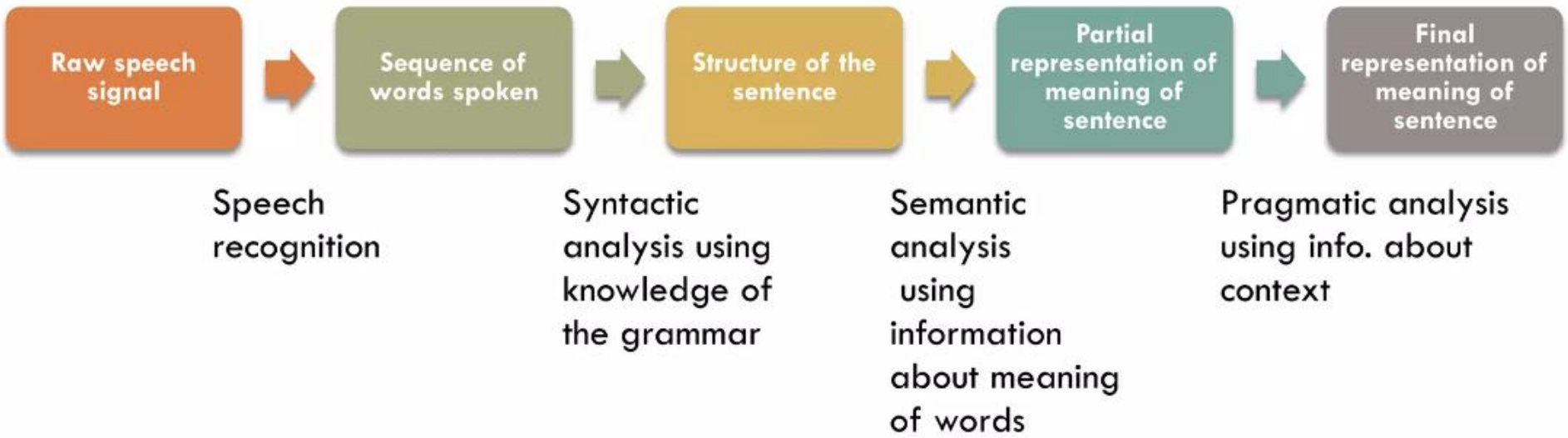
Natural Language Processing

Introduction

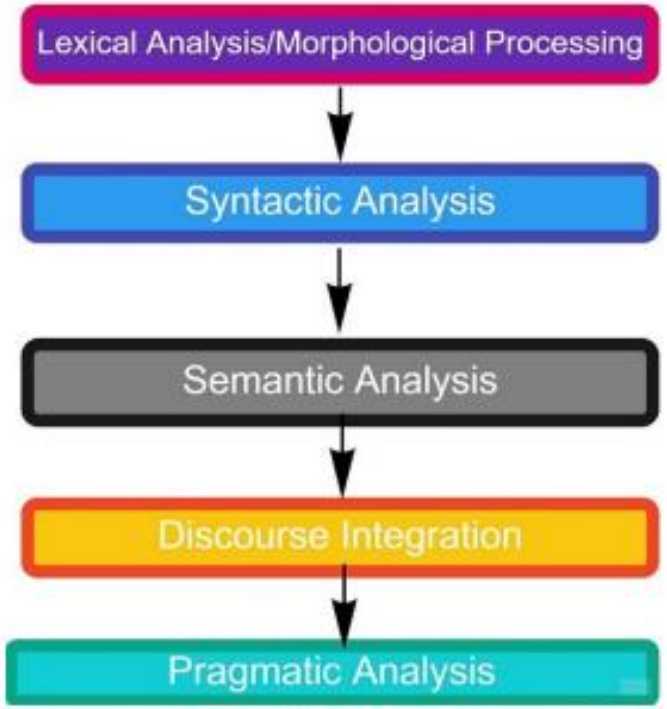
- Natural Language Processing(NLP) is defined as the branch of Artificial Intelligence that provides computers with the capability of understanding text and spoken words in the same way a human being can.
- It incorporates machine learning models, statistics, and deep learning models into computational linguistics i.e. rule-based modeling of human language to allow computers to understand text, spoken words and understands human language, intent, and sentiment

