International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217688

A Study on the journey of Natural Language Processing models: from Symbolic Natural Language Processing to Bidirectional Encoder Representations from Transformers

Rajarshi SinhaRoy
Department of Computer Science, St. Xavier's College Kolkata, Kolkata, West Bengal, India

Article Info
Volume 7, Issue 6
Page Number : 331-345

Publication Issue :
November-December-2021

Article History
Accepted : 12 Dec 2021
Published : 26 Dec 2021

ABSTRACT

In this digital era, Natural language Processing is not just a computational process rather it is a way to communicate with machines as humanlike. It has been used in several fields from smart artificial assistants to health or emotion analyzers. Imagine a digital era without Natural language processing is something which we cannot even think of. In Natural language Processing, firstly it reads the information given and after that begins making sense of the information. After the data has been properly processed, the real steps are taken by the machine throwing some responses or completing the work. In this paper, I review the journey of natural language processing from the late 1940s to the present. This paper also contains several salient and most important works in this timeline which leads us to where we currently stand in this field. The review separates four eras in the history of Natural language Processing, each marked by a focus on machine translation, artificial intelligence impact, the adoption of a logico-grammatical style, and an attack on huge linguistic data. This paper helps to understand the historical aspects of Natural language processing and also inspires others to work and research in this domain.

Keywords: Natural language Processing (NLP), Artificial Intelligence, Machine Learning, History of NLP, computational semantics, machine translation

I. INTRODUCTION

Computer technology now dominates a wide range of sectors, from our daily lives to advanced scientific techniques. Our PCs and smartphones have become our daily drives, and the online phase has taken over since the start of the COVID epidemic in 2020. Everything is done online these days, from attending meetings and schools to ordering groceries, which is only possible because of computers. We've made a lot of progress in AI Technology, and life will indeed be impossible without this. These results were achieved because of the ongoing efforts of AI researchers. The goal of Natural Language Processing or NLP is to narrow the gap between human language and computer comprehension. Researchers offered many methods for allowing a computer to process and comprehend human natural speech. Machine

translation was among the first NLP studies. Machine translation aims to create automatic computers that can analyse text, speech and convert it into different languages. We all speak, read, and write using language. We also think about the world in terms of words, we make plans in terms of words also we dream in terms of words and take decisions in terms of words. The impact of AI is just not limited to a single field, but instead extends to any field that can be imagined; this seems to be true for NLP. NLP is being used in a variety of sectors to make systems more durable and automated in order to meet future requirements. In 2011, Apple introduces Siri, a voice assistant that is the best achievement for the technology and beauty of NLP. After Siri, Google integrated "Google Assistant", a voice assistant to the operating system Android. This is the ground-breaking achievement of NLP. These voice assistants are as good as a human. If we say: "Hey Google, how is the weather?" Google will reply: "You may need sunglasses; it is 22°C (Sunny).". Not only that it can predict next week's weather. It is all happening because of the advanced research in NLP, AI, and Deep Learning. Recently, Deep learning is used to predict and enhance with NLP methods. Earlier deep learning algorithms failed to produce satisfactory results because of the significant processing power required for deep learning implementation. NLP researchers have been improved a lot because of high-performance computers by which we can perform a lot of complex calculations, a steady increase of data and potential, and good algorithms. Many NLP models' foundations were proposed before the 90's decade. Several modern NLP researchers look into various ways to improve deep learning methodologies used in NLP, like using recurrent neural networks (RNNs) to predict the article's theme and selecting the very next word in a phrase. The primary goal of this paper is to give a simple understanding of how NLP starts its journey, how researchers developed modern NLP algorithms, the evolution of these NLP models. In this paper first section covers the basics of NLP is discussed and then the journey phases of NLP. In the second section, I talk about some basic models to some advanced current models of NLP. The last section discusses the conclusion in NLP research and the future refinements.

II. NATURAL LANGUAGE PROCESSING

A. Classification of Natural Language Processing

Natural Language Processing (NLP) is an optimized computational approach, tract of Artificial Intelligence and Linguistics for evaluating and understanding the statements or words and also analyse in order to achieve human-like language processing. NLP is created to communicate with computers as humans interact with each other. Everyone does not know programming languages by which they can communicate with computers, but researchers want to create some algorithms so that computers can understand human language. This is the background of where NLP comes from.

Language consists of some discrete symbols. The base element for language is characters which is a representation of symbols. Words are formed by characters that imply some meaning of events, objects, actions, ideas, etc. All of these objects, events are based on some rules. NLP can be classified into some types shown in Figure. 1.

Natural Language Processing is based on understanding and generation. Natural language Understanding or Linguistics is the science of language, which is consists of Syntax, Phonology, Morphology, Semantics, Pragmatics. The syntax is the structure of sentences, Phonology refers to sound, Morphology means word formation, Pragmatic refers to understanding and semantics is syntax.

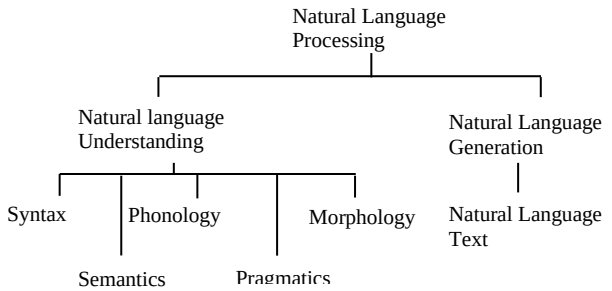


Figure 1. An illustration of classification of Natural Language Processing

B. Natural Language Understanding

Phonology could be a branch of linguistics that deals with the systematic organization of sound. Phonology is originated from the Greek prefix 'phono', which refers to sound, and the suffix 'logy', which simply means word. Russian linguist Nikolai Trubetzkoy stated in 1993 that the study of sound as it relates to a linguistic system. Although Lass argued in 1998 that phonology is properly involved with the function, behaviour, and organization of sounds as linguistic objects, it may well be understood as phonology legitimate is focused on the feature, behaviour patterns, and organization of sounds as linguistic elements. The semantic use of sound to encode meaning in any Human language is referred to as phonology.

