Module 1: Introduction to Natural Language Processing

- By Tina D’abreo

Content

- ➔ Origin & History of NLP
- ➔ Stages in NLP
- ➔ Ambiguities and its types in English and Indian Regional Languages
- ➔ Applications of NLP:
  - ◆ Machine Translation
  - ◆ Information Retrieval
  - ◆ Question Answering System
  - ◆ Sentiment Analysis
  - ◆ Text Categorization
  - ◆ Text Summarization
  - ◆ Named Entity Recognition

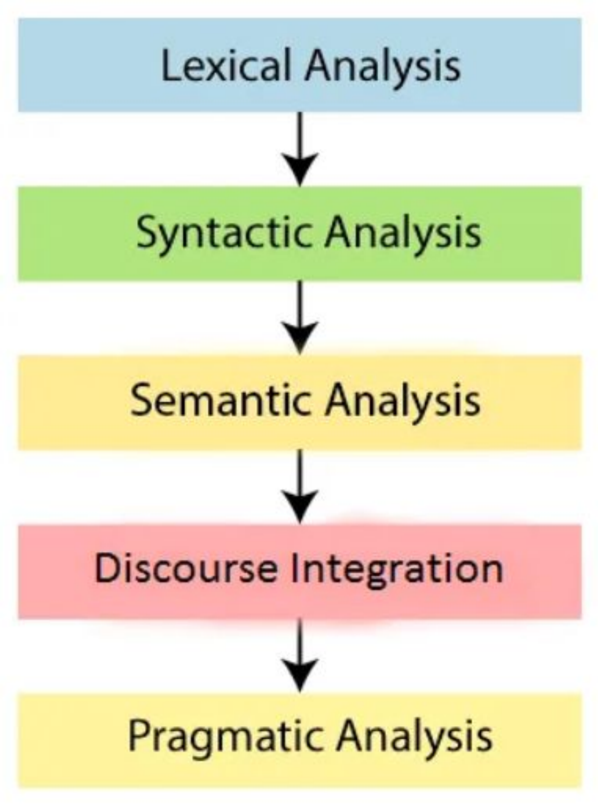
Origin & History of NLP

- **What is NLP?**
  - Natural language processing (NLP) is a subfield of computer science and artificial intelligence (AI) that uses machine learning to enable computers to understand and communicate with human language.
  - Its core objective is to enable computers to understand, analyze, and generate human language in a way that is similar to how humans do. This includes tasks like:
    - ***Understanding the meaning:*** Being able to extract the meaning from text, speech, or other forms of human language.
    - ***Analyzing structure:*** Recognizing the grammatical structure and syntax of language, including parts of speech and sentence construction.
    - ***Generating human-like language:*** Creating text or speech that is natural, coherent, and grammatically correct.

Stages in NLP | Phases of NLP

➔ ***First Phase: Lexical Analysis***

- The lexical phase in Natural Language Processing (NLP) involves scanning text and breaking it down into smaller units such as paragraphs, sentences, and words.
- This process, known as tokenization, converts raw text into manageable units called tokens or lexemes.
- Tokenization is essential for understanding and processing text at the word level.



Stages in NLP | Phases of NLP

- ***First Phase: Lexical Analysis***
  - Various data cleaning and feature extraction techniques are applied, including:
    - ***Lemmatization***: Reducing words to their base or root form.
    - ***Stopwords Removal***: Eliminating common words that do not carry significant meaning, such as “and,” “the,” and “is.”
    - ***Correcting Misspelled Words***: Ensuring the text is free of spelling errors to maintain accuracy.
- ***Second Phase: Syntactic Analysis***
    - Syntactic analysis, also known as parsing, is the second phase of Natural Language Processing (NLP).
    - This phase is essential for understanding the structure of a sentence and assessing its grammatical correctness.
    - It involves analyzing the relationships between words and ensuring their logical consistency by comparing their arrangement against standard grammatical rules.

