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

An n-gram is a sequence of n words. N-grams are often used to calculate the likelihood of a word following another word, or a sequence of words.
For example, the phrase "the quick brown fox" is a 4-gram. The word "fox" is a unigram.

List a few applications where n-grams could be used

N-grams are used in speech recognition, machine translation, and text summarization. They can be trained on large corpuses of text to predict the next word in a sentence.
N-grams can be used to identify languages or to detect plagiarism.

A description of how probabilities are calculated for unigrams and bigrams

Probabilities are calculated by counting the number of times a word appears in a corpus and dividing by the total number of words in the corpus.
The probability of the first word in a bigram is multiplied by the probability of the second word in the bigram to get the probability of the bigram.

The importance of the source text in building a language model

The source text is important because it is used to calculate the probabilities of words and bigrams. The more words and bigrams in the source text, the more accurate the probabilities will be.
Multiple source texts can be used to change or refine the language model. Larger and different source texts will produce more accurate models as there will be more variations in the text.

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

Smoothing is used to eliminate outliers in a probability set. For example, if a word only appears once in a corpus, it will have a probability of 1. This is not very useful as it is unlikely that the word will appear again. Smoothing is used to reduce the probability of rare words and bigrams.
A common simple approach to smoothing is laplace smoothing, which adds 1 to the count of each word and bigram in the corpus.

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 creating dictionaries of words and bigrams and their probabilities. The probabilities are used to generate a random number between 0 and 1. The word or bigram with the highest probability that is less than the random number is chosen. This process is repeated until a sentence is generated. The limitations of this approach are that the generated text will not make sense as the probabilities are not based on the context

of the sentence. For example, the word "the" will have a high probability of appearing in a sentence, but it will not be in the correct context.
This method does not take into account the context of the sentence, so it will not generate coherent text. It also does not account for grammar rules.

Describe how language models can be evaluated

Language models can be evaluated intrinsically or extrinsically. Extrinsic evaluation is done by comparing the language model to a human generated text. Intrinsic evaluation is done by comparing the language model to itself.

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

Google's n-gram viewer is used to visualize the relative frequency of words in a corpus. It has a large selection of corpuses to choose from, and will show the probability of each word showing up at different points in time.