Sanjana Sunil Agarwal

ASSIGNMENT #5

1 a) To decrypt the message we need to decrypt each alphabet since each alphabet here has a different encryption code.
We can have up to 26 states or more if we consider punctuations, and the white space.
The observed units will be the letters in the encoded word, that is, 'ucpcpc'
The hidden states are the letters in the actual word, that is, 'banana'
So we are trying to formulate an HMM problem where we have to come up with the best state sequence given the observed input.
The emissions that are framed are: p(encryption|English letter) where each letter has a particular encryption.
For a well-trained HMM, the emission probability(alphabet probability) should have one probability that is close to 1, of the word that is its true encryption and the others nearly close to 0.

So, by Cave and Neuwirth's experiment, we know that they experimented with different number of sets and found the optimal solution.
They had 12 sets and 8 classifiers, namely,
V - Vowel
SP - Space
C - Consonant
FL - First letter
LL - Last letter
VF - Vowel follower
VP - Vowel procceeder
CP - Consonant follower

This will give a simpler and more accurate network. We can also divide the sets into vowels and consonants for further simplicity.

1 b) Like I said above, the emission probabilities should be distributed such that the probability of the true encryption letter should be close to 1 and the remaining ones should be close to 0.
We can also arrange it based on the order of the frequency of the letters.
For the 26 states of alphabets, we can normalize the count to frequency and calculate the corresponding emission probabilities.
Using the Cornell English frequency model, which uses 40K words, we get,
https://www.math.cornell.edu/~mec/2003-2004/cryptography/subs/frequencies.html

The emission probabilities are nothing but the frequency value divided by 100.

	E	T	A	O	I	N	S	R	H	D	L	U	C	M	F	Y	W	G	P	B	V	K
Frequency	12.02	9.10	8.12	7.68	7.31	6.95	6.28	6.02	5.92	4.32	3.98	2.88	2.71	2.61	2.30	2.11	2.09	2.03	1.82	1.49	1.11	0.69

X	Q	J	Z
0.17	0.11	0.10	0.07

These values are calculated for a unigram model. Likewise, for a bigram model, we calculate frequencies of two letters appearing together, get the frequency chart and normalize the values.

1 c) Yes there are certain advantages and disadvantages to using the trigram model. Trigram model would definitely give better results but it will come at the cost of increased run time. For estimating the parameters, it would help as we would have b=more data (or context) and we will obtain a better evidence for one letter substituting another. However, if there is a strong probability of one translation, then more context will not really help. Thus we need to know our data well to decide which model should be used - bigram or trigram.

2) Initial vector = [0.45 0.35 0.15 0.05]

Transition matrix =

t / t+1	DT	JJ	NN	VB
DT	0.03	0.42	0.50	0.05
JJ	0.01	0.25	0.65	0.09
NN	0.07	0.03	0.15	0.75
VB	0.30	0.25	0.15	0.30

Emission matrix =

w / s	DT	JJ	NN	VB
a	0.85	0.05	0.03	0.05
myth	0.01	0.10	0.45	0.10
is	0.02	0.02	0.02	0.60
female	0.01	0.60	0.25	0.05
moth	0.12	0.13	0.25	0.20

Given sequence: A myth is a female moth

$\alpha_i = \alpha_{ik_1} * v_i$ for i = 1 to n

$\alpha(DT) = 0.45 * 0.85 = 0.3825$

$\alpha(JJ) = 0.35 * 0.05 = 0.0175$

$\alpha(NN) = 0.15 * 0.03 = 0.0045$

$\alpha(VB) = 0.05 * 0.05 = 0.0025$

$$\alpha_2(DT) = \sum_{i=1}^{n} \alpha_1(i) p(i, DT) * a_{DT,k_2}$$

$$\alpha_2(DT) = \alpha_1(DT) * p(DT, DT) * p(myth|DT) + \alpha_1(JJ) * p(JJ, DT) * p(myth|DT) + \alpha_1(NN) * p(NN, DT) * p(myth|DT) + \alpha_1(VB) * p(VB, DT) * p(myth|DT)$$

$$\alpha_2(DT) = (0.3825 * 0.03 * 0.01) + (0.0175 * 0.01 * 0.01) + (0.0045 * 0.07 * 0.01) + (0.0025 * 0.3 * 0.01)$$