Stages in NLP | Phases of NLP

➔ ***Second Phase: Syntactic Analysis***

- Parsing examines the grammatical structure and relationships within a given text.
- It assigns Parts-Of-Speech (POS) tags to each word, categorizing them as nouns, verbs, adverbs, etc.
- This tagging is crucial for understanding how words relate to each other syntactically and helps in avoiding ambiguity.
- Ambiguity arises when a text can be interpreted in multiple ways due to words having various meanings. For example, the word “book” can be a noun (a physical book) or a verb (the action of booking something), depending on the sentence context.
    - ○ Example :Correct Syntax: “John eats an apple.”
    - ○ Incorrect Syntax: “Apple eats John an.”
- ***POS Tags:*** John: Proper Noun (NNP), eats: Verb (VBZ), an: Determiner (DT), apple: Noun (NN)

Stages in NLP | Phases of NLP

➔ ***Third Phase: Semantic Analysis***

- Semantic Analysis is the third phase of Natural Language Processing (NLP), focusing on extracting the meaning from text.
- Unlike syntactic analysis, which deals with grammatical structure, semantic analysis is concerned with the literal and contextual meaning of words, phrases, and sentences.
- Semantic analysis aims to understand the dictionary definitions of words and their usage in context.
- It determines whether the arrangement of words in a sentence makes logical sense.
- This phase helps in finding context and logic by ensuring the semantic coherence of sentences.
- Named Entity Recognition(NER) and Word Sense Disambiguation(WSD) are key task of this phase

Stages in NLP | Phases of NLP

- ***Fourth Phase: Discourse Integration***
  - This phase deals with comprehending the relationship between the current sentence and earlier sentences or the larger context.
  - Discourse integration is crucial for contextualizing text and understanding the overall message conveyed.
  - Discourse integration examines how words, phrases, and sentences relate to each other within a larger context.
  - It assesses the impact a word or sentence has on the structure of a text and how the combination of sentences affects the overall meaning.
  - This phase helps in understanding implicit references and the flow of information across sentences.

Stages in NLP | Phases of NLP

➔ ***Fifth Phase: Pragmatic Analysis***

- This phase focuses on interpreting the inferred meaning of a text beyond its literal content.
- Human language is often complex and layered with underlying assumptions, implications, and intentions that go beyond straightforward interpretation.
- This phase aims to grasp these deeper meanings in communication.

Ambiguities and its types in English and Indian Regional Languages

- ➔ Natural language ambiguity refers to situations where a word, phrase, or sentence has multiple meanings, making it challenging to interpret correctly.
- ➔ Natural language ambiguity refers to the fact that human languages often have words and sentences that can have multiple meanings or interpretations.
- ➔ This ambiguity arises because language reflects the complexities and subtleties of human experience.
- ➔ Ambiguity is present in all the steps of NLP

Ambiguities and its types in English and Indian Regional Languages

- Types of Ambiguities in NLP:
  - Lexical Ambiguity
  - Syntactic Ambiguity
  - Semantic Ambiguity
  - Anaphoric Ambiguity
  - Pragmatic Ambiguity

Ambiguities and its types in English and Indian Regional Languages

➔ ***Lexical Ambiguity:***

- Lexical means relating to words of a language. During Lexical analysis given paragraphs are broken down into words or tokens.
- Each token has got specific meaning. There can be instances where a single word can be interpreted in multiple ways.
- The ambiguity that is caused by the word alone rather than the context is known as Lexical Ambiguity.
- Example: **I <u>saw</u> a <u>bat.</u>**
  - Saw - past tense of verb see and present tense of verb saw(sawing /cutting)
  - Bat - a nocturnal animal or a playing cricket bat.
- Handled by POS tagging and Word Sense Disambiguation

Ambiguities and its types in English and Indian Regional Languages

- ***Syntactic Ambiguity:***
  - Syntactic meaning refers to the grammatical structure and rules that define how words should be combined to form sentences and phrases.
  - A sentence can be interpreted in more than one way due to its structure or syntax such ambiguity is referred to as Syntactic Ambiguity.
    - Example: **"Old men and women"**
    - The above sentence can have two possible meanings:
      - "*All old men and young women*" and "*all old men and old women*"
    - Example : **"John saw the boy with telescope."**
    - In the above case, two possible meanings are
      - "*John saw the boy through his telescope*" and "*John saw the boy who was holding the telescope.*"

