edureka!
a Veranda Enterprise | ILLINOIS TECH

POST GRADUATE PROGRAM IN GENERATIVE AI AND ML

Natural Language Processing

Module Outline

edureka! a Veranda Enterprise | ILLINOIS TECH

- Introduction to NLP
- Text Processing and Feature Engineering
- Named Entity Recognition (NER) & Parsing
- Tokenization and Text Encoding
- Sentiment Analysis Essentials
- Advanced Sentiment Analysis
- Neural Language Models
- Machine Translation
- Speech and Multimodal NLP
- Building Chatbots

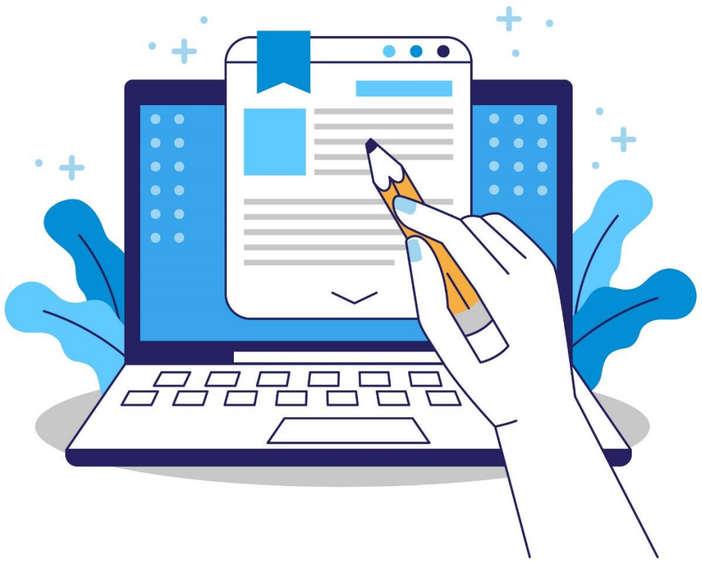
edureka!
a Veranda Enterprise

ILLINOIS TECH

Introduction to NLP

Topics

- Introduction and Scope of Natural Language Processing
- Core Tasks in NLP like Parsing, NER, and Sentiment Analysis
- NLP Approaches: Rule-Based, Statistical, and Hybrid Methods
- Morphology and Structure of Words in NLP
- Syntax and Sentence Structuring Techniques
- Semantics and Pragmatics in Understanding Language
- Text Preprocessing and Normalization Techniques in NLP



Learning Objectives

By the end of this lesson, you will be able to:

- Define the scope of NLP and its core components across language processing tasks.
- Differentiate between rule-based, statistical, and hybrid NLP approaches.
- Apply text preprocessing, morphological analysis, and sentence structuring techniques.
- Analyze semantic, pragmatic, and syntactic roles in understanding natural language.

What is NLP?

What is Natural Language?

Humans use Natural languages to communicate with each other, such as English, Spanish, Hindi, Mandarin, etc.

- Evolves Naturally
- Ambiguity
- Grammar and Structure
- Contextual

Why is Natural Language Used?

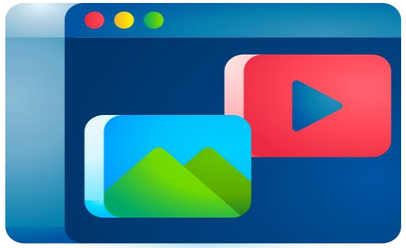
- Speech-to-Text Conversion
- Intent Recognition
- Context Understanding
- Task Execution

Natural Language Generation

NLG involves creating coherent text or speech from structured data and converting machine-readable information into natural language.

