THINK INDIA JOURNAL

ISSN:0971-1260

Vol-22-Issue-10-November-2019

Survey of Applications of Natural Language Processing: A Review

[1]VatslaGarg,[2]Shreshth Bansal,[3]Sonam Mittal

[1,2]Department of Computer Science, B K Birla Institute of Engineering and Technology, Pilani, Rajasthan

[1]vatslagarg99@gmail.com, [2]shreshthbansal2505@gmail.com, [3]sonam.mittal@bkbiet.ac.in

[3]Associate Professor (Computer Science Department), B K Birla Institute of Engineering and Technology, Pilani, Rajasthan

Abstract- Natural language processing (NLP) has recently gained much attention for representing and analyzing text by computerized means. The primary goal of NLP is to implement within computers the skill to understand a normal human language. One of the motivations of NLP is for the society whose access to web information is obstructed simply by their inability to use the keyboard and operating system. It has spread its applications in various fields such as machine translation, email spam detection, information extraction, summarization, medical and question answering, etc.

Keywords - Natural language processing, Phases of NLP, History and Future of NLP, Applications of NLP

I. INTRODUCTION

The main idea behind Natural Language Processing (NLP) is to develop the computer systems that can understand and synthesize human natural language. NLP was originated when the study of human-languages developed the concept of communication with non-human devices. According to the survey, many researchers explained NLP as an area of research and applications that explores how computers can be used to manipulate natural language text or speech to perform useful tasks [1].

The research and development in NLP over the last sixty years as stated by Church and Rau [2] can be categorized into the following five areas:

- Natural Language Understanding
- Natural Language Generation
- Speech or Voice recognition
- Machine Translation
- Spelling Correction and Grammar Checking

There are some terminologies used in NLP [3]

- Morphology – It is a study of the construction of words from primitive meaningful units.
- Syntax − It is used to arrange the words to make a sentence.
- Semantics − It is concerned with the meaning of words and how to combine words into meaningful phrases and sentences.
- Discourse − It deals with how the immediately preceding sentence can affect the interpretation of the next sentence.
- Pragmatics − It is used to understand the sentences in different situations and how the interpretation of the sentence is affected [4].

The increasing demands for software that process text of all kinds have tremendously been influenced by the advent of the Internet and the World Wide Web. Over a decade, Internet publishing has become a commonplace activity for private individuals, commercial enterprises, and government organizations, as well as traditional media companies, and the medium of most of these communications and transactions is primarily natural language. Various forms of keyword processing provide access to Web sites as well as organizational principles for retrieving, navigating and browsing web pages within those sites. Search engines and spam filters are now of everyday life and work well enough that their viability as products is not in question[5].

II. PHASES OF NATURAL LANGUAGE PROCESSING(NLP)

ISSN:0971-1260

Vol-22-Issue-10-November-2019

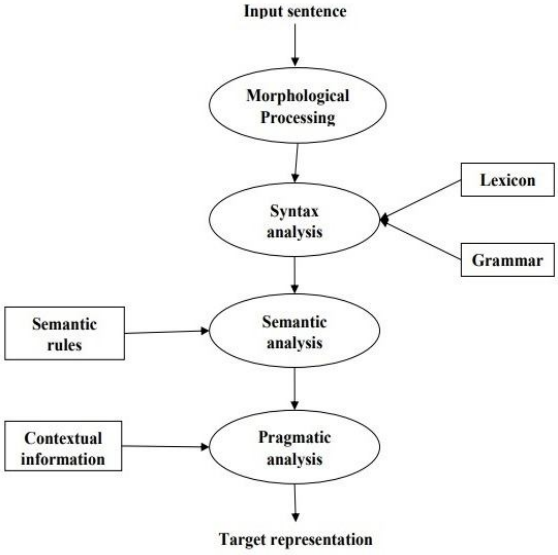

Fig. 1 - Four phases of compiler

1) MORPHOLOGICAL PROCESSING:

The first phase of Natural Language Processing is Morphological Processing. Morphology is the study of the structure of words or it can be said that it is a study of the construction of words from primitive meaningful units known as Morphemes.The purpose of this phase is to break chunks of language input into sets of tokens corresponding to paragraphs, sentences, and words.

