© July 2022| IJIRT | Volume 9 Issue 2 | ISSN: 2349-6002

A Survey Paper on Natural Language Processing

Reena Sahu[1], Mr. Kranti Kumar Dewangan[2]
[1,2]*Department of Computer Science and Engineering, Rawatpura Sarkar University, Raipur (C.G)*

***Abstract* - Processing Natural Language such as English has always been one of the central research issues of Artificial Intelligence, both because of the key-role language plays in human intelligence & because of the wealth of potential applications. Natural Language processing techniques can make possible the use of Natural Language to express programming ideas; this in turn increases the accessibility of programming to non-expert users. NLP holds a great promise for making computer interfaces easier to use for people. NLP is used to analyze text, hence allowing machines to understand how humans speak. In this paper, we give an overview of NLP from the scratch. We also briefly discuss some of its major applications.**

***Index Terms* - Natural Language Processing, Machine Learning, Natural Language Understanding.**

INTRODUCTION

Processing of natural language is branch of linguistics, artificial intelligence & computer science and its purpose is to have interaction among natural language of human beings and computers [9]. We can say it is related to field of computer–human interaction. There are different challenges in this field like understanding of natural language i.e. allowing machines to have understanding from natural language of human beings. Mostly available tasks of natural language processing are: analysis of discourse, morphological separation, machine translation, generation and understanding of natural language, recognition of named entities, part of speech tagging, recognition of optical characters, recognition of speech and analysis of sentiments etc. Current research in NLP is showing more interest on learning algorithms which are either unsupervised or semi-supervised in nature. These techniques of learning can perform this task of learning from data which is not annotated manually with required answers or by applying mixture of non-annotated & annotated data. Normally, this job is very hard as compared to learning which is supervised & usually shows little correct results for particular amount of data as input. But there is large quantity of data is available which is non annotated in nature i.e. whole contents available on world wide web and it normally produces less accurate results.

NLP is the field of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP can be defined as the automatic processing of human language. Alternate terms that are often preferred are, ‘Language Technology’ or ‘Language Engineering’. It is closely related to linguistics. To summarize, NLP is a discipline which is concerned with the interaction between natural human languages and computing devices. It is a way in which computers analyze, understand and derive meaning from human language in a smart and useful way. This human-computer interaction enables real-world applications like Search engines (Google), Translation systems (Google Translate), automated question answering, Text categorization, Spelling and Grammar checking and more. NLP is integrated widely in a large number of contexts such as evaluation systems, e-learning, research, machine translation, multilingual and cross-language information retrieval (CLIR), speech recognition.

TECHNIQUES OF NATURAL LANGUAGE PROCESSING

Machine Translation

It is the process of translation [5] of text automatically from any human language to other human language. It is very hard problem & is and associated with problems named as AI-complete. For completely solving the problems of translation, It demands to possess different kinds of knowledge which humans beings have (i.e. Knowledge of semantics, grammar and concepts related to real world).

Analysis of Discourse

The task of discourse analysis has a many related jobs to do. One such job is determining structure of discourse of text which is connected, that is kind of relationships of discourse among lines like: contrast and explanation. One more job is identifying & categorizing acts of speech in the particular text. For example content question , yes/no questions, assertion and statement etc.

Morphological Splitting

Split terms to separate morphemes & recognize category of corresponding morphemes. Main problem in this job is that it relies largely upon complication of term structures in that language which we are considering. There is very simple morphology of English language, particularly in case of morphology related to inflection & hence it is usually feasible of ignoring the job completely & normally makes different feasible forms of any term. For example treating opened, opens and opening as different terms. But in case of Turkish language, this technique is not feasible because every entry in Turkish dictionary can have large number of feasible forms of a word.

Generation and Understanding of Natural Language

Generation of natural language involves translate information into easily readable language of beings from computerized databases. Understanding of natural language involves changing text sections to much formal notations like structures related to logic in first order which are very easier to manipulate by programs. Moreover it deals with recognition of semantics using many feasible semantics obtained using expressions of natural language that is normally in form of organized notations present in concepts of natural languages. Generation of ontology & meta-model in language are very suitable solutions and are empirical in nature. Formalization of semantics of natural languages by making assumptions like assumption of closed term vs assumptions of open term, assumptions of objective vs subjective in absence of confusions is required in generation of formalizations of semantics.

