N-Grams Reflection

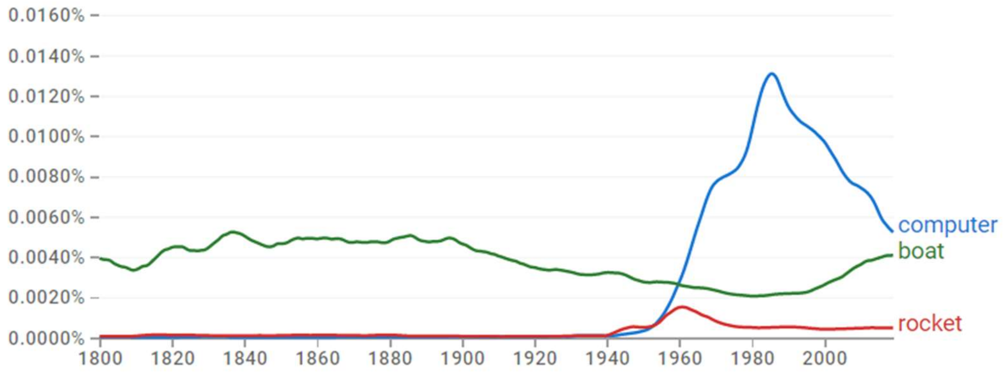
N-Grams are a type of probabilistic model that is trained on a given body of text. It specifically pays attention to what words occur in variable-length sequences. Using the given frequencies, it can make informed estimates of word appearances, especially in the context of the surrounding text. Therefore, it gradually "learns" a language based on how the training data uses the language. This approach can be applied to language recognition (as demonstrated), predictive speech, and speech recognition.

Probabilities of unigrams and bigrams can be found by counting the frequency of their appearance in the training data. Words or sequences of words that commonly appear are assumed to have a high probability of showing up again. As implied, it is important to have a reliable, large source text when building the model. Otherwise, words that did not show up in the source text but rightfully belong to the language will not be recognized by the model. In addition, the source text should have as little noise as possible, so that the model does not try to incorrectly identify a language.

Smoothing is an important concept in language model building because it accounts for words unseen in the source text. It works by saving some space in the model for unknown words so that when encountered during testing, they might be correctly identified by the model. This assignment made use of Laplace smoothing, which is also known as Add-1. It assumes that words seen in the testing data but not the training data were seen once. Its probability is still smaller than words actually encountered.

Language models can be used for text generation because it essentially identifies the context of words. Therefore, the model should know how words are used together, even for alternative definitions of a word. Once again, the usefulness depends on the goodness of an unbiased source text. Different people speak differently and may use a varying amount of stop words (noise.) Like any application of statistics, researchers should be wary of the generalizations made about a population based on a sample. Evaluation of a language model can be done extrinsically where human annotators check the output of the model against a predefined metric. Another way to evaluate a model is by intrinsically comparing models using a metric called perplexity. In this approach, some test data is set aside so that the researcher can observe the entropy. In other words, the model is not only accurate but precise. So, low entropy is more desirable.

Google's Ngram Viewer presents the frequency of phrases in a corpus of books over years. It attempts to display the prevalance of words and phrases over time for a specified language.


computer
boat
rocket