Text Summarization

Question Answering

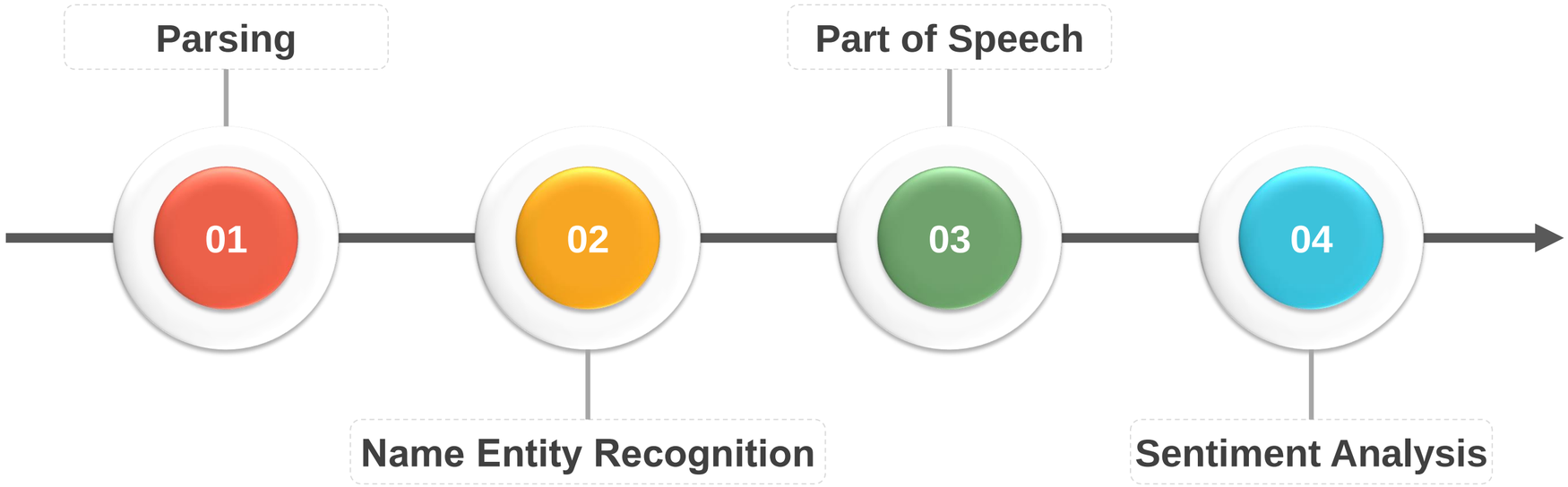


Content Creation

Dialogue Generation

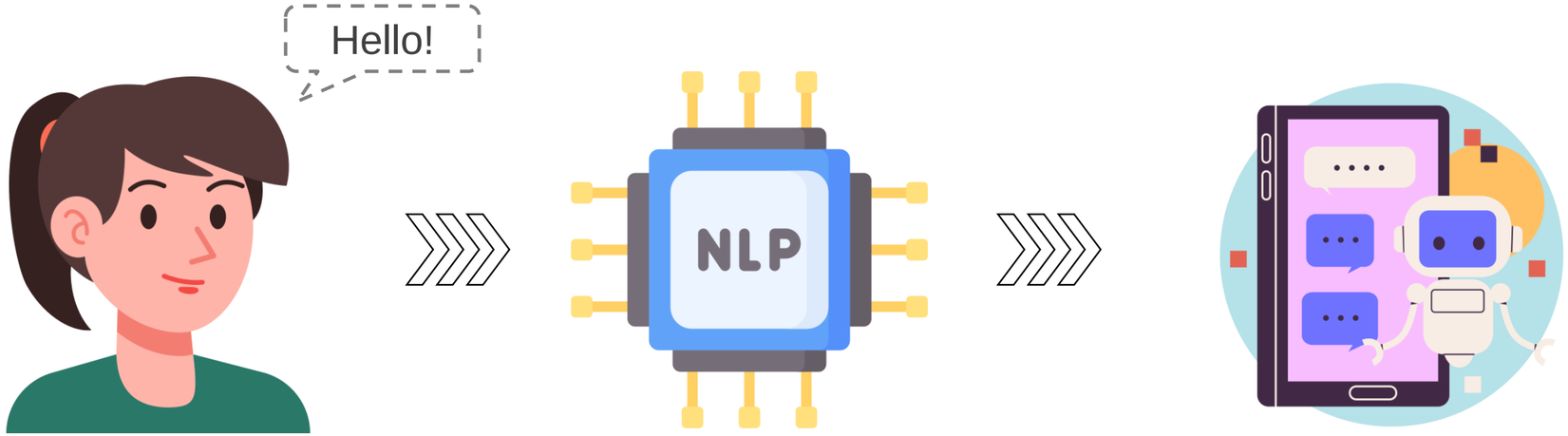
Natural Language Understanding

Natural Language Understanding (NLU) allows machines to comprehend human language by interpreting its meaning and context, enabling them to extract relevant information.



