Alexis Jennings
CS 4395.001
N-Grams Narrative

1. What are N-grams and how are they used to build a language model?
    a. N-grams are sliding windows of size N over text. They help build a probabilistic model of language, in which we can determine the probability of a sequence of N words.
2. List a few applications where N-grams could be used.
    a. N-grams can be used in applications that deal with speech recognition and machine translation.
3. Description of how probabilities are calculated for unigrams and bigrams:
    a. For unigrams, the equation is the number of occurrences of a token in a text divided by the total number of tokens. For bigrams, the equation is the probability of the first word as a unigram multiplied by the probability of the second word given the first word.
4. The importance of the source text in building a language model:
    a. The larger and more complex the source text, the better the language model can become. A simple corpus may only provide a fraction of the possibilities of the language, so more data is needed to ensure every possibility and to provide the most accurate model possible.
5. The importance of smoothing, and describe a simple approach to smoothing:
    a. N-grams from a real text may pose a sparsity problem where not every possible N-gram will be in the N-gram dictionary. Smoothing fills in the zero values with a bit of probability mass to smooth out the curve. A simplified approach to smoothing is the Good-Turing smoothing method, in which zero counts and probabilities are replaced with counts and probabilities of words that occur only once.
6. Describe how language models can be used for text generation, and the limitations of this approach.
    a. Language models can be used for text generation by giving a starting word and then iteratively finding the most likely word to come after the previous word, using the probabilities given by the N-grams. The limitations are that for a better model, larger values of N in the N-grams would be needed, as well as a large amount of source text.
7. Describe how language models can be evaluated:
    a. We can evaluate the quality of a language model by using an extrinsic method such as perplexity. Perplexity is the inverse probability of seeing the words we observe normalized by the number of words. The lower the perplexity, the more ideal the quality of the language model is.
8. Give a quick introduction to Google's N-gram viewer and show an example:
    a. Google's N-gram viewer is a large dictionary of N-grams found in Google Books' library. On the website, anyone can type in an N-gram and view its number of

occurrences in the Google Books library over time. Here is an example using three video game titles, “Ace Attorney,” “Animal Crossing,” and “Fire Emblem”: