Acquaintance with Natural Language Processing for Building Smart Society

Ravindra Parshuram Bachate[1*], *Ashok Sharma*[1]
[1]School of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab, India.

Abstract. Natural Language Processing (NLP) deals with the spoken languages by using computer and Artificial Intelligence. As people from different regional areas using different digital platforms and expressing their views in their spoken language, it is now must to focus on working spoken languages in India to make our society smart and digital. NLP research grown tremendously in last decade which results in Siri, Google Assistant, Alexa, Cortona and many more automatic speech recognitions and understanding systems (ASR). Natural Language Processing can be understood by classifying it into Natural Language Generation and Natural Language Understanding. NLP is widely used in various domain such as Health Care, Chatbot, ASR building, HR, Sentiment analysis etc.

1. Introduction

Languages which are spoken by human being called as a Natural Languages. There are many resources which are generating data in natural languages on daily basis. For example, Facebook, twitter, Instagram, blogs etc. are generating a big data which is very much difficult to process with traditional approaches. Today's world is data driven and we must work on techniques which will improve the living standard of human being. To achieve this, Natural Language Processing (NLP) plays a vital role because we will deal with user's expressions and feelings expressed on various platforms. A lot of research has been done on Natural Language Processing (NLP) but it is limited to some widely used languages only e.g. English. If we think about the India, around 70% of Indian population lives in rural area where they are not having good understanding of English language. If we want to make their life better using NLP, we need to work on regional languages like Marathi, Hindi, Punjabi and so on. In India there are 22 official languages are spoken and more than 1000 dialects are spoken [1]. These regional and dialect languages belongs to different language families like Indo-Aryan, Dravidian, Austric, Tibeto-Burman and other [1].

To work with these family of languages, we need to understand the representation of it. Natural Language can be described in two parts – Natural Language Processing and Natural Language Understanding. Natural Language Processing deals with many tasks such as Named Entity Recognition (NER), part of speech tagging (POS), text categorization, syntactic parsing, conference resolution, machine translation and question answering. Whereas Natural Language Understanding deals with relation extraction, semantic parsing, question and answering, sentiment analysis, summarization, paraphrase and natural language interface and dialogue agents. While building any ASR system or text to speech system or speech to text system, NLP and NLU is used.

2. Historical Review of Natural Language Processing

Historical review of natural language processing will help us to understand journey of progress of natural language processing in last 10 decades. In 1906, Professor Saussure thought about natural language as a science and since then people started doing research on it. He did a research about natural language and delivered lectures on it during 1906 to 1911. After his death, his colleagues Albert Sechehaye and Cherles bally written a paper entitled "Cours de Linguistique" in 1916. Alan Turing in 1950 written a revolutionary paper on Thinking Machine. In 1957, Noam Chomsky written a book on Syntactic structure[2]. In 1964, US established "Automatic Language Processing Advisory Committee (ALPAC)" and initiated research project in 1966 in collaboration with National Research Council on NLP and AI. Until 1980s, people were using handwritten rules for processing natural language, but it is difficult to write handwritten rules every time. To deal with this problem, statistical models were introduced in 1990s. N-gram became very powerful due to its use in recognizing and tracking clumps of linguistic data.

LSTM Recurrent Neural Network model introduced in 1997 and wide use of this algorithm for voice and text processing started around 2007. In 2001, first feedforward neural network was proposed by Yoshio Bengio. Apple

