ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Natural Language Processing Various Applications and Developments in Indian Languages: Review

Sakthi Vel S.
Lecturer in Computational Linguistics
Department of Linguistics, University of Kerala
Thiruvananthapuram, India.

Abstract: The entitled paper details about the very fundamentals of NLP technologies and its relevance in Indian languages along with a view of its current status. NLP is an inter-disciplinary field of Computer Science, Linguistics, Psychology and Logic, Which concerned with the interactions between User and Computer. It is the process in which computers analyze the input provided in human languages and convert the input into a useful form of representation. NLP technologies are becoming extremely relevant in satisfying the need for user friendly interactions with computer. It is popularly known in other term as Human Computer Interaction (HCI). The HCI applications are specifically required in the areas of automatic or Machine Translation (MT), Information Retrieval/ Extraction (IR/IE), automatic Text and Speech processing etc. The soul focus of the entitled review paper is to provide an overall up to date view about the research and developments of NLP technologies in the multilingual Indian context. The first phase of the paper summed up with the introduction to NLP, levels of Language analysis, different components of NLP and overall structure of NLP. The second phase allotted to detail about the present scenario of various researches and developments of NLP technologies and systems in Indian languages.

Index Terms— NLP, Linguistics, Computer Science, Text, Speech, Corpora, Research and Development.

I. INTRODUCTION

The discussions and researches to explore and bring new advancement in the area of NLP technologies has years of story to tell. Natural Language Processing or conveniently abbreviated as NLP encompasses everything that a system requires to sense and generate Natural Languages. Artificial intelligence paved the way for the beginning of NLP and later it took a multi- disciplinary face by collaborating with various other schools of studies such as Computer science, Linguistics, Psychology and Data mining. The applications of NLP extends from querying archives, accessing collections of texts, extracting information and report generation to automatic text and speech manipulation and many more [1].

NLP in its most simple sense dwelt with Human-Computer interactions [2]. It empowers the users to manipulate and process the naturally occurring text with the help of a set of computational techniques. These theoretically motivated type of NLP techniques and tools make it possible for system to indulge in various levels of Linguistic analysis. In theory, NLP is an attractive method for Man- Machine Communication [6]. But the present reality is that even though machines had achieved the ability to convert large matrices with speed and grace long ago, it yet fails to sense the fundamentals of our written and spoken languages [6].

The history of NLP took origin in the year 1950s. The article "Machine and Intelligence" authored by Alan Turing proposed what is later came to be known in the computer world as Alan Turing Test. The goal of Natural Language Processor is to formulate common software that will interface, analyze, understand, process automatically and foster languages which human can use naturally. If the above mentioned are come to the implementation state you will be able to address your computer as you are addressing a fellow being [6].

The various applications of NLP are; Automatic Summarization, Machine Translation (MT), Named Entity Recognition (NER), Computational Lexicography, automatic Morphological analyzer, Optical Character Recognition (OCR), Parsing, Sentence Breaking, Digital Documentation, Sentiment analysis, POS-Tagging, Text mining, Automatic Speech Recognition (ASR), Text To Speech (TTS), Speech To Text (STT), Speech To Speech (STS), Speaker/ Speech Identification, Information Extraction/ Retrieval (IE/ IR), Question Answering (QR), Words Sense Disambiguation (WSD) and so on [1].

II. DIFFERENT LEVELS OF NLP

The main task NLP is to make computers understand and to perform useful tasks in Natural Languages. The NLP can have input and output in the form of speech or

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

writing. Different levels of language analysis in NLP are: Phonological/ Sound, Morphological/ Lexical, Syntactic, semantic, Discourse and Pragmatics [4, 9]. The different levels of Language analysis are given below;

A. Morphological or Lexical analysis: Individual words are analyzed into their components and non word token such as punctuation are separated from the words. The process of Morphological and Lexical analysis consists of detection of word boundary, Root and stem, Dictionary creation, Morphological Parsing and etc.

B. Syntactic analysis: syntax implies the idea that, liner sequences of words are transformed into structures and these words are related to one another [9]. It consists of Syntactical Parsing, tokenization, Pos Tagging, Automatic sentence construction and so on.

