# NLP

*Prof. Deepak C Vegda*
*email.:deepakvegda.it@ddu.ac.in*

# History

- **(1940-1960) - Focused on Machine Translation (MT)**
  - The Natural Languages Processing started in the year 1940s.
  - **1948** - In the Year 1948, the first recognisable NLP application was introduced in Birkbeck College, London.
  - **1950s** - In the Year 1950s, there was a conflicting view between linguistics and computer science. Now, Chomsky developed his first book syntactic structures and claimed that language is generative in nature.
  - In 1957, Chomsky also introduced the idea of Generative Grammar, which is rule based descriptions of syntactic structures.

Conti..

- **(1960-1980) - Flavored with Artificial Intelligence (AI)**
  - **Augmented Transition Networks (ATN)**
    - Augmented Transition Networks is a finite state machine that is capable of recognizing regular languages.
  - **Case Grammar**
    - Case Grammar was developed by **Linguist Charles J. Fillmore** in the year 1968. Case Grammar uses languages such as English to express the relationship between nouns and verbs by using the preposition.
    - In Case Grammar, case roles can be defined to link certain kinds of verbs and objects.
    - **For example:** "Neha broke the mirror with the hammer". In this example case grammar identify Neha as an agent, mirror as a theme, and hammer as an instrument.

# Conti.

- **SHRDLU**
  - SHRDLU is a program written by **Terry Winograd** in 1968-70. It helps users to communicate with the computer and moving objects. It can handle instructions such as "pick up the green boll" and also answer the questions like "What is inside the black box." The main importance of SHRDLU is that it shows those syntax, semantics, and reasoning about the world that can be combined to produce a system that understands a natural language.
- **LUNAR**
  - LUNAR is the classic example of a Natural Language database interface system that is used ATNs and Woods' Procedural Semantics. It was capable of translating elaborate natural language expressions into database queries and handle 78% of requests without errors.

# Conti.

- **1980 – Current**
  - Till the year 1980, natural language processing systems were based on complex sets of hand-written rules. After 1980, NLP introduced machine learning algorithms for language processing.
  - Speech recognition
  - Machine translation
  - Machine text reading
  - Alexa
  - Eva

# Component of NLP

- **NLU**
  - It is used to map the given input into useful representation.
  - It is used to analyze different aspects of the language.
- **NLG**
  - Natural Language Generation (NLG) acts as a translator that converts the computerized data into natural language representation. It mainly involves Text planning, Sentence planning, and Text Realization.
  - **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.

# Applications of NLP

- Sentiment analysis
- Question answering
- Spam detection
- Machine translation
- Error correction
- Chatbot
- IE(survey analysis)
- Speech recognition
- Hiring and recruitment
- Search auto correct
- Social media monitoring
- Targeted advertising

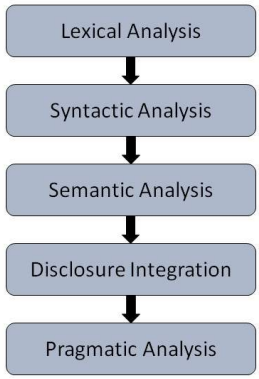
# What to be done in NLP

- **Sentence Segmentation**
  - It breaks the paragraph into separate sentences.
- **Word Tokenization**
  - Word Tokenizer is used to break the sentence into separate words or tokens.
- **Stemming**
  - Stemming is used to normalize words into its base form or root form.
- **Lemmatization**
  - It is used to group different inflected forms of the word, called Lemma. The main difference between Stemming and lemmatization is that it produces the root word, which has a meaning.
- **Identifying Stop Words**
  - **Stop words** might be filtered out before doing any statistical analysis.
- **Dependency Parsing**
  - Dependency Parsing is used to find that how all the words in the sentence are related to each other.
- **POS tags**
  - Dependency Parsing is used to find that how all the words in the sentence are related to each other.
  - POS stands for parts of speech, which includes Noun, verb, adverb, and Adjective
- **Named Entity Recognition (NER)**
  - Named Entity Recognition (NER) is the process of detecting the named entity such as person name, movie name, organization name, or location.
- **Chunking**
  - Chunking is used to collect the individual piece of information and grouping them into bigger pieces of sentences.

# NLP Terminology

- **Phonology** − It is study of organizing sound systematically.
- **Morphology** − It is a study of construction of words from primitive meaningful units.
- **Morpheme** − It is primitive unit of meaning in a language.
- **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.
- **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.

