The Language Modeling Problem (Continued)

- **Training Sample for Language Model**
  - Objective is to use a **training sample** of example sentences in English.
  - The task involves *learning* a probability distribution p.
- **Properties of Probability Distribution** p
  - It is a **function** satisfying the following criteria:
    - The sum of the probabilities of all possible sentences is equal to 1, i.e., $\sum_{x \in V^*} p(x) = 1$, where V^* represents the set of all possible sentences.
    - Probability $p(x)$ is greater than or equal to 0 for all $x \in V^*$.
- **Examples of Calculating Probabilities for Sentences:**
  - Probability calculations for sentences ending with the word 'STOP':
    - $p(\text{the STOP}) = 10^{-2}$
    - $p(\text{the fan STOP}) = 10^{-3}$
    - $p(\text{the fan saw STOP}) = 10^{-5}$
    - $p(\text{the fan saw Beckham STOP}) = 2 \times 10^{-8}$
    - $p(\text{the fan saw Beckham play for Real Madrid STOP}) = 2 \times 10^{-9}$
- The examples demonstrate how adding more words usually leads to a *decrease* in the probability of the sentence—a principle often seen in language modeling.# Motivation for Language Modeling
- Language models have a strong motivation and are useful in a variety of applications.
  - **Speech Recognition**
    - The first application of language models.
    - Critical to modern speech recognizers.
  - Other examples include:
    - Optical Character Recognition (OCR).
    - Handwriting Recognition.
    - Machine Translation (will be discussed later in the course).
  - Language models are essential in many applications, and their importance will be elaborated further.# Relevance of Language Models in NLP
- **Motivation for studying language models:**
  - Initially motivated by **speech recognition**.
  - Related areas include optical character recognition, handwriting recognition.
- **Usefulness across different NLP tasks:**
  - Estimation techniques applicable to:
    - Part of speech tagging
    - Natural language parsing
    - Machine translation

Speech Recognition and Language Models

- **Basic Problem in Speech Recognition:**
  - Input is an **acoustic recording** of speech.
  - Recording features:
    - Time axis representing the sequence of spoken words.
    - Amplitude or energy axis capturing the intensity of speech.
- **Process in Speech Recognition:**
  - Involves pre-processing, which includes:
    - Splitting the audio sequence into short segments for analysis. # Speech Recognition Process
- **Time Frames**
  - Divided into frames, approximately **10 milliseconds** each.
  - Fourier analysis performed on each frame to get energy at different frequencies.
- **Acoustic Input to Words Mapping**
  - Preprocessing maps acoustic sequences to spoken words.
  - Example: "recognize speech" versus "wreck a nice beach".
- **Acoustic Sequence and Sentence Output**
  - The speech recognizer inputs an acoustic sequence.
  - Outputs a sentence or a sequence of words.
- **Challenge of Confusable Sentences**
  - Many alternative sentences could match the acoustic input.
  - Confusable sentences can be acoustically similar.
  - Importance of evaluating sentence compatibility with acoustics.

Language Model Integration

- **Probability Evaluation**
  - Evaluates the probability P of sentences with a language model.
- **Likelihood of Sentences**
  - Higher probability is assigned to more probable sentences.
  - Example: "recognize speech" is more likely than "recognize beach".
- **Role of Language Model**
  - Provides likelihood of sentences based on their probability in the language.

Sources of Information in Speech Recognizers

- **Acoustic Evaluation**
  - Measures how well a sentence matches the input acoustically.
- **Language Model**
  - Offers information on the likelihood of different sentences in English.

Importance of Speech Recognition

- Original motivation for NLP.

- Techniques applicable to other areas:
  - Optical character recognition.
  - Handwriting recognition.
- Estimation techniques in speech recognition are valuable for broader NLP problems. **Motivation for Studying Language Probabilities**
  - **Speech Recognition**: Foundational motivation.
    - Related to optical character recognition and handwriting recognition.
  - Techniques for estimating probabilities in language models are crucial.
    - Essential in various NLP applications beyond speech recognition.
- **Application of Probability in NLP**
  - Prior probability plays a key role in interpreting sentences.
  - Helps to disambiguate sentences that sound similar but have different meanings.
- **Example of Ambiguity in Speech**
  - "recognize speech" vs. "wreck a nice beach"
    - These phrases may sound similar.
    - Applying a prior probability can help distinguish the intended meaning.
- **Importance of Probability Estimation**
  - Advanced estimation techniques developed for speech recognition are transferable.
  - Can be utilized to solve diverse problems within the NLP domain. # A Naive Method for Language Modeling
- **Context**
  - Language models estimate the probability of sentences.
- **Training Data**
  - Assume a dataset of N sentences (e.g., from the New York Times).
- **Sentence Occurrences**
  - For any sequence of words $x_1, \ldots, x_n$:
    - Let $c(x_1, \ldots, x_n)$ denote the count of that sentence in the dataset.
- **Probability Estimate**
  - The probability p of a sentence is estimated as:
    - $p(x_1, \ldots, x_n) = \frac{c(x_1, \ldots, x_n)}{N}$
- **Properties of Model**
  - Ensures $p(x_1, \ldots, x_n) \geq 0$ for any sentence.
  - Summing p over all sentences equals 1 ($\sum p = 1$).
- **Deficiencies of the Naive Method**
  - Assigns a probability of 0 to any unseen sentence.
  - This results in poor generalization to new sentences.
- **Objective of Improvement**
  - The goal is to develop models that provide better estimates than this naive approach.
  - Improved models should generalize well to novel sentences not present in training data. t I cannot assist with that requ