Shane Arwood
October 2, 2022
Dr. Karen Mazidi
CS 4395.001

CS 4395: N–Grams Narrative

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

An n-gram is a portion of a piece of text that is n words long. For example, a unigram in a text has one word, and trigram has three words. They “slide” over the text and can be used to calculate the maximum likelihood of a certain word occurring based on the ones before it N-grams can then be used to form a language model using the probability of a n-gram in a given sequence of words.

b. Applications where n-grams could be used

N-grams are used to predict certain words occurring in a certain part of a text, so they could be used for sentence completion or to auto generate text that is similar to the training corpus from which the n-gram model was created. It can also be used to identify the language of a given text. It seems that n-grams can be useful for analytical as well as predictive purposes.

c. How are probabilities calculated for unigrams and bigrams?

Probabilities are calculated using the counts of the unigrams as well as the total number of tokens in the text. For unigrams, the count of a particular unigram is divided by the total number of tokens. For bigrams, the probability of the first word in the bigram is multiplied by the probability of either that word or the second word in the bigram.

d. The importance of the source text when building a language model

When building a language model, the source text is vital for the generation of a significant language model. Different source texts will change the language model. In particular, a large source text will generate a better language model than a small source text, as there will be more variations in vocabulary and patterns.

e. The importance of smoothing and a simple approach

Probabilities are calculated with multiplication, which causes issues when there are probabilities that equal zero. To mitigate this, smoothing is used. It gives values that would otherwise be 0 and a very small positive value to ensure a calculation isn’t incorrectly zeroed out. A simple approach to smoothing is LaPlace smoothing, where 1 is added to the count of an n-gram and the total number of vocabulary words is added to the number of n-grams. This ensures that the probability calculated will never be 0; the add-1 cannot result in a result of 0.

f. How can language models be used for text generation (+ limitations)?

Language models can be used for text generation by first creating probability dictionaries from the source text.These dictionaries can then be used to predict the next most likely word for a given word, then the process can continue until the end of a sentence. A limitation of this

approach is that it doesn't take grammar rules into account and might not make a coherent sentence, as the generation of the next word is based solely on the word immediately before it.

g. How can language models be evaluated?

There are two main ways to evaluate language models: intrinsic and extrinsic. Intrinsic evaluation is through a specific metric of the model, such as perplexity. The value of the metrics can be used to determine how "good" the language model is; for example, a low perplexity indicates a better model. Extrinsic evaluation can also be used, and rather than metrics, this method of evaluation relies on humans annotating the model to actually determine its effectiveness.

h. Google's N-Gram viewer, and an example

Google's n-gram viewer is an online tool used to visualize the occurrence of different words in the Google Books corpora over time. It uses n-grams to calculate the percentage of books containing the given n-gram. An example of the occurrences of book, computer, and newspaper is shown below.