C. Semantic analysis: Semantic analysis indulges in the process of assigning relevant meanings to the words or sentences structure created by the syntactic analyzer. The process of finding/ identifying possible meaning in the given languages is called Semantics [9].

D. Discourse analysis: Discourse concerned with how the sentences preceded affect the interpretation of the following sentence [11].

E. Pragmatics analysis: It is about how sentences are used in different context and how this usage affects the interpretation of the sentence [11].

III. COMPONENTS OF NLP

NLP is classified into different components and they are connected to one another through the exchange of different Linguistics information. The following diagram represents various components of NLP. Let's have a glance on each components depicted in the diagram.

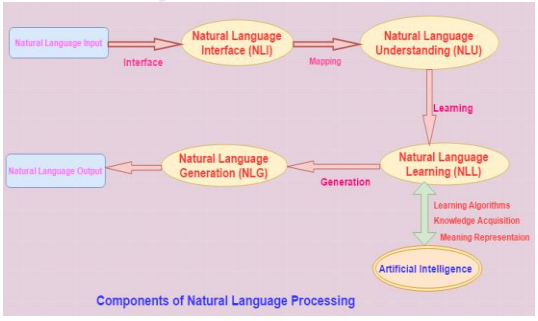

Figure 1. Levels of Natural Language Processing

a.Natural Language Interface (NLI): Natural Language Interface enables the users to communicate or interface with the computer or any other electronics devices through English, Tamil, Malayalam or any other natural languages.

b. Natural Language Understanding (NLU): Natural Language Understanding is a great challenge because in order to bring it into the level of application, system must possess not only a grammatical knowledge but also background knowledge including common sense knowledge [12]. NLU indulges in the process of mapping the given input in natural languages into useful representation and analyzing various aspects of the languages [5]. NLU also convert chunks of language text into more formal representations such as Predictive Logic and First Order Logic (FOL) structures that are easier for computer programs to manipulate. NLU makes out the required meanings from the multiple semantic data which can be derived from natural language expressions.

c.Natural Language Learning (NLL): Modern NLP learning algorithms are grounded in machine learning, specifically in statistical machine learning. The NLL researches in the modern statistical NLP algorithms requires an understanding of a number of different fields including Computer Science, Linguistics, Statistics particularly Bayesian statistics, Linear Algebra, theory of computation and Optimization theory. The aim of AI includes reasoning, knowledge acquisition, Logic, planning, learning, Language processing, communication, perception, the ability to shift and manipulate objects.

d. Natural Language Generation (NLG): NLG is the act or process to create meaningful sentences and phrases in the form of natural languages for some internal representation. It involves Text planning, Sentence planning and Text realization [5]. NLG requires extensive knowledge about languages and the ability to interpret it. This knowledge of interpretation requires high level support from the other acclaimed branch of Computer Science– The Artificial Intelligence [7].

IV. NLP IN INDIAN LANGUAGES: CURRENT SCENARIO

The current developments of Natural Language Processing in Indian languages context can be identified through the discussion of the advancement in the various areas of Language Technology such as Human Machine Interface, Corporal development, Multilingual E-content

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

generation, Language Tools and devices, Text processing and Speech processing.

1. Human Machine Interface level:

- ***Human Computer Interface (HCI):*** As a core area of Language Technology research, it increases the chances for non-technical people to quickly and successfully interact with computers. The finest example for Human Computer Interface is Bharat Operating System Solutions (BOSS). The joint effort of Centre for Development of Advanced Computing (C-DAC) and Centre for Free/ Open Source Software (NRCFOSS) Chennai brought BOSS into a reality. GNU/ Linux is the distributor of BOSS. Currently it is able to process18 Indian languages like Sanskrit, Assamese, Bengali, Tamil, Bodo, Gujarati, Hindi, Kannada, Punjabi, Kashmiri, Konkani, Manipuri, Maithili, Malayalam, Marathi, Oriya, Telugu and Urdu [5, 8].

