Introduction to Levels of Natural Language Processing

Assist. Prof. Manvinder Kaur
Dept. of Computer Science, B.Z.S.F.S. Khalsa Girls College, Morinda (Punjab)

Abstract- Natural language processing is a branch of computer science and artificial intelligence (AI) which is study of communication between computers and human beings. Natural language processing is an area of research and application which deals with how computers can be used to understand and manipulates natural language text or speech. Human have total five senses of hearing, touch, smell, taste and sightthrough which they perceive and communicate. There are number of natural languages present. They contain countless sentences. A person produced new sentences easily no matter how many sentences a person had heard or seen. There an ambiguity is present in Natural language. Many words have several meanings such as well, orange, fly and sentences have different meaning in different contexts. To gain deep knowledge of natural language processing first need to understand the basic steps or levels of Natural language processing. Natural language processing is mathematical and computational processing of various aspects of language. Natural language processing has a big role in computer science because many aspects of the field deal with linguistic features of computation.

Keywords- Understanding; Phonology; Morphology; Syntactic; Semantic; Discourse; Pragmatic

I. INTRODUCTION

The idea of computers having the ability to grasp standard languages and hold conversations with mortals has been a staple of fantasy since the primary half of the 21th century and was envision in classic paper by Alan Turing (1950) as an indicator of machine intelligence. Since the beginning of the 21st century this vision has been commencing to look acceptable: computer science techniques related with the scientific study of language have emerged from universities and analysis laboratories to tell a range of {commercial of business} and commercial applications. Several websites currently provide automatic translation; mobile phones will seem to grasp spoken queries and commands; search engines like Google use basic linguistic techniques for automatically finishing or 'correcting' your queries and for locating relevant results that are closely matched to your search terms. Automated systems arestill not performed full machine understanding of natural language. Machine-controlled translations still need of human translator's intervention. Developing programs to grasp natural language is very important in AI as a result of variety of communication with systems is important for user acceptance [1].Moreover, one in all the foremost crucial tests for intelligent behaviour is that the ability to speak effectively.AI program should be able to communicate with their human counterparts in an exceedingly natural approach, and language is one in all the foremost necessary medium for this purpose. A program understands a language if it behaves by taking an accurate or acceptable action in response to the input. The action taken needn't every time to give external response. It should merely be the creation of some internal information structures as would occur in learning some new fact but the structures created should be meaningful and correctly interact with the world model representation.

II. LEVELS OF NATURAL LANGUAGE PROCESSING

The most instructive methodology for presenting what truly happens at intervals a linguistic communication process system is by means of 'levels of language' approach that helps to come up with the informatics text by realizing Content designing, Sentence designing and Surface Realization phases. This is often conjointly cited because the synchronic model of language is distinguished from the sooner consecutive model that hypothesizes that the amount of human language process follows each other during a strictly consecutive manner. Cognitive psychology analysis suggests that language process is way a lot of dynamic, because the levels will act during a sort of orders. Reflection reveals that we regularly use information that have tendency to gained from what's usually thought of as high level of processing to help during a lower level's study.

