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`

Solution

We are given a sequence $\vec{w} = (w_1, w_2, ..., w_n)$ and vocabulary V with $w_i \in V$. The probability of any given sequence can be formulated as (chainrule):

$$P(\vec{w}) = P(w_1) * P(w_2|w_1) * P(w_3|w_1, w_2) * ... * P(w_n|w_1, ..., w_{n-1})$$

Since we are using a bigram model, we can apply the following Markov assumption:

$$P(w_i|w_1, ..., w_{i-1}) = P(w_i|w_{i-1})$$

that is, only the previous word in the sequence matters when predicting the next word. This yields:

$$P(\vec{w}) = P(w_1) * P(w_2|w_1) * P(w_3|w_2) * ... * P(w_n|w_{n-1})$$

Finally, we introduce the special tokens `<s>` or Start-of-sequence (SOS) token to start any given sequence, which allows us to impose that $P(w_1 = \texttt{<s>}) = 1$ and `</s>` or end-of-sequence (EOS) token to end any given sequence. This ensures that the model we use defines a single probability distribution over all possible sequences [1]. Finally, we can perform MLE to obtain the probability of each bigram [2]:

$$P(w_i|w_{i-1}) = \frac{\text{count}((w_{i-1}, w_i))}{\text{count}(w_{i-1})}$$

Now, we have everything we need to solve this exercise. However, to make our lifes a bit easier, we can already compute all bigram probabilities for our training corpus:

W_{i-1}/W_i	<s>	am	do	I	like	Sam	</s>
<s>	0	0	$\frac{1}{5}$	$\frac{1}{5}$	0	$\frac{3}{5}$	0
am	0	0	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$
do	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0
I	0	$\frac{2}{5}$	$\frac{1}{5}$	0	$\frac{2}{5}$	0	0
like	0	0	0	0	0	$\frac{1}{3}$	$\frac{2}{3}$
Sam	0	0	0	$\frac{3}{5}$	0	0	$\frac{2}{5}$

For example:

$$P(\texttt{Sam}|\texttt{<s>}) = \frac{\text{count}((\texttt{<s>}, \texttt{Sam}))}{\text{count}(\texttt{<s>})} = \frac{3}{5}$$

1. (a) The last word of the sequence is `''Sam''`. As established before, this is the only word we need to predict the next word of any given sequence. Using our table, we see that:

$$P(\texttt{I}|\texttt{Sam}) = \frac{3}{5} > P(\texttt{</s>}|\texttt{Sam}) = \frac{2}{5} > P(\text{"any other word"}\ w \in V|\texttt{Sam}) = 0$$

Under our model, the most probable next word is `I`.

[1] Further reading: Dropping the EOS-token would mean that only sequences of the same length share a probability distribution. See for instance Jurafsky, Martin: "Speech and Language Processing"

[2] Further reading: For a derivation, see for instance
https://leimao.github.io/blog/Maximum-Likelihood-Estimation-Ngram/

(b) $P(\texttt{like}|\texttt{do}) = P(\texttt{I}|\texttt{do}) = \frac{1}{2}$
Under our model, the most probable words are `like` and `I` (both are equally likely - its a tie).

(c) The last word in the sequence is `Sam`. Thus, the result is the same as in a)

(d) $P(\texttt{</s>}|\texttt{like}) = \frac{2}{3} > P(\texttt{Sam}|\texttt{like}) = \frac{1}{3}$
Under our model, the most probable next word is `</s>`.

2. (a)

$$\begin{aligned} P((\texttt{<s>, Sam, I, am, </s>})) &= P(\texttt{<s>}) * P(\texttt{Sam}|\texttt{<s>}) * P(\texttt{I}|\texttt{Sam}) * P(\texttt{am}|\texttt{I}) * P(\texttt{</s>}|\texttt{am}) \\ &= 1 * \frac{3}{5} * \frac{3}{5} * \frac{2}{5} * \frac{1}{2} \\ &\approx 0.0288 \end{aligned}$$

(b) $1 * \frac{3}{5} * \frac{3}{5} * \frac{1}{5} * \frac{1}{2} * \frac{2}{5} * \frac{2}{3} \approx 0.0096$

(c) $1 * \frac{1}{5} * \frac{1}{5} * \frac{1}{2} * \frac{1}{3} * \frac{3}{5} * \frac{2}{5} * \frac{1}{2} \approx 0.0008$

Thus, a) is the most likely sequence under our model.

