YENEPOYA
(DEEMED TO BE UNIVERSITY)
Recognized under Sec 3(A) of the UGC Act 1956
Accredited by NAAC with 'A' Grade

UNIT 1

Overview And Language Modeling

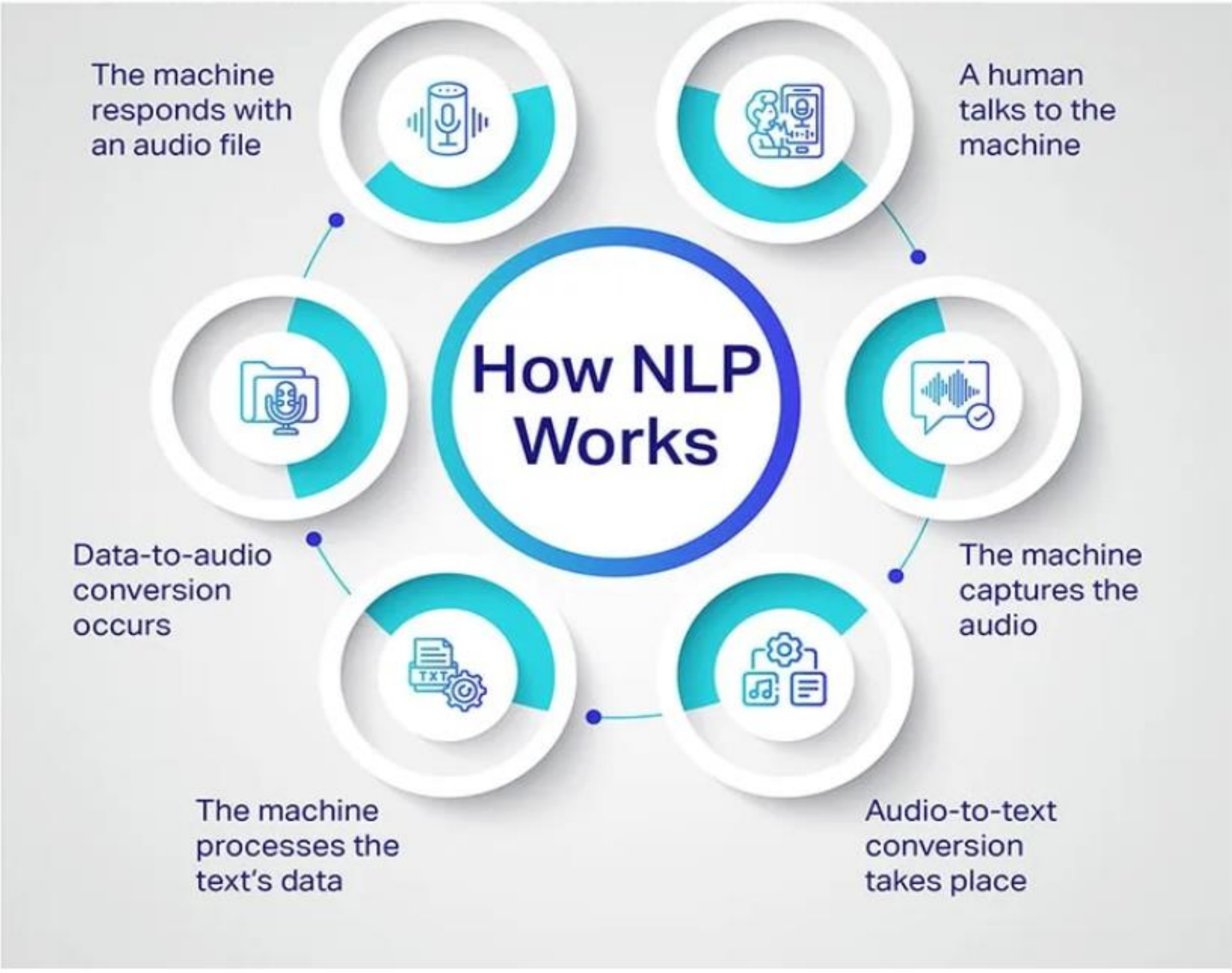
SYLLABUS

- **Overview: Origins and challenges of NLP**
- **Language and Grammar-Processing Indian Languages-**
- **NLP Applications-Information Retrieval.**
- **Language Modelling: Various Grammar based Language Models-Statistical Language Model.**

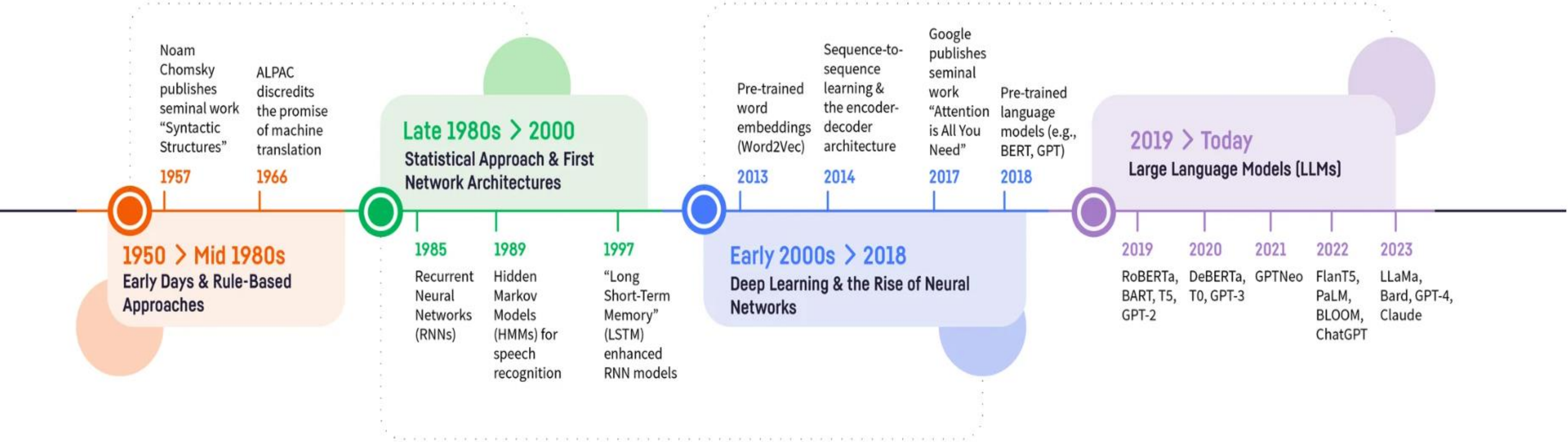
What is NLP?

- **Natural Language Processing (NLP)** is a branch of Artificial Intelligence that focuses on enabling computers to understand, interpret, and generate human language, both written and spoken

- It allows machines to process and analyze text and speech, enabling various applications like virtual assistants, translation, and sentiment analysis.


