NLP Assignment 5 : HMM POS tagger

Intent

As a part of this assignment , I implemented the following tasks

1. The required task to implemented , the viterbi algorithm and use laplace smoothing for the transition and the emission probabilities .
2. Changed the amount of training data and looked at the change in accuracy
3. Reduced the number of POS tags and measured accuracy
4. Sorted the test sentence based on their confidence and interpreted trends

HMM POS Tagger

As a part of this task , I first computed the transition and the emission probabilities from the word counts. I used laplace smoothing to obtain the probabilities . I experimented with different smoothing parameters and obtained interesting results

After obtaining the transition and the emission probabilities , I implemented the viterbi algorithm and predicted the POS tags for the different sentences using the “wsj15-18.pos” file.

Accuracy estimation :

In order to measure the accuracy of the POS tagger , I counted the number of correctly predicted tags and the total number of tags across all the test sentences and used these values to find the accuracy . I also estimated the precision , recall and the fscore for each tag as a part of the accuracy estimation

Experiment 1 :

For the test case where 200 sentences were used for testing and the rest of the sentences from the wsf15-18.pos file was used for training , the results were as follows

Results :

```
NNPS 23.08   75.0   35.29
JJS 29.63   100.0   45.71
JJR 60.98   92.59   73.53
Accuracy:89.237
```

For this case , we got an accuracy of around 90% and we also get to see the F1 score of the tags . On the analysis of the results , we see most of the tags have a Fscore of more than 80 . However a few tags like NNPS , PDT have an Fscore below 50 . In both the cases , we see that the precision is too low. The predeterminers like what,such,rather,quite can occur as other tags as well and hence there is a chance that from the training data they might have been tagged as other more probable types

	Precision	Recall	FScore
PRP$	67.24	95.12	78.79
VBG	92.45	77.78	84.48
VBD	97.32	88.96	92.95
VB	81.21	88.32	84.62
POS	91.43	91.43	91.43
''	70.59	100.0	82.76
VBP	89.86	93.94	91.85
WDT	70.59	85.71	77.42
JJ	94.89	79.64	86.6
WP	100.0	100.0	100.0
VBZ	86.3	84.0	85.14
DT	99.19	97.87	98.53
#	50.0	87.5	63.64
RP	8.33	100.0	15.38
$	89.47	100.0	94.44
NN	98.6	88.52	93.29
)	60.0	100.0	75.0
(	85.71	100.0	92.31
,	98.51	100.0	99.25
.	99.01	100.0	99.5
TO	99.21	99.21	99.21
PRP	89.16	98.67	93.67
RB	90.18	82.11	85.96
:	90.91	90.91	90.91
NNS	96.69	84.84	90.38
NNP	95.71	75.62	84.49
``	75.0	100.0	85.71
WRB	70.0	100.0	82.35
CC	100.0	99.07	99.53
PDT	16.67	100.0	28.57
RBS	40.0	100.0	57.14
RBR	44.44	66.67	53.33
VBN	90.0	75.9	82.35
EX	26.09	100.0	41.38
IN	97.96	97.96	97.96
CD	96.6	71.72	82.32
MD	95.56	100.0	97.73
NNPS	23.08	75.0	35.29

Experiment 2 :

I tried varying the smoothing parameter for laplace smoothing and got interesting results for it

Laplace Smoothing constant	Accuracy
1	58.96
0.1	76.565
0.01	89.237

Analysis :

This goes to show that adding to much weight to the unknowns can affect the probabilities drastically. It is very nice to see that as the laplace smoothing constant becomes smaller , the accuracies increase drastically.

Task 2 : Changing the input size :

Size of test and train case	Accuracy
Train : 200 , Test : 200	60.488
Train : 1000 , Test : 200	77.156
Train :2000 , Test : 200	86.49
Train:8736 , Test:200	89.237

Analysis :

With the increase in the input size , we see that the accuracy increase . We can also notice that however after a point , the increase in the input size does not affect the accuracy so much . Hence the increase in training size , increases accuracies but after a point , the accuracies become almost constant

Task 3 : Reduced the POS tags and measure accuracy :

On Composition	Accuracy
No Composition	89.237
Composing All Vbs(VBD/G/N/P/Z)	81.003
Composing Vbs,NNs	67.621
Composing Vbs,NNs,JJs,RBs	66.77
Composing Jjs,RBs	88.415

Analysis :

