# Natural Language Processing: History, Evolution, Application and Future Work

Prashant Johri, Mukul Kathait, Munish Sabharwal,
Ahmad T. Al-Taani and Shakhzod Suvanov

April 23, 2020

# Natural Language Processing: History, Evolution, Application and Future Work

Prashant Johri[1], Mukul Kathait[1], Munish Sabharwal[1], Ahmad T. Al-Taani[2], and Shakhzod Suvanov[3]

[1] Department of Computer Science, Galgotias University, India
[2] Department of Computer Science, Yarmouk University, Irbid, Jordan
[3] Faculty of Mathematics and Informatics, Samarkand State University, Samarkand, Uzbekistan

johri.prashant@gmail.com
kathaitmukul0699@gmail.com
mscheckmail@yahoo.com
ahmadta@yu.edu.jo
shakhzod_suvanov@yahoo.com

**Abstract.** It is quite hard to imagine a smart system like a voice assistant or a chat-bot or a recommender system without Natural Language Processing (NLP). It all starts with an initial unit that first interprets the data (audio or text) provided and then start making sense of the data and after proper processing of the data the actual steps are followed by the machine to throw some replies or get the work done. NLP does not fall under a discipline; rather it is a part of several different disciplines i.e. computer science, information engineering, Artificial Intelligence (AI), and linguistics. The concern of NLP is the interaction between a computer and human languages. NLP areas include Speech Recognition, Machine Translation, Automatic Text Summarization, Part-of-Speech Tagging, etc. Generally, NLP is used in many real time applications like smart homes, smart offices like Alexa, Cortana, Siri, and Google Assistant. The history of NLP generally started in the 1950s and has come a long way from then and improved a lot. This paper discusses the history of NLP, its evolution, its tools and techniques, and its applications in different fields. The paper also discusses the role of Machine Learning and Artificial Neural Networks (ANNs) to improve NLP.

**Keywords:** NLP, Machine Learning, ANNs, Deep Learning, ATNs

## 1 Introduction

In today's fast-moving world, almost everything is just a one click away. Hey Google, what the weather would be like next Monday? You may want sunglasses; it would be 18C° (Sunny). It is a quite basic command for the voice assistant, and it is a way below its capability. We have come a long way in AI and life without it will be difficult. This came as a result of the continuous efforts of researchers in the field of AI. NLP was proposed to bridge the gap between the human language and computer's understanding. Researchers proposed different approaches to enable the computer to process and understand the human language. One of the early researches in NLP was in machine translation. The goal of machine translation is to develop automated programs to trans- late text or speech from one natural language to another.

The reach of AI is not confined in a field rather it is expanded in every field that one can think of; this is somewhat the same for NLP. Several fields of study are incorporating the concept of NLP to make systems robust and automated to meet the needs of the future. Apple integration of Siri voice assistant on their iPhones in 2011, was the first groundbreaking achievement of AI especially NLP that was very visible even to with little experience in technology. The assistance was not too much from experienced staff whereas it opened the door for other manufacturers to do their experiments in the technology and came up with more accurate and advanced products. Later, different companies started proposing new approaches in NLP in different fields such as medical field.

Recent NLP approaches are based on deep learning. Earlier approaches that employed deep learning did not gave good results since the processing power needed for deep learning implementation is very high. Nowadays, we have very efficient computers that can perform complex tasks within a fraction of a seconds and the data required for training the machine learning model is abundant too. Most of the AI technologies use NLP as a crucial part. During the past decade, many NLP approaches have been proposed. Some of the ongoing research in NLP investigates various ways for improving deep learning approaches used in NLP, such as the use of Recurrent Neural Networks (RNNs) to guess the theme of the article and suggesting the upcoming word in a sentence.

The primary objective of this paper is to provide an understanding of the rise of NLP, its evolution, recent applications, and suggest future applications that can take advantage of this technology.

The paper is organized as follows. The second section discusses the history and evolution of NLP from merely a word translation to a real-time language processing. The third section discusses the implementation of NLP approaches. Section four explains the tools and techniques that NLP incorporates and the use of different components for improving its effectiveness. The fifth section presents the applications of NLP and how NLP is contributing in different fields. The last section discusses the gaps in NLP research and the areas where NLP still needs refinements for future work.