Ambiguities and its types in English and Indian Regional Languages

- ***Semantic Ambiguity:***
  - Semantics is nothing but “Meaning”. The semantics of a word or phrase refers to the way it is typically understood or interpreted by people.
  - Syntax describes the rules by which words can be combined into sentences, while semantics describes what they mean.
  - Semantic Ambiguity occurs when a sentence has more than one interpretation or meaning.
  - ***Example 1***: **“Seema loves her mother and Sriya does too.”**
    - The interpretations can be Sriya loves Seema’s mother or Sriya likes her mother.
  - ***Example 2:*** **“The dog has been domesticated for 1000 years.”**
    - The above sentence can be interpreted as either a particular dog is domesticated or the dog species is being domesticated.

Ambiguities and its types in English and Indian Regional Languages

- ***Anaphoric Ambiguity:***
  - A word that gets its meaning from a preceding word or phrase is called an anaphor.
  - Example: “Susan plays the piano. She likes music.”
    - In this example, the word she is an anaphor and refers back to a preceding expression i.e., Susan.
    - The linguistic element or elements to which an anaphor refers is called an antecedent.
    - The relationship between anaphor and antecedent is termed ‘anaphora’. ‘Anaphora resolution’ or ‘anaphor resolution’ is the process of finding the correct antecedent of an anaphor.

Ambiguities and its types in English and Indian Regional Languages

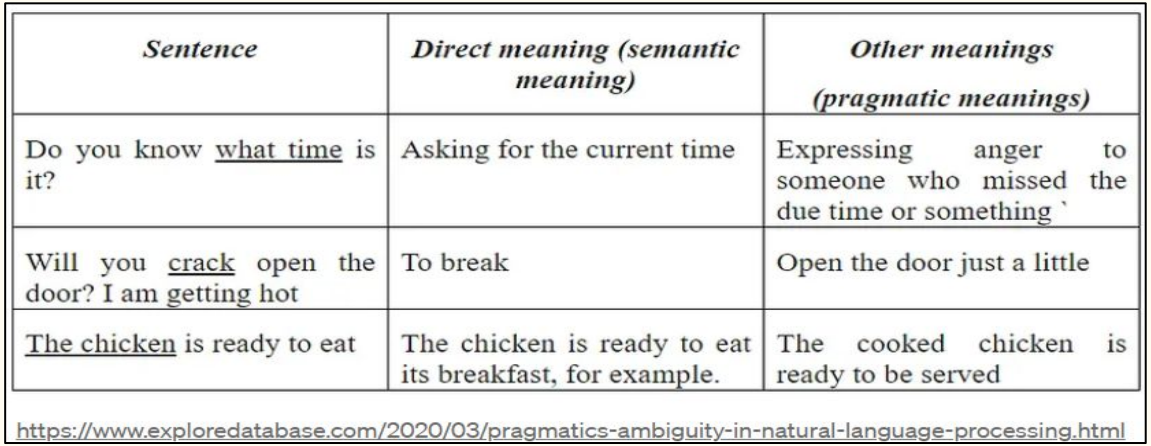
- ***Anaphoric Ambiguity:***
  - Ambiguity that arises when there is more than one reference to the antecedent is known as Anaphoric Ambiguity.
  - ***Example 1:*** “The horse ran up the hill. It was very steep. It soon got tired.”
    - In this example, there are two ‘it’, and it is unclear to which each ‘it’ refers, this leads to Anaphoric Ambiguity.
    - The sentence will be meaningful if first ‘it’ refers to the hill and 2nd ‘it’ refers to the horse.
    - Anaphors may not be in the immediately previous sentence. They may present in the sentences before the previous one or may present in the same sentence.

