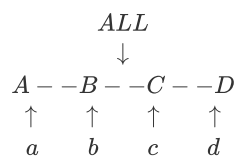
Abstract

In this task, we try to use Wapiti to do sequence annotation for AGAC corpus just like what is taught in class. Also, we will change the pat file in order to see how the difference of characteristic function will influence the sequence annotation result.

Principle

$$HMM \rightarrow MEMM \rightarrow CRF$$

CRF depends on and develops from MEMM and HEMM. In this developing process, algorithm mainly overcomes two hypothesis: Observation independence hypothesis and Markov hypothesis.



This graph can reflect the structure of linear-CRF in some degree. First, in this algorithm we try to calculate conditional probability not joint probability. Second, the relationship between different status is undirected, which means information or observation in the future can influence status in the past. These two characteristics correspond with two hypothesis which is overcome by CRF.

Measure

1) Configuration Environment(Ubuntu 20.04 LTS)

```
sudo apt update
sudo apt install build-essential
sudo apt install make
sudo apt install gcc
```

2) Install Wapiti and Run Demo(Ubuntu 20.04 LTS)

```

unzip wapiti-master.zip
cd wapiti-master
sudo make
sudo make install
wapiti train -a sgd-l1 -t 3 -i 10 -p pat/Tok321dis.pat <(cat
AGAC/train_split/*.txt) AGAC/mod/AGAC_train.mod
wapiti label -c -m AGAC/mod/AGAC_train.mod <(cat AGAC/test_split/*.txt)
AGAC/train_out.tab
perl conllevel.pl -d $'\t' <AGAC/train_out.tab | tee AGAC/train_out.eval

```

3) Change .pat File to Understand How Pattern Influence Sequence Annotation

In the demo pattern, they build relationship between one word and itself , two adjacent words and three adjacent words. Because we only learn some knowledge about liner-CRF model which only have relationship between one word and itself , two adjacent words, so we add '#' in the head of three adjacent words part.(We retain patterns which match the head and the tail of the sequence)

```

*

U:tok:1:-1:%X[-1,0]
U:tok:1:+0:%X[0,0]
U:tok:1:+1:%X[1,0]

U:tok:2:-1:%X[-1,0]/%X[0,0]
U:tok:2:+0:%X[0,0]/%X[1,0]

#U:tok:3:-2:%X[-2,0]/%X[-1,0]/%X[0,0]
#U:tok:3:-1:%X[-1,0]/%X[0,0]/%X[1,0]
#U:tok:3:+0:%X[0,0]/%X[1,0]/%X[2,0]

U:pre:1:+0:4:%M[0,0,"^.??.?.$"]

U:suf:1:+0:4:%M[0,0,".???.?.$"]

U:dis:1:-1:%X[-1,2]
U:dis:1:+0:%X[0,2]
U:dis:1:+1:%X[1,2]

```

4) Run Wapiti Again

```

wapiti train -a sgd-l1 -t 3 -i 10 -p pat/Tok321dis.pat <(cat
AGAC/train_split/*.txt) AGAC/mod/AGAC_train.mod
wapiti label -c -m AGAC/mod/AGAC_train.mod <(cat AGAC/test_split/*.txt)
AGAC/train_out.tab
perl conllevel.pl -d $'\t' <AGAC/train_out.tab | tee AGAC/train_out.eval

```

Result

1) Train Result of Demo Pattern

```
* Load patterns
* Load training data
  1000 sequences loaded
  2000 sequences loaded
* Initialize the model
* Summary
  nb train:    2530
  nb labels:   25
  nb blocks:   270490
  nb features: 6762875
* Train the model with sgd-11
  - Build the index
    Done
[  1] obj=NA act=133927 err= 9.26%/29.57% time=1.12s/1.12s
[  2] obj=NA act=143130 err= 8.47%/47.31% time=0.94s/2.06s
[  3] obj=NA act=84472  err= 9.48%/27.27% time=0.94s/3.00s
[  4] obj=NA act=80772  err= 6.44%/32.37% time=0.95s/3.94s
[  5] obj=NA act=73716  err= 6.67%/52.41% time=0.92s/4.87s
[  6] obj=NA act=69693  err= 4.34%/34.15% time=0.97s/5.84s
[  7] obj=NA act=67335  err= 5.52%/27.71% time=0.94s/6.78s
[  8] obj=NA act=65163  err= 3.05%/25.89% time=0.93s/7.71s
[  9] obj=NA act=62789  err= 4.13%/34.47% time=0.93s/8.64s
[ 10] obj=NA act=60795  err= 4.33%/34.70% time=0.93s/9.57s
* Save the model
* Done
```

