Processing of Natural Languages Semantically - A Detailed Survey

Rashmi S*, Dr. M Hanumanthappa**,
Regina L Suganthi***

*(Research Scholar, Department of Computer Science and Applications, Bangalore University, Bangalore-56.
** (Associate Professor, Department of Computer Science and Applications, Bangalore University, Bangalore-56
*** (Research Scholar, Department of Computer Science and Applications, Bangalore University, Bangalore-56

ABSTRACT
Semantic analysis is a method of representing knowledge. The goal is to reduce the syntactic structures and provide the meaning. English is a language that is geographically accepted all over the world. This is a boon as English is considered to be a language spread worldwide but it could also be a curse because even today people face problem in understanding and accessing this huge corpus of information which is available in English. There is a tremendous amount of information in the internet but all of which are in English. This causes problem for those users who do not understand English. When a word/sentence is translated into another language, it is very important to retain the original meaning even after the translation. This underlying idea has given exposure to Cross language Information Retrieval (CLIR) which is a process of information retrieval in a language different from the language of users query. CLIR is one of the techniques in Natural Language Processing (NLP). In this paper, we discuss about semantic analysis and explore different works that have been done on semantic analysis by different researchers.
Keywords: Corpus, Cross Language Information Retrieval (CLIR), Information Retrieval (IR), Natural Language Processing (NLP), Machine Translation, Morphological analysis, Semantic analysis, Syntactic analysis.

I. INTRODUCTION

Natural Language Processing (NLP) is the process of computer analysis of input provided in a human language (natural language), and conversion of this input into a useful form of representation. It is the language spoken by people like English, Hindi, Urdu, Kannada, Telugu and many more as opposed to programming language such as C, C++, Java etc. [1]. The field of NLP is primarily concerned with getting computers to perform useful and interesting tasks with human languages. The NLP is concerned with helping computer to better understand the human language. The input/output of NLP system can be written text or speech. To process written text, we need: Lexical, syntactic, semantic knowledge about the language, discourse information and real world knowledge**.** To process spoken language, we need everything required to process written text, signal processing, the challenges of speech recognition and speech synthesis and phonetic analysis.

1.1 Corpus: Large representation or collection of language material stored in computer processable form. There are Tagged corpuses in which all the words are assigned with the exact meaning depending on the given context and Raw corpus where all the words are assigned the actual meanings but without sense tagging. Example: Tata is a big industrial plant. Raw corpus gives us all the possible meaning of the word plant (herb, flower, vegetable, shrub, weed, factory, business unit) whereas Tagged corpus returns the actual meaning of the word plant in this sentence (factory, business unit).

1.2 Syntactic- The goal of syntactic analysis is to determine whether the text string on input is a sentence in the given (natural) language. It also checks for grammatical structure and part of speech tagging. If the text string belongs to a language, the result of the analysis contains a description of the syntactic structure of the sentence, for example in the form of a derivation tree or parse tree. Such type of formalizations aims at making computers "understand" relationships between words (and indirectly between corresponding people, things, and actions). Syntactic analysis can be utilized for instance when developing a punctuation corrector, dialogue systems with a natural language interface, or as a building block in a machine translation system.

1.3 Morphology- It is derived from the Greek word. Morph→shape or form, ology→study of something. Morphology in linguistics is the scientific study of forms and structure of words in a language [2]. Morphology as a sub-discipline of linguistics was named for the first time in 1859 by the German linguist August Schleicher who used the term for the study of the form of words.[3]. Morphology includes the following: Morphemes, word, simple words, complex words, affixes, bound (suffix and prefix), free morphemes.

1.4 Semantics –Semantics is a sub discipline of linguistics which focuses on the study of Meaning. Semantics analysis understands the meaning of every element in the given language and finds out how it is constructed, interpreted and negotiated both by speakers and listeners of the language. Meaning = Extension and Intention. So meaning in semantics, is defined as Extension: Which is the thing in the world that the word/phrase refers to, and Intention**:** Which concepts/mental images that the word/phrase evokes [4], or in simple terms semantic is concerned about what words mean and how these meaning combine in sentences to form sentence meaning. It is important because when CLIR is involved one sentence in one language might sound different in other language. When conversion or translation is concerned we need to ensure that the meaning of the sentence is retained even after the translation.