Ambiguities and its types in English and Indian Regional Languages

- ***Anaphoric Ambiguity:***
  - Anaphoric references may not be explicitly present in the previous sentence rather they might refer to the part of the antecedent.
  - ***Example 2:*** “I went to the hospital, and they told me to go home and rest.”
    - In this sentence, ‘they’ does not explicitly refer to the hospital instead it refers to the Dr or staff who attended the patient in the hospital.
  - Anaphors are mostly pronouns, or they can even be noun phrases in some instances.
  - ***Example 3:*** “Darshan plays keyboard. He loves music. “
    - In this case ‘He’ is a pronoun.
  - ***Example 4:*** “A puppy drank the milk. The cute little dog was satisfied.”
    - Here Anaphora is ‘cute little dog’ which is a noun phrase.

Ambiguities and its types in English and Indian Regional Languages

- ***Pragmatic Ambiguity:***
  - Pragmatics focuses on the real-time usage of language like what the speaker wants to convey and how the listener infers it.
  - Situational context, the individual's mental states, the preceding dialogue, and other elements play a major role in understanding what the speaker is trying to say and how the listeners perceive it.
  -



Sentence	*Direct meaning (semantic meaning)*	*Other meanings (pragmatic meanings)*
Do you know what time is it?	Asking for the current time	Expressing anger to someone who missed the due time or something `
Will you crack open the door? I am getting hot	To break	Open the door just a little
The chicken is ready to eat	The chicken is ready to eat its breakfast, for example.	The cooked chicken is ready to be served

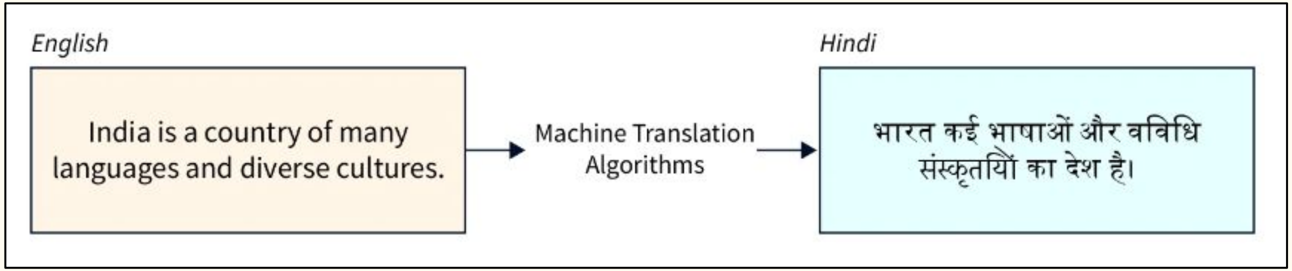
https://www.exploredatabase.com/2020/03/pragmatics-ambiguity-in-natural-language-processing.html

Ambiguities and its types in English and Indian Regional Languages

➔ ***Difference between Semantic and Pragmatic:***

- Semantics and pragmatics are two fields of linguistics. Both of them concern with study of meaning of humans speech signs.
- However, each of which tackles meaning from a different angle.
- Semantics pays attention to the literal meaning of words (dictionary meaning), whereas pragmatics concerns with the intended meaning of an utterance (what does the speaker mean?).

