Natural Language Processing
Dr. Ali Bahrainian
Published: 09.05.2024

University of Tübingen
Summer Term 2025

Assignment 3

Instructions

This assignment is ungraded, so there is no need to share your solution. Please complete this exercise sheet to practice what you learned in the lecture.
The assignment will be discussed in the tutorial on 16.05.2024.

Exercise 1 - Bigram Inference

You are given the following training corpus:

1. `<s> I am Sam </s>`
2. `<s> Sam I am </s>`
3. `<s> Sam I like </s>`
4. `<s> Sam I do like </s>`
5. `<s> do I like Sam </s>`

Assume now that you have trained a bigram language model on this corpus.

1. What is the most probable next word predicted by the model for the following word sequences?
   (a) `<s> Sam ...`
   (b) `<s> Sam I do ...`
   (c) `<s> Sam I am Sam ...`
   (d) `<s> do I like ...`
2. Which of the following sentences is better, e.g. it gets a higher probability with this model?
   (a) `<s> Sam I am </s>`
   (b) `<s> Sam I do I like </s>`
   (c) `<s>I do like Sam I am</s>`
3. Compute the perplexity of the model for the following sequence (note that, in general, start-of-sentence tokens are excluded when calculating perplexity):
   `<s> I do like Sam`

Exercise 2 - Character recognition using HMM

Given the structure of hidden states (see figure 1) and the learned HMM for character 'A' and the learned HMM for character 'B' as follows:

$$A^{(\text{letter A})} = \begin{bmatrix} .8 & .2 & 0 \\ 0 & .8 & .2 \\ 0 & 0 & 1 \end{bmatrix} \quad B^{(\text{letter A})} = \begin{bmatrix} .9 & .1 & 0 \\ .1 & .8 & .1 \\ .9 & .1 & 0 \end{bmatrix}$$

and similarly for letter "B":

$$A^{(\text{letter B})} = \begin{bmatrix} .8 & .2 & 0 \\ 0 & .8 & .2 \\ 0 & 0 & 1 \end{bmatrix} \quad B^{(\text{letter B})} = \begin{bmatrix} .9 & .1 & 0 \\ 0 & .2 & .8 \\ .6 & .4 & 0 \end{bmatrix}$$

For the transition matrices, rows denote the current state and columns the next state. For example, the probability of transitioning from state 1 to state 2 is given by:

$$P(S_{i+1} = s_2 | S_i = s_1) = A_{12}^{(\text{letter A})} = A_{12}^{(\text{letter B})} = 0.2$$

And similarly, the probability of observing "3" given the process is currently in state 2 is:

$$P(O_i = 3 | S_i = s_2) = B_{23}^{(\text{letter A})} = 0.1$$
$$P(O_i = 3 | S_i = s_2) = B_{23}^{(\text{letter B})} = 0.8$$

for letter "A" and letter "B", respectively.
Suppose that after character image segmentation the following sequence of island numbers in 4 slices was observed (see figure 2):

$$\vec{o} = (1, 3, 2, 1)$$

What HMM is more likely to generate this observation sequence, HMM for 'A' or HMM for 'B'?

Assume that each HMM is initially in state s_1, so:

$$\pi = (P(S_1 = s_1), P(S_1 = s_2), P(S_1 = s_3)) = (1, 0, 0)$$

and that no state-sequence repeats a state more than once, i.e. the following sequence is NOT possible since it repeats state 2 twice:

$$s_1 \rightarrow s_2 \rightarrow s_2 \rightarrow s_2$$

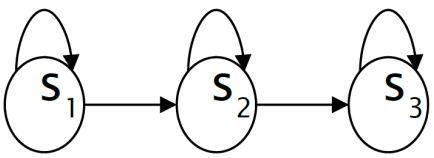

Figure 1: Structure of the hidden states

Figure 2: An example of a vertical slice for both characters.