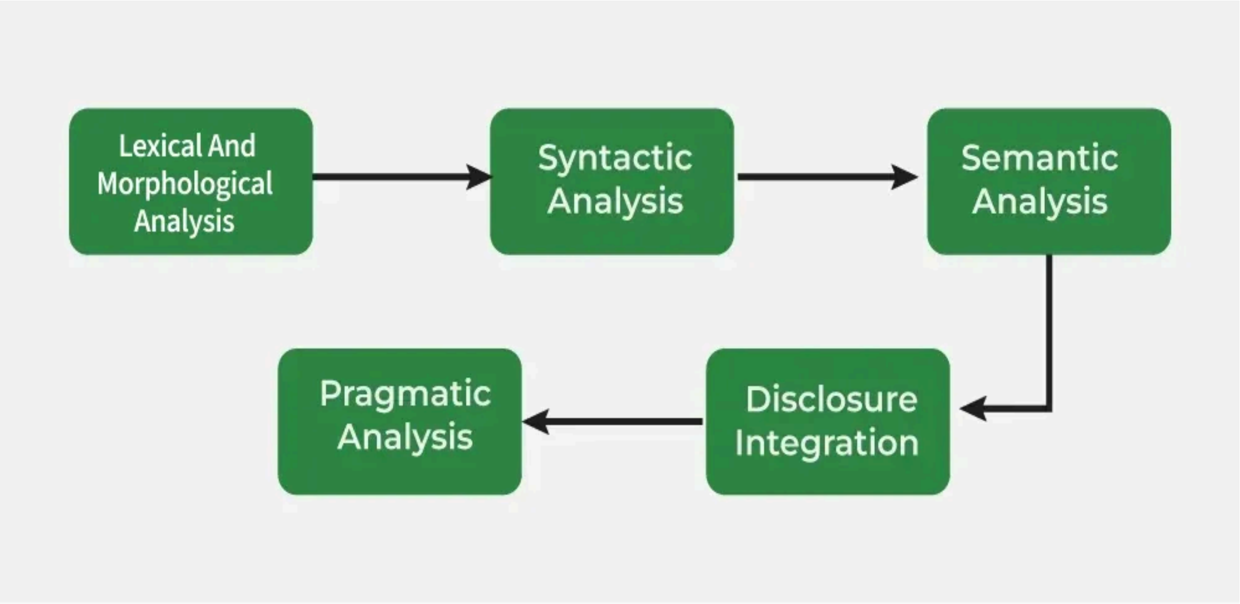
NATURAL LANGUAGE PROCESSING ALGORITHMS

PHASES - NATURAL LANGUAGE UNDERSTANDING - NATURAL LANGUAGE GENERATION

Phases of Natural Language Processing (NLP)

Natural Language Processing (NLP) helps computers to understand, analyze and interact with human language. It involves a series of phases that work together to process language and each phase helps in understanding structure and meaning of human language. In this article, we will understand these phases.



1. Lexical and Morphological Analysis

Lexical Analysis

It focuses on identifying and processing words (or lexemes) in a text. It breaks down the input text into individual tokens that are meaningful units of language such as words or phrases.

Key tasks in Lexical analysis:

1. Tokenization: Process of dividing a text into smaller chunks called tokens. For example the sentence "I love programming" would be tokenized into ["I", "love", "programming"].

2. Part-of-Speech Tagging: Assigning parts of speech such as noun, verb, adjective to each token in the sentence. This helps us to understand grammatical roles of words in the context.

Example: Consider the sentence: "I am reading a book."

- Tokenization: Sentence is broken down into individual tokens or words: ["I", "am", "reading", "a", "book"]
- Part-of-Speech Tagging: Each token is assigned a part of speech: ["I" → Pronoun (PRP), "am" → Verb (VBP), "reading" → Verb (VBG), "a" → Article (DT), "book" → Noun (NN)]

Importance of Lexical Analysis

- Word Identification: It breaks text into tokens which helps the system to understand individual words for further processing.
- Text Simplification: It simplifies text through tokenization and stemming which improves accuracy in NLP tasks.

Morphological Analysis

It deals with morphemes which are the smallest units of meaning in a word. It is important for understanding structure of words and their parts by identifying free morphemes (independent words like "cat") and bound morphemes (like prefixes or suffixes e.g. "un-" or "-ing").

