# A Dive into Natural Language Processing (NLP)

Moha
Assistant Professor
KCC Institute of Legal and Higher Education

***Abstract*- Natural Language Processing (NLP) is the field of study that focuses on interactions between the human language and the computers. It sits at the interaction of computational linguistics, computer science and artificial intelligence. NLP is gaining large attention nowadays for analysing and understanding human language. By using NLP, developers can structure and organize the knowledge to perform various task in varied fields like named entity recognition, machine translation, chatbot, speech recognition, automatic summarization, sentiment analysis, recommendation software etc. In this paper, a detailed introduction of NLP will be explained along with the applications of NLP. Also, a natural language toolkit (NLTK) will be discussed which is used for implementing the various tasks.**

***Keywords- NLP, machine translation, chatbot, NLTK, phonology***

## I. INTRODUCTION

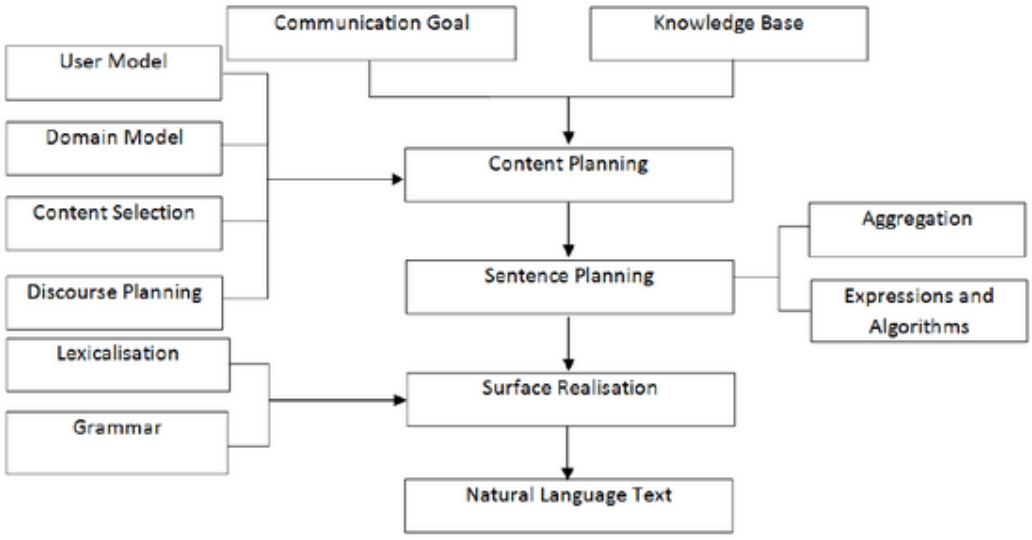
NLP is a subfield of computer science and artificial intelligence concerned with interactions between computers and human (natural) languages. It is used to apply machine learning algorithms to text and speech. NLP combines the power of linguistics and computer science to study the rules and structure of language, and create intelligent systems (run on machine learning and NLP algorithms) capable of understanding, analyzing, and extracting meaning from text and speech. For example, we can use NLP to create systems like speech recognition**,** document summarization**,** machine translation**,** spam detection**,** named entity recognition**,** question answering, autocomplete, predictive typing and so on. Nowadays, most of us have smartphones that have speech recognition. These smartphones use NLP to understand what is said. Also, many people use laptops which operating system has a built-in speech recognition.

NLP is used to understand the structure and meaning of human language by analyzing different aspects like syntax, semantics, pragmatics, and morphology. Then, computer science transforms this linguistic knowledge into rule-based, machine learning algorithms that can solve specific problems and perform desired tasks. Take Gmail, for example. Emails are automatically categorized as *Promotions*, *Social*, *Primary*, or *Spam*, thanks to an NLP task called keyword extraction. By "reading" words in subject lines and associating them with predetermined tags, machines automatically learn which category to assign emails. NLP is used to analyze text, allowing machines to understand how humans speak. This human-computer interaction enables real-world applications like automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts-of-speech tagging, relationship extraction, stemming, and more. NLP is commonly used for text mining, machine translation, and automated question answering. NLP is characterized as a difficult problem in computer science. Human language is rarely precise, or plainly spoken. To understand human language is to understand not only the words, but the concepts and how they're linked together to create meaning. Despite language being one of the easiest things for the human mind to learn, the ambiguity of language is what makes natural language processing a difficult problem for computers to master. NLP algorithms have a variety of uses. Basically, they allow developers and businesses to create a software that understands human language. Due to the complicated nature of human language, NLP can be difficult to learn and implement correctly. However, with the knowledge gained from this article, you will be better equipped to use NLP successfully, no matter your use case. Some of the examples of natural language processing are-

