IJISEM
INTERNATIONAL JOURNAL OF INNOVATIONS IN SCIENCE ENGINEERING AND MANAGEMENT

A systematic Survey of Natural Language Processing (NLP) Models & it's Application

Nikita Saxena

Research Scholar, RITS, Bhopal, M.P., India

OPEN ACCESS

Manuscript ID:

AG-2022-1005

Volume: 1

Issue: 1

Month: September

Year: 2022

P-ISSN: 2321-XXXX

E-ISSN: 2582-XXXX

Received: 29.08.2022

Accepted: 07.09.2022

Published: 13.09.2022

Citation:

Nikita Saxena. "A systematic Survey of Natural Language Processing (NLP) Models & it's Application." International Journal of Innovations In Science Engineering And Management, vol. 1, no. 1, 2022, pp. 25–31.

Abstract

Computers may be used to analyse texts using "Natural Language Processing" (NLP). NLP is concerned with learning about the ways in which people perceive and express themselves via language. Computer systems can interpret and modify the natural languages to execute a variety of desired activities if the right tools and procedures are developed. This paper will introduce you with the natural processing language and it brief about the Different models of NLP, it also explores the applications of NLP.

Keyword: Natural Language Processing, Application of NLP, NLP models

Introduction

PCs may be used to comprehend and manipulate the natural language text or voice in order to accomplish useful things in the field of Natural Language Processing. This research is aimed at gaining insights into human language comprehension and usage in order to provide adequate tools and frameworks for computer systems which can interpret as well as control the natural languages to carry out the needed activities. It has also become more sophisticated over time to comprehend and analyse natural language material. Learning acquisition, data retrieval, and the language translation are all the areas where NLP breakthroughs are becoming more important in the building of user-friendly decision-support frameworks for ordinary non-expert users. For this reexamination, we're looking at the present situation and potential implications of NLP breakthroughs and also the frameworks in the workplace.

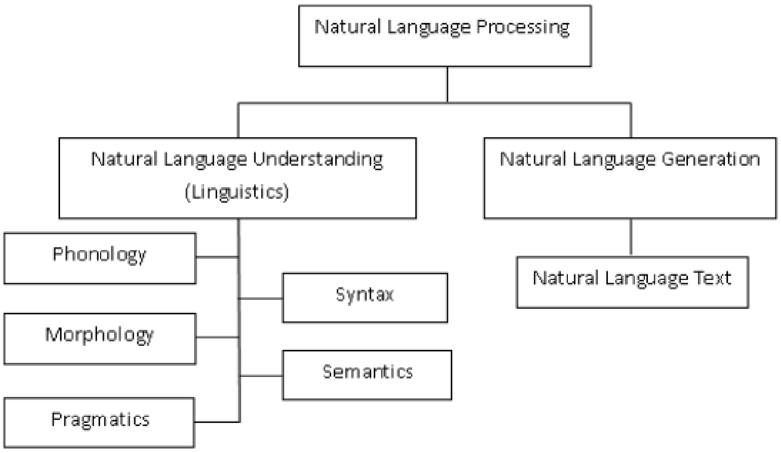
Natural Language Processing
Natural Language Understanding (Linguistics)
Natural Language Generation
Phonology
Morphology
Pragmatics
Syntax
Semantics
Natural Language Text

Figure 1: Broad Classification of NLP

The field of Artificial Intelligence as well as Linguistics known as Natural Language Processing focuses on teaching computers how to read and interpret the human language.

NLP was created to make the user's job easier and to meet their need to interact with the computer using the natural language. Because not every user is fluent in machine-specific language, NLP is designed for those users who lack the time or inclination to acquire new languages or improve their proficiency in their current ones. Rules and symbols may be used to define the language. Symbols are a mashup of letters, numbers, and other visual representations that may be used to communicate ideas. The Rules have sway over Symbols. It is possible to divide Natural Language Processing (NLP) into 2 parts: Natural Language Understanding (NLU) as well as Natural Language Generation (NLG) (Figure 1).