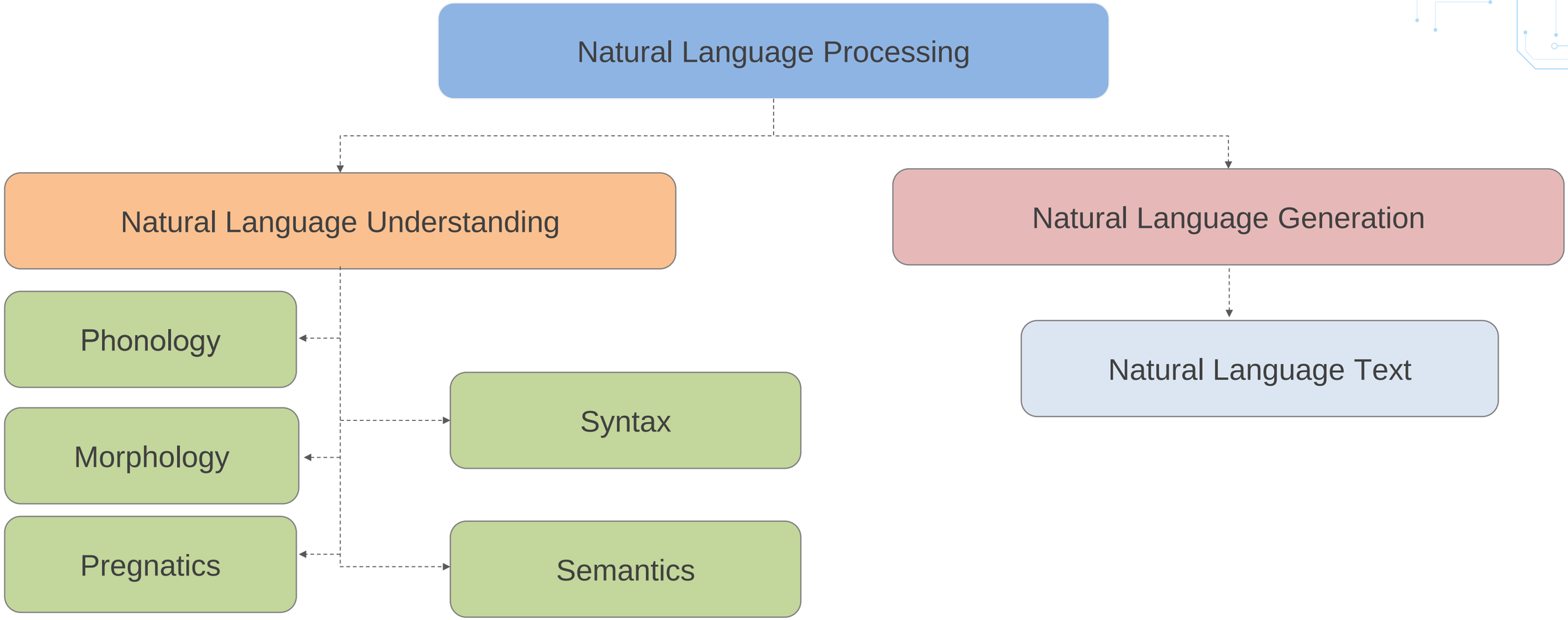
Natural Language Processing

Natural Language Processing (NLP) technology allows computers to comprehend human language.

