Natural Language Processing (NLP) Unit-I

1. Natural Language Processing – Introduction

- Humans communicate through some form of language either by text or speech.
- To make interactions between computers and humans, computers need to understand natural languages used by humans.
- Natural language processing is all about making computers learn, understand, analyze, manipulate and interpret natural(human) languages.
- NLP stands for Natural Language Processing, which is a part of Computer Science, Human languages or Linguistics, and Artificial Intelligence.
- Processing of Natural Language is required when you want an intelligent system like robot to perform as per your instructions, when you want to hear decision from a dialogue based clinical expert system, etc.
- The ability of machines to interpret human language is now at the core of many applications that we use every day - chatbots, Email classification and spam filters, search engines, grammar checkers, voice assistants, and social language translators.
- The input and output of an NLP system can be Speech or Written Text.

2. Applications of NLP or Use cases of NLP

1. Sentiment analysis

- **Sentiment analysis**, also referred to as **opinion mining**, is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text.
- This is a popular way for organizations to determine and categorize opinions about a product, service or idea.
- Sentiment analysis systems help organizations gather insights into real-time customer sentiment, customer experience and brand reputation.
- Generally, these tools use text analytics to analyze online sources such as emails, blog posts, online reviews, news articles, survey responses, case studies, web chats, tweets, forums and comments.
- Sentiment analysis uses machine learning models to perform text analysis of human language. The metrics used are designed to detect whether the overall sentiment of a piece of text is positive, negative or neutral.

2. Machine Translation

- **Machine translation**, sometimes referred to by the abbreviation **MT**, is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another.

- On a basic level, MT performs mechanical substitution of words in one language for words in another, but that alone rarely produces a good translation because recognition of whole phrases and their closest counterparts in the target language is needed.

- Not all words in one language have equivalent words in another language, and many words have more than one meaning.

- Solving this problem with corpus statistical and neural techniques is a rapidly growing field that is leading to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies.
- **Corpus:** A collection of written texts, especially the entire works of a particular author.

3. Text Extraction

- There are a number of natural language processing techniques that can be used to extract information from text or unstructured data.
- These techniques can be used to extract information such as entity names, locations, quantities, and more.
- With the help of natural language processing, computers can make sense of the vast amount of unstructured text data that is generated every day, and humans can reap the benefits of having this information readily available.
- Industries such as healthcare, finance, and e-commerce are already using natural language processing techniques to extract information and improve business processes.
- As the machine learning technology continues to develop, we will only see more and more information extraction use cases covered.

4. Text Classification

- Unstructured text is everywhere, such as emails, chat conversations, websites, and social media. Nevertheless, it’s hard to extract value from this data unless it’s organized in a certain way.
- Text classification also known as *text tagging* or *text categorization* is the process of categorizing text into organized groups. By using Natural Language Processing (NLP), text classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content.
- Text classification is becoming an increasingly important part of businesses as it allows to easily get insights from data and automate business processes.

5. Speech Recognition

- Speech recognition is an interdisciplinary subfield of computer science and computational linguistics that develops methodologies and technologies that enable the recognition and translation of spoken language into text by computers.
- It is also known as automatic speech recognition (ASR), computer speech recognition or speech to text (STT).
- It incorporates knowledge and research in the computer science, linguistics and computer engineering fields. The reverse process is speech synthesis.

Speech recognition use cases

- A wide number of industries are utilizing different applications of speech technology today, helping businesses and consumers save time and even lives. Some examples include:
- Automotive: Speech recognizers improves driver safety by enabling voice-activated navigation systems and search capabilities in car radios.
- Technology: Virtual agents are increasingly becoming integrated within our daily lives, particularly on our mobile devices. We use voice commands to access them through our smartphones, such as through Google Assistant or Apple’s Siri, for tasks, such as voice search, or through our speakers, via Amazon’s Alexa or Microsoft’s Cortana, to play music. They’ll only continue to integrate into the everyday products that we use, fueling the “Internet of Things” movement.
- Healthcare: Doctors and nurses leverage dictation applications to capture and log patient diagnoses and treatment notes.
- Sales: Speech recognition technology has a couple of applications in sales. It can help a call center transcribe thousands of phone calls between customers and agents to identify common call patterns and issues. AI chatbots can also talk to people via a webpage, answering common queries and solving basic requests without needing to wait for a contact center agent to be available. In both instances speech recognition systems help reduce time to resolution for consumer issues.

6. Chatbot

- Chatbots are computer programs that conduct automatic conversations with people. They are mainly used in customer service for information acquisition. As the name implies, these are bots designed with the purpose of chatting and are also simply referred to as “bots.”

- You’ll come across chatbots on business websites or messengers that give pre-scripted replies to your questions. As the entire process is automated, bots can provide quick assistance 24/7 without human intervention.