How NLP Works
A human talks to the machine
The machine captures the audio
Audio-to-text conversion takes place
The machine processes the text's data
Data-to-audio conversion occurs
The machine responds with an audio file

Origin / history of NLP



1950 > Mid 1980s
Early Days & Rule-Based Approaches

1957
Noam Chomsky publishes seminal work "Syntactic Structures"

1966
ALPAC discredits the promise of machine translation

Late 1980s > 2000
Statistical Approach & First Network Architectures

1985
Recurrent Neural Networks (RNNs)

1989
Hidden Markov Models (HMMs) for speech recognition

1997
"Long Short-Term Memory" (LSTM) enhanced RNN models

Early 2000s > 2018
Deep Learning & the Rise of Neural Networks

2013
Pre-trained word embeddings (Word2Vec)

2014
Sequence-to-sequence learning & the encoder-decoder architecture

2017
Google publishes seminal work "Attention is All You Need"

2018
Pre-trained language models (e.g., BERT, GPT)

2019 > Today
Large Language Models (LLMs)

2019
RoBERTa, BART, T5, GPT-2

2020
DeBERTa, T0, GPT-3

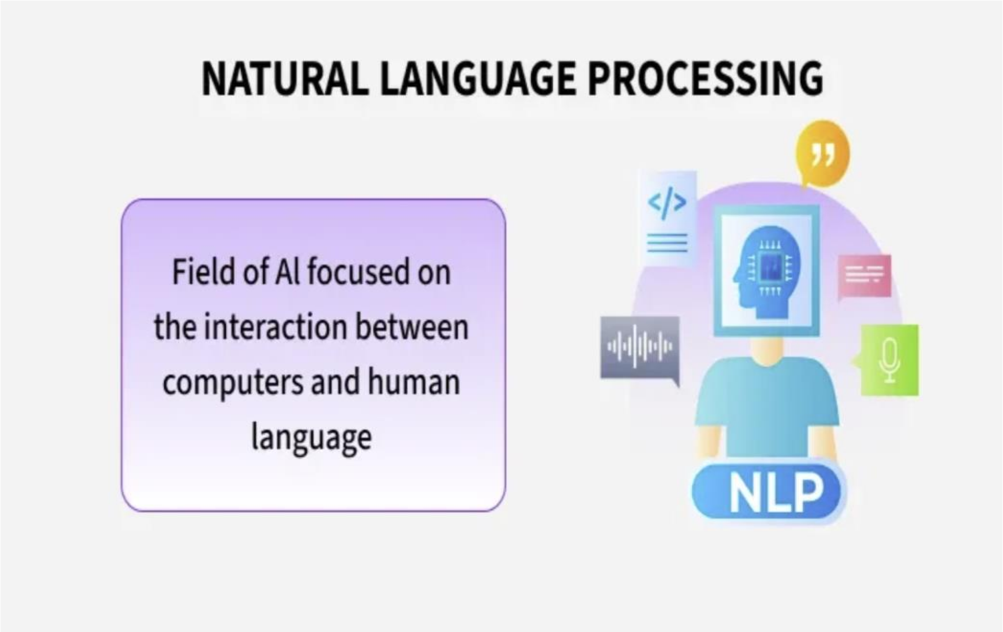
2021
GPTNeo

2022
FlanT5, PaLM, BLOOM, ChatGPT

2023
LLaMa, Bard, GPT-4, Claude

Natural Language Processing (NLP) - Overview

- Natural Language Processing (NLP) is a field that combines computer science, artificial intelligence and language studies. It helps computers understand, process and create human language in a way that makes sense and is useful.
- With the growing amount of text data from social media, websites and other sources, NLP is becoming a key tool to gain insights and automate tasks like analyzing text or translating languages.



NATURAL LANGUAGE PROCESSING

Field of AI focused on the interaction between computers and human language

Challenges in natural language processing

- **Misspellings:** It has typos and variations, like **"proces"**, **"process"**, or **"résumé"** vs **"resume"**.

- **Language Differences:** Different languages say the same thing in different ways. For example, **"I'm going home"** (English) and **"Je rentre à la maison"** (French) mean the same, but NLP needs translation to understand both.

- **Uncertainty and False Positives:** False positives happen when NLP thinks it understands but responds incorrectly. A good system should recognize this and ask questions to clarify.

- **Training Data:** NLP struggles with bad training data. More data helps, but if it's wrong or biased, the system learns poorly or incorrectly.

- **Innate Biases:** NLP uses human logic and data, so it can reflect biases from programmers or datasets, sometimes misinterpreting context and giving wrong results.

- **Words with Multiple Meanings:**NLP assumes language is clear, but it's often not. Words like **"bark"** can mean a dog sound or tree covering, causing confusion.

(NLP)language and grammar overview

- Natural Language Processing (NLP) is a field at the intersection of linguistics, computer science, and artificial intelligence

Here's how language and grammar play a central role in NLP:

What Is Grammar in NLP?

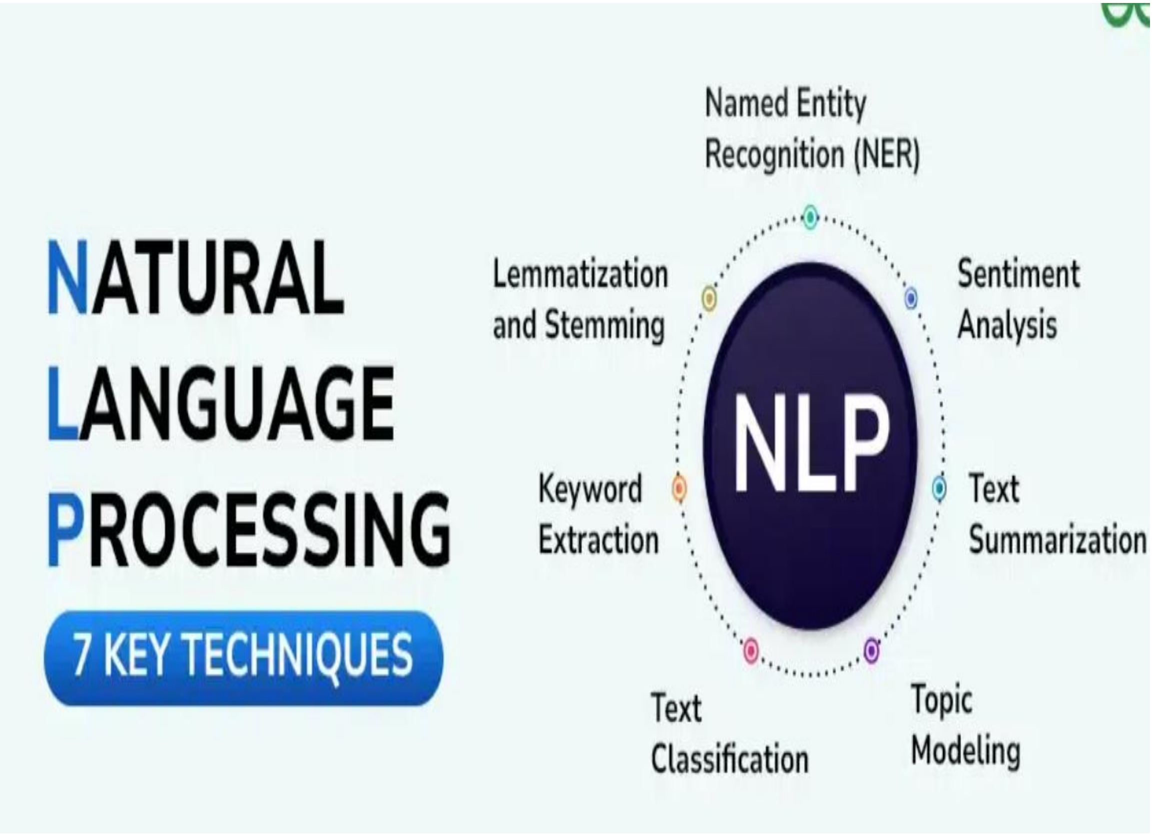
- Grammar in NLP refers to the **rules and structures** that govern how words combine to form meaningful sentences. It helps machines analyze and generate syntactically correct language.
- That enables machines to understand, interpret, and generate human language.