Syntax defined the format of the sentence, which explains how words and phrases are used properly to generate a correct sentence. A sentence is not correct or meaningful until it is syntactically correct. Noam Chomsky gave a great example to understand about the syntax: 'Colourless green ideas sleep furiously' does not give any proper meaning though every word has its own meaning.

Morphology is basically an analogy of word representation which is also known as Morphemes. Let's take the word "preoccupation", where the prefix is 'pre', the root is 'occupa', also the suffix is 'tion'. We understand the meaning of a word by breaking it into some morphemes. One word which has its own meaning is called Lexical morphemes and Lexical morphemes combined with words like 'ed', 'ing', etc. are called grammatical morphemes. The combinational occurrence of grammatical morphemes is called bound morphemes.

Semantic processing focuses on the relationships between the sentence's word-level meanings to discover the sentence's various meanings. Semantics can handle the inner meaning of the sentences by semantically distinguishing the words.

Pragmatics is a subset of linguistics. Conversational implicature is a procedure where the speaker suggests, and the listener infers. Pragmatics focuses on this process. It is basically an understanding between people to interact meaningfully. In real-life words does not have always a constant meaning rather it can be different in certain situations. Pragmatics consider these types of aspects of human language.

C. Natural Language Understanding

The process of constructing meaningful words, sentences, and paragraphs from an internal representation is known as Natural Language Generation (NLG). This is a branch of Natural Language Processing and comprises four phases: establishing goals, organizing how objectives might be attained by analysing the circumstances and accessible sources, and implementing the plans as a text Fig. 2.

Natural Language Generator has three components: Speaker and generator, Components and level of representation, Application and speaker. Speaker and generator basically generate the text. Language generation is done by component and level of representation which is also categorized into 4 types: Textual organization handles the grammar and format of the text, Content selection basically handles the selection of information by checking its representation, Realization is a process which checks if selected resources are realizable or not means the way we talk or write, Linguistic Resource basically chooses the particular word, idioms and syntactic constructs, etc., which supports the realization of information.

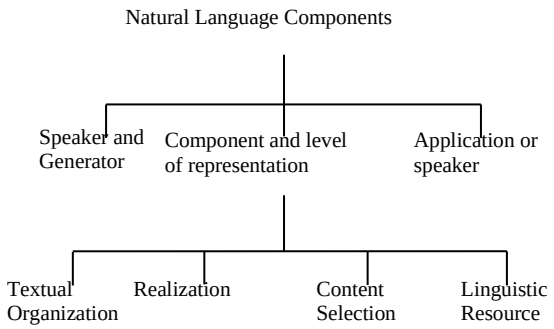


Figure 2. An illustration of a classification of Natural Language Generation

D. Natural Language Understanding

To represent the sentence and generate text 'level of language' is used by considering the sentence planning, surface realization and content planning Figure. 3. Communication goal and Knowledge base are the refectories of Content planning. User Model, Domain Model, Content Selection, and Discourse Planning are used for Content Planning. Sentence Planning derives from Content Planning which is two types: Aggregation and Expressions and Algorithms. Lexicalization and Grammar is the part of Surface Realization which helps to derive the Natural Language Text.

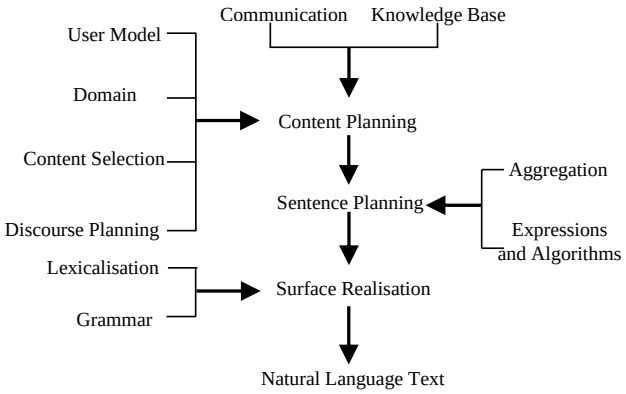


Figure 3. An illustration of Level of Language in Natural Language Processing

III. JOURNEY PHASES OF NLP MODELS

Language is a method of conveying feelings, information, and ideas, as well as feelings, flaws, and ambiguities. Now in mathematics when we ask, "What is Language?", we can say it as an equation of mathematics which is a combination of functions defined with sentences or phrases. As a result, forming logical combinations that can be applied to the computer, causing it to learn the message and provide the same results, becomes challenging.

A. Before the Beginning of Everything

From ancient times language went through various transitions. Linguistics is the scientific study of trying to find a real meaning that has been lost over time. In mid centuries, for more than 1000 years it is been Latin was the most common language in Europe. It is realist means generalizations exist in the real world, separate from and before physical existence also nominalist means various components exist in their very own right. Conceptual objects (symbols) are designated using universals. In modern times we are seeking a true language which can be Sanskrit: an ancient Indian language, also being an Inflective language so it has a high productivity rate. Swiss linguist Ferdinand de Saussure stated that ideas are like galaxies, a beyond thinking which can be represented using some arbitrary of signs. He also stated about the clarification of language and its structural relation. The majority of modern computer-assisted natural language processing is based on Saussure's hypothesis. The field of inquiry has been expanded to include parole, based on the study of language. Saussure believed that a tumultuous state of nature could only have been identified by using language to describe it. Linguistic autonomy was proposed, and language research evolved into a science unrelated to the natural world.

B. Beginning of the Natural Language Processing (1940-1960)

In 9th century Arabic cryptographer, Al-Kindi used several modern machine translations. After that in 17th-century French philosopher, mathematician,

and scientist René Descartes suggested a universal language, with similar thoughts in many languages using the same symbol. Again, in the middle of 1930, two researchers used the term “translating machine” in their research paper. Georges Artsrouni utilized a bilingual dictionary and paper tape to map words from one language to another. And a more extensive proposal, based on the Esperanto grammatical system, was given by Russian researcher Peter Petrovich Troyanskii, which featured both the bilingual dictionary and a technique for handling grammatical roles between languages. In World War II Germany attempted to use NLP using enigma to send secret encrypted messages. In 1946 Britain researchers came up with new technology named colossus to decrypt the messages sent by enigma. At that time every superpower country was focused on machine translation, the basics of NLP.