- ***On Screen Keyboard/ Indian Languages Virtual Keyboard:*** The modern desktop comes with an on-screen keyboard with Indian language support. The easy and Smart Common Inputting Method tool provides input mechanism for various Indian languages. In the present state Unicode 6.1 supports all these mechanism for Indian Languages. This allows the users to provide input in all Indian Languages with different keyboard layouts.

- ***Keyboard standards:*** The Bureau of Indian Standards has developed and published the Indian Standard- 'Enhanced Inscript keyboard layouts'- IS 16350:2016. It specifies the character codes set for representing Indian Languages and their required scripts on digital medium. The Enhanced Inscript key board layouts' provides covers Character, code charts, character names for 11 Indian scripts which cover 21 official languages of India. Along with that, the standard provides more enhanced versions of the Inscript keyboard layouts and mapping of the code points with the respective keyboard layouts [22].

- ***Robust Document Analysis & Recognition System Developments:*** Optical Character Recognition is a utility tool for digitizing the content and is essential for the creation of knowledge networks such as online language tools, digital libraries. There is facility in OCR technology to scan, read, classify and store the printed text. The basic aspects of OCR technology are to scan, to recognize, classify and to read language text.

- ***Development of Online Handwriting Recognition System (OHWR):*** Online Handwriting Recognition System is a tool which is useful to convert the written piece of an individual text into editable text.

- ***IndiX:*** it is an easily accessible Graphical User Interface of Linux Operating System in Indian languages. IndiX has Multilingual and Multimodal interface which can interface, access and communicate through any native languages.

2. Corpora Development:

- The corpus plays a vital role in language computing. The Aligned corpora provides the basis for extraction of various linguistic resources, which is inevitable for building translation memory, Cross Language Information Retrieval (CLIR), Information Retrieval/ Information Extraction systems, Terminology extraction, and etc. [5].
- C-DAC has developed speech corpora along with text for three Indian Languages viz. Bangla, Assamese and Manipuri. The above mentioned corpus contains Parts-Of-Speech (POS) annotation for the text and phoneme level annotation for the speech. Annotations may include structural mark-up, part-of-speech tagging (POS Tagging), parsing, and numerous other representations. An example of annotating a corpus is part-of-speech tagging or POS-tagging, in which information about each word's part of speech is included or added in the corpus in the form of tags [13].

- Part-of-speech (or POS) tagged: The most basic way of classifying a word is according to its part of speech. The basic parts of speech are nouns, verbs, adjectives and adverbs. POS tagged corpora will be developed in a bootstrapping manner. It is possible to manually tag some amount of text [18]. A POS tagger uses learning techniques learned from the already tagged data. After training, the tool will automatically tag another set of the raw corpus. Automatically tagged corpus can be checked

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

manually which will be used as additional training data for enhancing the performance of the tool.

- Corpora building developed by Central Institute of Indian Languages Linguistics Data Consortium- in Indian Languages Mysuru (CIIL-LDC-IL) [5]. The attempt indulges in Dialect mapping, Corpora creation and analysis of above with the help of theoretically motivated model or dataset.

- Handwriting Interface to Computers: The hurdle while dealing with Indian languages is that they are not suitable for keyboard-based entry because of its complex nature. Replacing the keyboard with a simpler and more natural interface based on handwriting would bring revolution in the world of technology and computers become more accessible to the lay man and educators. Imagine that the keyboard is replaced with a special writing pad for handwriting input. HWR technologies convert the manual writings into system readable form. Now HWR has to strive to get access for numerals, punctuation and editing gestures and functionally replace the keyboard [26].

3. Multilingual E-content generation:

- ***Multilingual Multimedia Content Development:*** The human communication is a blend of verbal and non-verbal (gestures) languages. The media mode of communication is also not an exception. Digital texts are combined with pictures and sounds. The movies also contain language in spoken and written form. The speech and text technologies overlapping and interacting with various technologies facilitate processing of multi- modal communication and multimedia documents [6].

- ***Web Based Education Systems***, On-line Tutoring system for Computer Aided Language Teaching (CALT) and Computer Aided Language Learning (CALL) for Indian Languages.