Goal

Composing computer models of the natural language in order to study and generate it is goal of the natural language processing. Machine translation frameworks, the natural language interfaces to the databases, human-machine interfaces to computers, voice understanding systems as well as other text investigation and also understanding systems are all examples of mechanically inspired PC frameworks. Secondly, there is the cognitive and etymological motivation to obtain a deeper understanding of how people interact with each other in natural language. Programming that analyses, understands, and creates languages which people use naturally is the goal of the Natural Language Processing. This means that one day you would be able to speak to your computer as if it were the another person.

Models and Approaches to the challenges:

LSTM

In computing, LSTM stands for the long short-term memory. The LSTM model is built on the recurrent neural network. Values are stored in the memory of the Recurrent Neural Network (RNN). Recurrent neural networks that can recall and anticipate the outcome of past input across arbitrary time intervals are referred to as long-term memory. It is used to train the machine by providing it with data. It is among the most widely used machine learning models in the field of NLP. As the learning process continues, the stored values remain unchanged. It is unable to alter input sets in LSTM model, but model is capable of learning from it by calculating its frequency as per the event by processing it multiple times. Data input is flushed out of network during first phase of LSTM. The forget gate layer "0" symbolizes "totally forget" and the "absolutely keep" represents 1 is what determines this. In the next phase, the input gate layer determines what to store in the cell state. The tanh layer, which comes after that, generates a fresh set of the candidate values, which are then used as inputs to state update. This "peephole" link allows gates to verify cell states before dumping data from networks, which is a feature of several LSTM models.

Seq2Seq model

The tradition seq2seq model contains two recurrent neural network i.e. encoder network and decoder network.

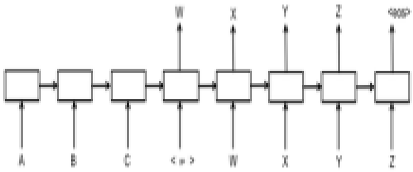

Figure 2 Recurrent neural network structure

Every box represents an RNN cell, the most popular implementation of which being the LSTM algorithm. Each input is encoded into the fixed-size vector in this approach, and the decoder is used to decode the vector later on.

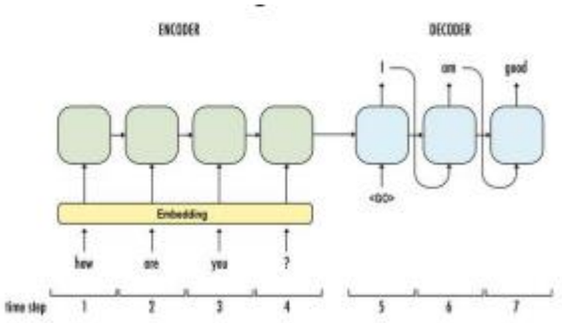

Figure 3 Encoder Decoder structure and working

In the beginning, the model is trained to recognize the right grammar syntax by building vocabulary list via embedding. The vocabulary collection is analyzed to find and categorize terms that appear often, seldom, and only in this context. As an alternative, the words are substituted with ids. A reply suggestion is encoded and produced depending on ids provided. During the input compilation process, the model makes use of the following tags:

EOS: End of sentence.

GO: Start decoding.

PAD: Filler.

UNK: Unknown; word not in vocabulary.

An illustration of how this model works is as follows:

Question: How are you?

Answer: I am fine. This pair has been repurposed as:

Question: "[PAD, PAD, PAD, PAD, PAD, PAD, "?", "you", "are", "How"]"

Answer: "[GO, "I", "am", "fine", ".", EOS, PAD, PAD, PAD, PAD]."

Named entity recognition Model

In order to detect relevant names and categories names according to entity they belong to, named entity recognition is utilised. Input data may be in text or audio form, but the NER model is able to identify names, locations, persons, and the other important elements. There are two stages to the NER model's operation. An NER model begins by breaking up the text into sections or chunks and classifying them. In order to categories these pieces, we need tokens like name of the person, organisation or place. Bolding as well as capitalization are omitted from formatting.

