netID: kh2383

Kexin Huang

HW6

Prof. Adam Meyers

Files:

response_WSj_23.chunk -> the output for WSJ_23

name_entity_test.ipynb -> program to find features for testing file (Jupyter Notebook)

name_entity_train.ipynb -> program to find features for training file (Jupyter Notebook)

Kexin_README_HW6 -> README

I ran 4 models. With each has different features. Result in the back.

These are the features:

POS tag

#p_p -> previous POS

#p_w -> previous word

#p_BIO -> previous BIO

#p2_p -> 2nd previous POS

#p2_w -> 2nd previous word

#p2_BIO -> 2nd previous BIO

#f_p -> forward POS

#f_w -> forward word

#f_BIO -> forward BIO

#f2_p -> 2nd forward POS

#f2_w -> 2nd forward word

#f2_BIO -> 2nd forward BIO

#fLC->firstLetterCapital

#w->word

#isN -> is a number

```
 60:  .. loglikelihood=-116465.39572620146      0.989063885665264
 61:  .. loglikelihood=-115499.32101985556      0.9891501998647393
 62:  .. loglikelihood=-114557.55880637318      0.9892438823495356
 63:  .. loglikelihood=-113639.10950708075      0.9893249334880673
 64:  .. loglikelihood=-112743.03000644328      0.9893996689534666
 65:  .. loglikelihood=-111868.4296307915       0.9894733518066773
 66:  .. loglikelihood=-111014.46646989015      0.9895386137623781
 67:  .. loglikelihood=-110180.34400734822      0.9895954548205691
 68:  .. loglikelihood=-109365.30802995569      0.9896459802056278
 69:  .. loglikelihood=-108568.64378915151      0.9897459783635565
 70:  .. loglikelihood=-107789.67339089955      0.9898091350948798
 71:  .. loglikelihood=-107027.75339252969      0.9898617657043159
 72:  .. loglikelihood=-106282.27258783905      0.9899196593746957
 73:  .. loglikelihood=-105552.64996308959      0.9899922896157176
 74:  .. loglikelihood=-104838.3328088901       0.9900417623885875
 75:  .. loglikelihood=-104138.79497406224      0.990093340385835
 76:  .. loglikelihood=-103453.53524917529      0.9901543918927809
 77:  .. loglikelihood=-102782.07586857928      0.9902143907875381
 78:  .. loglikelihood=-102123.96112085553      0.9902512322141435
 79:  .. loglikelihood=-101478.75605862054      0.9902964945382585
 80:  .. loglikelihood=-100846.04529934065      0.9903649143305255
 81:  .. loglikelihood=-100225.43190976152      0.9904059662058857
 82:  .. loglikelihood=-99616.536367126 0.9904533337543782
 83:  .. loglikelihood=-99018.9955910392        0.9905070169760031
 84:  .. loglikelihood=-98432.46204030867       0.9905480688513633
 85:  .. loglikelihood=-97856.60286977443       0.990587015502346
 86:  .. loglikelihood=-97291.0991423428        0.9906375408874047
 87:  .. loglikelihood=-96735.6450921116        0.9907017502309168
 88:  .. loglikelihood=-96189.9474345517        0.9907512230037868
 89:  .. loglikelihood=-95653.72472043434       0.9908017483888455
 90:  .. loglikelihood=-95126.70673005735       0.9908438528763944
 91:  .. loglikelihood=-94608.63390491532       0.990875431242056
 92:  .. loglikelihood=-94099.25681411548       0.9909185883417937
 93:  .. loglikelihood=-93598.33565298677       0.990974376787796
 94:  .. loglikelihood=-93105.63977166449       0.991019639111911
 95:  .. loglikelihood=-92620.94723152998       0.991073322333536
 96:  .. loglikelihood=-92144.0443875409        0.9911133215967074
 97:  .. loglikelihood=-91674.72549476508       0.9911470051867466
 98:  .. loglikelihood=-91212.79233740793       0.9911901622864842
 99:  .. loglikelihood=-90758.05387884215       0.9912227932643346
100:  .. loglikelihood=-90310.32593127602       0.9912912130566016
osamdeMacBook-Pro:MAX_ENT_FILES KexinHuang$ java -cp .:maxent-3.0.0.jar:trove.jar MEtag test.chunk model.chunk response.chunk
osamdeMacBook-Pro:MAX_ENT_FILES KexinHuang$ python score.chunk.py WSJ_24.pos-chunk response.chunk
31033 out of 32853 tags correct
  accuracy: 94.46
3378 groups in key
3072 groups in response
2424 correct groups
  precision: 81.83
  recall:    88.61
  F1:        85.09
osamdeMacBook-Pro:MAX_ENT_FILES KexinHuang$
```

Accuracy: 94.46 and F1:85.09

This model use least 5 features: 1. POS tag,2. previous word info 3. First Letter Capital

```
[172-17-83-132:MAX_ENT_FILES KexinHuang$ python score.chunk.py WSJ_24.pos-chunk r]
esponse.chunk
31106 out of 32853 tags correct
  accuracy: 94.68
8378 groups in key
8826 groups in response
7478 correct groups
  precision: 84.73
  recall:    89.26
  F1:        86.93
```

Accuracy: 94.68

F1: 86.93

This model use 8 features 1. POS tag 2. Previous word info (word, pos, bio) 3. First letter Capital 4. Forward word info (word, pos, bio)

```
172-17-83-132:MAX_ENT_FILES KexinHuang$ java -cp .:maxent-3.0.0.jar:trove.jar MEtag test_6.chunk model_6.chunk resp
onse_6.chunk
172-17-83-132:MAX_ENT_FILES KexinHuang$ python score.chunk.py WSJ_24.pos-chunk response_6.chunk
31495 out of 32853 tags correct                                                                                    ]
  accuracy: 95.87
8378 groups in key                                                                                                 ]
8926 groups in response
7609 correct groups
  precision: 85.25
  recall:    90.82
  F1:        87.94
```

Accuracy 95.87 F1:87.04

This model uses 14 features: 1. POS tag 2. Previous word info (word, pos, bio) 3. First letter Capital 4. Forward word info (word, pos, bio) 5. Previous second word info 6. Forward second word info

```
172-17-83-132:MAX_ENT_FILES KexinHuang$ python score.chunk.py WSJ_24.pos-chunk response_8.chunk
31273 out of 32853 tags correct
  accuracy: 95.19
8378 groups in key
8780 groups in response
7548 correct groups
  precision: 85.97
  recall:    90.09
  F1:        87.98
```

Accuracy: 95.19 F1: 87.98

This model uses 16 features: 1. POS tag 2. Previous word info (word, pos, bio) 3. First letter Capital 4. Forward word info (word, pos, bio) 5. Previous second word info 6. Forward second word info 7. Is a number 8. Word itself

From comparison of four models, we see the F1 measure is within the difference of 2%. And in the last three models, F1 and accuracy were not changed drastically. So, we can see there are some features are not relevant and maybe new features need to be introduced.