- ***Multilingual Digital Libraries for Education Purpose:*** A wealth of literature and other education material in Indian languages is stored in books, which require careful storage and are subject to physical decay. Online books on the other hand have no such problems; it can be made available to students all over in their schools, homes or hostels over the Internet. The creation of digital library can open a new window to the learner. The advantage of such digital library is that it can be accessed using the native languages or regional languages. It is possible to access the information in spoken (using Speech Recognition) or written (using Online HWR) form of queries. Results can be viewed on screen, and also read out using Text-to-Speech conversion. In addition, an annotation system will allow students to do annotations for their own sake in the book. This solution can be used by individual libraries to create district, state or national level online educational resources [26].

- ***Automatic Forms processing/ Educational Testing:*** Millions of application forms filled every year in Indian languages especially in the education sector. There is a clear need for tools to read out the manual entries from scanned images of forms and applications. As a result of a growing school-going population, manual evaluation of answer papers has become very difficult. The use of Offline HWR technology is a solution for automatically assessing at least the fill-in-the-blanks style of questions with few options [26].

4. Language Tools and Devices

- C-DAC has rendered enormous contribution in forming tools and devices for Indian languages. They have ***True Type Fonts (TTFs)*** and Open Font Format suitable for various languages of India. Open Type Fonts are available in various scripts of 22 scheduled languages. The amazing efforts of C-DAC are visible from the statistical statement that, they have 8000 fonts (consisting of True Type, Open Type and Bitmap) in the credit [15].

- ***Optical Character Recognition*** renders a vital part in NLP, it converts the scanned images of books, magazines, and news papers into machine-readable text. The research in C-DAC Gist Lab's attempts to develop an Optical Character Recognition engine, which can assure great amount of accuracy in converting Indian language images to text. The basic OCR for Devanagari script named 'Chitrankan' can be found in its product portfolio [17]. The complexities of Indian scripts like their cursive nature or the features like conjuncts and joint-characters (especially in the scripts like Devanagari, Gujarati, Bengali) make it difficult for the system to recognize

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

and segment the units. Along with that there is one more practical difficulty to handle the issue; text is printed in different font of various sizes. The paper quality, scanning resolution, images in texts and other related aspects turn image processing a challenging job.

- ***Handwriting Tutor:*** The solution described above can also adopted to improve writing skills of school children, increase literacy as part of adult education programs, or allow literate adults to learn new scripts [26].

- ***Bharatavani Multilingual Portal:*** The project opened with an aim of multimedia mode (i.e., text, audio, video, images) of delivering knowledge in and about all the languages in India through a portal /website. The portal offers very inclusive, interactive, dynamic and moderate way. In the era of Indian digitalization the idea is to make India an Open Knowledge Society. Bharatavani acts as a platform to showcase all the available and updated IT tools for Indian Languages. The portal coordinate with the Ministry of Communication and IT (MCIT), TDIL make such tools possible for Ministry. The portal offers access for language tools such as fonts, software, typing tools, mobile apps, multi language translation tools, text to speech, speech to text etc. [27].

- ***Unicode Converter:*** This tool will convert English characters or words to html UTF-8 characters. Unicode Transformation Format is abbreviated as UTF-8, it signifies 8-bit blocks to represent a character. UTF -8 is a Unicode standard format which is used to support multilingual environments web site that means it allows you to design local language website thereby your web site can reach local language speaking people. Advantage of utf-8 encoding is that it is a Unicode consortium standard so it supports all operating system and modern browsers to convert into Unicode characters [24].

5. Text Processing

- NLP based ***Information Extraction and Language Processing Retrieval System:*** It is capable of retrieving information based on a combination of keywords and conceptual matching which uses domain defined Ontology. It is a knowledge based Information Retrieval system.

- ***Cross Language Information Retrieval:*** irrespective of language and regional barrier it tries to make domain information accessible to end user One can ask a query in source language and can access documents available in the language of the query as well as the target language by using a Machine Translation System e.g. MANTRA [7].

- ***Consortium Project on Cross Lingual Information Access (CLIA):*** It allows Service user queries in regional languages like Tamil, Bengali, Punjabi, Hindi, Marathi and Telugu languages to get Information Retrieval (IR) in Hindi and English and displaying content in the given query language.

