2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) | 979-8-3503-8247-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/UPCON59197.2023.10434553

Developments in Natural Language Processing: Applications and Challenges

Sunil G
School of CS & AI, SR University,
Warangal, Telangana, India
g.sunil@sru.edu.in

Dr. Tamirat Tagesse Takore
Assistant Professor, Department of Electrical and Computer Engineering
Wachemo University,
Ethiopia
tame277@gmail.Com

Praseeda Ravuri
Computer Science Engineer, Oregon state university,
Corvallis, Oregon, USA 97331
praseeda.ravuri1@gmail.com

Amandeep Nagpal
Lovely Professional University,
Phagwara
amandeep.nagpal@lpu.co.in

Dr. Melanie Lourens
Deputy Dean Faculty of Management Sciences,
Durban University of Technology
South Africa
melaniel@dut.ac.za

K. Ganapathi Babu
Assistant Professor, Computer Science and Engineering, St. Martin's Engineering College,
Secunderabad -500100, Telangana
ganapathicse2@gmail.com

***Abstract*: Natural language is any language that a person uses or speaks, such as Hindi, English, French, Marathi, Bengali, Gujrati, and so forth.Computers can better comprehend, interpret, and manipulate human language with the help of the area of called NLP. NLP is dependent on a number of academic disciplines, notably software engineering and computational languages, in order to bridge the gap among human communication and machine understanding. Natural language processing (NLP), which can represent and interpret human language computer, has attracted a lot of attention lately.Its applications have grown to include, among other things, machine translation, spam email detection, information extraction, summarization, and question answering. The article breaks its topic into four sections: a discussion of different NLP levels and NLG components comes first, then the history and evolution of NLP, the state of the art, a rundown of all the different NLP applications, and a discussion of recent advancements and difficulties.**

Keywords: NLP, application, challenges, techniques, development.

I. INTRODUCTION

Natural language processing (NLP), a field of artificial intelligence, emphasizes the translation of computational linguistics. This field combines machine learning methods that assess text and audio in a range of circumstances quantitatively. Additionally, it includes the wide and potent field of computational linguistics' pragmatic study, which has grown via the use of several approaches [1] the increasing power and accessibility of NLP approaches, which daily enhance the accuracy and development of computational language.

The areas of natural language processing and machine learning are those with the most active research. Most of the inspirations on NLP are drawn from a variety of fields, includes psychological research, neuroscience, cognitive linguistics, and several more. Engineering computational models that solve challenges with linguistic and interaction between people are involved. Numerous software tools in the language modeling fields have been developed with this goal in mind, making it possible for people to easily grasp computer language.

A. Importance of NLP

Smaller subcategories may be created from the NLP's two primary sections, basic area and practical area, which deal with two separate areas of research.. Language modeling is one of the basic or fundamental issues that core NLP topics handle and research. It is possible to come across word relationships that naturally arise in language. In addition to these NLP fundamental domains, morphological processing also deals with the distinction of meaningful word parts. Syntactic parsing, which creates sentence diagrams, is used to try to analyze linguistic content appropriately. Words, sentences, phrases, and higher levels of abstraction in a text are decoded via semantic processing. NLP is a crucial component in overcoming phobias, anxiety, and other issues [2].

B. NLP and NLG

A set of rules or a set of symbols make up a language. A mixture of symbols is used to transmit or distribute information. The Rules impose their will on symbols. The two primary divisions of natural language processing are NLG (Natural Language Generation), which develops the task of comprehending and creating the text, and NLU (natural language understanding) (Figure 1).

Phonology, which studies sound, morphology, which studies word formation, syntax, which studies sentence structure, semantics, which studies syntax, and pragmatics, which studies understanding, are all included in the study of language.

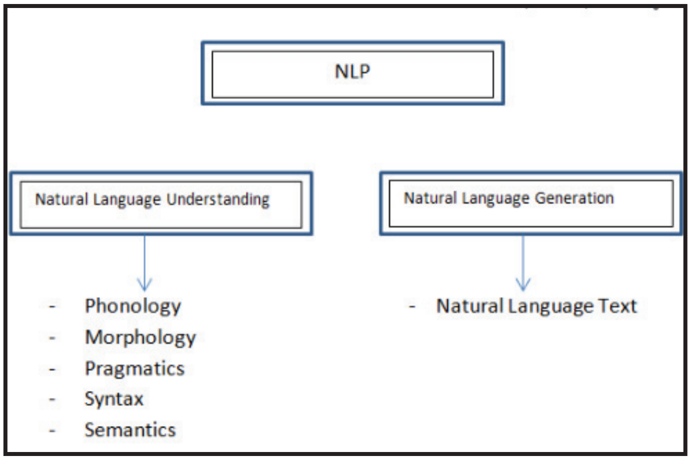

Fig. 1. NLP classification

The bulk of the research in natural language processing is done by computer scientists, although linguists, psychologists, and philosophers have all shown interest in the field. That NLP improves our grasp of human language is its most contradictory aspect. Diverse concepts and techniques in the field of natural language processing attempt to solve the challenge of utilizing natural language to interact with computers. Ambiguity is one of the primary problems with natural language, which often arises at the syntactic level and involves subtasks for word generation and lexical and morphological investigations.

