Narrative: (N-Grams)

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

A contiguous sequence of n items from a given sample of text or audio is known as "N-grams." Depending on the application, the elements may be phonemes, syllables, letters, words, or base pairs. The n-grams are typically gathered from a corpus of spoken or written language. In any set of words in the language, an N-gram language model forecasts the likelihood of a specific N-gram occurring. If our N-gram model is accurate, we may forecast p(w | h), which is the likelihood that a word will appear given a history of words h, where the history consists n-1 words. For the purpose of building an N-gram model, we must calculate this probability. This probability is calculated twice, In order to easily determine p(w1...ws), we must first use the chain rule of probability and then a very strong simplifying assumption.

List a few applications where n-grams could be used:

N-grams find use in several areas of computer science, computational linguistics, and applied mathematics. They have been used to: Create learning kernels for machine learning techniques like support vector machines. Find potential candidates for a word's right spelling. Enhance compression in algorithms where a little amount of data calls for n-grams with longer lengths. In pattern recognition systems, speech recognition, OCR (optical character recognition), Intelligent Character Recognition (ICR), machine translation, and similar applications, determine the likelihood that a particular word sequence would appear in text of a language of interest. Enhance retrieval in information retrieval systems when it is hoped to discover related "documents" given a single query document and a database of reference documents (a term whose conventional meaning is sometimes stretched, depending on the data set). Enhance retrieval performance for genetic sequence analysis systems like the BLAST family. Find out what language a text is written in or what species a little DNA sequence comes from. As in the detached press algorithm, predict letters or words at random to make text. And in many more cases.

A description of how probabilities are calculated for unigrams and bigrams:

The following assumptions are made by the unigram language model: each word's probability is unrelated to those that come before it. Instead, all that matters is how frequently this word appears overall in the training text. By dividing the total number of times, the word "prime" appears in the corpus by the number of times the string "prime minister" appears in the corpus, one may get the bigram likelihood.

The importance of the source text in building a language model

By examining text data, language models can calculate word probability. For language models to learn representations that consistently encode most semantic characteristics, only 10–100 million words are needed. In order to learn enough common-sense information to master typical downstream NLU tasks, a higher amount of data is required. If an adequate amount of unbiased

and clean source text is not available, that would hamper the overall accuracy of a language model.

The importance of smoothing, and describe a simple approach to smoothing:

Data smoothing is based on the notion that it can recognize simpler changes to assist in the prediction of various trends and patterns. It serves as a tool for statisticians or traders who must examine a large amount of data—which is frequently difficult to comprehend—in order to spot patterns, they might not otherwise see. Rectangular smoothing, often known as "unweighted sliding-average smooth," is the most basic smoothing algorithm. The "smooth width" of this approach is a positive integer, and it replaces each point in the signal with the average of "m" nearby points. M is typically an odd number.

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

Approach:

How can we produce text using a language model? The technique is iterative: choose a word based on the previous words in the sequence, add that word, and repeat. So, all we need to know is how to choose the following word. Several tactics exist, including:

Summarize, more likely to be chosen are words that suit the conditional word probability distribution better. We would choose the words "toys" with probability 14%, "aardvark" with probability 0.001%, and so on for the aforementioned case. Greedy, always choose the term with the greatest probability (akaargmax). Choose "food." Beam Search, the outcome with the highest overall probability is not always the one that results from the greedy strategy. In order to prevent being misled by local maxima, a beam search keeps track of numerous likely variants at each stage. Choose "food" and "toys," then decide which is superior after the addition of more words.

Describe how language models can be evaluated:

Perplexity, cross entropy, and bits-per-character are the traditional metrics used to assess the performance of language models (BPC). Since language models are frequently used as pre-trained models for other NLP tasks, their effectiveness on downstream tasks is frequently taken into account when evaluating them.

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

The great majority of public domain books that Google scanned to create its Google Books search engine are the source of the text to be studied in the instance of the Google Books n-gram Viewer. The body of text you will search for using Google Books n-gram Viewer is referred to by Google as the corpus. Although you may independently study British and American English or combine them, the n-gram Viewer aggregates by language. For example, any phrase or words that you want to evaluate should be typed. Use commas to separate each phrase. To get you started, Google proposes "Albert Einstein, Sherlock Holmes, and Frankenstein." Choose a range of dates. The range by default is 1800–2000. Select a corpus. You may see items like "English (2009)" or "American English (2009)" near the bottom of the list while searching for texts in English or other languages in addition to the usual options. Although Google has already updated

these earlier corpora, there may be some justification for you to compare your results to previous data sets. Most people may disregard these and concentrate on the newest corpora. The smoothing level is set. The term "smoothing" describes the final smoothness of the graph. Although smoothing level 0 is the most realistic portrayal, it may be challenging to read. The initial setting is 3. You often don't need to make any adjustments. Press “Search lots of books”. You may go further into the information using Google's n-gram Viewer. Using tags will allow you to search for the verb fish rather than the noun fish. In this instance, you would look up fish_VERB.