A Practical Approach TO

Natural Language Processing (NLP)

Hariom

19 October 2022

Table Of Content:

Motivation

We as a human communicate in various ways like; speaking, making gestures, listening, and using sign language or more important in a text form. So let me try another way. Every day we travel from one place to another place and we see traffic lights, how traffic police communicate with the drivers, and how different symbols and colors of lights control your driving.

Or Whenever I watch the news on Doordarshan (DD News) there is a person in the corner of the screen reading the news in sign language, even though I have never understood it but I know it is helpful for the deaf-persons.

Therefore, language is an essential tool for human survival. In the evolution of humans, there are some natural languages (i.e., language which evolved naturally like: Body Language) and some human build languages (i.e., which we use to write).

Now as a human, we can understand voice and text languages in different forms (e.g., phrase, one word, or story form) but computers do not work like this. Computers cannot understand these languages as a human.

To build these capabilities in machines humans started working. We as humans use mostly two mediums of communication: first is *text or written* and the other is *voice or spoken.* Now we have started to make machines that can interact with humans in natural languages.

This evolution drives us to an exciting and revolutionizing branch of *computer science* i.e., *Natural Language Processing*. So, the question is: what is Natural Language Processing?

Introduction

What are Natural Language Processes? Well, you have already identified if you are reading the above part carefully. So, in short, NLP is the branch of computer science more accurately NLP is the branch of *Artificial Intelligence (AI)* that provides the ability to analyze and reply the text and voice data machines. Here we have used the new term *Artificial Intelligence (AI)*, so let us define it firstly:

What is Artificial Intelligence (AI)? The Term Artificial Intelligence was first coined by *John McCarthy* in 1965. This is a branch of computer science. The goal of AI is to make capable machines with human-like intelligence.

Now coming to our topic after understanding the need for NLP the question is: How does NLP work? Before this, we have to understand what language means.

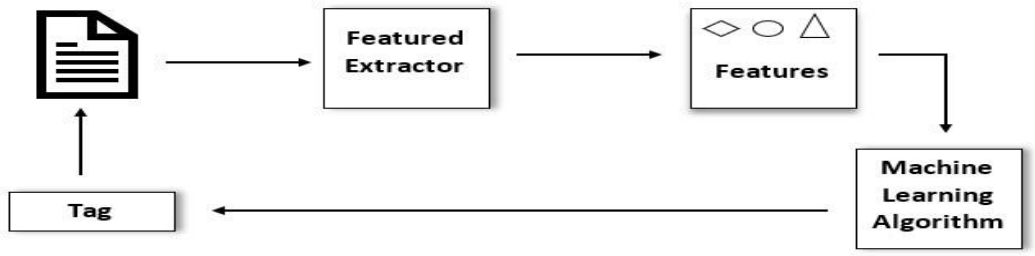
What is a language? A language is a system of communication. Every time we speak or write is just a set of sounds and written symbols that are used by us for talking and writing.

There are different types of languages. In India, we say *after every 15-20 KM language changes.* But for any language there are some sets of rules to combine the words and form the sentences called ***Syntax,*** the meaning of these words and sentences in a language is called ***Semantics*** and ***Pragmatics*** enables us to apply the correct meaning to the correct situation.

The study of the spoken sound system of language with its rules is called ***Phonology*** and the study of rules that includes the minimal meaningful units of language is called ***Morphology***.

We do the same tasks in NLP as described in these feature definitions for our machines to interact with humans. Now we are ready to explore, how does NLP work?

How does NLP work? To understand the process let us look at the simple diagram:



Let's take text as input and then use *text vectorization (i.e., NLP tool transforms text into machine language)*, the machine learning algorithms are fed training data (mostly 80% input data) and expected outputs (i.e., Tag) to train machines to make associations between a particular input and its corresponding output. Machines then use statistical analysis methods to build their own "knowledge bank" and discern which features best represent the texts, before making predictions for unseen data. Ultimately, the larger the data fed into NLP algorithms the better the text analysis result will be.

After understanding the basic overview of how NLP works for text data the next question arises which type of tasks NLP performs?