NLP algorithms are typically based on machine learning algorithms. Instead of hand-coding large sets of rules, NLP can rely on machine learning to automatically learn these rules by analyzing a set of examples (i.e. a large corpus, like a book, down to a collection of sentences), and making a statistical inference. In general, the more data analyzed, the more accurate the model will be. Cortana- The Microsoft OS has a virtual assistant called Cortana that can recognize a natural voice. You can use it to set up reminders, open apps, send emails, play games, track flights and packages, check the weather and so on. Siri- Siri is a virtual assistant of the Apple Inc.'s iOS, watchOS, macOS, HomePod, and tvOS operating systems. Again, you can do a lot of things with voice commands: start a call, text someone, send an email, set a timer, take a picture, open an app, set an alarm, use navigation and so on. Gmail- The famous email service Gmail developed by Google is using spam detection to filter out some spam emails.

## II. LEVELS OF NLP

The 'levels of language' are one of the most explanatory method for representing the Natural Language processing which helps to generate the NLP text by realising Content Planning, Sentence Planning and Surface Realization phases Linguistic is the science which involves meaning of language, language context and various forms of the language. The various important terminologies of Natural Language Processing are: -




1. Phonology

Phonology is the part of Linguistics which refers to the systematic arrangement of sound. The term phonology comes from Ancient Greek and the term phono- which means voice or sound, and the suffix logy refers to word or speech. In 1993 Nikolai Trubetzkoy stated that Phonology is “the study of sound pertaining to the system of language". Whereas Lass in 1998 wrote that phonology refers broadly with the sounds of language, concerned with the two lathe sub discipline of linguistics, whereas it could be explained as, "phonology proper is concerned with the function, behaviour and organization of sounds as linguistic items. Phonology include semantic use of sound to encode meaning of any Human language. (Clark et al.,2007) [6].

2. Morphology

The different parts of the word represent the smallest units of meaning known as Morphemes. Morphology which comprise of Nature of words, are initiated by morphemes. An example of Morpheme could be, the word pre-cancellation can be morphologically scrutinized into three separate morphemes: the prefix pre, the root cancella, and the suffixation. The interpretation of morpheme stays same across all the words, just to understand the meaning humans can break any unknown word into morphemes. For example, adding the suffixed to a verb, conveys that the action of the verb took place in the past. The words that cannot be divided and have meaning by themselves are called Lexical morpheme (e.g.: table, chair) 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 occurs in combination called bound morphemes (eg. -ed, -ing) Grammatical morphemes can be divided into bound morphemes and derivational morphemes.

3. Lexical

In Lexical, humans, as well as NLP systems, interpret the meaning of individual words. Sundry types of processing bestow to word-level understanding – the first of these being a part-of-speech tag to each word. In this processing, words that can act as more than one part-of-speech are assigned the most probable part-of speech tag based on the context in which they occur. At the lexical level, Semantic representations can be replaced by the words that have one meaning. In NLP system, the nature of the representation varies according to the semantic theory deployed.

4. Syntactic

This level emphasis to scrutinize the words in a sentence so as to uncover the grammatical structure of the sentence. Both grammar and parser are required in this level. The output of this level of processing is representation of the sentence that divulge the structural dependency relationships between the words. There are various grammars that can be impeded, and which in twirl, whack the option of a parser. 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 [1]. Syntax conveys meaning in most languages because order and dependency contribute to connotation. 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.

