Need to improve custom NER tagging output, using HMM

1 post by 1 author

me (Yogesh Kulkarni change) Sep 7

I am building a NER tagger for a new tag "LEGAL".

I have got training data with IOB format, sample below:

```
Training sentence: [('kurian', 'O'), ...('the', 'O'),
('proceedings', 'B-LEGAL'), ('by', 'O'), ('notification', 'B-
LEGAL'), ('dated', 'O'), ... ('the', 'O'), ('land', 'B-
LEGAL'), ('acquisition', 'I-LEGAL'), ('act', 'B-LEGAL'), ...
('of', 'O'), ('acquisition', 'B-LEGAL'), ('is', 'O'),
('residential', 'O'), ('and', 'O'), ('commercial', 'B-LEGAL'),
('for', 'O'),...('period', 'B-LEGAL'), ('of', 'O'), ('three',
'O'), ('months', 'O'), ('from', 'O'), ('today', 'O')]
```

Although not shown fully here, above, I have many IOB marked sub-sequences. And there are about 400 such texts used for training.

Code I am using is:

```
from nltk.tag import hmm

print("Training sentence: {}".format(train_data[0]))

trainer = hmm.HiddenMarkovModelTrainer()
tagger = trainer.train_supervised(train_data)
print(tagger)

for tst in test_x[:20]:
 test_sentence = " ".join(tst)
 print("Test sentence: {}".format(test_sentence))
 result = tagger.tag(test_sentence.split())
 print("Tagged sentence: {}".format(result))
 catchphrases = [ w for w,t in result if "LEGAL" in t]
 print("Catchphrases: {}".format(catchphrases))
```