Module-I (Part-1): Introduction to Natural Language Processing

by:
Dr. Soumya Priyadarsini Panda
Sr. Assistant Professor
Dept. of CSE, SIT, Bhubaneswar

Contents

- Introduction
- Need of Processing natural languages
- Applications of NLP
- A brief history of NLP application development
- Issues and processing complexities
- Phases of Natural Language Processing

Introduction

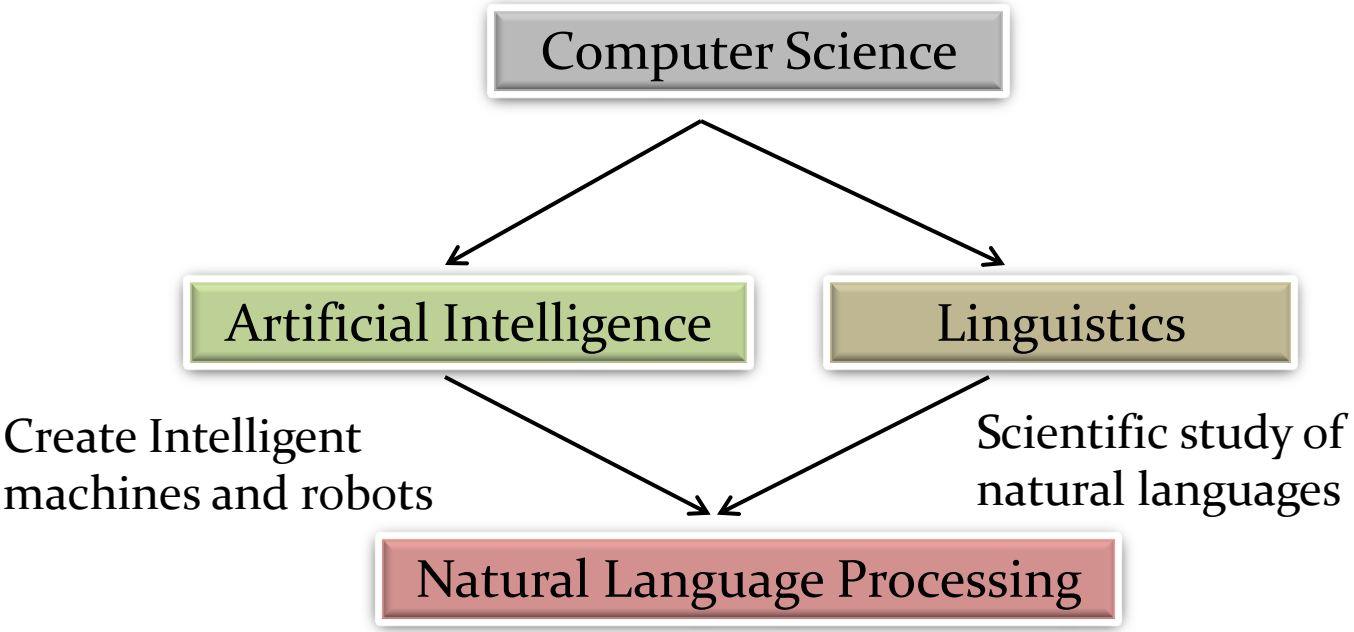
Natural Language:

- Language spoken and written by human for general purpose communication
  - Examples : Odia, Hindi, English, French, Chinese, ……, etc

Natural Language Processing (NLP):

- Focuses on processing, understanding and generating natural languages by machines.

Overview of the NLP Field



- Designing intelligent machines that understands and generates human languages

Need of Processing Human Languages

- To develop different Human-computer Interactive systems through natural languages.
- Allows to communicate with machines through human languages
- To design different information processing systems
  - to work on digital documents/data represented in natural languages

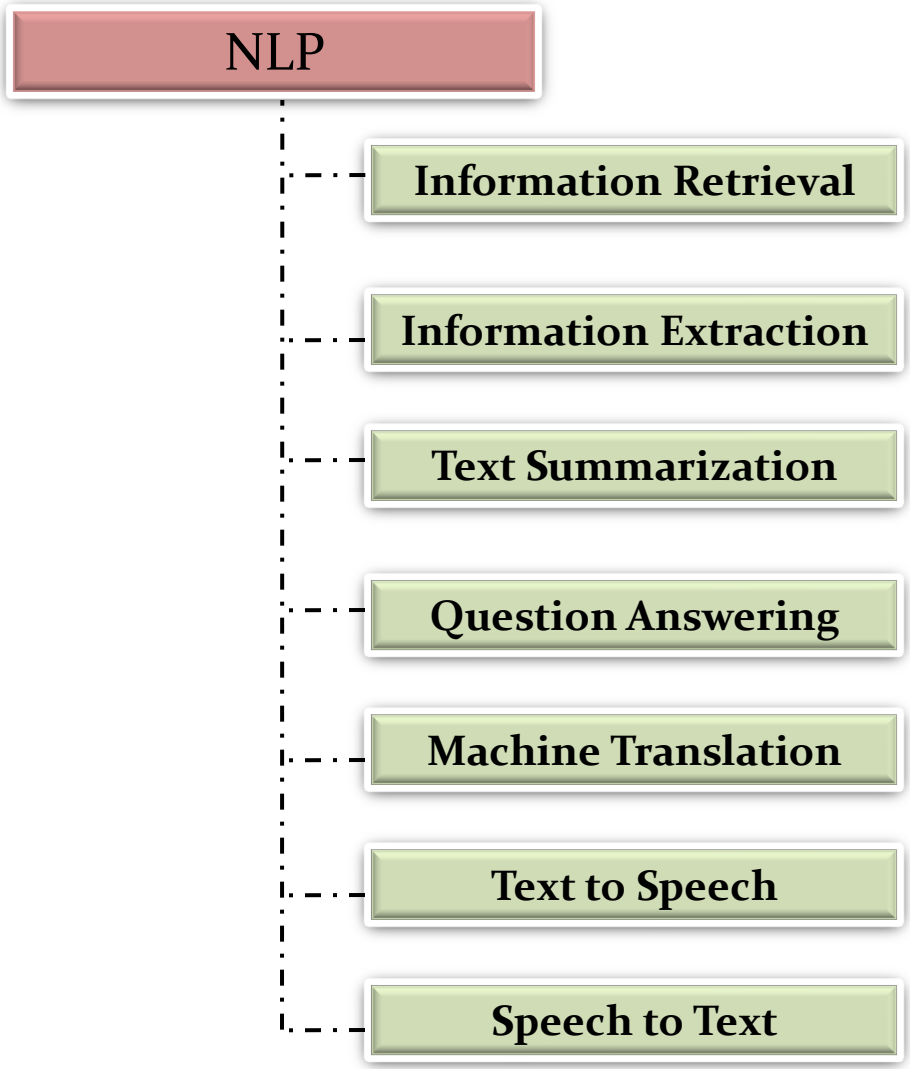
Applications of NLP

Example: The Search Engines

- The Web made it possible to access a large amount of information quickly

- Search engines provides the required information quickly at finger tip

Applications of NLP



Other Applications:

- Dictionary word suggestion
- Spelling error/ grammar correction
- Chat bots
- Sentiment analysis

…..

Cont…

Information Retrieval (IR):

- Focuses on retrieving the documents from a large document repository based on their relevance to a user's query.

- The IR technology is used in online search engines, library information retrieval, organizational data retrieval, etc

Information Extractions (IE):

- Refers to the automatic extraction of structured information from unstructured sources.

- The IE technology can be used to retrieve specific text such as place, organization, people, monetary values, etc from unstructured documents and present them in the form of a structured report.

Cont…

Text Summarization:

- The goal of automatic summarization is to take an information source, extract contents from it and provide the most important contents to the user in a concise form as a summary.

- It has its applications in obtaining document summaries, storylines of events, summarization of user-generated content, etc.

Question Answering:

- Concerns with building systems that can automatically answer questions posed by users in a natural language.

- It has its applications in designing experts question answering systems on different domains such as: medical, agricultural, legal, etc

Cont…

Machine Translation System:

- Translates from one human language text/speech into another language text/speech.

- It has its applications in designing language interpreters, website translation, document translation, speech translation applications, etc

Cont…

Text to Speech:

- A **text-to-speech (TTS)** system converts natural language text into speech

- It has its applications in designing speech synthesizers, screen readers, language learning apps, etc.

Speech to Text:

- Speech to text (STT) conversion is the process of converting spoken words into written texts.
- This is also called **speech recognition.**

- It has its applications in designing text dictation systems, command and control, audio document transcription, etc