- ***The Cross Language Information Access Portal for Indian Languages:*** It is an initiative of a group of academic team, research institutions and industry partners. The initiative brought a drastic change in the way to communicate with the system. The possibility to make query and get reply in any Indian language and English. The portal has given great importance to Hindi and English. The other interesting feature of the portal is the facility for snippet and summary translation. The languages accesses with this facility are Assamese, Tamil, Bengali, Hindi, Marathi, Punjabi, Telugu, Gujarati and Oriya.

- ***C-DAC Machine Aided Translation:*** C-DAC team for Translation is equipped with kit of tools to translate from English to major and minor Indian languages including Hindi, Telugu, Assamese, Malayalam, Bangla, Nepali, Punjabi, and Urdu. AnglaBharati, Mantra and MaTra are some of the acclaimed translation tools developed by C-DAC. Moreover the C-DAC team also successed in the creation of tool to translate among the cognate languages likes Urdu, Hindi and Punjabi. Some of these are developed as part of multi Institution consortium projects. These systems differ each other on the basis of their underlying approach to translation, use of language pairs and domains supported. Most of the systems follow machine aided approach due to the complexity of automated translation [7].

- ***Question Answering System:*** Question Answering (QA) System makes use of previously congregated

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

information to give answer to the questions. The QA is very useful tool to cater relevant information from the processed information along with that it is useful to develop insights and then provide the relevant answers. It would continuously improve its expertise to provide relevant answers.

- ***NLP Based Information Extraction and Retrieval:*** it accomplish the need of users to easily access relevant information. Information Extraction and Retrieval is a suitable tool for any kind of users like one having formal computer background or a layman. Applied Artificial Intelligence Group focuses on solving research problems in the areas of IR & IE, using Natural Language Processing based Semantic Search technique or algorithms for Data and text mining applications. To facilitate efficient and effective access of relevant information from the unstructured information or semi structured information sources. The process can make information structured for the benefit of end users. The other remarkable achievement of the group is Cross Lingual Information Retrieval (CLIR) which removes the barriers of language and makes information accessible to all users irrespective of language and region.

- ***Social Sentiment Analysis:*** this venture of C-DAC is commendable. The team is striving to develop an automated system to monitor, mine and analyze the citizen sentiments and opinions about social media, forums, blogs, e-commerce websites etc., which can be further used by Government agencies to make and implement policies. It can dig out the real picture of citizen's mindset and attitude. Such systems can improve the quality and quantity of public feedback. It will turn media as a channel to communicate with and about government.

- ***Consortium Project on Building Indo Wordnet:*** Hindi Wordnet is already available for public to download and use. Marathi Wordnet is on the way of completion. The Wordnet building efforts like;

- ***North-East WorldNet:*** incorporating languages like Assamese, Bodo, Manipuri and Nepali etc. Dravidnet: Kannad, Malayalam, Tamil, Telugu and Indradhanush; Konkani, Bangla, Punjabi, Guajarati, and Urdu are on the path of fulfillment. And all these words are incorporated together with English and Hindi leading to the formation of Indo Wordnet.

- ***A Text To Speech (TTS):*** is a form of speech synthesis that converts digital input text into spoken voice output. The TTS System has multi- mode of applications. TTS Software equipped with Screen Reader to help the visually challenged people to read the text on the computer screen and they would be able to perform the computer operations. Text to Speech screen reader application, for both Windows (NVDA) and Linux (OCRA) based operating systems available in six Indian Languages like Hindi, Marathi, Bengali, Tamil, Telugu and Malayalam. Text to speech as a screen reader application works with different editors like MS -Word, Notepad, Word Pad but it does not support PDF. And currently the software is monolingual. TTS system is independent of any font. TTS system simply readout what is written in the text editor and it does not support proof reading [21].

6. Speech Processing

- ***Audio-Video Search:*** The system aims to focus on extraction and retrieval of information from audio and video sources, not just by their meta- data, but by performing search on its content [7]. The audio and video files also contain multiple types of visual, text, and audio information which are not easy to extract like textual information. The major process involves in extracting or accessing information from such sources are automatically recognized Speech Transcripts, classifying audio and video date, image similarity matching etc. The application works on the transcribed text of the audio/video files. Transcription can also be achieved using Speech recognition engine, Automatic Speech Recognition Technique and speaker recognition, etc. Its key traits are: Storage of Audio and Video data, IProcessing of Audio & Video Data: Audio or Video Search through NLP based Retrieval Engine or search engine [15].

