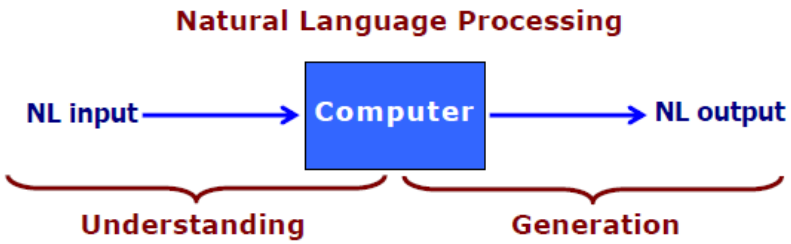
UNIT – 9 Natural Language Processing(NLP)

What is NLP?

- Natural Language Processing (NLP) is the capacity of a computer to "understand" natural language text at a level that allows meaningful interaction between the computer and a person working in a particular application domain.
- Processing of natural language is required when we want an intelligent system like robot to perform as per our instructions when we want to hear decision from a dialogue based expert system etc.
- Natural Language Processing (NLP) refers to AI method of communicating with intelligent systems using a natural language such as English. Natural languages are spoken by people.
- The field of 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
- A language is a set of system, a set of symbols and a set of rules (or grammars).
  - The symbols are combined to convey new information.
  - The rules govern the manipulation of symbols.
- NLP encompasses anything a computer needs to understand natural language (typed or spoken) and also generate the natural language.

Components of NLP

There are two components of NLP as given –



bit-papers.blogspot.com ©19171R

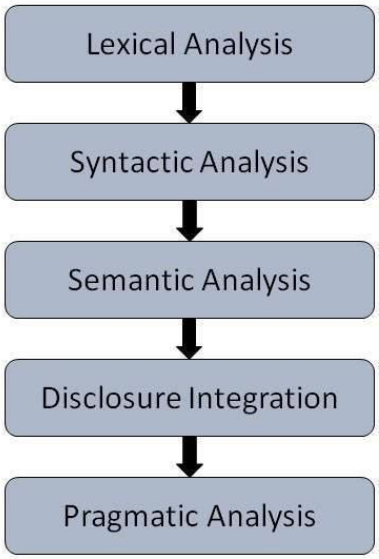
1. Natural Language Understanding (NLU)

- Taking some spoken/typed sentence and working out what it means.
- The NLU task is understanding and reasoning while the input is a natural language.
- Mapping the given input in the natural language into a useful representation.
- Understanding involves the following tasks –
  - Mapping the given input in natural language into useful representations.
  - Analyzing different aspects of the language.
- Different levels of analysis required:
  - Morphological analysis
  - Syntactic analysis
  - Semantic analysis
  - Discourse analysis

2. Natural Language Generation (NLG)

- Taking some formal representations of what you want to say and working out a way to express it in a natural language (e.g. English).
- NLG is a subfield of natural language processing NLP.
- It is the process of producing meaningful phrases and sentences in the form of natural language from some internal representation.
- Producing output in the natural language from some internal representation.
- Different level of synthesis required:
  - Deep planning(what to say)
  - Syntactic generation
- It involves –
  - **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.
- The NLU is harder than NLG.

Steps in NLP:



a. Lexical Analysis:

- The lexicon of a language is it's vocabulary that includes it's words and expressions.
- Lexical analysis involves dividing a text into paragraphs, words and the sentences.

b. Syntactic analysis:

- Syntax concerns the proper ordering of words and it's affect on meaning.
- This involves analysis of the words in a sentence to depict the grammatical structure of the sentence.
- The words are transformed into structure that shows how the words are related to each other.
- E.g. "the girl the go to the school". This would definitely be rejected by the English syntactic analyzer.

c. **Semantic Analysis** –

- Semantics concerns the (literal) meaning of words, phrases and sentences.
- This abstracts the dictionary meaning or the exact meaning from context.
- The structures which are created by the syntactic analyzer are assigned meaning.
- Example: "colorless blue idea". This would be rejected by the analyzer as colorless blue do not make any sense together.

d. **Discourse Integration** –

- Sense of the context.
- The meaning of any single sentence depends upon the sentences that precedes it and also invokes the meaning of the sentences that follow it.
- Example: the word "it" in the sentence "she wanted it" depends upon the prior discourse context.

e. **Pragmatic Analysis** –

- Pragmatic concerns the overall communicative and social context and it's effects on interpretation.
- It means abstracting and deriving the purposeful use of the language in the situations.
- The main focus is on what was said is reinterpreted on what it actually means.
- Example: "close the window?" should have been interpreted as request rather than an order.

Applications of NLP

- text processing - word processing, e-mail, spelling and grammar checkers
- interfaces to data bases - query languages, information retrieval, data mining, text summarization
- expert systems - explanations, disease diagnosis
- linguistics - machine translation, content analysis, writers' assistants, language
- Companies using AI chat bots that give you suggestions to locate the nearest grocery store, book a movie ticket, order food, etc.
- Sentiment analysis during a political campaign to take informed decisions by monitoring trending issues on social media
- Analyzing lengthy text reviews by users of products on an e-commerce website
- Call centers using NLP to analyze the general feedback of the callers

Linguistic Organization of NLP

- Grammar and lexicon - the rules for forming well-structured sentences, and the words that make up those sentences
- Morphology - the formation of words from stems, prefixes, and suffixes

E.g., eat + s = eats

- Syntax - the set of all well-formed sentences in a language and the rules for forming them
- Semantics - the meanings of all well-formed sentences in a language
- Pragmatics (world knowledge and context) - the influence of what we know about the real world upon the meaning of a sentence. E.g., "The balloon rose." allows an inference to be made that it must be filled with a lighter-than-air substance.
- The influence of discourse context (E.g., speaker-hearer roles in a conversation) on the meaning of a sentence
- Ambiguity
    - lexical - word meaning choices (E.g., *flies*)
    - syntactic - sentence structure choices (E.g., *She saw the man on the hill with the telescope*.)
    - semantic - sentence meaning choices (E.g., *They are flying planes.*)

Parse tree representation in natural language

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 describes what tree structures are legal, must be available. These rules say that a certain symbol may be expanded in the tree by a sequence of other symbols.

