Nick Johnson
CS 4395.001
N-Grams

Narrative

a. What are n-grams and how are they used to build a language model?

   N-grams are sequences of n words taken from a certain part of a text. They can be used to predict how likely a certain word is to follow another word.

b. List a few applications where n-grams could be used.

   N-grams can be used in things like predictive text, sentiment analysis, and spell check.

c. Describe how probabilities are calculated for unigrams and bigrams

   For unigrams, probability is calculated by taking the number of times the unigram appears and dividing it by the number of unigrams in the source text. For bigrams, probability is calculated by taking the probability of the first word appearing and multiplying it by the count of the bigram divided by the count of the first word. So for bigrams, this looks like $P(w1, w2) = P(w1)P(w2|w1)$.

d. Describe the importance of the source text in building a language model.

   As seen in the previous answer, the source text is used to calculate probabilities for unigrams and bigrams. As the source text gets bigger, you'll start seeing more accurate probabilities since your sample size will be bigger.

e. Describe the importance of smoothing as well as a simple approach to smoothing.

   Smoothing is used to help make distributions smoother by assigning non-zero probabilities to words that haven't been seen yet. A simple approach to smoothing is known as Laplace smoothing. This is where each word gets 1 added to its count. This approach does not perform very well since this is too aggressive of a probability adjustment.

f. Describe how language models can be used for text generation, and the limitations of this approach.

   You can use language models to create probability dictionaries of n-grams. To generate text, the probabilities are used by finding the bigram with the start word in the first position that has the highest probability. This will keep happening until a period is reached and the sentence is formed. The limitation here is that context is not considered when doing this text generation. Mathematically the sentence(s) will be correct in terms of probabilities but will not necessarily be coherent.

g. Describe how language models can be evaluated.

Extrinsic evaluation is when human annotators evaluate the text that is generated by using some sort of predefined metric. This kind of evaluation is only occasionally used since it is both expensive and time-consuming. Intrinsic evaluation is when an internal metric is used such as perplexity to compare models.

h. Give a quick introduction to Google's n-gram viewer and show an example.

Google's n-gram viewer allows users to enter phrases and see how often those phrases have appeared in a chosen corpus of books over a certain period of time.