Key tasks in morphological analysis:

1. Stemming: Reducing words to their root form like "running" to "run".
2. Lemmatization: Converting words to their base or dictionary form considering the context like "better" becomes "good".

Importance of Morphological Analysis

1. Understanding Word Structure: It helps in breaking the composition of complex words.
2. Improving Accuracy: It enhances accuracy of tasks such as part-of-speech tagging, syntactic parsing and machine translation.

By identifying and analyzing morphemes system can identify text correctly at the most basic level which helps in more advanced NLP applications.

2. Syntactic Analysis (Parsing)

Syntactic Analysis helps in understanding how words in a sentence are arranged according to grammar rules. It ensures that the sentence follows correct grammar which makes the meaning clearer. The goal is to create a parse tree which is a diagram showing the structure of sentence. It breaks the sentence into parts like the subject, verb and object and shows how these parts are connected. This helps machines understand the relationships between words in the sentence.

Key components of syntactic analysis include:

- POS Tagging: Assigning parts of speech (noun, verb, adjective) to words in a sentence as discussed earlier.
- Ambiguity Resolution: Handling words that have multiple meanings (e.g "book" can be a noun or a verb).

Examples Consider the following sentences:

- Correct Syntax: "John eats an apple."
- Incorrect Syntax: "Apple eats John an."

Despite using same words only the first sentence is grammatically correct and makes sense. The correct arrangement of words according to grammatical rules is what makes the sentence meaningful. By analyzing sentence structure NLP systems can better understand and generate human language. This helps in tasks like machine translation, sentiment analysis and information retrieval by making the text clearer and reducing confusion.

3. Semantic Analysis

Semantic Analysis focuses on understanding meaning behind words and sentences. It ensures that the text is not only grammatically correct but also logically coherent and contextually relevant. It aims to understand dictionary definitions of words and their usage in context and also find whether the arrangement of words in a sentence makes logical sense.

Key Tasks in Semantic Analysis

1. Named Entity Recognition (NER): It identifies and classifies entities such as names of people, locations, organizations, dates and more. These entities provide important meaning in the text and help in understanding the context. For example in the sentence "Tesla announced its new electric vehicle in California," NER would identify "Tesla" as an organization and "California" as a location.
2. Word Sense Disambiguation (WSD): Many words have multiple meanings depending on the context in which they are used. It identifies the correct meaning of a word based on its surrounding text. For example word "bank" can refer to a financial institution or the side of a river. It uses context to identify

which meaning applies in a given sentence which ensures that interpretation is accurate.

Example of Semantic Analysis
"Apple eats a John." while grammatically correct this sentence doesn’t make sense semantically because an apple cannot "eat" a person. Semantic analysis ensures that the meaning is logically sound and contextually appropriate. It is important for various NLP applications including machine translation, information retrieval and question answering.

4. Discourse Integration

It is the process of understanding how individual sentences or segments of text connect and relate to each other within a broader context. This phase ensures that the meaning of a text is consistent and coherent across multiple sentences or paragraphs. It is important for understanding long or complex texts where meaning focuses on previous statements.

Key aspects of discourse integration:

- Anaphora Resolution: Anaphora refers to the use of pronouns or other references that depend on earlier parts of the text. For example in the sentence "Taylor went to the store. She bought groceries" pronoun "She" refers back to "Taylor." It ensures that references like these are correctly understood by linking them to their antecedents.
- Contextual References: Many words or phrases can only be fully understood when considered in the context of following sentences. It helps in interpreting how certain words or phrases focuses on context. For example "It was a great day" is clearer when you know what event or situation is being discussed.

Example of Discourse Integration

1. "Taylor went to the store to buy some groceries. She realized she forgot her wallet." Understanding that "Taylor" is the antecedent of "she" is important for understanding sentence's meaning.
2. "This is unfair!" helps in understand what "this" refers to we need to identify following sentences. Without context statement's meaning remains unclear.

It is important for NLP applications like machine translation, chatbots and text summarization. It ensures that meaning remains same across sentences which helps machines to understand context. This enables accurate and natural responses in applications like conversational AI and document translation.

5. Pragmatic Analysis

Pragmatic analysis helps in understanding the deeper meaning behind words and sentences by looking beyond their literal meanings. While semantic analysis looks at the direct meaning it considers the speaker's or writer's intentions, tone and context of the communication.

