**TASK – 1**

**Project Name: Natural language processing (NLP)**



**About Natural Language Processing**

NLP refers to Natural language processing, which is a part of Computer Science, Human Language, and Artificial intelligence. It is the Technology used by Machines to understand, analyses, manipulate, and Interpret Human’s Language. It means training computers, for example, computers can understand text and speak the same as human beings speak. It can perform task such as translation, automatic summarization, named entity recognition, and topic segmentation.

Natural language processing covers a diverse range of tasks, methods, and linguistic phenomena. Advocates of “Natural Language Processing from Scratch” propose to use machine learning to train end-to-end systems that mute raw text into any output structure Example: a summary, database, or translation. The core work of Natural language processing is Somes taken to be transforming text into a stack of general-purpose linguistic structure. Building machine learning models whose architectures are inspired by linguistic theories. For Example, thes organization of sentences is often described as compositional, with meaning of larger units gradually contracted from the meaning of their smaller constituents.

**History of Natural Language Processing**

* The Natural Language Processing in (1940-1960) Machine Translation (MIT) came into existence. Natural Language Processing started in the 1940s.
* In the year 1948, the first NLP application was introduced by Birbeck college, London.1950s-In the year 1950s, there was a conflicting view between linguistics and computer science. Chomsky interduce is first book of syntactic structures and claimed the language of generative in nature. In 1957, Chomsky also introduced the idea of generative grammar, which is rule-based descriptions of syntactic structures.
* In (1960-1980) flavored with Artificial intelligence. The year 1960 to 1980, the key developments were:
* Augmented Transition Networks (ATN): Augmented Transition Networks is a finite state machine that can recognize regular languages.
* Case grammar: Case grammar was developed and introduced by Linguist Charles J. In 1968. Case Grammer use language such as English to certain kinds of verbs and objects. For Example: “Priya broke the mirror with the hammer". In this example case grammar identifies Priya as an agent, mirror as a theme and hammer is an instruction.
* LUNAR: Lunar is the classic example of Natural language Processing database interface system.It us ATNNs and Wood’s procedural semantics. Till the year 1980, natural language processing systems were based on complex sets of hand-written rules. After 1980, NLP introduced machine learning algorithms for language processing.

**Natural Language Processing (1950s – 1990s)**

The symbolic NLP is well-explained by John Searle’s Chinese room experiment: He gave the collection of rules (For Example a Chinese phrasebook, with questions and matching answers), the computer examines natural language understanding or other NLP tasks by applying the rules to the data it confronts.

* **1950s**: In 1954 The Georgetown experiment involved fully automatic translation of more than sixty Russian sentences into English. The authors claimed that within three or five years, machine translation would be a solved problem. Real progress was much slower, and after the ALPAC report in 1966, which found that ten-year-long research had failed to fulfill their expectations, funding for machine translations was dramatically reduced. Little further research in machine translation was conducted until the late 1980s. The first statistical machine translation systems were developed and came into Existence.
* **1960s**: The successful natural language processing systems developed in the year 1960 were SHRDLU a natural language system working was restricted blocks worlds with restricted vocabularies, and ELIZA, simulation of a Rogerian psychotherapist, written by Joseph Weizenbaum between the year 1964 and 1966. Using almost no information about human thought or emotion, ELIZA sometimes provided a startlingly human-like interaction. When the "patient" exceeded the very small knowledge base, ELIZA might provide a generic response.
* **1970s**: In the year 1970, many programmers began to write "conceptual ontologies", which structured real-world information data into computer language-understandable language data. Examples are QUALM (Lehnert, 1977), Politics (Carbonell, 1979), and Plot Units (Lehnert 1981). At this time the first chatterbots were written.
* **1980s**: At 1980s and early 1990s mark the day of symbolic methods in NLP. Focused areas of that time included research on rule-based parsing.