NLP Language And Grammar Overview

- **Syntax**: The arrangement of words and phrases to create well-formed sentences.
- **Morphology**: The study of word structure (e.g., prefixes, suffixes, root words).
- **Parts of Speech (POS)**: Identifying nouns, verbs, adjectives, etc.
- **Parse Trees**: Visual representations of sentence structure.
- **Grammar Types**:
- **Context-Free Grammar (CFG)**: Rules that define sentence structure hierarchically.
- **Constituency Grammar**: Breaks sentences into nested phrases (e.g., noun phrases).
- **Dependency Grammar**: Focuses on relationships between words (e.g., subject-verb).

NLP Techniques:

- NLP encompasses a wide array of techniques that aimed at enabling computers to process and understand human language.
- These tasks can be categorized into several broad areas, each addressing different aspects of language processing.



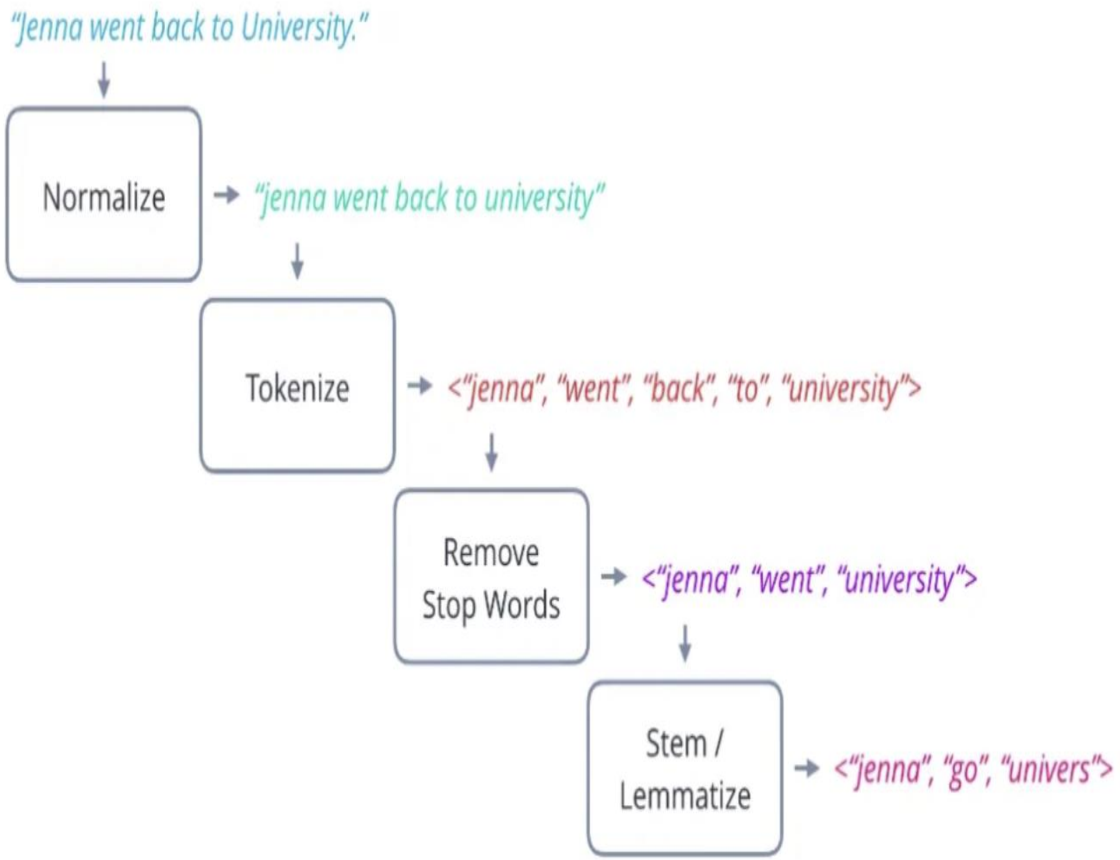
1. Text Processing And Preprocessing

•Tokenization: Dividing text into smaller units, such as words or sentences.

•Stemming and Lemmatization: Reducing words to their base or root forms.

•Stopword Removal: Removing common words (like "and", "the", "is") that may not carry significant meaning.

•Text Normalization: Standardizing text, including case normalization, removing punctuation and correcting spelling errors.

