Gal Cesana 318510633
*Authorization has been granted to submit the assignment individually

Natural Language Processing – Exercise 3

$Q1$:

Given Data:

Hidden states – {H, L}

Transition Probabilities		
	H	**L**
H	0.5	0.5
L	0.4	0.6

Emission Probabilities				
	A	**T**	**G**	**C**
H	0.2	0.2	0.3	0.3
L	0.3	0.3	0.2	0.2

Given sequence: **S = ACCGTGCA**

Starting state = H.

Viterbi Algorithm:

K=0:

$$\pi(0,H) = 1$$

$$\pi(0,L) = 0$$

$$bp(0,H) = bp(0,L) = None$$

We know that the starting state is H, and there are no back pointers yet.

K=1: A nucleotide was emitted

$$\pi(1,H) = \max\big(\pi(0,H) \cdot P(H|H) \cdot P(A|H)\,, \pi(0,L) \cdot P(H|L) \cdot P(A|H)\big)$$

$$= \max(1 \cdot 0.5 \cdot 0.2\,, 0 \cdot 0.4 \cdot 0.2) = 0.1$$

$$\pi(1,L) = \max\big(\pi(0,H) \cdot P(L|H) \cdot P(A|L)\,, \pi(0,L) \cdot P(L|L) \cdot P(A|L)\big)$$

$$= \max(1 \cdot 0.5 \cdot 0.3\,, 0 \cdot 0.4 \cdot 0.3) = 0.15$$

$$bp(1,H) = H$$

$$bp(1,L) = H$$

<u>K=2:</u> C nucleotide was emitted

$$\pi(2,H)=\max\big(\pi(1,H)\cdot P(H|H)\cdot P(C|H)\,,\pi(1,L)\cdot P(H|L)\cdot P(C,H)\big)$$

$$=\max(0.1\cdot 0.5\cdot 0.3\,,0.15\cdot 0.4\cdot 0.3)=\max(0.015{,}0.018)=0.018$$

$$\pi(2,L)=\max\big(\pi(1,H)\cdot P(L|H)\cdot P(C|L)\,,\pi(1,L)\cdot P(L|L)\cdot P(C|L)\big)$$

$$=\max(0.1\cdot 0.5\cdot 0.2\,,0.15\cdot 0.6\cdot 0.2)=\max(0.01{,}0.018)=0.018$$

$$bp(2,H)=L\ \ ,\ \ bp(2,L)=L$$

<u>K=3:</u> C nucleotide was emitted

$$\pi(3,H)=\max\big(\pi(2,H)\cdot P(H|H)\cdot P(C|H)\,,\pi(2,L)\cdot P(H|L)\cdot P(C,H)\big)$$

$$=\max(0.018\cdot 0.5\cdot 0.3\,,0.018\cdot 0.4\cdot 0.3)=\max(0.0027{,}0.00216)=0.0027$$

$$\pi(3,L)=\max\big(\pi(2,H)\cdot P(L|H)\cdot P(C|L)\,,\pi(2,L)\cdot P(L|L)\cdot P(C|L)\big)$$

$$=\max(0.018\cdot 0.5\cdot 0.2\,,0.018\cdot 0.6\cdot 0.2)=\max(0.0018{,}0.00216)=0.00216$$

$$bp(3,H)=H\ \ ,\ \ bp(3,L)=L$$

<u>K=4:</u> G nucleotide was emitted

$$\pi(4,H)=\max\big(\pi(3,H)\cdot P(H|H)\cdot P(G|H)\,,\pi(3,L)\cdot P(H|L)\cdot P(G,H)\big)$$

$$=\max(0.0027\cdot 0.5\cdot 0.3\,,0.00216\cdot 0.4\cdot 0.3)=\max(0.000405{,}0.0002592)$$
$$=0.000405$$

$$\pi(4,L)=\max\big(\pi(3,H)\cdot P(L|H)\cdot P(G|L)\,,\pi(3,L)\cdot P(L|L)\cdot P(G|L)\big)$$