NLP Tasks

What are the NLP tasks? Before answering this question just recall the purpose of NLP, have you remembered? Let me write; To *provide the capabilities to machines so that they can understand human languages text and voice mainly.*

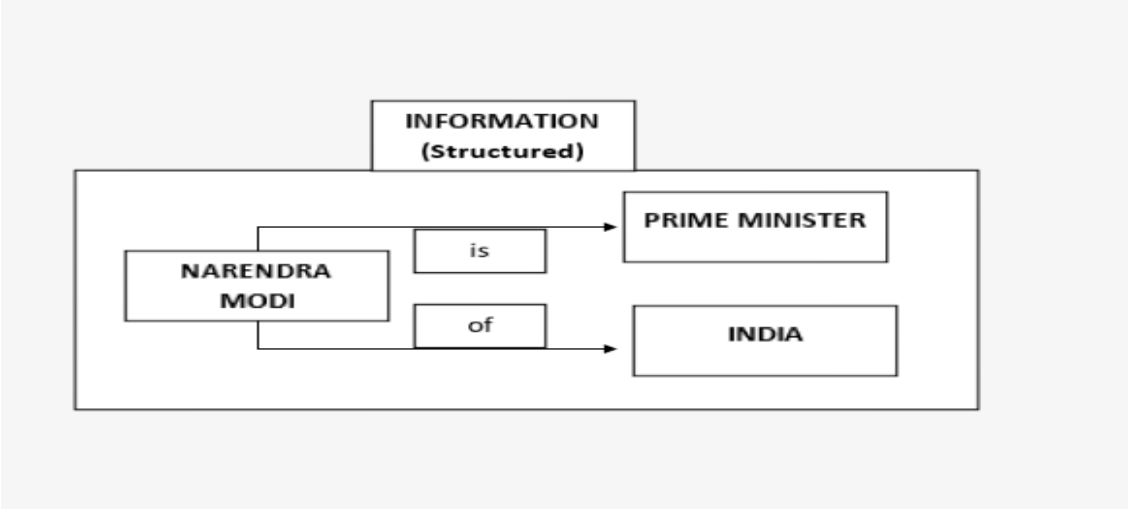
Now the main task is to break down the text and voice data in a way that helps the machines make sense of what it's ingesting. Some of the tasks are given below:

- ***Information Extraction:*** As the name suggested it is a task to extract information from the data automatically with the help of machines. It seems interesting but yes, we use NLP to provide this type of feature to a machine.

For Example:

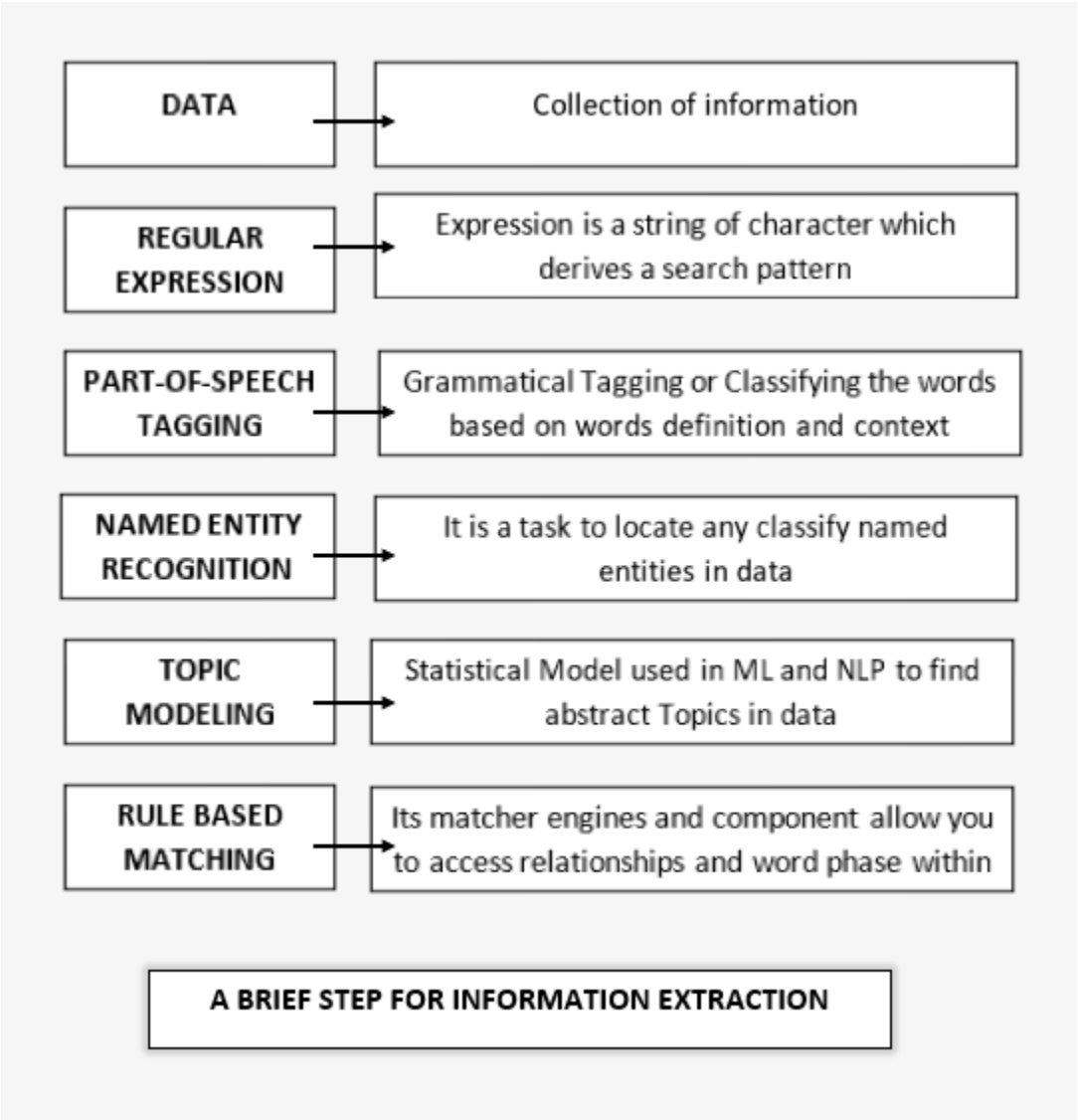
Text- **“Narendra Modi is the prime minister of India.”**

Information Extraction-



But the question is how it is useful in the real world. In short, businesses have a large amount of data about their customers and they are always keen to know their customer’s behavior, the number of satisfied and unsatisfied customers, and the reason behind this.

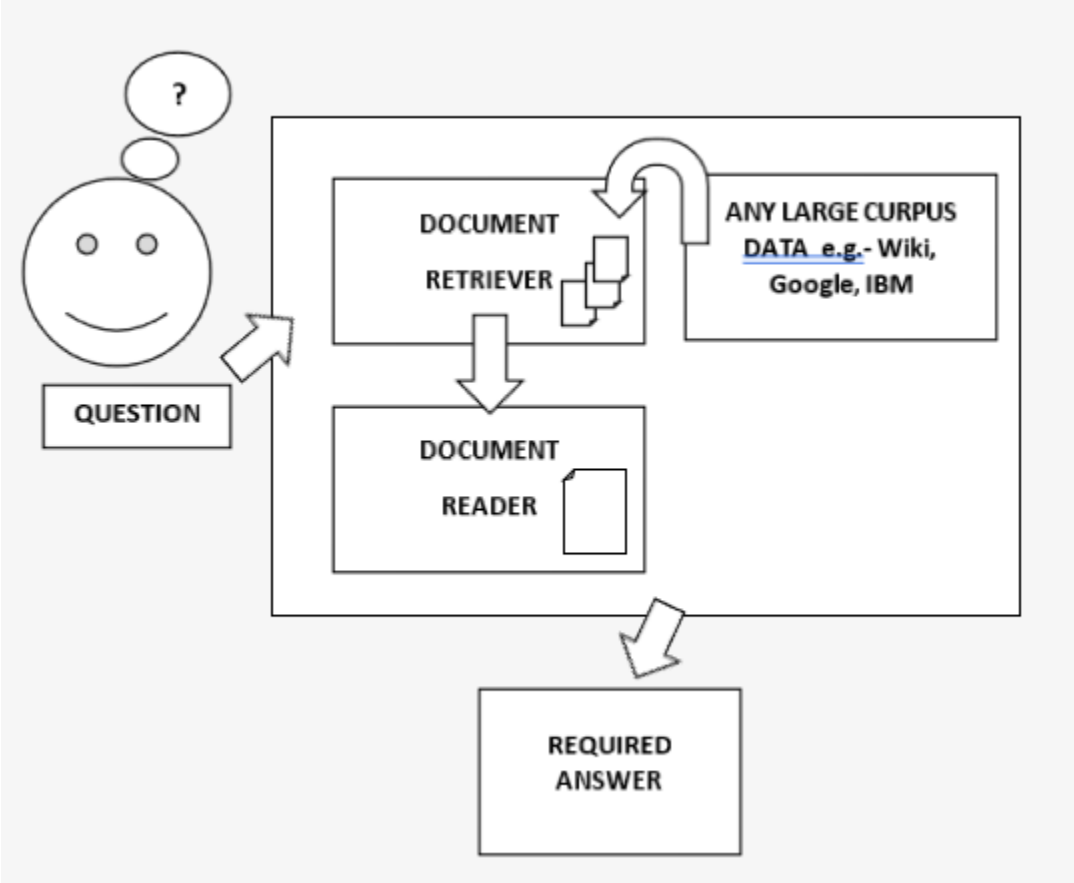
Well, here is some sentimental analysis required but most of the part is information extraction and we can use NLP to solve this problem. Let us see the simple steps:



See these steps look easy but in practice, it comes with many issues:

- ➔ Combining data from multiple sources.
- ➔ Context and relationship of pieces of data cannot be inferred at all in this process.
- ➔ Unstructured data is a very big issue when it comes to information extraction.
- ➔ Quality and veracity of data matters when it comes from social media etc.

- ***Question-Answering (QA)*:** Every time I use any reputable websites, I always find an assistant asking me to help. One day I was using the IRCTC (*Indian railway site*) website to book my tickets I want to save my details forever in my IRCTC account for the future but I don't know how. So, I asked the IRCTC assistant (***AskDisha2.0***) and she literally provided me with exact steps to save my information. It was an interesting experience but the question is how it works. Well, it's a system of *question-answering,* built using NLP. It allows us to ask questions in our languages and provides immediate answers. Here are simple steps on how it works:



It looks like a simple design where you are asking questions and getting answers but the system has to deal with vast data and it's critical to find the shortest fragment of text in the question that answers the query. It's critical to model such a system to deal with unstructured questions and provide the best answer from all potential spans.

- ***Machine Translation:*** We always see when one country's leader goes to another country. There is a translator with them to translate the conversation and help in communicating with each other. A translator is a crucial link between these types of meetings. Now we see in the ***United Nations (UN)*** or other forums where approximately all countries' world leaders come to participate. It's hard and costly to provide a translator for each country's leaders but it was happening in the past now NLP has solved this problem and there are systems that help to translate different languages into the required language. Well, there is still a need for improvement and advancement and we are continuing to do this for now widely used software is ***Google Translator (by Google)*** to translate human languages.
- ***Speech Recognition*****:** This is a voice-to-text transformation task. It helps follow voice commands or answer spoken questions. Well, it's challenging when it comes to humans' natural talk because it's quick, with varying emphasis, different accents, or maybe incorrect grammar.
- ***Sentimental Analysis*****:** Here we work on subjective qualities like attitude, emotion, confusion, suspicion, etc. from the text. It is mostly used on social media platforms like *Twitter, Instagram, Facebook, etc.*
- ***Natural Language Generation*****:** It is just opposite to speech recognition i.e. here the task is to convert the text into speech. We can see it in *Siri (Apple), Alexa (Amazon), and Google duo (Google)*. These are the most popular devices revolutionizing the human experience.
- ***Part-Of-Speech (POS)*****:** This one is also called grammatical tagging. It is a process of identifying the part of speech of a particular word or sentence based on its use and context.

There are many more tasks but now we have to move to our next curiosity i.e. which tools or methods are used to make these tasks possible?

NLP Tools

What are the NLP tools? Well, this question is very subjective means it depends on the projects on which you are working or the problem you are trying to solve, or the system you are trying to build. I prefer *Python (Jupyter Notebook)* because of my work domain but you can use different tools for working. Here is a list of tools:

1. Natural Language Toolkit (NLTK)
2. Spacy
3. TextBlob
4. Google Cloud
5. Amazon Comprehend
6. IBM Waston
7. Aylien
8. Stanford Core NLP, etc.

