Natural Language Processing: An approach to Parsing and Semantic Analysis

Shabina Dhuria

Department of Computer Science, DAV College, Sector-10, Chandigarh

Abstract: Natural language processing is the study of mathematical and computational modelling of various aspects of language and the improvement of a wide range of systems. Natural language is any language that arises as an innate facility for language possessed by the human intellect; it may be spoken, signed or written. Machine learning algorithms are used in conjunction with language models to recognize text in natural language processing systems, which may also employ speech models and hardware/software specialized to process and recognize speech or even signed (gesture-based) language. Natural language processing provides a potential means of gaining access to the information inherent in the large amount of text made available through the internet. This paper presents the basic concepts of Natural Language Processing, levels of linguistic analysis, Parsing techniques, Semantic analysis and applications of natural language in real-world.

Index Terms: Natural Language Processing, Information Extraction, Machine Translation, Linguistic Analysis, Semantic Analysis, Parsing.

I. INTRODUCTION

Natural language processing (NLP) is a field of computer science, artificial intelligence (also called machine learning), and linguistics concerned with the interactions between computers and human (natural) languages. It is the process of a computer extracting meaningful information from natural language input and/or producing natural language output. It is analysis of human language based on semantics and various parsing techniques as mentioned below in figure 1. The goal of NLP is to identify the computational machinery needed for an agent to exhibit various forms of linguistic behavior (i.e. Scientific Goal). It also design, implement, and test systems that process natural languages for practical applications (i.e. Engineering Goal).

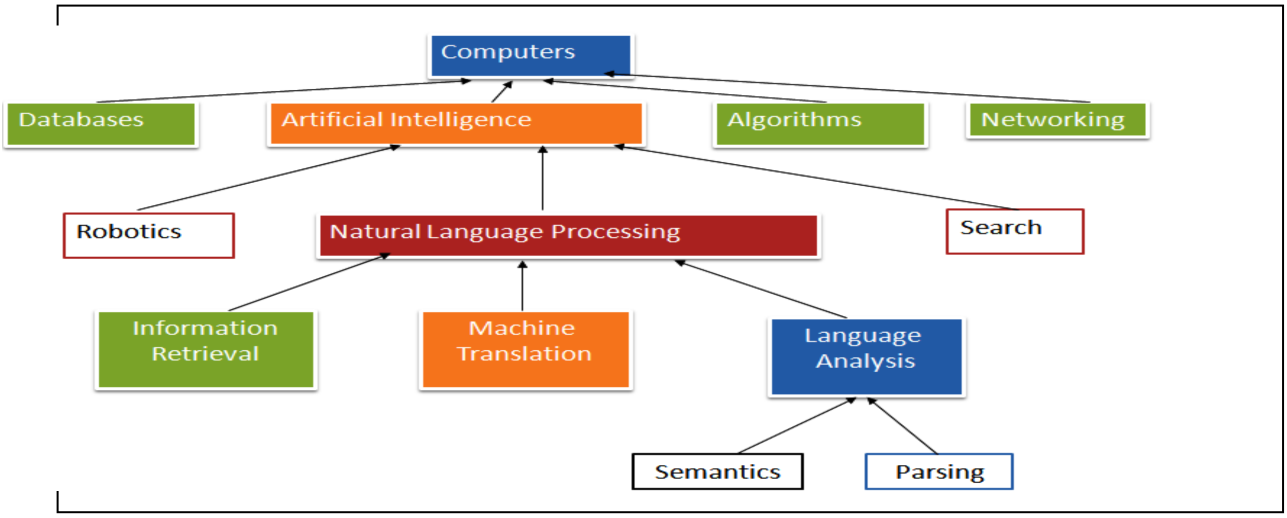

Fig. 1: NLP in Computer Science

Natural Language Processing (NLP) is a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty [1]. The main task of it is to construct programs in order to process words and texts in natural language. The main aspects of NLP are:

- **Information Retrieval (IR):** It is concerned with storing, searching and retrieving of information from text documents as shown in figure 2. It is a field within computer science closer to databases and relies on some of the NLP methods [2]. Ex. stemming.