## 2 History and Evolution

Language is a medium of delivering thoughts, information, and ideas along with emotions, imperfections, and ambiguity. It is difficult to think of a language as a combination of mathematical rules that works simultaneously to form a sentence or a phrase. Thus, making it difficult to form logical criteria that can be applied to the machine, to make it understand the language and return the results in the same way.

The first time when the world came across the term ‘translating machine’ was in the mid-1930s when the first patent for ‘translating machine’ was appeared. There were two patents for the same technology. First was for Georges Artsrouni, who used a bilingual dictionary to map the words of one language directly to another using paper tape. It was basic approach and did not provide a way to deal with grammar for a language. The second approach was given by a Russian, named Peter Troyanskii, who gave a detailed strategy to tackle the grammar of a language. He used the bilingual dictionary along with a method to deal with grammar of the language established over Esperanto (originally titled, the international language). Both were the disruptive approaches toward the technical domain, but they fail to provide the working model and remain the conceptual approaches only.

The first attempt of using NLP was by the Germans in World War II. **Enigma**, the name of the machine that was used for encrypting the secret message of Germans' into the secret code, and is used also to transfer the message to the field commanders and military units of Germans placed in Europe, to further arrange the attacks and met the requirements of the army. It was one of the great achievements of Germans to be able to communicate secretly even in the most precarious situations. Later in 1946, Britain comes up with, **Colossus**, a machine that was capable of successfully decrypting the secret code generated by **Tunny** (code name given to Enigma, by the British). It was the turning point of World War II since the British were able to know the position of the German Military, their strategies, army conditions, strengths, and could take prior action accordingly.

During World War II, the British government cryptographic establishment was placed in Bletchley Park. Bletchley Park was the place where Alan Turing, along with other intelligence agents, decoded the German message and provides insights about the German military. This was indeed the ground-breaking achievement in World War II, but it was also the first wave for the advancement in the field of modern-day computing.

In 1950, a Turing test (imitation game) is proposed by Alan Turing in order to determine whether a computer can think like a human or not. The test includes three players: a man (player-1), a woman (player-2), and an interrogator (player-3). Player-3 tries to find out the gender of player-1 and Player-2, with the help of written conversation only. But the catch of the game was, Player B will lead the interrogator to the correct solution and Player A will try to trick the interrogator and lead it to the wrong solution. Now, Turing suggested replacing Player A by a machine. If Player C successfully able to identify the gender of both the player it the machine fails the test, otherwise it wins the test [1]. This test was not to make a machine solve the problem but to identify whether the machine can perform tasks that humans can do indistinguishably or in other words, can machine think the way humans think.

Any Language is incomplete without the use of proper grammar. The true sense of a sentence is made only when the proper utilization of grammar is done. Until 1957, there was no way to incorporate grammar into the machines and make them understand its true sense. The true evolution in the field of NLP comes in the year 1957 when Noam Chomsky introduced the *syntactic structures*. Chomsky emphasis on “formalized theory of linguistic structure” [2]. He works on refining the set of rules to form vigorous linguistics based on universal grammar. But later, Charles Hockett found out several drawbacks to Chomsky’s approach. The major one was, Chomsky, considered language as a well-defined, stable structure and a formal system, which was possible only in an exceptionally ideal condition [3].

One of the early areas of NLP was machine translation. The goal of machine translation is to develop automated programs to translate text or speech from one natural language to another. For example, an automatic text translation program was developed through a joint project between Georgetown University and IBM Company in 1954. Experimental results showed that the proposed program was successfully translated sixty Russian sentences into English.

It was done by directly mapping the sentence and make use of the dictionary, which was explicitly maintained by the author, and the author claimed that within the coming few years, the machine would become capable of translating consistently. But the progress comes way slower than expected which hit the funding of the project and it reduces dramatically.

In the late 1960s, a machine was developed by Terry Winograd at Massachusetts Institute of Technology (MIT) named SHRDLU. It was the first NLP computer program that was able to perform the tasks like moving objects, determining the current state, and remembering names. It performs all these tasks in the “blocks world” environment. It became the first major achievement of AI and lures the attention of the researchers to optimize this technology. But the success of the program gets halted when it comes to interpreting the more real-world situations which were ambiguous and complex [4].

In the year 1969, Roger Schank introduced the concept of tokens that provide a better grasp of the meaning of a sentence. These tokens include real-world objects, real-world actions, time, and locations. For every sentence, these tokens provide the machine with better insights into the things happening and the object involves in it [5].