= 0.00011475 + 0.00000175 + 0.00000315 + 0.0000075
= **0.00012715**

$$\alpha_2(JJ) = (0.3825 * 0.42 * 0.1) + (0.0175 * 0.25 * 0.1) + (0.0045 * 0.03 * 0.1) + (0.0025 * 0.25 * 0.1)$$

= 0.016065 + 0.0004375 + 0.0000135 + 0.0000625
= **0.0165785**

$$\alpha_2(NN) = (0.3825 * 0.5 * 0.45) + (0.0175 * 0.65 * 0.45) + (0.0045 * 0.15 * 0.45) + (0.0025 * 0.15 * 0.45)$$

= 0.0860625 + 0.00511875 + 0.00030375 + 0.00016875
= **0.09165375**

$$\alpha_2(VB) = (0.3825 * 0.05 * 0.1) + (0.0175 * 0.09 * 0.1) + (0.0045 * 0.75 * 0.1) + (0.0025 * 0.3 * 0.1)$$

= 0.0019125 + 0.0001575 + 0.0003375 + 0.000075
= **0.0024825**

$$\alpha_3(DT) = \alpha_2(DT) * p(DT, DT) * p(is|DT) + \alpha_2(JJ) * p(JJ, DT) * p(is|DT) + \alpha_2(NN) * p(NN, DT) * p(is|DT) + \alpha_2(VB) * p(VB, DT) * p(is|DT)$$

$$\alpha_3(DT) = ((0.00012715 * 0.03) + (0.0165785 * 0.01) + (0.09165375 * 0.07) + (0.0024825 * 0.3)) * 0.02$$

= **0.00014660224**

$$\alpha_3(JJ) = ((0.00012715 * 0.42) + (0.0165785 * 0.25) + (0.09165375 * 0.03) + (0.0024825 * 0.25)) * 0.02$$

= **0.00015136531**

$$\alpha_3(NN) = ((0.00012715 * 0.5) + (0.0165785 * 0.65) + (0.09165375 * 0.15) + (0.0024825 * 0.15)) * 0.02$$

= **0.00049920075**

$$\alpha_3(VB) = ((0.00012715 * 0.05) + (0.0165785 * 0.09) + (0.09165375 * 0.75) + (0.0024825 * 0.3)) * 0.02$$

= **0.042590091**

$$\alpha_4(NN) = \alpha_3(DT) * p(DT, DT) * p(a|NN) + \alpha_3(JJ) * p(JJ, DT) * p(a|NN) + \alpha_3(NN) * p(NN, DT) * p(a|NN) + \alpha_3(VB) * p(VB, DT) * p(a|NN)$$

$\alpha_4(NN) = ((0.00014660224 * 0.5) + (0.00015136531 * 0.65) + (0.00049920075 * 0.15) + (0.042590091 * 0.15)) * 0.03$

= 0.000002199036 + 0.000002951623545 + 0.000002246403375 + 0.0001916554095

= **0.0001990524724**

Backward Probability:

$\beta_i(t+1) = 1$ for 1≤ i ≤n

$$\beta_i(t) = \sum_{j=1}^{n} a_{ij} b_{ijO_t} \beta_j(t+1)$$

1≤ t ≤ T, 1≤ i ≤ N

Total:

$$P(O|U) = \sum_{i=1}^{N} \pi_i \beta_i(1)$$

a	myth	is	a	female	moth
(1)	(2)	(3)	(4)	(5)	(6)

$\beta_7(DT) = 1$

$\beta_7(JJ) = 1$

$\beta_7(NN) = 1$

$\beta_7(VB) = 1$

$\beta_6(DT) = 0.12$

$\beta_6(JJ) = 0.13$

$\beta_6(NN) = 0.25$

$\beta_6(VB) = 0.2$

$\beta_5(DT) = \beta_6(DT) * p(DT, DT) * p(female|DT) + \beta_6(JJ) * p(JJ, DT) * p(female|DT) + \beta_6(NN) * p(NN, DT) * p(female|DT) + \beta_6(VB) * p(VB, DT) * p(female|DT)$

