ISSN:0975-9646
Mallamma V. Reddy et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 8 (1) , 2017, 27-30

Constructs of Natural Language Processing

Mallamma V. Reddy[#1], Hanumanthappa M.[*2]

Department of Computer Science, Rani Channamma University, Vidyasangam, Belgaum-591156, India
[*]Department of computer science, Bangalore University, Jnanabharathi Campus, Bangalore-560056, India

***Abstract*— Natural language processing (NLP) is a process in which it is concerned with human languages where computers to perform useful and understanding tasks. It process sentences in a natural languages such as English or any Indian officially recognized languages: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu, and Urdu, rather than in specialized artificial computer languages such as BASIC,JAVA. In the era of World Wide Web, a rich source of information is growing at an enormous rate and the diversity of languages makes it even more necessary to have sophisticated Systems for Natural Language Processing, some of the requirements are Classify text into categories, Index and search large texts, Automatic translation, Speech understanding, Information extraction, Automatic summarization, Question answering, Knowledge acquisition, Text generations / dialogues. This paper describes the constructs of Natural Languages process and its methods for processing human languages to easily communicate by the system.**

***Keywords*—Part of speech, Phonetics, Semantics, Tokenization, Natural Language processing (NLP).**

I. INTRODUCTION

Natural language is typically used for communication either in the form of written, signed or spoken. The text in Internet is available in great number of languages other than English. One of the major issues raised in any application of automatic processing of these digital documents is that of multilingualism, since we want to perform linguistic processing. The following natural language terminologies are useful in natural language processing they are:

A. Phonology

It studies the sound systems of language. The Table 1. shows typical examples of the occurrence of consonant phonemes in words.

TABLE I
EXAMPLE OF PHONEMES

/f/	fan	/v/	van
/m/	man	/c/	can

B. Morphology

It studies the structure or forms of words, and how words are constructed from more basic meaning units called morphemes. A morpheme is the primitive unit of meaning in a language such as root words, affixes, parts of speech, and intonation/stress. It has three branches such as inflectional, derivational and compounding.

C. Syntax

It studies of the rules that govern the structure of sentences. The symbols used in syntax are shown in Table 2.

TABLE III
SYMBOLS USED IN SYNTAX

Symbol	Meaning	Example
S	Sentence	Apple is sweet
NP	Noun Phrase	An apple
VP	Verb Phrase	saw an apple tree
PP	Prepositional Phrase	with a telescope
Det	Determiner	the
N	Noun	Cat

D. Semantics

It studies the relation between signifiers, like words, phrases, signs, and symbols, and what they stand for, their denotation. An example is shown in Fig. 1.

Fig. 1 Denotation

The denotation of this example is a red rose with a green stem. The connotation is that it is a symbol of passion and love, this is what the rose represents.

E. Pragmatics

It studies how sentences are used in different situations and how it affects the interpretation of the sentence.

The sentence "You have a green light" is ambiguous. Without knowing the context, the identity of the speaker, and his or her intent, it is difficult to infer the meaning with confidence. For example, it could mean that:

- You have green ambient lighting.
- You have a green light while driving your car.
- You can go ahead with the project.
- Your body has a green glow.
- You possess a light bulb that is tinted green.

F. Discourse

It concerns how the immediately preceding sentences affect the interpretation of the next sentence. For example, the expected answer to the question "Do you know what the time is?" is something like "9 AM", and not just "Yes", though the latter is lexically, syntactically and semantically accurate.

G. Stylistics

It studies the linguistic factors (rhetoric, diction, stress) that place a discourse in context.

H. Semiotics

It studies the signs and sign processes (semiosis), indication, designation, likeness, analogy, metaphor, symbolism, signification, and communication.

I. World Knowledge

It studies general knowledge about the world, which each language user must know about the other's beliefs and goals.

II. CONSTRUCTS OF NATURAL LANGUAGE PROCESS

The constructs [1] of natural language processing is shown in Fig. 2. The detailed description is given below:

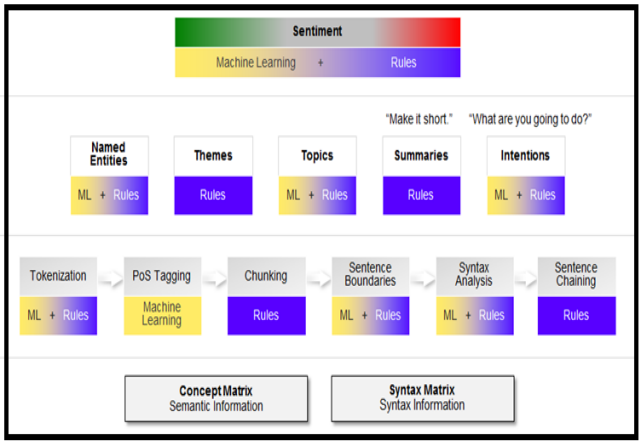

Fig. 2 Constructs of Natural Language Process.