Now the question is how these whole efforts are beneficial for us i.e., what is the use of NLP in our real world?

NLP Use Cases

Where does the NLP use? Simply it's a never-ending process, once you start using NLP you can be equally creative as an artist in the tech world. Here are some used cases:

- ***Virtual Assistance*:** Have you heard about Apple's Siri or Amazon's Alexa or Google Duo? These are all 21st-century devices, revolutionizing the world and human experiences. These devices use an NLP method called *speech recognition* to recognize patterns in the voice command and natural language generation to reply humanely.
- ***Spam Detection*:** Well, you can think about how NLP is used in spam detection but the truth is Google and other companies use NLP's text classification capabilities to scan phishing.
- ***Text Summary Generation*:** When it comes to reading books most people fear the huge number of pages but the good thing is NLP techniques to digest the huge pages into short summaries without losing the actual context of the book.

Importance of NLP

After reading all this we can feel the importance of NLP but for a short note just imagine:

The software speaks any language in just one click which you are not comfortable means NLP can work as a personal assistant for you. You just need to speak in your language and it will translate into any language according to you. You can imagine how it revolutionizes business operations and customer experience.

But, there are a lot of challenges too.

Challenges in NLP

Here we are explaining some of the challenges:

- ***Ambiguity*:** we always use some words which have two or more meanings this is called ambiguity. **E.g. "I saw the boy on the beach with my binoculars."**
  This could mean that I saw a boy through my binoculars ***or that*** the boy had my binoculars with him. There are different types of ambiguity:

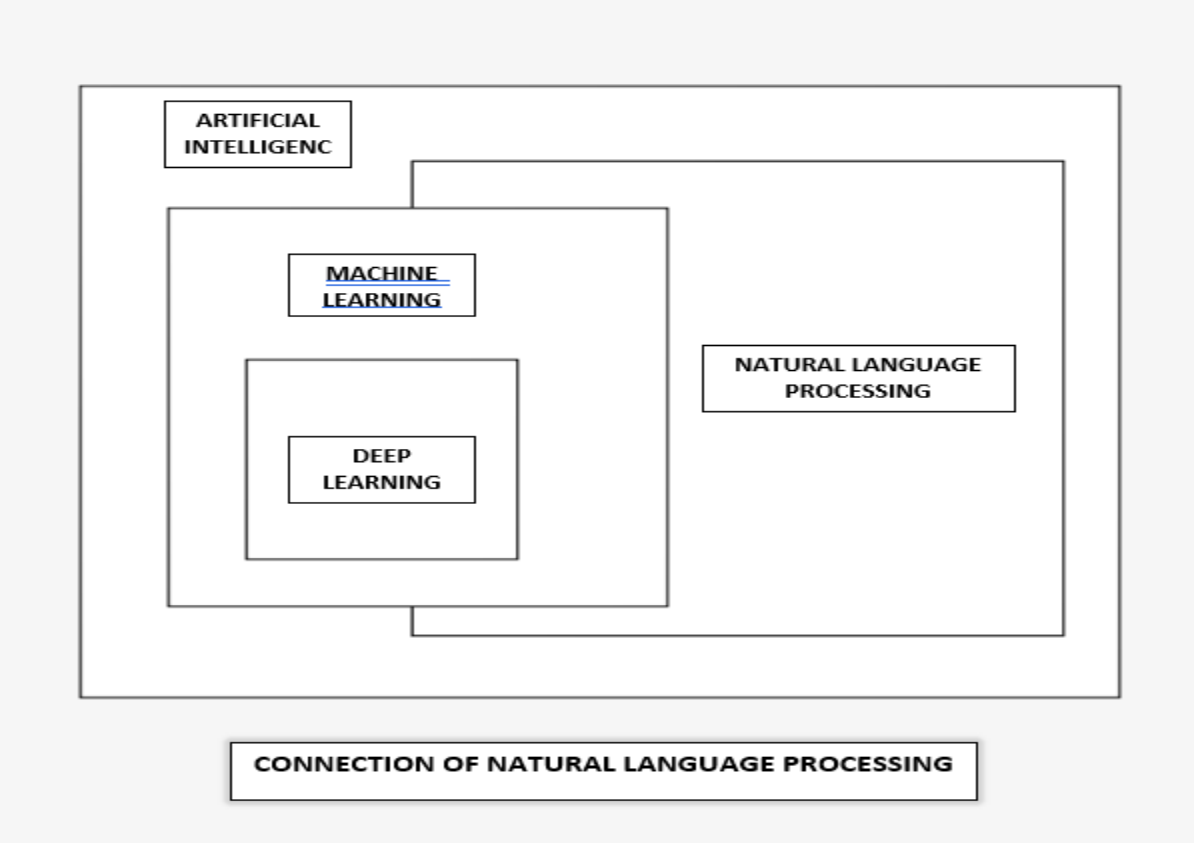
  - ➢ **Semantic Ambiguity:** There are two ways to read a sentence that is called semantic ambiguity. E.g.- *"John kissed his wife, and so did Sam"* - (Sam kissed John's wife or his own)
  - ➢ **Anaphoric Ambiguity:** A word refers to something previously but there is no more than one possibility. E.g.- *"I went to the hospital, and they told me to go home and rest"* (They = Hospital staff)
  - ➢ **Syntactic Ambiguity:** A sentence has multiple parse trees called syntactic ambiguity. E.g.- *"He puts the ketchup on himself."*
  - ➢ **Lexical Ambiguity:** A word has multiple meanings called lexical ambiguity. E.g. - *"I saw a bat."* (bat = flying mammal / wooden club) etc.