The previous research of Booth and Richens, and the discussion on translation at Weaver's seminal in 1949, the NLP study began in earnest in the 1950s. The IBM-Georgetown Experiment in 1954 featured machine translations from Russian to English in a quite primitive form and with minimal testing. In 1952 a conference was held on Machine Translation then a new journal MT is published in 1954. After 2years again a conference was held. In 1958, at Washington International Conference Auto-abstracts (really excerpts) for one session's papers were provided by Luhn, Artificial intelligence was brought up by Minsky and while using a thesaurus, language processing has been linked to information retrieval. In the middle of 1957, Noam Chomsky proposed the syntactic structure of language. Proper use of grammar is important to form a language. After that, an International Conference was again held in 1961 at Teddington on Machine translation. This discussion was mainly focused on MT of languages and applied language analysis. Also, in these conferences, many researchers from different countries worked on various aspects of NLP.

In this era, there was no high-level programming languages and no advanced algorithm available. Also, programming was done by assemblers and machine accessibility was not so common back then, less and slow memory was available at that time. NLP researchers from all over the world managed it. In this era, researchers focused on word-to-word processing, syntax, syntactic processing, syntax-driven processing topics of NLP. Also, some of them focused on semantically driven processing. By ALPAC report published in 1966, the USA stopped funding for MT research because researchers can find the true point of MT. Notably, in terms of syntactic analysis and effectiveness in parsing and characterizing phrases, NLP researchers tried to focus on computational language processing. And some of them focused on polysemy and processing, generation. In this phase, they developed first-stage tools. Although in the last quarter of the 1960s, sending output to their clients MT production systems are used, the research of this era did not generate any technologies of scope.

We cannot say this era developed NLP but, in this era, the computer was used for language study. It was the beginning phase of NLP though there was a huge number of misconceptions, a lack of concern was present.

C. The Cold phase of Natural Language Processing (1960-1980)

NLP research was heavily influenced by Artificial Intelligence (AI) in the second phase. With the BASEBALL the automatic question-answering system, AI researchers found the solution to the difficulties of addressing and knowledge bases in 1961. In comparison to typical MT processing, the practical input to such systems was limited, and the language processing required was relatively rudimentary. Terry Winograd of the Massachusetts Institute of Technology (MIT) created SHRDLU, a machine that allowed users to communicate with English terms. This was the first NLP computer program capable of

tasks such as identifying objects and moving them, assessing current status, and recalling names. Roger Schank established the notion of tokens in 1969, allowing for a better understanding of the meaning of a phrase. Real-world things, real-world behaviours, time, and locations are among the tokens. For each phrase, these tokens supply the computer with a clearer understanding and further information about what is going on and the objects involved.

Transformation grammar was basically not suited for computing and analysis, specifically. Around the late 1960s, the mainstream linguistic theory was Transformation grammar, but TG did not focus on the concept of semantics. At the beginning of the 70's era, something new was introduced, Augmented transition networks were a concept to represent the human language. And this was first proposed by W.A. Woods. ATNs is a kind of graph hypothetical structure of transition network capable of taking actions and performing some tasks. For storing information ATNs use a set of registers. ATNs has a finite number of states which supports some actions or rule in order to move from one state to another. Basically, feature-based rule support and feature-based value assignment are included for ATNs. To tackle the problem of not having enough information and deferring a decision till additional information is available, ATNs used recursion.

In this era, NLP researchers had delivered a lot of new thoughts about general world models by filling up unique inputs and deep representations of models. But NLP researchers understood that NLP models are more complex to develop. Finding the fluency and construction of the proper sentences is not only hard but very time-consuming for the high-speed processors available at that time. Though Yale Group proposed a model to find the goal and plans by analysing the language, this became trendier in this era. Underline words meaning, indirect words meaning, story or report analysing, dialogue analysing were also becoming trendy at that time.

D. The Return of Natural Language Processing (1980-2000)

In this era, researchers understood that a real-world NLP model building was not so easy which can be accessed by a heavy application or system but if they focus on the utilitarian MT then it can lead them to a hope of the future of NLP. Before this era, researchers were finding different models to enhance the power of NLP models but in this era, they were more interested in the grammatical-logical phase to understand the language grammar more systematically. The development of Grammatical theory leads NLP developers to a clear idea of grammar used in natural language. This development helped researchers to represent the knowledge base and the reasoning. ATNs were used in grammar to simplify the version of the conversational approach. Linguists developed different types of grammar. This thought lighted up the abstract model of Context-free-parsing.

After mid 80's everyone was interested in Computational grammar theory making it a hot topic for research. This theory helped develop to understand the narrator or the speaker's intention and plans by working on the knowledge representation of this theory. Although it was the best topic in that decade this theory did not cover the concept of mood and expressions of the text. NLP researchers again worked on a Grammatical logical approach and introduce a new calculus-type approach for representing the meaning of language. This approach leads them to differentiate between 'Semantic' and 'Pragmatic' meaning. It also helped them to design an abstract model for the heavy systems.

In the late 1980s, a dramatic change happened in the research of NLP. Before that, all major approaches were dependent on complex handwritten rules or algorithms, but the increase of computational power and Machine learning algorithms make NLP research more efficient. For this statistical approach prediction and understanding of models became more efficient and probabilistic. At the end of the '80s and begging

of the '90s, a heavy application developer used the understandable grammar and parsing algorithm. European and Japanese researchers interested in MT. Also, they started a project named "Eurota research project". Japanese multinational companies are interested to work in this field, so they started helping them financially.