5. Semantic

In semantic most people think that meaning is determined, however, this is not it is all the levels that bestow to meaning. Semantic processing determines the possible meanings of a sentence by pivoting on the interactions among word-level meanings in the sentence. This level of processing can incorporate 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. For example, amongst other meanings, ‘file’ as a noun can mean either a binder for gathering papers, or a tool to form one’s fingernails, or a

line of individuals in a queue (Elizabeth D. Liddy,2001) [7]. The semantic level scrutinizes words for their dictionary elucidation, but also for the elucidation they derive from the milieu of the sentence. Semantics milieu that most words have more than one elucidation but that we can spot the appropriate one by looking at the rest of the sentence. [8]

6. Discourse

While syntax and semantics travail with sentence-length units, the discourse level of NLP travail with units of text longer than a sentence i.e, it does not interpret multi sentence texts as just sequence sentences, apiece of which can be elucidated singly. Rather, discourse focuses on the properties of the text as a whole that convey meaning by making connections between component sentences (Elizabeth D. Liddy,2001) [7]. The two of the most common levels are Anaphora Resolution - Anaphora resolution is the replacing of words such as pronouns, which are semantically stranded, with the pertinent entity to which they refer. Discourse/Text Structure Recognition - Discourse/text structure recognition sway the functions of sentences in the text, which, in turn, adds to the meaningful representation of the text.

7. Pragmatic:

Pragmatic is concerned with the firm use of language in situations and utilizes nub over and above the nub of the text for understanding the goal and to explain how extra meaning is read into texts without literally being encoded in them. This requisite much world knowledge, including the understanding of intentions, plans, and goals. For example, the following two sentences need aspiration of the anaphoric term 'they', but this aspiration requires pragmatic or world knowledge (Elizabeth D. Liddy,2001) [7].

## III. INTRODUCTION TO THE NLTK LIBRARY FOR PYTHON

NLTK (Natural Language Toolkit) is a **l**eading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to many corpora and lexical resources. Also, it contains a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. Best of all, NLTK is a free, open source, community-driven project.

**Features:**

- Tokenization.
- Part Of Speech tagging (POS).
- Named Entity Recognition (NER).
- Classification.
- Sentiment analysis.
- Packages of chatbots.

## IV. CONCLUSION

NLP's role in the modern world is skyrocketing. With the volume of unstructured data being produced, it is only efficient to master this skill or at least understand it to a level so that you as a data scientist can make some sense of it. In this paper, the importance and a brief introduction of NLP is given. Natural Language Processing helps machines automatically understand and analyze huge amounts of unstructured text data, like social media comments, customer support tickets, online reviews, news reports, and more. Natural language processing tools can help machines learn to sort and route information with little to no human interaction – quickly, efficiently, accurately, and around the clock. Natural language processing algorithms can be tailored to your needs and criteria, like complex, industry-specific language – even sarcasm and misused words.

## REFERENCES

1. Chomsky, Noam, 1965, Aspects of the Theory of Syntax, Cambridge, Massachusetts: MIT Press.
2. Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe,I., Rigau, G., Soroa, A., Ploeger, T., and Bogaard, T.(2016). Building event-centric knowledge graphs from news. Web Semantics: Science, Services and Agents on the World Wide Web, In Press.
3. Shemtov, H. (1997). Ambiguity management in natural language generation. Stanford University.
4. Emele, M. C., & Dorna, M. (1998, August). Ambiguity preserving machine translation using packed representations. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1 (pp. 365-371). Association for Computational Linguistics.
5. Knight, K., & Langkilde, I. (2000, July). Preserving ambiguities in generation via automata intersection. In AAAI/IAAI (pp. 697-702).
6. Nation, K., Snowling, M. J., & Clarke, P. (2007). Dissecting the relationship between language skills and learning to read: Semantic and phonological contributions to new vocabulary learning in children with poor reading comprehension. Advances in Speech Language Pathology, 9(2), 131-139.
7. Liddy, E. D. (2001). Natural language processing.
8. Feldman, S. (1999). NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. ONLINE-WESTON THEN WILTON-, 23, 62-73.
9. "Natural Language Processing." Natural Language Processing RSS. N.p., n.d. Web. 25 Mar. 2017
10. Hutchins, W. J. (1986). Machine translation: past, present, future (p. 66). Chichester: Ellis Horwood.
11. Hutchins, W. J. (Ed.). (2000). Early years in machine translation: memoirs and biographies of pioneers (Vol. 97). John Benjamins Publishing.
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 IRE-AIEE-ACM computer conference (pp. 219-224). ACM.
13. Woods, W. A. (1978). Semantics and quantification in natural language question answering. Advances in computers, 17, 1-87.
14. Hendrix, G. G., Sacerdoti, E. D., Sagalowicz, D., & Slocum, J. (1978). Developing a natural language interface to complex data. ACM Transactions on Database Systems (TODS), 3(2), 105-147.
15. Alshawi, H. (1992). The core language engine. MIT press.