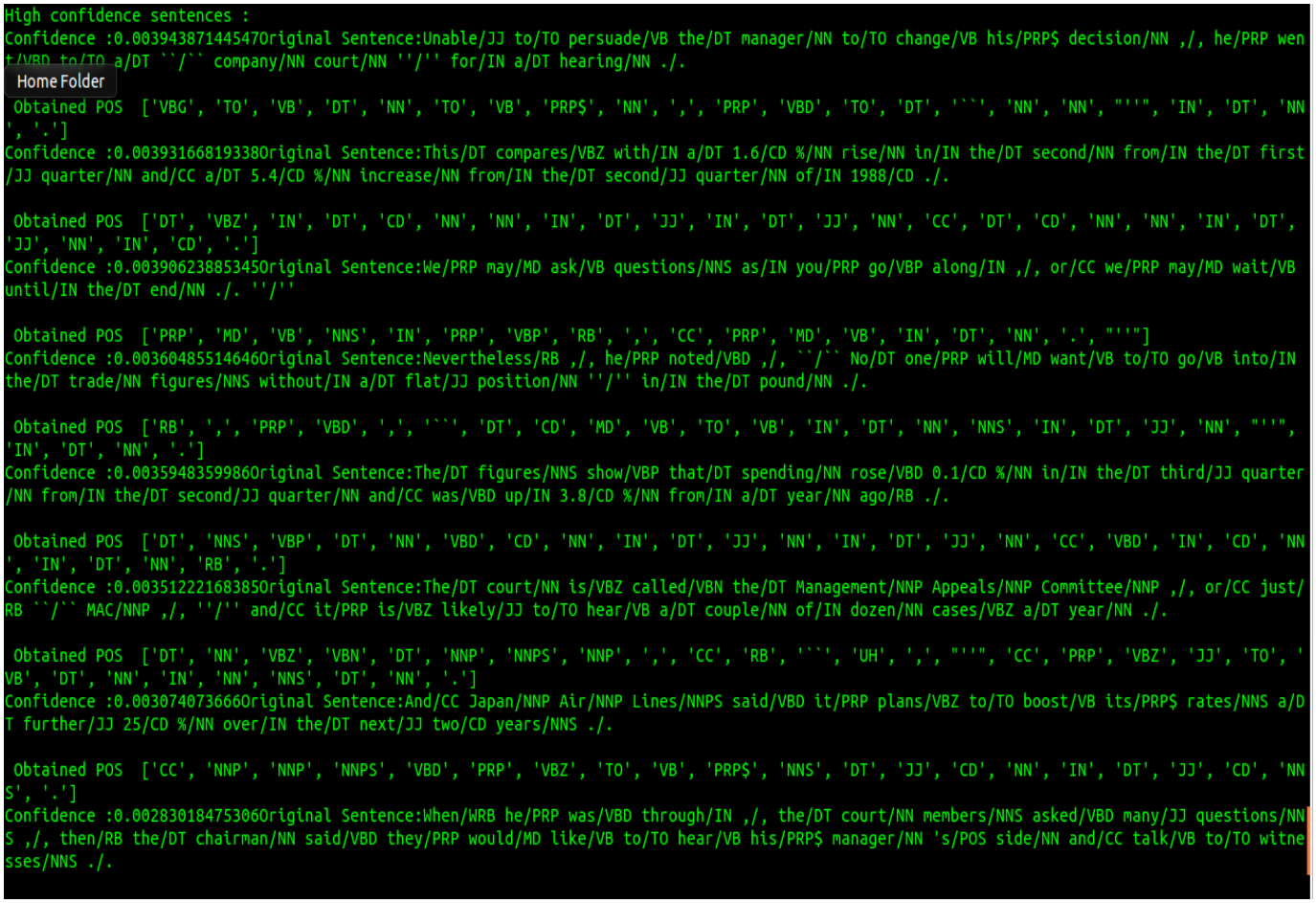
The above table has a number of interesting results. The first notable fact is that tag composition reduces the accuracy in most of the cases . When tags like verb are composed (past tense verb, gerund verb , past participle , 3rd person singular present and non-3rd person singular present) , the accuracy drops quite a bit . This is because , there exists pattern in the text such that a given type of word is mostly followed by a given tag and hence if the tags are composed it affects the accuracy severly . When both the nouns and the verb are composed , we see that the accuracy drops very badly and the composition of adjectives and adverbs makes the situation worse.

Another interesting point to note is when the tags like the adjectives and the adverbs are composed , the accuracy does not drop so much . This goes to show that these words do not affect their successor so much .

