***Abstract - Natural language processing (NLP) as of late has received a ton of press for its computational portrayal and investigation of human language. It has a wide scope of utilization, including machine interpretation, email spam identification, information extraction, rundown, clinical, and question addressing, among others. The article is partitioned into four areas, starting with a conversation of various degrees of NLP and parts of Natural Language Generation (NLG), then, at that point, continuing to the set of experiences and development of NLP, the best in class, latest things and troubles and the future scope.***

***We will also be discussing the tools and techniques use in Natural Language processing and how these techniques come into action when we apply them. Comparison between techniques and how each technique works individually. Natural language processing hasn't reached perfection so far, but continuous improvements in this area can certainly touch the line of perfection. Today, various AIs use natural language processing algorithms to recognize and process voice commands from users.***

**INTRODUCTION**

Natural Language Processing refers to the area of computational linguistic which combines human language with statistical data and deep learning models. It allows computers to process human language to the full of its extent, that is, understanding the human emotions and intent behind the words, the speaker or the writer has said or written.

**Fig. 1. Broad Classification of NLP**

NLP does not treat text like mere sequence of symbols. It understands the thought process of the speaker and why such words are used in an order. It digs a little deeper than just the surface.

Natural Language Processing (NLP) is a part of AI and semantics worried about causing PCs to get explanations or words written in human dialects. Natural language processing was made to make clients' lives more straightforward and to satisfy their craving to associate with PCs in regular language. NLP obliges those clients who need more of an ideal opportunity to learn new dialects or educate them, as not all clients are knowledgeable in machine specific language.

An assortment of rules or a bunch of images can be utilized to characterize a language. Images are blended and used to send or communicate data. The rules have a domineering grasp over images. Natural Language Processing is separated into two sections: Natural Language Understanding and Natural Language Generation, which advances crafted by grasping and creating text.

Phonetics is the investigation of language, and it contains Phonology, which manages sound, Morphology, which manages word creation, Syntax, which manages sentence structure, Semantics punctuation, and Pragmatics, which manages perception.

**RELATED WORK**

Many researchers worked on NLP, building tools and systems which makes NLP what it is today. Tools like Sentiment Analyser, Parts of Speech (POS) Taggers, Chunking, Named Entity Recognitions (NER), Emotion detection, Semantic Role Labelling made NLP a good topic for research.

Jeonghee et al., 2003 [26] Sentiment analysis works by extracting emotions about a particular problem theme. Sentiment analysis consists of subject-specific feature terminology extraction, emotions extraction and association by relationship analysis.

(Namrata Tapswi, Suresh Jain ., 2012) [27] Parts of speech taggers for the languages like European languages. The most procedures for part of speech can work efficiently on European languages, but it won’t on Asian languages or middle eastern languages.

McDonald et al., 2005 Chunking – it is also known as Shadow Parsing, it works by labelling segments of sentences with syntactic correlated keywords like Noun Phrase and Verb Phrase. Every word has a unique tag often marked as Begin Chunk (B-NP) tag or Inside Chunk (I-NP) tag.

(Alan Ritter., 2011) [33] Usage of Named Entity Recognition in places such as Internet is a problem as people don’t use traditional or standard English. This degrades the performance of standard natural language processing tools substantially. By annotating the phrases or tweets and building tools trained on unlabelled, in domain and out domain data. It improves the performance as compared to standard natural language processing tools.

**LEVELS OF NLP**

The 'levels of language' are a basic method for communicating Natural Language Processing, which helps with the production of NLP message by finishing the Content Planning, Sentence Planning, and Surface Realization stages.



**Fig. 2. Phases of NLP Architecture**

Phonetics is a part of study that concentrates on the significance of language, its specific situation, and its many structures. Coming up next are a portion of the critical terms in Natural Language Processing: -

1. **Phonology**

Phonology is a part of etymology that arranges with the deliberate association of sound. Phonology comes from the Ancient Greek prefix phono-, which alludes to voice or sound, and the addition - logy, which alludes to word or discourse.

1. **Morphology**

Morphemes are the littlest units of importance addressed by the different components of the word. Morphemes are the beginning stages for morphology, which is the investigation of the idea of words. The word precancellation, for instance, might be separated morphologically into three morphemes: the prefix pre, the root cancellation, and the addition. People can part any obscure word into morphemes to appreciate the significance since the understanding of morpheme is something very similar across all words.