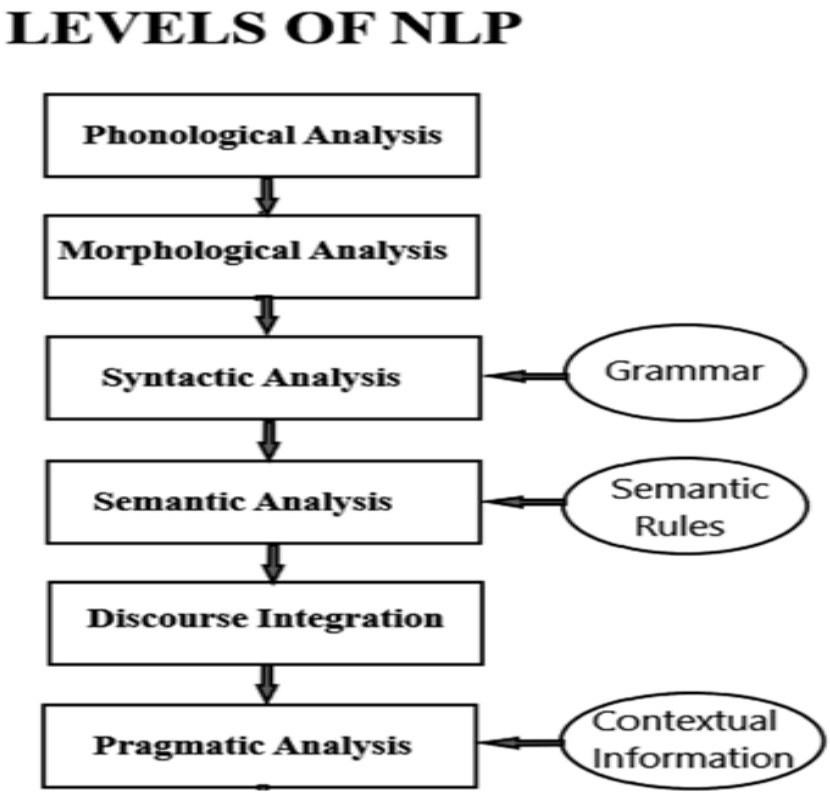

Fig.1: Levels of NLP

A.PHONOLOGICAL ANALYSIS

Phonology refers to the systematic arrangement of sound. The term phonology is an Ancient Greek term and theterm phono means the voice or sound, and the suffix- logy refers to word or speech. In 1993 Nikolai Trubetzkoyexplicit that descriptive linguistics is "the study of sound referring to the system of language". Whereas Lass in 1998 wrote that descriptive linguistics refers generally with the sounds of speaking language and explained as, "Phonology describes the function, behaviour and organization of sounds as linguistic things. It includes linguistics use of sound to encrypt the Human language [2]. This level deals with the interpretation of speech sounds at intervals and across words. There are 3 sorts of rules employed in synchronic linguistics analysis: 1) phonetic rules – for sounds at intervals betweenwords; 2) sound rules – for variations of pronunciation that is produced when words are spoken and; 3) delivery rules – for fluctuation in stress given to words and intonation across a sentence. In associate informatics system that accepts spoken input, the sound waves are analysed and encoded into a digitized signal for interpretation by varied rules or by comparison to the actual language model being used.

a. Transcription of sound

Phonetic transcription is the visual illustration of speech sounds. It's sometimes written within the International Phonetic Alphabet (IPA), in which every English sound has its own image. For example, *there* is transcribed as /ðeəʳ/.

b. Articulatory phonetics

The field of articulatedphonetics could be a subfield of Phonetic. In learning articulation, phoneticians make a case for however humans turn out speech sounds via the interaction of various physiological structures.

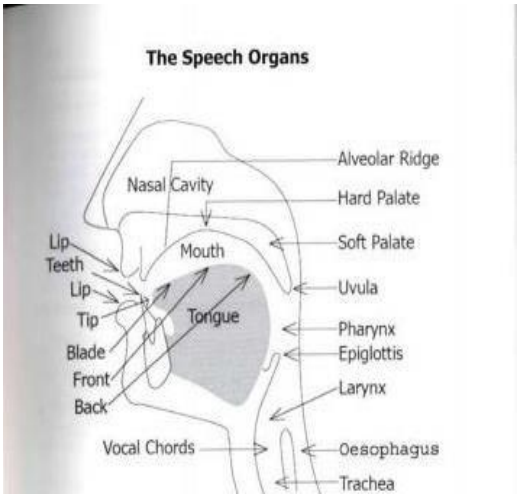

Fig.2: The speech production chain from intention to acoustics.

c. Acoustic phonetics

In acoustic phonetics the study of the physical properties of speech is performed to analyse the sound wave signalsshown in fig. 3 that occur within speech through varying frequencies, amplitudes and duration.

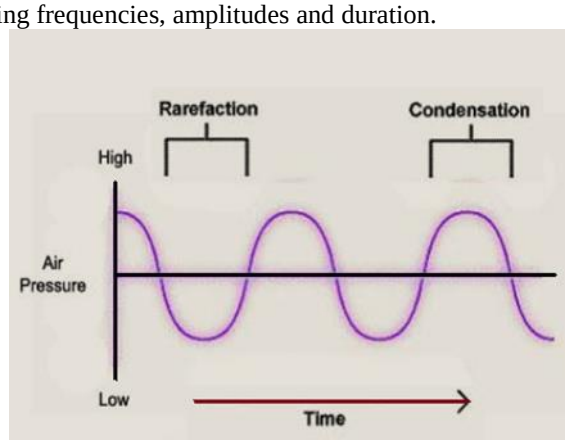

Fig.3: Sound Wave Signal

d. Auditory Phonetics:

It focuses on the perception of sounds. This tellshow sounds are heard and interpreted by nervous system and brain.