Key tasks in pragmatic analysis:

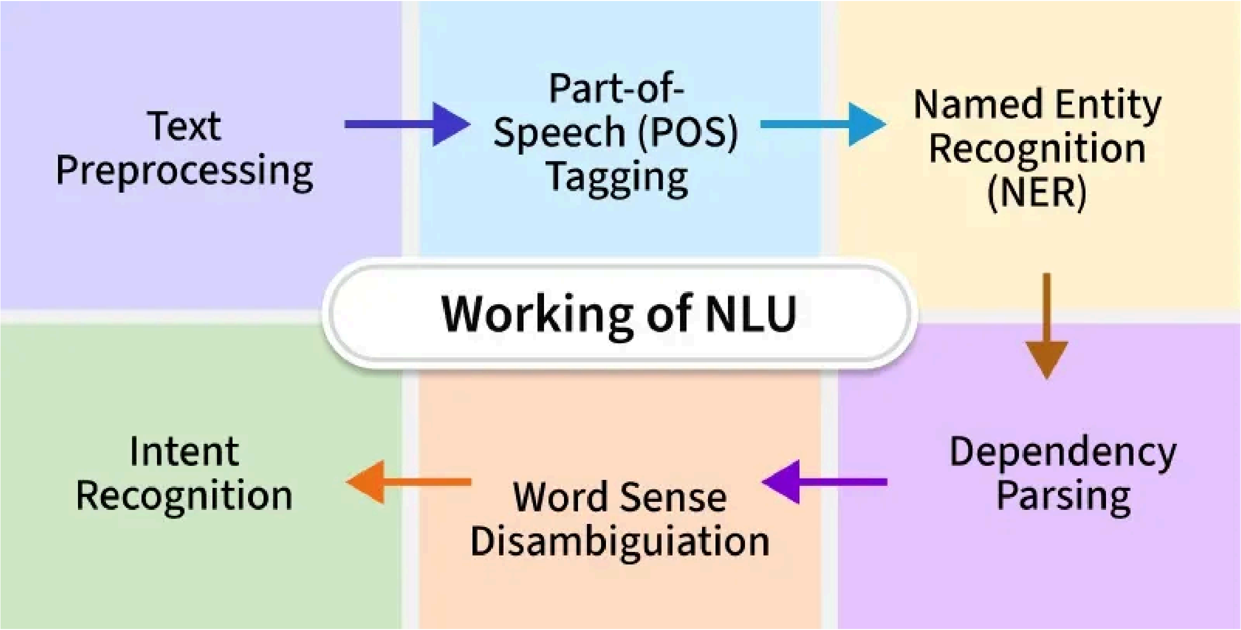
- Understanding Intentions: Sometimes language doesn’t mean what it says literally. For example when someone asks "Can you pass the salt?" it's not about ability but a polite request. It helps to understand true intention behind such expressions.
- Figurative Meaning: Language often uses idioms or metaphors that can’t be taken literally.

Examples of Pragmatic Analysis

- "Hello! What time is it?" here it might be a straightforward request for the current time but it could also imply concern about being late.
- For example "I’m falling for you" means "I love you" not literally falling. It helps to interpret these non-literal meanings.

It is important for NLP tasks like sentiment analysis, chatbots and conversation-based AI. It helps machines to understand the speaker's intentions, tone and context which go beyond the literal meaning of words. By identifying sarcasm and emotions this helps systems to respond naturally which improves human-computer interaction. By combining these phases NLP systems can effectively interpret, analyze and generate human language making more intelligent and natural interactions between humans and machines.

NLU: Step-by-Step Breakdown


Text Preprocessing
Part-of-Speech (POS) Tagging
Named Entity Recognition (NER)
Working of NLU
Dependency Parsing
Word Sense Disambiguiation
Intent Recognition

To understand how Natural Language Understanding processes input, consider the sentence:

"A new mobile will be launched in the upcoming year."

1. Text Preprocessing: The first step is to clean and normalize the input. This involves breaking the sentence into individual words (tokenization), removing common stopwords and also reducing words to their root forms (such as converting “launched” to “launch”). This results in a simplified and more meaningful representation of the sentence.
Output: At this stage, non-essential elements are removed and the text is transformed into a basic list of meaningful words.