Till now all the rules that were used to make sure that machine could understand a sentence are based on the phrase structuring. One way or another, researchers try to make well-defined set of rules that a machine could follow to deduce the meaning of the phrase. In the year 1970, William Woods introduced the concept of Augmented Transition Networks (ATNs) for the representation of natural languages. He used finite set automata with recursion to reach the final state and conclude the meaning of the phrase according to the information available. The concept provides the result possible for information available (part of a sentence available) and makes changes to the meaning as per the information further provided. It comes up with a guess when the phrase provided is leading to an ambiguous situation. ATNs uses recursion to solve the problem where enough information is not provided and postpone the decision until more information is provided [6].

Most of the approaches used until now to make the machine understands natural languages were based on the handwritten rules that machine used to match. But in the 1980s, another emerging field of computer science was holding its grasp in the area of computing, that was, machine learning. Machine learning was the major shift from the handwritten rules to the concept making. Machine learning algorithms provide more advanced way to interpret the ambiguity and further provide acceptable evidence for the consideration of a decision. Algorithms like decision trees use if-then rules to better deduce the optimal result and probabilistic algorithms back up the decision opted by the machine by providing enough confidence.

Currently, a shift has been made to deep learning. This was due to the dominance of deep learning to perform tasks that are difficult to perform by simply using rules and fixed criteria. When it comes to NLP, there is no way to solve the ambiguity of the language. Single word can have multiple meanings according to its neighboring situation and it is not possible to write a rule or to use a decision tree to represent every possible meaning. Deep learning solves this problem efficiently as it does not require a programmer to provide the rules for deciding rather an algorithm itself deduces the process of mapping an input to an output.

## 3 Language Implementation

Processing human language is one of the toughest tasks for a machine to tackle. It's not a single form or concept that contributes to sentence formulation, rather there are several forms like nouns, verbs, adjectives, determinants, etc., that are used in association with each other to come up with the meaning of the sentence. Moreover, the ambiguity in the sentence meanings, like in the sentence: “he will be *running* a marathon soon", and in a sentence: "the software is *running* faster than expected," adds another layer of complexity. In both phrases '*running'* is used, but the meaning of the *running* is different. The meaning of a word depends on the reference to which it is used. Parsing these types of sentences containing the ambiguous words is difficult for computers. The task becomes more complex when the grammar is considered. The grammar can be extended to handle sub-categorization by proposing additional rules or constraints (e.g., would is often used in conditional sentences with a clause beginning with "If"). But the use of rule-based approaches may give inefficient results due to the ambiguity and complexity of the language.

Addressing all these problems is an essential task to move forward and it gave rise to statistical NLP [7].Parsing-rule is addressed with the probabilistic approach in statistical parsing. Probability is assigned to an individual rule, which is determined through machine learning algorithms. Numerous detailed rules are collected, and broader rules are defined. These broader rules help in building decision trees and statistical parser always parse a sentence with the maximum likelihood. Thus, statistical approach produces better results.

NLP can be comprised into five major components: morphological analysis, syntactic analysis, semantic analysis, discourse analysis, and pragmatic analysis [Jurafsky & Martin 2000]. Parsing is the main step in syntactic analysis; it is the process of analyzing the constituent words in a text, based on an underlying grammar, to determine its syntactic structure. The five components include some sub-tasks like: text decomposition, spelling, morphological parsing, stemming. The parsing process results in a parse tree having sentence as root, noun-phrase, verb-phrase, etc. as intermediate nodes, and individual words as leaves.

Semantics in NLP is a process of deducing the meaning of the text. It performs the process like semantic analysis (check meaningfulness of a text), word sense disambiguation (determining the correct sense of the word, in case of ambiguity in meaning), lexical semantics (checks for the synonyms, antonyms, homonyms, etc.)

Pragmatic deals with the extraction of information from a piece of text [Manning and Schutze]. It is further divided into three sub-fields: reference resolution (detecting reference) discourse analysis (determining the structure of the text), dialog interpretation (interpreting the information from the text).

## 4 NLP Tools and techniques

NLP is concerned with making computers to interpret, understand, and to manipulate human languages. Traditionally, the interaction of humans to computers is done through a programming language. When it comes to human language interaction with the machine, achieving this interaction is quite challenging, as human language is highly ambiguous, contains slangs with unusual meanings, and contains social contexts. The task becomes more challenging when the accent is taken into consideration, as people from different regions have a different accent.