- Humans communicate with each other using words and text. The way that humans convey information to each other is called Natural Language. Every day humans share a large quality of information with each other in various languages as speech or text.
- However, computers cannot interpret this data, which is in natural language, as they communicate in 1s and 0s. The data produced is precious and can offer valuable insights. Hence, you need computers to be able to understand, emulate and respond intelligently to human speech.
- Natural Language Processing or NLP refers to the branch of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages

Natural language understanding



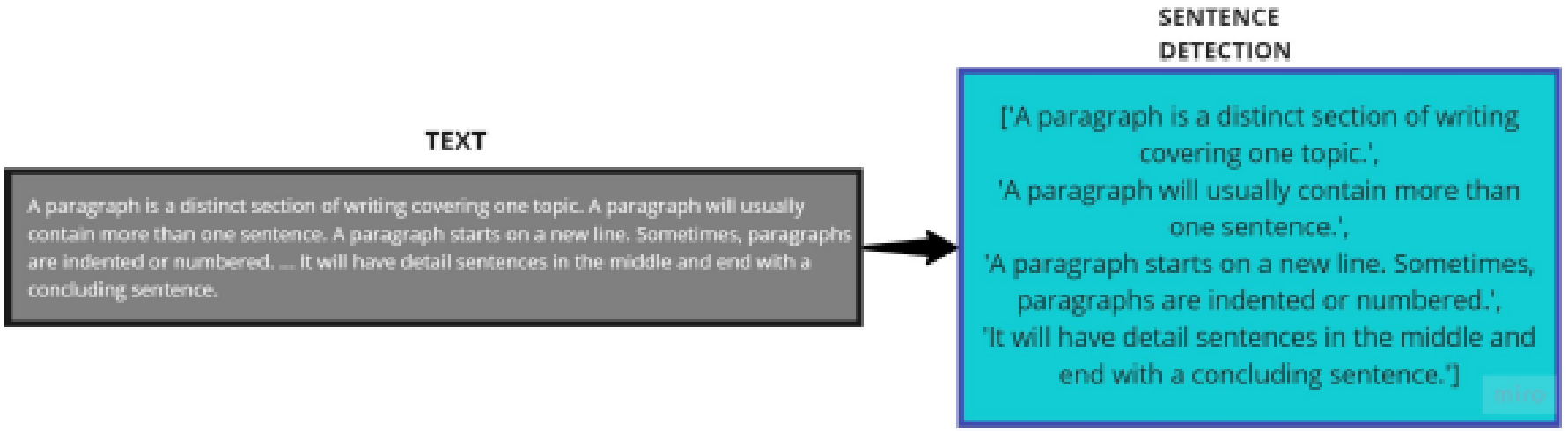
Phases


Lexical Analysis/Morphological Processing
Syntactic Analysis
Semantic Analysis
Discourse Integration
Pragmatic Analysis

Lexical Analysis

- The first phase is lexical analysis/morphological processing. In this phase, the sentences, paragraphs are broken into tokens.
- These tokens are the smallest unit of text. It scans the entire source text and divides it into meaningful lexemes.
- For example, The sentence “He goes to college.” is divided into [‘He’ , ‘goes’ , ‘to’ , ‘college’, ‘.’] .
- There are five tokens in the sentence. A paragraph may also be divided into sentences