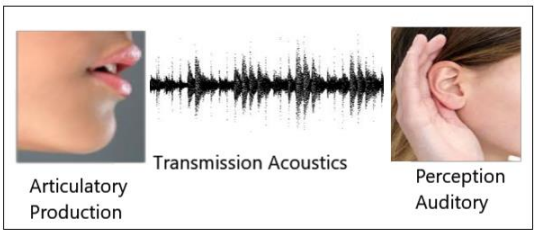

Fig.4: Perception of Sounds

B.MORPHOLOGICAL ANALYSIS

Morphology considerations word formation. It's a study of the patterns of formation of words by the mixing of sounds into smallest possible distinctive units of meaning referred to as morphemes. In morphologicalinformation considerations the focus is on how morphemes create words.

Morphology is initiated by morphemes. A simple example of morphologyis the word preregistration. According to morphological analysis this word divided into three separate morphemes: the prefix pre, the basisregistra, and into the suffix -tion. The same method of interpretation of morpheme applies across all the words. So that it can be understand that how humans can break any unknown word into morphemes. For instance, when the suffix –ed is added to a verb then it shows that action of the verb has been occurred in the past. The words that can't be divided and have meaning by themselves are referred to as Lexical morpheme (e.g.:

pen, car,).The words like -ed, -ing, -est, -ly, -ful etc. arewhen combined with the lexical morphemes are referred to as Grammatical morphemes (e.g. Worked, Going, Highest, Steeply, Faithful). Morphological analysis comprised of lexical analysis.

a. Lexical analysis

It deals with signification of a word. In Lexical analysis, whether humans or language processing systemsboth interpret the meaning of individual words. One amongst the basic operations that may be applied to a text is tokenising: ending a stream of characters into words, punctuation marks, numbers and alternative distinct things. Thus, for instance the character string "Mr Watson, Mr David", said Steve, introducing us. is tokenised as within the following example, wherever every token is bound within single quotation marks:

'"' 'Mr' 'Watson' ',' 'Mr' 'David' '"' ',' 'said' 'Steve' ',' 'introducing' 'us' '.'

At this level, words haven't been classified into grammatical classes and we have little indication of syntactical structure. Still, a good quantity of data is also obtained from comparatively shallow analysis of tokenised text.

When the method of tokenisation has performed, the found tokens got to be labelledwith their part-of-speech. In this stage analysing text is associates each token with part of speech (POS). During this process, words that may perform more than one part-of-speech are tagged by most likely part-of speech tag and supported the context in which they occur.

A number of various POS classifications are developed inside linguistics. The subsequent is the list of classes that are usually encountered in linguistics.

Noun: car, book, house, pencil, table, language
Proper noun: David, John, Italy, Berlin
Verb: loves, hates, studies, sleeps, thinks, is, has
Adjective: beautiful, sleepy, happy
Adverb: slowly, quickly, now, here, there
Pronoun: I, you, he, she, we, us, it, they
Preposition: in, on, at, by, around, with, without
Conjunction: and, but, or, unless
Determiner: the, a, an, some, many, few, 100

Nouns 'generally indicate individuals, places, things or concepts' whereas verbs 'describe events or action. One will simply realize or construct examples wherever constant thought is expressed by a noun or a verb, or by associate adjective or associate adverb. And on the opposite hand, there are several words that may take completely different parts of speech counting on what they are doing during a sentence:

1. She is well.
2. The wellis full of water.
3. Give me a stamp.
4. Stamp this paper.

Additionally, some forms of verbs don't correspond to any specific action however serve a strictly grammatical function: these embody the auxiliary verbs like did, shall etc. Thus, in outline, we will usually assigna part of speech to a word counting on its function in context instead of however it relates to real things or events within the world. In lexical level, one meaning is assigned to the semantic representation of the word.