NLP incorporates two major tasks: syntax analysis and semantic analysis. Syntax analysis is used to make the arrangement of the words in the sentence in such a way that it starts to make grammatical sense. It helps NLP to assess the meaning of the sentence based on the grammatical rules. Semantic analysis is done to discover the meaning behind the words and their use in a sentence. It is applied by NLP for understanding the structure and meaning of a sentence.

Recently, deep learning has received big attention by researchers in NLP due to the availability of huge (big) amount of texts (data) for natural languages. Standard NLP tools and techniques are fine to use but deep learning has altogether revolutionized the entire process. The massive success of deep learning is due to two major reasons:

1. The increase of the amount of data available:

Today, the data is producing at a tremendous rate and the pace is only accelerating with the growth of the Internet of Things (IoT). With this amount of data, the training process of the deep learning model can be done with greater precision and the model can work on a variety of examples. Previously, deep learning fails to get this much exposer because of the limited amount of data available.

2. The high processing power available:

Deep learning model requires high-end hardware with faster processing capabilities to perform the training and testing phases. Earlier, the machine was not that capable of providing this high processing power to meet the requirement. But now, the hardware is easily available to carry out the operation of deep learning appropriately.

I. Deep Learning Challenges:

Deep Learning approach requires a huge amount of labelled data for training the model and getting this amount of labelled data is one of the main hurdles to NLP. Labelled data is not readily available, and to overcome this hurdle the data needs to be labelled explicitly which is a highly expensive and time-consuming process.

To make this process faster, deep learning engineers came up with a suggestion of using a semi-supervised approach. Semi-supervised learning is a combination of supervised and unsupervised learning. In this learning approach, a small amount of labelled data is first used to train the model to label the remaining unlabeled data and make the labelling process much faster. It makes the labelling process more efficient and less expensive.

Earlier NLP research was based on rule-based approach, i.e. the machine was not trained to identify the meaning of the sentence or phrase, but it is trained to look for certain words or collection of words in certain patterns and respond with a specific action when the word or pattern is encountered. Deep learning has the advantage because it uses a more intuitive and flexible approaches in which many examples were used to identify the speakers' intent and then provide its own response.

II. NLP Tools

There are several open-source tools available for NLP. These include:

1. Python Tools: Natural Language Toolkit (NLTK), TextBlob, PyTorch-NLP, Textacy, SpaCy.
2. Java Tools: OpenNLP, StanfordNLP, CogCompNLP
3. Node Tools: Retext, Compromise, Natural, NlP.js

## 6 NLP Applications

NLP has many applications in various fields. Deep learning has opened the door for the research in the field of NLP. NLP is widely used in the IoT devices. Voice assistant like Cortana Alexa, Siri, and Google Assistant have made their own market lately. It can be seen in homes and offices quite often. Voice assistant is not confined to these areas only rather it has also started implementing in automobiles too.

There are tones load of comments, ideas, suggestions been posted on the internet from social media. It became quite hard for an enterprise to go through all these comments to monitor the performance of their business. That is where the sentimental analysis, another primary use case of NLP comes to rescue. Sentiment analysis helps businesses to filter out the essential information from a massive amount of text data and make an optimal decision in the right direction in order to max out their profits.

NLP can be used to analyze a large amount of textual data and make the process faster and efficient. A large amount of information is stored in the form of text and it is time-consuming to go through this text manually and come up with a solution. Before NLP, this was done manually but the entire process gets revolutionized once deep learning-based NLP arise. Take the medical records of patients, for example, there are a huge amount of medical reports generated and specialized doctors need to go through all these reports to suggest a diagnose. This is a slow process, but it can be made efficient by using NLP for text analysis.

Radiology is a medical discipline concerned with diagnose and treatment of diseases in animals and human bodies using medical imaging. The report generated is highly specific and requires specialized lookup for further diagnosis and treatment. NLP can be used with machine learning in medical applications to medical reports with commendable accuracy. NLP can be applied to extract real-life concepts from unstructured data, sometimes leveraging advancement in machine learning to perform classification tasks [8].