User preference graph

Create a list of options for the user using the user preference graph. This graph is formed whenever the user uses the same adjectives, tenses, conjunctions, as well as prepositions in the same phrases over and over again. When this happens, the model offers next words by assessing their chance of being used. As a result, the preference graph is built for every user based on the mapping of these terms. Users that have the similar user preferences are the grouped together in order to provide a broad range and diversity of recommendations in large scale implementations of this paradigm. The smart reply, typing suggestions, and auto-respond systems may all benefit from this.

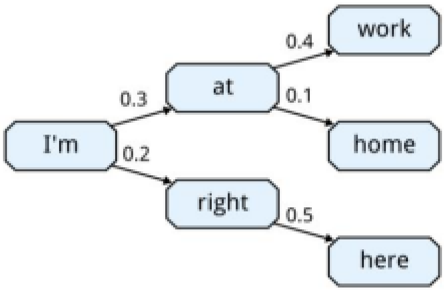

Figure 4 User preference graph example

Word Embedding:

When words and the phrases in the natural language are mapped into vectors of the real number of the preference graph, this technique is known as the word embedding.

Phrase Based Machine Translation:

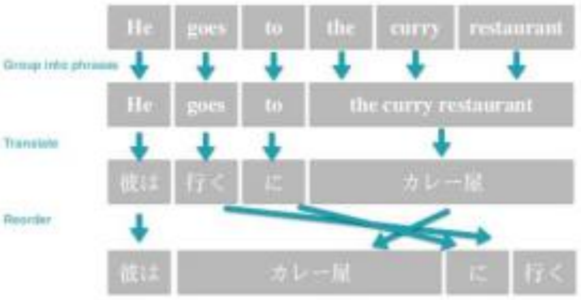

Figure 5 Phrase-based translation model

Statistical Machine Translation (SMT) includes PBMT as one of its methods. Text is translated via the use of predictive modelling. Bilingual unstructured texts are used to generate or learn such models. The most likely outcome is produced with their support.

According to this algorithm, PBMT works like this:

1. Dissecting the Original Sentence into Tiny Chunks
2. All translations for every chunk are found using the corpora, which tell us how people have interpreted a certain piece of writing in the context of everyday speech.
3. Find the most probable sentence by generating as many as possible: We may produce more than 5000 distinct phrases by combining translations in step 2).
4. By comparing it to training set, you may estimate the score. Text from many books, journals, newspapers, and other sources makes up the instructional materials in this instance. We provide the probability score to every other combination of the aforementioned steps by comparing it to training dataset . We'll choose the one which has the most plausible chunk translation as well as the high probability score after experimenting with other combinations and running them through the training data set.

Because PBMT is so complex to construct and maintain, it has a drawback. An additional language must only be added if the bilingual corpora of that language are already available. Translation sacrifices are made for less well-known language pairs. If Gujarati to Georgian translation doesn't really involve a complicated pipeline. To avoid this, it may first translate it to English and subsequently to Georgian.

Neural Machine Translation (NMT):

The newest approach to Machine Translation is neural. In comparison to Statistical Machine Translation, it produces substantially more accurate translations. The input is sent through many "layers" of processing before being produced via Neural Machine Translation. Learning language rules on its own from the Statistical Models is possible thanks to NMT's ability to employ an algorithm. Using attentional encoders and decoders, the NMT system processes sub word units. Back-translations of monolingual News corpus are employed as extra training data in order to further enhance the accuracy of the system. It is ideal for the translations in both directions. Among the benefits of NMT are its prowess in dealing with various verb order patterns and its ability to completely eliminate verb omissions. English nouns are supported. NMT also has a good grasp of the structure of sentences and the use of articles. NMT's drawbacks are confusing translations into German. In addition, there is a problem with generating continuous tenses. Prepositions provide the most difficult challenge for NMT.