C. SYNTACTIC ANALYSIS

This level emphasises to inspect the words in a sentence so as to reveal the grammatical structure of the sentence. At syntactic level we study how words combine to form phrases, phrases combine to form clauses and how clauses put together to make sentences. Syntactic analysis deals with the correctsentence formation. It study the structural role played by each word in the sentence and what phrases are subparts of what other phrases. It involves grammatical analysis of words in sentences.The sentence such as "The road runs on bus" is rejected by English syntactic analyser.At this level of processing the output isthe representation of asentence that discloses the structural dependency relationships between the words. Not all NLP applications require a full parse of sentences, therefore the abide challenges in parsing of prepositional phrase attachment and conjunction audit no longer impede that plea for which phrasal and clausal dependencies are adequate [3]. In most languages syntax conveys the meaningof sentences because order and dependency contribute to connotation. For example, the two sentences: 'The police chased the thief.' and 'the thief chased the police.' differ only in terms of syntax but these convey quite different meanings.

Almost all the systems that are literally have 2 main components:

A declarative illustration, known as a descriptive linguistic, of the syntactical facts regarding the language.
A procedure, known as parser that compares the descriptive linguistic against input sentences to provide parsed structures.

Syntactic Analysis is well-developed space of natural language processing, deals with the syntax of language. In syntactical Analysis, a grammar is used to find out the legal sentences. The descriptive linguistics is being applied by a parsing rule to provide a structure illustration, or break down tree. The syntactic analysis mainly used
i. Context-Free Grammar
ii. Top-Down Parser

Context-free grammars will generate context-free languages. This is done by taking a collection of variables that are outlined recursively, in terms of one another, by a collection of production rules. Context-free descriptive linguisticsare named intrinsically as a result of any of the assembly rules within the grammar.

Noam Chomsky, an American linguist in 1957, used the subsequent notation referred to as productions, to outline the syntax of English [4]. The terms used here, sentence, phrase etc. and the subsequent rules describe a little set of English sentences. The articles a, and the are classified as adjectives for simplicity.
<sentence> --><noun phrase><verb phrase>
<noun phrase> --><adjective><noun phrase> | <adjective><singular noun>
<verb phrase> --><singular verb><adverb>
<adjective> --> a | the |little
<singular noun> --> boy
<singular verb> --> ran
<adverb> --> quickly
Here, the arrow, -->, might be read as "is defined as" and the vertical bar, "|", as "or". .

Input sentences are parsed by the parser.A parser performs the procedural interpretation of the grammar. During parsing the parser follows the productions rules of a grammar. The result of parsing is the construction of the tree structures and these structures are according to the grammar.

For Example: Creation of grammar to parse a sentence − "The cow munched the grass"
Articles (DET) − a | an | the
Nouns − cow | cows | grass | grasses
Noun Phrase (NP) − Article + Noun | Article + Adjective + Noun
= DET N | DET ADJ N
Verbs − munch| munching | munched
Verb Phrase (VP) − NP V | V NP

As a result of parsing numbers of parse trees are constructed. To choose the optimal tree structure a set of rewrite rule is constructed. The rewrite rules for the sentence are as follows −
S → NP VP
NP → DET N | DET ADJ N
VP → V NP
Lexicon:
DET → a | the
N → cow | cows | grass | grasses
V → munch | munching | munched
The parse tree can be created as shown −

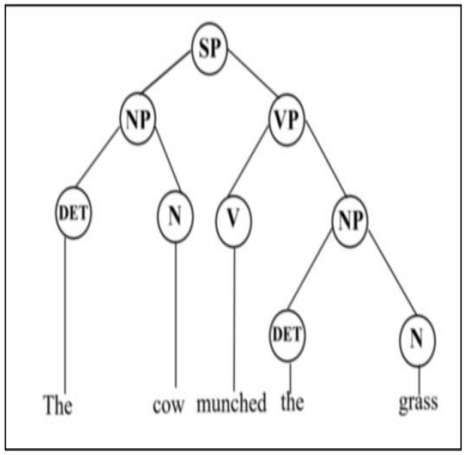

Fig.5: parsing of sentence

The parser starts with symbol S and rewritesthe rules and matched these with the input sentence to describe the structure of sentences. If these are not matched then the process is started again and is repeated until a correct rule is not found.

