Cairo University
Faculty of Engineering
Computer Engineering Dept.

Sheet 2: Language Models

1) Write out the equation for trigram probability estimation. Then, write out all the non-zero trigram probabilities for the corpus:
I am Sam. Sam I am. I do not like green eggs and ham.

2) Calculate the probability of the sentence "I want Chinese food"
   a. using the given bigram probabilities.
   b. using the given add-1 smoothed bigram probabilities.
   c. Which of the two probabilities you computed is higher, unsmoothed or smoothed? Explain why.

Bigram Probabilities:

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

P(i|<s>)=0.25 and P(</s>|food)=0.68

Add-1 Smoothed Bigram Probabilities:

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

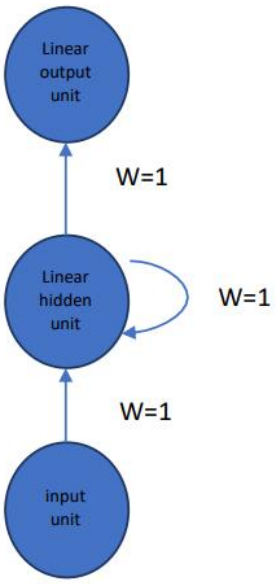
P(i|<s>)=0.19 and P(</s>|food)=0.40

3) Suppose we didn't use the end-symbol </s>. Train an unsmoothed bigram grammar on the following training corpus without using the end-symbol </s >
<s> a b
<s> b b
<s> b a
<s> a a
Demonstrate that your bigram model does not assign a single probability distribution across all sentence lengths by showing that the sum of the probability of the four possible 2 word sentences over the alphabet {a,b} is 1.0, and the sum of the probability of all possible 3 word sentences over the alphabet {a,b} is also 1.0.

4) We are given the following corpus:
<s> I am Sam </s>
<s> Sam I am </s>
<s> I am Sam </s>
<s> I do not like green eggs and Sam </s>
If we use linear interpolation smoothing between a maximum-likelihood bigram model and a maximum-likelihood unigram model with $\lambda_1 = ½$ and $\lambda_2 = ½$, what is P(Sam|am)? Include <s> and </s> in your counts just like any other token.

5) You are given a training set of 100 numbers that consists of 91 zeros and 1 each of the other digits 1-9. Now we see the following test set: 0 0 0 0 0 3 0 0 0 0. What is the unigram perplexity?

6) For the following RNN, we have the following input sequence:
x(t=0)=2, x(t=1)=-0.5; x(t=2)=1.
   a. Write the equations and compute all the network values.
   b. Draw the "unrolled" network.
   c. What is this RNN learning?



7) Given an RNN character-level language model, assume the very small vocabulary {‘h’,‘e’,‘l’,‘o’} and tokens are single letters represented in the input with a one-hot encoded vector (note that in practice instead of one-hot encoded vectors we will have word embeddings). The model with the numbers is shown in the following figure.
   a. Compute the final output using the **softmax** function.
   b. For each output mention what the model predicts and whether the model did a correct prediction or not specifying what the correct prediction should be.