Automatic text summarization using the template-based method.

Feature-based algorithms evaluates the content of the sentences depending on a set of predefined criteria or features, as well as the sentences which are found in input data are added as they are in output summary, whereas the template-based extraction algorithms modify the content of the extracted text to make it more grammatical and human-sounding, as shown in the examples. Two stages are involved in the implementation of template-based technique for automated text summarization.

A. Text pre-processing
B. Information extraction

A. Text pre-processing

Included in this phase of implementation is module seen in the following image:

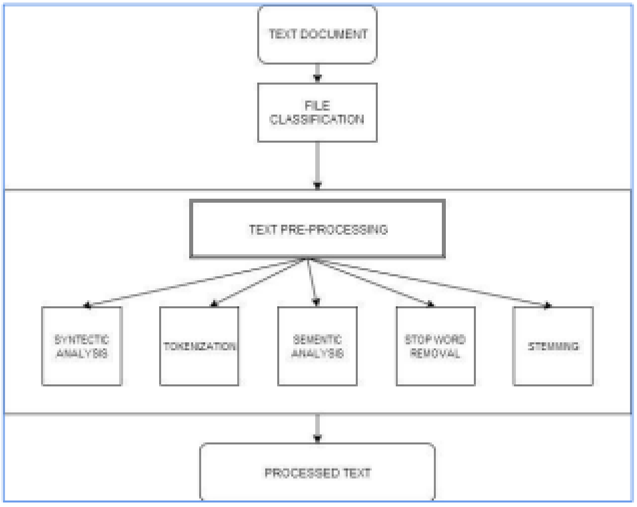

Figure 6 Text preprocessing model

1. **Syntactic analysis**

An input document's beginning as well as ending points are determined by syntactic analysis module. The full stop sign is currently used as the end of the sentence by the algorithm. That includes any sequence of characters up to as well as including the full stop.

2. **Tokenizer**

It is the task of tokenizer to split syntactic analysis result into the tokens. Words, numerals, and punctuation marks are all examples of sentence fragments.

3. **Semantic**

Every word in the sentence has specific function in “semantic analysis subphase". Then, each word that is the noun, adjective, verb, etc. is assigned a tag. Part-of-Speech tagging, or POS tagging, is an act of allocating and separating the word into multiple classifications.

4. **Stop Word Removal**

Even if they appear more often in a natural language text, certain words have a low impact when it comes to determining the overall meaning of a phrase. For this reason, the use of these terms is referred to as "stop words.

5. **Stemming**

An input text document's word stemming job is to determine the most basic form of that word. To prevent this, stemming is done as well as such words with the very same meaning but various tenses are turned into the fundamental simple tense, thereby avoiding any confusion.

B. Information extraction

Modules in this section of algorithm include:

1. Training the dialogue control
2. Knowledge based discovery
3. Dialogue management
4. Template based summarization

a) Training the dialogue control

For example, while a system is being trained, it learns the names of people and locations as well as rules that govern their use. Every training set improves the algorithm's intelligence as well as efficiency. To put it another way, every time we feed algorithm a piece of data, it saves results in the data storage and afterwards utilizes those findings to guide its subsequent analysis of new data. The system's knowledge base contains the ideas learned during training.

b) Knowledge based discovery

Extracting intelligent information as well as storing it in the unstructured text form is known as the knowledge discovery here. Thus, lowering the need to establish various storage structures to hold distinct words of different categories, therefore reducing search time and therefore enhancing the overall speed of algorithm.

c) Dialogue management

The conversation management module is indeed a tool that facilitates communication between people and computers. Natural language text may be used as a means of asking for the information utilising this module. In training model as experience data set, a dialogue control module takes user request, interprets it and then uses its knowledge base to provide replies that may include the requested information.

1. Template based summarization

Compiling all the relevant text in source data or document into a little package is what text mining is all about. The algorithm also enables the user to create a template with options for specifying locations, events, named entities, and so on. User may also select any number of the POS patterns of their choice.