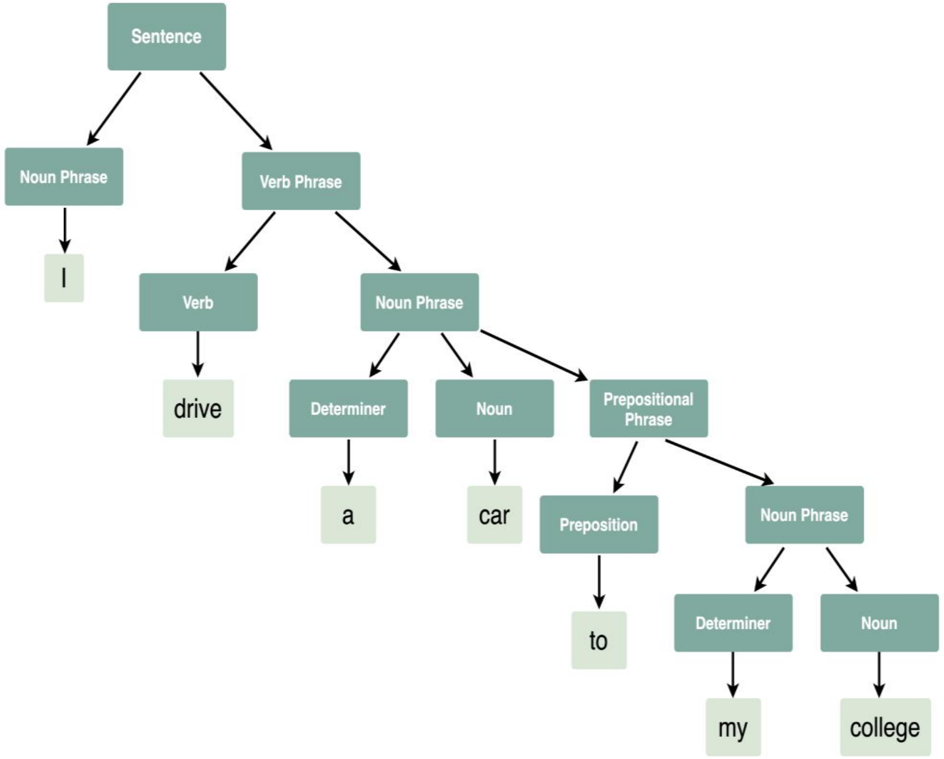

"Jenna went back to University."
Normalize
"jenna went back to university"
Tokenize
<"jenna", "went", "back", "to", "university">
Remove Stop Words
<"jenna", "went", "university">
Stem / Lemmatize
<"jenna", "go", "univers">

2. Syntax And Parsing

- •Part-of-Speech (POS) Tagging: Assigning parts of speech to each word in a sentence (e.g., noun, verb, adjective).

- •Dependency Parsing: Analyzing the grammatical structure of a sentence to identify relationships between words.

- •Constituency Parsing: Breaking down a sentence into its constituent parts or phrases (e.g., noun phrases, verb phrases).



3. Semantic Analysis

- **Named Entity Recognition (NER):** Identifying and classifying entities in text, such as names of people organizations, locations, dates, etc.

- **Word Sense Disambiguation (WSD)**: Determining which meaning of a word is used in a given context.

- **Coreference Resolution**: Identifying when different words refer to the same entity in a text (e.g., "he" refers to "John").

4.Information Extraction

- **Entity Extraction:** Identifying specific entities and their relationships within the text.

- **Relation Extraction**: Identifying and categorizing the relationships between entities in a text.

5. Text Classification In NLP

- Sentiment Analysis: Determining the sentiment or emotional tone expressed in a text (e.g., positive, negative, neutral).

- Topic Modelling: Identifying topics or themes within a large collection of documents.

- Spam Detection: Classifying text as spam or not spam.

6. Language Generation

- **Machine Translation:** Translating text from one language to another.
- **Text Summarization:** Producing a concise summary of a larger text.
- **Text Generation:** Automatically generating coherent and contextually relevant text.

7. SPEECH PROCESSING

- **Speech Recognition**: Converting spoken language into text.

- **Text-to-Speech (TTS) Synthesis**: Converting written text into spoken language.

8. Question Answering

- **Retrieval-Based QA:** Finding and returning the most relevant text passage in response to a query.

- **Generative QA:** Generating an answer based on the information available in a text corpus.

9. Dialogue Systems

- **Chatbots and Virtual Assistants:** Enabling systems to engage in conversations with users, providing responses and performing tasks based on user input

10. Sentiment And Emotion Analysis In NLP

- **Emotion Detection**: Identifying and categorizing emotions expressed in text.
- **Opinion Mining**: Analyzing opinions or reviews to understand public sentiment toward products, services or topics.

Components Of NLP

Components Of NLP

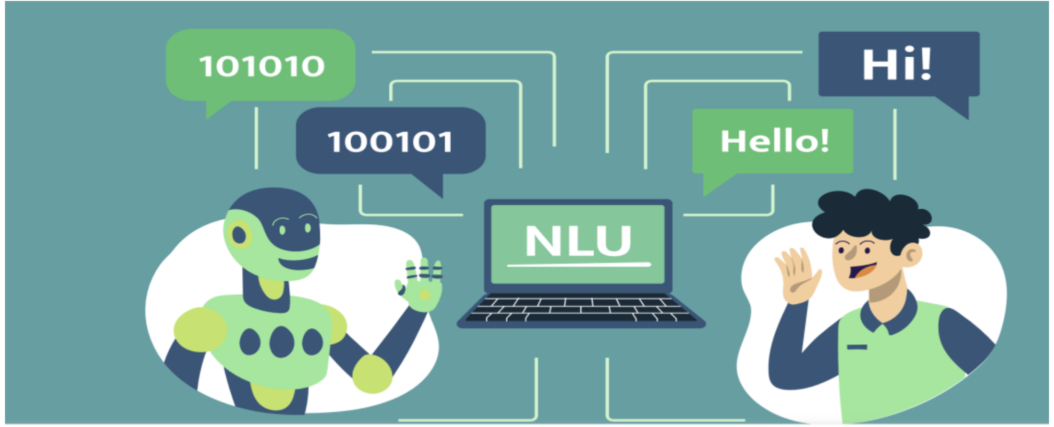
- **There are two components of NLP:**

1) Natural Language Understanding (NLU) and

2) Natural Language Generation (NLG).

Natural Language Understanding (NLU)

- It involves transforming human language into a machine-readable format.
- It helps the machine to understand and analyse human language by extracting the text from large data such as keywords, emotions, relations, and semantics.



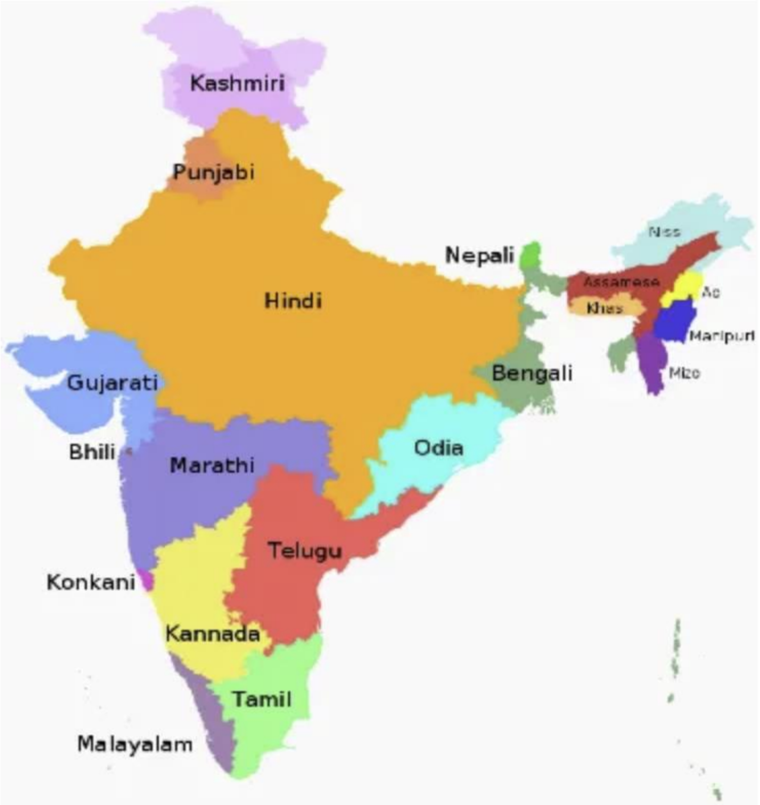
Natural Language Generation (NLG)

- It acts as a translator that converts the computerized data into natural language representation.
- It mainly involves Text planning, Sentence planning, and Text realization.
- The NLU is harder than NLG.

