International Journal of Computer Science and Information Technology for Education
Vol. 4, No. 2 (2019), pp.25-32
http://dx.doi.org/10.21742/ijcsite.2019.4.2.04

A Detailed Review on The Methods to Process an Application in Natural Language Processing

Debnath Bhattacharyya[1], N. Thirupathi Rao[2]

[1,2]*Department of Computer Science and Engineering, Vignan's Institute of Information Technology (A), Visakhapatnam 530049, AP, India.*
[1]*debnathb@gmail.com,* [2]*nakkathiru@gmail.com*

Abstract

Natural language processing is one of the most important applications and working methods in the recent days of research. The current method of processing is having a various number of applications and techniques that can be used such that to process the applications. The utilization of these methods and techniques had increased a lot in recent years. Several new techniques and methods had introduced in recent years for the development of this technology and some of them are getting famous day by day by using them. The methods that were being used in the current day of technology were most important and it can be sued for other applications also. In the process of implementing the various applications and that may be either new or existing applications, the user can write a vast number of applications and the users had started using such technologies with the help of these methods. Hence, in the current article an attempt has been made to give a brief note on the various methods that can be sued for the processing of an application under natural language processing methods.

Keywords: *Natural Language Processing, Stemming, Lemma, Tokens, normalization, stop words, semantic analysis, sentiment analysis.*

1. Introduction

Natural language processing is considered as the capacity of a machine or a computer to understand the language or the words being used by a person in a particular language while speaking. For implementing or processing such big tasks, several devices are required for storing the data, processing of data and for further processing of data with a computer machine. For the processing of any language, the processing of natural language processing applications are key factors. The natural language processing and its related topics and its sub applications are always inked with the Artificial Intelligence and its related topics. The processing of such applications with languages is very important in various aspects. Collecting the data from the various sources for the machines to understand the words which were pronounced by the humans and the voice levels which were being generated by these people is also important. Syntax analysis of words and languages and semantic analysis of languages are the two important processing tasks of any NLP application. [1]

The other methods that can be used are the collection of words that were spoken by people in any language to be identified and the segmentation of those words for easy identification and the other methods are to identify such words by using the current morphological methods. The other approaches or the methods to be used are the stemming, segmentation, morphological

Article history:
Received (August 10, 2019), Review Result (October 3, 2019), Accepted (November 6, 2019)

Print ISSN: 2205-8370, eISSN: 2207-5372 IJCSITE

methods, language generation, entity identification etc,. Some of the areas that were being used for these methods of identification are the applications of deep learning, neural networks machine learning and other rule based systems of artificial intelligence. Some of the tools available for the processing of these tasks or performance of these tasks by using tools like the NLTK stands for Natural language tool kit, Intel NLP Architect and Gensim etc,. All these tools are available in the internet for the public to use. The people can download the software's from various sources and can use them. Some of these available software's are paid software's and some other tools are freely available tools. The people who can need of urgent usage can download these tools and can use them.

