Phases in Natural Language Processing

There are phases in NLP which need to be performed in order to extract meaningful information from the text corpus.

Corpus is a collection of machine-readable text collected according to certain criteria. Representative collection of text.

Once these phases are completed, you are ready with your refined text and then you can apply some machine learning/deep learning model to predict something.

1. **Lexical analysis** is a crucial phase in natural language processing (NLP) where text data is processed to break it down into smaller units (tokens) and analyse the structure and content of words.

Steps in Lexical Analysis

a. **Tokenization:**

- **Definition:** Breaking down a text into smaller units (tokens) such as words or sentences.
- **Purpose:** Provides a structured representation of text that can be analysed or processed by algorithms.
- **Types:**
  - **Word Tokenization:** Dividing text into words.
  - **Sentence Tokenization:** Dividing text into sentences.
  - **Tweet Tokenization:** Specific tokenization for social media posts like tweets, which may include hashtags and mentions.

Input Text: "Natural language processing is fun and challenging."

` **Tokenization Output:**

Tokens: ["Natural", "language", "processing", "is", "fun", "and", "challenging", "."]

b. **Stop Word Removal:**

- **Definition:** Removing commonly used words (stop words) that do not contribute significant meaning to the text.
- **Purpose:** Improves efficiency by **reducing noise in the data** and focusing on important words.
- **Example:** Removing words like "and," "of," "the," etc., which are frequent but usually not informative.

Original Sentence: "The weather in the city is beautiful today."

Stop Words (Assumed List): ["the", "in", "is"]

After Stop Word Removal: "weather city beautiful today."

c. **Stemming:**

- **Definition:** Reducing words to their root or base form by removing suffixes like "-ing," "-es," "-s," etc.
- **Purpose:** Normalizes words with similar meanings to a common base form, reducing variation in the text data.
- **Example:** Converting "running" to "run," "cats" to "cat."

d. **Lemmatization:**

- **Definition:** Similar to stemming, but lemmatization considers the context and converts words to their base or dictionary form (lemma).
- **Purpose:** Provides a more accurate base form of words by considering grammar and meaning.

- **Example:** Converting “better” to “good”.
  Input Text: "The cats are chasing mice around the house."
  Lemmatization Output:
  Lemmatized Text: "The cat be chase mouse around the house."

Search Engines: **Lemmatization improves the relevance of search results by matching user queries with base forms of words.**

2. **Syntactic Analysis**, also known as syntax parsing or syntactic parsing, is a fundamental task in natural language processing (NLP) that involves **analysing the grammatical structure of sentences to understand the relationships between words and their roles within the sentence**.

Techniques in Syntactic Analysis:

a. Dependency Parsing:

Definition: Dependency parsing involves analysing the grammatical structure of a sentence to establish relationships between "head" words and words that modify those heads.

Purpose: It determines how each word relates to other words in terms of grammatical relationships such as subject, object, modifier, etc.

Example: In the sentence "The cat chased the mouse," dependency parsing would identify that "chased" depends on "cat" as the subject and "mouse" as the object.

b. Parts of Speech (POS) Tagging:

Definition: POS tagging assigns grammatical categories (tags) to words in a sentence, such as noun, verb, adjective, adverb, etc.

Purpose: It provides information about the syntactic role of each word, which helps in understanding the sentence structure.

Example: In the sentence "She runs quickly," POS tagging would tag "She" as a pronoun, "runs" as a verb, and "quickly" as an adverb.

Example: "The dog chased the cat."

Tokenization and POS Tagging:
Tokenization: Breaks down the sentence into individual words or tokens.
Tokens: ["The", "dog", "chased", "the", "cat", "."]
POS Tagging: Assigns parts of speech tags to each token.
Tags: ["DT", "NN", "VBD", "DT", "NN", "."]
Explanation: DT (determiner), NN (noun), VBD (past tense verb).

Dependency Parsing:
Dependency Relations: Establishes relationships between words.
Example Dependency Tree:

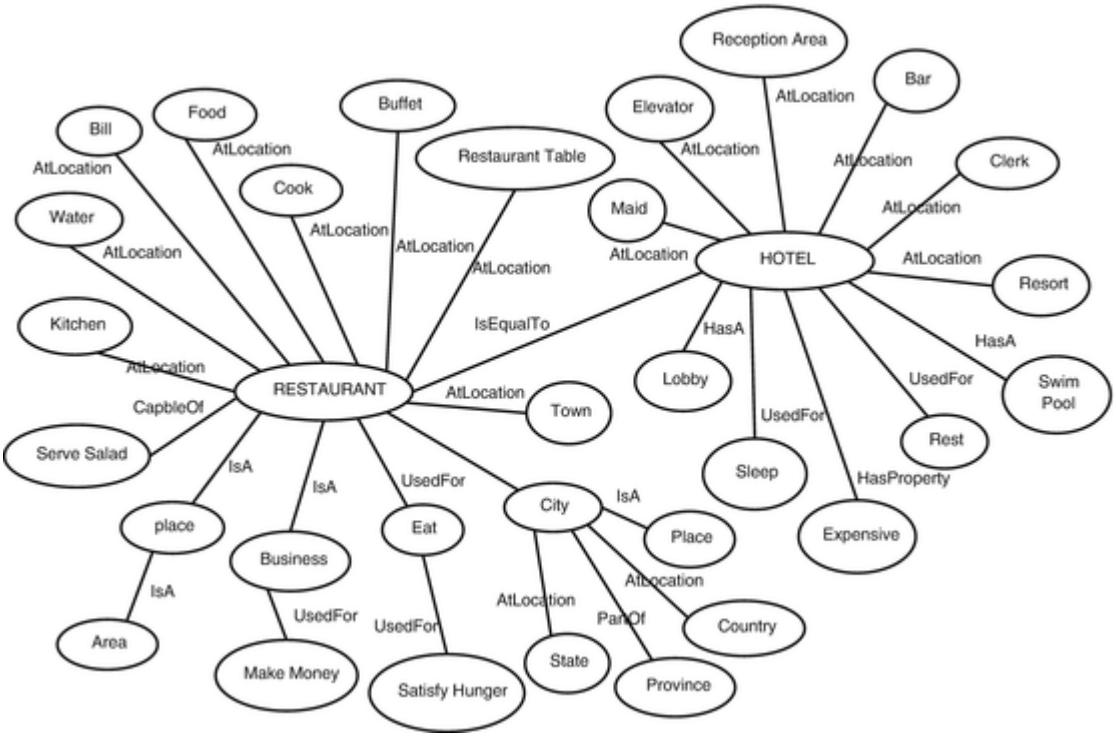
```
ROOT - chased
  |
  +-- dog (subject)
  |
  +-- cat (object)
```

3. Semantic analysis in natural language processing (NLP) involves understanding the meaning of text beyond its syntactic structure.
Semantic analysis examines relationships between individual words and analyses the meaning of words that come together to form a sentence.
This analysis provides a clear understanding of words in context. For example, it provides context to understand the following sentences:

Input Sentence: "Truck is eating oranges."
Semantic analysis identifies the sentence "Truck is eating oranges." as nonsensical because it evaluates the meaning and relationships between the words in context.
Steps in Semantic Analysis

Before semantic analysis, the sentence undergoes syntactic analysis to parse its grammatical structure. Followed by given steps are carried out.

- **Semantic Role Labelling**:
  Semantic role labelling assigns roles to different parts of the sentence, such as agent, action, and object.
  "Truck" is identified as the agent (doer of the action).
  "is eating" is identified as the action.
  "oranges" is identified as the object (recipient of the action).
- **World Knowledge and Ontology**:
  World knowledge or ontologies provide information about the typical properties and capabilities of entities.



Ontologies or databases (such as WordNet or ConceptNet) contain knowledge about what kinds of actions are possible for different entities.
Trucks are categorized as vehicles, which are typically associated with actions like driving, transporting, or carrying.
Eating is an action typically associated with living beings, particularly animals and humans.

- **Logical Consistency Check**:
  The semantic analysis checks for logical consistency between the agent and the action.
  Here, it evaluates whether it makes sense for a "truck" (a non-living, inanimate object) to perform the action "eating."
  It finds that eating is not an action that trucks (vehicles) can perform based on world knowledge and typical capabilities.
- **Contextual Reasoning**:
  If there were additional context, the analysis would consider the broader context to see if there might be a metaphorical or non-literal interpretation.
  In the absence of such context, the sentence is evaluated at face value.
- **Semantic Anomaly Detection**:
  The system identifies a semantic anomaly: the action (eating) is incompatible with the agent (truck).
  This inconsistency leads the system to determine that the sentence is semantically nonsensical or wrong.
  Example of Semantic Knowledge Use:
  Ontology Data:
  Truck - Category: Vehicle, Capabilities: Transport, Carry, Move, Incapabilities: Eat, Speak, Think
  Eating - Typical Agents: Humans, Animals, Incompatible Agents: Vehicles, Objects
  Analysis:

Agent (Truck): Vehicle, incapable of eating.
Action (is eating): Requires a living being as the agent.
Conclusion: The action "eating" is incompatible with the agent "truck," making the sentence semantically invalid.

` **Tasks in Semantic Analysis**:-

a. Word Sense Disambiguation (WSD):

Definition: Identifying the correct meaning of words that have multiple meanings based on context.

Example:

"The boy ate the apple" defines an apple as a fruit.

"The boy went to Apple" defines Apple as a brand or store.

b. Named Entity Recognition (NER):

Definition: Identifying and classifying entities such as names of persons, organizations, locations, dates, etc., within text.

Example:

- Recognizing "New Bombay" as a location entity in the sentence "I live in New Bombay."

- **"Apple released the iPhone 12 in October 2020."**
  Organization: Apple
  Product: iPhone 12
  Date: October 2020

c. Semantic Role Labelling (SRL):

Definition: Identifying the relationships between words in a sentence and their semantic roles (e.g., agent, patient, instrument).

Example: Identifying that in "She opened the door with a key," "She" is the agent, "door" is the patient, and "key" is the instrument.