2) Label Result of Demo Pattern

```
* Load model
* Label sequences
  1000 sequences labeled    10.31%/45.90%
  2000 sequences labeled    9.68%/45.10%
  Nb sequences   : 2506
  Token error    : 9.48%
  Sequence error: 44.89%
* Per label statistics
  O      Pr=0.94 Rc=0.96 F1=0.95
  B-Var   Pr=0.35 Rc=0.39 F1=0.37
  B-Gene  Pr=0.33 Rc=0.06 F1=0.10
  B-Enzyme Pr=0.00 Rc=0.00 F1=-nan
  B-PosReg Pr=0.30 Rc=0.39 F1=0.34
  B-MPA   Pr=0.19 Rc=0.12 F1=0.14
  I-MPA   Pr=0.15 Rc=0.08 F1=0.11
  B-Disease Pr=0.24 Rc=0.25 F1=0.25
  I-Disease Pr=0.32 Rc=0.34 F1=0.33
  I-Var   Pr=0.23 Rc=0.47 F1=0.31
  B-Interaction Pr=0.12 Rc=0.14 F1=0.13
  B-Protein Pr=0.50 Rc=0.02 F1=0.03
  B-Pathway Pr=0.00 Rc=0.00 F1=-nan
  I-Pathway Pr=0.67 Rc=0.05 F1=0.10
  B-NegReg Pr=0.57 Rc=0.39 F1=0.46
  B-Reg    Pr=0.66 Rc=0.10 F1=0.17
  I-PosReg Pr=0.46 Rc=0.92 F1=0.61
  B-CPA    Pr=0.12 Rc=0.14 F1=0.13
```

```

I-CPA    Pr=0.12 Rc=0.08 F1=0.09
I-NegReg Pr=0.76 Rc=0.83 F1=0.79
I-Gene   Pr=0.35 Rc=0.17 F1=0.23
I-Reg    Pr=0.79 Rc=0.17 F1=0.28
I-Protein Pr=-nan Rc=0.00 F1=-nan
I-Interaction Pr=0.00 Rc=0.00 F1=-nan
I-Enzyme Pr=0.00 Rc=0.00 F1=-nan

```

* Done

3) Analyze Wapiti Result of Demo Pattern

```

processed 62559 tokens with 2416 phrases; found: 1617 phrases; correct: 466.
accuracy: 90.52%; precision: 28.82%; recall: 19.29%; FB1: 23.11
    CPA: precision: 11.94%; recall: 14.29%; FB1: 13.01 67
    Disease: precision: 20.96%; recall: 22.01%; FB1: 21.47 439
    Enzyme: precision: 0.00%; recall: 0.00%; FB1: 0.00 4
    Gene: precision: 28.05%; recall: 4.94%; FB1: 8.39 82
    Interaction: precision: 12.50%; recall: 14.29%; FB1: 13.33 8
    MPA: precision: 14.52%; recall: 8.96%; FB1: 11.08 124
    NegReg: precision: 56.82%; recall: 39.06%; FB1: 46.30 88
    Pathway: precision: 0.00%; recall: 0.00%; FB1: 0.00 2
    PosReg: precision: 30.48%; recall: 39.02%; FB1: 34.22 105
    Protein: precision: 50.00%; recall: 1.67%; FB1: 3.23 2
    Reg: precision: 62.30%; recall: 9.55%; FB1: 16.56 61
    Var: precision: 31.97%; recall: 35.80%; FB1: 33.78 635

```

4) Train Result of Changed Pattern

```

* Load patterns
* Load training data
    1000 sequences loaded
    2000 sequences loaded
* Initialize the model
* Summary
    nb train: 2530
    nb labels: 25
    nb blocks: 104635
    nb features: 2616500
* Train the model with sgd-l1
    - Build the index
    Done
[ 1] obj=NA act=72180 err= 9.51%/32.21% time=0.99s/0.99s
[ 2] obj=NA act=75331 err= 9.19%/32.37% time=0.93s/1.92s
[ 3] obj=NA act=48057 err= 9.21%/27.59% time=0.95s/2.87s
[ 4] obj=NA act=46324 err= 8.26%/30.87% time=0.91s/3.78s
[ 5] obj=NA act=42695 err= 9.19%/61.98% time=0.92s/4.70s
[ 6] obj=NA act=40814 err= 8.26%/29.01% time=0.91s/5.61s
[ 7] obj=NA act=39386 err= 6.51%/46.09% time=0.92s/6.53s
[ 8] obj=NA act=38274 err= 5.82%/40.40% time=0.90s/7.43s
[ 9] obj=NA act=37113 err= 5.05%/35.97% time=0.95s/8.39s
[10] obj=NA act=36397 err= 5.16%/34.74% time=0.92s/9.31s
* Save the model
* Done