Lexical Analysis



Syntactic Analysis/Parsing

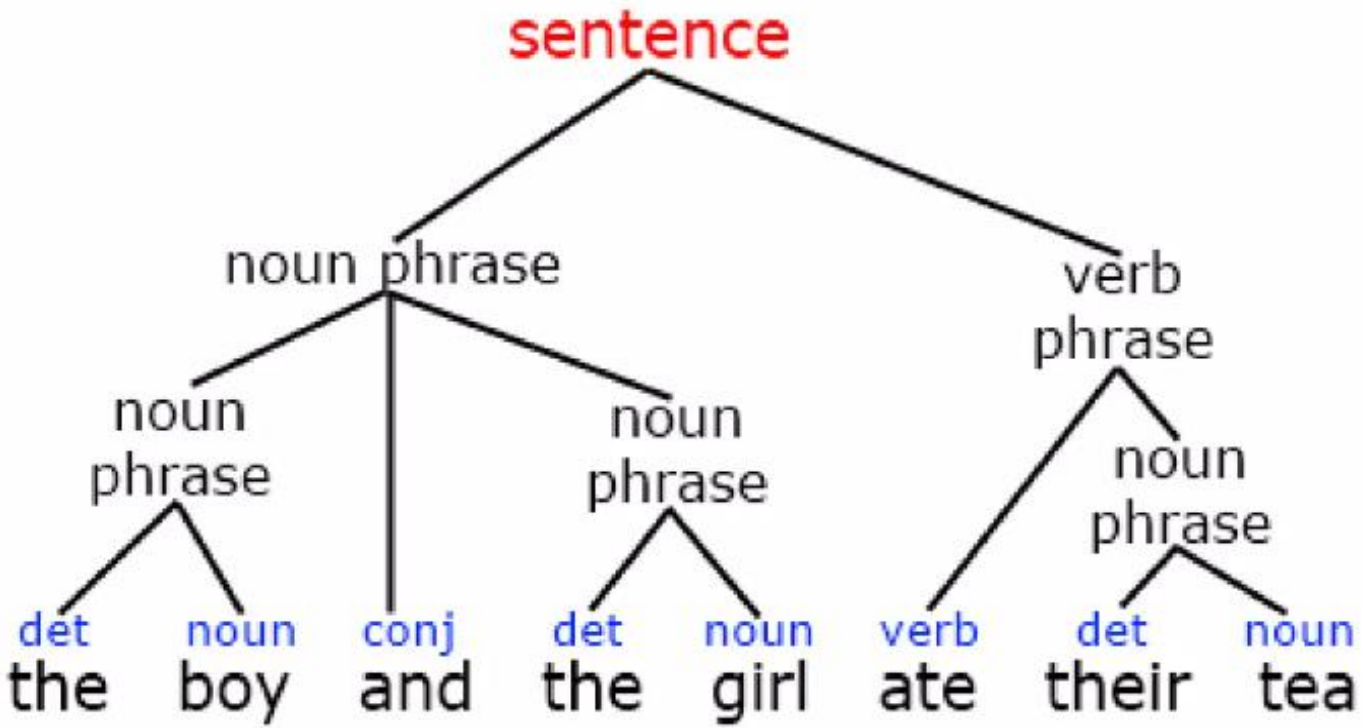
- The second phase is Syntactic analysis. In this phase, the sentence is checked whether it is well formed or not.
- The word arrangement is studied and a syntactic relationship is found between them. It is checked for word arrangements and grammar.
- For example, the sentence “Delhi goes to him” is rejected by the syntactic parser

Symbols in grammar

- S - Sentence
- NP- Noun Phrase
- PN- Proper Noun
- N-Noun
- VP-Verb Phrase
- Adv-Adverb
- V-Verb
- Adj-Adjective
- Prep-Preposition
- Art-Article

- Pro-Pronoun
- PP-Prepositional Phrase
- * Ungrammatical Sentence
- → Consists of / rewrites as
- ( ) Optional Constituent
- { } Only one of these constituents must be selected

Syntactic Analysis - grammar



Semantic Analysis

- The third phase is Semantic Analysis. In this phase, the sentence is checked for the literal meaning of each word and their arrangement together.
- For example, The sentence “I ate hot ice cream” will get rejected by the semantic analyzer because it doesn’t make sense.
- E.g.. “colorless green idea.” This would be rejected by the Symantec analysis as colorless Here; green doesn’t make any sense.

Parts of Semantic Analysis

Semantic Analysis of Natural Language can be classified into two broad parts:

1. Lexical Semantic Analysis: Lexical Semantic Analysis involves understanding the meaning of each word of the text individually. It basically refers to fetching the dictionary meaning that a word in the text is deputed to carry.

2. Compositional Semantics Analysis: Although knowing the meaning of each word of the text is essential, it is not sufficient to completely understand the meaning of the text.