NLG

Processing Indian Languages

Processing Indian languages in Natural Language Processing (NLP) is a fascinating and complex challenge due to their linguistic diversity, rich morphology, and varied scripts.



Why Indian Languages Are Unique In NLP

Multiple Language Families: Includes Indo-Aryan, Dravidian, Tibeto-Burman, and Austroasiatic.

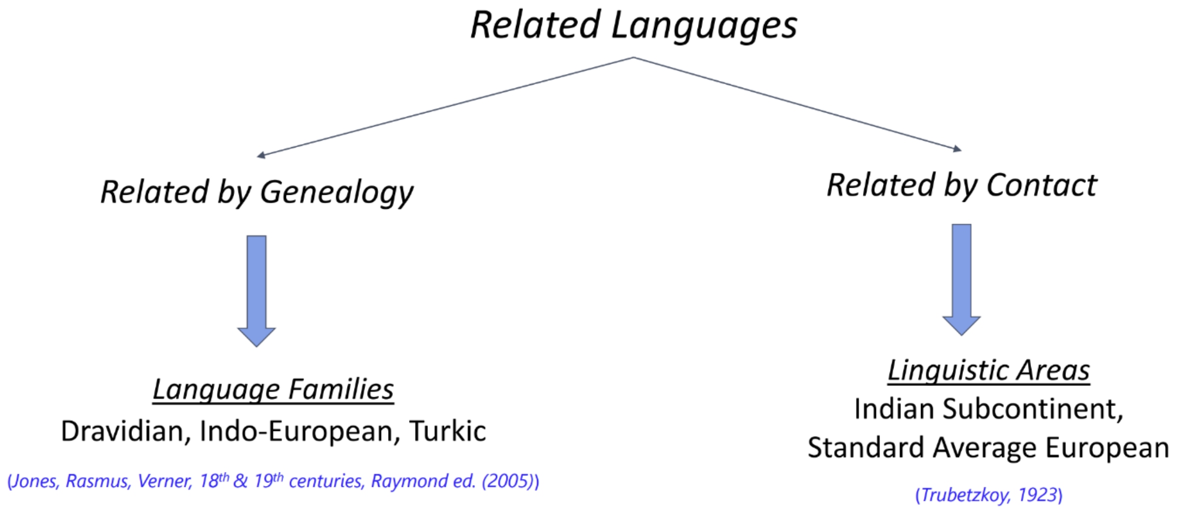
Morphological Richness: Words often carry extensive grammatical information through inflections.

Free Word Order: Many Indian languages allow flexible sentence structures, unlike English.

Script Diversity: Languages use distinct scripts (e.g., Devanagari, Tamil, Telugu, Bengali), complicating tokenization and transliteration.

Language Relatedness:

Why are Indian languages related?



Related Languages

Related by Genealogy

Related by Contact

__Language Families__
Dravidian, Indo-European, Turkic

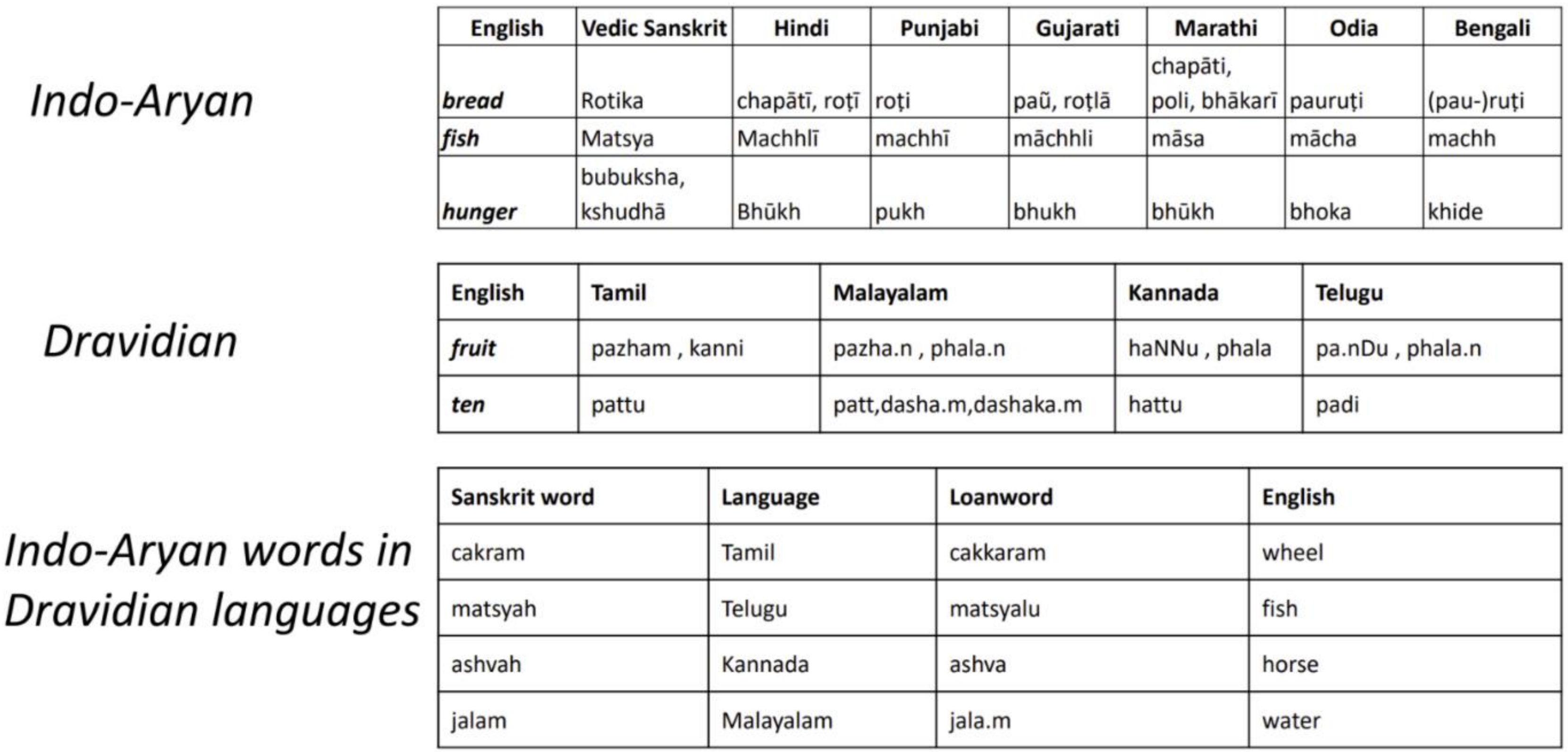
(Jones, Rasmus, Verner, 18th & 19th centuries, Raymond ed. (2005))

__Linguistic Areas__
Indian Subcontinent,
Standard Average European

(Trubetzkoy, 1923)

Related languages may not belong to the same language family!