# Phases of NLP




# Lexical Analysis

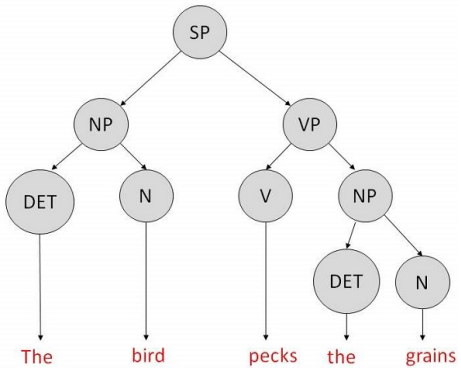
- It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of txt into paragraphs, sentences, and words.

Example:-

"***Natural language processing*** *(**NLP**) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data."*

## Syntactic Analysis (Parsing)

- Syntactic Analysis is used to check grammar, word arrangements, and shows the relationship among the words.
- Algorithm
  - Context-Free Grammar
    - The parse tree breaks down the sentence into structured parts so that the computer can easily understand and process it. In order for the parsing algorithm to construct this parse tree, a set of rewrite rules, which describe what tree structures are legal, need to be constructed.
  - Top-Down Parser
    - The parser starts with the S symbol and attempts to rewrite it into a sequence of *terminal symbols* that matches the classes of the words in the input sentence until it consists entirely of terminal symbols.




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 and objects in the task domain.
- It allows computers to understand and interpret sentences, paragraphs, or whole documents, by analyzing their grammatical structure, and identifying relationships between individual words in a particular context.

**Discourse Integration**

- Discourse Integration depends upon the sentences that proceeds it and also invokes the meaning of the sentences that follow it.
- 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.

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.
- It helps you to discover the intended effect by applying a set of rules that characterize cooperative dialogues.

Problems in NLP

**Ambiguity: -**Ambiguity and Uncertainty exist in the language

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

  For example,

  Kabir is looking for a match.

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

  For example,

  - The burglar threatened the student with the knife**.**

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

  For example,

  - Kiran went to Sunita. She said, "I am hungry."

## Lexical ambiguity

For eg: The word silver can be used as a noun, an adjective, or a verb.
She bagged two silver medals.
She made a silver speech.
Lexical ambiguity can be resolved by Lexical category disambiguation i.e, parts-of-speech tagging.

## Lexical Semantic ambiguity

The type of lexical ambiguity, which occurs when a single word is associated with multiple senses.
Eg: bank, pen, fast, bat, cricket etc
For eg:
The tank was full of water.
I saw a military tank.
Lexical Semantic ambiguity resolved using word sense disambiguation (WSD) techniques,where WSD aims at automatically assigning the meaning of the word in the context in a computational manner.

## Syntax Level ambiguity

Scope Ambiguity: Scope ambiguity involves operators and quantifiers.
E.g. Old men and women were taken to safe locations.
Attachment Ambiguity:Attachment ambiguity arises from uncertainty of attaching a phrase or clause to a part of a sentence.
E.g. The man saw the boy with the telescope

## Semantic ambiguity

This occurs when the meaning of the words themselves can be misinterpreted.
Eg: The car hit the pole while it was moving.

## Anaphoric Ambiguity

Anaphoric Ambiguity: Anaphoras are the entities that have been previously introduced into the discourse.
E.g. The horse ran up the hill. It was very steep. It soon got tired.

## Pragmatic Ambiguity

It refers to a situation where the context of a phrase gives it multiple interpretation.
E.g.
Tourist (checking out of the hotel): Waiter, go upstairs to myroom and see if my sandals are there; do not be late; I have tocatch the train in 15 minutes.
Waiter (running upstairs and coming back panting): Yes sir,they are there.
Clearly, the waiter is falling short of the expectation of thetourist, since he does not understand the pragmatics of the situation.

## NLP Libraries

- **Scikit-learn:** It provides a wide range of algorithms for building machine learning models in Python.
- **Natural language Toolkit (NLTK):** NLTK is a complete toolkit for all NLP techniques.
- **Pattern:** It is a web mining module for NLP and machine learning.
- **TextBlob:** It provides an easy interface to learn basic NLP tasks like sentiment analysis, noun phrase extraction, or pos-tagging.
- **Quepy:** Quepy is used to transform natural language questions into queries in a database query language.
- **SpaCy:** SpaCy is an open-source NLP library which is used for Data Extraction, Data Analysis, Sentiment Analysis, and Text Summarization.
- **Gensim:** Gensim works with large datasets and processes data streams

Expert System

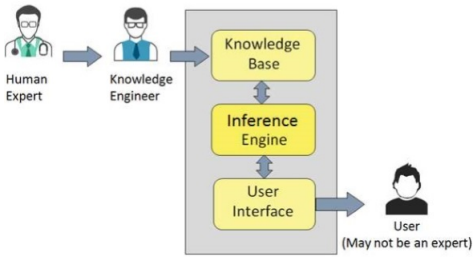
## What is it?