For example, consider the following two sentences:

- **Sentence 1:** Students love chatgpt.
- **Sentence 2:** chatgpt loves Students.

Tasks involved in Semantic Analysis

1. Word Sense Disambiguation :- In Natural Language, the meaning of a word may vary as per its usage in sentences and the context of the text. Word Sense Disambiguation involves interpreting the meaning of a word based upon the context of its occurrence in a text.

For example, the word ‘Bark’ may mean ‘the sound made by a dog’ or ‘the outermost layer of a tree.’

- The ability of a machine to overcome the ambiguity involved in identifying the meaning of a word based on its usage and context is called Word Sense Disambiguation.

2. Relationship Extraction :- It involves firstly identifying various entities present in the sentence and then extracting the relationships between those entities.

Elements of Semantic Analysis

- Hyponymy: Hyponymys refers to a term that is an instance of a generic term. They can be understood by taking class-object as an analogy. For example: ‘Color‘ is a hypernymy while ‘grey‘, ‘blue‘, ‘red‘, etc, are its hyponyms.
- Homonymy: Homonymy refers to two or more lexical terms with the same spellings but completely distinct in meaning. For example: ‘Rose‘ might mean ‘the past form of rise‘ or ‘a flower‘, – same spelling but different meanings; hence, ‘rose‘ is a homonymy.
- Synonymy: When two or more lexical terms that might be spelt distinctly have the same or similar meaning, they are called Synonymy. For example: (Job, Occupation), (Large, Big), (Stop, Halt).
- Antonymy: Antonymy refers to a pair of lexical terms that have contrasting meanings – they are symmetric to a semantic axis. For example: (Day, Night), (Hot, Cold), (Large, Small).
- Polysemy: Polysemy refers to lexical terms that have the same spelling but multiple closely related meanings. It differs from homonymy because the meanings of the terms need not be closely related in the case of homonymy. For example: ‘man‘ may mean ‘the human species‘ or ‘a male human‘ or ‘an adult male human‘ – since all these different meanings bear a close association, the lexical term ‘man‘ is a polysemy.
- Meronomy: Meronomy refers to a relationship wherein one lexical term is a constituent of some larger entity. For example: ‘Wheel‘ is a meronym of ‘Automobile‘

Discourse Integration

- The fourth phase is discourse integration. In this phase, the impact of the sentences before a particular sentence and the effect of the current sentence on the upcoming sentences is determined.
- For example, the word “that” in the sentence “He wanted that” depends upon the prior discourse context

Cont…

- While processing a language there can arise one major ambiguity known as referential ambiguity. Referential ambiguity is the ambiguity that can arise when a reference to a word cannot be determined. For example,

Ram won the race.
Mohan ate half of a pizza.
He liked it.

In the above example, “He” can be Ram or Mohan.

- This creates an ambiguity. The word “He” shows dependency on both sentences. This is known as *disclosure integration*. It means when an individual sentence relies upon the sentence that comes before it. Like in the above example the third sentence relies upon the sentence before it. Hence the goal of this model is to remove referential ambiguity.

Pragmatic Analysis

- The last phase of natural language processing is Pragmatic analysis. Sometimes the discourse integration phase and pragmatic analysis phase are combined.
- The actual effect of the text is discovered by applying the set of rules that characterize cooperative dialogues.
- E.g., “close the window?” should be interpreted as a request instead of an order

Cont…

- The pragmatic analysis means handling the situation in a much more practical or realistic manner than using a theoretical approach. As we know that a sentence can have different meanings in various situations. For example, The average is 18.

  The average is 18. (average may be of sequence)

  The average is 18. (average may be of a vehicle)

  The average is 18. (average may be of a mathematical term)

- Same input but different perceptions.
- To interpret the meaning of the sentence we need to understand the situation. To tackle such problems we use pragmatic analysis. The pragmatic analysis tends to make the understanding of the language much more clear and easy to interpret.

Why NLP is difficult?

NLP is difficult because Ambiguity and Uncertainty exist in the language.