Cognates & Borrowed Words In Indian Languages:



Indo-Aryan

English	Vedic Sanskrit	Hindi	Punjabi	Gujarati	Marathi	Odia	Bengali
bread	Rotika	chapātī, roṭī	roṭi	paũ, roṭlā	chapāti, poli, bhākarī	pauruṭi	(pau-)ruṭi
fish	Matsya	Machhlī	machhī	māchhli	māsa	mācha	machh
hunger	bubuksha, kshudhā	Bhūkh	pukh	bhukh	bhūkh	bhoka	khide

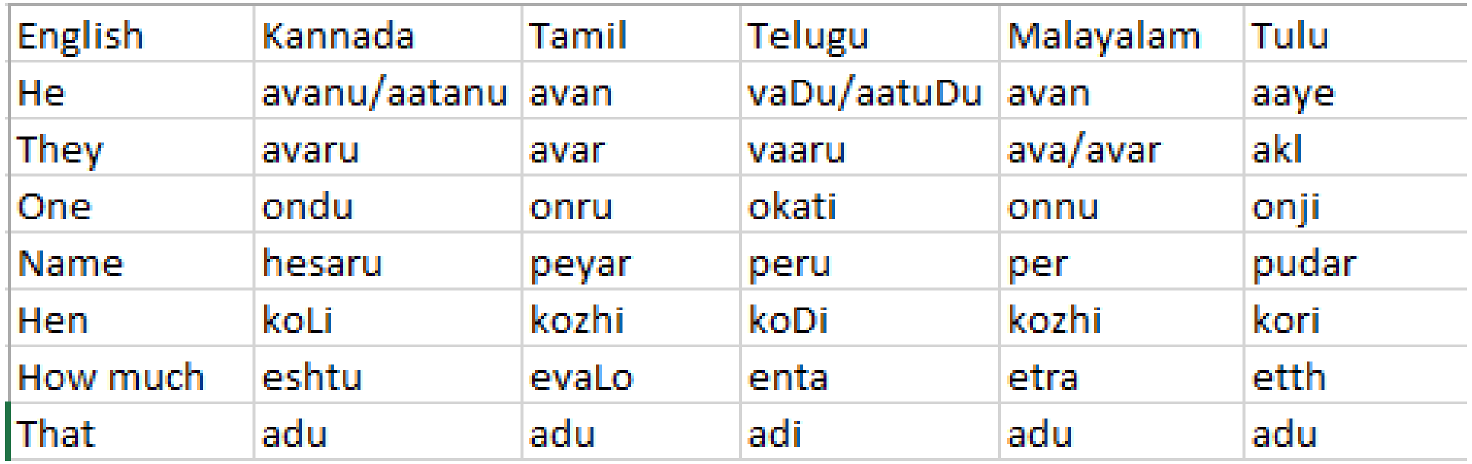
Dravidian

English	Tamil	Malayalam	Kannada	Telugu
fruit	pazham , kanni	pazha.n , phala.n	haNNu , phala	pa.nDu , phala.n
ten	pattu	patt,dasha.m,dashaka.m	hattu	padi

Indo-Aryan words in Dravidian languages

Sanskrit word	Language	Loanword	English
cakram	Tamil	cakkaram	wheel
matsyah	Telugu	matsyalu	fish
ashvah	Kannada	ashva	horse
jalam	Malayalam	jala.m	water

Key Similarities Between Related Languages:



English	Kannada	Tamil	Telugu	Malayalam	Tulu
He	avanu/aatanu	avan	vaDu/aatuDu	avan	aaye
They	avaru	avar	vaaru	ava/avar	akl
One	ondu	onru	okati	onnu	onji
Name	hesaru	peyar	peru	per	pudar
Hen	koLi	kozhi	koDi	kozhi	kori
How much	eshtu	evaLo	enta	etra	etth
That	adu	adu	adi	adu	adu

Example: Using Hindi and Bengali:

English: He is learning the language.

Hindi: वह भाषा सीख रहा है। (*vah bhāṣā sīkh rahā hai*)

Bengali: সে ভাষা শিখছে। (*se bhāṣā śikhche)*

Applications Of NLP (Natural Language Processing)

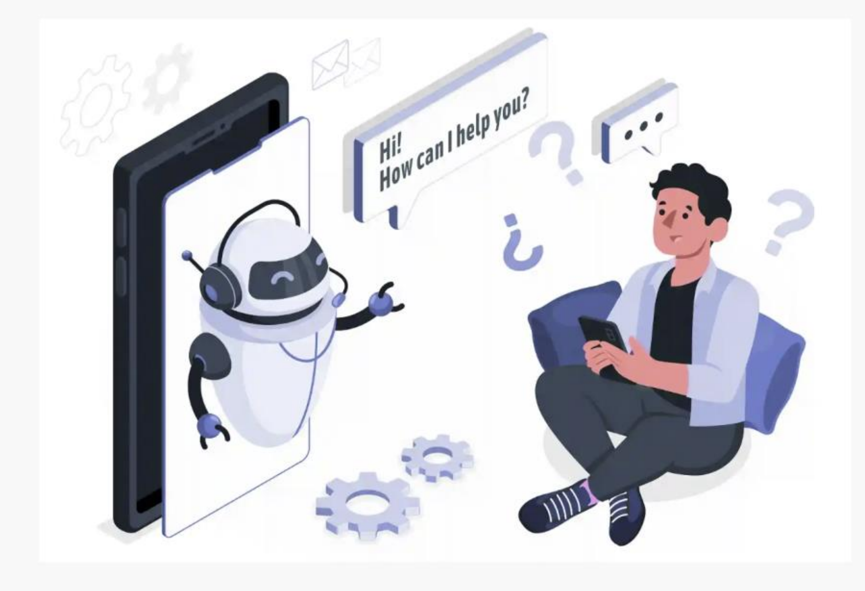
1. Chatbots:

•Chatbots are a form of artificial intelligence that are programmed to interact with humans in such a way that they sound like humans themselves.

•They can either just respond to specific keywords or they can even hold full conversations that make it tough to distinguish them from humans.

•Chatbots are created using Natural Language Processing and Machine Learning, which means that they understand the complexities of the English language and find the actual meaning of the sentence.

•Chatbots work in two simple steps. First, they identify the meaning of the question asked and collect all the data from the user that may be required to answer the question. Then they answer the question appropriately.



Applications Of NLP (Natural Language Processing)

2. Voice Assistants

- These days voice assistants are all the rage! Whether its Siri, Alexa, or Google Assistant, almost everyone uses one of these to make calls, place reminders, schedule meetings, set alarms, surf the internet, etc.
- These voice assistants have made life much easier.
- They use a complex combination of speech recognition, natural language understanding, and natural language processing to understand what humans are saying and then act on it.
- The long term goal of voice assistants is to become a bridge between humans and the internet and provide all manner of services based on just voice interaction.