1. **Lexical**

People and NLP frameworks both comprehend the significance of individual words in Lexical. Word-level perception is helped by an assortment of handling procedures, the first is the expansion of a grammatical feature tag to each word. Semantic portrayals can be subbed by words with a solitary significance at the lexical level. The idea of the portrayal in an NLP framework change relying upon the semantic hypothesis utilized.

1. **Syntactic**

This level accentuates inspecting the expressions of a sentence to decide the expression's linguistic design. This level requires the utilization of both language structure and a parser. The portrayal of the expression that uncovers the underlying reliance joins between the words is the result of this degree of handling. There are assortment of sentence structures that might be discouraged, and which, thusly, confine the parser choice.

1. **Semantic**

The vast majority accept that importance is chosen in semantics, yet this isn't true. Semantic handling recognizes the different implications of an expression in view of the connections between the sentence's statement-level implications. This degree of handling can incorporate semantic disambiguation of words with various implications, like how syntactic disambiguation of expressions can be confounded as various grammatical features is refined at the syntactic level.

1. **Discourse**

While grammar and semantics work with sentence-length units, NLP's talk level works with message units that are longer than a sentence, for example it doesn't comprehend multi-sentence messages as a progression of single-sentence sentences

**NLP TOOLS AND TECHNIQIUE**

NLP can be used through SaaS (Software as a Service) tools or using open-source libraries.

SaaS tools are powerful, out-of-the-box, cloud-based solutions that can be implemented with little or no code. SaaS platforms often provide pre-trained NLP models and APIs. These are for users who need more flexible low-code options. A professional developer or programmer who wants to simplify their work.

Open-source libraries, on the other hand, are free and flexible, allowing you to fully customize your NLP tools. However, because they are aimed at developers, they are very complex to understand and require machine learning experience to build open-source NLP tools. Fortunately, however, most are community-driven frameworks, so you can count on a lot of support.

|  |  |  |
| --- | --- | --- |
| **Technique** | **Example** | **Output** |
| Sentence Segmentation | Mark met the president. He said: ”Hi! What’s up -Alex?” | Sentence 1 - Mark met the president.  Sentence 2 - He said: “Hi! What’s up - Alex?” |
| Tokenization | My phone tries to ‘charging’ from ‘discharging’ state. | [My] [phone] [tries] [to] [‘] [charging] [‘][from] [‘][discharging] [‘] [state][.] |
| Stemming/Lemmatization | Drinking, Drank, Drunk | Drink |
| Part-of-Speech tagging | If you build it, he will come. | **IN** - prepositions and subordinating conjunctions.  **PRP** - Personal Pronoun  **VBP** - Verb Noun 3rd person singular present form.  PRP- Personal pronoun  **MD** - Modal Verbs  **VB** - Verb base form |
| Parsing | Mark and Joe went into a bar. | (S(NP(NP Mark) and (NP(Joe))  (VP(went (PP into (NP a bar)))) |
| Named Entity Recognition | Let’s meet Alice at 6 am in India. | Let’s meet Alice at 6 am in India  Person (Alice)  Time (6am)  Location (India) |
| Coreference resolution | Mark went into the mall. He thought it was a shopping mall. | Mark went into the mall. He thought it was a shopping mall. |

The Natural Language Toolkit (NLTK) using Python is one of the leading tools for NLP modelling. NLTK focuses on research and education in the field of NLP and is supported by an active community and a variety of language processing tutorials, sample datasets, and resources, including comprehensive manuals on language processing and Python.

This library takes some time to master but is considered a great playground for hands-on experience with NLP. The modular structure of NLTK provides numerous components for NLP tasks such as tokenization, tagging, stemming, parsing, and classification.



**Fig. 3. NLP Toolkit**

There are different techniques in NLP that we can use to extract text from a given text snippet:

* **Sentence segmentation** - Defines sentence boundaries in the given text. That is, where one sentence ends and another begins. Sentences are often marked ended with the punctuation mark ‘.’.
* **Tokenization** - Identifies various words, numbers, and other punctuation mark and treat them individually.
* **Stemming** - It eliminates the endings from words, for example, 'eating,' which is diminished to 'eat.'
* **Part of speech (POS) tagging** - Assign a unique part-of-speech tag to each word in the sentence. Designating a word as a noun or adverb.
* **Parsing** - The specified text falls into various categories. To answer a question like this part of the sentence, modify another part of the sentence.
* **Named Entity Recognition** - Identifies people, places, times, and other entities in a document.
* **Co-Reference resolution** - This is to define the relationship between a particular word in a sentence and the previous and next sentences.

