CS673A: Machine Translation
Assignment #2

Due on: 29-Feb-2020, 23:59

This is a group assignment. Attempt both parts of the assignment: 1A and 1B.

Read the following before getting started:

- Please note that part 1A carries 50% and part 1B carries 50%.
- There will be some more hidden test cases over which your program will be evaluated. The evaluation will be done over the test cases similar to the test cases given in the assignment.
- All the files (code, output, etc. except training and testing files) must be in a single zipped (.zip) folder. Name the zip file as < GroupNo >.zip e.g. 3.zip. Also make separate directories for Part 1A and Part 1B to include the corresponding files.
- Submit a README file having the details about the programming language, packages/software, steps used in the assignment, and a brief description of your approach. Also, mention the commands for compiling/executing the program.
- There will be a demo of your work. Each group member should be acquainted with the whole work and explain one's contribution.
- For part 1A you have to come up with your own code. But for part 1B if you use an open source parser, then you have to mention this in the README file and 20% marks will be deducted for using open source code (or executable). If you fail to mention that you have used open source code (or executable) and used it, then it will come under cheating. However, it is allowed to depend on some pseudo code or algorithm for implementing your work.
- There will be no extension for submitting the assignment (mid semester exam is already considered in the deadline).
- The solutions must be written on your own and any sources must be clearly referenced. Any instance of plagiarism will be severely punished.
  https://cse.iitk.ac.in/pages/AntiCheatingPolicy.html

Part 1A

You are given a training corpus (train.txt) containing POS tagging for English sentences. Use the corpus to learn a POS tagging model and then give precision, recall and F1 Score for each POS tag on the test corpus (test1.txt). The test corpus is similar to the training corpus and contains human tagging, so that you can use it with your predicted tagging to calculate the above mentioned scores. You will be evaluated on another test corpus (test2.txt) during the demo.

Tasks:

- Find the bigram and lexical probabilities
- Use either Viterbi algorithm or your own method for tagging

The performance measures (precision, recall and F1 Score) should be reported in a text file (output.txt). Format of train.txt, test1.txt and test2.txt files is:

- Each line represents a sentence
- '#' symbol: separates a word from its POS tag
- '@' symbol: separates one word from another
- Example:
  Corpus:
  John#NOUN@entered#VERB@the#DET@room#NOUN@from#ADP@the#DET@hallway#NOUN@to#ADP@the#DET@kitchen#NOUN@.#.
  Sentence: John entered the room from the hallway to the kitchen.
  POS tags: NOUN VERB DET NOUN ADP DET NOUN ADP DET NOUN .

Note: F1 score will be used for grading part 1A so make sure you get good results.

Part 1B

Given the following grammar:
S → NP VP
NP → ART N | ART N PP | PRON | N
VP → V NP PP | V NP
PP → P NP
ART → a | an | the
N → boy | telescope | football | jam | book | saw | play
PRON → I | we | you | they
V → saw | play | eat | study | jam
P → with | for

Parse the following sentences (you can use any parsing algorithm):

- I saw a boy with a telescope
- We play football
- They eat the jam
- I play with you
- You saw a play

Please mention if the sentence is parsable or not, and if it is parsable then output the parsing of each sentence in bracketed parse (tree) format in a text file (output.txt) as shown below. If there are multiple parses for a sentence show at least two parses.

Example:
The sentence “I saw a boy with a telescope” has two possible parses:

- (S (NP (PRON “I”)) (VP (V “saw”)(NP (ART “a”) (N “boy”)) (PP (P “with”) (NP (ART “a”) (N “telescope”)))))
- (S (NP (PRON “I”)) (VP (V “saw”) (NP (ART “a”) (N “boy”) (PP (P “with”) (NP (ART “a”) (N “telescope”))))))