- The expert systems are the computer applications developed to solve complex problems in a particular domain, at the level of extra-ordinary human intelligence and expertise.

- Characteristics of Expert Systems
  - High performance
  - Understandable
  - Reliable
  - Highly responsive

## Components of Expert Systems

- Knowledge Base
- Inference Engine
- User Interface




## Knowledge Base

- It contains domain-specific and high-quality knowledge.
- Knowledge is required to exhibit intelligence. The success of any ES majorly depends upon the collection of highly accurate and precise knowledge.
- What is Knowledge?
  - The data is collection of facts. The information is organized as data and facts about the task domain. **Data, information,** and **past experience** combined together are termed as knowledge.
- Components of Knowledge Base
  - **Factual Knowledge** – It is the information widely accepted by the Knowledge Engineers and scholars in the task domain.
  - **Heuristic Knowledge** – It is about practice, accurate judgement, one's ability of evaluation, and guessing.

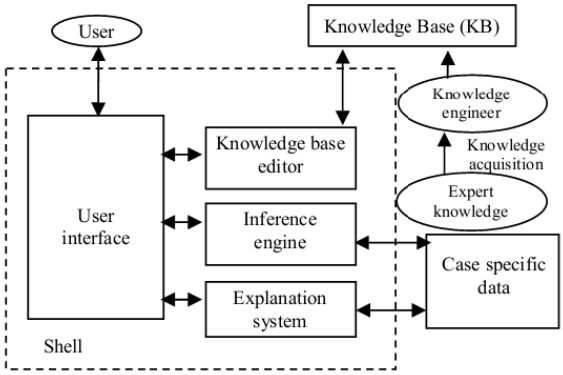
- Knowledge representation
  - It is the method used to organize and formalize the knowledge in the knowledge base. It is in the form of IF-THEN-ELSE rules.
- Knowledge Acquisition
  - The success of any expert system majorly depends on the quality, completeness, and accuracy of the information stored in the knowledge base.
  - The knowledge base is formed by readings from various experts, scholars, and the **Knowledge Engineers**. The knowledge engineer is a person with the qualities of empathy, quick learning, and case analyzing skills.

# Inference engine

- Use of efficient procedures and rules by the Inference Engine is essential in deducting a correct, flawless solution.
- In case of knowledge-based ES, the Inference Engine acquires and manipulates the knowledge from the knowledge base to arrive at a particular solution.
- In case of rule based ES, it −
  - Applies rules repeatedly to the facts, which are obtained from earlier rule application.
  - Adds new knowledge into the knowledge base if required.
  - Resolves rules conflict when multiple rules are applicable to a particular case.
- To recommend a solution, the Inference Engine uses the following strategies −
  - Forward Chaining
  - Backward Chaining

# User Interface

- User interface provides interaction between user of the ES and the ES itself. It is generally Natural Language Processing so as to be used by the user who is well-versed in the task domain. The user of the ES need not be necessarily an expert in Artificial Intelligence.
- Requirements of Efficient ES User Interface
  - It should help users to accomplish their goals in shortest possible way.
  - It should be designed to work for user's existing or desired work practices.
  - Its technology should be adaptable to user's requirements; not the other way round.
  - It should make efficient use of user input.




ure 6: Expert System Architecture illustrating the cur

# Limitation

- Limitations of the technology
- Difficult knowledge acquisition
- ES are difficult to maintain
- High development costs

| Application | Description |
|---|---|
| Design Domain | Camera lens design, automobile design. |
| Medical Domain | Diagnosis Systems to deduce cause of disease from observed data, conduction medical operations on humans. |
| Monitoring Systems | Comparing data continuously with observed system or with prescribed behavior such as leakage monitoring in long petroleum pipeline. |
| Process Control Systems | Controlling a physical process based on monitoring. |
| Knowledge Domain | Finding out faults in vehicles, computers. |
| Finance/ Commerce | Detection of possible fraud, suspicious transactions, stock market trading, Airline scheduling, cargo scheduling. |

## Some examples of ES

- **DENDRAL:** It was an artificial intelligence project that was made as a chemical analysis expert system. It was used in organic chemistry to detect unknown organic molecules with the help of their mass spectra and knowledge base of chemistry.
- **MYCIN:** It was one of the earliest backward chaining expert systems that was designed to find the bacteria causing infections like bacteraemia and meningitis. It was also used for the recommendation of antibiotics and the diagnosis of blood clotting diseases.
- **PXDES:** It is an expert system that is used to determine the type and level of lung cancer. To determine the disease, it takes a picture from the upper body, which looks like the shadow. This shadow identifies the type and degree of harm.
- **CaDeT:** The CaDet expert system is a diagnostic support system that can detect cancer at early stages.