Jackie Le

CS4395

Mazidi

N-Grams Narrative

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

N-grams are sliding windows of size n over text. Unigrams are a sequence of one word, bigrams are a sequence of 2 words and trigrams are a sequence of 3 words. They are used to build a language model by training on a large corpus to predict the likelihood of a word to appear in text through the use of n-grams.

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

N-grams can be used in spelling correction, machine translation, speech recognition, auto suggestion, and features in machine learning applications.

1. Description of how probabilities are calculated for unigrams and bigrams

Unigrams are calculated by dividing the probability of a word by the total number of words or tokens in the text. Bigrams are calculated by computing the probability of a word given the previous word divided by the occurrence of the previous word.

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

The source text is very important in building a language model. The language model is greatly influenced by the type of text being used to train the model. For example, text from a science textbook is going to look different compared to text from a William Shakespeare play. The larger and more diverse a given source text is, the language model will perform better than a small one.

1. The importance of smoothing, and describe a simple approach to smoothing

When computing n-grams, an issue called the sparsity problem arises because our data cannot contain every sequence of possible words so there may be zero probabilities. Smoothing is the remedy to this issue by filling in zero values with a bit of probability mass. An approach of smoothing is Laplace smoothing which adds 1 to the 0 count so that there are no zeroes.

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

Language models can be used for text generation by predicting the next likely or highest probable word to appear next. The limitation of this approach however is that the program can output a word of the highest probable n-gram that does not make sense with the previous word(s)

1. Describe how language models can be evaluated

A way of evaluating a language model is by doing an intrinsic evaluation and using perplexity as the metric to compare models. For higher confidence/ a better model, a low perplexity score is ideal.

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

Google’s n-gram viewer is a web tool that is used to show an analysis of the frequency of words or phrases in a large corpus using the Google Books database. The tool allows for the input of a specified term(s), language, and time period (in years). Below, I inserted the terms artificial intelligence, machine learning, and natural language processing.

Chart

Description automatically generated