- ***Universal Speech To Speech Translation:*** Speech-to- Speech translation system is an outcome of the effort of International Consortium and Universal Speech Translation Advanced Research (U-STAR) going among the various Languages of the world.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

- ***Speech-To-Text (STS):*** Continuous speech recognition is the next task that Indian Language's STS systems have to attain now. The reported systems are under development. IBM Research Laboratory's (IRL), Indian Languages speech recognition system is based on HMM-based acoustic recognizer. It uses a trigram language model. The aim is to make desktop as well as telephone-based recognition systems, with preferably 90% plus accuracy. The systems are basically adaptation of IBM's through Voice recognition system for Indian languages. The biggest challenge here may be to map English phonemes into Indian languages, as the later has many phonemes not known to the former [23].

- Dhvani is a Text To Speech (TTS) system designed for Indian Languages. The aim of the project is to ensure that knowledge of English is not necessary for using Computer. We hope that it will reduce the digital divide and will be helpful for the visually challenged users as a screen reader in their mother tongue. Currently Dhvani is capable of generating intelligible speech for the following Languages such as Bengali, Hindi, Kannada, Marathi, Malayalam Oriya, Panjabi, Gujarati, Tamil and Telugu.

- Linguistic Annotated Speech Corpora for Speech processing in Indian languages [5]. The annotated corpus is more useful for doing various Linguistic researches and analysis. Parts-of-speech tagging is a way of annotating Corpus in which information about each word's part of speech like verb, noun, adjective, adverb, etc, is added to the corpus in the form of tags [13].

- Speech/ Speaker recognition systems of Indian Languages: Speech recognition system is a mode of inter-communication between human and machine. Automatic Speech Recognition (ASR) is an advance way to operate computer through speech without much efforts. It offers user friendly interface. The thought often haunt the people with less educational background and physical disabilities how to get familiarize with computer in the technology ruling era. Here Speech recognition system acts as a helping tool. The possibility having Speech recognition in native languages provide convenient and comfortable environment to the user. The factors like noisy environment, different grammar rules, dialect variations, varying pronunciations of speakers make the performance of Speech Recognition System a challenging one. [20].

- Voice User Interfaces for IT applications and services have become more and more common for languages like English, and are valued for their ease of access, especially in telephony-based applications. The relevance to develop such tool in Indian context become vivid with the help of following facts: less knowledge of English among a major share of population, varying literary rates among the states and linguistic diversity. All this hinders the access and use of computers among the masses. The availability of the above mentioned technology can solve the problem to some extent. The execution of such tool will benefit heavily to the common mass dwelling in semi-urban and rural areas. However, to make it a reality, system must be capable to comprehend speech input in the user's language/ native languages and provide speech output. In India speech technology is packed with the possibility of translation among various Indian languages, therefore services and information can be provided across languages more easily [25].

V. FUTURE OF NLP

When the Natural Language Understanding or readability improves, computers or devices will get access of the knowledge of various human languages therefore they can apply what they gained or learned from actual world. If NLP is supplemented with Natural Language Generation computers will become more and more competent in receiving, understanding, providing useful and valuable information and data. The NLP's current applications are needed to redefine in the coming future as it faces new technological challenges and a push from the market to create more user-friendly systems [4]. NLP plays a crucial

part in the design and development of successful Web portals. Multi lingual Web portal services interface are becoming increasingly user-friendly. As the universal platform of the Web and the user for portals broadens, the search tool must be appealing to all types of users. And most interestingly searching need not require an education in SQL, Boolean logic, lexical analysis, or the underlying structures of information repositories [12].