Advanced Applications

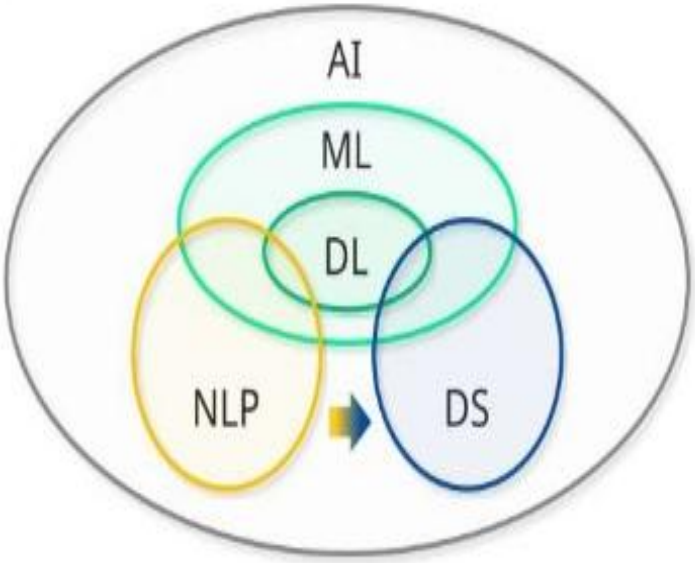
- **Virtual Assistants:**
  - Apple- ***Siri***
  - Amazon- ***Alexa***
  - Microsoft- ***Cortana***
  - Google: ***Google Assistant***

A Brief History of NLP Application Development

- **1950:** Mathematical model of computation (Turing machine)
- **1957:** Rule based syntactic structures (Chomsky's Syntactic Structures)
- **1969:** Conceptual dependency theory
- **1970:** Finite State Automata concepts
- **Up to 1980s:** NLP systems based on ***complex sets of hand-written rules***
- **1980 onwards:** NLP with **machine learning** algorithms
- **1990s:** NLP systems based on **Statistical models**
- **2000 onwards:** NLP systems applied with a**dvanced machine learning techniques**

Cont…

- Current trends in information processing technologies uses: AI, Data Science, NLP, Machine Learning for-
  - designing intelligent machines that-
    - Understands and generates human languages
    - Becomes more intelligent by learning from its experience



Goals of NLP

- **Scientific Goal**
  - Identify the computational machinery needed for an agent to exhibit various forms of linguistic behavior

- **Engineering Goal**
  - Design, implement, and test systems that process natural languages for practical applications

Issues and Processing Complexities

1. Ambiguity:

- Natural languages are highly ambiguous
- Words in a natural language may have a number of different meanings.

Examples:

  - Bank (River bank/ Financial Institution),
  - Bat(cricket bat/ species)

- For many NLP task, the proper sense of each ambiguous word in a sentence must be determine to interpret correct meaning.

Cont…

2. Language Variability:

- There are various ways to express meaning
- A large number of languages are available world wide
  - most of the languages use different character set, structure and grammar rules.
- It is difficult to design a language processing model that can capture all language variability.

Cont…

3. Difficult to incorporate human cognition over machine:

- It is difficult to capture all required knowledge human use to process natural languages
- People have no trouble understanding language as they have- common sense knowledge, reasoning capacity and experience
- Computers have- no common sense knowledge and no reasoning capacity

How can a machine understand these differences?

Examples:

- Decorate the cake **with** **the kids**.
- Throw out the cake **with** **the kids**.
- The man saw the girl **with** a telescope **in** a park.
- Stolen painting found **by tree.**
- **The old man finally kicked the bucket.**

Issues in Processing Indian Languages

1. Unlike English, Indic scripts have a nonlinear structure.

- Example:

Language	Script
English	English language
Hindi	हिन्दी भाषा
Odia	ଓଡ଼ିଆ ଭାଷା
Bengali:	বাংলা ভাষা

Cont…

2. English uses SVO (Subject-Verb-Object) format while Indian languages uses SOV (Subject-Object-Verb) format

Example:

English: pooja plays veena

(S) (V) (O)

Hindi: pooja veena Bajati hai

(S) (O) (V)

Cont…

3. Indian languages have a free word order

- i.e. words can be moved freely within a sentence without changing the meaning of the sentence

Example:

Usne khaanaa khaya

or

Usne khaaya Khanaa

Cont…

4. Have rich set of morphological variants

- Example: variants of the word ‘horse’

Hindi:

ghodda, ghodde, ghoddi, ghoddon……

Cont…

5. Extensive and productive use of complex predicates

Example:

हिन्दी

शब्द

सम्पूर्ण

Cont…

6. Ambiguity:

Example:

सोना

Language and Grammar

- Automatic processing of language requires the rules and exceptions of a language to be explained to the computer.
- Grammar defines the language
- It consists of a set of rules that allows parse and generate sentences in a language

Major Approaches of NLP

- **Rationalist approach(Symbolic approach)**
- **Empiricist approach**

Rationalist approach (Symbolic approach)

- Assumes existence of some language faculty in human brain
  - i.e. significant part of the knowledge in human mind is not derived by sense. It is fixed in advance by genetic inheritance.
- Example: children can’t learn complex things from limited sensory inputs.
- Machines can be made to function like human brain by giving some basic knowledge and reasoning mechanisms.
- Linguistic knowledge is explicitly encoded in rule or other forms of representation