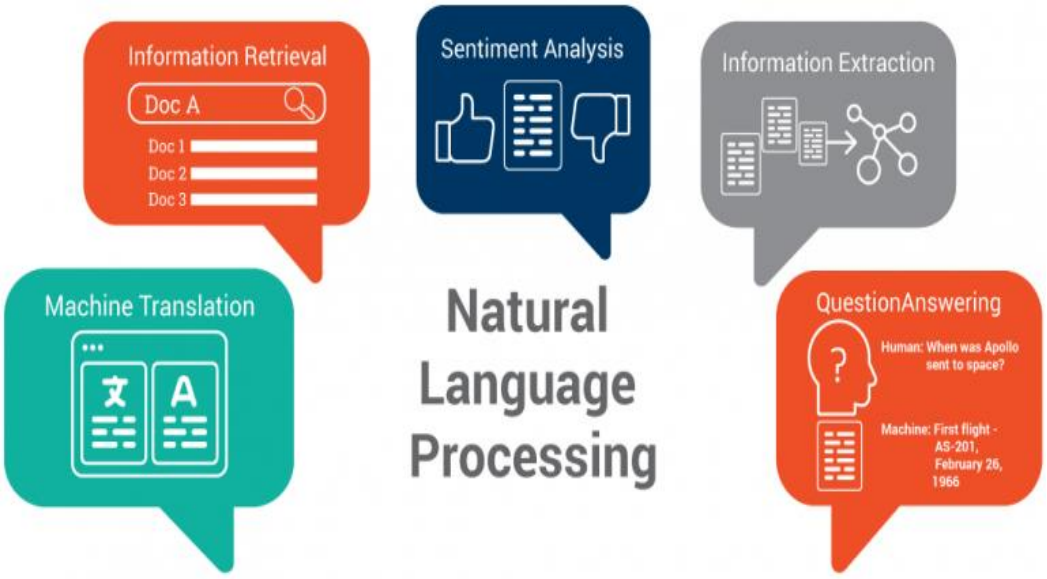

Figure.1 An Applications of NLP[2]

Now a days, the prominent and very interesting research topics, areas and groups of researchers are working on this area for improvement in the standards of the utilization of these applications. For all these research groups, the main task were to address and refine the search process more interestingly and more sophisticated. The methods to be developed and remodified in such a way to get more accurate results and improved results when the system and the people can be benefitted more and more. The major topic or the task all these researchers are being concentrated was the enterprise search. A group of people are allowed to raise some questions related to a common topic or a different topic and all those topics are to be formed in a single group such that to process the application and can find the results more interestingly and more accurately as per the expectations of the users. Some of the best examples are the patient's data where some huge amount data can be stored at different hospitals at various locations in the country. Collecting all these data formats and making a common format and implementing the same methods or the different methods to achieve the same is also an important task and challenge for the public and also for the researchers. In order to process such huge data, the other method or the other mode of operation can be considered is the sentiment analysis. In this method a group of people with their interests and ideas can be considered and all those ideas and interests can be considered to form as a group that can be

processed easily for the researchers. Based on the likes and dislikes of the people with different locations, different age groups and different interests might use for the analysis methods.

Some of the methods to be used in natural language processing to classify and normalize the text contents and data are as follows,

2. TOKENIZATION

One of the early steps in NLP is the tokenization. This process stands for the splitting of a word or a sentence into parts based on some logic as tokens. This process can be achieved as one word can be considered as a token in a sentence and some sentences can be considered as tokens in a paragraph for better and easy processing.

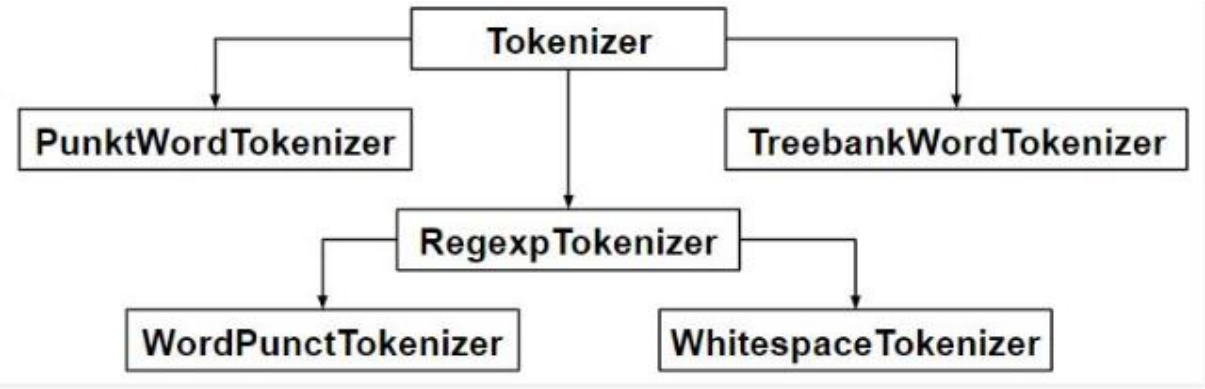

Figure.2 Tokenization process example 错误!未找到引用源。

In the process of tokenizing the words or the sentences present in a an application the major question arrives at that juncture was which tokens to be considered as important to use and which tokens to be considered as no important or not use in the sentences or paragraphs.