3. Preplexity of a given sequence $W = w_1, ...$ with length N under a bigram model is defined as:

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}}$$

and so for W = `<s> I do like Sam`:

$$\begin{aligned} PP(W) &= \sqrt[4]{\prod_{i=1}^{4} \frac{1}{P(w_i|w_{i-1})}} \\ &= \sqrt[4]{\frac{1}{\frac{1}{5}} * \frac{1}{\frac{1}{5}} * \frac{1}{\frac{1}{2}} * \frac{1}{\frac{1}{3}}} \\ &= \sqrt[4]{150} \end{aligned}$$

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:

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

and similarly for letter "B":

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

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:

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

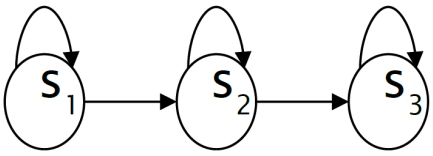

Figure 1: Structure of the hidden states

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

Solution

We want to find $P(O = \vec{o})$ under models $M = (A^{(\text{letter A})}, B^{(\text{letter A})}, \pi)$ and $M = (A^{(\text{letter B})}, B^{(\text{letter B})}, \pi)$. Consider:

$$\begin{aligned} P(O = \vec{o}) &= \sum_{\vec{s} \in \text{Seq}} P(O = \vec{o}, S = \vec{s}) \\ &= \sum_{\vec{s} \in \text{Seq}} P(O = \vec{o} | S = \vec{s}) P(S = \vec{s}) \\ &= \sum_{\vec{s} \in \text{Seq}} P(O_1 = o_1 | S_1 = x_1) P(S_1 = x_1) \prod_{i=2}^{4} P(O_i = o_i | S_i = x_i) P(S_i = x_i | S_{i-1} = x_{i-1}) \end{aligned}$$

where Seq is the set of possible sequences of length $|\vec{o}| = 4$ under M, with $\forall \vec{s} \in S : \vec{s} = (S_1 = x_1, S_2 = x_2, S_3 = x_3, S_4 = x_4)$, and $\forall x_i \in \vec{s} : x_i \in \{s_1, s_2, s_3\}$. Thus, we first need to find all possible transition sequences under M, which are as follows:

1. $s_1 \to s_1 \to s_1 \to s_1$
2. $s_1 \to s_1 \to s_1 \to s_2$
3. $s_1 \to s_1 \to s_2 \to s_2$
4. $s_1 \to s_1 \to s_2 \to s_3$
5. $s_1 \to s_2 \to s_2 \to s_2$
6. $s_1 \to s_2 \to s_2 \to s_3$
7. $s_1 \to s_2 \to s_3 \to s_3$

Note that this already excludes all sequences containing impossible transitions such as $s_2 \to s_1$ since

$$P(S_i = s_2 | S_{i-1} = s_1) = 0$$

under both models. Further, we can observe that

$$P(O_2 = 3 | S_2 = s_1) = 0$$

for states other than 2 under both models. This narrows down the possible transition sequences further, as we can now exclude all sequences that are not in state 2 at step 2, which leaves us with:

$$\begin{aligned} \vec{s_1} &= s_1 \to s_2 \to s_2 \to s_2 \\ \vec{s_2} &= s_1 \to s_2 \to s_2 \to s_3 \\ \vec{s_3} &= s_1 \to s_2 \to s_3 \to s_3 \end{aligned}$$

Finally, we need to filter all sequences that repeat states more than once:

$$\begin{aligned} \vec{s_1} &= s_1 \to s_2 \to s_2 \to s_3 \\ \vec{s_2} &= s_1 \to s_2 \to s_3 \to s_3 \end{aligned}$$

Now, all that's left to do is to find and sum the probabilities:
For instance, the probability of observing $\vec{o}$ under model "A" if it follows the first transition sequence $\vec{s}_1$ is given by:

$$\begin{aligned} P(O=\vec{o}|S=\vec{s}_1)P(S=\vec{s}_1) = {} & P(O_1=1|S_1=s_1)P(S_1=s_1)* \\ & P(O_2=3|S_2=s_2)P(S_2=s_2|S_1=s_1)* \\ & P(O_3=2|S_3=s_2)P(S_3=s_2|S_2=s_2)* \\ & P(O_4=1|S_4=s_3)P(S_4=s_3|S_3=s_2) \\ = {} & 0.9*1*0.1*0.2*0.8*0.8*0.9*0.2 \\ = {} & 0.0020736 \end{aligned}$$

and similarly:

$$\begin{aligned} P(O=\vec{o}|S=\vec{s}_2)P(S=\vec{s}_2) &= 0.9*1*0.1*0.2*0.1*0.2*0.9*1 \\ &= 0.000324 \end{aligned}$$

and thus under model "A":

$$\begin{aligned} P(O=\vec{o}) &= 0.0020736 + 0.000324 \\ &= 0.0023976 \end{aligned}$$

and for model "B":

$$\begin{aligned} P(O=\vec{o}|S=\vec{s}_1)P(S=\vec{s}_1) &= 0.9*1*0.8*0.2*0.2*0.8*0.6*0.2 \\ &= 0.0027648 \\ P(O=\vec{o}|S=\vec{s}_2)P(S=\vec{s}_2) &= 0.9*1*0.8*0.2*0.4*0.2*0.6*1 \\ &= 0.006912 \\ P(O=\vec{o}) &= 0.00027648 + 0.0006912 \\ &= 0.0096768 \end{aligned}$$

and so, the observed sequence is more likely under model "B".