Natural Language Processing

Contents

Natural Language?

Refers to the **language spoken by people**,

i.e English, Spanish, japanese as opposed

to artificial languages like C++, Java etc.

Natural Language Processing?

Natural Language Processing is the science of teaching **Machines** how to understand the language we **Human** speak and write.

Aim of NLP?

To build intelligent system that can interact

With **Human Being** like a **Human Being**.

Where is NLP?

NLP is Everywhere even if we don't know it. And although NLP applications can rarely achieve performances of 100%, still they are part of our lives and provide a precious help to all of us.

- ❑ Machine translation
- ❑ Automatic summarization
- ❑ Sentiment analysis
- ❑ Text classification
- ❑ Language detection
- ❑ Information Extraction & lot more

Input & Output of NLP

NLP involves making computers to perform useful tasks with the natural languages humans use. The input and output of an NLP system can be :

- Speech
- Written Text

Components of NLP

Natural Language Understanding

NLU involves the following tasks :

- **Analyzing** different aspects of the language.

- **Mapping** the given input in natural language into useful representations.

- The NLU is harder than NLG.

Natural Language Generation

NLG involves the following tasks :

- **Text planning** : It includes retrieving the relevant content from knowledge base.

- **Sentence planning** : It includes choosing required words, forming meaningful phrases, setting tone of the sentence.

- **Text Realization** : It is mapping sentence plan into sentence structure.

NLP Terminology

❑ **Phonology** : It is study of organizing sound patterns and their meanings.

❑ **Morphology** : It is a study of construction of words from primitive meaningful units.

❑ **Morpheme** : A meaningful morphological unit of a language that cannot be further divided (e.g. *in*, *come*, *-ing*, forming *incoming*)

❑ **Syntax** : It refers to arranging words to make a sentence. It also involves determining the structural role of words in the sentence and in phrases.

- **Semantics** : It is concerned with the meaning of words and how to combine words into meaningful phrases and sentences.

- **Pragmatics** : It deals with using and understanding sentences in different situations and how the interpretation of the sentence is affected. (e.g. Young men and women)

- **Discourse** : It deals with how the immediately preceding sentence can affect the interpretation of the next sentence.

- **World Knowledge** : It includes the general knowledge about the world.

