IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

NATURAL LANGUAGE PROCESSING

Pratibha Dahiya[#1], Rama Raman[#2], Shubham Tiwari[#3], Abhishek Kumar Singh[#4]

[1,2,3&4]DEPARTEMT OF MCA, IIMT COLLEGE OF ENGINEERING, GREATER NOIDA, U.P.

ABSTRACT

This paper shares its views on Natural Language Processing (NLP) from the beginning of the birth of computer till now. As the human intelligence is measured on the basis of their capabilities of understanding and speaking a variety of languages so the same applies on the technical world. We have come through many generations of computers and in each we have seen a variety of modifications as per the technologies need. Among which NLP plays a major role which involves the use of computer algorithms to analyze, interpret and generate human-like language. It's a subfield of artificial intelligence that focuses on understanding and processing natural language text or speech. NLP applications include chatbots, language translations, sentiment analysis, and more.

1. KEYWORDS:

Natural language processing, human intelligence, Algorithms, Artificial intelligence, Chatbots.

2. INTRODUCTION

NLP is basically a methodology or it is a set of methods that makes human language accessible to computers. To understand it on a further level, we will have to dig into its past from where it actually begins.

The main goal of Natural Language Processing is to acquire one or more features of an algorithm or system. It is commonly used in multilingual event detection. The system mainly incorporates modular set of foremost polyglottic Natural Language Processing (NLP) tools. The pipeline integrates modules for basic NLP processing as well as more advanced tasks such as cross-lingual named entity linking, semantic role labelling and time normalization.

Thus, the cross-lingual framework allows for the interpretation of events, participants, locations and time, as well as the relations between them. Output of these individual pipelines is intended to be used as input for a system that obtains event centric knowledge graphs. Every module function similarly to a UNIX pipe: it accepts standard input, performs some annotation, and then outputs a standard output that serves as the input for the subsequent module.

Pipelines are constructed using a data-centric architecture to allow for module adaptation and replacement. Modular architecture also permits dynamic distribution and a variety of combinations. The majority of the research in Natural Language Processing is done by computer scientists, although linguists, psychologists, philosophers, and other experts have also expressed interest.

The difficulty of using natural language for computer communication is addressed by a variety of ideas.

3. A BRIEF HISTORY OF NLP

- **Level 1: late 50's-** Just after the computers came into existence Scientists started to explore the possibilities of making human languages understandable to the machines(computers). ELIZA is an early NLP program introduced by Joseph weizenbum at MIT, whose main aim is to establish real time communication between machines and human beings. Basically, ELIZA is considered as the first chatbot designed to stimulate conversation with a psychotherapist. Even though its methodology was totally based on recognising certain patterns in human behaviour and it's all responses were prescript, yet people found it really amazing taking it as if they were actually interacting with a physical being.

- **Level 2**: Early 70's- In 1968, a new approach to NLP understanding was introduced by Terry Winograd that allows human interaction using English terms. SHRDLU worked on the simple idea of using predefined rules to examine and process text. There were approx. 50 words: that include nouns like "block and cone", adjectives like" big, blue" and verbs like "move to, place on "etc. That allowed users to issue commands like "move the red block onto the green block" and the execution is done accordingly.

- **Level 3:** Between 80's to 90's- later in 1980's Machine learning and statistical approach initiated a level up in NLP. One of the major developments of that time is HMM (hidden Markova model). It describes the probabilities of relations, a sequence of hidden states and sequence of observations. It is generally used to classify the sequences to predict the future observations.

  The HMM uses two types of variables - hidden states and observations. Probability distribution is used to model the relationship between the observations and hidden states. There were two probability approaches used to define that relationship-:

  The transition probabilities define the transitioning probabilities from one hidden state to another and the emission probabilities provides an observation of an output given in a hidden state.

- **Level 4: 2000s to 2010s-** Deep learning and Neural networks propelled the NLP to a new height. Statics in NLP leads to the birth of computer linguistics and since then many technologies have been applied to machine learning including: HMM, decision trees, random forests, naive Bayes, k-nearest neighbour, and other support vector machines. The Neural networks came into actual process at the end of 2010's and till then it has transformed the whole internal structure of NLP even replacing the old methods.

  Traditionally, features of NLP were often Handcrafted, incomplete and even time consuming, but neural networks provided the facility to learn multilevel features automatically and provides a better result.

  Word embeddings and NN architectures are the main two innovations that enabled the use of neural networks in NLP.

  - **Word embedding-** Represents word as a real-valued vector in a small dimensional space area which made it much easier to identify the similar words because of their closeness in the vector space.
  - **NN Architecture-** This is applied to language modelling and later moved on to the morphology, conference resolution, POS tagging, semantics. And from these fields it is applied to the applications like information retrieval/extraction, text classification, summarization, Q/A, machine translation.