3. NORMALIZATION

In order to process the data or the sentences more easily and more accurately, the most useful process that can be implemented was the normalization. In this process a list of words are selected and those words are converted to a more uniform model of sentences and sequences. These models or the data that was processed by using these sorts of techniques will result us by using such data in the next level or later data processing units. By implementing this sort of conversions, the processing of data can be done more sophisticated and easier. If the data t o be processed is converted to a single format that is like either in upper case of letters for entire document or lower case of letter for the entire document. By using such methods, the entire document can be made easy to search any operation or any function for further processing..[7]

Table 1. An example of Normalization 错误!未找到引用源。

Generator	From	To
Leave intact	Good	Good
Edit distance	Bac	Back
Lowercase	Need	Need
Capitalize	It	It
Google spell	Disspaear	Disappear

Contraction	Wouldn’t	Would not
Slang language	Ima	Am going to
Insert punctuation	ε	.
Duplicated punctuation	!?	!
Delete filler	Lmao	ε

The normalization process can improve text matching. For example, there are several ways that the term "modem router" can be expressed, such as modem and router, modem & router, modem/router, and modem-router. By normalizing these words to the common form, it makes it easier to supply the right information to a shopper. Understand that the normalization process might also compromise an NLP task. Converting to lowercase letters can decrease the reliability of searches when the case is important.

4. STEMMING

Whenever there is a need of identifying the similar words or similar set of objects to be identified for the given words to search in a search bar, the methods that can be used mostly are the stemming and lemmatization in natural language processing. If the users or the people want to identify the similar set of objects or the search items in the internet whenever the users want to identify the similar objects, these methods can be used based on the logic or the requirement of the users. These operations and tasks can be successfully achieved by using these two methods. Two methods will work on two different concepts. As the working of these two methods are different, the results that the people may arise also different.

Table 2. Example of stemming process[5]

Form	Suffix	Stem
Studies	-es	Studi
Studying	-ing	Study
Ninas	-as	Nin
Ninez	-ez	nin

The stemming algorithm functioning was different when compared to the lemma algorithm. This algorithm works on by cutting the beginning letters of the words or at the letters of the ending of the word by considering the suffixes and prefixes of the words to be processed. By applying this cutting methods may give results in a better way in some situations and some other times they may not give better results.

5. LEMMATIZATION

This process is different when compared to the other process of stemming in processing the applications. In this process, the morphological relations of the words are considered and processed for the better results. For performing of such tasks the algorithms and applications needs many tasks and requirements for the processing of these tasks. Several databases are required for the processing of such tasks.

Table 3. Examples of Lemma process [5]

Form	Morphological Information	Lemma
Studies	Third person, present tense of the verb, singular in number	Study
Studying	Gerund of the verb study	Study
Ninas	Feminine gender, plural number fo the noun	Nino
Ninez	Singular number of the noun ninez	ninez

The words and their meanings should be treated carefully for the further processing and better understanding of the entire process.

6. STOP WORDS

Whenever a particular data is being prepared to process with the help of a computer, the major important consideration to be considered was to create the data amore clean and neat. In order to make the data more clean and neat, the unwanted words or the contents or special characters in the data to be removed from time to time such that to make the machine to process the selected data more easily and fastly and results can be observed from such applications was good for the easy processing and easy analysis.

Table 4. An example for stop words[6]

Sample text with stop words	Without stop words
GeeksforGeeks-A computer science portal for Geeks	Geeksfor Geeks, Computer Science, Portals, Geeks
Can Listening be exhausting?	Listening, Exhausting
I like reading, so I read?	Like, Reading, Read

These unwanted words in the data can be known as stop words. If these stop words cannot be removed from the data, the processing of data for the machine was very much problem.

7. SYNTACTIC ANALYSIS

The syntactic analysis can also is named as parsing. Among various steps in the processing of natural language processing, it is the third phase of the total method to be considered.