**Applications of NLP**

Machine interpretation, email spam discovery, data extraction, synopsis, and question addressing are only a couple of the spaces where Natural Language Processing might be utilized.

1. **Machine Translation**

Because the internet connects the majority of the globe, the task of making knowledge open and available to everyone is a challenge. The language barrier is a significant hurdle in making information available. There are several dialects, each with its own sentence structure and punctuation. Machine Translation is the process of interpreting sentences from one language to the next with the use of a translation engine like Google Translate. The challenge with machine interpretation advancements aren't only understanding words but maintaining the relevance of sentences, as well as syntax and tenses. The genuine AI gathers as a good deal records as they could that appears to be equal throughout dialects and crunches it to decide the chance that something in Language is comparable. Language B has anything to do with A. In September 2016, Google announced a new machine interpretation framework based on Artificial Neural Organizations and Deep Learning. Different methodologies have recently been presented to measure the quality of machine interpretation by evaluating theory with reference interpretations. Word blunder rate and position-autonomous word blunder rate are examples of such techniques (Tillmann et al., 1997) [37], age string exactness (Bangalore et al., 2000) [38], multi-reference word error rate (Nießen et al., 2000) [39], multi-reference word mistake rate (Nießen et al., 2000) [40], NIST score (Papineni et al., 2002) [39], BLEU score (Papineni et al., 2002) (Doddington, 2002) [41] All of these criteria seek to estimate human evaluation and frequently achieve a startling level of similarity to human emotional evaluations of familiarity and sufficiency (Papineni et al., 2001; Doddington, 2002)

1. **Text Categorization**

Classification frameworks take a large amount of data, such as authority archives, military loss reports, market data, and newswires, and divide it into preset groups or lists. For example, The Carnegie Group's Construe framework (Hayes PJ, Westein ; 1991)[44] imports Reuters articles and saves time by doing the work that would otherwise be completed by staff or human indexers. Order frameworks have been used by a few businesses to categorise annoyance complaints or objection requests and direct them to the appropriate task groups. Email spam channels are another use of text order. Spam channels are becoming increasingly important as the primary line of defence against unwanted messages. A fictitious negative and fictitious positive Spam channels are at the heart of NLP innovation, and it's been reduced to a test of eliminating significance from text strings. A sifting setup for an email system employs a number of rules to determine which of the incoming messages are spam and which are not. There are several types of spam channels available. Content channels: Examine the content of the message to determine whether it is spam or not. Header channels: Look over the email header for any fake material. All messages from boycotting beneficiaries are blocked via the general blacklist channels. Client-defined models are used in Rules-Based Filters. For example, blocking explicit individual sends or stopping mail containing a certain term. Filters for consent: Require that anybody creating an impression be pre-approved by the recipient. Challenge Response Filters: Requires everybody who makes an impression to input a code in order to get permission to send email.

1. **Spam Filtering**

It uses text order, and several AI algorithms, such as Rule Learning, have recently been used to message categorization or Anti-Spam Filtering (Cohen 1996) [45], Bayes, Nave (Sahami et al., 1998 ;Androutsopoulos et al.,2000b ;Rennie .,2000) [46] [47] Memory-based Learning (48),Memory-based Learning (48),Memory-based Learning (48), (Androutsopoulos et al.,2000b) Support vector machines (n.d.) (n.d.) (n.d (Druker et al., 1999) Decision Trees [49], (Carreras and Marquez , 2001) [50] Model of Maximum Entropy (Berger et al. 1996) [51]. Consolidating various pupils every now and again is a good idea (Sakkis et al., 2001) [52]. It is preferable to use these techniques since the classifier is obtained by preparing information rather than by supplying it. Despite its effortlessness, the credulous bayes is preferred due to its display (Lewis, 1998) [53] Two types of models have been used in text categorization (McCallum and Nigam, 1998) [54]. Both modules assume that a correct vocabulary is present. However, in the first approach, a record is created by first selecting a subset of jargon and then repeatedly using the chosen words, to some extent once regardless of request. The model is known as the Multi-variate Bernoulli model. It records which words are used in an archive, independent of the amount of words or the request. A record is constructed in the second model by selecting a group of word events and organising them in any request. In addition to the Multi-variate Bernoulli model, this model is known as the multi-nomial model, and it captures data on how frequently a term is used in an archive. The majority of anti-spam text arrangement methods The multivariate Bernoulli model has been used in email filtering (Androutsopoulos et al., 2000b) [47].