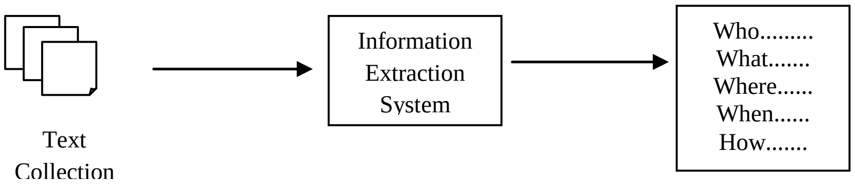

Fig 2: Information Extraction

- *Machine Translation***:** It is related to automatic translation from one human language to another as mentioned in figure 3. Ex. Deluxe Universal Translator [3].

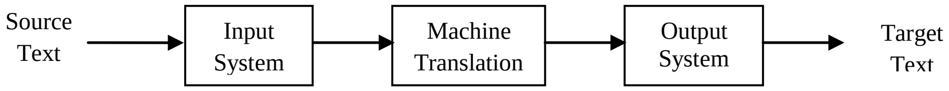

Fig 3: Machine Translation System

- *Language Analysis***:** It is concerned with parsing of an input sentence to construct syntactic tree and further sentiment analysis is done to find meaningful words in a sentence as shown in figure 4.

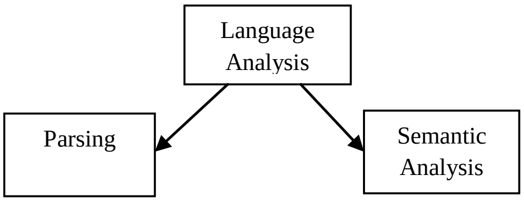

Fig 4: Language Analysis Process

II. Levels of Linguistic Analysis

Linguistic is the science of language. It study includes Sounds (phonology), Word formation (morphology), Sentence structure (syntax), Meaning (semantics) and Understanding (pragmatics) as shown in figure 5 [4]:

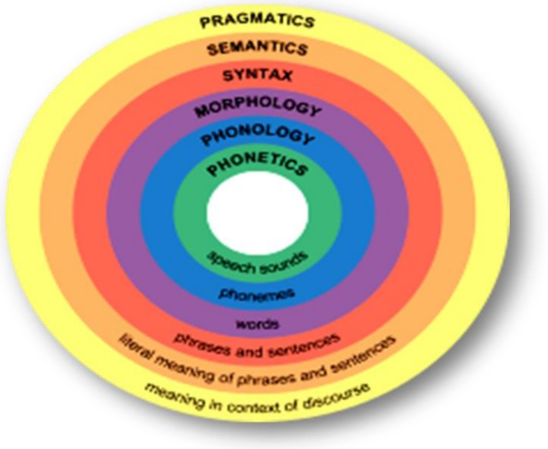

Fig 5: Levels of Linguistic Analysis

- *Phonological Analysis:* relates sounds to the words we recognize. Phoneme is smallest unit of sound, and the phones are aggregated into word sounds.
- *Morphological Analysis***:** Morphology is a sub discipline of linguistics that studies word structure. It is concerned with derivation of new words from existing ones. In NLP, words are known as lexicon items and a set of words form a lexicon. Lexicon is a module that tells what words there are and what properties they have [5].

- *Syntactic Analysis***:** is analysis of words in a sentence to know the grammatical structure of a sentence and these words are transformed into structures that show how the words relate to each others.
- *Semantic Analysis***:** It is concerned with the meaning of the language. The first step in semantic processing system is to look up the individual words in a dictionary (or lexicon) and extract their meanings [6].
- *Pragmatic Analysis***:** to reinterpret what was said to what was actually meant. It concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

III. PARSING TECHNIQUES

Parsing and generation are sub-divisions of NLP dealing respectively with taking language apart and putting it together. To parse a sentence, it is necessary to find a way in which that sentence could have been generated from the start symbol. Parsing uses knowledge about word and the word meanings (lexicon) based upon the reasoning processes. It exploits the pre-defined legal structures (grammar) i.e. set of rules as shown in figure 6.

