CS329 hw7

Alisa Starr

- Explain why it is difficult to measure the joint probability of a long word sequence.
    - There are a near infinite number of possible combinations/sequence of words which can create different possibilities of sentences. Because of this, it is not possible to create a large enough corpus of data to compute the statistics for all of these long prefixes and sentences. This means that when we are trying to compute the joint probabilities for a long word sequence, we will often times not be able to at all, because there are no occurrences of that specific ordering of words within the corpus. This leads to problems, like sparse probabilities in a chain rule calculation for long sentences.
- Explain the intuition behind the Markov assumption and why it is useful for language modeling.
    - Many sentences can be very long, and a long word sequence has too many conditional probabilities to estimate. Thus use the Markov Assumption to closely approximate chain rule maximum likelihood. For Markov Assumption we assume that the probability of a word is only dependent on a small number of previous word(s) (so not dependent on the entire prefix as is assumed by chain rule). This makes probabilities for long word sequences much easier to estimate. Reduces amount of work and avoids the problem of lack of data on long sequences.
- Explain why it is useful to apply the chain rule to estimate maximum likelihood.
    - Chain rule is used to decompose the (complicated) joint probability of a sequence of words into a product of conditional probabilities. It is very straight forward and with Markov's assumption it gives reliable results even for long word sequences.
- Describe two advantages of using log likelihood instead to estimate maximum

likelihood.

- The chain rule, even with Markov assumption, is very costly due to the many costly multiplication computations. The log likelihood is much more efficient for the computer to do. This is because the log of a product of probabilities is equal to the summation of the logs of the probabilities, so we can do addition calculations which are less costly than multiplication.
- Another issue with the chain rule is that each probability is a fractional value <=1 so as these values are multiplied, the final product approaches zero. For many sequences (especially long sequences), the chain rule product will have so many decimal places that the computer will have rounding/display issues. The log likelihood outputs values always greater than 0, so those issues are solved. Also, log is a monotonically increasing function

- Explain how Laplace smoothing is applied to estimate the unigram probability of unknown words.
    - If a word x1 is unknown (it does not appear in the dataset/corpus) then in a bag of words model (where we are counting the occurrence of each word in the corpus) the count of x1 is equal to 0, thus P(x1)=0. When P(x1)=0, the chain rule will equal 0. Even for log likelihood this is a problem because log(0) is undefined. To prevent these issues, we use Laplace smoothing. In Laplace smoothing for estimating unigram probability, the smoothing parameter alpha (nonzero, usually 1) is added to all the counts. Assuming that alpha=1, Laplace smoothing ensures that even when C(x1) = 0, we will have P(x1) = (0+1)/(N+1|X|) = 1/(N+|X|) != 0. Thus, we can more accurately calculate the maximum likelihood when dealing with unknown words.