Ambiguity

There are the following three ambiguity -

- **Lexical Ambiguity**

Lexical Ambiguity exists in the presence of two or more possible meanings of the sentence within a single word.

Example:

Manya is looking for a **match**.

In the above example, the word match refers to that either Manya is looking for a partner or Manya is looking for a match. (Cricket or other match)

- **Syntactic Ambiguity**

Syntactic Ambiguity exists in the presence of two or more possible meanings within the sentence.

Example:

I saw the girl with the binocular.

In the above example, did I have the binoculars? Or did the girl have the binoculars?

- **Referential Ambiguity**

Referential Ambiguity exists when you are referring to something using the pronoun.

Example: Kiran went to Sunita. She said, "I am hungry."

In the above sentence, you do not know that who is hungry, either Kiran or Sunita.

Difference between Natural language and Computer Language

Natural Language	Computer Language
Natural language has a very large vocabulary.	Computer language has a very limited vocabulary.
Natural language is easily understood by humans.	Computer language is easily understood by the machines.
Natural language is ambiguous in nature.	Computer language is unambiguous.

NLP Implementation

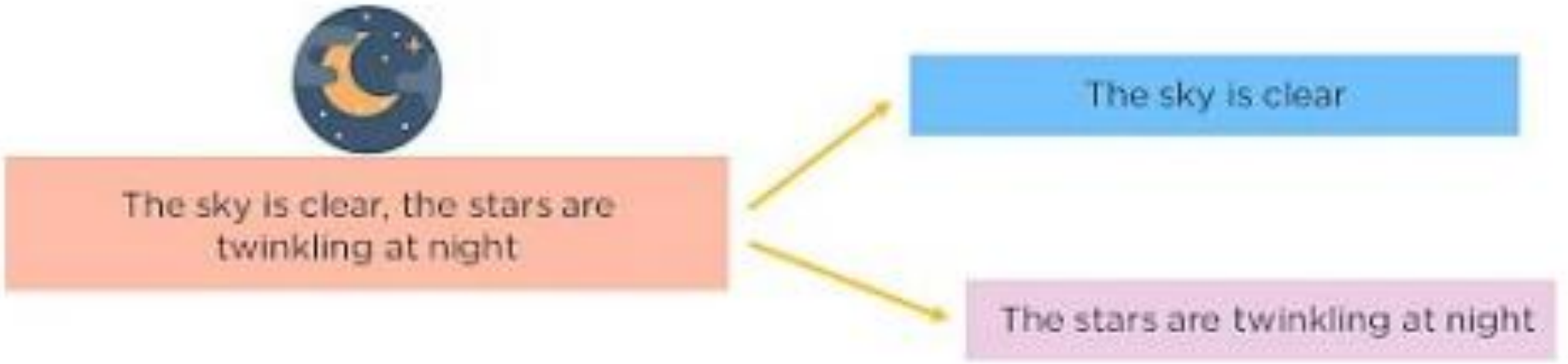
- Below, given are popular methods used for Natural Learning Process: – Machine learning: The learning nlp procedures used during machine learning. It automatically focuses on the most common cases. So when we write rules by hand, it is often not correct at all concerned about human errors. Statistical inference: NLP can make use of statistical inference algorithms. It helps you to produce models that are robust. e.g., containing words or structures which are known to everyone.

NLP Steps

- How to Perform NLP?
  - – Segmentation
  - – Tokenizing
  - – Removing Stop Words:
  - – Stemming
  - – Lemmatization
  - – Part of Speech Tagging
  - – Named Entity Tagging