An example of Morpheme could be, the word precancellation can be morphologically structured into three separate morphemes: the prefix pre, the root cancella, and the suffix –tion [6]. The interpretation of morpheme stays the same across all the words, just to understand the meaning humans can break any unknown word into morphemes. For example, adding the suffix –ed to a verb conveys that the action of the verb took place in the past. The words that cannot be further divided and have their individual meanings are called Lexical morpheme (e.g.: bat, car) The words (e.g. -ed, -ing, -est, -ly, -ful) that are combined with the lexical morpheme are known as Grammatical morphemes (eg. Worked, Consulting, Smallest, Likely, Use). Those grammatical morphemes that occur in a combination called bound morphemes (eg. -ed, -ing) Grammatical morphemes can be divided into bound morphemes and derivational morphemes.

2) SYNTAX ANALYSIS:

The second phase of NLP is Syntax Analysis. The main purpose of this phase is to check whether the sentence is correctly formed or not and to break it up in such a structure that it shows syntactic relationships between different words. A syntactic analyzer performs this using a dictionary of word definitions (the lexicon) and a set of syntax rules (the grammar). 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. Syntax conveys meaning in most languages because order and dependency contribute to connotation [7]. For example, the two sentences: ‘The cat chased the mouse.’ and ‘the mouse chased the cat.’ differ only in terms of syntax, yet convey quite different meanings.

3) SEMANTIC ANALYSIS:

The third phase of NLP is Semantic Analysis. The main purpose of this phase is to find the exact meaning or the dictionary meaning of the text.For example, the semantic analyser would reject a sentence like “Hot ice-cream” because in this phase text is checked for its meaningfulness.

This level of processing can include the semantic disambiguation of words with multiple senses; in a cognate way to how syntactic disambiguation of words that can errand as multiple parts-of-speech is adroit at the syntactic level. The semantic level scrutinizes words for their dictionary elucidation, but also for the elucidation they derive from the milieu of the sentence.

4) PRAGMATIC ANALYSIS:

The fourth phase of NLP is Pragmatic analysis.Pragmatic analysis simply fits the actual objects/events, which exist in a given context with object references obtained during the last phase (semantic analysis).

The main purpose of this phase is using and understanding the sentences in different situations and determining how the interpretation of the sentence is affected. For analysis in this phase therequisite is much world knowledge, including the understanding of intentions, plans, and goals. The

ISSN:0971-1260

Vol-22-Issue-10-November-2019

structure representing what was said is reinterpreted to determine what was actually meant. E.g. "close the door?" should have been interpreted as a request rather than an order.

III. HISTORY OF NATURAL LANGUAGE PROCESSING(NLP)

The history of NLP is divided into four phases. The phases have distinctive concerns and styles:

1. First Phase (Machine Translation Phase) - Late 1940s to late 1960s:

This phase focused on the work done on machine translation (MT). It was the first computer-based application related to natural language. This phase was a period of enthusiasm and optimism.

- The research on NLP started in the early 1950s after Booth &Richens' investigation and Weaver's memorandum on machine translation in 1949.
- Weaver and Booth started one of the earliest MT projects in 1946, on computer translation which was based on breaking enemy codes during World War II.
- According to Liddy [7], the earliest works in MT followed the basic view, that the only difference between languages was vested in their vocabularies and the permitted word orders. Hence systems that were made from this perspective used dictionary-lookup.
- This was done without considering the lexical ambiguity inherent in natural language. This generated poor results which made researchers to come up with a more sufficient theory of language.
- It was Chomsky's 1957 publication of the syntactic structures which introduced the idea of generative grammar, which gave the linguistic a better understanding of how they could help the machine translation [8]. Subsequently, other NLP application areas began to emerge, such as speech recognition.

2. Second Phase (AI Influenced Phase) – Late 1960s to late 1970s:

The work was majorly related to world knowledge and its role in the construction and manipulation of meaning representations. That is why this phase is also called the AI-flavored phase.

- In early 1961, the work began on the problems of addressing and constructing data or knowledge base and focused on the issue of how to represent meaning and developing computationally tractable.
- Examples areChomsky's 1965 transformational model of linguistic [8]; case grammar of Fillmore[9], semantic networks of Quillian [10] and conceptual dependency theory of Schank, which explained syntactic anomaliesand provided semantic representations.
- Formalisms representation which included Wilks' preference semanticsand Kay's functional grammar; Augmented transition networks of Woods which extended the power of phrase-structure grammar by incorporating mechanisms from programming languages [11].
- In the same year, a BASEBALL question-answering system [12] was also developed. The input to this system was restricted and the language processing involved was a simple one.

