Luna McBride

NLP Homework 1 Part 2

1) What is the conditional probability of a word occurring given its previous word and how does this probability change when you use laplace smoothing?

The conditional probability of a word occurring given its previous word is the number of times the two occur together over the number of times the previous word is the first word in the bigram in this scenario. It is the probability that if the previous word shows up, the specified word will follow. This probability is changed in laplace smoothing by adding 1 over the number of values it can take on, in this case the size of the vocabulary. This method adds the assumption that this one word out of the whole vocabulary was seen more often than was actually seen, hence the one over vocabulary being added. This assumption has a relatively small impact given its use of the whole vocabulary in the denominator, thus the change is very small for larger vocabularies. The trade-off is seen in smaller vocabularies, where the small subset increases the probability significantly.

2) What do you expect will happen when you use trigrams instead of bigrams to predict the next word in a sentence? Does the perplexity increase or decrease? Why?

Trigrams will likely follow the idea of rays when it comes to the probability, changing the probability based on what is further down the line rather than just the next word in the sequence. An example would be with the sentences “I like to walk my dog”, “I like to walk to Starbucks”, and “I like to bus to the store”. Looking two forward from the “to” in “I like to”, the bigram approach would see walk as more likely next since two of the sentences have walk after to. The trigram approach may make different considerations here, given that “to __ to” is also just as common.

3) What do you expect will happen when you use a larger corpus to predict the next word in a sentence?

Not much will happen differently besides increased processing time. The greedy approach to the algorithm creates a loop where the same input will result in the same output, which was shown by running the brown corpus a couple times with the same prompt of only “<s>”. This approach has no chance of learning beyond what has the best probability, so the size of the corpus will not change anything besides the immediate first answer. This could be at least partially fixed with a random chance to give a random next word (like in Q-Learning) or having a memory gate and user input to artificially add more/change probability bias based on conversation (like with ChatGPT).