In the 1990s, every researcher got interested in statistical models for NLP. Also, in this time Commercial NLP system research became more popular. Speech technology, Dialogue interference, Text Analytics, etc., every sector was developing rapidly. Basis Technology, NetOwl, Refinitiv Company were developed in the mid-90s for text analytics. In NLS CoGenTex had been founded. In 1997, Microsoft developed their own syntax-based grammar checker and Unix developed their UnixWriter's Workbench. Also, in the 90's era, finite-state modelling is used in a voice dialogue system. Also, in this era LSTM RNN models were proposed, and the N-grams concept model was developed.

E. The Current phase of Natural Language Processing

After a massive impact of NLP in the commercial industry and research field in early 2001, a Deep feed-forward neural network model was proposed where this model takes n-words input as a format of vector and looked up in a table. This is the first concept and real-life implementation of word embedding. All embedding happens in the hidden layer and the output is generated from the last layer. In 2008 multi-task learning method was developed. This algorithm worked by sharing parameters on different models. Two embedded matrices are shared for two unique tasks which help them for developing low-level information handling. Now In 2013, other word embeddings models were introduced and that helps to convert words into a vector representation. Along with that several RNN and CNN implementations are occurred. Also, development in those algorithms also takes place between 2013 to 2016. In 20014 sequence to sequence model was also introduced. This model basically checks symbol by symbol of a sentence and then converts it into a vector. After that LSTM was also enhanced with the combination of sequence to sequence model. With this, the encoder and decoders concept were also introduced at the same time. In 2015 attention model was introduced which covers the issue of sequence to sequence mode by not converting all symbols. After Attention model 2018 some pre-trained models are introduced. These models help NLP to deal with billions of tokens, different text, languages, and many more problems. These models can also learn from a few amounts of information and predict an outstanding output. Also, Reinforcement learning also helped these models to train and select the data.

IV. TEXT REPRESENTATION TECHNIQUES

Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might

A. BAG OF WORDS

In-text classification problems Bag of Words is a widely used classical text representation technique. To solve a sentimental analysis problem, sending only text to the model is not a proper use although we can send it to a bag of words and convert it into numerical format or vector. The idea behind this technique is that a bag is consists of words where every text is a class in the bag of words and two similar words get the same class on a bag of words. Bag of words maps every word with a unique number between 1 and |V|. Every text inside the corpus would then be turned into a matrix with |V| dimensions, with the i_{th} component, $i=w_{id}$, basically representing the number of times the word w appears in the text.

If there are 3 sentences – "He is good boy.", "She is good girl.", "Boy and girl are good.". Then Bag of Words first checks the frequency of each word and

then it generates the embedded matrix. The frequency of the words is- Good(3), Boy(2), Girl(2).

TABLE I
BAG OF WORDS

	Good(f_1)	Boy(f_2)	Girl(f_3)
Sent 1	1	1	0
Sent 2	1	0	1
Sent 3	1	1	1

Bag of Words is two types: Binary Bag of Words and Bag of Words. In binary bag of words checks if the feature is present or not. If present, then 1 else 0. But for bag of words, we just increment the no of features if it is present multiple times.

B. TF-IDF

In Bag of Words, we cannot distinguish the words between two words: 'Which value is more than other?' – this question's solution we cannot get from a bag of words. In semantic analysis the value of a word is more important, i.e. 'He is an intelligent boy.' Here 'intelligent' word has more value than the 'boy' word. But in the Bag of Words, both the values give us the same weightage value of (1,1).

Term frequency of TF means in a given text: the measurement of repetition words appears. Because the lengths of the texts in the corpus vary, a word may appear more frequently in a longer text than in a shorter one. We divide the number of occurrences by the document's length to normalize the numbers.

$$TF(t,d) = \frac{(Number\ of\ repetition\ of\ words\ t\ in\ text\ d)}{(Total\ number\ of\ words\ in\ the\ text\ d)}$$

The term's relevance in a corpus is measured using IDF (inverse document frequency). All terms are given equal weight when computing TF (weightage). Stop words such as is, are, am, and others, on the other hand, are well-known for being unimportant, despite their widespread use. To adjust for these situations, IDF weights the terms that are relatively common across a corpus down and the rare terms up.

$$IDF(t) = \log_e \frac{Total\ number\ of\ sentences}{Number\ of\ sentences\ containing\ words\ t}$$

The frequency value of each word or TF-IDF score is the multiplication of TF and IDF.

If there are 3 sentences – "He is good boy.", "She is good girl.", "Boy and girl are good.". Frequency of the words are- Good(3), Boy(2), Girl(2).

TABLE III
TF

TF	Sent1	Sent3	Sent3
Good	1/2	1/2	1/3
Boy	1/2	0	1/3
Sent 3	0	1/2	1/3

TABLE IIII
IDF

Words	IDF
Good	$\log_e$ 3/3
Girl	$\log_e$ 3/2
Boy	$\log_e$ 3/2

TABLE IVI
TF-IDF

TF-IDF	Good(f_1)	Boy(f_2)	Girl(f_3)
Sent 1	0	1/2 × $\log_e$ 3/2	0
Sent 2	0	0	1/2 × $\log_e$ 3/2
Sent 3	0	1/3 × $\log_e$ 3/2	1/3 × $\log_e$ 3/2

C. WORD2VEC

Word2Vec is one of the most used models in Word Embeddings. It is basically a mathematical approach to make a relationship between some similar words. Each word is basically represented as a vector of 32 or more dimensions instead of a single number.

$$\text{King} - \text{Man} + \text{Woman} = \text{Queen} \quad (1)$$

Let a 2D representation example where Man is in (3,6) and Woman is in (3.2,6.2) and word Play is in (6,4). So, here we can understand that the position of man and woman is very near rather than play which implies that man and woman are a respectively similar word than word play. Now if King has a position of (4,5) and Queen is (4.3,5.3) then Word2Vec can generate an equation like – KING (4,5) – MAN (3,6) + WOMAN (3.2,4.2) which implies the position (4.2,5.2) which is very close to QUEEN (4.3,5.3). This is how the WORD2VEC works.

WORD2VEC has low dimensions and high dense which helps ML models to work perfectly.