Identification of Named Entities

With input text, identify terms can be labeled as named entities like places names, people names & also to identify to which types these named entities belong for example organization, location or person. Capitalization is although helpful for identifying the names present in languages like English, but it is not helpful in identifying type of names. Moreover capitalization is not the sufficient criteria for identifying the names because, 1st character in a line is also written in capital case and also these names usually span many terms any few of them are written as capitalized. Moreover many languages which are non western like Hindi, Arabic, Punjabi and Chinese etc. are not possessing feature of capitalization. Also many languages having capitalization feature can not throughout apply it for identifying the names e.g. In German language all noun terms are capitalized irrespective of if they point to names or not, & Spanish and French also not use capitalization for names which treat like adjectives.

Marking Part of Speech

Input a line, identify and mark part of speech [3] in case of every term. Many terms, usually common terms may be treated with more than one parts of speech e.g. a term book might be treated as noun or can be treated as verb. Another term set might be treated as verb, noun or adjective. Also many languages can show much of this type of ambiguity. English Language having very less inflectional morphology is very much prone to this ambiguity. Chinese language also show this ambiguity as this language is of type a tonal language while performing verbalization.

Optical Character Recognition

Optical character recognition deals with identifying text from images denoting text in printed form. Success of a OCR for any language depends on quality of images denoting the printed text in same language.

Recognition of Speech

Recognition of speech deals with recognizing textual notation of any speech by listening to sound clip of any person. It is very hard problem and is entirely opposite to the task of text to speech conversion. Moreover in case of any natural speech there will be very less number of pauses among consecutive terms & we can say that segmentation of speech is important sub-step of recognition of speech. In many spoken languages, the utterance of sounds denoting consecutive words mix with each other and this process is called as co articulation that is why changing analog signal of sound into discrete textual characters is very difficult job.

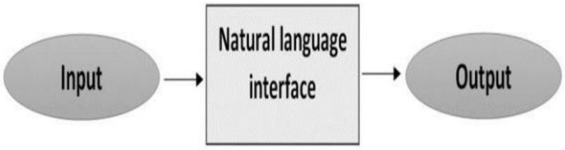
Analysis of Sentiments

Recognition Sentiment analysis [2] deals with retrieving information subjective in nature normally from collection of text documents like online reviews for finding polarity of particular objects. This process is very much applied for determining It is very much applied in marketing for determining sentiments or reviews of public opinion about social media.

NLP System Architecture

NLP system includes:

- User input
- natural language interface
- Output obtained in a language that is understood by the application program



Pragmatic analysis: it involves deriving those aspects of language which require real world knowledge. During this, what was said is re-interpreted on what it actually meant.

The input and output of an NLP system can be of two types:

Components of NLP

There are two components of NLP.

Natural Language Understanding (NLU): It involves mapping the input into useful representations and analyzing different aspects of the language.

Natural Language Generation (NLG): It involves producing meaningful sentences in natural language form from the representations.

NLP Terminology

Phonology: Study of how sounds are organized and used in natural languages.

Morphology: Study of words, how they are formed and their relationship to other words in the same language.

Syntax: Arrangement of words and phrases to create well-formed sentences in a language.

Semantics: Study of meanings of words and phrases in a language. Has two main areas: lexical semantics and logical semantics

Pragmatics: Deals with using and understanding sentences in different situations and how the interpretation of sentence is affected.

Tools for NLP

Previously, no one but specialists could be a piece of natural language processing ventures that necessary unrivaled information on arithmetic, AI, and linguistics. Presently, engineers can utilize instant apparatuses that streamline content preprocessing with the goal that they can focus on building AI models. There are numerous instruments and libraries made to take care of NLP issues. Peruse on to learn progressively 8 astonishing Python Natural Language Processing libraries that have throughout the years helped us convey quality tasks to our customers.

