Tianrungroj Yossathorn
03200437

Assignment 2

1. Data

The data that being used in this assignment is *84. Penn Treebank Sample* (nltk.org/data). Inside the 199 files under 'tagged' directory, are sentences with each word labeled with its part of speech. Here is an example from file 'wsj_0003.pos'.

```
…

[ The/DT asbestos/NN fiber/NN ]
,/,
[ crocidolite/NN ]
,/, is/VBZ unusually/RB resilient/JJ once/IN
[ it/PRP ]
enters/VBZ
[ the/DT lungs/NNS ]
,/, with/IN
[ even/RB brief/JJ exposures/NNS ]
to/TO
[ it/PRP ]
causing/VBG
[ symptoms/NNS ]

[ that/WDT ]
show/VBP up/IN
[ decades/NNS ]
later/JJ ,/,
[ researchers/NNS ]
said/VBD ./.

[ Lorillard/NNP Inc./NNP ]
,/,
[ the/DT unit/NN ]
of/IN
[ New/JJ York-based/JJ Loews/NNP Corp./NNP ]

[ that/WDT ]
makes/VBZ Kent/NNP
[ cigarettes/NNS ]
,/, stopped/VBD using/VBG
[ crocidolite/NN ]
in/IN
[ its/PRP$ Micronite/NN cigarette/NN filters/NNS ]
in/IN
[ 1956/CD ]
./.

…
```

As you can see, each word has their slash-separated corresponding part of speech behind it. However, after I inspected a few files, I found that only some of them have a clear sentence separator. Luckily, I also found out that two newline characters always follow the end of the sentence, although it's not guarantee that there is only one sentence between these indicators. The reason that a period ('.') cannot be used as a sentence-ending indicator is because they also appear in acronyms and in the middle of a sentence that contains quotation marks.

Tianrungroj Yossathorn
03200437

2. Implementation

To implement a Viterbi algorithm, first I need to clean the data. I extracted sentences from the files and accumulated the labels. The word frequency is also being counted in this step. Next is pre-processing the data and preparing the probability tables. In this step, I used the word frequency to mark a word that only appear once as 'UNK' and also mark numbers as 'NUM'. Then I created two probability tables using the number of unique labels and unique words. The two probability tables are the table indicating the probability that a label would follow by another and the table indicating the probability that a label would be assigned to a word. Next step is filling the table. I divided the data into 80% training and 20% testing. The frequencies are counted, and the probabilities are calculated and filled into the tables using only the sentences in training data. Lastly, by using to the probability tables, the sentences in the test data is being predicted, likelihoods and accuracies are also calculated. Here I used three different metrics, accuracy by sentence, accuracy by word and average correct prediction ratio.

The full implementation can be found in 'IS10_Kadai2.pdf'.

3. Result

Train data size: 1460 sentences
Test data size: 273 sentences
Elapsed training time: 1.72 seconds

metrics	value
Accuracy by sentence	0.0659
Accuracy by word	0.8647
Average correct prediction ratio	0.8702

```
Example sentence (normalized):
-- dorothy l. sayers , `` the nine tailors '' aslacton , england -- of all scenes that
 evoke rural england , this is one of the loveliest : an ancient stone church stands a
mid the fields , the sound of bells cascading from its tower , calling the faithful to
 evensong .

Labels:
: NNP NNP NNP , `` DT CD NNP '' NNP , NNP : IN DT NNS WDT VBP JJ NNP , DT VBZ CD IN DT
 JJS : DT JJ NN NN VBZ IN DT NNS , DT NN IN NNS VBG IN PRP$ NN , VBG DT NN TO NN .

Prediction:
: NNP NNP NNP , `` DT CD NN '' NNP , NNP : IN DT NN IN NNP NNP NNP , DT VBZ CD IN DT N
N : DT NN NN NN VBZ VBN DT NNS , DT JJ IN NNP NNP IN PRP$ NNP , NNP DT NN TO VB .

Likelihood=6.500328067697114e-114
```

4. Discussion

I want to point out to the first metrics. The accuracy by sentence is merely 6.5%. This is definitely because of the length of the sentences I extracted from the data. The reason is because the model did very good in word-level metrics which are more than 80% accuracy on both metrics. The evidence is also in the example above, for a 30-words long sentence like this, the predictions are mostly correct.