Level 5: PRESENT DAY

Open AI and transformer models like GPT (Generative pre trained transformer) have made significant strides in NLP. These models can process and generates human like texts by capturing the contextual dependencies within large amount of training data. GPT-3 released in 2020 served as one of the major programs of NN. It has come out Clean over 175 billion machine learning parameters to put things into scale.

2. WHAT IS THE DIFFICULTY OF NLP?

Speech recognition constructs the major implementation in NLP where speech signal is served as its input and the output is word string. The input can be in the form of a single sentence or multiple sentence at a time or it might not be

sentences at all in the sense of complete grammatical units and could be the fragments of languages. And sometimes the input might contain some useful cues like punctuations and capitalisations or there might be some sentence boundaries that could be missing. The output from a system that incorporates NLP. It can be an answer from database, or a command to change some data in database, spoken response and some other actions on part of system. But these are the output of the system as the whole not the output of the NLP components.

A further illustration of the difficulty of language comprehension comes from a recent Calvin and Hobbes cartoon:

Calvin: I enjoy using verbs.

Hobbes: How come?

Calvin: I make use of nouns and adjectives as verbs.

When was "access" a thing, you ask? It's something you do now. It became verbed.

Calvin: Verbing makes words strange.

Hobbes: Perhaps in the end we could render language completely incomprehensible.

For many years, the word error rate—which accounts for replacements, insertions, and deletions—has been the standard metric. It is widely used, simple to use, and so effective that there isn't much need for the speech research community to alter it. It is relatively simple to determine if an SR system is performing well or not because the SR task is so clearly defined. It is also very simple to determine which of two distinct SR systems, given identical input, performs better.

An NLP can receive inputs from a wide variety of sources.

Language can take many different forms. Examples include text paragraphs that may contain some non-natural language, commands entered directly into a computer system, the (perhaps imperfectly recognized) output of an SR system, etc.

A system that uses natural language processing (NLP) may ultimately produce a spoken response, a database response, an order to modify data in a database, or some other system activity. However, it is crucial to note that they represent the system's overall output rather than just the NLP component's output.

- **LEVELS OF NLP**

One of the most illustrative ways to depict Natural Language Processing is through the "levels of language," which help to realize the phases of Content Planning, Sentence Planning, and Surface Realization in order to produce NLP text.

The study of language meaning, language context, and language varieties is known as linguistics. The following are some of the key terms used in natural language processing:

1. Phonology

Phonetics The branch of linguistics known as phonology deals with the orderly arrangement of sound. The word "phonology" originates in Ancient Greek, where "phono-" denotes voice or sound, and the suffix "-logy" denotes words or speech. Phonology is "the study of sound pertaining to the system of language," according to Nikolai Trubetzkoy in 1993. Contrary to Lass's 1998 assertion that phonology is a broad term referring to language's sounds and a subfield of linguistics, phonology proper can be defined as "the function, behaviour, and organization of sounds as linguistic items." The semantic use of sound to encode meaning in any human language is known as phonology.

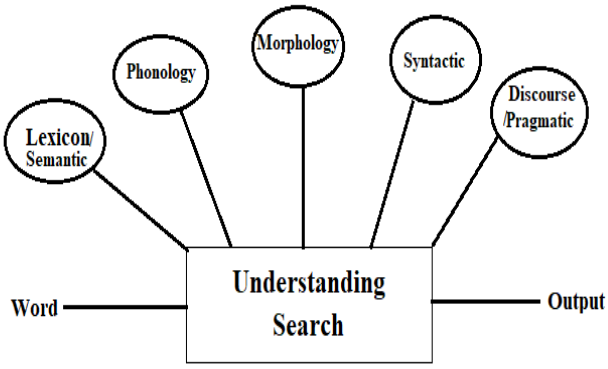
2. Morphology

The word's constituent parts stand for the smallest meaning units, or morphemes. Morphemes are the starting point for morphology, which is the nature of words. The word precancellation, for instance, can be morphologically examined to

reveal three distinct morphemes: the prefix pre, the root cancella, and the suffix -tion. This is an example of morpheme. Humans can break down every unknown word into its constituent morphemes in order to understand its meaning, as morphemes have the same interpretation across all words. For instance, adding the suffix -ed indicates that a verb's action occurred in the past. Lexical morphemes are words that have meaning all by themselves and cannot be separated (e.g.: table, chair) Words ending in -ed, -ing, -est, -ly, or -ful, for example, Grammatical morphemes (e.g. Worked, Consulting, Smallest, Likely, Use) are coupled with the lexical morpheme. Bound morphemes are those grammatical morphemes that appear in combinations (example. -ed, -ing) Bound morphemes and derivational morphemes are two categories of grammatical morphemes.

3. Lexical

Both humans and NLP algorithms decipher word meanings in lexical systems. Word-level understanding is granted via various processing kinds; the first of these involves assigning a part-of-speech tag to every word. When a word can function as more than one part of speech, it is assigned the most likely part of speech tag according to the context in which it appears. Single-meaning words can take the place of semantic representations at the lexical level. The type of representation used in an NLP system varies depending on the semantic theory that is applied.