3. Third Phase (Grammatical-logical Phase) – Late 1970s to late 1980s:

It was the grammatical-logical phase. Due to the failure of practical system building in the last phase, the researchers moved towards the use of logic for knowledge representation and reasoning in AI.

- The grammatical-logical approach, towards the end of the decade, helped us with powerful general-purpose sentence processors like SRI's Core Language Engine and Discourse Representation Theory [13], which offered a means of tackling more extended discoursewithin the grammatical-logical framework.
- In this phase, we got some practical resources & tools like parsers, e.g. Alvey Natural Language Tools (Briscoe et al., 1987) [14] along with more operational and commercial systems, e.g. for database query.

ISSN:0971-1260

Vol-22-Issue-10-November-2019

- The work on lexicon in the 1980s also pointed in the direction of the grammatical-logical approach.

4. Fourth Phase (Lexical & Corpus Phase) – The 1990s:

This was a lexical & corpus phase. In the early 1980s, Computational grammar theory became a very active area of research linked with logic for meaning and knowledge's ability to deal with the user's beliefs and intentions and with functions like emphasis and themes.This phase had a lexicalized approach to grammar that appeared in the late 1980s and became an increasing influence. There was a revolution in natural language processing in this decade with the introduction of machine learning algorithms for language processing.

IV. CURRENT AND FUTURE PROGRESS OF NATURAL LANGUAGE PROCESSING(NLP):

Some of the active researches on NLP phenomena include the:

- Syntactic phenomena: those that pertain to the structure of a sentence and the order of words in the sentence, based on the grammatical classes of words rather than their meaning (e.g. discriminative models for scoring parses, coarse to fine efficient approximate parsing, dependency grammar); Machine translation (e.g. models and algorithms, low- resource and morphological complex language).
- Semantic phenomena: those that pertain to the meeting of a sentence relatively independent of the context in which the language occurs (e.g. sentiment analysis, summarization, information extraction, slot filling, discourse analysis, textual entailment).
- Pragmatic phenomena such as Speech: those that relate the meaning of a sentence to the context in which it occurs. This context can be linguistic (such as the previous text or dialogue) or, non-linguistic (such as knowledge about the person who produced the language, about goals of the communication, about the objects in the current visual field, etc. (e.g. language modelling-syntax and semantics, models of acoustics, pronunciation) [15].
- Speech recognition and information retrieval have finally gone commercial and there is a ton of text and speech on the Internet, cell phones, etc. It is now clear that studies regarding anything about a language are possible, e.g. formalizing some insights e.g. discrete knowledge (what is possible) and continuous knowledge (what is likely) studying the formalism mathematically; developing and implementing algorithms and testing on real data.
- The current and on-going future changes or improvements that need to be done to NLP are: to add features to existing interfaces, back end processing should be fully implemented (e.g. information extraction and normalization to build databases). Another anticipated improvement is of having handheld devices with translators and personal conversation recorder with topical searches [15].

V. APPLICATIONS OF NATURAL LANGUAGE PROCESSING:

In the present scenario, Natural Language Processing (NLP) is applied to various areas like Machine Translation, Email Spam detection, Information Extraction, Summarization, Question Answering, etc.

i. MACHINE TRANSLATION:

Nowadays most of the world is online, therefore the task of making data available and accessible to all is a major challenge. The main barrier in making data accessible is language as there are different languages with different sentence structures and grammar. Machine Translation is generally translating phrases from one language to another with the help of a statistical engine like Google Translate. The challenge with machine translation technology is not only keeping the words or sentences but also keeping its meaning and grammar along with tenses. The statistical machine learning gathers large data that they find the parallel between two languages and they crunch their data to find the likelihood that something in Language. In recent years, various methods have been proposed to

ISSN:0971-1260

Vol-22-Issue-10-November-2019

automatically evaluate machine translation quality by comparing hypothesis translations with reference translations. Examples of such methods are word error rate, position-independent word error rate *(Tillmann et al., 1997)*, generation string accuracy *(Bangalore et al., 2000)*, multi-reference word error rate [15].

ii. TEXT CATEGORIZATION:

In Categorization systems there are inputs of large data like official documents, military casualty reports, market data, newswires, etc. and assign them to predefined categories or indices. Many companies have been using categorization systems to categorize trouble tickets or complaint requests and routing to the appropriate desks. One of the latest used applications of text categorization is email spam filters. Spam filters are becoming important to defense against unwanted emails. The major challenge is to extract meaning from strings of text. A filtering solution that is applied to an email system uses a set of protocols to determine which of the incoming messages spams are and which are not. There are different types of spam filters available like [3]:

- Content filters: Review the content within the message to determine whether it is a spam or not.
- Header filters: Review the email header looking for fake information.
- General Blacklist filters: Stops all emails from blacklisted recipients.
- Rules-Based Filters: It uses user-defined criteria. Such as stopping mails from specific persons or stopping mail including a specific word.
- Permission Filters: Require anyone sending a message to be pre-approved by the recipient.
- Challenge-Response Filters: Requires anyone sending a message to enter a code in to gain permission to send email.

iii. INFORMATION EXTRACTION:

The main aim of Information extraction is to identify phrases of interest in textual data. In many applications, extracting entities such as names, places, events, dates, times and prices is a powerful way to summarize the information relevant to a user's needs. Domain-specific search engineincreases the accuracy and efficiency of a directed search as important information is automatically identified.

For example, noticing the popup ads on any websites showing the recent items you might have looked at an online store with discounts. In Information Retrieval two types of models have been used. Both models assume that a fixed vocabulary is present. But in the first model, a document is generated by first choosing a subset of vocabulary and then using the selected words any number of times, at least once without any order. This is called the Multi-variate Bernoulli model. It takes the information of which words are used in a document irrespective of the number of words and order. In the second model, a document is generated by choosing a set of word occurrences and arranging them in any order. this model is called the multinomial model, in addition to the Multi-variate Bernoulli model, it also captures information on how many times a word is used in a document

Extracted information can be applied for a variety of purposes, for example, to prepare a summary, to build databases, identify keywords, classifying text items according to some predefined categories, etc.It has been suggested that many IE systems can successfully extract terms from documents, acquiring relations between the terms is still a difficulty. PROMETHEE is a system that extracts lexico-syntactic patterns relative to a specific conceptual relation [16]. IE systems should work at many levels, from word recognition to discourse analysis at the level of the complete document. There's a system called MITA (Metlife's Intelligent Text Analyzer) [17]that extracts information from life insurance applications.

iv. SUMMARIZATION:

In this modern era, the most valuable thing is data or information. The major problem occurring is overloading of information that causes improper access to useful data, therefore we are in a serious need to summarize the information because the flood of information over the net is not going to stop.Text summarization may be defined as the technique to create a short, accurate summary of longer text documents. So, an ability to summarize the data while keeping the meaning intact is highly required.

ISSN:0971-1260

Vol-22-Issue-10-November-2019

Text summarization is important not only to recognize the large set of data but also to understand its deeper meaning.The types of text summarization depend on the basis of the number of documents and the two important categories are single-document summarization and multi-document summarization.

A large amount of annotated data is needed for learning techniques. Few techniques are as follows–

- Bayesian Sentence based Topic Model (BSTM) uses both term-sentences and term-document associations for summarizing multiple documents.
- Factorization with Given Bases (FGB) is a language model where sentence bases are the given bases and it utilizes document-term and sentence term matrices. This approach groups and summarizes the documents simultaneously [18].
- Topic Aspect-Oriented Summarization (TAOS) is based on topic factors. These topic factors are various features that describe topics such as capital words that are used to represent an entity. Various topics can have various aspects and various preferences of features are used to represent various aspects [19].

v. MEDICINE:

NLP is applied in the medical field as well. The Linguistic String Project-Medical Language Processor is one of the large-scale projects of NLP in the field of medicine. The LSP-MLP helps to enable physicians to extract and summarize information of any signs or symptoms, drug dosage, and response data with aim of identifying possible side effects of any medicine while highlighting or flagging data items. The Specialist System developed by The National Library of Medicineis expected to function as an Information Extraction tool for Biomedical Knowledge Bases, particularly Medlineabstracts. The lexicon was created using MeSH (Medical Subject Headings), Dorland's Illustrated Medical Dictionary and general English Dictionaries. The Centre d'InformatiqueHospitaliere of the Hospital Cantonal de Geneve is working on an electronic archiving environment with NLP features. In the first phase, patient records were archived. At a later stage, the LSP-MLP has been adapted for French, and finally, a proper NLP system called RECITHas been developed using a method called Proximity Processing. Its task was to implement a robust and multilingual system able to analyze/comprehend medical sentences and to preserve a knowledge of free text into a language-independent knowledge representation. The Columbia University of New York has developed an NLP system called MEDLEE (MEDical Language Extraction and Encoding System) that identifies clinical information in narrative reports and transforms the textual information into structured representation.