V. MACHINE LEARNING AND DEEP LEARNING MODELS FOR NATURAL LANGUAGE PROCESSING

A. NAÏVE BAYES

On the basis of Bayes theorem, this algorithm was proposed. This algorithm is based on the conditional probability where P(A | B) = (P(A∩B))/(P(B)). Thomas Bayes just changed the probability equation for the same variables but different aspects and generate the relation between them which turns into P(A | B) = P(B | A) □ (P(A))/(P(B)). In a classification problem if the dataset has some features {f1, f2, …, fn} and output is y then the corresponding values are {x1, x2, …, xn} and the output value is y1. For this kind of dataset, the Bayes' theorem will be.

$$P(y \mid x1, x2, \dots, xn) = \frac{P(x1 \cap y)P(x2 \cap y)\dots P(xn \cap y) \times P(y)}{P(x1)P(x2)\dots P(x_y)} = \frac{P(y)\prod_{i=1}^{n} P(x_i)}{P(x1)P(x2)\dots P(x_y)}$$

If we consider P(x1) P(x2) … P(xy) as a constant, then,

$$P(y \mid x1, x2, \dots, xn) \propto P(y)\prod_{i=1}^{n} P(x_i)$$

then it implies that y = the maximum probability value of

$$P(y)\prod_{i=1}^{n} P(x_i)$$

B. Recurrent Neural Network (RNN)

Recurrent Neural Network is a deep learning neural model which takes different dimensions of inputs and generates the output with respect to time by using the hidden layers of neurons. In a use case of a sentimental analysis if the sentence is X1 = < X11, X12, … > with d dimension of vector, then in the first case it X11 will pass the hidden layer and the output will pass for the next part by t time of gap. Let's consider the X11, X12, X13, X14□□ Rd inputs analyzed by 100 hidden neurons. Out1 = f (X11 × w') for X11 input then Out2 = f(X12 w'' + Out1w') and so on and give the final output using Sigmoid function. This is called the Forward Propagation of RNN. By this, the sequence of the wards is maintained which is the disadvantage of TF-IDF and BOW. In Backpropagation the process of calculation starts with ∂L/(∂y ˆ) where L is Loss and y ˆ is the final output. After Calculating this the second part comes updating the value of wf'. To update this the equation: the last updated weight wf' □ wf' □ ∂L/(∂w'f); where,

$$\frac{\partial L}{\partial w'_{f-1}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial out_{f-1}} \times \frac{\partial out_{f-1}}{\partial w'_f}$$

After several iterations, the global optimal solution will be available through this method.

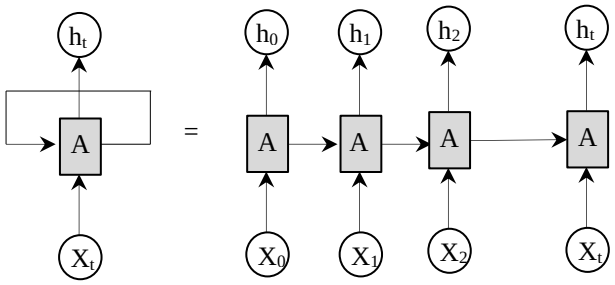


Figure 4. An illustration of the process of Recurrent Neural Network

C. LONG SHORT-TERM MEMORY (LSTM)

The main problem of RNN is the dependency of an output weight which is input at the very initial state. At the point of backpropagation when the weight became very less than the derivative of that weight is near tense to 0 which leads to an error. To solve this the LSTM or long short-term memory is used.

It is a type of RNN but used with some advanced features.

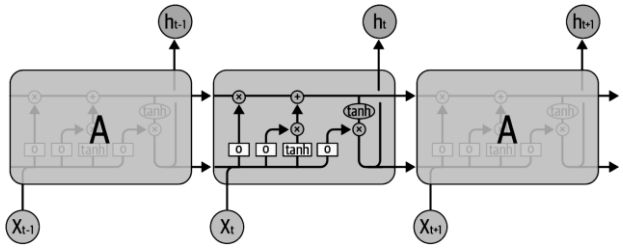


Figure 5. An illustration of the process of Recurrent Neural Network

Fig. 5. has four main parts: Memory Cell, Forget Gate, I/P Gate, O/P Gate. Memory Cell is just a tool to memorize and forget something. There are two operations: Pointwise Operation is [1,2,3,4] ⊙ [1,1,0,0] => [1,2,0,0]. Here the 0 means forget the data. Additional Operation is used for adding some data to memorize by memory cell. Walk Through of LSTM: at first two inputs x_t and h_{t-1} are used to find the output by passing from a sigmoid function. $f_t = \sigma (W_f \odot [h_{(t-1)}, x] + b_f)$ and by passing this value using sigmoid function we will get a vector of 0's and 1's. Next step is to add the new data using additional operator: $it=\sigma (Wi \times [h(t-1), x_t]+b_i)$ and $C_t' = \tanh (W_C \odot [h_{(t-1)}, x_t] + b_C)$. And at the next step just add the change value using the equation: $C_t = f_t \odot C_{(t-1)} + i_t \odot C_t'$ and at the last step $Out_t = \sigma (W_O \odot [h_{(t-1)}, x_t] + b_O)$ and $h_t = Out_t \odot \tanh(C_t)$ is calculated to find the final Output (Out_f). LSTM has three popular varieties: Sequence to Sequence, Vector to Sequence, Vector to Vector.

D. ENCODER-DECODER MODEL

Before the Sequence to Sequence model, basically we are taking a sequence of inputs and generating one output. To remove this disadvantage the new model is proposed name 'Sequence to Sequence Model' where this model can take a sequence of inputs and give a sequence of outputs. Now, this model has two parts: Encoder, Decoder. So, it is also called as 'Encoder-Decoder Model'. In Encoder, there are n number of Input taking options but do not have an option for output. Only the last Neuron box generate the output which is again used in the decoder to generate the output. LSTM, RNN, GRU models are used but LSTM is the most used neuron model. Encoders take the input sequence up to <EOS> after that it generates the Context Vector which is used as input in Decoder and after t time of operation, it generates the Output value.