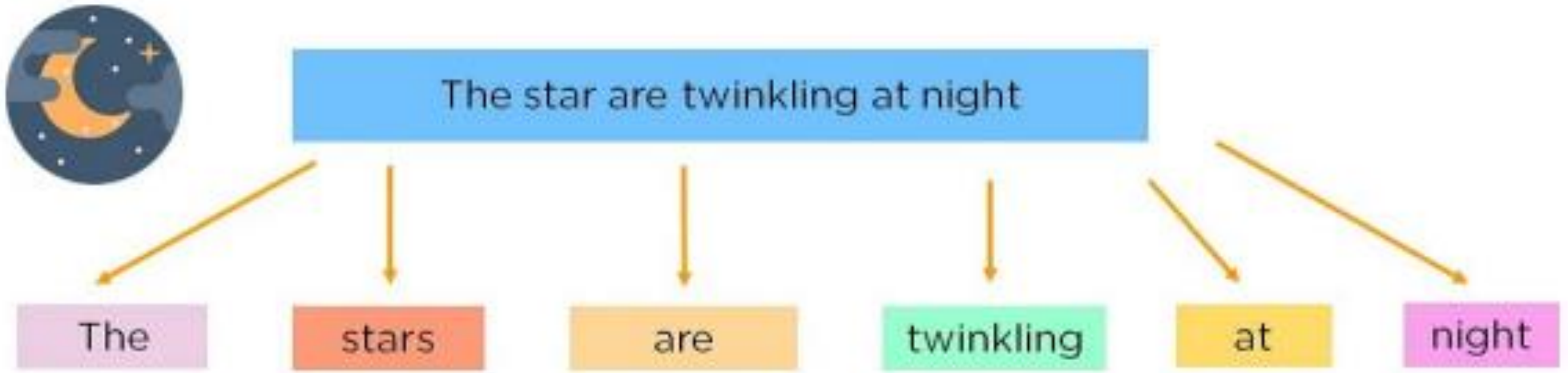
Segmentation

- You first need to break the entire document down into its constituent sentences. You can do this by segmenting the article along with its punctuation like full stops and commas.



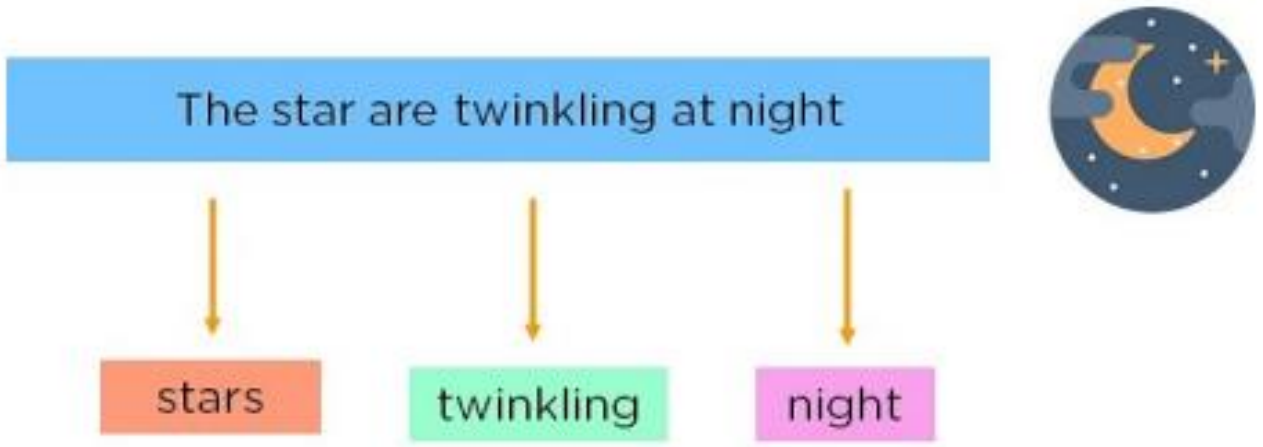
Tokenizing

- For the algorithm to understand these sentences, you need to get the words in a sentence and explain them individually to our algorithm.
- So, you break down your sentence into its constituent words and store them. This is called tokenizing, and each world is called a token



