Harry Wada <hcw170000@utdallas.edu> Mar 4, 2023
CS 4395.001

N-grams

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

N-grams refer to sequences of words grouped together, i.e. unigrams are single words, bigrams are pairs of words, etc. These can be generated from corpora to be applied in probabilistic language models. Intuitively, these are useful because they represent how common (or uncommon) certain word sequences are in relation to a corpus.

b. List a few applications where n-grams could be used.

N-grams may be used to group pairs of words in traditional languages, or they may also find application in recognizing patterns in DNA or protein sequences, or even more abstract data like clustered satellite images.

c. A description of how probabilities are calculated for unigrams and bigrams.

Probabilities are calculated first be generating a list of unigrams and bigrams from a corpus. Once this is done, each unique instance of every unigram and bigram is counted and compared as a ratio to the total number of unigrams and bigrams, respectively.

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

Since the source text is really the only thing the language model will have as a reference to compare novel data to, it is important that the source text be representative of the input you expect to encounter. Otherwise, its output will be largely unreliable. The size of the source text is also of utmost importance, as larger corpora will be more likely to contain representative samples of the entire space.

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

A very simple approach to this consists of creating probabilities for each bigram based on its frequency (use the frequency of the first word) in the corpus. Then, use this to probabilistically determine one word after another, appending to the string until a termination point.

f. Describe how language models can be evaluated.

Language models can be evaluated intrinsically and extrinsically. Extrinsically entails the use of human evaluators to measure certain metrics of a model, which can be tedious and time-consuming. Intrinsic measurements are therefore more convenient. One example of an intrinsic measurement is perplexity, which excludes a relatively small amount of training data so that the model can later be evaluated on this data and see how consistent it is to data from the same corpus that it has never seen before.

g. Give a quick introduction to Google’s n-gram viewer and show an example

Google’s n-gram viewer uses a corpus of books to show the popularity of n-grams over time. The date range, corpus language, along with the n-grams themselves, can all be configured separately. One of the first examples that it shows me is "Sherlock Holmes", a bigram. This character first appeared in print in 1887, and the graph reflects this: beginning at essentially 0% in 1885, it rises rapidly to 0.0000310453% by 1900, which may not seem like much, but compared to the entire corpus is relatively high. By 2015, it’s reached 0.0001374117%, more than 10 times the previous value.