Each of these stages has the risk of posing inquiries, some of which may be resolved by comprehending the whole assertion. The problem may be solved using a variety of strategies, such as Minimizing Issue, Preserving Ambiguity, Interactive Disambiguation, and Weighting Ambiguity [3]. One method some academics have proposed to eliminate it is to maintain ambiguity [3] [4] [5]. Their objectives are very comparable to the last one; a wide variety of ambiguities are covered, and a statistical element is included in their technique.

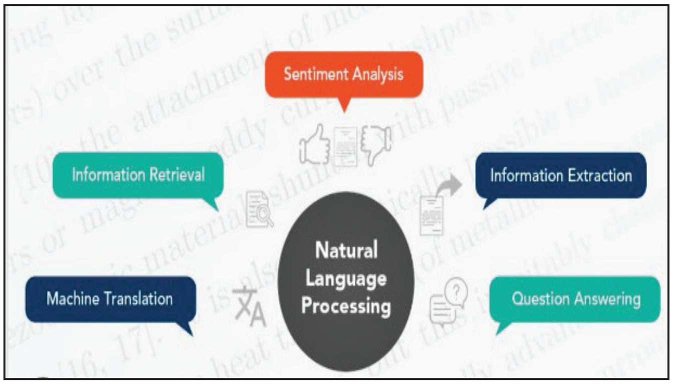

Fig. 2. NLP techniques

II. LITERATURE REVIEW

The study of computational language, AI, and interaction between humans and computers using natural languages like English, Arabic, Urdu, etc. is known as natural language processing, or NLP. NLP has a human component and is used by people to interface with computers [6].In addition, artificial neural networks model nonlinear processes and use tools to address NLP-related issues like classification, which involves clustering, regression, pattern recognition from various angles, decision-making, visualization, and computer vision [7].

A. Machine translation

The study of using translation software to transliterate text or speech from one script to another is known as machine translation (MT), which is a subfield of computational linguistics.Due to the adoption of NLP applications, which allow computers to communicate with several individuals at once without requiring human effort, computers and humans are becoming increasingly similar. The automated conversion of text into spoken or written language is known as machine translation.

1) Translation using rules-based software (RBMT)

The first useful method of machine translation was RBMT, developed many years ago. It works by identifying the text in a source sentence, analyzing the kind of text it is, and then translating it into the proper language using linguistic rules. In reality, there exist guidelines for language translations. Statistical machine learning or a hybrid system has superseded rule-based translation [8].

2) SMT (statistical machine translation)

It is a kind of machine translation that involves training on massive amounts of data from bilingual, multilingual, or monolingual corpora. Source text and transliteration have a systemic relationship; SMT uses both features to produce end outputs that mimic translation of the provided source text.The SMT allows for the measurement of the proportion of translation and helps to eliminate manual translation [9].

3) EBMT (example-based machine translation system)

The example-based machine translation system (EBMT) is used by a translator to translate a sentence from any language. It is simple to translate a sentence, and the outcome will almost certainly be correct. Otherwise, a large quantity of material is translated with only one click, resulting in a great deal of ambiguity in the text. It takes a long time to carry out a lot of sentences [10].

4) Hybrid

In the first scenario, the RBMT engine first translates the text before it is processed by a machine and mistakes are fixed as they arise. Second, although the SMT engine uses input data to help the RBMT engine, the latter does not translate the text [11].

B. Automatic summarization

Text is divided into digestible chunks for text summarization. The most pervasive problem in NLP is automatic text summarization. There are two primary methods for automatically summarizing text in NLP. One is based on extraction, while the other is based on abstraction [11].

C. Sentimental Analysis

Another name for sentiment analysis is opinion mining. Systems are being created to try to detect and extract opinions from text in the area of NLP referred to as sentiment analysis

[12, 13]. Even if the system recognizes the views, these systems extract the statement's properties. In text summary, opinions and views may be divided into two categories: facts and opinions and views. Facts are assertions about text that are not biased. Opinions and opinions are often subjective phrases that show how the general public feels about a text. According to some researchers, subjective classification and polarity classification are two of the two sub problems of sentiment analysis [14].To categorize a statement as subjective or objective uses the term "subjective classification." Expressing positive, neutral, and negative opinion/views is known as polarity classification [15].