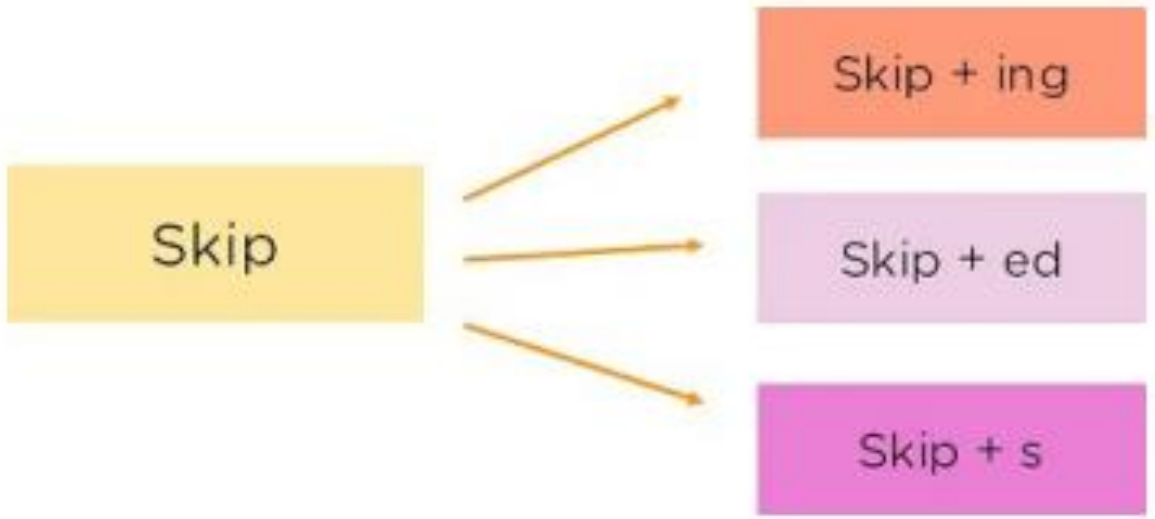
Removing Stop Words

- You can make the learning process faster by getting rid of non-essential words, which add little meaning to our statement and are just there to make our statement sound more cohesive. Words such as was, in, is, and, the, are called stop words and can be removed.



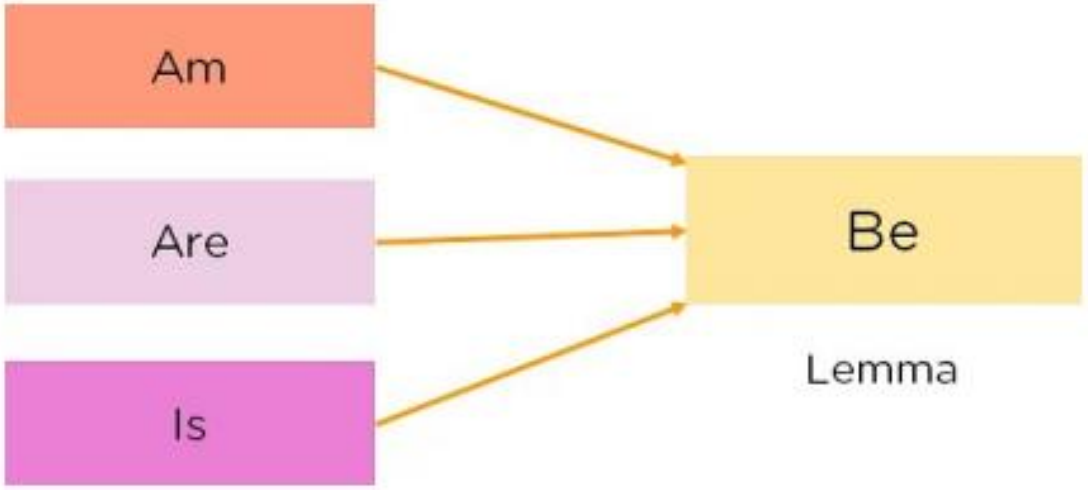
Stemming

- It is the process of obtaining the Word Stem of a word. Word Stem gives new words upon adding affixes to them