4. Syntactic

This degree of attention places on carefully examining the words of a sentence to determine its grammatical structure. In this level, parser and grammar are both necessary. A representation of the sentence that reveals the structural dependencies between the words is the result of this level of processing. In most languages, syntax carries meaning because order and dependency add to connotation. The meanings of the two statements, "The cat chased the mouse" and "The mouse chased the cat," for instance, are very different despite just having little syntactical differences.

5. Semantic

Most people mistakenly believe that meaning is predetermined when it comes to semantics, yet meaning is actually conferred by all of the levels. Semantic processing focuses on the relationships between the sentence's word-level meanings to ascertain the sentence's potential meanings.

6. Discourse

Discourse focuses on the elements of the text as a whole that connect individual words to convey meaning.

7. Pragmatic

The pragmatic approach to language use focuses on the firm application of language in specific contexts. It employs nuance beyond the literal meaning of the text to help grasp the purpose and clarify how additional meaning is interpreted from texts without being encoded in them. This necessitates having a broad awareness of the world, including plans, goals, and intents.

- **APPLICATIONS OF NLP**

1. Machine Translation

Since the majority of the world's population uses the internet, it might be difficult to make data widely available. One of the biggest obstacles to data accessibility is the language barrier. Numerous languages exist, each with its own unique syntax and sentence structure. The difficulty with machine translation technology is not so much in translating words as it is in maintaining sentence meaning, syntax, and tenses.

2. Spam Filtering

Text classification is how it operates, and in more recent times, machine learning techniques such as Rule Learning have been used for text categorization and Anti-Spam Filtering.

3. Dialogue System

Dialogue systems, which concentrate on a tightly defined applications (like refrigerator or home theatre systems), are arguably the most ideal application of the future. Dialogue systems now use the phonetic and lexical levels of language in the systems envisioned by significant providers of end user applications. It is thought that these dialogue systems provide the possibility for fully automated dialog systems when they make use of all language processing levels.

4. Medicine

The realm of medicine also uses NLP. One of the major NLP efforts in the world of medicine is the Linguistic String Project-Medical Language Processor. In order to identify potential side effects of any medication, the LSP-MLP assists in enabling doctors to extract and summarize information about any signs or symptoms, drug dosage, and response data while highlighting or flagging data items.

5. Information Extraction

Finding interesting terms in textual material is the focus of information extraction. In numerous applications, the ability to extract entities like names, locations, events, dates, times, and prices is a potent means of condensing the data pertinent to the user's requirements. When using a domain-specific search engine, directed searches can be more accurate

• APPROACHES

The rationalist or symbolic approaches presuppose that a significant portion of the knowledge stored in the human mind is fixed beforehand, most likely due to genetic inheritance, rather than coming from sense perception. The person who supported this strategy the most was Noam Chomsky. It was believed that by providing basic knowledge and a reasoning mechanism, a computer might be programmed to perform similar functions to those of a human brain.

- **Hidden Markov Model**

An HMM is a system that switches between many states, producing workable output symbols at each switch. Although they may be vast, the sets of feasible states and distinct symbols are known and finite. The core workings of the system are hidden, yet we may criticize its results. Covert Markov in voice recognition, models are widely employed to match the sequence of individual phonemes in the output sequence. It's been said that Frederick Jelinek, a proponent of statistical-NLP who spearheaded the use of HMMs at IBM's Speech Recognition Group, quipped, "Every time a linguist leaves my group, the speech recognizer performs better."

CONCLUSIONS:

To sum up, natural language processing is in the vanguard of technology advancement, providing a viable path towards improving communication, streamlining processes, and encouraging a more profound comprehension of the intricacies present in human language. A judicious application of NLP across a wide range of fields will surely create a more intelligent and connected world as we move forward. Advances in natural language processing are a critical step in closing the gap between machine comprehension and human communication, which will enable a wide range of applications in numerous fields.

Ultimately, this research paper's journey has shown the difficulties and achievements in Natural Language Processing, highlighting the necessity of ongoing investigation, creativity, and moral reflection to guarantee its responsible application and full potential.

REFRENCES

1. Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544-551.

2. Eisenstein, J. (2019). *Introduction to natural language processing*. MIT press.

3. Jones, K. S. (1994). Natural language processing: a historical review. *Current issues in computational linguistics: in honour of Don Walker*, 3-16.

4. King, M. (1996). Evaluating natural language processing systems. *Communications of the ACM*, *39*(1), 73-79.

5. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, *82*(3), 3713-3744.

6. Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603-649.

7. Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

8. Liddy, E. D. (2001). Natural Language Processing.

9. Reshamwala, A., Mishra, D., & Pawar, P. (2013). Review on natural language processing. *IRACST Engineering Science and Technology: An International Journal (ESTIJ)*, *3*(1), 113-116.

10. Smeaton, A. F. (1992). Progress in the application of natural language processing to information retrieval tasks. *The computer journal*, *35*(3), 268-278.