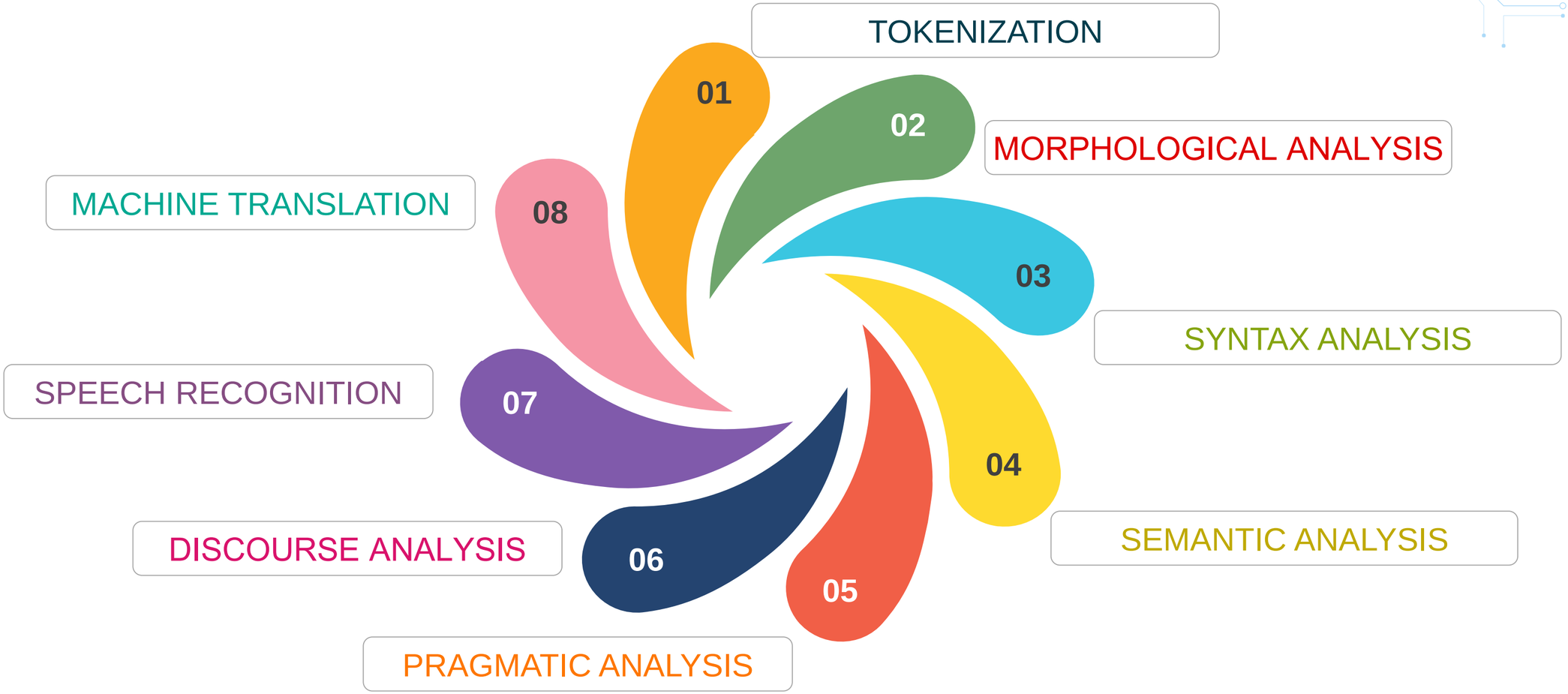


Classification of NLP

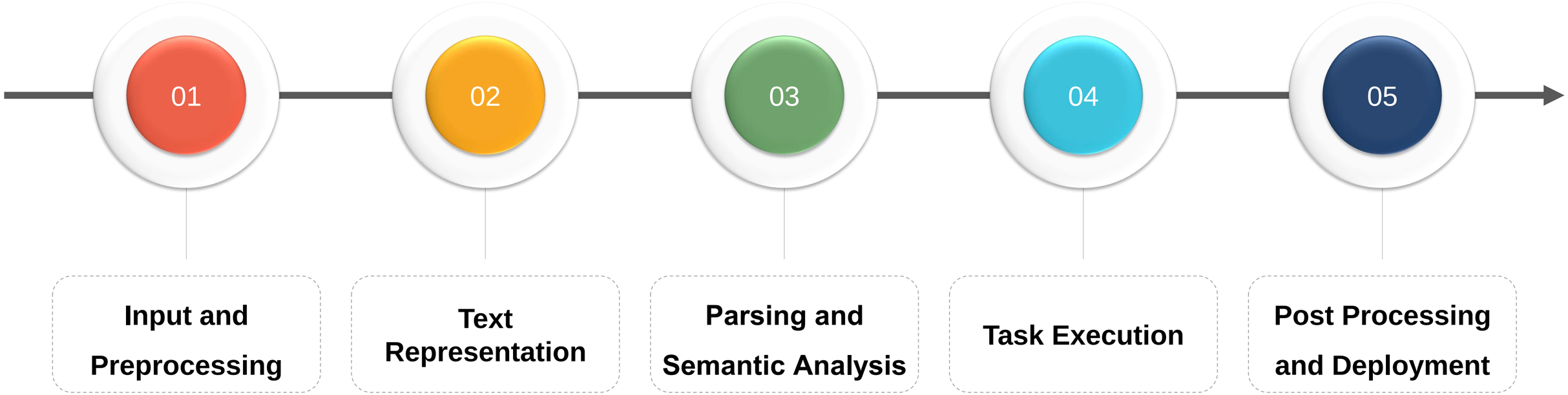
Classification of NLP



Components of NLP

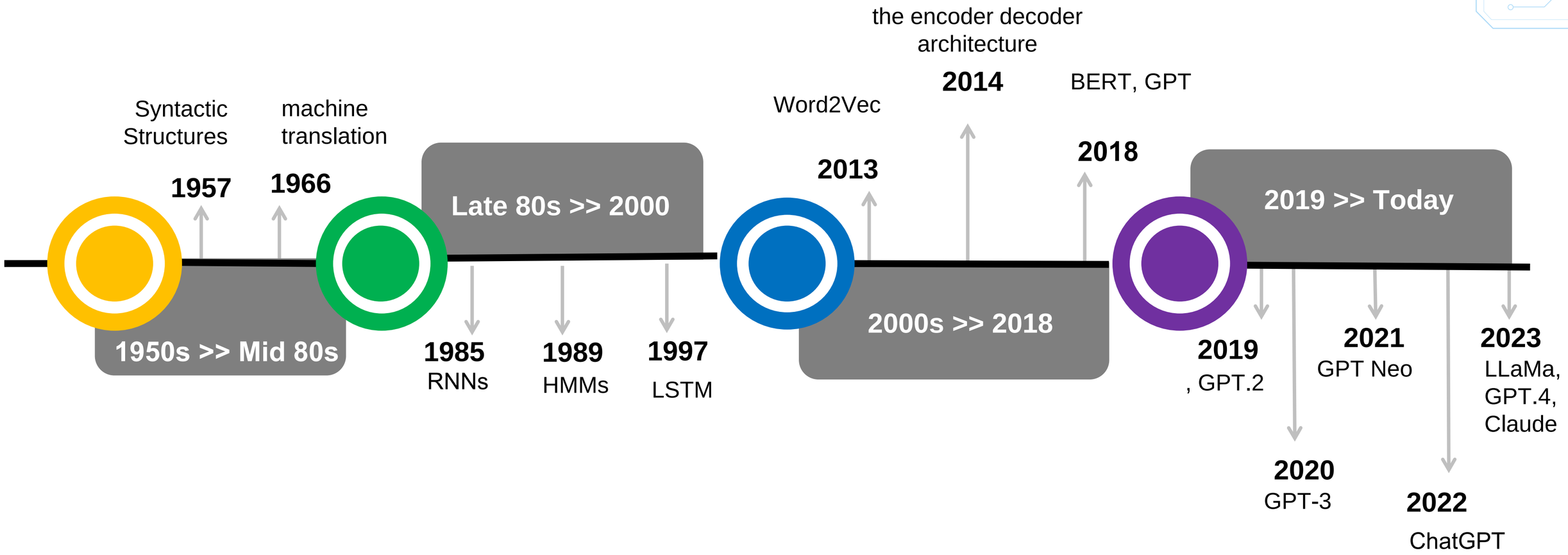


Working Process of NLP



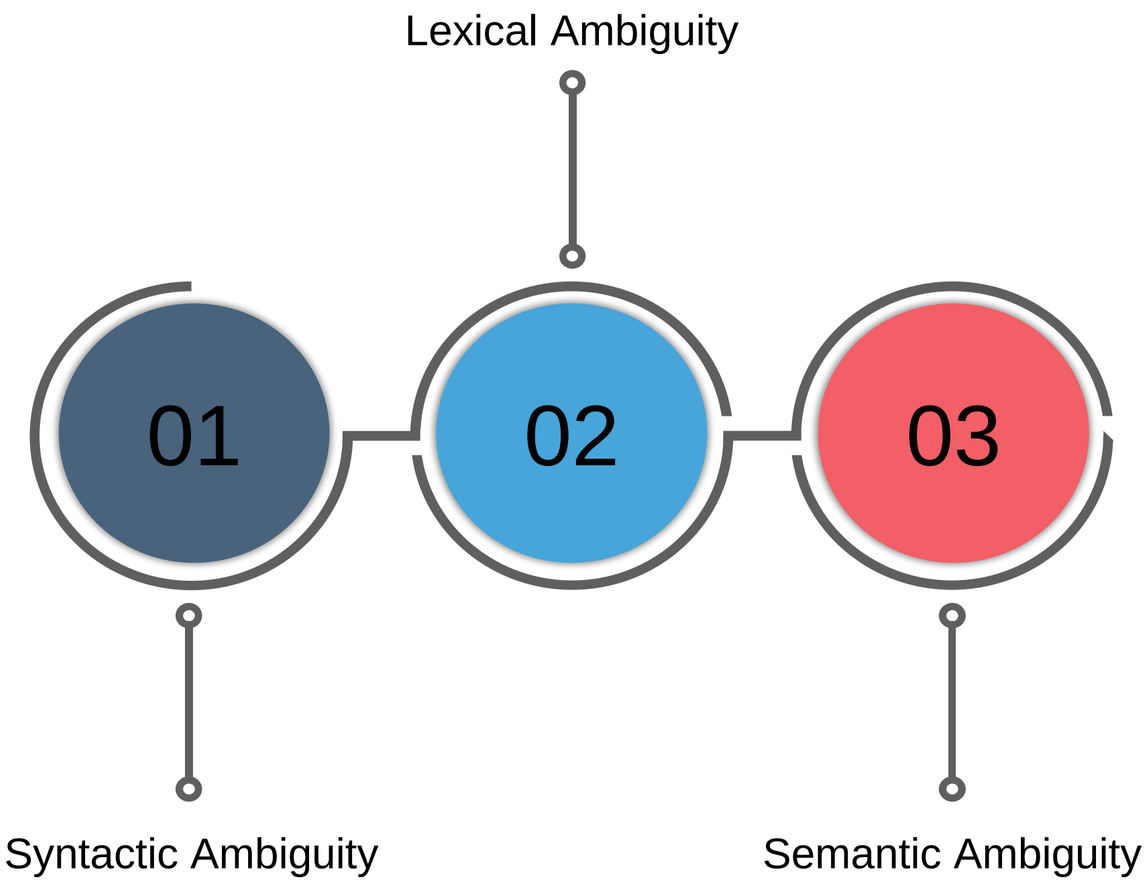
History of NLP Development

History and Evolution of NLP



Key Challenges in NLP

Ambiguity in NLP



"Bank" – Financial Institution vs Riverbank

I saw the man with the Telescope

Star

Contextual Understanding

Context Sensitivity

“Bark” as Tree and Dog Sound

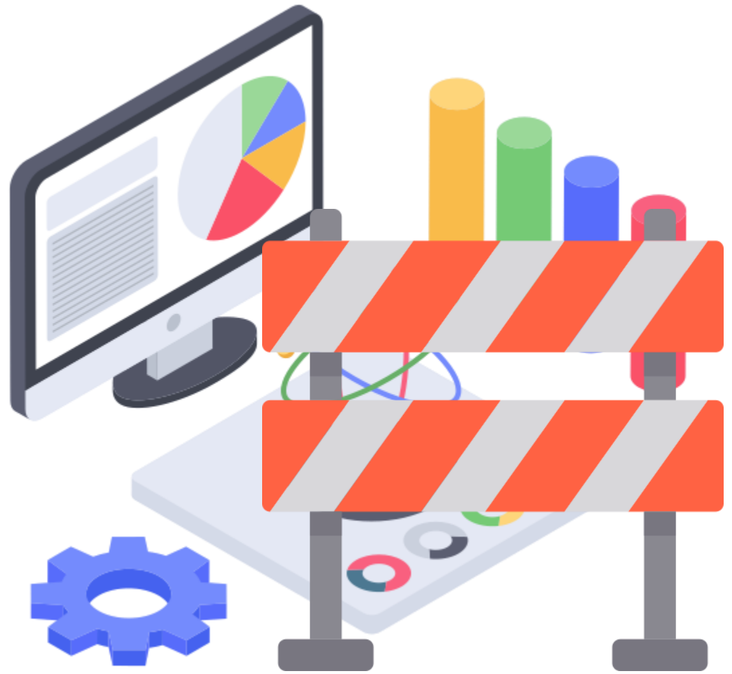
Long-Range Dependencies

Pronouns referring to previous names

