Neal Gandhi, Nina Rao
CS4395.001

N-Grams Narrative

N-grams is a popularly used statistical language modeling technique in NLP that is applied to a wide range of uses, such as sentiment analysis, speech recognition, spelling correction, and the most important, language prediction. To create an n-gram model you must examine a large dataset of text and calculate the frequency of each sequence of n-number of words. This means that a unigram counts the frequency of 1 word sequences but a trigram counts the frequency of 3 word sequences

As mentioned, one of the foremost applications of n-gram modeling is to predict the next word in a given phrase or sentence. To achieve this, the model must calculate the probability of each of the possible “next words”, and chooses the word with the highest probability of being the next word. This approach can be limiting, as it’s only being applied to text from the training data, and will likely make mistakes when handling unseen word combinations that were not present in the initial dataset.

Determining the probabilities for a unigram (where n is equal to 1) is fairly straightforward, as all you must do is divide the frequency count of each of the individual words divided by the total number of words in the dataset. Bigrams and other higher order n-grams have a more complex method of calculating probabilities, one of which is the maximum likelihood estimation. This involves dividing the frequency count of each n-gram by the frequency count of the previous n-1-grams. This can lead to the model being too dependent on just the training dataset, making it incompatible with other datasets or having it return incorrect estimations.

The selection of the source text used for building the n-gram model is crucial in determining the model's accuracy. If the corpus used to construct the model is not representative of the text being analyzed, the language model may fail to predict the likelihood of a given sequence of words with precision. To combat the limitations of n-gram models, smoothing techniques come into play to adjust the probabilities of rare n-grams. One such approach to smoothing is the utilization of add-k smoothing, wherein a small constant value (k) is added to

the frequency count of each n-gram in the corpus. This approach minimizes the impact of rare n-grams in the model, thus helping to prevent overfitting.

N-gram models may be evaluated using metrics such as perplexity, which gauges the model's ability to predict a test set of text with accuracy. Other metrics, such as precision, recall, and F1 score, may be employed to evaluate the performance of n-gram models in specific applications. Overall, n-gram models have found significant success in a diverse range of NLP applications, with current research aimed at addressing the limitations of n-gram models and developing more advanced language models capable of better capturing the intricate patterns and dependencies in natural language.

One prominent application of N-grams is Google's n-gram viewer, which is an online tool that provides people with the ability to research and understand the overall frequency of words and phrases against a large amount of data. This dataset incorporates books, articles, research papers, and many other documents all in varying languages and from different time periods. Using the tool, you can input a word or a phrase and generate a custom graph showing the frequency and usage of those words over a time span. An example for this is the phrase "climate change", which reveals that between 1900 and 2008, the phrase had a drastic spike in usage, revealing the increase in frequency. Google's n-gram viewer is an amazing tool that has been provided to researchers and linguists which can help to provide insights on how language has evolved over the course of time.