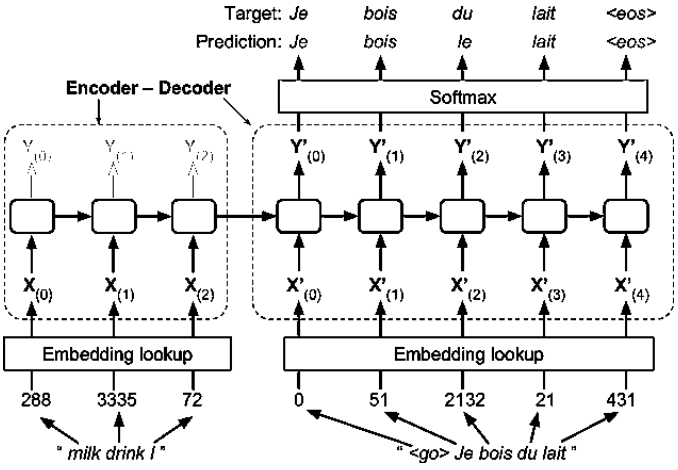


Figure 6. An illustration of Encoder and Decoder Model [32]

E. TRANSFORMERS

In sequence to sequence model long range of dependencies are really challenging and also this model cannot even handle parallelization. To overcome these problems Transformer Model is proposed.

In encoder block has two parts: Multi-head Attention, Feed Forward Neural Network, and the decoder has all the parts of the encoder and also a Masked Multi-Head Attention. Multiple identical encoders and decoders are layered on top of each other in the encoder and decoder blocks. The number of units in both the encoder and decoder stacks is the

same. A hyperparameter is the number of encoder and decoder units. Six encoders and decoders were employed in the study.

The first encoder receives the input sequence's word embeddings. After that, the data is converted and sent to the next encoder. All of the decoders in the decoder-stack get the result from the last encoder in the encoder-stack.

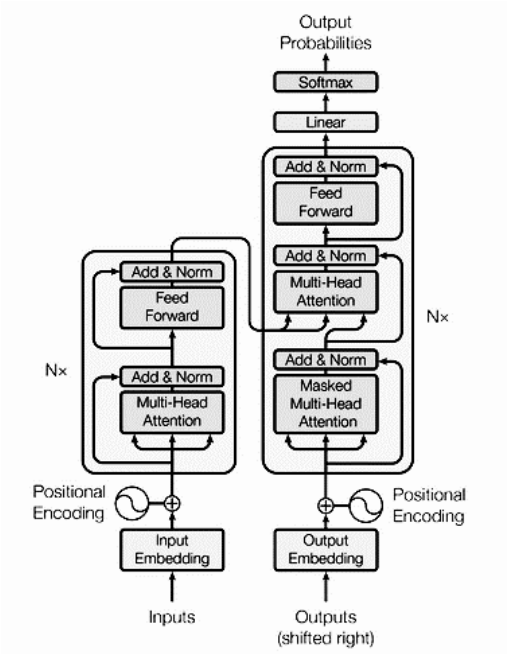


Figure 7. An illustration of a working process of Transformers Model[31]

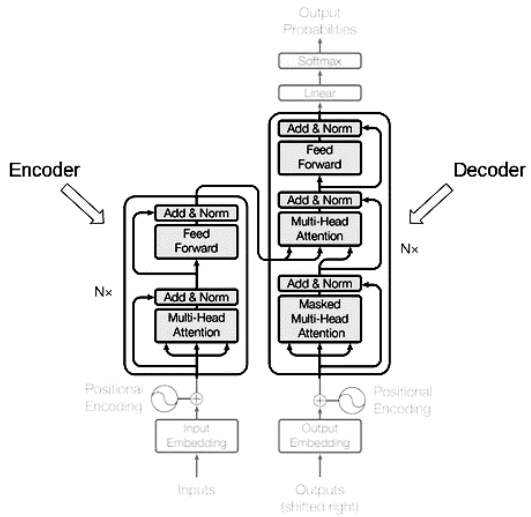


Figure 8. An illustration of Encoder and Decoder working process of Transformers Model[31]

In encoder block has two parts: Multi-head Attention, Feed Forward Neural Network, and the decoder has all the parts of the encoder and also a Masked Multi-Head Attention. Multiple identical encoders and decoders are layered on top of each other in the encoder and decoder blocks. The number of units in both the encoder and decoder stacks is the same. A hyperparameter is the number of encoder and decoder units. Six encoders and decoders were employed in the study.

The first encoder receives the input sequence's word embeddings. After that, the data is converted and sent to the next encoder. All of the decoders in the decoder-stack get the result from the last encoder in the encoder-stack.

It's worth noting that, in addition to the self-attention and feed-forward layers, the decoders contain an additional layer called Encoder-Decoder Attention. This allows the decoder to concentrate on the relevant bits of the input sequence.