*Corresponding author: **bachateravi@gmail.com**

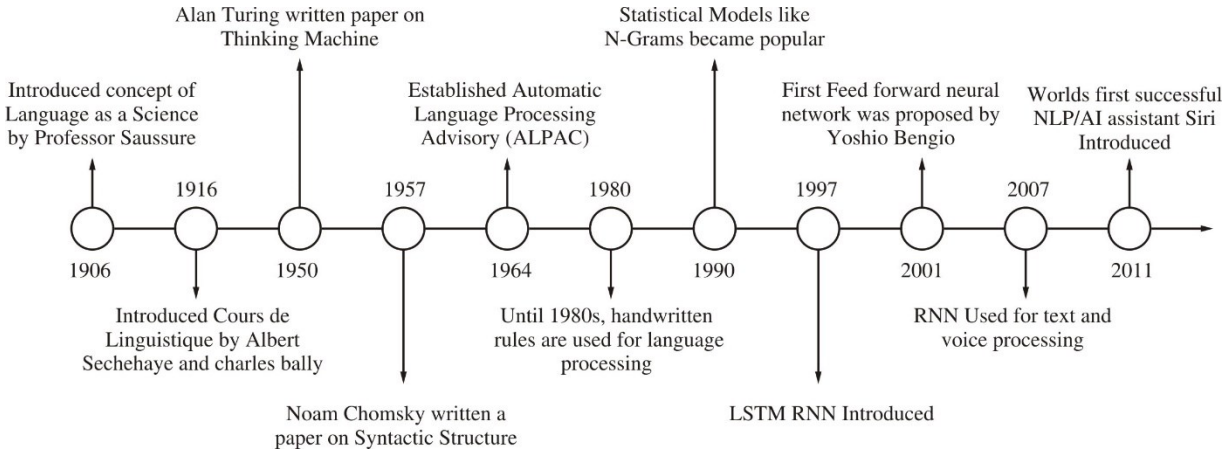

Fig.1: NLP History

introduced Siri which became world's first NLP/AI based assistant. Later other companies also introduced assistant application such as Googles Assistant, Amazon's Alexa, Microsoft's Cortona etc. Artificial Intelligence and Machine learning branches adds significant contribution in expanding application area, effectiveness and efficiency of natural language processing for the society. Automatic Speech Recognition and Understanding is a branch of Natural Language Processing which is widely used in applications such as Siri, Alexa etc.

3. NLP Architecture

Understanding Natural Language is not easy due to various kind of ambiguity exists while processing natural language as many factors influence on it such as gender, surrounding, voice clarity, mic used etc. Natural Language Processing is carried out in various phases or tasks which are described in Figure 2. Each phase having its own problems occurred while processing natural languages due to ambiguity present in input.

3.1 Lexical Analysis / Phonetical Analysis

Input for this phase can be text or voice. Based on type of input, respective analysis will be done in this phase. Text is taken as an input for lexical analysis and it generates tokens for the next phase. For phonetic analysis, sound is taken as an input and it classifies sound based on its physical properties which considers transmission, articulation and reception of speech sounds. Phonetic analysis is required to convert speech into text. Phonetic variation is a big challenge as it varies as user changes[3].

3.2 Morphological Analysis

In morphological analysis, basically it analyses both words and non-words. After analysing, it is segregated into two parts i.e. words and non-words such as semi colon. Morphology is used to describe the relationship between word forms and lexical forms[4].

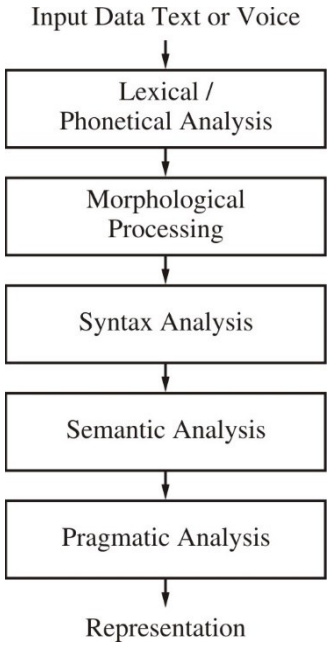

Fig. 2: NLP Architecture

3.3 Syntax Analysis

Syntax analysis involves two tasks-grammar checking and parsing input[4].The first task i.e. grammar checking involves checking correctness of input and second task deals with parsing input statement and arranging their relationship among the word. Parser generates three kind of structure.

3.4 Semantic Analysis

Semantic analysis deals with finding the meaning of statements. It helps to understand the meaning of natural language so that it can be used for other domains such as Artificial Intelligence, Machine Learning, Data Analysis, Speech Understanding etc. Same statement can be interpreted more than one meaning while doing semantic analysis.

3.5 Pragmatic Analysis

The same statement may have more than one meaning found in semantic analysis while interpreting its meaning. Now it is difficult to choose which meaning is applicable into respective context of that statement. The last phase of Natural Language Processing (NLP) chose the correct meaning of the statement with respect to context of use.

4. NLP for Smart Society

Due to NLP's strong ability of making machines to work independently, it is widely used in all possible areas or domains in the world. The applications of NLP are shown in Figure 3.