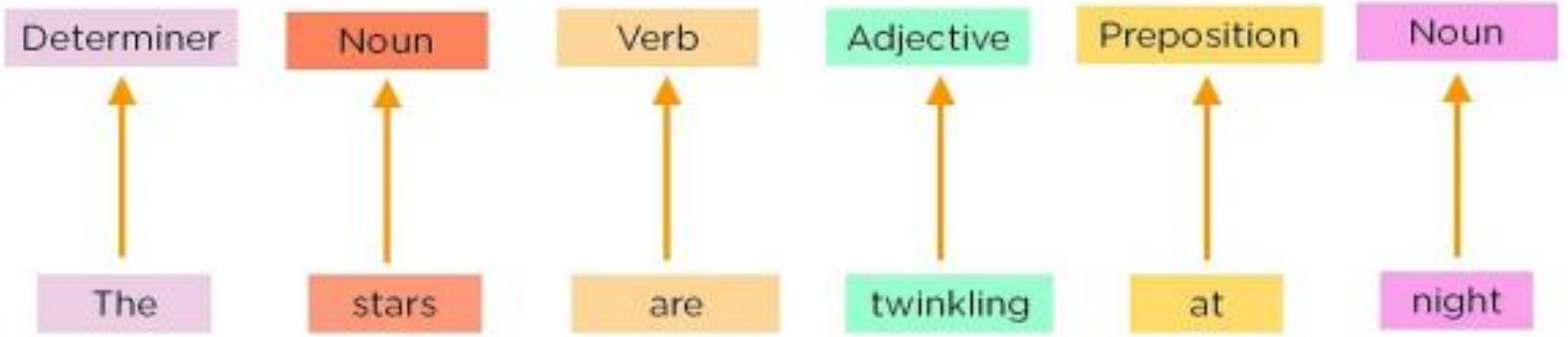
Lemmatization

- The process of obtaining the Root Stem of a word. Root Stem gives the new base form of a word that is present in the dictionary and from which the word is derived. You can also identify the base words for different words based on the tense, mood, gender,etc



Part of Speech Tagging

- Now, you must explain the concept of nouns, verbs, articles, and other parts of speech to the machine by adding these tags to our words. This is called ‘part of’.



Named Entity Tagging

- Next, introduce your machine to pop culture references and everyday names by flagging names of movies, important personalities or locations, etc that may occur in the document.
- You do this by classifying the words into subcategories. This helps you find any keywords in a sentence. The subcategories are person, location, monetary value, quantity, organization, movie.
- After performing the preprocessing steps, you then give your resultant data to a machine learning algorithm like Naive Bayes, etc., to create your NLP application.

Applications of NLP

- NLP is one of the ways that people have humanized machines and reduced the need for labor. It has led to the automation of speech-related tasks and human interaction. Some applications of NLP include :
  - Translation Tools: Tools such as Google Translate, Amazon Translate, etc. translate sentences from one language to another using NLP.
  - Chatbots: Chatbots can be found on most websites and are a way for companies to deal with common queries quickly
- Virtual Assistants: Virtual Assistants like Siri, Cortana, Google Home, Alexa, etc can not only talk to you but understand commands given to them

- Targeted Advertising: Have you ever talked about a product or service or just googled something and then started seeing ads for it? This is called targeted advertising, and it helps generate tons of revenue for sellers as they can reach niche audiences at the right time.
- Autocorrect: Autocorrect will automatically correct any spelling mistakes you make, apart from this grammar checkers also come into the picture which helps you write flawlessly.
- Information retrieval & Web Search: Google, Yahoo, Bing, and other search engines base their machine translation technology on NLP deep learning models. It allows algorithms to read text on a webpage, interpret its meaning and translate it to another language

- Grammar Correction:
  - NLP technique is widely used by word processor software like MS-word for spelling correction & grammar check.

Advantages of NLP

- Users can ask questions about any subject and get a direct response within seconds.
- NLP system provides answers to the questions in natural language
- NLP system offers exact answers to the questions, no unnecessary or unwanted information
- The accuracy of the answers increases with the amount of relevant information provided in the question.
- NLP process helps computers communicate with humans in their language and scales other language-related tasks
- Allows you to perform more language-based data compares to a human being without fatigue and in an unbiased and consistent way.
- Structuring a highly unstructured data source

Disadvantages of NLP

- Complex Query Language- the system may not be able to provide the correct answer it the question that is poorly worded or ambiguous.
- The system is built for a single and specific task only; it is unable to adapt to new domains and problems because of limited functions.
- NLP system doesn’t have a user interface which lacks features that allow users to further interact with the system