INTRODUCTION TO NLP

Lecture 1

Index

- **What is Natural Language Processing?**
- **Core Components of NLP**
- **How Does Natural Language Processing Work?**
- **Challenges in Natural Language Processing**
- **Introduction to Arabic NLP**

What is Natural Language Processing?

What is Natural Language Processing?

- Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on the interaction between computers and humans through natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of human languages in a manner that is valuable. It involves teaching computers to seamlessly interpret and process human languages, enabling them to perform tasks such as translation, sentiment analysis, and topic extraction.

Core Components of NLP

Core Components of NLP

- Syntax
- Semantics
- *Pragmatics*

Syntax

Syntax in linguistics refers to the set of rules, principles, and processes that govern the structure of sentences in a given language, specifically the order of words and phrases and how they combine to form sentences. When studying syntax, one focuses on the formal patterns of language without considering the meanings of words and phrases. Syntax is crucial because it provides a clear structure that helps in understanding and processing language efficiently.

Syntax

- For example, in English, a basic syntactic rule is that a typical sentence follows a Subject-Verb-Object order. "The cat (subject) chased (verb) the mouse (object)."

Semantics

- Semantics is the branch of linguistics that studies the meanings of words, phrases, and sentences. It delves into how people understand and interpret language in various contexts. Semantics covers a range of topics including the meanings of individual words, the changes in meaning that occur when words combine, and the way meaning can shift based on context.

Semantics

- For instance, the word "bank" can mean the edge of a river or a financial institution, depending on the context in which it's used. Semantics helps to analyze these meanings to ensure clear communication.

Pragmatics

Pragmatics is the study of how context influences the interpretation of meaning in language. It examines how speakers use language in social situations and how interpretations can vary depending on factors such as the speaker's intentions and the listener's perceptions. Pragmatics is concerned with aspects of meaning that aren't solely derived from the linguistic elements themselves but are about how the context and use of language contribute to meaning.

Pragmatics

- For example, the phrase "Can you pass the salt?" is typically understood as a request, not just a question about one's ability to pass the salt. This understanding comes from interpreting the social context and the speaker’s intent. Pragmatics explores these subtleties of language use that are crucial for effective communication.

How Does Natural Language Processing Work?

How Does Natural Language Processing Work?

- **Text Preprocessing Techniques**
  - **Stemming and Lemmatization**
  - **Part-of-Speech Tagging**
- **NLP Algorithms and Models**
  - **Machine Learning Models**
  - **Deep Learning Models**

Stemming and Lemmatization

Stemming and Lemmatization

- Stemming and lemmatization are two foundational techniques used in the field of natural language processing (NLP) to reduce words to their base or root form. Stemming involves cutting off the ends of words in the hope of achieving this goal correctly most of the time. It is a somewhat crude approach that chops off word prefixes and suffixes. For example, the stem of the words "connection," "connections," "connective," "connected," and "connecting" is "connect."

Stemming and Lemmatization

- Lemmatization, on the other hand, involves a more sophisticated approach where words are reduced to their lemma or dictionary form. Unlike stemming, lemmatization considers the context and converts the word to its meaningful base form. For instance, "is," "are," and "am" are all lemmatized to "be." Lemmatization uses vocabulary and morphological analysis of words, which makes it slower but more accurate than stemming.

Part-of-Speech Tagging

Part-of-Speech Tagging

- Part-of-Speech (POS) tagging is an essential process in natural language processing that involves assigning a part of speech to each word in a given text, based on both its definition and its context. This can include labels for nouns, verbs, adjectives, adverbs, etc. POS tagging is crucial for syntactic parsing and word sense disambiguation, helping machines understand the grammatical structure of sentences and the roles of each word.

- Modern POS taggers use machine learning algorithms, particularly those that take context into account, such as Hidden Markov Models (HMMs) or more advanced deep learning models. These tools are trained on large corpora of annotated text where the correct part-of-speech tags are already assigned, allowing them to learn and predict the tags for new texts.

Challenges in Natural Language Processing

Challenges in Natural Language Processing

- **Handling Ambiguity and Context**
- **Language Diversity and Adaptability**
- **Computational Complexity**