Homework 5: Sequence Labeling

Due March 8, 2021 (11:59 PM)

1 Introduction

In this homework, you will be using a bigram hidden Markov model (HMM) for part-of-speech (POS) sequence labeling. We will explore two decoding algorithms: greedy (implemented) and Viterbi (you will implement this).

For a recap of HMMs and the Viterbi algorithm for POS sequence labeling, you may wish to review the Lecture 10 from February 18, 2021. Recall that the bigram HMM assumption is that the label for a particular time step (word in a sentence) only depends on the current and previous time step (current word and previous word).

The data we will use for this assignment is from the Universal Dependencies English Web Treebank, which consists of sentences (sequences of words and punctuation, sourced from weblogs, newsgroups, email, reviews, and question-answers on the internet) along with POS labels for each part of the sequence.

The Colab notebook for this homework is located here:
`https://github.com/dbamman/nlp21/blob/main/HW5/HW_5.ipynb`.

2 Deliverable: `viterbi_decode` function implementation

Your task is to implement the function `viterbi_decode`. This function takes in the following arguments:

- `transitions`: a matrix where the entry `transitions[s1, s2]` is the log probability of observing label `s2` after label `s1` in a sequence.
- `emissions`: a matrix where the entry `emissions[s, t]` is the log probability of observing token `t` with label `s`.
- `y_vocab`: a dictionary mapping each POS tag label to an index of `0` thru the number of POS tag labels.
- `x_vocab`: a dictionary mapping each vocab word to an index of `0` thru the number of vocab words.
- `sequence`: a list of (string) words/tokens.

The `viterbi_decode` function should output a string with the POS tag labels (separated by spaces) for the words in the given sequence (ex: `"NN VBZ IN DT NN"`).

Hint: The last slide of Lecture 10 from February 18, 2021 provides a pseudocode for the Viterbi algorithm that you may find helpful.

3 How to Submit

1. Download your Colab notebook as an .ipynb file (File $\rightarrow$ Download .ipynb)
2. Submit **HW_5.ipynb** to the Homework 5 assignment on Gradescope. The files must be named **HW_5.ipynb** for the Gradescope autograder.