D. *Classifying texts and responding to questions*

A technique for classifying texts according to context is text classification. Free text is accessible from a variety of sources, including newspapers, social media, chat rooms, internet forums, and more. The text might be formatted appropriately using text categorization. Classifying text into different categories, such as documents, paragraphs, sentences, words, and letters, is the fundamental work of natural language processing (NLP) from figure 3.

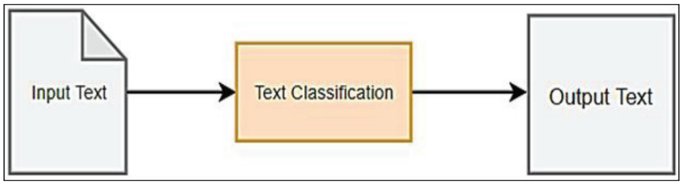

Fig. 3. Classifying text

III. METHODOLOGY

A. *Technique for Machinery Translation*

The technique of translating from one natural language into another while maintaining meaning and creating coherent output is known as machine translation. Machine Transition Techniques are built on a variety of models, including:

1) *Bilingual machine translation*

The ability of a bilingual machine translation system to translate between additional languages cannot be adjusted.

2) *Transfer-based machine translation*

The three modules that make up this translation paradigm are Analysis, Transfer, and Generation.

3) *Interlingual-based machine translation*

The two primary components of this translation model are the Analysis module and the Generation module.

4) *Memory-based machine translation*

This paradigm of translation is based on the concept of "translation memory." It uses a corpus-based methodology. Without actually examining the original material, the system simply reuses translations that the professional translator had already saved. A dictionary (terminology support) is then utilized to assist the expert in translating the portions of texts that have not yet been translated.

5) *Machine translation using statistics*

It uses a corpus-based translation methodology. One of the first machine translation strategies is the use of statistical ideas.

B. *Technique of Morphological splitting*

Split morphology is a theory that requires specific data on the Derivation and Inflection that are distinct parts of grammar. While lexical rules and syntactic rules, respectively, regulate Inflections and Derivations.

C. *Language Disambiguation*

Word sense disambiguation is a notion in which a word that is used any number of times in a certain context is assigned a proper or "sense" to the corresponding phrase, which is often unknown to people. This context's characteristic offers the evidence needed to classify. The study on this topic has consistently and without a shadow of a doubt provided accurate or precise findings. In order to oversee machine learning strategies that provide a collection of manually sense-annotated instances, they conducted research on a number of techniques employing dictionary-based methodologies and the information that was stored in lexical resources.

Clusters of repeated words with word meaning have also been produced by study using unsupervised approaches. But only the supervised learning strategy has been deemed effective out of all these learning strategies.

D. *Taging and Identification*

The technique of assigning a word in a body to a relevant part of speech tag based on its meaning and context is known as POS tagging. Building parse trees, which are used to construct NERs, requires the usage of part of speech (hence referred to as POS) tags. Building lemmatizers, which are tools for breaking down words to their basic forms, need POS Tagging as well. There are several POS Taggings, including:

- Lexical Based Method
- Rule-BasedMethods
- Probabilistic Methods
- Deep Learning Methods

The capacity of electronic equipment to identify spoken words is known as speech recognition. A person's voice is captured by a microphone, and the hardware transforms the analog sound waves in the signal into digital audio.Software then processes the audio input, deciphering the sound as distinct words.

IV. RESULTS AND DISCUSSION

A. *Natural language processing application issues*

1) *Text format*

The biggest difficulties in machine learning are words and phrases that have syntactic and semantic translation issues. A single term might have several meanings in various languages. Humans are able to distinguish between different word meanings and use them appropriately. The machine is not able to comprehend human language just by reading some text; sometimes, text has concealed meaning, in which case a translator is not necessary. Multiple linguistic differences contribute to syntactic issues [10].

2) *Quality*

The usage of words for machine translation presents the most issues. Since a single word may have multiple meanings based on the circumstance in which it is spoken, software is unable to understand the nature of words and circumstances. People can grasp context, particularly in various settings, and can identify emotions, nonverbal cues, and other things [1].

3) *Lexical Gap*

The same meaning in natural language processing might be expressed in several ways. Lexical gaps considerably increase the number of questions that are answered the system since a question can often only be addressed if every referenced thought is understood.