Applications Of NLP (Natural Language Processing)

3.Language Translator

- Want to translate a text from English to Hindi but don’t know Hindi? Well, Google Translate is the tool for you! While it’s not exactly 100% accurate, it is still a great tool to convert text from one language to another.
- Google Translate and other translation tools as well as use Sequence to sequence modeling that is a technique in Natural Language Processing.

Applications Of NLP (Natural Language Processing)

4.Autocomplete in Search Engines:

- Have you noticed that search engines tend to guess what you are typing and automatically complete your sentences? For example, On typing “game” in Google, you may get further suggestions for “game of thrones”, “game of life” or if you are interested in maths then “game theory”.
- They use Natural Language Processing to make sense of these words and how they are interconnected to form different sentences.

Applications Of NLP (Natural Language Processing)

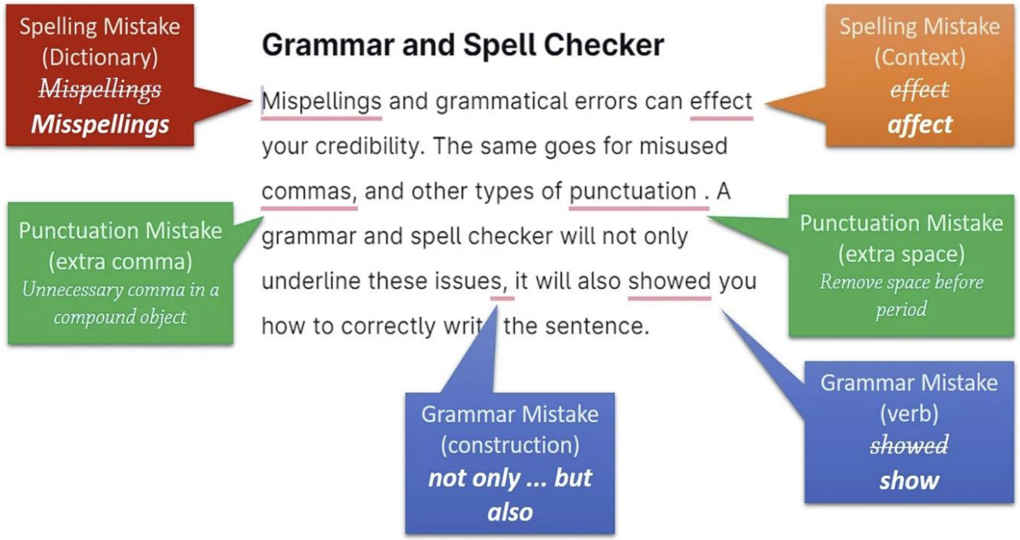
5. Sentiment Analysis

- Almost all the world is on social media these days! And companies can use sentiment analysis to understand how a particular type of user feels about a particular topic, product, etc.
- They can use natural language processing, computational linguistics, text analysis, etc. to understand the general sentiment of the users for their products and services and find out if the sentiment is good, bad, or neutral.
- Companies can use sentiment analysis in a lot of ways such as to find out the emotions of their target audience, to understand product reviews, to gauge their brand sentiment, etc.

Applications Of NLP (Natural Language Processing)

6.Grammar Checkers

- Grammar and spelling is a very important factor while writing professional reports for your superiors even assignments for your lecturers.
- After all, having major errors may get you failed! That's why grammar and spell checkers are a very important tool for any professional writer.
- They can not only correct grammar and check spellings but also suggest better synonyms and improve the overall readability of your content.
- And guess what, they utilize natural language processing to provide the best possible piece of writing!
- The NLP algorithm is trained on millions of sentences to understand the correct format.



Grammar and Spell Checker

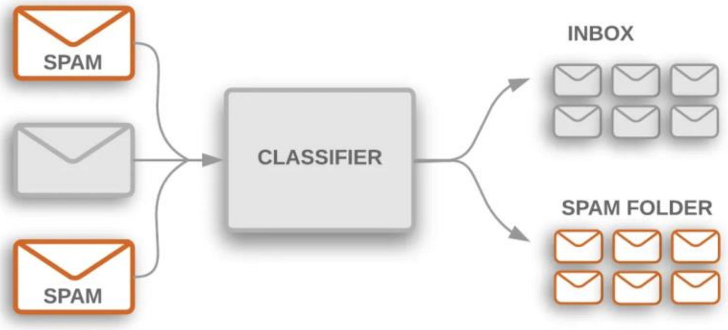
Mispellings and grammatical errors can effect your credibility. The same goes for misused commas, and other types of punctuation . A grammar and spell checker will not only underline these issues, it will also showed you how to correctly writ the sentence.

Spelling Mistake (Dictionary)
~~Mispellings~~
Misspellings
Spelling Mistake (Context)
~~effect~~
affect
Punctuation Mistake (extra comma)
Unnecessary comma in a compound object
Punctuation Mistake (extra space)
Remove space before period
Grammar Mistake (construction)
not only ... but also
Grammar Mistake (verb)
~~showed~~
show

APPLICATIONS OF NLP (NATURAL LANGUAGE PROCESSING)

7.Email Classification and Filtering:

- Emails are still the most important method for professional communication. However, all of us still get thousands of promotional Emails that we don’t want to read.
- Emails are automatically divided into 3 sections namely, Primary, Social, and Promotions which means we never have to open the Promotional section! But how does this work?
- Email services use natural language processing to identify the contents of each Email with text classification so that it can be put in the correct section.