Grammars and parsing

Syntactic categories (common denotations) in NLP

- np - noun phrase
- vp - verb phrase
- s - sentence
- det - determiner (article)
- n - noun
- tv - transitive verb (takes an object)
- iv - intransitive verb
- prep - preposition
- pp - prepositional phrase
- adj - adjective

A *context-free grammar (CFG)* is a list of rules that define the set of all well-formed sentences in a language. Each rule has a left-hand side, which identifies a syntactic category, and a right-hand

side, which defines its alternative component parts, reading from left to right.
Context-Free Grammars are simply grammars consisting entirely of rules with a single symbol on the left-hand side of the rewrite rules. The obvious advantage of CFG is that it is simple to define. Many of the grammars used for NLP systems are CFG, as such they have been widely studied and understood and hence highly efficient parsing mechanisms have been developed to apply them to their input.

However, CFG also have some severe disadvantages. Consider the following rewrite rules, since *V* can be replaced by both "eat" or "eats", sentences like "The cat eat the rice" would be allowed. Therefore, additional sets of grammar would have to be implemented for both singular and plural sentences. Moreover, completely different sets of rules would also be needed for passive sentences, e.g. "The rice was eaten by the cat". This means that an extremely large set of rules would have to be created which makes it difficult to handle. Many different grammar formalisms like the unification grammar and the categorical grammar have been developed to capture the rules of syntax more concisely, but we won't go into them.

Q. Parse tree for “The cat eats the rice”
The rewrite rules of this example is as follows:
S -> NP VP
NP -> DET N | DET ADJ N
VP -> V NP
DET -> the
ADJ -> big | fat
N -> cat | cats | rice
V -> eat | eats | ate

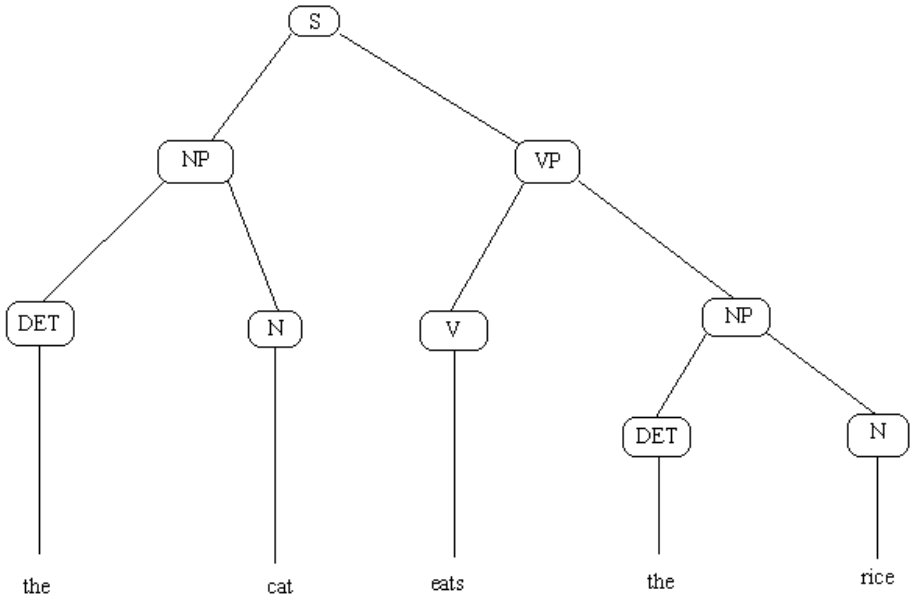


Figure 1. A grammar and a parse tree for "the giraffe dreams".

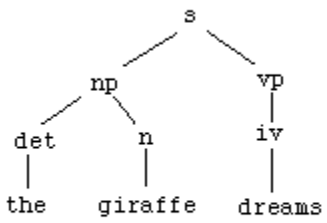
```
s   ⟶ np vp

np  ⟶ det n
vp  ⟶ tv np
    ⟶ iv

det ⟶ the
    ⟶ a
    ⟶ an

n   ⟶ giraffe
    ⟶ apple

iv  ⟶ dreams
tv  ⟶ eats
    ⟶ dreams
```



bit-papers.blogspot.com **©19171R**

NLP vs PLP (Programming Language Processing)

There are some parallels, and some fundamental distinctions, between the goals and methods of programming language processing (design and compiler strategies) and natural language processing. Here is a brief summary:

	NLP	PLP
domain of discourse	broad: what can be expressed	narrow: what can be computed
lexicon	large/complex	small/simple
grammatical constructs	many and varied - declarative - interrogative - fragments etc.	few - declarative - imperative
meanings of an expression	many	one
tools and techniques	morphological analysis syntactic analysis semantic analysis integration of world knowledge	lexical analysis context-free parsing code generation/compiling interpreting