vi. QUESTION ANSWERING:

Question answering is another main application of natural language processing (NLP). Search engines provide almost every information of the world on our fingertips but it lacks in answering the questions written by human beings in their natural language. Question answering is a computer science discipline within the fields of AI and NLP. It mainly focuses on developing the systems which can automatically answer the questions posted by human beings in their natural language.A computer system that understands the natural language has the capability of a program system to translate the sentences written by humans into an internal representation so that the valid answers can be generated by the system. The exact answers can be generated by doing syntax and semantic analysis of the questions. Lexical gap, ambiguity, and multilingualism are some of the challenges for NLP in building a good question answering system.

VI. CONCLUSION:

As a computerized approach to analyzing text, NLP is continually striving forward. Researchers are continuously working to gather knowledge on how human beings understand and use different languages. From the early 1940s, NLP is serving human beings by making their tasks easy by developing several applications. The main reason behind developing appropriate tools and techniques is to make computer systems understand and manipulate natural language so that they can perform several tasks and help human beings in each and every area.Technologies, such as string matching, keyword search, glossary lookup are now on the past as, to more forward-looking technologies such as

ISSN:0971-1260

Vol-22-Issue-10-November-2019

grammar checkers, conceptual search, and event extraction, interlingual ongoing and striving forward.

REFERENCES:

[1] N. Kaur1, V. Pushe and R Kaur,"Natural Language Processing Interface for Synonym", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.7, July- 2014, pp. 638-642 ,ISSN 2320–088X.

[2] K. W Church and L.F Rau," Commercial applications of Natural Language Processing". Communication of the ACM, vol 38, No. 11,November 1995

[3] SamtaTembhekar, Monika Kanojiya, "A Survey Paper on Approcahes of Natural Language Processing (NLP), International Journal of Advance Research, Ideas and Innovations in Technology, 2017

[4] Ann Copestake – "Natural Language Processing". http://www.cl.cam.ac.uk/users/aac/. Copyright Ann Copestake, 2003– 2004(7-6-17)

[5] P. Jackson and I. Moulinier," Natural Language Processing for Online Applications": Cambridge University press, New York.2012, page 7-9.

[6] DikshaKhurana, Aditya Koli, KiranKhatter, Sukhdev Singh, Natural Language Processing: State of The Art, Current Trends and Challenges, August 2017

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

[8] N. Chomsky.Syntaticstructures.The Hague: Mouton & Co. Reprinetd 1978,Peter Lang Publisjing.

[9] C. J. Fillmore," The Case for Case". In Bach and Harms (Ed.): niversals in Linguistic Theory. New York: Holt, Rinehart, and Winston, 1-88, 1968.

[10] R. Quillian, "A notation for representing conceptual information: An application to semantics and mechanical English para- phrasing", SP-1395, System Development Corporation, Santa Monica, 1963.

[11] W.C. Mann & S. Thompson, "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization", 1988. Text 8 (3). Pp. 243-281.

[12] Green Jr, B. F., Wolf, A. K., Chomsky, C., &Laughery, K. (1961, May). Baseball: an automatic question-answerer. In Papers presented at the May 9-11, 1961, western joint IREAIEE-ACM computer conference (pp. 219-224). ACM.

[13] Kamp, H., &Reyle, U. (1993). Tense and Aspect. In from Discourse to Logic (pp. 483689). Springer Netherlands.

[14] Lea, W.A Trends in speech recognition, Englewoods Cliffs, NJ: Prentice Hall, 1980.

[15] Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000, May). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In LREC

[16] Morin, E. (1999, August). Automatic acquisition of semantic relations between terms from technical corpora. In Proc. of the Fifth International Congress on Terminology and Knowledge Engineering-TKE'99.

[17] Glasgow, B., Mandell, A., Binney, D., Ghemri, L., & Fisher, D. (1998). MITA: An information-extraction approach to the analysis of free-form text in life insurance applications. AI magazine, 19(1), 59.

[18] Wang, D., Zhu, S., Li, T., Chi, Y., & Gong, Y. (2011). Integrating document clustering and multidocument summarization. ACM Transactions on Knowledge Discovery from Data (TKDD), 5(3), 14.

[19] Fang, H., Lu, W., Wu, F., Zhang, Y., Shang, X., Shao, J., &Zhuang, Y. (2015). Topic aspect-oriented summarization via group selection. Neurocomputing, 149, 1613-1619.