4) *Ambiguity*

The repeated use of the same sentences might have different structural and syntactic meanings. When a single word, phrase, sentence, or paragraph has many meanings, it is said to be ambiguous. The vague term is often used in natural linguistic. Lexical analysis, syntactic analysis, semantic analysis, discourse analysis, and pragmatic analysis are a few different sorts of ambiguity.

B. Analysis of NLP techniques in comparison

After reading many research articles on the uses and methods of NLP, a comparison analysis is conducted. On the basis of many criteria, including accuracy, performance, decision-making, applicability, and approaches/techniques, a correlation table is created below.

TABLE I. NLP TECHNIQUES AND THE FACTORS THAT INFLUENCE THEM

Parameter	Techniques			
	Grammar Evaluation	Graphical Representation	Extraction of Features	Sentiment Analysis
Application	Translation of SRS written in natural language in formal language covering dynamicaspect of software system	Developing Test Cases	Detection of spam Character detection from backdrop facilitates engagement between many stakeholders by collecting needs from online discussion forums	Product Evaluations
Accuracy	more	more	more	a lesser degree than others
Performance	high	medium	high	medium
Schemas for implementing NLP or implementation approaches	Sentence Diagramming System by Reed Kellog	Analysis of the Boundary Values for the Parser/Parsed Tree POS Tagging	K -mean Clustering Algorithm CANARY tool.	
Decision Making	yes	no	yes	yes

V. CONCLUSION

This essay offers a summary of the recommended NLP methodology and approaches via a comparative examination of several NLP techniques. The efficacy of various suggested approaches is defined based on many characteristics. This study aids in comprehending the extent of the NLP discipline, the best approach for a given application, and possible future research projects. It is concluded that NLP may be combined with other approaches to create a platform that makes it simple for people to connect with computers. NLP has a broad variety of applications in the fields of medicine, security, social media, software development, robotics, and more.

REFERENCES

[1] Li, J., Monroe, W., &Jurafsky, D. (2016). Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

[2] Colnerič, N., & Demšar, J. (2018). Emotion recognition on twitter: Comparative study and training a unison model. *IEEE transactions on affective computing*, *11*(3), 433-446.

[3] Seligman, M. (2019). The evolving treatment of semantics in machine translation. *Adv. Empir. Transl. Stud. Dev. Transl. Resour. Technol.*, 53.

[4] Luong, M. T., Nakov, P., & Kan, M. Y. (2019). A hybrid morpheme-word representation for machine translation of morphologically rich languagees. *arXiv preprint arXiv:1911.08117*.

[5] Palmkvist, V., Castegren, E., Haller, P., & Broman, D. (2021, March). Resolvable ambiguity: principled resolution of syntactically ambiguous programs. In *Proceedings of the 30th ACM SIGPLAN International Conference on Compiler Construction* (pp. 153-164).

[6] Olex, A., Maffey, L., & McInnes, B. (2019, June). Nlp whack-a-mole: challenges in cross-domain temporal expression extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 3682-3692).

[7] Li, I., Fabbri, A. R., Tung, R. R., & Radev, D. R. (2019, July). What should i learn first: Introducing lecturebank for nlp education and prerequisite chain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 6674-6681).

[8] Jung, N., & Lee, G. (2019). Automated classification of building information modeling (BIM) case studies by BIM use based on natural language processing (NLP) and unsupervised learning. *Advanced Engineering Informatics*, *41*, 100917.

[9] Zabin, A., González, V. A., Zou, Y., & Amor, R. (2022). Applications of machine learning to BIM: A systematic literature review. *Advanced Engineering Informatics*, *51*, 101474.

[10] Alishahi, A., Chrupała, G., &Linzen, T. (2019). Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop. *Natural Language Engineering*, *25*(4), 543-557.

[11] Al-Makhadmeh, Z., & Tolba, A. (2020). Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. *Computing*, *102*, 501-522.

[12] Hassan, A., & Mahmood, A. (2018). Convolutional recurrent deep learning model for sentence classification. *Ieee Access*, *6*, 13949-13957.

[13] Rameshbhai, C. J., & Paulose, J. (2019). Opinion mining on newspaper headlines using SVM and NLP. *International journal of electrical and computer engineering (IJECE)*, *9*(3), 2152-2163.

[14] Liu, Z., Zhu, H., & Chong, T. Y. (2019, July). An NLP-PCA Based Trading Strategy On Chinese Stock Market. In *4th International Conference on Humanities Science, Management and Education Technology (HSMET 2019)* (pp. 80-89). Atlantis Press.

[15] Wang, J., Tuyls, J., Wallace, E., & Singh, S. (2020). Gradient-based analysis of NLP models is manipulable. *arXiv preprint arXiv:2010.05419*.