SKILRY

NATURAL LANGUAGE PROCESSING

NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) refers to AI method of communicating with an intelligent systems using a natural language such as English.

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 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**

COMPONENTS OF NLP

There are two components of NLP as given:

NATURAL LANGUAGE UNDERSTANDING (NLU)

Understanding involves the following tasks:

- Mapping the given input in natural language into useful representations.
- Analyzing different aspects of the language.

NATURAL LANGUAGE GENERATION (NLG)

It is the process of producing meaningful phrases and sentences in the form of natural language from some internal representation.

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.

The NLU is harder than NLG.

DIFFICULTIES IN NLU

- **NL has an extremely rich form and structure.**
- **It is very ambiguous. There can be different levels of ambiguity:**

1. **Lexical ambiguity: It is at very primitive level such as word-level.**
2. **For example, treating the word “board” as noun or verb?**
3. **Syntax Level ambiguity: A sentence can be parsed in different ways.**
4. **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?**
5. **Referential ambiguity: Referring to something using pronouns. For example, Rima went to Gauri. She said, “I am tired.” - Exactly who is tired?**
6. **One input can mean different meanings.**
7. **Many inputs can mean the same thing.**

NLP TERMINOLOGY

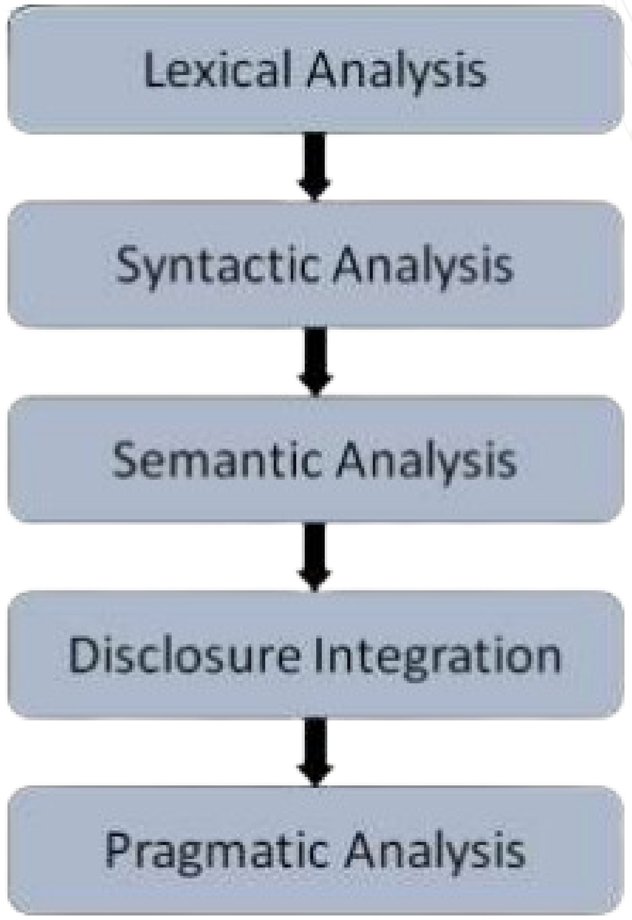
- **Phonology: It is study of organizing sound systematically.**
- **Morphology: It is a study of construction of words from primitive meaningful units.**
- **Morpheme: It is primitive unit of meaning in a language.**
- **Syntax: It refers to arranging words to make a sentence. It also involves determining 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.**

STEPSIN NLP

There are general five steps:

1.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.

2.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. The sentence such as "The school goes to boy" is rejected by English syntactic analyzer.



Lexical Analysis

Syntactic Analysis

Semantic Analysis

Disclosure Integration

Pragmatic Analysis

3. **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. The semantic analyzer disregards sentences such as “hot ice cream”.

4. **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 the immediately succeeding sentence.

5. **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.

IMPLEMENTATION ASPECTS OF SYNTACTIC ANALYSIS

There are a number of algorithms researchers have developed for syntactic analysis, but we consider only the following simple methods:

- **Context-Free Grammar**
- **Top-Down Parser**

Let us see them in detail:

CONTEXT-FREE GRAMMAR

It is the grammar that consists rules with a single symbol on the left-hand side of the rewrite rules. Let us create grammar to parse a sentence –
“The bird pecks the grains”

Articles (DET): a | an | the.

Nouns: bird | birds | grain | grains

Noun Phrase (NP): Article + Noun | Article + Adjective + Noun
= DET N | DET ADJ N

Verbs: pecks | pecking | pecked
Verb Phrase (VP): NP V | V NP
Adjectives (ADJ): beautiful | small | chirping

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.