$$=\max(0.0027\cdot 0.5\cdot 0.2\,,0.00216\cdot 0.6\cdot 0.2)=\max(0.00027{,}0.0002592)$$
$$=0.00027$$

$$bp(4,H)=H\ \ ,\ \ bp(4,L)=H$$

<u>K=5:</u> T nucleotide was emitted

$$\pi(5,H)=\max\big(\pi(4,H)\cdot P(H|H)\cdot P(T|H)\,,\pi(4,L)\cdot P(H|L)\cdot P(T,H)\big)$$

$$=\max(0.0000405{,}0.0000216)=0.0000405$$

$$\pi(5,L)=\max\big(\pi(4,H)\cdot P(L|H)\cdot P(T|L)\,,\pi(4,L)\cdot P(L|L)\cdot P(T|L)\big)$$

$$=\max(0.00006075{,}0.0000486)=0.00006075$$

$$bp(5,H)=H\ \ ,\ \ bp(5,L)=H$$

K=6: G nucleotide was emitted

$$\pi(6,H) = \max\big(\pi(5,H)\cdot P(H|H)\cdot P(G|H)\,,\pi(5,L)\cdot P(H|L)\cdot P(G,H)\big)$$

$$= \max(6.075\cdot 10^{-6}, 7.29\cdot 10^{-6}) = 7.29\cdot 10^{-6}$$

$$\pi(6,L) = \max\big(\pi(5,H)\cdot P(L|H)\cdot P(G|L)\,,\pi(5,L)\cdot P(L|L)\cdot P(G|L)\big)$$

$$= \max(4.05\cdot 10^{-6}, 7.29\cdot 10^{-6}) = 7.29\cdot 10^{-6}$$

$$bp(6,H) = L\ ,\ \ bp(6,L) = L$$

K=7: C nucleotide was emitted

$$\pi(7,H) = \max\big(\pi(6,H)\cdot P(H|H)\cdot P(C|H)\,,\pi(6,L)\cdot P(H|L)\cdot P(C,H)\big)$$

$$= \max(1.093\cdot 10^{-6}, 8.748\cdot 10^{-7}) = 1.093\cdot 10^{-6}$$

$$\pi(7,L) = \max\big(\pi(6,H)\cdot P(L|H)\cdot P(C|L)\,,\pi(6,L)\cdot P(L|L)\cdot P(C|L)\big)$$

$$= \max(7.29\cdot 10^{-7}, 8.748\cdot 10^{-7}) = 8.748\cdot 10^{-7}$$

$$bp(7,H) = H\ ,\ \ bp(7,L) = L$$

K=8: A nucleotide was emitted

$$\pi(8,H) = \max\big(\pi(7,H)\cdot P(H|H)\cdot P(A|H)\,,\pi(7,L)\cdot P(H|L)\cdot P(A,H)\big)$$

$$= \max(1.093\cdot 10^{-7}, 6.998\cdot 10^{-8}) = 1.093\cdot 10^{-7}$$

$$\pi(8,L) = \max\big(\pi(7,H)\cdot P(L|H)\cdot P(A|L)\,,\pi(7,L)\cdot P(L|L)\cdot P(A|L)\big)$$

$$= \max(1.64\cdot 10^{-7}, 1.57\cdot 10^{-7}) = 1.64\cdot 10^{-7}$$

$$bp(8,H) = H\ ,\ \ bp(8,L) = H$$

As we can see the best end state is L as shown in K=8: $\pi(8,L) > \pi(8,H)$.

We will follow the back pointers and reach to the best state-sequence:

hidden state	**L**	**H**	**H**	**H**	**L**	**H**	**H**	**L**
nucleotide emitted	A	C	C	G	T	G	C	A

Best sequence probability calculation:

$$P(S|best\ state\ seq) = P(A|L)\cdot P(C|H)\cdot P(C|H)\cdot P(C|H)\cdot P(G|H)\cdot P(T|L)\cdot P(G|H)\cdot P(C|H)\cdot P(A|L) = 0.3^{8} = 6.561\cdot 10^{-5}$$

$Q2$:

<u>Pseudo-code for the Four-gram Viterbi Algorithm:</u>

Input:

- An integer n.
- Parameters $q(w|t,u,v)$ and $e(x|s)$.