1. **Information Extraction**

The recognition of indications of interest in text-based information is a concern for data extraction. Extraction of components such as names, locations, events, dates, times, and expenses is a powerful approach for certain applications to summarise the data relevant to a client's requirements. The programmed identifiable evidence of crucial facts may increase precision and productivity of a coordinated quest thanks to a space explicit internet searcher. To extract the important fields of examination papers, stowed away Markov models (HMMs) are used. These deleted text sections are utilised to allow for the examination of specific areas, the display of indexed lists, and the matching of references to papers. For example, you may have seen the spring up advertisements on any sites showcasing the new products you may have looked at on an online based store with restrictions. Two types of models have been used in information retrieval (McCallum and Nigam, 1998) [55]. The two modules recognise the existence of a valid jargon. Under any event, in the first approach, a record is created by first selecting a subset of jargon and then repeatedly using the chosen words, once with almost no structure. The model is known as the Multi-variate Bernoulli model. It records which terms are used in a report, independent of the amount of words or the request. A report is created in the second model by selecting a group of word events and coordinating them in any request. Apart from the Multi-variate Bernoulli model, this model is known as the multi-nomial model, and it captures data on how frequently a word is used in a record.

Over the next few years, information disclosure will be a hot topic of discussion. Information disclosure research employs a variety of techniques to extract important data from source documents, including

Stop-words, Chunking or Shadow Parsing, and Grammatical Features (POS) labelling (Keywords that are utilised and should be eliminated prior to handling reports), Word reference-based stemming and Porter style stemming (Porter, 1980) [55] are two ways for mapping words to some basis. The first alternative has higher precision but higher execution costs, whereas the last option has reduced execution costs but is often insufficient for IR). Phrases that are compound or statistical (Compounds and factual expressions record multi token units rather than single tokens.) Word Sense Disambiguation (Word sense disambiguation is the process of determining what a word means in context.) When used for data recovery, words are replaced in the record vector by their faculties.)

Its divided data may be used in a variety of ways, for example, to create an outline, to assemble data sets, to recognise catchphrases, to organise text items according to a few pre-defined classifications, and so on. For example, CONSTRUE was developed for Reuters and is used to sort stories (Hayes, 1992) [57]. Although it has been suggested that several Internet Explorer frameworks can efficiently extract words from records, finding relationships between the terms remains a challenge. PROMETHEE (Morin,1999) [58] is a framework for extracting lexico-syntactic instances in comparison to a certain computed link. IE frameworks should function on a variety of levels, from word recognition to conversation investigation at the complete record level. The Blank Slate Language Processor is being used (BSLP) (Bondale et al., 1999) [59] a method for examining a true regular language corpus made up of responses to open-ended surveys in the field of marketing.

MITA (Metlife's Intelligent Text Analyzer) (Glasgow et al. (1998) [60]) is a framework that extracts data from disaster-prevention apps. Ahonen et al. (1998) [61] proposed a standard structure for text mining that incorporates commonsense and conversational level text analysis.

**CONCLUSION**

As explained in the context above, NLP has a fairly long history of research dating back to the 1950s, but many of its applications have only recently emerged. With the introduction of Google as the leading search engine, the increasingly digital world, and the rise in employment, NLP has sneaked into our lives almost unnoticed. But this is the reason for multiple conveniences in our daily lives. Developers can use NLP to perform tasks such as speech recognition, sentiment analysis, translation, automatic grammar correction when typing, and automatic answer generation. NLP is a challenging field because it deals with human languages ​​that are very diverse and can be spoken in different ways. NLP's various algorithms and techniques give developers a wider range of improvements and accurate results. NLP adoption is expected to gain momentum over the next few years with the introduction of more personal assistants, smartphone enhancements, and the development of big data to automate everyday human tasks.

1. Huang, Rachel, Jonathan Pedoeem, and Cuixian Chen. "YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers." *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018