A. Sentiment analysis

It is also called as opinion mining [2], extracting non trivial subjective information from database. It is widely useful in reviews and social media for a wide variety of applications, such as marketing and customer service. The main aim of sentiment analysis is to identify attitude towards a particular topic, product, etc. is positive, negative, or neutral. There are many sentiment analysis tools available few of them are LIWC used for identifying psychological concerns, POMS used for mood analysis, OPNOIN FINDER used for positive vs. negative mood from text content, POMS(Google-POMS) used for identifying six dimensions such as calm, alert, sure, vital, kind and happy.

B. Machine learning

Machine Learning (ML) is a set of statistical techniques for identifying some aspect of text such as parts of speech, named entities, sentiment, etc. There are three types of machine learning techniques they are:

1) Supervised machine learning: that can be expressed as a model that is then applied to other text is called supervised machine learning some of the methods applied for supervised machine learning are Support Vector Machines, Bayesian Networks, Maximum Entropy, Conditional Random Field, Neural Networks/Deep Learning. We use these methods for a number of natural language processing tasks such as Tokenization, Part of Speech tagging, Named Entity Recognition, Sentiment, classification.

2) Unsupervised machine learning: it could be a set of algorithms that work across large sets of data to extract meaning is called unsupervised machine learning some of the methods applied for unsupervised machine learning are Clustering, Latent Semantic Indexing, and Matrix Factorization.

3) Lexalytics: It is the hybrid form of supervised and unsupervised machine learning. It uses unsupervised learning to produce some "basic understanding" of how language works. In order to interpret the meaning of a set of words, three things are required they are: semantics, syntax, and context. To process the methods used are Concept Matrix, Syntax Matrix.

C. Automatic Text Summarization

It is a process of generating a summary of a given document and retaining most important points in the original document. It is the part of machine learning and data mining. Automatic text summarization [3] can be done in different ways:

1) *Extraction-based*: In this system extracts objects from the entire collection, without modifying the objects themselves, where the goal is to select individual words or phrases to "tag" a document.

2) *Abstraction-based*: In this system extracts abstraction involves paraphrasing sections of the source document for example, key clauses, sentences or paragraphs.

3) Maximum entropy-based: In this system extracts sentence for multi-document summarization from the news domain.

4) *Aided summarization:* In this system Machine learning techniques from closely related fields such as information retrieval or text mining have been successfully adapted to help automatic summarization [4],[5],[6]. Apart from this there are systems that aid users with the task of summarization they are Machine Aided Human Summarization, for example by highlighting candidate passages to be included in the summary, and Human Aided Machine Summarization depend on post-processing by a human.

D. Tokenization

Extracting words [7] from text may appear to be simple task. The top-down method breaks text on whitespace

characters such as a space, Tab, or a punctuation character. Non whitespace characters are concatenated to form a word or token. The bottom-up method builds tokens one character at a time from a text stream until a non token character is encountered. The simplest definition of a token is any consecutive string of alphanumeric characters. Between tokens, we find one or more non token characters.

E. Part of speech tagging

Any language comprises of number of blocks. The syntax of a language contains a phrase structure (PS) component and a transformational component [8]. In phrase structure the assumed largest unit of grammar, the sentence [S] is progressively expanded by the application of rules into 'strings' of smaller units because in Transformational grammar (TG) sentence is the basic unit of the syntactic system. Instead of beginning with actual sentences, directions for generating structural descriptions of sentences are set forth in phrase structure rules. Each rule provides a symbol representing a constituent of a sentence to the left of an arrow and a symbol or series of symbols to the right. The Table 3. depicts the symbols used in phrase structure rules which are used as part of speech tagger. The main aim of Part-of-speech tagging is the task of assigning a syntactic category to each word in a text, thereby resolving some ambiguities. E.g., the tagger decides whether the word ships are used as a plural noun or a third person singular present tense verb.

F. Syntax analysis

Linear sequences of words are transformed into structures that show how the words relate to each other. While analyzing the language Word order in languages differs. Some classification can be done by naming the typical order of subject (S), verb (V) and object (O) in a sentence and one can talk, for instance, of SVO or VSO languages. e.g. The sentence “Rama threw the ball” depicts the structure of English language (i.e. Subject-verb-object) and its equivalent Kannada sentence “rAma chendannu Esedanu”(“ರಾಮ ಚೆಂಡನ್ನು ಎಸೆದನು”.), depicts the structure of Kannada language (i.e. Subject-Object-Verb). The syntactic parser [9] recognizes the sentence by solving lexical and attachment ambiguities and assigns a syntactic structure to it in the form of a parse tree as shown in Figure 3.