Steps in NLP
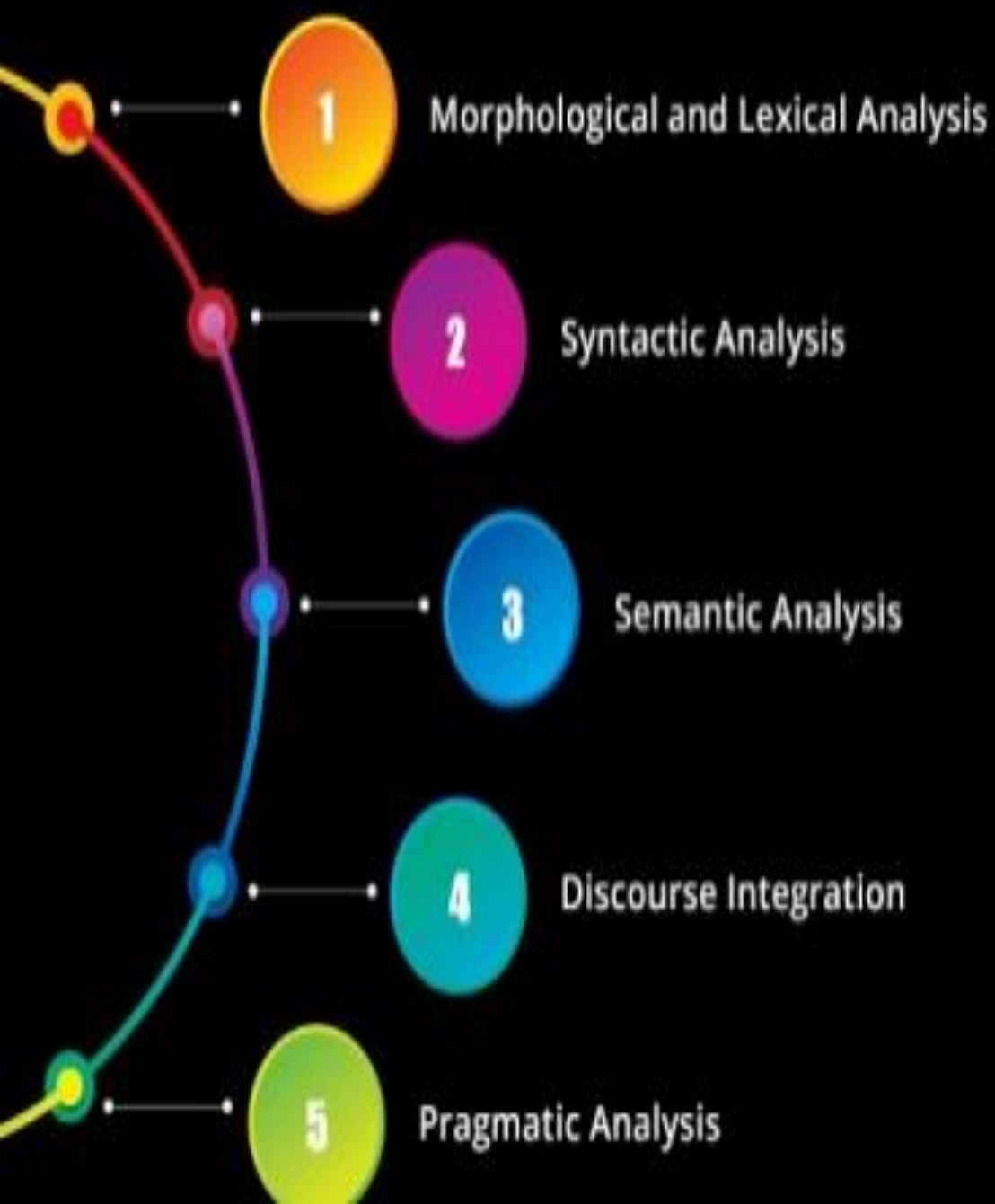
Detailed study Of NLP Process
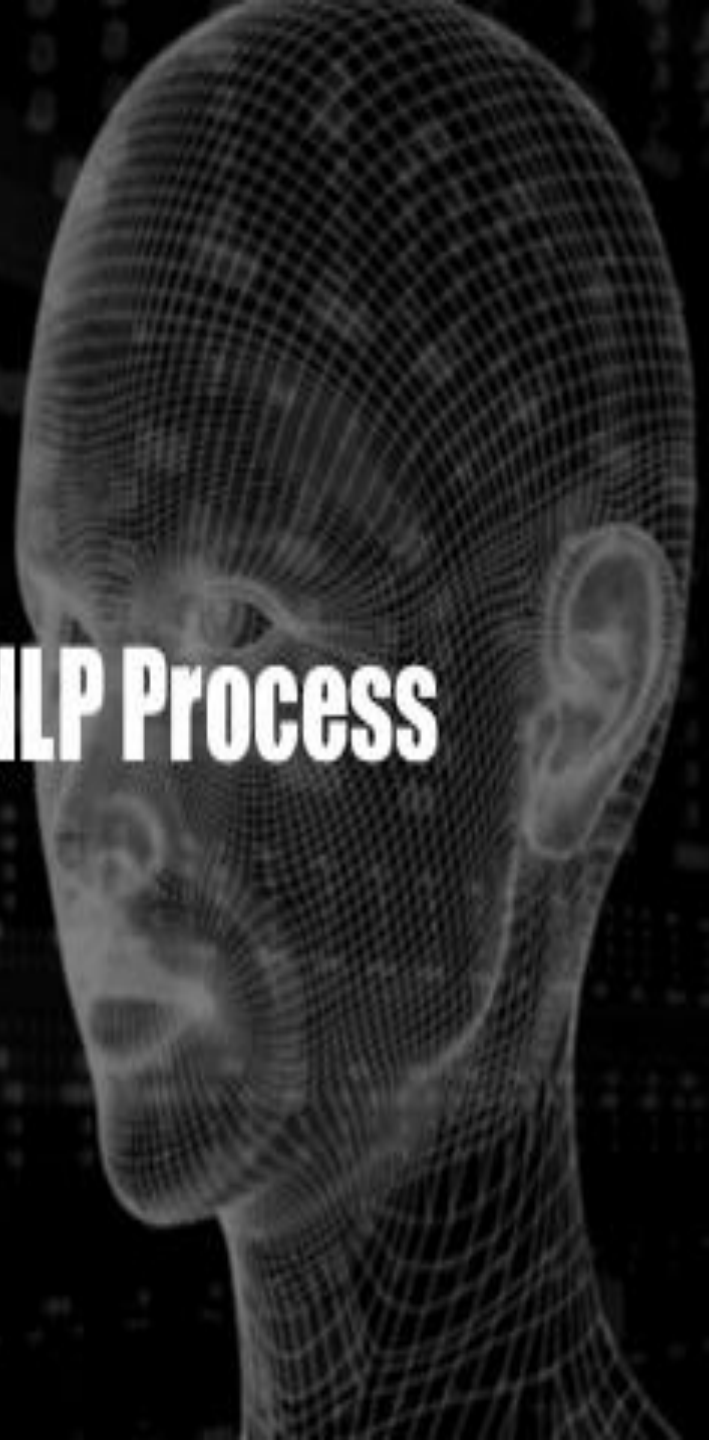
❏ **Lexical Analysis** : Lexical analysis is dividing the whole chunk of text into paragraphs, sentences, and words.

❏ **Syntactic Analysis (Parsing)** : It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. i.e, Boy the go the to store

❏ **Semantic Analysis** : It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures. i.e, colourless green idea

❏ **Discourse Integration** : The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

I.e, Kate wanted it.

❏ **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.

I.e, Do you know what day it is?

Developers Use of NLP Algorithms?

- By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition etc.

- Algorithms are available with various useful applications on many platforms using : Ruby, Python, Java, JavaScript, NodeJs, Swift & so on.