Data Sparsity & Domain-Specific Language

Lack of Data

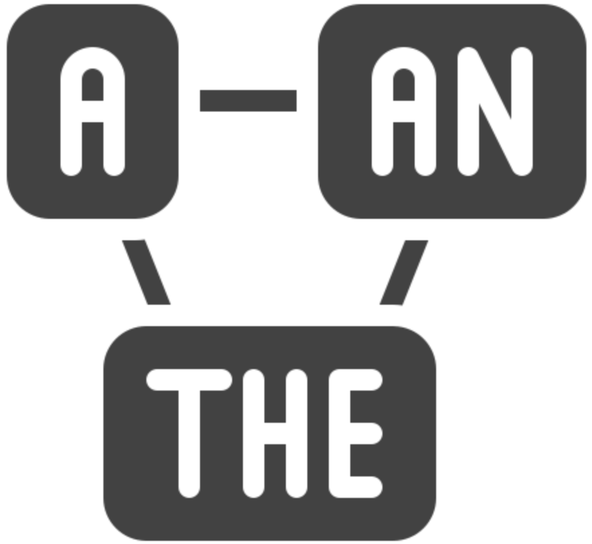
Domain Specific Terms



Multilingualism

Language Diversity

Low-Resource Languages



Sentiment and Emotion Recognition

Sarcasm/Irony: Difficulty detecting tone (e.g., "Great, another flat tire").

Mixed Sentiments: Sentences expressing multiple emotions (e.g., "Good food, bad service")

Named Entity Recognition (NER)

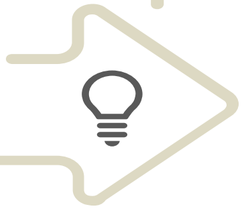
Entity Ambiguity: Identifying names, dates, and locations that may have multiple meanings (e.g., "Apple").

Context-Dependent Entities: Entities whose interpretation varies based on context.

