CSE556 - Natural Language Processing
Assignment 4 Report

Aditya Chetan (2016217)

October 17, 2018

1 Assumptions

- I have considered that all the utterances will be in lower cases in the test file.
- I have considered that the test file will be in the format as prescribed in the assignment problem statement.
- I have implemented Add-k smoothing for both bigram transitional probabilities and emission probabilities. By default, `k = 1`
- In emission probabilities, I have also implemented replacement of words for which frequency is less than a threshold with `UNK`. By default, this threshold is set to 1.
- I have considered "." to be the final tag, q_f

2 Methodology

2.1 Training the HMM

- The function `get_HMM()` can either load a saved HMM or train a new HMM from a given training corpus with appropriate normalization for words with less frequency.
- It returns a dictionary with keys `A, B & pi`
- `A`: It contains the transitional counts for the hidden states and will be later used to get transitional probabilities.
- `B`: Used to store emission counts
- `pi`: Used to store initialization probabilities for the hidden states

2.2 Decoding

Here, Viterbi algorithm is implemented with help of functions `max & argmax` from the `numpy` library in python.

3 Example

For an input file that looks like the following:

```
i
'd
wan
na
eat
food
.

some
french
restaurants
please
.
```

We get the following output:

```
i PRP
'd MD
wan VB
na TO
eat VB
food NN
. .

some DT
french JJ
restaurants NNS
please UH
. .
```