**Natural Language Processing Systems**

Natural Language Processing is a combination of computational linguistics-based modeling of human language with statistical, machine learning, and deep learning models. Together, these technologies make computers understand Human language if it is in text form or voice form to understand it fully complete with the speaker or writer’s intent. There are two main parts to natural language processing: data preprocessing and algorithm development. There are many different natural language processing algorithms, but two main types are commonly used:

* Rules-based system. This system uses carefully designed linguistic rules. This approach was used early in the development of natural language processing, and it is still used.
* Machine learning –based system. Machine learning algorithms use statistical methods.

**Natural language processing tasks**

Human language is filled with ambiguities that make it incredibly difficult to write software that accurately determines the intended meaning of text or voice data. Homonyms, homophones, sarcasm, idioms, metaphors, grammar and usage exceptions, variations in sentence structure these just a few of the irregularities of human language that take humans years to learn, but that programmers must teach natural language-driven applications to recognize and understand accurately from the start, if those applications are going to be useful.

* Speech recognition
* Part of speech tagging
* Word sense disambiguation
* Named entity recognition
* Co-reference resolution
* Sentiment analysis
* Sentiment analysis

**Natural language processing use cases**

Natural language processing is the driving force behind machine intelligence in many modern real-world applications. Here are a few examples:

* Spam detection: you may think of spam detection as an NLP solution, but the best spam detection technologies use NLP’s text classification capabilities to scan emails that indicate spam or phishing. These indicates include overuse of financial terms, bad grammar, threating languages and more. Spam detection is the main use full of NLP problems.
* Machine translation: Google translator is the best example of widely available NLP technology at work. Truly useful machine translation involves more than replacing words in one language with words of another. Machine translation tools are making good progress in terms of accuracy
* Virtual agents and chatbots: Virtual agent's best examples are Amazon's Alexa use speech recognition to recognize patterns in voice commands and natural language generation to respond with appropriate action or helpful comments. Chatbot performs the same in response to typed entries.
* Social media sentiment analysis: NLP has become an essential business tool nowadays for uncovering hidden data insights from social media channels. Sentiments analysis can analyze language used in social media posts, reviews and more.
* Text summarization: Text summarization uses NLP techniques to digest huge volume of digital text and creates summaries and synopses index, research databases and more. The best text summarization application uses semantic reasoning and natural language generation (NLG) to add useful context and conclusions to summaries.

**Advantages of Natural Language Processing**

* The Natural Language Processing system offers exact answers to the questions, and no need of unwanted information.
* The accuracy of the answer increases with the amount of relevant information increases.
* Structuring high unstructured is possible in data source, this is the main advantage.
* Users can ask direct questions about an given data and get direct response for the given information in seconds. This is the best method for time saving of NLP.
* It's easy to implement and user friendly.
* A program is less costly than hiring a person. A person can take more time compared to a machine to execute the given task.
* The Natural Language Processing system provides best solutions from the given data in natural language.
* Natural Language Processing allows you to do more language-based data comparison to a human being without fatigue.
* Natural Language Processing process helps in computer communication with a human in their language and scales the other related languages tasks.
* It is a faster customer service response time.

**Disadvantages of Natural Language Processing**

* Natural Language Processing may not show context
* Unpredictable
* This requires more keystrokes
* Natural Language Processing has a limited function
* Natural Language Processing is unable to adapt to the new domain.
* Natural Language Processing is built for a single and specific task.

**Natural Language Processing Methods and TOOLS**

In the early days, many language-processing systems were designed by symbolic methods the hand-coding of a set of rules, coupled with a dictionary More recent systems based on [machine-learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms have many advantages over hand-produced rules

* The learning procedures used during machine learning automatically focus on the most common cases, whereas when writing rules by hand it is often not at all obvious where the effort should be directed.
* Automatic learning procedures can make use of statistical inference algorithms to produce models that are robust to unfamiliar input. Generally, handling such input gracefully with handwritten rules, or, more generally, creating systems of handwritten rules that make soft decisions, is extremely difficult, error-prone and time-consuming.
* Systems based on automatically learning the rules can be made more accurate simply by supplying more input data. However, systems based on handwritten rules can only be made more accurate by increasing the complexity of the rules, which is a much more difficult task.