Bias and Fairness

Bias in Data

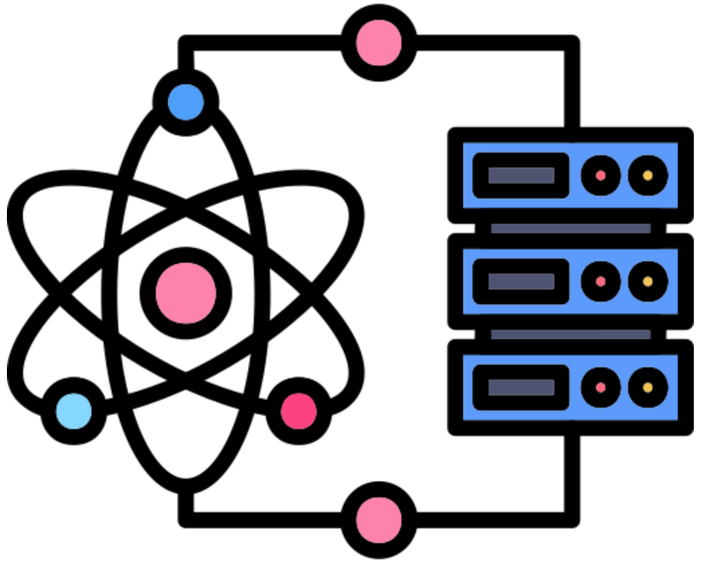
NLP models can inherit societal biases from training data, leading to unfair outcomes.



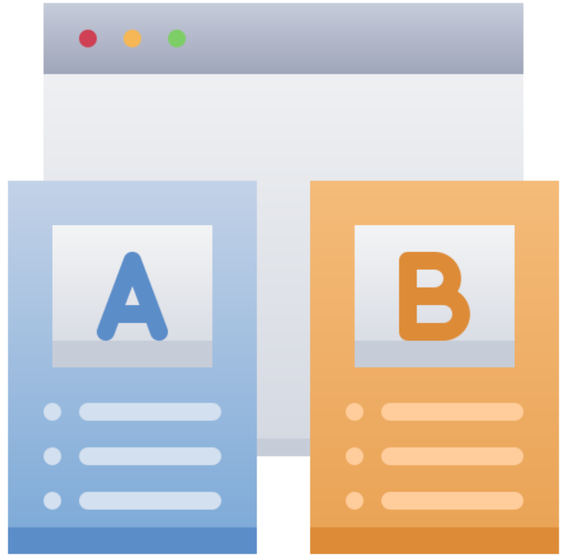
Biases can impact critical sectors like hiring, healthcare, and law enforcement.

Sensitive Applications

Computational Complexity & Evaluation



High Resource Demand



Evaluation Challenges

Real-World NLP Applications

NLP Applications

Health Care

- Medical Chatbots
- Medical Record Analysis

Finance

- Sentiment Analysis
- Fraud Detection

Retail

- Social Media Monitoring
- Product Recommendation

Education

- Automated Grading
- Multilingual Learning

Entertainment

- Multilingual Subtitles
- Content Recommendation

Rule-Based and Statistical Approaches

Rule-Based NLP

Experts define **grammar rules, lexicons, and pattern-based heuristics**.

It works well when language structures are fixed and predictable.

Lexicon-Based Sentiment Analysis:

- Wordlist: ["good," "excellent," "bad," "terrible"].
- If "good" → Positive sentiment, If "bad" → Negative sentiment.

Grammar-Based Chatbots (ELIZA):

- User: "I feel sad.“
- Rule: If the input contains "feel [emotion]," respond with "Why do you feel [emotion]?"

Named Entity Recognition (NER) with Handcrafted Rules:

- Rule-based pattern: (Dr.| Mr. | Ms.)\s[A-Z][a-z]+ → Identifies names.

Statistical Based NLP

Uses probability and statistical techniques to model language patterns.

Learns from large datasets instead of relying on fixed rules.

n-Gram Language Models:

- Predicts the next word based on previous words.
- Example (Bigram model):"New York is a [great]" vs. "New York is a [banana]""great" is more probable.

HMMs for Part-of-Speech (POS) Tagging:

- Uses probabilities to assign POS tags. "She (PRONOUN) plays (VERB) football (NOUN)."

Naïve Bayes for Spam Detection:

- Classifies emails as spam/not spam based on word probability.

Hybrid Based Approach

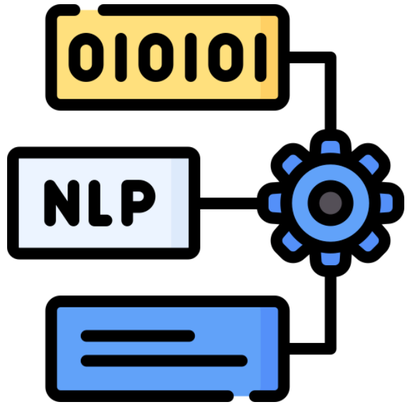
Modern NLP uses all the approaches

Google Search Engine:

- Rule-Based: Keyword-based search algorithms.
- Statistical: Ranking based on click-through rates.

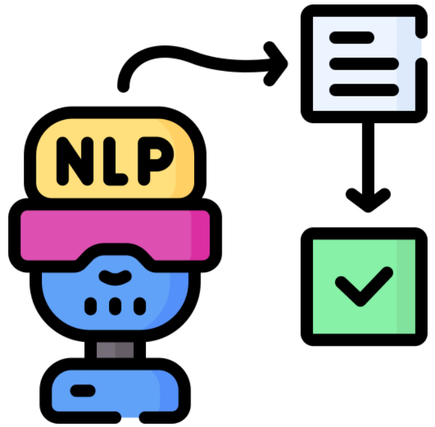
Speech Assistants (Siri, Alexa, Google Assistant):

- Rule-Based: Commands like "Turn on the light.“
- Statistical: Probabilistic language modeling.



