ChatGPT markdown Copy code # Natural Language Processing Course Notes

Language Modeling

- **Definition**: Language modeling is a fundamental problem in statistical natural language processing (NLP).
- **Importance**:
  - It's one of the oldest problems studied in NLP.
  - Essential for a wide range of natural language applications.

Trigram Language Models

- **Overview**:
  - A significant class of language models.
  - Extremely widely used in various NLP tasks.

Evaluating Language Models

- **Perplexity**:
  - A standard measure for evaluating the effectiveness of language models.

Estimation Techniques

- **Linear Interpolation**:
  - One of the primary estimation techniques for language modeling.
- **Discounting Methods**:
  - Another fundamental technique used in language modeling.
- **Relevance**:
  - Both techniques are not only crucial for language modeling but are also applicable in other NLP problems.
  - Their application extends beyond language modeling to other areas in NLP.

The Language Modeling Problem

- **Vocabulary** (V):
  - Defined as a finite set containing all words in a given language.
  - Examples include words like 'the', 'a', 'man', 'telescope', 'Beckham', 'two', etc.
  - The size of V can range from thousands to tens of thousands of unique words.
- **Set of Strings** ($V^\dagger$):
  - Represents an (infinite) set of all possible sentences or strings constructed from V.

  - A well-formed sentence is one that comprises zero or more words from V, followed by a special 'STOP' symbol.
  - The 'STOP' symbol denotes the end of a sentence and is a crucial part of the language modeling process.
- **Sentence Formation**:
  - Can include any sequence of words from V ending with 'STOP', even if nonsensical.
  - Includes edge case with only the 'STOP' symbol, representing a zero-length sentence.

The Language Modeling Problem (Continued)

- **Training Sample**:
  - Consists of a collection of example sentences from a specific language, such as English.
  - Can be derived from various sources, like newspapers or the web.
- **Learning a Probability Distribution (p)**:
  - The goal is to learn a distribution p over sentences.
  - p is a function that assigns probabilities to sentences, ensuring two conditions:
    - For any sentence x in the set of possible sentences $V^{\dagger}$, $p(x) \geq 0$.
    - The sum of $p(x)$ for all x in $V^{\dagger}$ equals 1, i.e., $\sum_{x \in V^{\dagger}} p(x) = 1$.
- **Examples**:
  - A sentence composed only of 'the STOP' may be assigned a probability of 10^{-12}.
  - A more complex sentence like 'the fan saw Beckham STOP' could have a probability of 2×10^{-8}.
  - The probabilities reflect the likelihood of encountering the sentence in the language.