- ***Synonyms*:** We know many words express the same idea. When it comes to training the machines for this type of challenge it is a bit tricky and not easy. So, building an NLP system e needs to include all possible synonyms.
- ***Errors In Text and Speech*:** Humans are known for their mistakes and when it comes to text or speech it is very much possible to misspell and that causes big confusion for the machine. So, building an NPL system we need to take care of this.
- ***Sarcasm*:** It can create problems for machines' learning models. We always try to feed extra meaning to words or sentences or phrases in the system but sarcastic words can create troubles for the system.

We have discussed almost every required point. Now let's suppose that we trained the machines such that they can learn by themselves the required tasks to perform and give us the result. It is the same when we learn any skill and start exploring and modifying as we want. Here comes Machine Learning. Now we see how ML strengthens the NLP.

The Role of Machine Learning

Let us see the broader picture of different domains and how they are inter connected:



Using ML, we provide machines the ability to automatically learn and improve from experience without being explicitly programmed. ML helps to improve NLP by automating the process and delivering the required outputs. There are mainly four types of ML processes in NLP:

1. **Supervised Machine Learning:** Here we train the machines using the "labeled" dataset, and based on the training, the machine predicts the output.
   The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y). Some real-world applications of supervised learning are Risk Assessment, Fraud Detection, Spam filtering, etc.

2. **Unsupervised Machine Learning:** Here machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision.

The main aim of the unsupervised learning algorithm is to group or categorize the unsorted dataset according to the similarities, patterns, and differences. Machines are instructed to find the hidden patterns from the input dataset.

3. **Semi-Supervised Machine Learning:** It is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning.
   The main aim of semi-supervised learning is to effectively use all the available data, rather than only labeled data like in supervised learning. Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labeled data. It is because labeled data is a comparatively more expensive acquisition than unlabeled data.