Spam Detection (Gmail):

- Rule-Based: Keywords like "free money" flagged.
- Statistical: Bayesian spam filters.



Morphology: Words, Stems and Lemmas

What is Morphology?

Morphology is the study of how words are formed, structured, and related to each other in a language. It helps NLP understand word variations, meaning, and grammar.

Inflectional Morphology

"play" → "played"

(verb tense changes, meaning remains)

Derivational Morphology

"happy" → "happiness"

(new word with different meaning)

Word

A word is the smallest unit of meaning in a language that can function independently in speech and writing.

1. **Root Words**

   Example: "friend" (the root of "friendship," "friendly").

2. **Inflected Words**

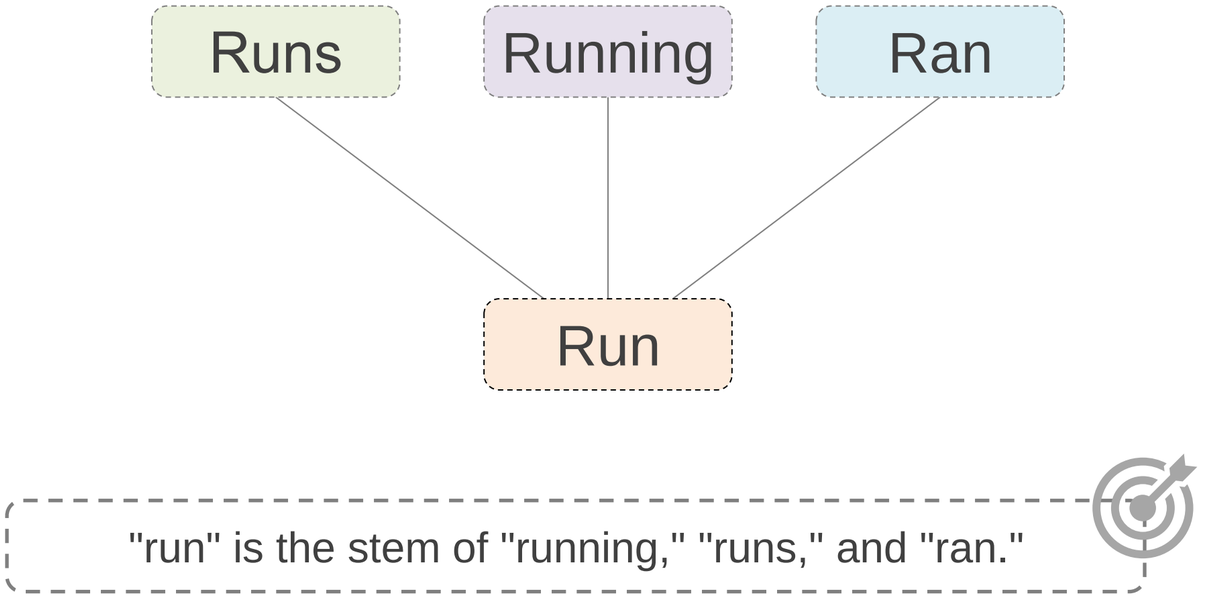
   Example: "write" → "writes," "writing," "written.“

3. **Derived Words**

   Example: "happy" → "unhappy" (prefix added), "act" → "actor" (suffix added).

Stem

A stem is the base form of a word, serving as the foundation for its grammatical variations.



Lemma

A lemma is the dictionary-based form of a word used as the standard reference in linguistic analysis.

Inflected Forms	Lemma
running, ran, runs	run
better, best	good
mice	mouse

What is Morphological Analysis

Morphological analysis studies word structure and formation, essential for understanding and using language effectively.

01

Input the Word.

02

Split the root in Prefixes and Affixes.

03

Identify the Grammatical and semantic function of each part.