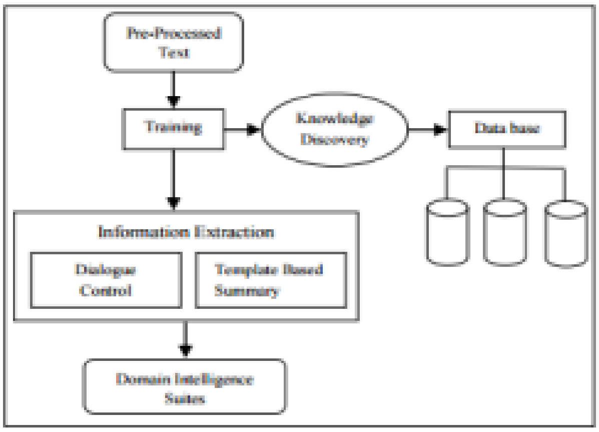

Figure 7 Information Extraction Module

d) Application of NLP

When it comes to NLP, summarization is a common use case. In addition, we'll talk about as well as compare the two of the finest methods for summarizing material in order to reach a conclusion. But first, let's take a closer look at what automated text summarization entails. The goal of the automatic text summarizing is to condense a big volume of text into the meaningful short summary that helps reader to comprehend what information the document includes in a brief descriptive form so that it saves user's effort as well as time.

Automatic text summarization may be accomplished in two major methods. Extraction as well as abstraction are two examples of these. This kind of summarizing relies on a selection of phrases, words, or sentences from the original source to create an extracted summary. The semantic internal representation is built through abstractive approaches, and the natural language generation methods are used to construct meaningful summaries by mimicking the brain function of the computer as the human brain would. To construct a summary that really is closer to what the human would really extract and give as the summary of text, this method is used. Verbal innovations are included in this computed summation. For picture collection summarizing, text summarization as well as video summarization, extractive approaches have been the primary focus of research to date.

Literature Review

(Pandey & Rajput, 2020) The study of how computers as well as human languages interact is known as "natural language processing," a subfield of the computer science and also the artificial intelligence. Mathematical and the computational models are used to analyse many elements of language and construct a broad variety of systems in the natural language processing (NLP). Included in this category are spoken language systems which incorporate speech with natural language. Natural language processing plays an important part in the computer science since the area deals with many elements of computing that have to do with linguistics in some way. Natural language processing is the branch of computer science that studies how computers may be used to comprehend and modify the natural language text or voice in order to accomplish meaningful tasks. Artificial intelligence (AI) and voice recognition are only a few of the many applications for the natural language processing that can be found in a variety of industries like natural language text processing, machine translation, and the user interfaces.

(Chithra & Henila, 2019) Nowadays, the issue of the natural language processing is the hotly debated and actively explored one. One of the very oldest areas of the machine learning research, it is employed in a wide range of applications, including voice and also the text processing as well as machine translation. The study of computing and artificial intelligence has made tremendous strides thanks to advances in the natural language processing. The recurrent neural network is at the heart of many of the algorithms used in the natural language processing. In this review study, many text and audio processing algorithms are explored and their workings are illustrated with illustrations. As can be seen from the results of several algorithms, this discipline has advanced considerably over the last decade or so. As well as attempting to distinguish between distinct algorithms, we've also attempted to determine the potential future study areas for each of them. Several algorithms are discussed in this paper, as well as how they might be used in different scenarios. The area of the natural language processing has yet to achieve perfection, but with constant advancement, it is certain to do so in the near future. Various artificial intelligences currently make use of algorithms for processing natural language to identify and interpret user-supplied voice commands.

(Solangi et al., 2019) Opinion mining has emerged as a primary method for sifting through such vast amounts of data on Internet, and it now accounts for bulk of all online networking. Modern applications may be found in a wide range of fields. There are a number of different pronunciations that might complicate the investigation of a certain topic. Opinion mining has been a vibrant study area in recent years because of research problems. NLP approaches for opinion mining as well as the sentiment analysis are discussed in this work. First, the basics of NLP are discussed, followed by an explanation of some of the most typical and practical pre-processing procedures. This study examines and critique’s opinion mining on a variety of levels.