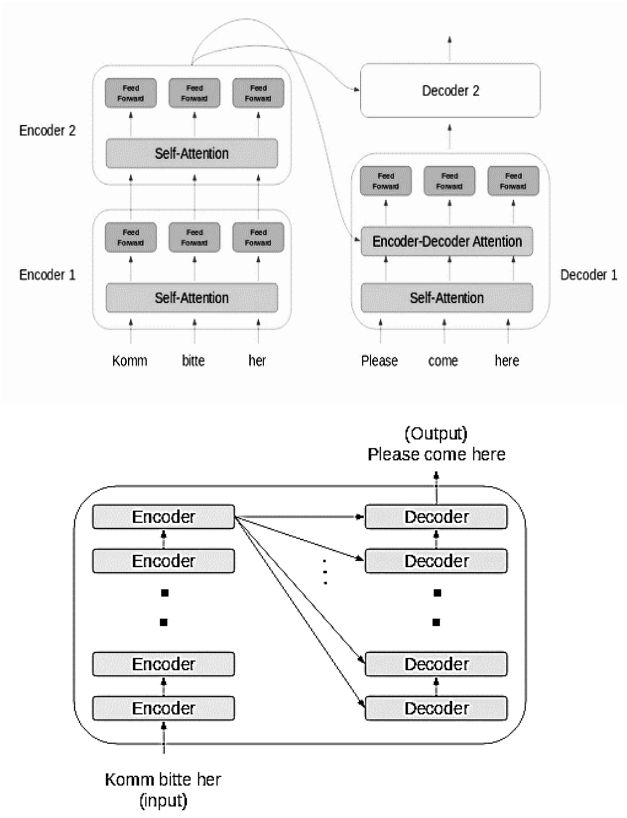
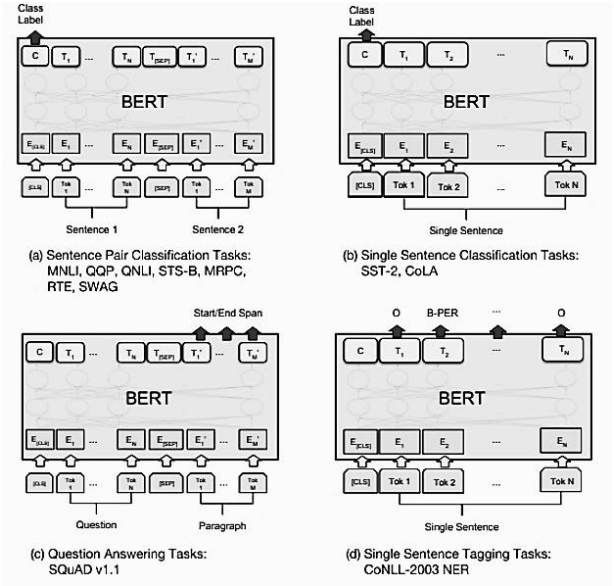


Figure 9. An illustration of Encoder-Decoder Stack and Self-Attention Model [32]

F. BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS

Pre-training and fine-tuning are used in the BERT framework, a new language representation model from Google AI, to develop state-of-the-art models for a variety of tasks. Question answering systems, sentiment analysis, and linguistic inference are examples of these tasks. BERT has two different techniques: Masked Language Modeling (MLM), Next Sentence Prediction. The goal of the masked language model is to predict the original vocabulary id of the masked word based only on its context after randomly masking some of the tokens from the input. The MLM aim, unlike left-to-right language model pre-training, allows the representation to integrate the left and right contexts, allowing us to pre-train a deep bidirectional Transformer.



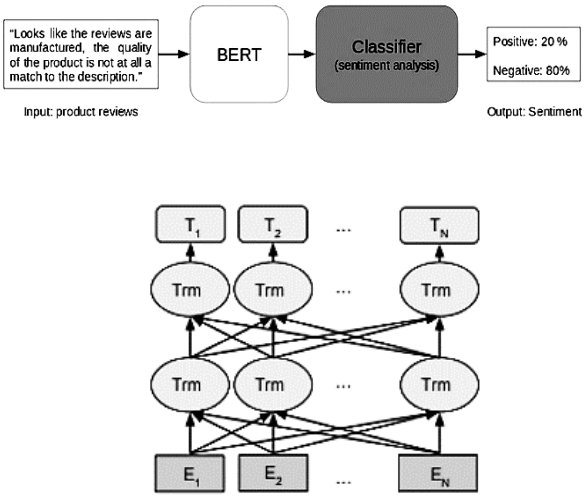


Figure 10. An illustration of BERT Architecture, Sentimental Analysis Task, NLP tasks for BRET [32]

VI. CONCLUSION

After Eighty years of a long journey, Natural Language Processing comes through several Ups and Downs. Several failures and lack of proper machines could not stop the progress of NLP research. Without NLP it is impossible to think about today's world and thanks to the researchers who believed in NLP. After BERT Natural Language Processing is in another level with leads researchers more interest in this field as a major area of research and development. This Glorious journey not only gives us the historical aspects but also gives us the hope for future development of NLP.

Natural language processing is used by most AI systems today, including Google Assistant, Netflix, Apple's Siri, and Grammarly. There is a theory that as Natural Language Processing and Biometrics improve, computers, such as humanoid robots, will be able to read facial expressions, body language, and speech, such as when a person is speaking face to face. Because they can serve as the physical body for a programmed artificial soul, humanoid robots are required for this form of communication. NLP and Biometrics will be able to take Humanoid robot

research to a whole new level as their popularity and accuracy grows, allowing them to communicate themselves through movement, postures, and expressions. As a result, despite any current limitations, we can expect many of these barriers to be dismantled in the coming years as new techniques and technology emerge on a daily basis.

VII. REFERENCES

[1] Schank, R. C., Conceptual Information Processing, Amsterdam, North Holland, 1975.

[2] Siemens AG (ed) Verbmobil: Mobiles Dolmetschgerät; Studie, Siemens AG, München, 1991.

[3] Sparck Jones, K. "Natural language processing: an overview", International encyclopedia of linguistics (ed W. Bright), New York: Oxford University Press, Vol. 3, pp. 53-59, 1992.

[4] Sparck Jones, K. "Thesaurus", Encyclopedia of artificial intelligence (ed Shapiro), 2nded, New York: Wiley, pp. 1605-1613, 1992.

[5] Walker, D.E. "SAFARI: an on-line text-processing system", Proceedings of the American Documentation Institute Annual Meeting, pp. 144-147, 1967.

[6] Winograd, T. "A procedural model of language understanding", pp. 249-266, 1973.

[7] Woods, W.A. "Semantics and quantification in natural language question answering", pp. 205-248, 1978.

[8] Yngve, V.H. "MT at MIT", pp. 451-523, 1967.

[9] Hutchins, W.J. Machine translation, Chichester, England: Ellis Horwood, 1986.

[10] Hutchins, W.J. and Somers, H.L. An introduction to machine translation, London: Academic Press, 1992.

[11] Jacobs, PS. (ed) Text-based intelligent systems, Hillsdale, NJ: Lawrence Erlbau Associates, 1992.

[12] Joshi, A.K., Webber, B.L. and Sag, I. A. (eds.) Elements of discourse understanding, Cambridge: Cambridge University Press, 1981.

[13] Kay, M., Gawron, J.M. and Norvig, P. Verbmobil: a translation system for faceto- face dialogue, CSLI, Stanford University, 1991.