4.1 Question & Answering

Question and Answering is one of best application of Natural Language Processing. It will be greatly useful for smart societies to have such system to whom we can ask any question and get a answer in fraction of second. Applications such as Siri, Google Assistant, Cortana and Alexa are the best examples of developing QA system. Developing QA framework needs intense NLP work to be done to achieve the greater accuracy[5]. Q & A Systems belongs to AI and NLP domains. Such systems are designed to answer humans' questions by machine. It is useful for building Human Computer Interaction Systems.

4.2 Automatic Speech Recognition

Building Automatic Speech Recognition (ASR) Systems requires natural language processing. To analyses and understand the speech is a challenging task. To deal with this, natural language processing is used in ASR systems. Applications like Siri uses ASR for recognizing the speech of user while developing Human Machine Interaction Systems[6]. ASR systems are used in various products where commands to that product is supposed to given by spoken language such as operating music systems, smart home systems, handling different car operations etc.

4.3 Machine Translation

Machine translation is an application of NLP which translates one language text into another language e.g. Marathi to English and vice versa by keeping meaning of statements same[7]. Basically, there are two types of Machine Translations – Bilingual and Multilingual. Bilingual Machine Translation systems translates between two languages only whereas Multilingual systems translates text into more than one language. Such translations help people of different regions who speak different languages but still wants to communicate with each other.

4.4 Chatbot

A chatbot can be defined as a software program implemented using Artificial Intelligence and Natural Language Processing intended to interact with human like a human. Chatbot gives humans answer on chat and user doesn't feels like that it is a computer program. Many domains including banking, e-learning, telephone uses chatbot to give 24 * 7 support to their customer. Home automation can be done using IoT and Chatbot such as Amazon Alexa, Mi which internally uses Natural Language Processing[8].

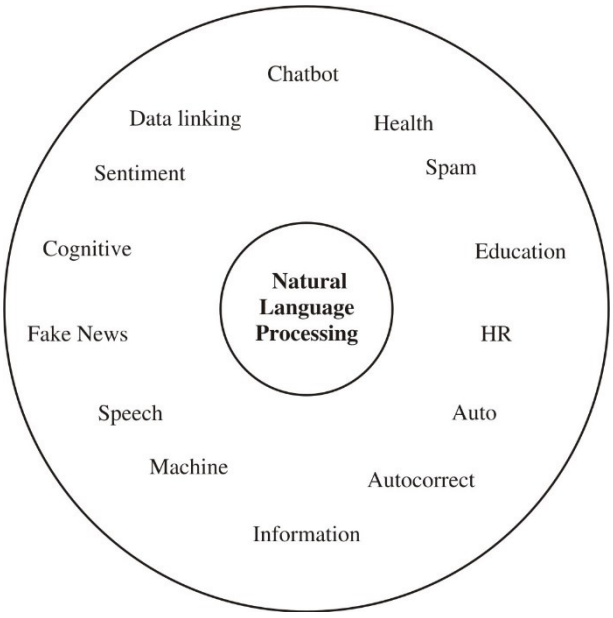

Fig. 3: NLP Applications

4.5 Education

Natural Language Processing can bring a very positive change in education system[9]. As knowledge in the world is exist in various languages, NLP can make that knowledge available and understandable to students with the help of NLP's different applications. To improve e-learning systems in education, NLP will help a lot as knowledge will be available to seekers in their own spoken language. And it is proved that seekers can understand and remember the things, if they learn it in their own language.

4.6 Fake News Detection

Due to tremendous growth of using social media, reaching to people gets more easier than ever. It has both pros and cons too. People having negative mindset posts fake news using social media platforms such as Facebook, twitter, Instagram etc which spreads hatred[10]. Finding fake news on social media is a priority task for social media companies as they are partially responsible for it. So fake news detection is one of the applications of NLP which is used in all social media companies to detect and remove fake news from their platforms.

4.7 Sentiment Analysis

As name suggest, this is used to identify the sentiment of people on various platforms. To improve any business or wants to understand users view about the product or system, Sentimental analysis plays a vital role. Natural language processing is used to understands the reviews of user and then sentiments of user are identified. Sentimental analysis is used to judge whether people liked movie or not. Also, it can be used for identifying satisfaction level of users.

4.8 Health Care

NLP can be widely used in health care sector. To learn adverse effect of drugs or to understand the purposes of drug Natural language processing can be used[11].NLP also can be used for various purposes in health care domain such as clinical trial matching, clinical documentation, clinical decision support, to develop computer assistance system for patients etc.