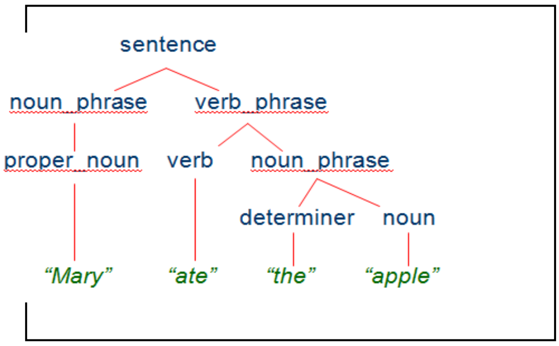

Fig 6: Parsing Tree [7]

Parsing has two main components [8]:

- *Grammar*: a declarative representation describing the syntactic structure of sentences in the language.
- *Parser:* an algorithm that analyzes the input and outputs its structural representation (its parse) consistent with the grammar specification. To construct the parsing tree the below mentioned techniques are used:

- *Top-down Parsing:* It begins with start symbol and apply the grammar rules forward until the symbols at the terminals of the tree correspond to the components of the sentence being parsed as shown in figure 7. (i.e. Top-down parsing starts with the S symbol and tries to rewrite it into the sentence).

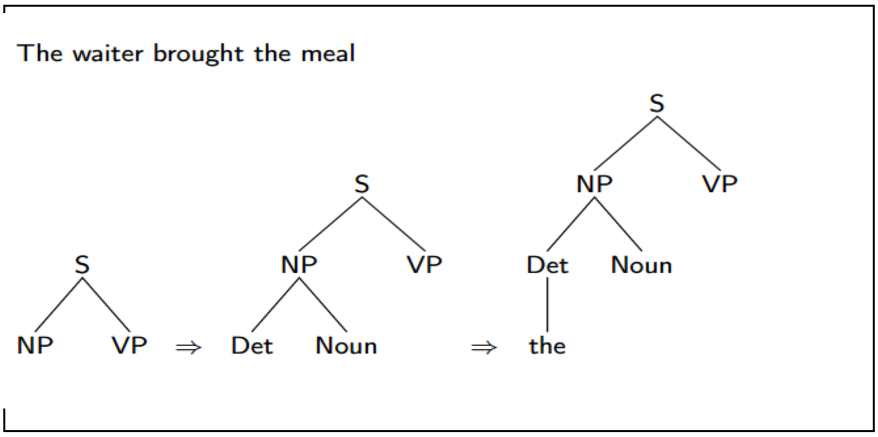

Fig 7: Top-down Parsing [12]

- *Bottom-up parsing:* It begin with the sentence to be parsed and apply the grammar rules backward until a single tree whose terminals are the words of the sentence and whose top node is the start symbol has been produced as shown in figure 8. (i.e. Bottom-up parsing starts with the words and tries to find symbols that generate them).

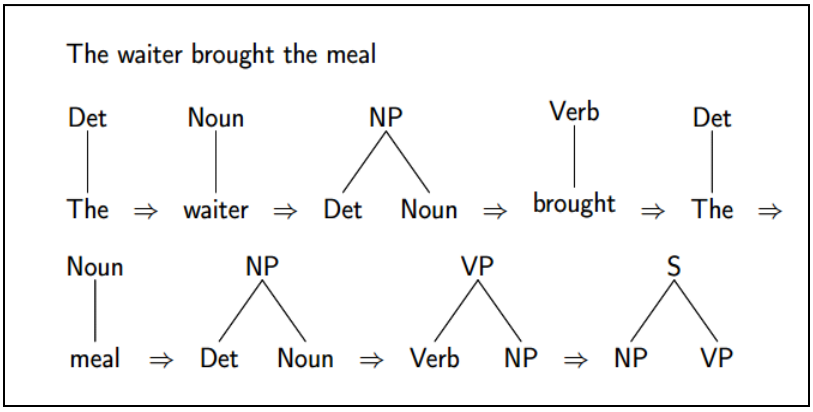

Fig 8: Bottom-up Parsing [12]

These two approaches are similar to the choice between forward and backward reasoning in other problem-solving tasks.