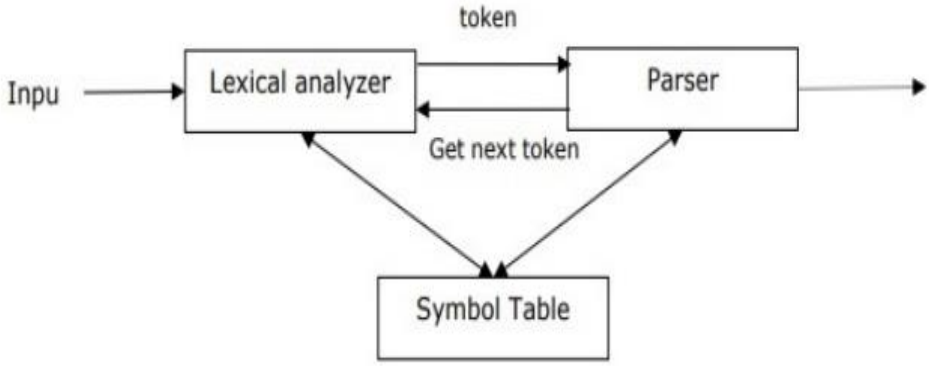

Figure 3. Syntactic analysis example

The current method is mainly use for identifying the meaning of a word either from the text or from a phrase. The syntactic analysis can also be named as the syntax analysis.

8. SEMANTIC ANALYSIS

It is the process of verifying whether the declarations done in a document or in a statement and statements given in a program or a source code of lines. The meaning of the statements and the lines used in the document are to be verified and studied by using this method of processing. The working of the current method involves three major processes. They are the Label checking, type checking and flow control checks. The labels existing in the document to be verified and needs to done by the label checking method. Type checking will be done by using various data types and their related topics to be verified by using the current type of the documents. The flow of the program and the related items flow in the execution and processing the applications to be verified by suign these sorts of applications.

9. SENTIMENT ANALYSIS

One of the most interesting and fast developing technology and methods related to the business type of applications was the sentiment analysis. The major concern in the area of the current model was the identification of various methods and utilization of various techniques such that to identify the applications of various customers data and trying to identify the interests and willingness of the customers for choosing various products in the market today. As the competition in the business market today are growing in a rapid fast day by day, the companies and the industries are looking for various techniques and methods with the help f latest technologies such that to identify the customers with similar requirements and similar tastes such that to sell and supply their products. The other major source for the people to use the technology was the utilization of various social networking websites like the Facebook, Twitter and other set of sources. More number of people is located and using such network groups and the people are using and expressing their interests and non-interests from which the users or the technical teams can identify the people's interest and non-interests. As a result, the companies can easily reach to the customers whenever they need to purchase some products.

The development of the recent technologies like the machine learning and deep learning techniques, the utilization and observing the good results and good outputs had became more and more famous and productive. Proper better utilization of these advanced techniques will

give us more and better results for the better execution and for better marketing strategies and good results.

10. CONCLUSION

In the current article an attempt has been made to give a brief note on the natural language processing and its applications and the methods to be implemented in NLP to achieve better results and better output from the set of people. The article will be better helpful for the new readers and the people who can try to learn about the basics of natural language processing and its working methods. Some methods had given in brief and some are given as an example with better models to easy understanding of the readers.[8][9]

References

[1] https://www.kdnuggets.com/2017/02/natural-language-processing-key-terms-explained.html

[2] Srishti Sawla, "Introduction to Natural Language Processing", **(2018)**.

[3] GeeksforGeeks, “NLP, How tokenizing text, sentence, words works” https://www.geeksforgeeks.org/nlp-how-tokenizing-text-sentence-words-works/ **(2019)**

[4] Congle Zhang et. Al., "Adaptive Parser-Centric Text Normalization", in ACL, **(2013)**.

[5] Bitext, “What is the difference between stemming and lemmatization?” https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/ **(2018)**

[6] https://www.geeksforgeeks.org/removing-stop-words-nltk-python/**(2019)**

[7] A. Gelbukh et. Al., "Natural language processing", Fifth International Conference on Hybrid Intelligent Systems (HIS'05), 6-9 Nov. **(2005)**, DOI: 10.1109/ICHIS.2005.79.

[8] Partha Mukherj et.al., "Development of GUI for Text-to-Speech Recognition using Natural Language Processing" 2018 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), 4-5 May **(2018)**, DOI: 10.1109/IEMENTECH.2018.8465238.

[9] Rusian Posivikin, "Translation of natural language queries to structured data sources", 2015 9th International Conference on Application of Information and Communication Technologies (AICT), 14-16 Oct. **(2015)**, DOI: 10.1109/ICAICT.2015.7338516.