4.9 Spam Detection

Now days, people are getting many spam mails in their mail box which distracts the user form important mails and sometimes users get trapped in fraudulent activities. Spam mails can be defined as a bulk mail sent to large number of audiences in an intention to do the marketing of their products or to do fraud by collecting users' details[12]. So, it is a big challenge for all mail service providers to protect their users from such spam mails. By implementing natural language processing in mail program, mail service provider can find and segregate spam mails from important mails.

4.10 Text Auto Correction

Natural language processing has another application which corrects the spelling if typed wrong. It helps people to write the content without any fear and tension as it reduces the risks of doing spelling mistakes. It improves the correctness of document. NLP also can be used for auto completion of statement based on the analysis of previous typing history. It reduces cognitive load and typing work of user.

5. NLP Tools

There are many NLP Tools available for its implementation. These tools are listed technology wise in figure 4. Main three languages considered here to make a list of NLP tools i.e. Python, Java and Node which are most popular and widely accepted programming languages in the world. Python includes NLP tools such as NLTK, Spacy, TextBlob, PyTorch, Texacy. Java has CoreNLP, StanfordNLP, OpenNLP and CogCompNLP whereas Node has Compromise, Nlp.js, Retext and Natural NLP tools.

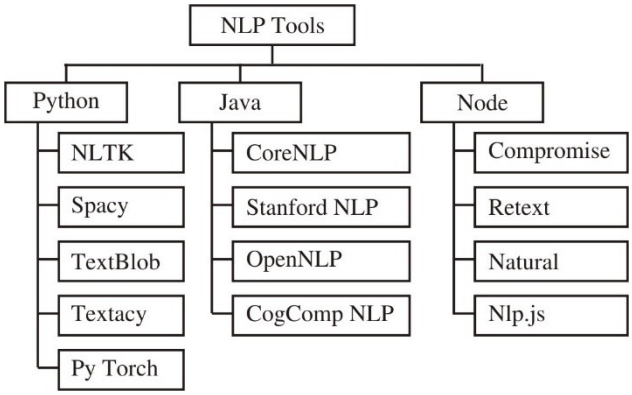

Fig. 4: NLP Tools

6. Conclusion

This paper discusses the Natural Language Processing in all possible aspects of it for understanding its use for making our society smart. Natural Language Processing research work and its implementation for regional languages which are spoken in respective part of India will help to improve the different domains associated with respected area. The various applications of NLP will help to build smart society or city with its effective and efficient use while developing products for it.

References

1. MHRD, Indian Linguistic. (1995).
2. N. Chomsky, *Syntactic Structures.* PARIS: MOUTON PUBLISHERS, (1957).

3. D. Williams and P. Escudero, “Detecting phonetic variation versus phonemic differences,” pp. 265–269.
4. A. Reshamwala, D. Mishra, and P. Pawar, “Review on natural language processing,” no. February, (2013).
5. G. Bouma, I. Fahmi, J. Mur, and G. Van Noord, “Linguistic knowledge and question answering,” vol. **46**, pp. 15–40, (2005).
6. R. Weiscbedel, J. Carbonell, W. Lehnert, M. Marcus, and R. Perrault, “White Paper on Natural Language Processing,” pp. 481–493.
7. M. T. Makwana and D. C. Vegda, “Survey: Natural Language Parsing For Indian Languages,” (2015).
8. C. J. Baby, F. A. Khan, and J. N. Swathi, “Home automation using IoT and a chatbot using natural language processing,” 2017 Innov. Power Adv. Comput. Technol. i-PACT 2017, vol. 2017-January, pp. 1–6, (2018).
9. K. M. Alhawiti, “Natural Language Processing and its Use in Education,” vol. **5**, no. 12, pp. 72–76, (2014).
10. F. Cardoso, A. Cristina, and B. Garcia, “Can Machines Learn to Detect Fake News ? A Survey Focused on Social Media,” vol. **6**, pp. 2763–2770, (2019).
11. T. Ly et al., “Evaluation of Natural Language Processing (NLP) systems to annotate drug product labeling with MedDRA terminology,” J. Biomed. Inform., vol. **83**, no. April, pp. 73–86, (2018).
12. D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural Language Processing: State of The Art, Current Trends and Challenges,” no. August (2017).