```

5) Label Result of Changed Pattern

```
* Load model
* Label sequences
    1000 sequences labeled      8.62%/36.60%
    2000 sequences labeled      8.33%/35.25%
Nb sequences   : 2506
Token error    :  8.15%
Sequence error: 34.84%
* Per label statistics
O              Pr=0.93 Rc=0.98 F1=0.96
B-Var          Pr=0.41 Rc=0.23 F1=0.29
B-Gene         Pr=0.62 Rc=0.04 F1=0.08
B-Enzyme       Pr=-nan Rc=0.00 F1=-nan
B-PosReg       Pr=0.31 Rc=0.28 F1=0.29
B-MPA          Pr=0.13 Rc=0.04 F1=0.07
I-MPA          Pr=0.11 Rc=0.03 F1=0.05
B-Disease      Pr=0.25 Rc=0.13 F1=0.17
I-Disease      Pr=0.37 Rc=0.16 F1=0.22
I-Var          Pr=0.25 Rc=0.20 F1=0.22
B-Interaction  Pr=0.20 Rc=0.14 F1=0.17
B-Protein      Pr=1.00 Rc=0.02 F1=0.03
B-Pathway      Pr=-nan Rc=0.00 F1=-nan
I-Pathway      Pr=-nan Rc=0.00 F1=-nan
B-NegReg       Pr=0.62 Rc=0.08 F1=0.14
B-Reg          Pr=0.49 Rc=0.06 F1=0.11
I-PosReg       Pr=0.56 Rc=0.83 F1=0.67
B-CPA          Pr=0.12 Rc=0.07 F1=0.09
I-CPA          Pr=0.06 Rc=0.01 F1=0.02
I-NegReg       Pr=1.00 Rc=0.04 F1=0.08
I-Gene         Pr=0.78 Rc=0.12 F1=0.20
I-Reg          Pr=0.69 Rc=0.06 F1=0.11
I-Protein      Pr=-nan Rc=0.00 F1=-nan
I-Interaction  Pr=-nan Rc=0.00 F1=-nan
I-Enzyme       Pr=-nan Rc=0.00 F1=-nan
* Done
```

6) Analyze Wapiti Result of Changed Pattern

```
processed 62559 tokens with 2416 phrases; found: 818 phrases; correct: 250.
accuracy:  91.85%; precision:  30.56%; recall:  10.35%; FB1:  15.46
    CPA: precision:  12.12%; recall:   7.14%; FB1:   8.99  33
    Disease: precision:  20.56%; recall:  10.53%; FB1:  13.92 214
    Enzyme: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
    Gene: precision:  59.38%; recall:   4.08%; FB1:   7.63  32
    Interaction: precision:  20.00%; recall:  14.29%; FB1:  16.67  5
    MPA: precision:  10.81%; recall:   3.98%; FB1:   5.82  74
    NegReg: precision:  62.50%; recall:   7.81%; FB1:  13.89 16
    Pathway: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
    PosReg: precision:  29.73%; recall:  26.83%; FB1:  28.21  74
    Protein: precision: 100.00%; recall:   1.67%; FB1:   3.28  1
    Reg: precision:  42.86%; recall:   5.28%; FB1:   9.40  49
    Var: precision:  37.50%; recall:  21.16%; FB1:  27.06 320
```

Discussion

The first problem we meet in this task is International Workers' Day. We can not use classroom on vacation, so we have to install sub-system on Windows to complete this task. Because of less experience in this field, it takes nearly a whole day to configurate environment on my PC.

We analyze the difference of Wapiti result of different patterns. Firstly, CRF calculate the integrals of probability by calculate sum of ϕ for every maximal clique, so the change of pattern will change the linkage of nodes in CRF model. Because of this, the maximal clique and the characteristic function will be different.

It is obviously that the demo pattern has more information than the changed pattern. This reflect in the result, for example, because demo pattern concern relationship of three adjacent words, demo pattern will produce more characteristic function to describe the data we load in and the result is also better than liner-CRF model. However, better result also means more time consuming and more calculation consuming. We also try to add relationship for four adjacent words in pattern file in order to get better result but Wapiti even cannot output the training result. We do not know which algorithm Wapiti choose to solve the training problem. But we can still guess what happened when we add relationship for four adjacent words:

A huge number of characteristic function → A huge number of parameter need to train → Time consuming

So, when using Wapiti it's important to keep the balance between accuracy and calculate amount. Further more, some questions still exist in .pat file, maybe it's fine to ask teacher in the next class.