Example: naanu yaaru? In Kannada, When translated to English: me who. But in reality the correct conversion is "WHO AM I?" However semantic analysis is not an easy task to achieve because of the following reasons. 1) Language is ambiguous in nature [5]. 2) Language is dynamic in nature- new words generation and changes in the language require updates even in the controlled vocabularies. 3) It is a challenging task to gain common sense knowledge for example; there was a heavy thunder and lightning last night. Farmers feared for their plantations.

II. METHODS TO ACHIEVE SEMANTIC ANALYSIS

2.1 Latent semantic analysis

According to Landauer and Dumais Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text [6]. They also say that the articles of the three research Foltz, Kintsch & Landauer, Rehder, et al., and Wolfe, et al., exploits the new theory of knowledge induction and representation that provides a method for determining the similarity of meaning of words and passages by analysis of large text corpora. After processing a large sample of machine-readable language, Latent Semantic Analysis (LSA) represents the words used in it, and any set of these words—such as a sentence, paragraph, or essay—either taken from the original corpus or new, as points in a very high (e.g. 50-1,500) dimensional "semantic space". LSA is closely related to neural net models, but is based on singular value decomposition, a mathematical matrix decomposition technique closely akin to factor analysis that is applicable to text corpora approaching the volume of relevant language experienced by people [6].

2.2 Probabilistic semantic analysis

In the year 1999, Thomas Hofmann, argued that Probabilistic semantic analysis is more principled way to achieve semantic analysis than latent semantic analysis. Hofmann also presented a paper with the same theory [7]. Probabilistic Latent Semantic Analysis is a novel statistical technique for the analysis of two-mode and co-occurrence data, which has applications in information retrieval and filtering, natural language processing, machine learning from text, and in related areas. Compared to standard Latent Semantic Analysis which stems from linear algebra and performs a Singular Value Decomposition of co-occurrence tables, the proposed method is based on a mixture decomposition derived from a latent class model. This results in a more principled approach which has a solid foundation in statistics.

III. LITERATURE SURVEY

India is country of unity in diversity. Unity in diversity is a concept of "unity without uniformity and diversity without fragmentation"[8]. India has wide variety of languages nearly 30 however the official language of the Union Government of Republic of India is Standard Hindi, while English is the secondary official language [9]. There is a big scope for CLIR in a multi-lingual country like India. Because even today most of the government offices still process the data in regional languages. NLP and its techniques play a vital role in order to have an effective communication. Machine translation was started in the early eighties which were proposed using Sanskrit as Interlingua for translation to and from Indian languages by are research and development projects at Indian Institute of

Technology (IIT) Kanpur, International Institute of Information Technology (IIIT) Hyderabad.

Semantic analysis had its inception in the early years of 1890's. Ludwig Wittgenstein stated that language has a single underlying logic, which can be explained by analyzing language and the world and their (picturing) relation but later the author went on to define language as a vast collection of different practices. Ludwig debated that language is a 'game' where it is based on agreed-upon rules [5]. The author found out the importance of meaning (Semantics) in a language. Poroshin.V.A says that 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. Primarily one needs to complete morphological and syntactical analysis before trying to solve any semantic problem. Formalization of NL leads us to solutions of all these problems. There are a lot of theories and opinions about how to do it [10]. Poroshin.V.A also talks about the theories of semantics proposed by Professor Tuzov V.A which can analyse Russian newspapers with high precision. These theories can be found in the above reference that has been made here.

There are a good number of researches done on LSA. Thomas K Landauer [6], have given a thorough discussion on LSA briefing about different ideologies. Landauer brought up two theories for the construction of LSA (1) simply as a practical expedient for obtaining approximate estimates of the contextual usage substitutability of words in larger text segments and similarities among words and text segments that such relations may reflect, or (2) as a model of the computational processes and representations underlying substantial portions of the acquisition and utilization of knowledge. Latent semantic analysis is a process of analyzing relationship between a set of documents and the terms that the documents contain. This analysis is produced based on the set of concepts related to the documents and terms. It assumes words that are close in meaning will occur in similar pieces of text. Later a matrix of words count per paragraph is constructed where a row contains unique word and columns represent each paragraph. This leads to Single Value Decomposition (SVD), a mathematical structure to reduce the number of columns. Here the constructed matrix is broken into three parts. First consists of original row entities as vectors of derived orthogonal factor values and second done the same way for column and third is by constructing diagonal matrix containing scaling values such that when all the three components are matrix-multiplied, original matrix is constructed [6]. In the next step, words are compared by taking the cosine of the angle between vectors formed by any two rows. If similar words are encountered, a value 1 is given and value 0 in case of dissimilar words. They were successful in educating us with the complete knowledge of LSA however they have not presented any tool that practically support their theory and for semantic analysis, syntactic knowledge is also very important and they have not high lightened on this.