**Statistical methods**

Statistical revolution was began in 1980s-1990s, much natural language processing research has been heavily on machine learning the machine-learning paradigm is called instead for using statistical inference to automatically learn such rules through the analysis of large corpora*.* Many different classes of machine-learning algorithms have been applied to natural-language-processing tasks. These algorithms take as input a large set of "features" that are generated from the input data. the relative certainty of many different possible answers rather than only one, producing more reliable results when such a model is included as a component of a larger system. Statistical methods in Natural Language Processing research have been largely replaced by neural networks. However, they continue to be relevant for contexts in which statistical interpretability and transparency is required.

**Neural networks**

A major drawback of statistical methods is that they require elaborate feature engineering. The field has thus largely abandoned statistical methods and shifted to neural networks for machine learning. In some areas, this shift has entailed substantial changes in how Natural Language Processing systems are designed, such that deep neural network-based approaches may be viewed as a new paradigm distinct from statistical natural language processing

Natural Language Processing is a powerful tool with more benefits, but there are still a number of Natural Language Processing limitations and Problems.

* Contextual words and phrases and homonyms
* Synonyms
* Irony and sarcasm
* Ambiguity
* Erros in text or speech
* Colloquialisms and slang
* Domain-specific language
* Low-resource languages

**Natural Language vs. Computer Language**

|  |  |  |
| --- | --- | --- |
| Parameter | Natural Language | Computer Language |
| Ambiguous- They are ambiguous in nature | They are ambiguous in nature | They are designed to unambiguous |
| Redundancy | Natural languages employ lots of redundancy | Formal languages are less redundant |
| Literalness | Natural languages are made of idiom & metaphor | Formal languages mean exactly what they want to say |

**Machine Learning for Natural Language Processing**

The Most important thing, “machine learning” refers to “machine teaching” We know that machine needs to learn, so we make our task to create a learning framework and provide properly formatted, relevant, good clean data for the machine to learn.

Machine learning for NLP and text analytics involves many sets of statistical techniques to identify parts of speech, entities, sentiment, and other aspects of text. The techniques can be expressed as a model that is then applied to another text, also known as supervised machine learning. It also could be a set of algorithms that work across large sets of data to extract meaning, which is known as unsupervised machine learning. It’s important to understand the difference between supervised and unsupervised learning, and how you can get the best of both in one system.

The most popular supervised NLP machine learning algorithms are:

* Support Vector Machines
* Bayesian Networks
* Maximum Entropy
* Conditional Random Field
* Neural Networks/Deep Learning

Supervised Machine Learning for Natural Language Processing and Text Analytics

In supervised machine learning, a batch of text documents are tagged or annotated with examples of what the machine should look for and how it should interpret that aspect. These documents are used to “train” a statistical model, which is then given un-tagged text to analyze

**Summary**

* Natural Language Processing is a branch of AI which helps computers to understand, interpret and manipulate human language
* NLP started when Alan Turing published an article called “Machine and Intelligence”.
* NLP never focuses on voice modulation; it does draw on contextual patterns
* Five essential components of Natural Language Processing in Artificial Intelligence are

1)Morphological and Lexical Analysis

2)Syntactic Analysis

3)Semantic Analysis

4) Discourse Integration

5) Pragmatic Analysis

* Three types of the Natural process writing system are

1)Logographic

2) Syllabic

3) Alphabetic.

* Machine learning and Statistical inference are two methods to implementation of Natural Process Learning
* NLP is ambiguous while open-source computer language is designed to unambiguous
* The biggest drawback of the NLP system is built for a single and specific task only so it is unable to adapt to new domains and problems because of limited functions
* The biggest advantage of the NLP in Artificial Intelligence system is that it offers exact answers to the questions, no unnecessary or unwanted information

FROM,

VISHNU KUMAR. D.S

Machine Learning Group 37.0

Machine Learning Internship Task

Gmail-vishnukumardsvishnu@gmail.com