Sorting sentences based on confidence and interpreting trends

As a part of this task , I decided to find the confidence of the tasks , sort them based on their confidence and look at the most confident and the least confident results. For confidence I adopted the same method as the one given in the hint . I took the most likely path and took the Nth root of it , where N is the length of the sentence

```
Low confidence sentences :
Confidence :0.0001563406410170riginal Sentence:A.P./NNP Green/NNP currently/RB has/VBZ 2,664,098/CD shares/NNS outstanding/JJ ./.

 Obtained POS  ['NNP', 'NNP', 'RB', 'VBZ', 'RP', 'NNS', 'JJ', '.']
Confidence :0.0001605999439810riginal Sentence:Smith-Kline/NNP Beecham/NNP goes/VBZ further/JJ and/CC sometimes/RB features/VBZ its/PRP$ grieva
nce/NN procedure/NN in/IN closed-circuit/JJ TV/NN programs/NNS ./.

 Obtained POS  ['SYM', 'FW', 'FW', 'FW', 'CC', 'RB', 'VBZ', 'PRP$', 'JJS', 'NN', 'IN', 'PRP$', 'NN', 'NNS', '.']
Confidence :0.0001705339666010riginal Sentence:Fujisawa/NNP added/VBD 80/CD to/TO 2,010/CD and/CC Mochida/NNP advanced/VBD 230/CD to/TO 4,400/C
D ./.

 Obtained POS  ['PRP', 'VBD', 'CD', 'TO', 'VB', 'CC', 'EX', 'VBD', 'CD', 'TO', 'VB', '.']
Confidence :0.0001839674911 2Original Sentence:The/DT FT/NNP 30-share/JJ index/NN closed/VBD 11.0/CD points/NNS lower/JJR at/IN 1761.0/CD ./.

 Obtained POS  ['DT', 'RBS', 'JJ', 'NN', 'VBD', 'PRP', 'VBZ', 'JJR', 'IN', 'FW', '.']
Confidence :0.0001874199003760riginal Sentence:Kyocera/NNP advanced/VBD 80/CD yen/NN to/TO 5,440/CD ./.

 Obtained POS  ['PRP', 'VBD', 'CD', 'NN', 'TO', 'VB', '.']
Confidence :0.0002000002095560riginal Sentence:4/CD ./. Make/VB your/PRP$ due-process/NN system/NN visible/JJ ./.

 Obtained POS  ['CD', '.', "''", 'PRP$', 'FW', 'FW', 'FW', '.']
Confidence :0.0002279651095730riginal Sentence:Sapporo/NNP gained/VBD 80/CD to/TO 1,920/CD and/CC Kirin/NNP added/VBD 60/CD to/TO 2,070/CD ./.

 Obtained POS  ['PRP', 'VBD', 'CD', 'TO', 'VB', 'CC', 'EX', 'VBD', 'CD', 'TO', 'VB', '.']
Confidence :0.0002432421040160riginal Sentence:Misawa/NNP Homes/NNP was/VBD up/IN 20/CD at/IN 2,960/CD ./.

 Obtained POS  ['``', 'EX', 'VBD', 'IN', 'CD', 'IN', 'FW', '.']
Confidence :0.0002457194411130riginal Sentence:No/DT lawyers/NNS or/CC tape/NN recorders/NNS were/VBD present/JJ ./.

 Obtained POS  ['DT', 'NNS', 'CC', 'NN', 'WDT', 'VBD', 'JJ', '.']
Confidence :0.0002545518956290riginal Sentence:Kajima/NNP advanced/VBD 40/CD to/TO 2,120/CD and/CC Ohbayashi/NNP added/VBD 50/CD to/TO 1,730/CD
 ./.

 Obtained POS  ['PRP', 'VBD', 'CD', 'TO', 'VB', 'CC', 'EX', 'VBD', 'CD', 'TO', 'VB', '.']
Confidence :0.0002754661675Original Sentence:Winners/NNS outpaced/VBD losers/NNS ,/, 572/CD to/TO 368/CD ,/, while/IN 181/CD issues/NNS remaine
d/VBD unchanged/JJ ./.

 Obtained POS  ['SYM', 'FW', 'NNS', ',', 'VBG', 'TO', 'CD', ',', 'IN', 'PRP$', 'NNS', 'VBD', 'JJ', '.']
```



Analysis :

In general it looks like the sentence with high confidence are far more accurate than the sentences with lesser confidence. The high confidence sentence seem to have almost all the tags in place . The low seem to be shorter in general and they seem have misunderstood a number of tags . For example in case of sentence 1 in low confidence , the number has been choosen as a particle. This could because they are more probable after a verb when compared to a number . We also see in that in a number of sentences , just one tag is misplaced like in case the of high confidence sentences and we cannot come to a clear conclusion that the low confidence sentences have a large number of misplaced tags