TABLE IIIII RULES USED AS POS

POS Tagger	Meaning	POS Tagger	Meaning
S	Sentence	Be	The verb Be
NP	Noun phrase	Pred	Predicate(noun,adjective,adverb)
VP	Verb phrase	Vt	Transitive Verb
N	Noun	Vi	Intransitive verb
VB	Verb	Vl	Linking Verb
T,art or D	Determiner	Comp	Complement(noun or adjective)
Pron	Pronoun	Adj	Adjective
Aux	Auxiliary	Adv	Adverb
M	Model Auxiliary	PP	Prepositional phrase

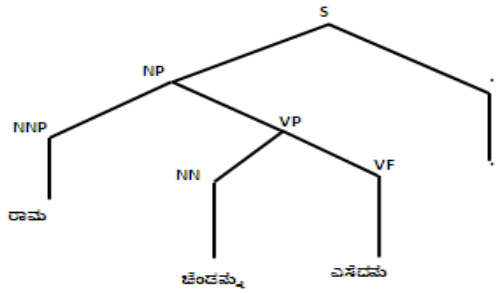

Fig. 3: Syntactic tree structure

There are also additional differences in word orders, for instance, where modifiers for nouns are located, or where the same words are used as a question or a statement. Some models will be used to represent linguistic knowledge they are:

- *State Machines:* FSAs, FSTs, HMMs, ATNs, RTNs.
- *Formal Rule Systems:* Context Free Grammars, Unification Grammars, Probabilistic CFGs.
- *Logic-based Formalisms:* first order predicate logic, some higher order logic.
- *Models of Uncertainty:* Bayesian probability theory.

The NLP laboratory is developing the synt syntactic analyser [10]. According to tests performed on large corpora, the performance of synt reaches the recall of 92 % and precision of 84 %. For educational purposes we have a simple syntactic analyzer Zuzana [11], which is capable of visualizing several types of derivation trees.

G. Named Entities

A named entity has a real-world existence, such as persons, locations, organizations, products, etc., that can be denoted with a proper name. Named Entities recognition is also known as entity identification, entity chunking and entity extraction. Machine transliteration [12], which is the conversion of a character or word from one language to another without losing its phonological characteristics. It is an orthographical and phonetic converting process. Therefore, both grapheme and phoneme information should be considered. Accurate transliteration of named entities plays an important role in the performance of machine translation and cross-language information retrieval processes. Whereas Machine translation, which is the conversion of a character or word from one language to another preserving its meaning.

III. CONCLUSIONS

In this paper we have explained the various constructs of natural language processing which will help for the researchers to understand the natural language processing for low-, mid-, and high-level text functions. Low-level text functions (Tokenization, PoS Tagging, Chunking, Sentence Boundaries, Syntax Analysis) are the initial processes any text input is run through. These functions are the first step in turning unstructured text into structured data; thus these low-level functions form the base layer of information from which our mid-level functions (Entities: Rules to determine “Who, What, Where”, Themes: Rules “What’s the buzz?”, Topics: Rules “About this?”, Summaries: Rules “Make it short”, Intentions) draw on. Those mid-level text functions

involve extracting the real content of a document of text, determining who is speaking, what they are saying, and what they are talking about. The high-level function of sentiment analysis is the final step, determining and applying sentiment on the entity, theme, and document levels.

REFERENCES

[1] http://lexalytics.com/lexablog/wp-content/uploads/2012/02
[2] https://en.wikipedia.org/wiki/Sentiment_analysis
[3] https://people.dsv.su.se/~hercules/textsammanfattningeng.html
[4] Lehmam, Abderrafih (2010). Essential summarizer: innovative automatic text summarization software in twenty languages - ACM Digital Library., Published in Proceeding RIAO'10 Adaptivity, Personalization and Fusion of Heterogeneous Information, CID Paris, France
[5] Mani, Inderjeet (2001). Automatic Summarization. ISBN 1-58811-060-5.
[6] Huff, Jason (2010). AutoSummarize., Conceptual artwork using automatic summarization software in Microsoft Word 2008.
[7] Mallamma V Reddy, Dr. Hanumanthappa. M, "Natural Language Identification and Translation Tool for Natural Language Processing" is published in the "International Journal of Science and Applied Information Technology" Volume 1, No.4, September –October 2012, ISSN No. 2278-3083. Available Online at http://warse.org/pdfs/ijsait03142012.pdf.
[8] Noam Chomsky's book on "syntactic structures" in 1957
[9] Roxana Girju, "Introduction to Syntactic Parsing", 2004.
[10] http://nlp.fi.muni.cz/projekty/zuzuna/
[11] http://nlp.fi.muni.cz/projekty/wwwsynt/
[12] Mallamma.V. Reddy, Dr. Hanumanthappa. M, "English to Kannada/Telugu Name Transliteration in CLIR: A Statistical Approach" Presented In the "3rd International Conference on Cognition & Recognition (ICCR)-2011" held on 9th and 10th December,2011 at Mysore University, Mysore.