(Devi & Ponnusamy, 2018) Automated Syntax Grammar Synthesis There is a lengthy history of study in both the languages and computer science, making them established fields. With the help of frameworks for “natural language processing", it is possible to convey concepts using natural language. Natural Language Processing (NLP) in the educational system provides solutions in a variety of different domains relating to social as well as cultural contexts in which language learning is conducted. A wide range of educational linkages, including research, linguistics, science, e-learning, as well as evaluations systems, are incorporated into Natural Language Processing, which adds to beneficial results in places like higher education, schools, and universities. NLP Frameworks may reduce the cost and increase quality of the electronic healthcare systems in the healthcare industry. This study discusses NLP approaches, their usefulness in, healthcare, education and their applications and also their limits, on the basis of this foundation. It is the purpose of this review to summarise and also report on the existing and future status of NLP advancements in various business settings.

(Khurana et al., 2017) Natural language processing (NLP) has lately attracted great interest for representing as well as analysing the human language computer. It has broadened its applications in numerous domains like email spam detection, machine translation, summarization, information extraction, medical, as well as the question answering etc. The study divides four parts by examining distinct levels of NLP as well as components of the Natural Language Generation followed by providing the history and development of NLP, state of art presenting the many applications of NLP and also the current trends and difficulties.

(Litman, 2016) Advances in the natural language processing (NLP) and also the educational technology, and also an availability of unprecedented volumes of the educationally-relevant text and voice data, have led to an increased interest in employing NLP to serve the requirements of instructors and also students. Educational applications vary in many respects, however, from the sorts of applications for which the NLP systems are normally created. This paper shall organise and present an outline of research in this field, concentrating on potential as well as problems.

Conclusion

When it comes to information technology, natural language processing (NLP) is the relatively new field of study and application, but there have indeed been enough achievements thus far to indicate that it would continue to be an important focus for researchers for years to come. Text processing methods rely on categorization as well as preference networks depending on entities, as the preceding context indicates. Text processing algorithms typically used in many applications to decrease the user's burden as well as time and provide relevant and also efficient output. Wise reply and wise recommendations are two examples. While the challenge of speech processing is far from being solved, it has made significant progress over the last decade.

References

Chithra, P., & Henila, M. (2019). International Journal of Computer Sciences and Engineering Open Access. *International Journal of Computer Sciences and Engineering*, *6*(10), 628–632. https://doi.org/10.26438/ijcse/v6i1.161167

Devi, N. V., & Ponnusamy, R. (2018). *A Systematic Survey of Natural Language Processing (NLP) Approaches in International Journal of Computer Sciences and Engineering Open Access A Systematic Survey of Natural Language Processing (NLP) Approaches in Different Systems*. *January*.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2017). Natural Language Processing : State of The Art , Current Trends and Challenges Department of Computer Science and Engineering Accendere Knowledge Management Services Pvt . Ltd ., India Abstract. *Sentiment Analysis Has Become One of the Most Profound Research Areas with the Increasing Growth of Social Media on Web. Nowadays, Millions of Users Exchange Their Views, Ideas, Expressions, Feelings, Opinions on Social Media like Twitter and Facebook. Se*, *Figure 1*.

Litman, D. (2016). Natural language processing for enhancing teaching and learning. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 4170–4176.

Pandey, V. K., & Rajput, P. (2020). Review on natural language processing. *Journal of Critical Reviews*, *7*(10), 1170–1174. https://doi.org/10.31838/jcr.07.10.230

Solangi, Y. A., Solangi, Z. A., Aarain, S., Abro, A., Mallah, G. A., & Shah, A. (2019). Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis. *2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences, ICETAS 2018*, *November*, 1–4. https://doi.org/10.1109/ICETAS.2018.8629198