Empiricist Approach

- No language faculty
- Believes in existence of some general organization principles:
  - Pattern reorganization
  - Generalization
  - Association
- Learning of detailed structures can takes place through the applications of these principles on sensory inputs available
- Focused on use of large amount of data and procedures involving statistical manipulation

Classification of Computational Models

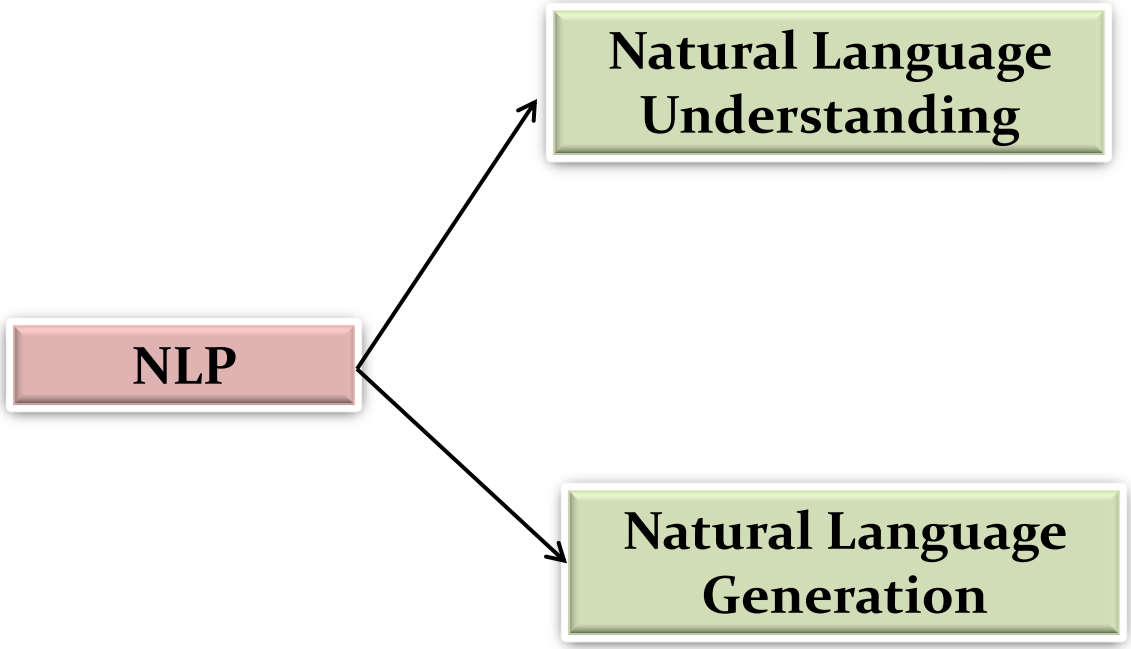
Knowledge driven:

- Rely on explicitly coded linguistic knowledge expressed as a set of hand crafted grammar rules.
- Constrained by the lack of sufficient coverage of domain knowledge (acquiring and encoding such knowledge is difficult)

Data driven:

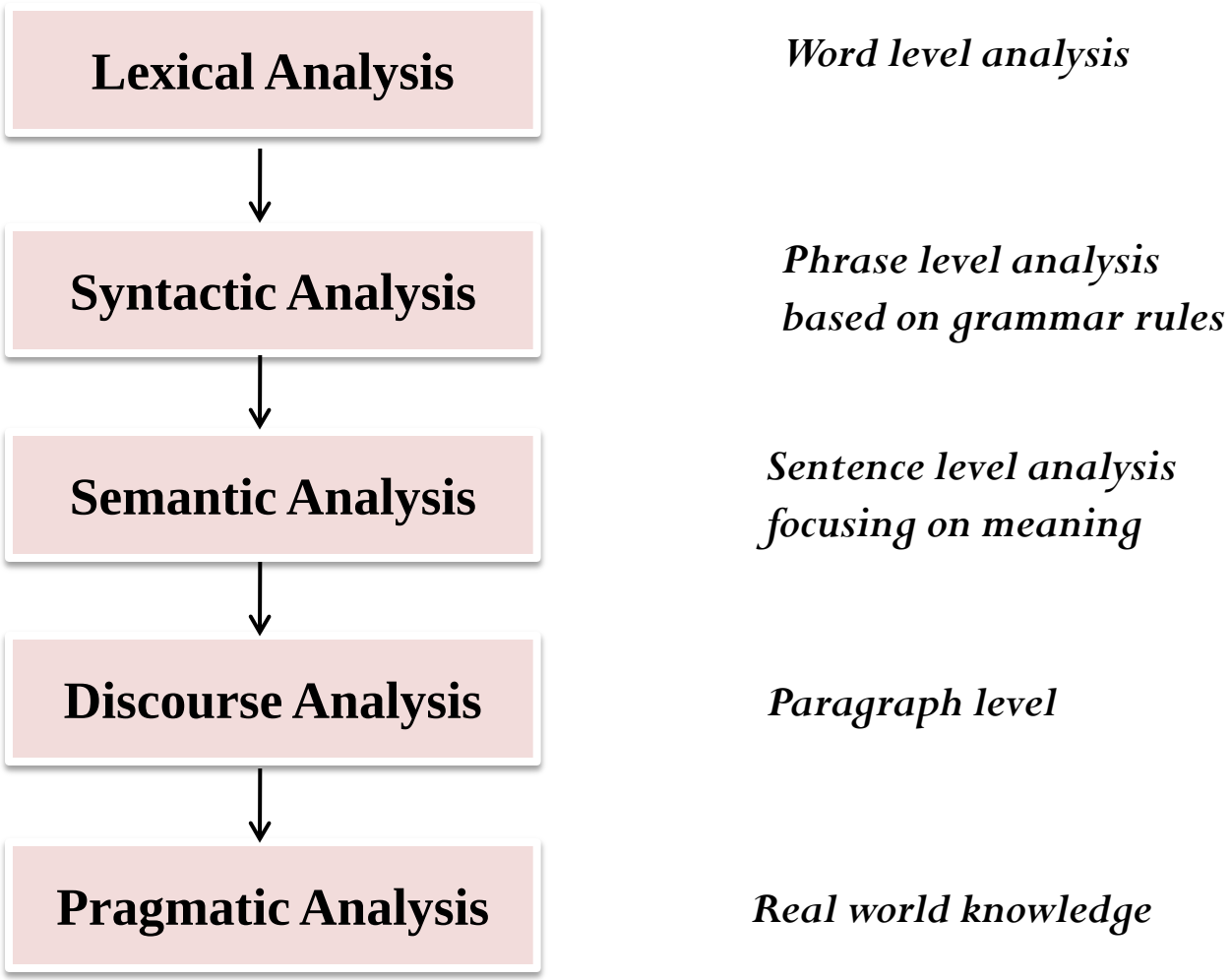
- Assumes the existence of large amount of data and techniques to learn syntactic pattern.
- Requires less human effort
- Performance depends on quantity of data
- Adaptive to noisy data

2 Main Components of NLP



- Mapping given inputs in a language into useful representations
- Analyzing different aspects of the language
- Involves producing meaningful phrases and sentences to convey some internal representation

Phases of Natural Language Processing



Lexical Analysis

- Analysis of **words** (most fundamental units of any natural language)

- Word level processing requires **Morphological Knowledge**
- i.e. knowledge about the structure and formation of words from basic units (morphemes)

- A **morpheme** is the smallest meaningful unit in a language

Example:

- cat (1 morpheme)
- cats (2 morphemes: ‘cat’, ‘s’)

- Rules for forming words are language specific

Syntactic Analysis

- Analysis of sequence of words as a unit(i.e. a sentence) and find its structure.

- It involves decomposes a sentence in to its constituents (or words) and identifies the relation between them

- Requires Syntactic Knowledge (how words are combined to form large units and constructs imposed on them)

Example:

“she is going to the market”	valid sentence
“she are going to the market”	invalid sentence

Semantic Analysis

- Concerned with creating meaningful representation of linguistic inputs.
- Example: “colorless green ideas sleep furiously”

  (Syntactically correct but semantically anomalous)
- Requires semantic knowledge
  - What words mean
  - How word meanings combined in sentences to form sentence meanings

Discourse Analysis

- Attempts to interpret structure and meaning of even larger units

- The meaning of any sentence depends upon the meaning of the sentence just before it.

- Requires **Discourse Knowledge**
  - Knowledge of how the meaning of a sentence is determined by processing sentences

Example:

She forgot her book.

To understand to whom- ‘she’, ‘her’ refers, processing of previous sentences are required

Pragmatic Analysis

- The meaning of a sentence can’t be always derived based on the meaning of its words. Multiple interpretations of a sentence can be possible.
- Syntactic structure and compositional semantics fail to explain these interpretations.
- Pragmatic analysis deals with the purposeful use of sentences in situations.
- It requires real world knowledge along with language knowledge

Example:

Do u know who am I?

(different context of use can be possible)