IV. Approaches to Semantic Analysis

Semantic analysis (SA) finds out the meaning of linguistic input and constructs meaning representations. It provides common-sense knowledge about the world [9]. To extract data and construct models of the world SA follows various approaches as mentioned below:

- *Predicate logic*

  Ex. "a cafe that serves South Indian food near TUT" corresponds to the meaning representation

  ∃x cafe(x) Λ Serves(x, SouthIndianFood) Λ Near (Location Of (x), Location Of (TUT))

- *Statistical approach*

  It is related to statistical machine translation that translates by matching source fragments against a database of real examples, and identifying the corresponding translation fragments, and then recombining these to give the target text.

- *Information retrieval*
- *Domain knowledge driven analysis*

Semantics and its understanding as a study of meaning covers most complex tasks like: finding synonyms, word sense disambiguation, constructing question-answering systems, translating from one NL to another, populating base of knowledge [11].

V. Real-world NLP Applications

Most NLP applications such as information extraction, machine translation, sentiment analysis and question answering, require both syntactic and semantic analysis at various levels.

- Text Mining
- Spelling correction
- Document Summary Systems
- Information Visualization. Ex. Cartia's Themescape [10]
- Grammar Checking Systems. Ex. MS Word Grammar Checker
- Information Retrieval / NL interface. Ex. Buzzcity
- Speech Recognition Systems / Speech Synthesizers. Ex. Siri (Apple, SRI, Nuance).
- Machine (Assisted) Translation - translating sentences from one language to another. Ex. Google Translator.
- Question answering (Question answering aims to give a specific response to the formulated query i.e. Who is the first prime minister of India?). Ex. IBM Watson.

VI. Conclusion

In this paper the concept of Natural language processing (NLP) as a study of mathematical and computational modelling for various aspects of language and the development of a wide range of systems is discussed as an interdisciplinary field which involves the parsing, semantic, linguistic analysis and machine translation process. The real world application areas of NLP, Parsing techniques and approaches to semantic analysis are also discussed which exploits the complexity at each point in using the knowledge paradigms.

VII. Future Directions

Some of the key research problems in NLP are: techniques for improving the efficiency of the parsing systems by exploiting lexical dependencies, techniques for exploiting certain regularities in specific domains, e.g., particular sentence patterns tend to appear more often in specific domains, systematic techniques for computing partial parses, systematic techniques for integration of parsing with semantic interpretation and translation, investigation of parallel processing techniques for parsing and experiments with large grammars.

References

[1] Radev, R., and D., "Natural Language Processing FAQ", Columbia University, Dept. of Computer Science, NYC, 2001.

[2] Carolina Ruiz, "Natural Language Processing", Computer Science WPI.

[3] Manning, C. D. and Schutze, H., "Foundations of Statistical Natural Language Processing", MIT Press, Cambridge, MA, pp. 680, 1999.

[4] K.R. Chowdhary, "Natural Language Processing", M.B.M. Engineering College, Jodhpur, India April 29, 2012.

[5] Xiaoyong Liu, "Natural Language Processing", School of Information Studies at Syracuse University.

[6] Russell and Norvig, "Artificial Intelligence: A Modern Approach", Prentice Hall, 2003.

[7] T. Dean, J. Allen, and Y. Aloimonos, "Artificial Intelligence: Theory and Practice", The Benjamin / Cummings Publishing Company, 1995.

[8] Dan W. Patterson, "Introduction to Artificial Intelligence and Expert System", PHI, 2001, Chapter 12.

[9] Eugene Charniak and Drew Mcdermott, "Introduction to Artificial Intelligence", Pearson, 1998, Chapter 4.

[10] H.Taneja, S.Dhuria and K.Sukhija "Natural Language Processing: A Backbone for Computational Linguistics", DHE Sponsored National Conference on Computational Sanskrit – Issues and Challenges, Dec. 2013, pp. 187-190.

[11] Poroshin V.A, "Semantic analysis of Natural Language", International Conference on Computational Linguistics, pp 16-23, Jan. 2014.

[12] H. Loftsson, H. Hogni, "Natural Language Processing Parsing techniques", School of Computer Science, Reykjavik University, 2008.