Donald E. Walker [11] has proposed a system called Pintle. Pintle—Procedures for Syntactic and Semantic Analysis. Pintle, the syntactic and semantic component of this version of the speech understanding system, is based on the Winograd "Computer Program for Understanding Natural Language". It is a top-down system for linguistic analysis in which syntax, semantics, and inference are combined to direct the processing of questions, statements, and commands. Initially all the words and corresponding sounds are stored in the database. If the Pintle encounters a new word that is not there in the word verifier, then it will use the appropriate word functions and if the word that is typed in is not present in the word verifier then the word is rejected. The author has explained the working of Pintle in a very organized and in a detailed way. However the word verifier collects the results for each word in a set, eliminates the impossible words, and constructs a list ordering the rest of the words according to confidence level. The word with the highest ranking is returned to Pintle, any others are saved on a backup list to be used successively if their predecessor does not lead to the prediction of a new set of words, one or more of which can be found in the utterance. The ending position of the accepted word is used as the starting point for testing words in this new set. It also requires the generation of the parser in an organized manner

Hiroyuki Yamauchi has developed a system called KAUS which is abbreviated as Knowledge Acquisition and Utilization System. This tool was based on the theories of axiomatic structures which has the capability of deductive inference and automatic program generation for database access. It can be also applied to the logic programming of the semantic processing of natural language. The writer has given the complete picture of axiom set, deductive inference rule. The main idea which the author has adapted is direct correspondence between the basic

sentence patterns (syntax) of natural language and the extended atomic formulas in KAUS, and that the pattern matching method can be used together with the deductive inference rule. A more characteristic is that words in a clause are put into four groups preserving the word order, each of which contains subjects, the main verb, direct objects/complements and indirect objects/complements respectively [12]. Mallamma V. Reddy presented a tool called CLIR which converts the given English word to Kannada and Telugu. There were few challenges in translating names. It is also been discussed about the issues concerned with the transliteration of different sources namely Goutham vs. Goutam, Soumya vs. Sowmya, Bharathi vs. Bharati [13]. This needs to be appreciated a lot as they were the pioneers for such an attempt in Kannada and were successful in bringing the output. However there was no semantic analysis done. This gives me the opportunity to introduce my proposed plan of work in order to overcome this limitation.

3.1 Key researchers in the field of semantics

Richard Montague (1930-1971), was an American mathematician and philosopher Studied at The University of California, Berkeley, He pioneered a logical approach to natural language semantics, which became known as the Montague Grammar.

Donald Davidson (1917-2003), was an American philosopher born in Springfield, Massachusetts. Developed an inspired approach to truth-conditional semantics Famously said "there is no such thing as a language, not if a language is anything like what many philosophers and linguists have supposed. There is therefore no such thing to be learned, mastered, or born with."

Ray Jackendoff (1945), an American linguist and professor of philosophy at Tufts University. Jackendoff mainly developed the semantic framework of Conceptual semantics which aim is to provide a characterization of the conceptual elements by which a person understands words and sentences to provide an explanatory semantic representation (title of a Jackendoff 1976 paper) Built on the work of Noam Chomsky and his ideas on the acquisition of language.

Geoffrey Leech (1936), a Professor of Linguistics and Modern Languages at Lancaster University from 1974-2002. One main area of his academic interests in English Linguistics is semantics (also very influential in the field of pragmatics) Wrote Semantics (1974; 2nd edition. 1981) [4]

IV. PROPOSED PLAN OF WORK

4.1 Identify different approaches to the analysis of meaning in language namely at lexical, clausal and discourse levels

Lexical semantics, a subfield of linguistic semantics that studies how and what the words of a language denote. These are those words which the human learn throughout his life as contradicted to the grammar which they learn when they are young. Clausal: A group of words that contains a subject and a predicate [14]. Discourse: Discourse can be defined as language beyond the level of a sentence, i.e. the language behaviours linked to social practices. Here the Language can be implied as a system of thought. Discourse Analysis (DA) is a modern discipline of the social sciences that covers a wide variety of different sociolinguistic approaches. Analysis of discourse looks not only at the basic level of what is said, but takes into consideration the surrounding social and historical contexts [15].

4.2 Identify and apply semantic theories/models that are relevant to the analysis of a specific linguistic feature or domain.