Email Spam Classifier

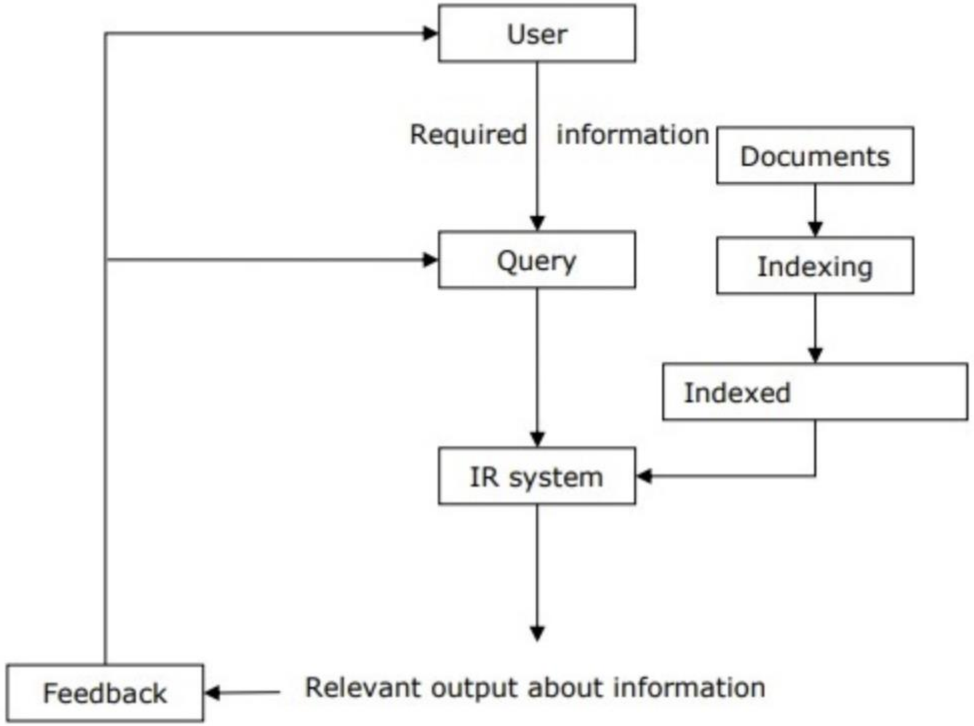
INFORMATION RETRIEVAL IN NLP

Information Retrieval In NLP:

- Information retrieval (IR) may be defined as a software program that deals with the organization, storage, retrieval and evaluation of information from document repositories particularly textual information.
- The system assists users in finding the information they require but it does not explicitly return the answers of the questions.
- It informs the existence and location of documents that might consist of the required information. The documents that satisfy users requirement are called relevant documents.
- A perfect IR system will retrieve only relevant documents.

Information Retrieval In NLP:

With the help of the following diagram, we can understand the process of information retrieval (IR) –



It is clear from the above diagram that a user who needs information will have to formulate a request in the form of query in natural language.

Then the IR system will respond by retrieving the relevant output, in the form of documents, about the required information.

Information Retrieval In NLP:

Classical Problem in Information Retrieval (IR) System

- The main goal of IR research is to develop a model for retrieving information from the repositories of documents. Here, we are going to discuss a classical problem, named ad-hoc retrieval problem, related to the IR system.
- In ad-hoc retrieval, the user must enter a query in natural language that describes the required information.
- Then the IR system will return the required documents related to the desired information. For example, suppose we are searching something on the Internet and it gives some exact pages that are relevant as per our requirement but there can be some non-relevant pages too.
- This is due to the ad-hoc retrieval problem.

Information Retrieval In NLP:

Information Retrieval (IR) Model

Mathematically, models are used in many scientific areas having objective to understand some phenomenon in the real world. A model of information retrieval predicts and explains what a user will find in relevance to the given query.

IR model is basically a pattern that defines the above-mentioned aspects of retrieval procedure and consists of the following –

1. **A model for documents.**
2. **A model for queries.**
3. **A matching function that compares queries to documents.**

Types Of Information Retrieval (IR) Model

1.Classical IR Model

- It is the simplest and easy to implement IR model.
- This model is based on mathematical knowledge that was easily recognized and understood as well. Boolean, Vector and Probabilistic are the three classical IR models.

2.Non-Classical IR Model

- It is completely opposite to classical IR model.
- Such kind of IR models are based on principles other than similarity, probability, Boolean operations. Information logic model, situation theory model and interaction models are the examples of non-classical IR model.

3.Alternative IR Model

- It is the enhancement of classical IR model making use of some specific techniques from some other fields. Cluster model, fuzzy model and latent semantic indexing (LSI) models are the example of alternative IR model.

Language Models In NLP:

- Language models are a fundamental component of natural language processing (NLP) and computational linguistics.
- A language model is the core component of modern Natural Language Processing (NLP).
- It’s a statistical model that is designed to analyze the pattern of human language and predict the likelihood of a sequence of words or tokens.
- They are designed to understand, generate, and predict human language.
- These models analyse the structure and use of language to perform tasks such as machine translation, text generation, and sentiment analysis.
- NLP-based applications use language models for a variety of tasks, such as audio to text conversion, speech recognition, sentiment analysis, summarization, spell correction, etc.

Various Grammar Based Language Models:

Overview

- Grammar in NLP is a set of rules for constructing sentences in a language used to understand and analyze the structure of sentences in text data.

- This includes identifying parts of speech such as nouns, verbs, and adjectives, determining the subject and predicate of a sentence, and identifying the relationships between words and phrases.

Types Of Grammar In NLP:

Three types of Grammar:

1. Context Free Grammar
2. Constituency Grammar and
3. Dependency Grammar

Context-Free Grammar (CFG)

- A formal system where **rewrite rules** expand non-terminal symbols into sequences of terminals/non-terminals.
- Follows the structure: A→α$A \rightarrow \alpha$, where AA is a non-terminal and αα is any combination of symbols.
- Used in **parsing** to generate hierarchical structures (e.g., parse trees).
- **Key Idea**: Sentences can be broken down recursively into nested components.

Constituency Grammar (Phrase-Structure Grammar)

- A type of **CFG** that organizes sentences into **constituents** (phrases like NP, VP).
- Represents **hierarchical grouping** (e.g., a noun phrase contains a determiner and noun).
- **Key Idea**: Language has a **part-whole structure**, where smaller units combine into larger meaningful chunks.

Dependency Grammar

- Focuses on **binary relationships** between words (head-dependent links) rather than phrase structure.
- Represents **direct connections** (e.g., a verb governs its subject and object).
- **Key Idea**: Sentences are structured based on **word-to-word dependencies**, not nested phrases.

Types Of Language Models:

There are two types of language models:

1. **Statistical Language Models**
2. **Neural Language Models**

1. Statistical Language Models

These models use probability and statistics to predict the next word in a sequence.

How it works: They calculate the probability of a sentence by looking at the frequency of word sequences in a large dataset.

Example Techniques:

- **Unigram Model:** Considers single words independently.
- **Bigram/Trigram Models:** Considers 2-word or 3-word combinations.
- **n-gram Models:** Generalized version where "n" words are used.

Limitations:

- Struggle with long-term dependencies (can't remember context far back).
- Require large storage for probabilities.

NEURAL LANGUAGE MODELS (NLMS)

These models use **neural networks** (deep learning) to learn word representations and predict word sequences.

How it works: Instead of just counting word frequencies, they learn **word embeddings** (vector representations) and capture deeper relationships between words.

Example Techniques:

Feedforward Neural Network LMs

Recurrent Neural Networks (RNNs) and LSTMs

Transformers (BERT, GPT, etc.)

Advantages:

Capture long-term dependencies and context.

Better at handling unseen word combinations.

More accurate and powerful than statistical models.

Case Study:

Some ABC Pvt. Ltd. wants to build a multilingual NLP-based academic search platform for English and multiple Indian languages (Hindi, Tamil, Kannada, Bengali). As an NLP consultant, design a conceptual solution addressing the role of language and grammar, methods for processing Indian languages, suitable grammar-based and statistical language models, and an information retrieval pipeline. Include possible applications, challenges, and future enhancements.

Thank You