# Homework 4: Sequence labeling (CS 2731 Fall 2023)

**Due 2023-11-09, 11:59pm**. *Instructions last updated 2023-11-03.*

In this assignment, you will manually decode the highest-probability sequence of part-of-speech tags from a trained HMM using the Viterbi algorithm. You will also fine-tune BERT-based models for named entity recognition (NER).

## 1. POS tagging with an HMM

Consider a Hidden Markov Model with the following parameters: postags = {NOUN, AUX, VERB}, words = {'Patrick', 'Cherry', 'can', 'will', 'see', 'spot'}

Initial probabilities:

| | $\pi$ |
|---|---|
| NOUN | 0.7 |
| AUX | 0.1 |
| VERB | 0.2 |

Transition probabilities: The format is P(column_tag | row_tag), e.g. P(AUX | NOUN) = 0.3.

| | NOUN | AUX | VERB |
|---|---|---|---|
| NOUN | 0.2 | 0.3 | 0.5 |
| AUX | 0.4 | 0.1 | 0.5 |
| VERB | 0.8 | 0.1 | 0.1 |

Emission probabilities:

| | Patrick | Cherry | can | will | see | spot |
|---|---|---|---|---|---|---|
| NOUN | 0.3 | 0.2 | 0.1 | 0.1 | 0.1 | 0.2 |
| AUX | 0 | 0 | 0.4 | 0.6 | 0 | 0 |
| VERB | 0 | 0 | 0.1 | 0.2 | 0.5 | 0.2 |

Using the Viterbi algorithm and the given HMM, find the most likely tag sequence for the following 2 sentences.

1. "Patrick can see Cherry"
2. "will Cherry spot Patrick"

To get you started on the Viterbi tables, here are the first 2 columns for the first sentence. You'll also want to include the backtraces.

| POS state | Patrick | can | see | Cherry |
|---|---|---|---|---|
| NOUN | 0.21 | 0.0042 | | |
| AUX | 0 | 0.0252 | | |
| VERB | 0 | 0.0105 | | |

### Deliverables for part 1

In your report, show your work for calculating the Viterbi tables or lattices for both example sentences. Report the most likely tag sequences for these 2 sentences.

## 2. Fine-tune BERT-based NER models

In this section, you will fine-tune multiple pretrained BERT-based models on Spanish NER data. Specifically, you will fine-tune at least one model pretrained on masked language modeling (MLM) on Spanish data, and at least one model pretrained on NER in a language other than Spanish.

Copy this skeleton Colab notebook, run the cells, and fill in the places that are specified.

### Deliverables for part 2

In your report, include:

1. The F1 score on the CoNLL-2003 Spanish test set for
   1. the model pretrained on MLM in Spanish, and
   2. the model pretrained on NER in another language
2. A brief discussion of which model performs better and any choices you made about hyperparameters in training
3. A link to your copied and filled out Colab notebook

### Submission

Please submit the following items on Canvas:

- Your report with results and answers to questions in Part 1 and Part 2, named `report_{your pitt email id}_hw3.pdf`. No need to include @pitt.edu, just use the email ID before that part. For example: `report_mmy29_hw3.pdf`.
- A `README.txt` file explaining
  - any additional resources, references, or web pages you've consulted
  - any person with whom you've discussed the assignment and describe the nature of your discussions
  - any generative AI tool used, and how it was used
  - any unresolved issues or problems

This homework assignment is worth 45 points.

### Acknowledgments

Part 1 of this assignment is based on homework assignments by Prof. Hyeju Jang and Prof. Diane Litman.