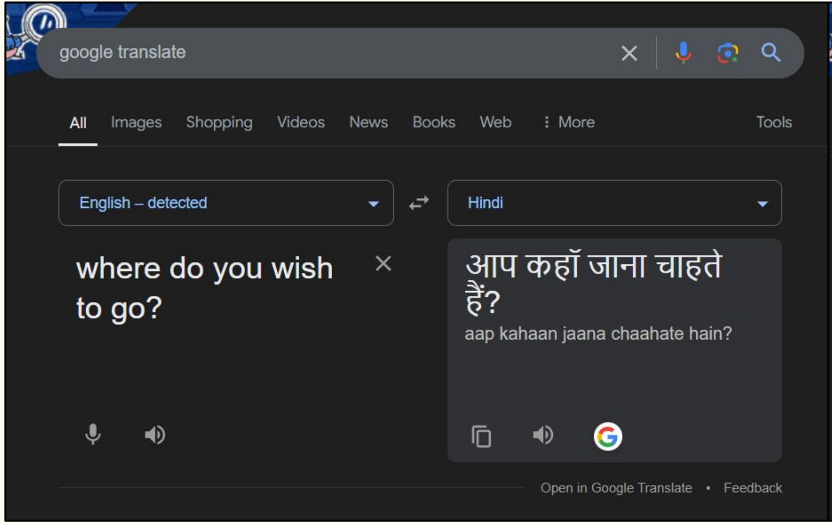
Applications of NLP

- ***Machine Translation:***
  - Machine Translation (MT) is a domain of computational linguistics that uses computer programs to translate text or speech from one language to another with no human involvement with the goal of relatively high accuracy, low errors, and effective cost.
  - The demand for translation is majorly due to the exponential rise increase in the exchange of information between various regions using different regional languages.
  - Examples such as access to web documents in non-native languages, using products from across countries, real-time chat, legal literature, etc. are some use cases.
  - 
English
India is a country of many languages and diverse cultures.
Machine Translation Algorithms
Hindi
भारत कई भाषाओं और वविधि संस्कृतयिों का देश है।

Applications of NLP

➔ ***Machine Translation:***

- Examples of Application: Google Translate, IBM Watson
-



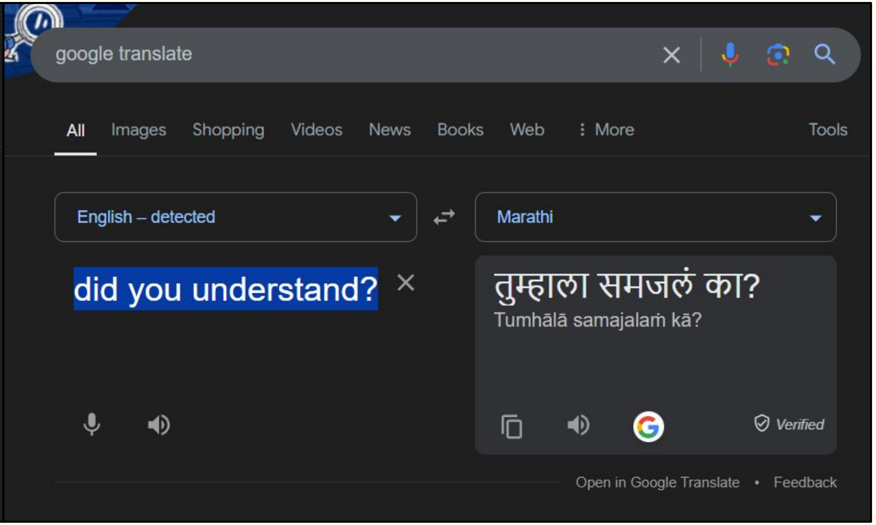
google translate

English – detected → Hindi

where do you wish to go?

आप कहाँ जाना चाहते हैं?

aap kahaan jaana chaahate hain?



google translate

English – detected → Marathi

did you understand?

तुम्हाला समजलं का?

Tumhālā samajalaṁ kā?

Applications of NLP

- ***Information Retrieval:***
  - Information retrieval (IR) may be defined as a software program that deals with the organization, storage, retrieval and evaluation of information from document repositories particularly textual information.
  - The documents that satisfy user's requirement are called relevant documents. A perfect IR system will retrieve only relevant documents.
  - Examples:
    - Digital Libraries
    - Media Search
    - Search Engines