4.3 Critically discuss an analysis by relating it to the current literature

Finally we focus on building the system that will parse through the above tasks automatically with less human intervention. The system thus will be designed and developed which encompasses the following features:

Scalability: The system can handle large volume of data

Adaptability: The system can also be used by organizations other than educational

institutions that organize events frequently

Feasibility: The system undoubtedly ensure ease of operation and effective retrieval

V. CONCLUSION

From all the discussion I have had so far, it is evident that in order to process any natural language syntactic, morphology and semantics are essential. Without syntactic or morphological analysis we cannot afford to do semantic analysis. It is observed that signal processing also plays a vital role because it transforms the spoken words into text however this signal processing can be neglected to some extent as computer takes the input directly as text, no matter whether it is from keyboard, file or some other source. Later the text is taken for syntactic analysis where the structures of the words, associated grammar rules are substantiated. Following which we have the semantic

analysis which profoundly deals with 'Meaning'. It also indicates how the words and sentences relate to the real world (Pragmatics). All of this can be summarized with an interpretation process that maps natural language sentences to the formal language, or from one formal language to others. However interpretation processes are of different types, depending on which formal language and stage is being considered. For example, a parser is an interpretation process that maps natural language sentences to their syntactic structure or representation (result of syntactic analysis) and their logical form (result of semantic analysis). The parser uses the rules of grammar and word meanings (in a lexicon). This mapping could be sequential or simultaneous. A contextual interpretation maps the logical form to its final knowledge representation. It is significant instead of directly leaping into NLP techniques, one should make an attempt of knowing "understanding", "generation", "processing". This survey emphasizes on semantic analysis, methods to achieve semantic analysis, efforts of different researchers and their work.

REFERENCES

[1] "*Natural Language Processing and Information Retrieval"* by Rada Mihalcea and Dragomir Radev www.cse.unt.edu/~rada/CSCE5290/Lectures/ Intro. 7th International Conference, CICLing 2006, Mexico City, Mexico, February 19-25, 2006. Proceedings. pp 319-330

[2] Aronoff, M. and Fudeman, K., (Date unknown). "*What is Morphology?"* [Pdf] Oxford: Blackwell Date: October 25, 2010 | ISBN-10: 1405194677 | ISBN-13: 978-1405194679

[3] Booij, G. E., (2007). *"The Grammar of Words: An Introduction to Linguistic Morphology"*. 2nd edition. Oxford: Oxford University Press

[4] Wood, G.C., (2011). Lecture on Introduction to Semantics at the University of Sheffield.

[5] *"Semantic analysis of text and speech SGN-9206 Signal processing graduate seminar II, Fall 2007"*Anssi Klapuri Institute of Signal Processing, Tampere University of Technology, Finland

[6] Landauer, T. K., Foltz, P. W., & Laham, D. (1998).*"Introduction to Latent Semantic Analysis. Discourse Processes"*, 25, 259-284.

[7] Thomas Hofmann, *"probabilistic semantic analysis"*, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 50-57

[8] Roxanne, Lalonde (April 1994). *"Edited extract from M.A. thesis"*. Unity in Diversity: Acceptance and Integration in an Era of Intolerance and Fragmentation (Thesis). Ottawa, Ontario: Department of Geography, Carleton University

[9] Schwartzberg, Joseph E., 2007. Encyclopaedia Britannica, India—Linguistic Composition

[10] *"Semantic analysis of Natural Language"*, by Poroshin.V.A, Saint-Petersburg State University, 2004.

[11] *"Speech understanding through syntactic and semantic analysis"*, Donald E. Walker Artificial Intelligence Center, Stanford Research Institute, Menlo Park, California., DE Walker, AI center, advanced papers of the conference

[12] Hiroyuki Yamauchi, *"Processing of syntax and semantics of natural language by predicate logic",* institute of space and aeronautical science, university of Tokyo, proceedings of the 8th conference on computational linguistics, pages 389-396

[13] Mallamma V Reddy, *"English to Kannada/Telugu name transliteration in CLIR: a statistical approach"*, India IJMI International Journal of Machine Intelligence ISSN: 0975–2927 & EISSN:0975–9166, Volume 3, Issue 4, 2011, pp-340-345

[14] Richard Nordquist, professor emeritus of rhetoric and English at Armstrong Atlantic State University, "*clausal meaning"*.

[15] Johnstone B., (2008). *Discourse Analysis*, 2nd edition, Oxford: Blackwell, June 2008 — Volume 12, Number 1 ISBN 978-1-4051-4427-8(paper)