7. Email Filter

- One of the most fundamental and essential applications of NLP online is email filtering. It began with spam filters, which identified specific words or phrases that indicate a spam message. But, like early NLP adaptations, filtering has been improved.
- Gmail's email categorization is one of the more common, newer implementations of NLP. Based on the contents of emails, the algorithm determines whether they belong in one of three categories (main, social, or promotional).
- This maintains your inbox manageable for all Gmail users, with critical, relevant emails you want to see and reply to fast.

8. Search Autocorrect and Autocomplete

- When you type 2-3 letters into Google to search for anything, it displays a list of probable search keywords. Alternatively, if you search for anything with mistakes, it corrects them for you while still returning relevant results. Isn't it incredible?

- Everyone uses Google search autocorrect autocomplete on a regular basis but seldom gives it any thought. It's a fantastic illustration of how natural language processing is touching millions of people across the world, including you and me.
- Both, search autocomplete and autocorrect make it much easier to locate accurate results.

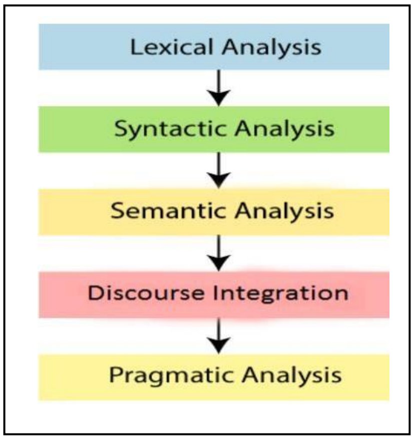
3. Components of NLP

- ➢ There are two components of NLP, Natural Language Understanding (NLU)and Natural Language Generation (NLG).
- ➢ Natural Language Understanding (NLU) which involves transforming humanlanguage into a machine-readable format.It helps the machine to understand and analyze human language by extracting the text from large data such as keywords, emotions, relations, and semantics.
- ➢ Natural Language Generation (NLG) acts as a translator that converts thecomputerized data into natural language representation.
- ➢ It mainly involves Text planning, Sentence planning, and Text realization.
- ➢ The NLU is harder than NLG.

4. Steps in NLP

There are general five steps :

- 1. Lexical Analysis
- 2. Syntactic Analysis (Parsing)
- 3. Semantic Analysis
- 4. Discourse Integration
- 5. Pragmatic Analysis



Lexical Analysis:

- ➢ The first phase of NLP is the Lexical Analysis.
- ➢ This phase scans the source code as a stream of characters and converts it into meaningful lexemes.
- ➢ It divides the whole text into paragraphs, sentences, and words.
- ➢ Lexeme: A lexeme is a basic unit of meaning. In linguistics, the abstract unit of morphological analysis that corresponds to a set of forms taken by a single word is called lexeme.
- ➢ The way in which a lexeme is used in a sentence is determined by its grammatical category.

- Lexeme can be individual word or multiword.
- For example, the word talk is an example of an individual word lexeme, which mayhave many grammatical variants like talks, talked and talking.
- Multiword lexeme can be made up of more than one orthographic word. For example,speak up, pull through, etc. are the examples of multiword lexemes.

Syntax Analysis (Parsing)

- Syntactic Analysis is used to check grammar, word arrangements, and shows therelationship among the words.
- The sentence such as “The school goes to boy” is rejected by English syntactic analyzer.

Semantic Analysis

- Semantic analysis is concerned with the meaning representation.
- It mainly focuses on the literal meaning of words, phrases, and sentences.
- The semantic analyzer disregards sentence such as “hot ice-cream”.
- Another Example is “Manhattan calls out to Dave” passes a syntactic analysis because it’s a grammatically correct sentence. However, it fails a semantic analysis. Because Manhattan is a place (and can’t literally call out to people), the sentence’s meaning doesn’t make sense.

Discourse Integration

- Discourse Integration depends upon the sentences that precedes it and also invokesthe meaning of the sentences that follow it.
- For instance, if one sentence reads, “Manhattan speaks to all its people,” and the following sentence reads, “It calls out to Dave,” discourse integration checks the first sentence for context to understand that “It” in the latter sentence refers to Manhattan.

Pragmatic Analysis

- During this, what was said is re-interpreted on what it actually meant.
- It involves deriving those aspects of language which require real world knowledge.
- For instance, a pragmatic analysis can uncover the intended meaning of “Manhattan speaks to all its people.” Methods like neural networks assess the context to understand that the sentence isn’t literal, and most people won’t interpret it as such. A pragmatic analysis deduces that this sentence is a metaphor for how people emotionally connect with place.

5. Finding the structure of Words

Words and Their Components

- Words are defined in most languages as the smallest linguistic units that can form acomplete utterance by themselves.
- The minimal parts of words that deliver aspects of meaning to them are called **morphemes**.

Tokens:

Suppose, for a moment, that words in English are delimited only by whitespace and punctuation (the marks, such as full stop, comma, and brackets)

- Example: Will you read the newspaper? Will you read it? I won’t read it. If we confront our assumption with insights from syntax, we notice twowords here: words ***newspaper*** and ***won’t***.