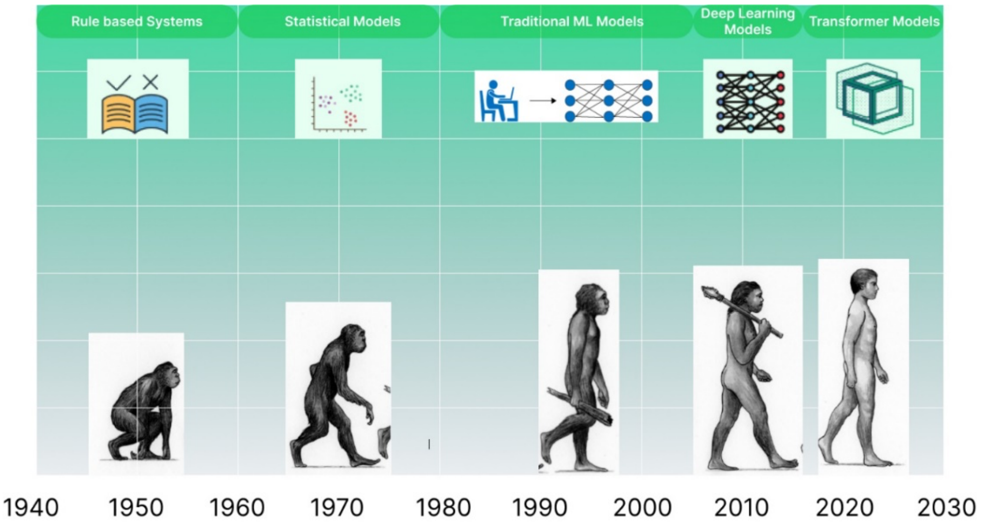
What is Natural Language Processing and Why is it Important

Natural Language Processing (NLP) is a subfield of AI that deals with interactions between computers and human languages. NLP algorithms help computers understand, interpret, and generate natural language. (ref: https://www.linkedin.com/pulse/from-rulesets-transformers-journey-through-evolution-sota-yeddula/)



NLP's inception dates back to the 1950s, and these early models were based on rule-based systems. And these models relied on manually crafted rules to process and translate language. The primary application of these systems was mainly translation services. In 1954, IBM achieved a milestone by utilizing a computer to translate 60 Russian sentences into English through a set of meticulously designed rules.

Moving into the 1970s and 1980s, a paradigm shift occurred with the rising prominence of statistical models and machine learning algorithms. During this period, one noteworthy model was the Hidden Markov Model (HMM), which found early success in speech recognition applications. This transition marked a pivotal moment in the evolution of NLP, paving the way for more dynamic and data-driven approaches in language processing.

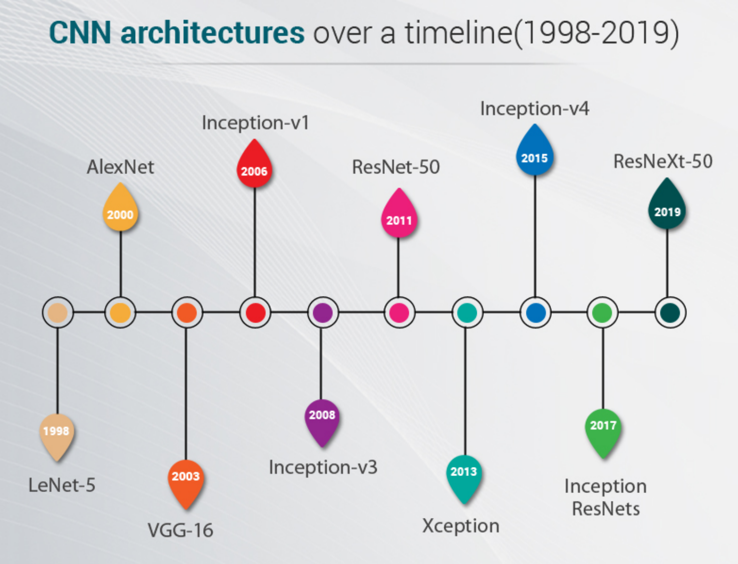
The rise of the internet and the explosion of new digital data created new challenges and opportunities for NLP. More robust models were required which could process large amounts of data, and then index and search it efficiently. Google, Yahoo and Meta all pioneered research in this field. The primary use cases included web searches, information retrieval and text mining. Although these were still basically statistical models, a transitional was occurring, with statistical methods dominating but gradually giving way to the rise of machine learning and neural-based approaches.

The Birth of Neural Networks

In the late 2000s, there was a need to solve more complex and non-linear tasks, and hence the development for machine learning evolved. The birth of neural networks was initiated which took an approach akin to the human brain, by structuring the way to solve problems with algorithms modelled on the brain. With the rise of deep learning (deep learning means multiple levels of neural networks) and neural networks, models such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) began to be used in NLP.

RNNs are a type of neural network that can handle sequential data by processing inputs one at a time while also maintaining an internal memory of previous inputs, whereas CNNs are a type of neural network that are particularly effective for processing images and other two-dimensional data by using convolutional layers to learn features and patterns in the data.

AlexNet was one of the pioneering CNNs that was introduced by Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton in 2012 and was the winner of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) that year. Having 8 layers (5 convolutional layers and 3 fully connected layers), it was one of the first deep neural networks. AlexNet was trained on two GPUs, which was novel at the time and allowed for faster training and achieved a top-5 error rate of 15.3%, which was a significant improvement over the previous state-of-the-art at the time. Its success demonstrated the potential of deep learning for computer vision tasks and sparked a resurgence of interest in convolutional neural networks.



One of the main use cases was Named Entity Recognition (NER), which is a form of natural language processing (NLP) that involves extracting and identifying essential information from text. The information that is extracted and categorized is called entity and can be any word or a series of words that consistently refers to the same thing.

Introduction of Transformer Models

The Transformer Model was introduced in 2017, in a paper titled “Attention is All You need”, published by Google researchers. Neural networks represented a major obstacle where traditional RNNs still processed inputs sequentially, which can be slow and computationally expensive. Thus, was born the Transformer model, which improved upon the existing RNNs by processing the input text all at once, rather than one word at a time like older RNN models.

The Transformer model is known for its attention mechanism, which allows it to capture relationships between words in a sequence more effectively compared to previous recurrent or convolutional neural network architectures. This attention mechanism enables the model to process input sequences in parallel, making it highly efficient and scalable. This makes it highly suitable for these ChatBOTs and language modelling.