2. Part-of-Speech (POS) Tagging: Each word is then assigned a grammatical category such as noun, verb or adjective. This helps identify the function of each word in the sentence.
Output: POS tagging helps the system understand which words serve as subjects, actions, or descriptors, contributing to sentence structure comprehension.

3. Named Entity Recognition (NER): In this phase specific types of information like names of products, dates or locations are identified. In the example sentence, “mobile” may be recognized as a product and “upcoming year” as a time reference.
Output: NER highlights the most informative parts of the sentence, enabling the system to grasp what and when something is being discussed.

4. Dependency Parsing: Dependency parsing examines how words are connected. It identifies which words depend on others to convey meaning. For instance, “mobile” depends on “launched,” and “upcoming year” provides a time reference for that action.

Output: This parsing shows relationships between words, allowing the system to interpret how the sentence is structured semantically.

5. Word Sense Ambiguity: Some words can have multiple meanings depending on context. Here, the word "mobile" could refer to a phone or a moving object. By checking the surrounding words like "launched" and "year" the system shows that the sentence is about a product release, specifically a smartphone.
Output: This step ensures that words are interpreted correctly based on their usage in context.

6. Intent Recognition: Intent recognition identifies the purpose behind the input. In this case, the sentence is likely meant to inform about a product launch. Determining intent is particularly important in dialogue systems where understanding user goals is essential.
Output: The system categorizes the input under an intent like inform_product_release, guiding appropriate actions or responses.

7. Output Generation: Once the sentence is understood, the system formulates a suitable response or action. For instance, it might respond with a confirmation or ask for more details, depending on the context of the conversation.
Output: A response is produced based on the extracted meaning and recognized intent, which helps to maintain a meaningful interaction.

NLG: Reverse of NLU
How does NLG work : A typical NLG pipeline consists of the following stages:

1. Content Determination: The system decides which information from the input data is relevant and should be mentioned. This involves filtering facts based on context or importance.
2. Document Structuring: The content is organized into a continuous structure. Decisions are made about the order in which topics or facts should be presented.
3. Aggregation: Facts are grouped to improve fluency and reduce redundancy. This ensures the text reads naturally, like how a human would summarize multiple data points.
4. Lexicalization: Appropriate words and expressions are chosen to represent the facts.
5. Referring Expression Generation: The system generates references to entities such as "it", "they" or proper names to maintain clarity and consistency across the text.
6. Linguistic Realization: Grammar rules are applied to construct well-formed sentences.

TRANSFORMER MODEL - ENCODER - DECODER - ATTENTION MECHANISM (SENTENCE)

Transformers - link

Why only Transformers: Transformers processes whole sentence using Attention mechanism by reduced layers unlike GRU - Word by word with more number of layers resulting in vanishing gradient descent and in LSTM - though including more layers could eliminate the vanishing gradient descent problem due to gate mechanism but the model can not understand the whole sentence (only word by word).

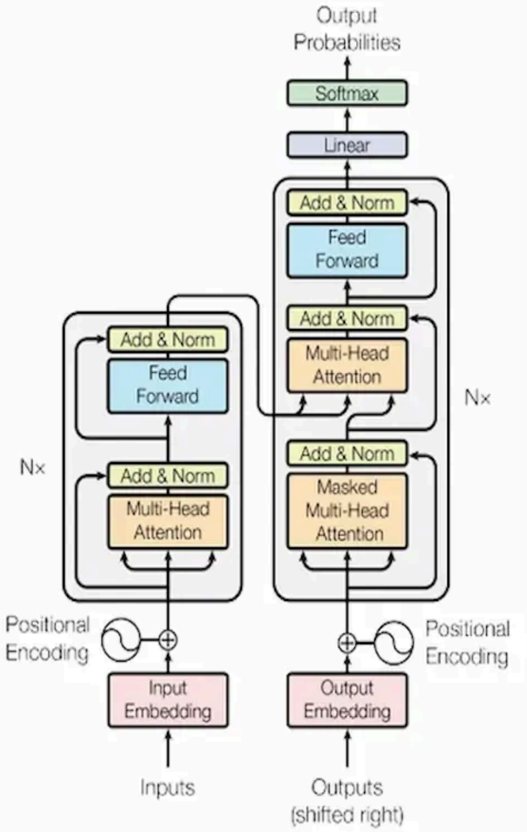


Figure 1: The Transformer - model architecture.

Transformer Model from Scratch using TensorFlow