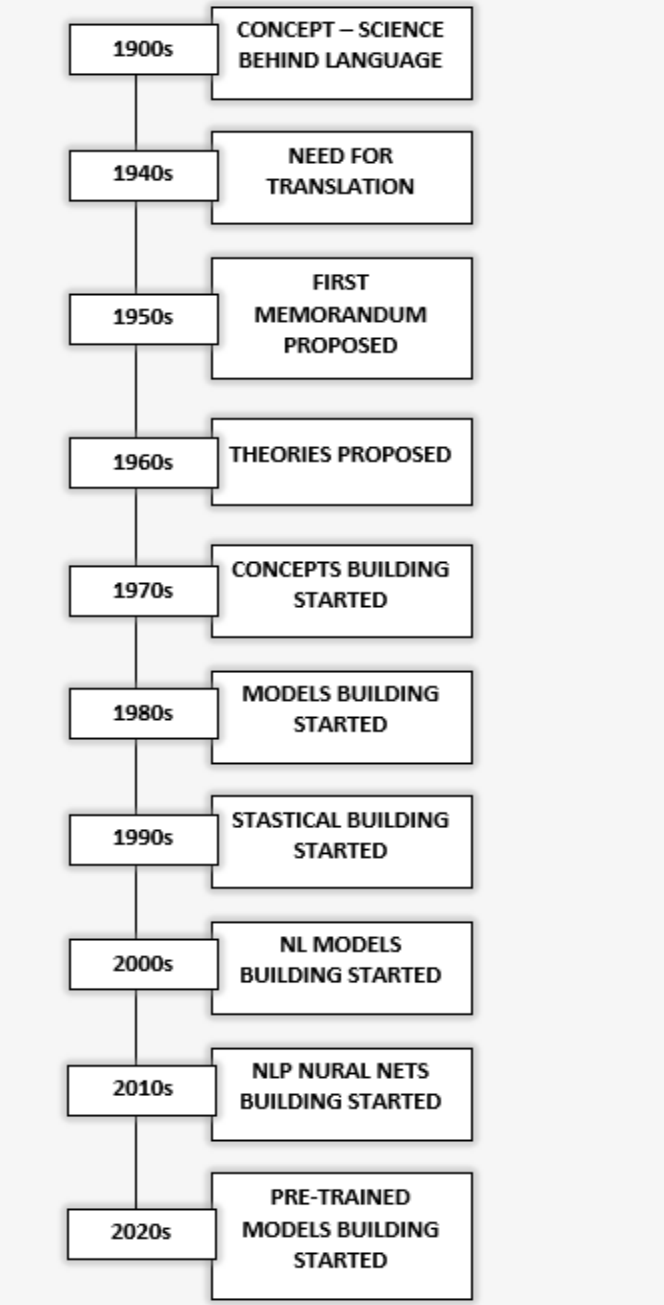
4. **Reinforcement Learning:** It works on a feedback-based process, in which an AI agent (A software component) automatically explores its surroundings by hitting & trail, taking action, learning from experiences, and improving its performance.

There is a lot to know about these topics but now let's explore the history of NLP in brief.

History Of NLP

It all started in the early 1900s when a Swiss professor named ***Ferdinand De Saussure*** of linguistics proposed courses at the *University of Geneva* titled ***"Languages as a Science".*** It was just the beginning as a subject but the golden era of NLP started after the Second *World War (the 1940s)* when people recognized the importance of translation from one language to another. They started experimenting and exploring this field in the hope that they could create a machine to translate these languages automatically.

Well, it looked imaginary at that time but now it's reality and we used far beyond the translation in different domains to solve different real-world problems. Let see the evolution through the diagram given below:


1900s
CONCEPT – SCIENCE BEHIND LANGUAGE
1940s
NEED FOR TRANSLATION
1950s
FIRST MEMORANDUM PROPOSED
1960s
THEORIES PROPOSED
1970s
CONCEPTS BUILDING STARTED
1980s
MODELS BUILDING STARTED
1990s
STASTICAL BUILDING STARTED
2000s
NL MODELS BUILDING STARTED
2010s
NLP NURAL NETS BUILDING STARTED
2020s
PRE-TRAINED MODELS BUILDING STARTED
EVOLUTION OF NATURAL LANGUAGE PROCESSING

Now it’s time to conclude our exploration:

Conclusion

We have seen the need for NLP and explored its different expectations of it. Well, It has limitations but still, it offers a midrange of benefits to any domain in the real world.
Thank You!

References:

- Natural Language Processing tutorial by Tutor Point
- Handbook of Natural Language Processing by Nitin Indurkhya and Fred J. Damerau
- NLP thesis on https://www.academia.edu
- Concept building with the https://archive.nptel.ac.in/courses
- Wikipedia: wikipedia.org
- Google Search engine: www.google.com
- Google Doc: docs.google.com
- GitHub Resources: github.com
- ScienceDirect Topics: www.sciencedirect.com
- harry Clark Translation Service: harryclarktranslation.co.nz
- Stemming and lemmatization: nlp.stanford.edu
- Linguistic Knowledge bank
- Typesetting with MS word