D. SEMANTIC ANALYSIS

Semantic analysis is concerns with the meaning of the language. It assigns the meaning to the structures created by the syntactic analyser.In other words it made the mapping syntactic structures and objects in the task domain. The structures for which the mapping is not possible may be rejected. For example the: sentence such as "Colourless yellow thoughts sleep fast" is rejected by sematic anomalies. Semantic analysis consists of the lexical processing.

Lexical processing: Its main job is to look the individual word in a dictionary and extract the meaning. Sometimes the word has several meaning so this level deals with the semantic ambiguity of words. For example, the word, 'file' as a noun can have different meanings like it can either be a bundle of papers, or a fingernails shaper [3]. The semantic level scrutinizesthe dictionary meaning of words, but also for the meaning they derive from the domain of the sentence.In semantic knowledgebase most words have more than one meaning but that we can choose the appropriate one by looking at the sentence carefully[5].Thus the process of determining the correct meaning of the individual word is called lexical processing or word sensing.

E. DISCOURSE INTEGRATION

The word discourse meanshow propositions fit together in a conversation. Discourse analysis deals with the multi-sentence which involves the interpretation of text. This analysis includes morphemes, n-grams, tenses, verbal aspects, page layouts, and so on. The syntax and semantics work with sentence-length units, the discourse level of NLP works with units of text longer than a sentence. That is, it does not interpret multisentence texts as just concatenated sentences and each sentence is interpreted separately. Rather, discourse focuses on the properties of the text as a whole that convey meaning by making connections between component sentences [3]. Several types of discourse processing can occur at this level, two of the most common are

(i)Anaphora resolution

(ii)Discourse/ text structure recognition.

Anaphora resolution is the replacing of words such as pronouns, which are semantically vacant, with the appropriate entity to which they refer. Discourse/text structure recognition determines the functions of sentences in the text, which, in turn, adds to the meaningful representation of the text. For example, company employs can be deconstructed into discourse components such as: Manager,Financer, Clerk, Software developer, Software tester, and Software analyst. Grosz, Joshi and Weinstein (1995) provide a broad-based discussion of the nature of discourse, clarifying what is involved beyond the sentence level, and how the syntax of the sentences support the structure of the discourse. In their analysis, discourse contains linguistic structure (syntax, semantics), attention structure (focus of attention), and intentional structure (plan of participants) and is structured into coherent segments. During discourse processing one important task for the hearer is to identify the referents of noun phrases [6].Discourse language concerns inter-sentential links that is how the immediately preceding sentences affect the interpretation of the next sentence.

Consider a discourse **David went to the mall on Saturday. He met harry**." Here, **He** refers to David.

F. PRAGMATIC ANALYSIS

Pragmatic analysis finds the actual meaning of the natural language's input. For this re-interpretation of input is performed. This level requires real world knowledge. This analysis is concerned with analysis, how language is used in different situations and

how context is applied on the contents of the text for understanding how extra meaning is extract from texts without actually encoded in them. This requires much world knowledge to understand the real situations, intentions, ideas, plans and goals. For this NLP applications utilize knowledge bases and inference rules. For example, the following two sentences require resolution of the anaphoric term 'they', and this resolution requires pragmatic or world knowledge.

The leaders of national party refusedto organize a rally because they feared violence. The leaders of national party refused to organize a rallybecause they advocated revolution.

The word "they" has different meaning in these two sentences. In order to figure out the difference, world knowledge in knowledge bases and inference rules areutilize.

III. CONCLUSION

Understanding and generating human language is difficult problem. It requires knowledge of grammar and language, of syntax and semantics, of what people know and believe their goals, the contextual setting, pragmatics, and world knowledge. This section has provided examples of some of the problems associated with analysing human languages and has described the most important stages in Natural Language Processing.

IV. REFRENCES

[1]. Dan W Patterson, 1990. Introduction to Artificial Intelligence and Expert Systems.

[2]. Nation, K., Snowling, M. J., & Clarke, P. (2007). Dissecting the relationship between language skills and learning to read: Semantic and phonological contributions to new vocabulary learning in children with poor reading comprehension.

[3]. Liddy, E. D. (2001). Natural language processing.

[4]. Chomsky, Noam, 1965, Aspects of the Theory of Syntax, Cambridge, Massachusetts: MIT Press.

[5]. Feldman, S. (1999). NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. *Online-weston then wilton-*, *23*, 62-73.

[6]. Grosz, B. J., A. K. Joshi and S. Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*21.2.203-225