Classification is one of the broadly used techniques in different business markets. Classification is used to filter the important e-mails from the spam ones. It is easy to classify a mail as spam mail or not because the size of the data is large. For big data, the chance of detection increases. But when the data is small, like in reviews, it is difficult to differentiate between the real ones and the fake ones, which hinders the credibility of the entire review collection. Identification and removal of such fake review requires automation since it is a difficult task for humans to perform manually. Machine learning algorithms can be used with NLP to produce good results [9].

NLP is an important discipline to make electronic health records a supportive source of data to meet needs and helping in main activities for researchers and clinicians while reducing their needs for data and charts review. NLP can be used to extract and process clinical data or information for both structured and unstructured forms. NLP also could be used for patient’s classification and supporting critical clinical tasks like clinical decision-making and producing quality reports [10].

A clinical decision-making system can be developed using patient's behavior toward a product, medicine, or treatment using NLP techniques. For instance, aspect-based sentiment analysis can be used in clinical decision-making for backing personalized therapy analytically [11].

NLP makes it easy for people to write down essays, emails, articles, etc. For example, the software Grammarly can detect the theme of the article once it is written, can check the grammar, and can suggest better ways to write a sentence.

As most of the information deals with digital documents and contracts nowadays, there is a need for a mechanism that protects the privacy of an organization. Named Entity Recognition can be used to monitor and detect privacy violations in online contracts by automatically monitoring Personally Identifiable Information [12].

Search Engines like Google, Bing, DuckDuckGo, etc. use machine translation tools to translate the content of a webpage into another language while holding the meaning and message intact.

Many organizations use chat-bots in their websites as a primary source of interaction between site and the user. Chat-bots process the query that the user enters and shows the results and suggestions accordingly. Chat-bots are considered much faster than humans to process the information and come up with a result of suggestions, thus, reducing the friction between the problems and the solutions.

## 7 Future work

NLP came a long way when it comes to its implementations and applications. From bilingual dictionaries to handwritten rules, then from phrase structuring to ATNs, and finally machine learning, NLP has evolved in different stages with better and more advanced processing capabilities. It is hard to deny the rise of AI in the future to come. There are a lot of applications of NLP through the door of deep learning and it is performing as per human expectations and in some fields, it is outperforming humans too. Today, in our surroundings, most of the things we use are automated, as machine learning has overwhelmed the market with advanced devices but there is still a continuous need for improvement. One may think that NLP had its effect on every area which can take advantage of, but still several areas can utilize this technical attainment and help in the progress toward a better future for everyone.

Nowadays, NLP is performing quite well in textual and audio data. NLP has achieved a commendable speed to process these data. But still, there is a set-back for NLP, while processing the sarcasm, irony, and idiom in data. More research is needed in NLP to process various clinical data. This could help in the identification of early symptoms and targeting the root cause by suggesting the potential drug in the pharmaceutical field. Further, the extension can be made to this field using deep learning methods to train the machine over various structured and unstructured data.

## References

1. A. M. Turing (1950) Computing Machinery and Intelligence. Mind 49: 433-460.
2. Lees, Robert(1957), "Review od Syntactic Structures", Language, 33 (3): 375-408, doi: 10.2307/411160, JSTOR 411160.
3. Hockett, Charles (1966), "Language, mathematics and linguistics", Current Trends in Linguistics, 3, Theoretical Foundations, The Hague: Mouton, pp. 155–304.
4. SHRDLU
5. Roger Schank, 1969, A conceptual dependency parser for natural language Proceedings of the 1969 conference on Computational linguistics, Sång-Säby, Sweden pages 1-3.
6. Wanner, Eric (1980). "The ATN and the Sausage Machine: which one is baloney?".
7. Nadkarni PM, Ohno-Manchado L, Chapman WW. Natural Language Processing: an introduction. J Am Med Inform Assoc 2011: 18:544-51.
8. Po-Hao Chen "Essential Elements of Natural Language Processing: What Radiology should know."
9. Vijayakumar B, Muhammad Fuad MM, "A New Method to Identify Short-Text Authors Using Combinations of Machine Learning and Natural Language Processing Techniques"
10. Juhu Y and Liu H "Artificial intelligence approaches using natural language processing to advance EHR-based clinical research".
11. Silva P, Goncalves C, Godinho C, Antunes N, Curado M "Using Natural Language Processing to Detect Privacy Voilation in Online Contracts".
12. Hiremath BN, Patil MM "Enhancing Optimized Personalized Therapy in Clinical Decision Support System using Natural Language Proccessing."