# Report

## HMM

The size of training sentences and testing sentences can be defined as arguments of the program. For more understanding, the combination of arguments is explained in the Readme file.

### Training

In training there are 4 steps needed to be done in the training:

1. Initialize all of the necessary variables such as training sentences and words &tags in the sentences.
2. Replace all of the words that occur only one in the training sentences with ‘UNK’.
3. Construct an emission probability table.
4. Applying smoothing technique which could be either Good-Turing or Laplace and construct a transition probability table. The smoothing technique will be selected beforehand as an argument of the program.

### Testing

In Testing there are 3 steps needed to be done in the training:

1. Initialize testing sentences and other variables (calculating accuracy purpose).
2. Predict tags of each word in the testing sentences by using Viterbi algorithm. While running the algorithm the program will keep track the accuracy of prediction.
3. Display the confusion matrix showing prediction accuracy on terminal. An extra csv file will be written out too because it will be easier to read when there are too many tags.

## Accuracy

The overall accuracy is approximately 80-95% which depends on the selected corpus and amount of training sentences. For instance, With Brown corpus, Smoothing = Laplace Tag set = ‘universal’, Training size = 10000 and Testing size = 500, the achieved prediction accuracy is 93.93%, and the confusion matrix is shown in the Table 1.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ADV | NOUN | NUM | ADP | PRT | DET | . | PRON | VERB | X | CONJ | ADJ | Correct | Total | Accuracy |
| ADV | 462 | 29 | 0 | 25 | 4 | 12 | 0 | 0 | 8 | 0 | 0 | 42 | 462 | 582 | 79.38% |
| NOUN | 0 | 2584 | 0 | 1 | 0 | 21 | 1 | 0 | 22 | 0 | 0 | 24 | 2584 | 2653 | 97.39% |
| NUM | 0 | 19 | 135 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 135 | 158 | 85.44% |
| ADP | 9 | 3 | 0 | 1573 | 3 | 2 | 0 | 2 | 2 | 0 | 0 | 1 | 1573 | 1595 | 98.62% |
| PRT | 3 | 1 | 0 | 59 | 205 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 205 | 267 | 76.77% |
| DET | 0 | 1 | 0 | 8 | 0 | 1496 | 0 | 1 | 0 | 0 | 0 | 0 | 1496 | 1506 | 99.33 |
| . | 0 | 0 | 0 | 0 | 0 | 0 | 1299 | 0 | 0 | 0 | 0 | 0 | 1299 | 1299 | 100% |
| PRON | 0 | 0 | 0 | 26 | 0 | 3 | 0 | 313 | 0 | 0 | 0 | 0 | 313 | 342 | 91.52% |
| VERB | 0 | 133 | 0 | 25 | 0 | 25 | 0 | 0 | 1525 | 0 | 0 | 8 | 1525 | 1716 | 88.86% |
| X | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 5 | 0% |
| CONJ | 2 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 408 | 0 | 408 | 413 | 98.78% |
| ADJ | 12 | 123 | 0 | 0 | 0 | 25 | 1 | 0 | 4 | 0 | 0 | 848 | 848 | 1013 | 83.71% |

Table : Confusion Matrix

It is possible to see that the tag that has the highest accuracy is DET and the lowest is PRT (without considering tag X and ‘.’).

## Observation

This program will be tested with several corpuses besides Brown corpus which are Conll2000, Conll2002, Alpino, Floresta, and Treebank. The accuracy of each one results as follow:

* Conll2000, with setting, Tag set = ‘universal’, Smoothing = Good-Turing, Training size = 10000 and Testing size = 500.

The prediction accuracy is 94.97%.

* Conll2002, with setting, Smoothing = Good-Turing, Training size = 10000 and Testing size = 200.

The prediction accuracy is 93.98 %.

* Treebank, with setting, Smoothing = Good-Turing, Training size = 3500 and Testing size = 400. The highest accuracy tag is PUNCT with 100%, and the lowest is FIXED with 4.1%.

The prediction accuracy is 91.07 %. There are several tags that can reach 100% accuracy such as PRP$, WP, $, TO, PRP, -LRB-, RBS, -NONE-, EX, etc. Besides tags that have 0% accuracy, the lowest tag accuracy is NNPS.

Conll2000, Conll2002 and Treebank can achieve high accuracy because they are still English which it does not have a complex structure, and the number of training sentences is quite high except Treebank.

* Alpino, with setting, Smoothing = Good-Turing, Training size = 6500 and Testing size = 636.

The prediction accuracy is 86.46 %. The highest accuracy tag is PUNCT with 100%, and the lowest is FIXED with 4.1%.

* Floresta, with setting, Smoothing = Good-Turing, Training size = 9000 and Testing size = 50.

The prediction accuracy is 81.072210 %. The number of training sentences may not be large compare to the other conditions due to the slowness in computing, but it is enough to see estimated result. Without considering tags with 100% and 0%, the highest accuracy tag is H+n with 95.27%, and the lowest accuracy tag is P<+n with 15.38%.

It seems that for language like Dutch and Portuguese this program could not achieve high accuracy unlike English. The reason for that might be the fact that both of them have complex structure or gramma (can be seen from number of tags).

All of the confusion matrix of each setting condition above can be found in the Confusion Matrix folder included along with the submission files.