Definitions:

- K: Set of possible tags.
- $K_{-2} = K_{-1} = K_0 = \{*\}$.
- $K_k = K\ for\ k = 1,\dots,n$.
- V: Set of possible words.

Initialization:

1. Define $\pi(-2,*,*) = 1$ (Base probability).
2. For $k = -1,0$: $\pi(k,*,*) = 0$.

Algorithm:

```
1. Iterate over positions $i = 1$ to $n$:
    o For each tag $y_{i-3}$ in $K_{i-3}$:
        ▪ For each tag $y_{i-2}$ in $K_{i-2}$:
            • For each tag $y_{i-1}$ in $K_{i-1}$:
                o Compute the maximum probability:
                  $\pi(i, y_{i-2}, y_{i-1}) = \max_{y_i \in K_i}(\pi(i-1, y_{i-3}, y_{i-2}) \cdot q(y_i|y_{i-3}, y_{i-2}, y_{i-1}) \cdot e(x_i|y_i)$
                o Store the corresponding $y_i$ in a back-pointer $bp[i][y_{i-2}][y_{i-1}]$.
2. Final Step (n+1):
    o Set $\pi(n+1, STOP)$ to:
      $\max_{y_{n-2}, y_{n-1}, y_n \in K}(\pi(n, y_{n-2}, y_{n-1}) \cdot q(STOP|y_{n-2}, y_{n-1}, y_n)$
    o Track back-pointers for the best sequence.
3. Return:
    a. Use the back-pointers to reconstruct the optimal tag sequence $y_1, \dots, y_n, STOP$.
    b. Return the sequence and the maximum probability.
```

Practical Part:

Results:

Baseline Results:

Known Error Rate = 0.07130937812882412

Unknown Error Rate = 0.7739463601532567

Total Error Rate = 0.14442340277085616

HMM Tagger Results:

Known Error Rate = 0.06530203582155969

Unknown Error Rate = 0.6408045977011494

Total Error Rate = 0.12518688328515898

HMM Tagger with Add-one Smoothing Results:

Known Error Rate = 0.1551896762709979

Unknown Error Rate = 0.5852490421455939

Total Error Rate = 0.19994019734874913

Error rates with pseudo-words and maximum likelihood estimation:

Known Error Rate: 0.07586367880485527

Unknown Error Rate: 0.5160409556313993

Total Error Rate: 0.140137546097877

Error rates with pseudo-words and Add-One smoothing:

Known Error Rate: 0.1469421101774043

Unknown Error Rate: 0.5365187713310581

Total Error Rate: 0.20382736968005583

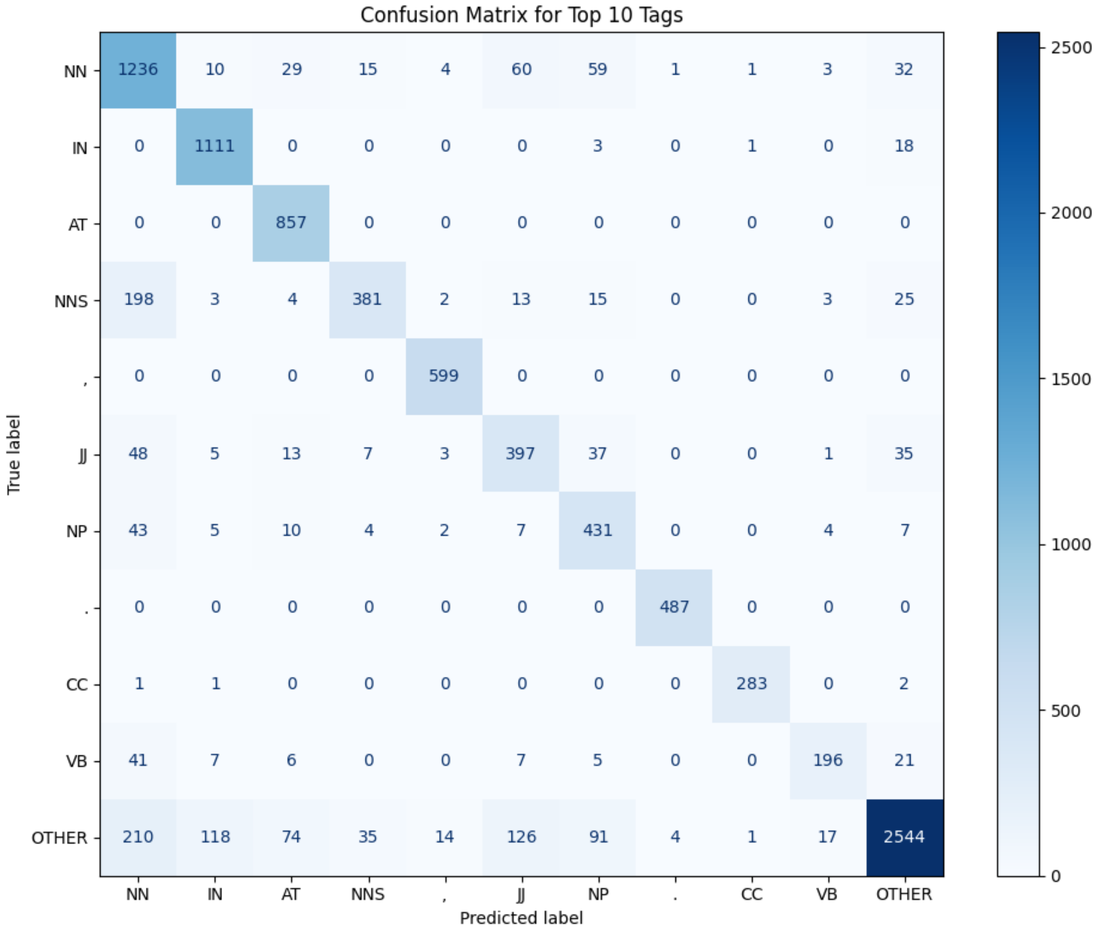
The HMM Tagger without smoothing achieves the best overall performance with the lowest total error rate (12.52%) and performs well for both known and unknown words.

For unknown words, using pseudo-words with maximum likelihood estimation achieves the best error rate (51.60%), highlighting the effectiveness of pseudo-words in handling rare or unseen words.

Add-One smoothing improves unknown word tagging compared to the baseline but increases known word errors due to over-regularization, resulting in higher total error rates.

Overall, the combination of pseudo-words and maximum likelihood estimation strikes a balance between improving unknown word tagging and maintaining competitive performance on known words.

Confusion Matrix:



Confusion Matrix for Top 10 Tags

True label \ Predicted label	NN	IN	AT	NNS	,	JJ	NP	.	CC	VB	OTHER
NN	1236	10	29	15	4	60	59	1	1	3	32
IN	0	1111	0	0	0	0	3	0	1	0	18
AT	0	0	857	0	0	0	0	0	0	0	0
NNS	198	3	4	381	2	13	15	0	0	3	25
,	0	0	0	0	599	0	0	0	0	0	0
JJ	48	5	13	7	3	397	37	0	0	1	35
NP	43	5	10	4	2	7	431	0	0	4	7
.	0	0	0	0	0	0	0	487	0	0	0
CC	1	1	0	0	0	0	0	0	283	0	2
VB	41	7	6	0	0	7	5	0	0	196	21
OTHER	210	118	74	35	14	126	91	4	1	17	2544

The confusion matrix shows that the model performs well overall, particularly for frequent and distinct tags like "NN," "AT," and "IN."

However, it struggles with ambiguities between similar tags, such as singular/plural nouns ("NN" vs. "NNS") and nouns/adjectives ("NN" vs. "JJ").

Additionally, a significant number of tokens in "OTHER" indicate challenges with rare or unseen tags, despite the use of pseudo-words and smoothing.