[14] Kittredge, R. and Lehrberger, J. (eds.) Sublanguage: studies of language in restricted semantic domains, Berlin; Walter de Gruyter, 1982.

[15] Kobsa, A. and Wahlster, W. (eds.) User modelling in dialogue systems, Berlin: Springer-Verlag, 1989.

[16] Lea, W.A. (ed) Trends in speech recognition, Englewood Cliffs, NJ: Prentice-Hall, 1980.

[17] Locke, W.N. and Booth, A.D. (eds.) Machine translation of languages, New York: John Wiley, 1955.

[18] McKeown, K.R. Text generation, Cambridge: Cambridge University Press, 1985.

[19] Minsky, M. (ed.) Semantic information processing, Cambridge, 1968.

[20] Minsky, M., "A framework for representing knowledge," (ed Winston, P.), The psychology of computer vision, McGraw-Hill, 1975.

[21] Nagao, M. (ed) A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC, USA, Japan Electronic Industry Development Association, 1989.

[22] Plath, W "Multiple path analysis and automatic translation", in Booth, pp. 267-315, 1967.

[23] Reifler, E. "Chinese-English machine translation, its lexicographic and linguistic problems", in Booth, pp. 317-428, 1967.

[24] Rumelhart, D.E., McClelland, J.L. and the POP Research Group, Parallel distributed processing, in Cambridge, vol. 2, 1986.

[25] Rustin, R. (ed) Natural language processing, New York: Algorithmics Press, 1973.

[26] HLT: Proceedings of the ARPA Workshop on Human Language Technology, March 1993; San Mateo, CA: Morgan Kaufmann, in press.

[27] Dahl D, “Natural Language Processing: Past, Present and Future”, Springer New York, pp. 49-73, 2013.

[28] Diksha Khurana, Aditya Koli, Kiran Khatter, Sukhdev Singh, "Natural Language Processing: State of The Art, Current Trends and Challenges"

[29] Nation, K., Snowling, M. J., & Clarke, "Dissecting the relationship between language skills and learning to read: Semantic and phonological contributions to new vocabulary learning in children with poor reading comprehension. Advances in Speech Language Pathology", pp. 131-139, 2007.

[30] Feldman, "NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval.", ONLINE-WESTON THEN WILTON-, vol. 23, pp. 62-73, 1999.

[31] https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/

[32] https://medium.com/analytics-vidhya/neural-machine-translations-implementing-encoder-decoder-658c3facd530

[33] Hutchins, W. J., "Early years in machine translation: memoirs and biographies of pioneers", Vol. 97, 2000.

[34] Woods, W. A., "Semantics and quantification in natural language question answering.", vol. 17, pp. 1-87, 1978.

[35] Kamp, H., & Reyle, "Tense and Aspect. In From Discourse to Logic", Springer Netherlands, pp. 483-689, 1993.

[36] Eugene Charniak and Drew McDermott, Introduction to Artificial Intelligence , Pearson, 1998, Chapter 4.

[37] K.R. Chowdhary Professor & Head CSE Dept. M.B.M. Engineering College, Jodhpur, India. April 29, 2012 Natural Language Processing.

[38] M. Li, S.J. Liu, D.D. Zhang and M. Zhou, Machine Translation, Beijing:Higher Education Press, 2018.

[39] Y. Wang, "Natural language processing and applications in machine learning", Modern Chinese, vol. 5, pp. 187-191, 2019.

[40] Y.F. Song, "The development history and current situation of natural language processing", China High-Tech, vol. 3, pp. 64-66, 2019.

[41] M. Wang, S.W. Yu and X.F. Zhu, "Natural language processing and its applications in education", Mathematics in Practice and Theory, vol. 40, no. 20, pp. 151-156, 2015.

[42] K.B. Hu and Y. Li, "The features of machine translation and its relationship with human translation", Chinese Translators Journal, vol. 37, no. 5, pp. 10-14, 2016.

[43] Z.W. Feng, "Parallel development of machine translation and artificial intelligence", Journal of Foreign Languages, vol. 41, no. 6, pp. 35-48, 2018.

[44] Z.W. Feng, "Computational linguistics: its past and present", Journal of Foreign Languages, vol. 34, no. 1, pp. 9-17, 2011.

[45] Z.H. Zhou, Machine Learning, Beijing:Tsinghua University Press, 2016.

[46] W.J. Hutchins, Machine Translation: Past Present Future, Chichester:Ellis Horwood Limited, 1986.

[47] Rajarshi Sinharoy, Swarnendu Sarkhel, "Air Quality Index Prediction in Realtime Using SVM based model in Machine Learning", International Journal of Innovative Research in Physics, vol. 3, pp. 43-49, 2021.

[48] Q.P. Jiang, "Challenges and future of natural language processing", China Computer & Communication, vol. 14, pp. 219-221, 2013.

[49] B. Manaris, "Natural language processing: a human-computer interaction perspective", Advances in Computers, vol. 47, pp. 1-66, 1998.

[50] Y. Bar-Hillel, "The present status of automatic translation of languages", Advances in Computers, vol. 1, pp. 91-163, 1960.

[51] I. Sutskever, V. Oriol and V.L. Quoc, "Sequence to sequence learning with neural networks", Advances in Neural Information Processing Systems, vol. 4, pp. 3104-3112, 2014.

[52] P.E. Brown, J.D. Vincent, A.D. Stephen and L.M. Robert, "The mathematics of statistical machine translation: parameter estimation", Computational Linguistics, vol. 19, no. 2, pp. 263-311, 1993.

[53] Young, S. J., & Chase, "Speech recognition evaluation: a review of the US CSR and LVCSR programmes", pp. 263-279, 1998.

Cite this article as :

Rajarshi SinhaRoy, "A Study on the journey of Natural Language Processing models: from Symbolic Natural Language Processing to Bidirectional Encoder Representations from Transformers", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 6, pp. 331-345, November-December 2021. Available at doi : https://doi.org/10.32628/CSEIT217688
Journal URL : https://ijsrcseit.com/CSEIT217688