A6 – Part I

Writing assignments – Assignments should be typed with a 12-point font such as Times New Roman and should be double spaced.

Write a lab report that describes the results of your experiments on the TIMIT corpus. Guidelines for writing scientific reports can be found on the assignments page. If you wish to perform additional experiments to the ones that you did this week, you may do so, but do not spend more than a day or two doing so (resubmit your code and indicate EXPERIMENT REVISIONS next to the affidavit in your driver so that I know not to count the revisions as late). You can of course show differences in your results based on architectural differences (e.g., depth, width, regularization) or feature inputs, but you should think some about the types of errors that your system makes. Does it tend to confuse phonemes that are similar to one another (e.g., plosives that differ in voice or place of articulation)? It would be interesting (but not required) to see the impact on your phoneme error rate if on phoneme boundaries you considered either phoneme correct.

Your lab report should not exceed 5-6 pages.

Part II –Questions are 20 points each.

1. Estimate the bigram probabilities using a frequentist model (relative counts, slides 16 & 24, not using discounted probability) for the following quote from Geisel's (aka Dr. Seuss, 1960) *Green Eggs and Ham* which has been prepared for analysis.:

   <s> i would not like them here or there </s>
   <s> i would not like them anywhere </s>
   <s> i do not like green eggs and ham </s>
   <s> i do not like them sam i am </s>

   Show tables with your unigram and bigram counts in addition to the bigram probabilities.

2. For the same text, compute histograms showing c vs. N_c for unigrams (e.g. slide 35). Then compute Good-Turing estimates of c* (you need not worry about any type of smoothing operation as you will be reading about in your assignment below). For this exercise, do not include <s> and </s> in your unigram counts.

3. Estimate the discounted probabilities for a Katz bigram model where C(like, y) > 0. As we are working with a very small corpus, our estimates are likely to be poor and we will rely on some estimates of smoothed values proposed by Church and Gale (Figure 3.9 of the Jurafsky & Martin text, repeated her for your convenience):

Bigram count in training set	Bigram count in heldout set
0	0.0000270
1	0.448
2	1.25
3	2.24
4	3.23
5	4.21
6	5.23
7	6.21
8	7.21
9	8.26

   Using these counts for your revised c*, find the discounted probability for the prediction words of phrases given the context word "like." What is the leftover probability?
4. Each input in a transformer is transformed by the dot product of a value weight vector W^V with each input. These inputs are scaled by a distribution of similarities between the current feature vector associated with the output all other feature vectors in the block. Explain the role of the query and key vectors and how this leads to a self-attention mechanism.
5. In class, we discussed the difference between causal and non-causal transformer networks. For a company's virtual assistant that routes people to an appropriate agent based on a prompt "Welcome to NewTech, how may I help you?" would we be better off using a causal or non-causal system. Justify your answer.
6. Block length is a limitation for transformer architectures. Why is this the case?

Geisel, T. S. (1960). Green eggs and ham. New York,: Beginner Books; distributed by Random House.