Narrative

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

N-grams are continuous sequences of words or symbols or tokens in a document. N-grams are used to build language models which can further be used to predict the probability of a particular sequence of words occurring. The key idea is to evaluate the frequency of appearances of n-grams in a given corpus and to predict the probability of the sequence of words in the actual text. Once the probabilities of n-gram are calculated, the next step is to utilize them in building a language model. Such models can generate new text or evaluate the likelihood of a particular sequence of words. Such models can be used for text generation which can be used in text editors, email applications, etc. Speech recognition is another application of such models.

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

N-grams have various applications:
- N-grams could be used in the auto-completion of sentences, auto spell check, and semantic analysis.
-N-grams could be used in DNA sequencing.
-Another application is text compression, a dictionary look-up feature in text editors.

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

Unigrams and bigrams are types of n-grams. For unigrams, the probability of a word w is calculated as:

$$P(w) = count(w) / N$$

Where, N is the total number of words within a corpus of data and count(w) is the frequency of word w in a given corpus.

Bigrams represent a pair of adjacent words. the probability of a word pair (w1, w2) is calculated as:

$$P(w2 \mid w1) = count(w1, w2) / count(w1)$$

Where the probability of seeing word w2 provided w1 has already occurred in a corpus is calculated by dividing the count of appearances of the number of times the word pair (w1, w2) occur in the corpus, to the total number of times word w1 appears in a corpus.

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

A language model is a probability distribution over words or a sequence of words. A language model learns from text. A model's ability to predict accurate results are directly impacted by the quality and diverseness of the source text. A language model performs probability distribution over words which can further be used to predict the most likely next word within a sentence. Since subsequent predictions are based on learning from texts, source text plays a vital role in building a language model.

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

Data smoothing is the process of removal of noise from a data set. This allows data patterns to be easily viewable. There are different approaches to performing smoothing - simple exponential, moving average, random walk, and exponential moving average. Simple exponential smoothing is one of the most popular methods, wherein it uses average calculation for assigning the exponentially declining weights starting from the most recent observation. This approach is fairly reliable as the difference between real projects and what actually happens is taken into consideration.

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

Text generation is the task of producing new text. Text generation is used to fill in incomplete text. In order to use language models for text generation, one needs to provide seed text which would then be used to generate text further. Such text generations are based on probabilities that are learned from the training data provided. This can further be fine-tuned by altering the parameters of the model or supplying the model with background information and context which help aid the text generation process. One such application for generating text is for generating news articles.

Limitations to this approach:

1. Quality of the text: Generated text can sometimes produce syntactically and grammatically correct text but it may lack relevance to the subject matter. This could be a result of a lack of proper context in the model.

2. Biases in response: These models could be prone to detecting biases in data and favoring its response based on it. This depends on the quality of the training data.

3. Large training data set is required: Large training data is required to train such models. This can be challenging in certain cases.

g. Describe how language models can be evaluated.

There are various measures to evaluate language models including but not limited to:

1. Accuracy - This is one of the important criteria. How accurately a model can analyze the context at hand and give the most accurate results.

2. Robustness - This indicates how well the model can deal with and ignore external noise which can distort its results.

3. Coherence of response - This refers to how relevant the responses are when compared to human responses.

4. Degree of diversity - How many different responses a model can provide including the number of unique phrases and words the model uses in its responses.

Although the above measures are commonly used, the measure of a language model might differ based on its application or background.

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

Google's n-gram viewer gives an analysis of the frequency of words in books and articles which have been digitized by Google. For example, if you enter the phrase "climate change" and select the English corpus, you would be able to view an increase in the frequency of the term over recent years. This shows the growing popularity and concern towards climate change in recent years.