These rules say that a certain symbol may be expanded in the tree by a sequence of other symbols. According to first order logic rule, ff there are two strings Noun Phrase (NP) and Verb Phrase (VP), then the string combined by NP followed by VP is a sentence. The rewrite rules for the sentence are as follows:

S -> NP VP

NP -> DET N | DET ADJ N

VP -> V NP

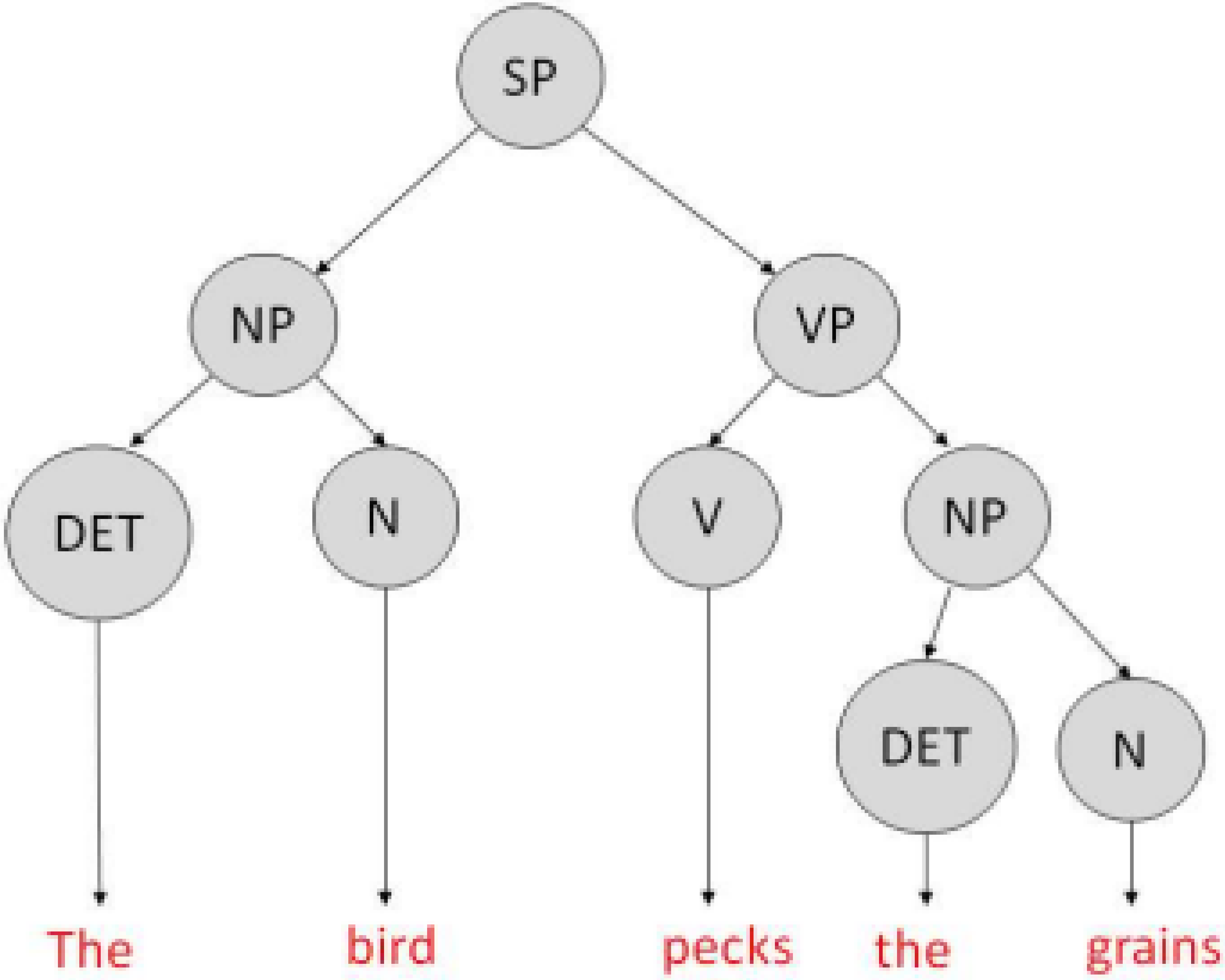
Lexocon:

DET -> a | the

ADJ -> beautiful | perching

N -> bird | birds | grain | grains

V -> peck | pecks | pecking



SKILRY

Now consider the above rewrite rules. Since V can be replaced by both, "peck" or "pecks", sentences such as "The bird peck the grains" can be wrongly permitted. i. e. the subject-verb agreement error is approved as correct.

Merit: The simplest style of grammar, therefore widely used one.]

Demerits:

- They are not highly precise. For example, “The grains peck the bird”, is syntactically correct according to the parser, but even if it makes no sense, the parser takes it as a correct sentence.

- To bring out high precision, multiple sets of grammar need to be prepared. It may require a completely different sets of rules for parsing singular and plural variations, passive sentences, etc., which can lead to creation of huge set of rules that are unmanageable.

TOP-DOWN PARSER

Here, 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.
These are then checked with the input sentence to see if it matched. If not, the process is started over again with a different set of rules. This is repeated until a specific rule is found which describes the structure of the sentence.

Merit: It is simple to implement.

Demerits:

- It is inefficient, as the search process has to be repeated if an error occurs.
- Slow speed of working.