$\beta_5(DT) = ((0.12 * 0.03) + (0.13 * 0.42) + (0.25 * 0.5) + (0.2 * 0.05)) * 0.01$

= **0.001932**

$\beta_5(JJ) = ((0.12 * 0.01) + (0.13 * 0.25) + (0.25 * 0.65) + (0.2 * 0.09)) * 0.6$

= **0.12852**

$$\beta_5(NN) = ((0.12 * 0.07) + (0.13 * 0.03) + (0.25 * 0.15) + (0.2 * 0.75)) * 0.25$$

= **0.04995**

$$\beta_5(VB) = ((0.12 * 0.30) + (0.13 * 0.25) + (0.25 * 0.15) + (0.2 * 0.30)) * 0.05$$

= **0.0083**

$$\beta_4(DT) = \beta_5(DT) * p(DT, DT) * p(a|DT) + \beta_5(JJ) * p(JJ, DT) * p(a|DT) + \beta_5(NN) * p(NN, DT) * p(a|DT) + \beta_5(VB) * p(VB, DT) * p(a|DT)$$

$$\beta_4(DT) = ((0.001932 * 0.03) + (0.12852 * 0.42) + (0.04995 * 0.5) + (0.0083 * 0.05)) * 0.85$$

= **0.067512406**

$$\beta_4(JJ) = ((0.001932 * 0.01) + (0.12852 * 0.25) + (0.04995 * 0.65) + (0.0083 * 0.09)) * 0.05$$

= **0.003268191**

$$\beta_4(NN) = ((0.001932 * 0.07) + (0.12852 * 0.03) + (0.04995 * 0.15) + (0.0083 * 0.75)) * 0.03$$

= **0.0005312502**

$$\beta_4(VB) = ((0.001932 * 0.30) + (0.12852 * 0.25) + (0.04995 * 0.15) + (0.0083 * 0.30)) * 0.05$$

= **0.002134605**

$$\beta_3(DT) = \beta_4(DT) * p(DT, DT) * p(is|DT) + \beta_4(JJ) * p(JJ, DT) * p(is|DT) + \beta_4(NN) * p(NN, DT) * p(is|DT) + \beta_4(VB) * p(VB, DT) * p(is|DT)$$

$$\beta_3(DT) = ((0.067512406 * 0.03) + (0.003268191 * 0.42) + (0.0005312502 * 0.50) + (0.002134605 * 0.05)) * 0.02$$

= **0.000075407355**

$$\beta_3(JJ) = ((0.067512406 * 0.01) + (0.003268191 * 0.25) + (0.0005312502 * 0.65) + (0.002134605 * 0.09)) * 0.02$$

= **0.0000405919778**

$$\beta_3(NN) = ((0.067512406 * 0.07) + (0.003268191 * 0.03) + (0.0005312502 * 0.15) + (0.002134605 * 0.75)) * 0.02$$

= **0.0001300911086**

$$\beta_3(VB) = ((0.067512406 * 0.30) + (0.003268191 * 0.25) + (0.0005312502 * 0.15) + (0.002134605 * 0.30)) * 0.60$$

= **0.013074503148**

$$\beta_2(NN) = \beta_3(DT) * p(DT, DT) * p(myth|NN) + \beta_3(JJ) * p(JJ, DT) * p(myth|NN) + \beta_3(NN) * p(NN, DT) * p(myth|NN) + \beta_3(VB) * p(VB, DT) * p(myth|NN)$$

$$\beta_2(NN) = ((0.000075407355 * 0.07) + (0.0000405919778 * 0.03) + (0.0001300911086 * 0.15) + (0.013074503148 * 0.75)) * 0.45$$

= **0.0044243492856633**

Thus, to summarize the final answers of the asked questions:

a) $\alpha_4(NN)$ = **0.0001990524724**
b) $\alpha_3(VB)$ = **0.042590091**
c) $\alpha_1(DT)$ = **0.3825**
d) $\beta_4(NN)$ = **0.0005312502**
e) $\beta_2(NN)$ = **0.0044243492856633**