Basic NLP Concepts

The Pipeline Towards Understanding

Ayush (250247)

Introduction

Natural Language Processing (NLP) relies on a specific pipeline to transform raw text into a format that computers can understand. This process involves breaking down language into smaller units and analyzing their grammatical and contextual roles. This report outlines the six primary stages of this pipeline.

1 Sentence Segmentation

Understanding a full paragraph at once is a complex task for a computer. To simplify this, the first step involves breaking a paragraph down into individual sentences. This allows the system to process ideas one at a time.

Typically, code separates sentences by identifying punctuation marks such as periods and semicolons. While standard punctuation is the primary method, more advanced techniques are often required to ensure reliability across different writing styles.

2 Word Tokenization

Once the text is segmented into sentences, the next step is **Tokenization**. This involves splitting sentences into separate words or units called "tokens."

- **Method:** The system generally splits text wherever there is a space. Punctuation marks are also treated as separate tokens because they carry semantic meaning.
- **Purpose:** These tokens become the fundamental units the computer works with to derive meaning and establish relations within the text.

3 Stop-Words Removal

After tokenization, the dataset is analyzed to identify words that appear frequently but convey little unique information. These are known as "stop words."

Common examples include connectors like "and," "the," and "a." Removing these words serves two main purposes:

1. It isolates the significant content of the paragraph (keywords).
2. It reduces the size of the dataset, thereby decreasing the processing time required by the computer.

4 Lemmatization

Words often appear in various forms depending on the grammar of a sentence (e.g., “breaking,” “broke,” “break”). If a computer treats these as completely different words, it misses the connection between them.

Lemmatization is the process of reducing these variations to their simplified, common base form, or “lemma.”

- **Example:** Reducing “breaking” and “broke” to “break.”

This normalizes the text, ensuring that different grammatical forms are understood as the same concept.

5 Part-of-Speech (POS) Tagging

This stage involves assigning a grammatical label to each token, such as noun, adjective, or adverb.

The model uses statistics to guess the part of speech based on similar sentences and words it has processed previously. It does not “understand” the meaning in a human sense but relies on probability to determine the role a word plays in a sentence. This classification is crucial for drawing relations between words later in the process.

6 Named Entity Recognition (NER)

The final step discussed is extracting the real-world context of the text. Named Entity Recognition classifies specific words into defined real-world categories.

Typical objects tagged by an NER system include:

- **People’s Names**
- **Company Names** (Corporations)
- **Geographic Locations** (Physical and political)
- **Dates and Events**

This classification allows the machine to move beyond grammatical structure and begin identifying the actual entities discussed in the text.