- Instead of hand-coding large sets of rules, NLP can rely on machine learning to automatically learn these rules by analyzing a set of examples.

- The more data analyzed, the more accurate the model will be.

Let us look into
some practical Applications

1. Summarize blocks of Text

Summarizer extracts the most important and central ideas while ignoring irrelevant information.

❏ **INPUTS :**

(Required): Large block of text.

(Optional): Number of sentences (default=3)

❏ **OUTPUT :**

Summarized block of text.

2. Auto tag

Automatically generate keyword tags : A technique that discovers topics contained within a body of text.

- ❏ INPUTS :

  (Required): Raw text.

- ❏ OUTPUT :

  List of extracted tags.

3. Named Entity Recognition

Identify the type of entity extracted, such as it being a person, place, or organization using Named Entity Recognition.

❏ INPUTS :

```
{
    "document": text
}
```

❏ OUTPUT :

```
{
    "sentences": List
    [{
        "detectedEntities": List
        [{
            "word": String,
            "entity": String
        }]
    }]
```

4. Sentiment Analysis

Identify and extract sentiment in given string. This algorithm takes an input string and assigns a sentiment rating in the range [-1 t0 1] (very negative to very positive).

❏ INPUTS :

```
{
    "document": text
}
```

❏ OUTPUTS :

```
[
    {
        "sentiment": 0.474,
        "document": "I really like
eating ice cream in the morning!
    }
]
```

5. Summarize URL

This is a algorithm for summarizing content in webpages. This algorithm utilizes two algorithm: util/Html2Text and nlp/Summarizer. It retrieves the html content in a smart wa by using heuristics with Html2Text algorithm, and summarizes the content using the Summarizer algorithm.

❏ INPUTS :

(Required): String URL.

(Optional): Number of sentences. (default=3)

❏ OUTPUT :

List of extracted tags.

Some Other Applications

- Spelling Correction, Suggests spelling corrections (aka. Autocorrect) for English words.

- Detect profanity in text automatically. It is word-based only. It may miss some words, miss certain misspellings, or double entendres and other material that is offensive in context.

- Create a chat bot using Parsey McParseface, a language parsing deep learning model made by Google that uses Point-of-Speech tagging.

- Reduce words to their root, or stem, using PorterStemmer, or break up text into tokens using Tokenizer.

NLP Libraries

These libraries provide the algorithmic building blocks of NLP in real-world applications.

- **Algorithmia** provides a free API endpoint for many algorithms.

- **Apache OpenNLP**: a machine learning toolkit that provides tokenizers, sentence segmentation, part-of-speech tagging, named entity extraction, parsing and more.

- **Natural Language Toolkit (NLTK)**: a Python library that provides modules for processing text, classifying, tokenizing, tagging, parsing, and more.

- **Standford NLP**: a suite of NLP tools that provide part-of-speech tagging, the named entity recognizer, coreference resolution system, sentiment analysis, and more.

- **Google SyntaxNet** a neural-network framework for analyzing and understanding the grammatical structure of sentences.

Difficulties Of NLP

Human language is rarely precise, or plainly spoken. To understand human language is to understand not only the words, but the concepts and how they're linked together to create meaning.

Despite language being one of the easiest things for humans to learn, the ambiguity of language is what makes NLP a difficult problem for computers to master.

❏ **Lexical ambiguity** : It is at very primitive level such as word-level.

For example, treating the word "board" as noun or verb?

- **Syntax Level ambiguity** : A sentence can be parsed in different ways.

  For example, "He lifted the beetle with red cap." Did he use cap to lift the beetle or he lifted a beetle that had red cap?

- **Referential ambiguity** : Referring to something using pronouns.

  For example, Rima went to Gauri. She said, "I am tired." – Exactly who is tired?

- One input can mean different meanings.

- Many inputs can mean the same thing.

Innovations

- Last year Google had announced the public beta launch of its Cloud Natural Language API which has, sentimental analysis, syntax detection, entity recognition etc,

  i.e the Vision API and the Translate API.

- Vision API has features : Label Detection, Face Detection, Explicit Content Detection, Image Attributes, Logo Detection, Landmark Detection, Optical Character Recognition.

- Translate API feature : Text Translation, Language Detection.

- Vision API & Translation API pricing is based on usage.

- A new foundational machine learning framework used across Apple products, including Siri, Camera, and QuickType has made available for developers via Core ML.

- Core ML delivers fast performance with easy integration using just a few lines of code.

- Supported features include face tracking, face detection, landmarks, text detection, rectangle detection, barcode detection, object tracking, and image registration.

- Core ML Tools is a python package.

- Core ML Model provides Places205-GoogLeNet, ResNet50, Inception v3, VGG16.

- Requirements Xcode 9 and iOS 11

Expansion Domains of NLP

- Medical

- Forensic Science

- Advertisement

- Education

- Business Development

- Marketing & where ever we use language

Research Areas

- Text Processing
- Speech Processing
- Text to Speech
- Automatic Speech Recognition
- Speech to Speech Translation
- Commercial machine translation systems

- Search Engines
- Advanced text Editors
- Information extraction
- Fraud detection
- Opinion mining
- Sentimental analysis