VI. CONCLUSION

The natural language systems are still very complicated to design even though the NLP research and development has been on working mode for more than sixty years. Multitude of models and algorithms exist today, even the NLP systems are not attained its full perfection because of the complexity of natural languages. And it is not easy as we think to capture the entire linguistics knowledge with hundred percent accuracy in processing. The upcoming NLP work has to emphasis on the analysis and development of specific NLP tools such as efficient text to speech tools, speech to text tool, tools for corpora analysis, tools for OCR, or machine translation tool and tool for automatic analyzing text and speech etc.

REFERENCES

[1] Bharati, Akshar., Chaitanya, Vineet, and Sangal, Rajeev. Natural Language Processing A Paninian Perspective. New Delhi: Prentice-Hall of India. Print.
[2] Bose Ranjit. Natural Language Processing: Current state and future directions. International Journal of the Computer, the Internet and Management Vol. 121. pp 1– 11. (January–April, 2004). Print.
[3] Deshpande, M., Avanti. A Survey: Structure of Machine Readable Dictionary. International Journal of Engineering and Innovative Technology (IJEIT), Volume 1, Issue 4, ISSN: 2277-3754. April 2012. Print.
[4] Hirsehberg, Julia., and D., Christopher. Review Advances in natural language processing. http://science.sciencemag. org.Vol 349 Issue 6245. 17 July 2015. Web.
[5] http://www.ldcil.org/areasOfWorkSpeech.aspx. Web.
[6] http://www.dfki.de/hansu/HLT-Survey.pdf. Web.
[7] https://cdac.in/index.aspx?id=mlingual. Web.
[8] http://bosslinux.in. Web.
[9] http://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm. Web.
[10] Kaur, Gaganpreet. Usage of Regular Expressions in NLP. IJRET: International Journal of Research in Engineering and Technology, eISSN: 2319-1163, pISSN: 2321-308. Web.
[11] KaurSidhu, Gurleen., and Kaur, Navjot. Role of Machine Translation and Word Sense Disambiguation in Natural Language Processing. International Journal of computer science and communication Engineering IJCSCE Special issue on “Recent Advances in Engineering & Technology, ISSN 2319-7080, NCRAET- 2013. Print.
[12] Kurian, Cini. A review on the progress of natural language processing in India. International Journal of Advances in Engineering & Technology (IJAET). ISSN: 22311963. Nov., 2014. Print.
[13] M.S., Bindu. and Idicula Mary Sumam. Named Entity Identifier for Malayalam Using Linguistic Principles Employing Statistical Methods. International Journal of Computer Science Issues, Vol.8, Issue 5, No 3, ISSN (Online):1694-0814. Web. September 2011. Print.
[14] Reddy V., Mallamma and Hanumanthappa, M., Dr. Natural Language Processing: A Statistical Machine Translation Approach. Pune: International Conference on Computer Science & Engineering (ICCSE), ISBN: 978-93-82208-74-7,17th March 2013. Print.
[15] S., Vel, Sakthi. Rule-based Model for Sentence level Automatic segmentation of continuous speech of Tamil using Supra-segmental Features. Unpublished M.Phil. Dissertation. April 2014. Print.
[16] Proctor, Francis. Encyclopedia of Advances in Corpus Linguistics History, Methods and

Analysis (Volume 1). Nyx Academics LLC. ISBN: 978-1-62158-222-9.2012. Print.
[17] https://cdac.in/index.aspx?id=mlc_gist_ocr.Web.
[18] http://www.ldcil.org/areasOfWorkCorpCreat.aspx.Web.
[19] http://www.ldcil.org/areasOfWorkCorpCreat.aspx. Web.
[20] https://pdfs.semanticscholar.org/c8b7/321eecdfff88cb8cd965db75f84bee91ca4.pdf.Web.
[21] http://www.tdil.meity.gov.in/pdf/Vishwabharat/41/8.pdf
[22] http://tdil-dc.in/index.php?option=com_vertical&parentid=1&Itemid=488&lang=en. Web.
[23] http://www.tdil.meity.gov.in/pdf/Vishwabharat/16/3.2.pdf.
[24] http://www.texttranslater.com/html-unicode-converter/index.php Web.
[25] http://www.ldcil.org/areasOfWorkSpeech.aspx Web.
[26] http://www.ldcil.org/areasIOfWorkCharRec.aspx Web.
[27] http://english.bharatavani.in/. Web.