4. Discourse Integration: Its scope is not only limited to a word or sentence, rather discourse integration helps in studying the whole text.

Example: "Mallon got ready at 9 AM. Later he took the train to Kerala"

Here, the machine is able to understand that the word "he" in the second sentence is referring to "Mallon". (Anaphora resolution)

5. Pragmatic Analysis: It is a complex phase where machines should have knowledge not only about the provided text but also about the real world. There can be multiple scenarios where the intent of a sentence can be misunderstood if the machine doesn't have real world knowledge.

Pragmatic analysis focuses on understanding the intended meaning, context, and implications of a sentence beyond its literal meaning.

Example:

"Thank you for coming so late, we have wrapped up the meeting" (Contains sarcasm)

"Can you share your screen?" (here the context is about computer's screen share during a remote meeting)

Example Sentence

"Can you pass the salt?"

Literal Interpretation

Literal Meaning: A question asking if the person is capable of passing the salt.

Pragmatic Analysis

Speaker's Intent: The speaker is not actually questioning the listener's ability to pass the salt but is politely requesting that the listener pass the salt.
Implied Meaning: "Please pass the salt."

Ambiguity in Natural Language Processing (NLP) refers to situations where a sentence or phrase can be interpreted in multiple ways due to the inherent complexities of language.

Types of Ambiguity in NLP

1. **Lexical Ambiguity:**

- **Definition:** Occurs when a word has multiple meanings.
- **Example:**
    - The word "bank" can mean the side of a river or a financial institution.
      Financial Institution: A bank is an organization where people can deposit money, withdraw money, take loans, and perform other financial transactions.

      Example: "I need to go to the bank to deposit my paycheck."

      Riverbank: A bank refers to the land alongside or sloping down to a river or lake.

      Example: "We had a picnic on the bank of the river."

      Storage: A bank can also mean a collection or supply of something, typically kept for future use.

      Example: "We have a bank of information on the new project."

1. Silver as a Noun

Meaning: Refers to the chemical element known for its shiny appearance, conductivity, and use in making coins, jewelry, and other items.

- **Example:** "The Olympic medal was made of silver."
    - In this sentence, "silver" is used as a noun to describe the material of which the medal is composed.

2. Silver as an Adjective

Meaning: Describes something that resembles the color of polished silver, typically a shiny grey-white shade.

- **Example:** "She wore a beautiful dress in silver."
    - Here, "silver" is used as an adjective to describe the color of the dress, indicating it is shiny and grey-white in appearance.

3. Silver as a Verb

Meaning: To cover something with a thin layer of silver or to make something shiny like silver.

- **Example:** "He silvered the mirror to make it more reflective."
    - In this sentence, "silvered" is used as a verb, describing the action of coating the mirror with a layer of silver to enhance its reflectivity.

2. **Syntactic Ambiguity:**

D**efinition:**

3. **Syntactic ambiguity** occurs when a sentence or phrase can be parsed (structured) in more than one way due to the arrangement of words and phrases.
   Cause:
4. It arises from the **grammatical structure** of the sentence and how words are grouped into phrases, clauses, and sentences.

Example: "She saw the man with the telescope."

- **Interpretation 1:** I used a telescope to see the man.
  - **Parse Tree:** (She saw (the man (with the telescope)))
- **Interpretation 2:** The man had a telescope.
  - **Parse Tree:** (She (saw the man) (with the telescope))

5. **Semantic Ambiguity:**

- **Definition:** Involves ambiguity in the meaning of words or phrases.
  - "The chicken is ready to eat."
  This could mean either that the cooked chicken is ready to be eaten, or that the live chicken is about to eat something.
  - "I gave her the book with the cover."
  This could mean either that the speaker gave a covered book to someone, or that he gave a book to someone, but it was accompanied by a separate cover.
  - "I saw her duck."
  Did you see her lower her head or did you see a duck that belonged to her?

6. **Referential Ambiguity:**

Referential ambiguity occurs when it is unclear which person, thing, or concept a pronoun or noun phrase refers to within a sentence or discourse. This ambiguity arises when there are multiple possible antecedents (preceding words or phrases) to which the pronoun or noun phrase could refer.

Example: "John told Mark that he was going to the store."

- **Referential Ambiguity:** The pronoun "he" could refer to either "John" or "Mark."

1. **Interpretation 1:** "He" refers to John.
   - **Sentence Interpretation:** John told Mark that John himself was going to the store.
   - **Sentence Breakdown:** "John told Mark" (John told Mark something) and "that he was going to the store" (he = John).
2. **Interpretation 2:** "He" refers to Mark.
   - **Sentence Interpretation:** John informed Mark that Mark himself was going to the store.
   - **Sentence Breakdown:** "John told Mark" (John told Mark something) and "that he was going to the store" (he = Mark).

Contextual Cues: Understanding the context or additional information provided before or after the sentence can help clarify who "he" refers to.

7. **Pragmatic Ambiguity:**

- **Definition:** Relates to ambiguity in how context affects the interpretation of a sentence.
- **Example:** "It's cold in here, isn't it?"
  Could be a statement about the temperature or a subtle request to adjust the environment.