Natural Language Toolkit (NLTK)

Link: https://www.nltk.org/

NLTK is a basic library underpins errands, for example, classification, stemming, labeling, parsing, semantic thinking, and tokenization in Python. It's essentially your primary apparatus for regular language handling and AI. Today it fills in as an instructive establishment for Python designers who are plunging their toes in this field (and AI). The library was created by Steven Bird and Edward Loper at the University of Pennsylvania and assumed a key job in advancement NLP explore. Numerous colleges around the world presently use NLTK, Python libraries, and different devices in their courses. This library is really flexible, however we should concede that it's likewise very hard to use for Natural Language Processing with Python. NLTK can be somewhat moderate and doesn't coordinate the requests of snappy paced creation use.

TextBlob

Link: https://textblob.readthedocs.io/en/dev/

TextBlob is an unquestionable requirement for designers who are beginning their excursion with NLP in Python and need to take advantage of their first experience with NLTK. It fundamentally gives tenderfoots a simple interface to assist them with learning most essential NLP undertakings like sentiment analysis, pos-labeling, or thing phrase extraction.

spaCy
Link: https://spacy.io/

spaCy is a generally youthful library was intended for creation use. That is the reason it's a great deal more open than other Python NLP libraries like NLTK. spaCy offers the quickest syntactic parser accessible available today. In addition, since the toolbox is written in Cython, it's likewise extremely rapid and productive.

Rapid Miner
Link: https://rapidminer.com/

RapidMiner is an information science programming stage created by the organization of a similar name that gives an incorporated domain to information arrangement, AI, profound learning, content mining, and prescient investigation. It is utilized for business and business applications just as for explore, instruction, preparing, quick prototyping, and application advancement and supports all means of the AI procedure including information arrangement, results perception, model approval and optimization. RapidMiner is created on an open center model. The RapidMiner Studio Free Edition, which is constrained to 1 legitimate processor and 10,000 information columns, is accessible under the AGPL license, by relying upon different non-open-source segments. Business evaluating begins at $5,000 and is accessible from the engineer.

CONCLUSION

Processing of natural language is branch of linguistics, artificial intelligence & computer science and its purpose is to have interaction among natural language of human beings and computers. We can say it is related to field of computer–human interaction. Mostly available tasks of natural language processing are: analysis of discourse, morphological separation, machine translation, generation and understanding of natural language, recognition of named entities, part of speech tagging, recognition of optical characters, recognition of speech and analysis of sentiments etc. Current research in NLP is showing more interest on learning algorithms which are either unsupervised or semi-supervised in nature.

In this paper, we have explored all the aspects of Natural language processing. The introduction part gave the decent introduction of NLP and its history. Then various applications of NLP has also discussed with proper description. NLP is very vast field to explore but we have shown essential applications in software market. There are some tools which are useful for implementations and research-oriented work.

REFERECNE

[1] Ms. Rijuka pathak et al. Natural language processing approaches, application and limitations. In Natural language processing approaches, application and limitations (Vol. 1 Issue 7, September – 2012) Julia Hirschberg et al. Advances in natural language processing. In SCIENCE (17 JULY 2015 • VOL 349 ISSUE 6245)

[2] Anjali M K1 et al. Ambiguities in Natural Language Processing. In International Journal of Innovative Research in Computer and Communication Engineering (Vol.2, Special Issue 5, October 2014)

[3] http://www.expertsystem.com/natural-language-processing-applications/

[4] http://language.worldofcomputing.net/understanding/applicati ons-of-natural-language-understanding.html

[5] Diksha Khurana,” Natural Language Processing: State of The Art, Current Trends and Challenges”,2017.

[6] R. Kibble,” Introduction of Natural language”, subject guide by university of london,2013.

[7] Sudhir K Mishra,” Artificial intelligence and Natural language processing”, book, 2018.

[8] Yasir Ali Solangi,”Review on Natural language processing and its toolkits for opinion mining & sentiment Analysis”, International conference on technologies & applied science, 2018.