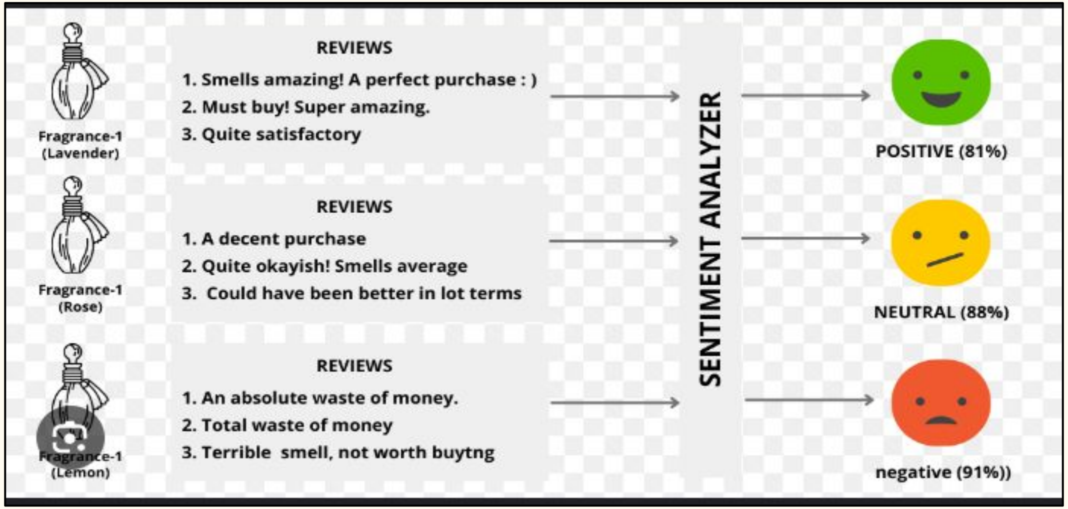
Applications of NLP

➔ ***Question Answering Systems:***

- Question answering is a critical NLP problem and a long-standing artificial intelligence milestone.
- QA systems allow a user to express a question in natural language and get an immediate and brief response.
- For example, after being asked, “how warm is it going to be today?” your ***Siri*** can extract raw information about today’s temperature from a weather service. In addition, instead of showing it to you as is, it processes the data and presents it to in proper English (or in any other supported language). Similarly, Alexa
- Two notable examples of earliest question answering systems were ***LUNAR and BASEBALL***. LUNAR answered questions about rocks that were analyzed during the Apollo Lunar missions. BASEBALL on the other hand answered questions about baseball league over a period of one year. LUNAR and BASEBALL were good at their respective domains. Both of the systems used the techniques used in chatterbot systems.

Applications of NLP

- ***Sentiment Analysis:***
  - Sentiment analysis is a popular task in natural language processing. The goal of sentiment analysis is to classify the text based on the mood or mentality expressed in the text, which can be positive negative, or neutral.
  - Sentiment analysis, also known as opinion mining, is an important business intelligence tool that helps companies



REVIEWS

1. Smells amazing! A perfect purchase :)
2. Must buy! Super amazing.
3. Quite satisfactory

REVIEWS

1. A decent purchase
2. Quite okayish! Smells average
3. Could have been better in lot terms

REVIEWS

1. An absolute waste of money.
2. Total waste of money
3. Terrible smell, not worth buytng

Applications of NLP

- ***Text Categorization:***
  - Text categorization is one of the most common tasks in NLP. It is the process of assigning a label or category to a given piece of text.
  - For example, we can classify emails as spam or not spam, tweets as positive or negative, and articles as relevant or not relevant to a given topic.
  - Examples:
    - Spam detection in emails
    - Sentiment analysis of online reviews
    - Topic labeling documents like research papers
    - Language detection like in Google Translate
    - Age/gender identification of anonymous users
    - Tagging online content

Applications of NLP

➔ ***Text Summarization:***

- ◆ Text summarization is the process of generating short, fluent, and most importantly accurate summary of a respectively longer text document. The main idea behind automatic text summarization is to be able to find a short subset of the most essential information from the entire set and present it in a human-readable format.
- ◆ Examples: Text Summary - TLDR Summarize, Text Summarizer, etc. in the domain of business, news, academics , etc.

Applications of NLP

➔ ***Named Entity Recognition:***

- A named entity is basically a real-life object which has proper identification and can be denoted with a proper name. Named Entities can be a place, person, organization, time, object, or geographic entity.
- For example, named entities would be Roger Federer, Honda city, Samsung Galaxy S10. Named entities are usually instances of entity instances.
- For example, Roger Federer is an instance of a Tennis Player/person, Honda City is an instance of a car and Samsung Galaxy S10 is an instance of a Mobile Phone.
- Also, chatbots, sentiment analysis tools and search engines.