04

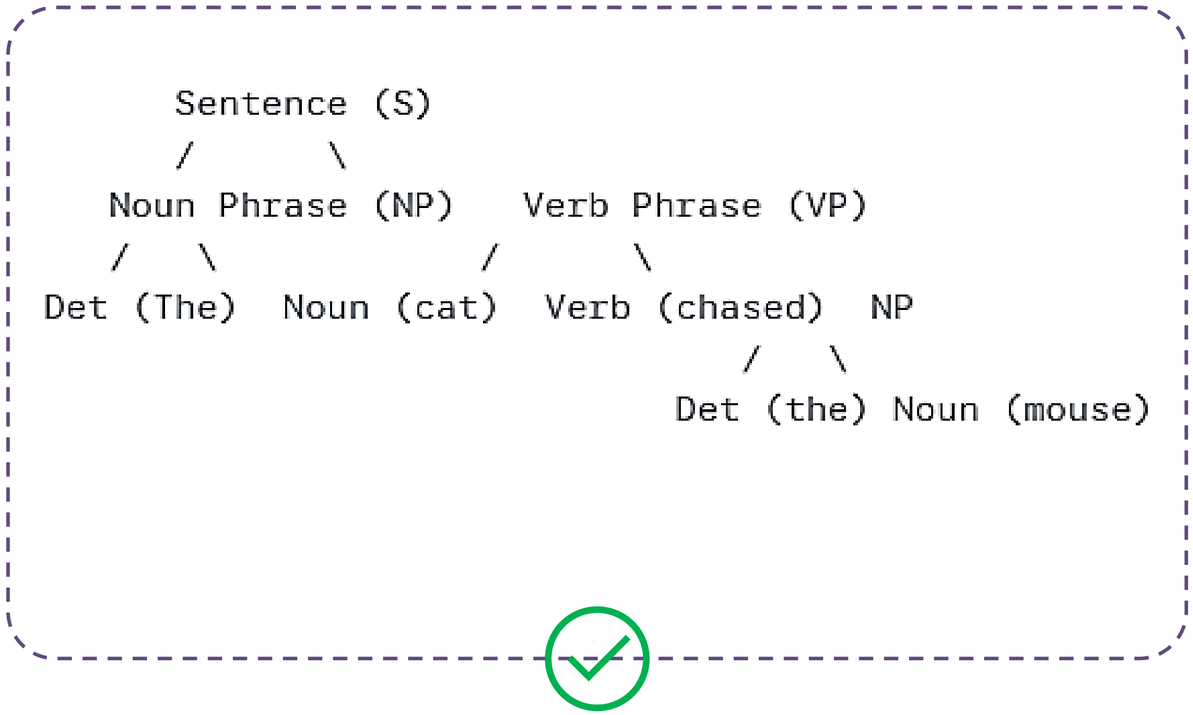
Output the root or lemma

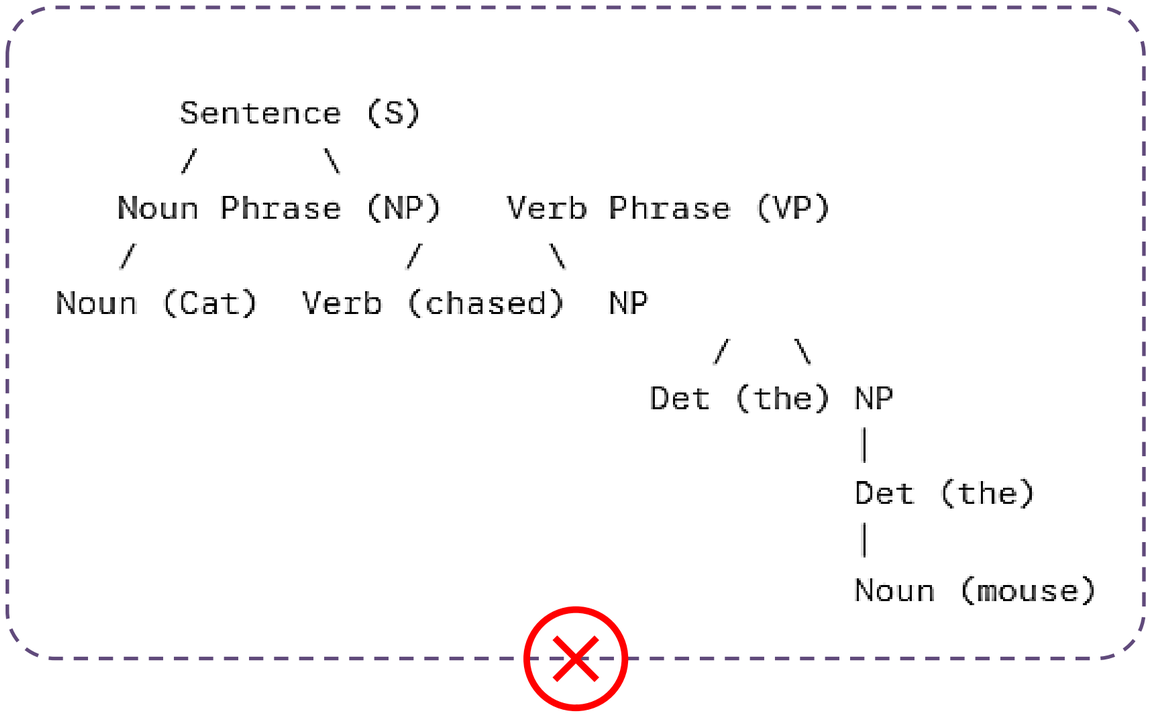
Sentence Structuring

What is Syntax?

Syntax is the study of rules governing the structure and arrangement of words in a sentence to convey meaning.

The cat chased the mouse





What is Syntax Tree?

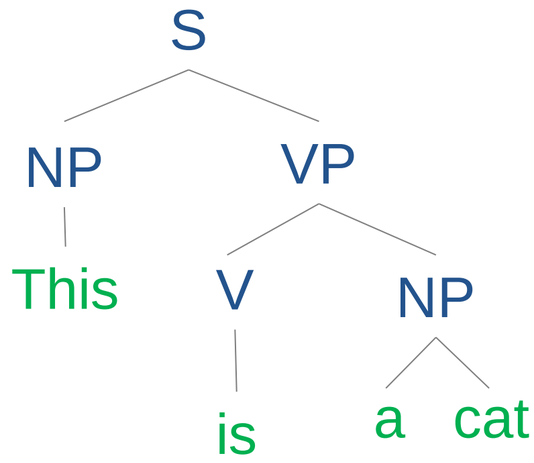
A Syntax tree or a parse tree is a tree representation of different syntactic categories of a sentence. It helps us to understand the syntactical structure of a sentence.

S (Sentence)

NP (Noun Phrase)

VP (Verb Phrase)

V (Verb)



What is Sentence Structure?

Sentence structure refers to the grammatical organization of words and phrases in a sentence.

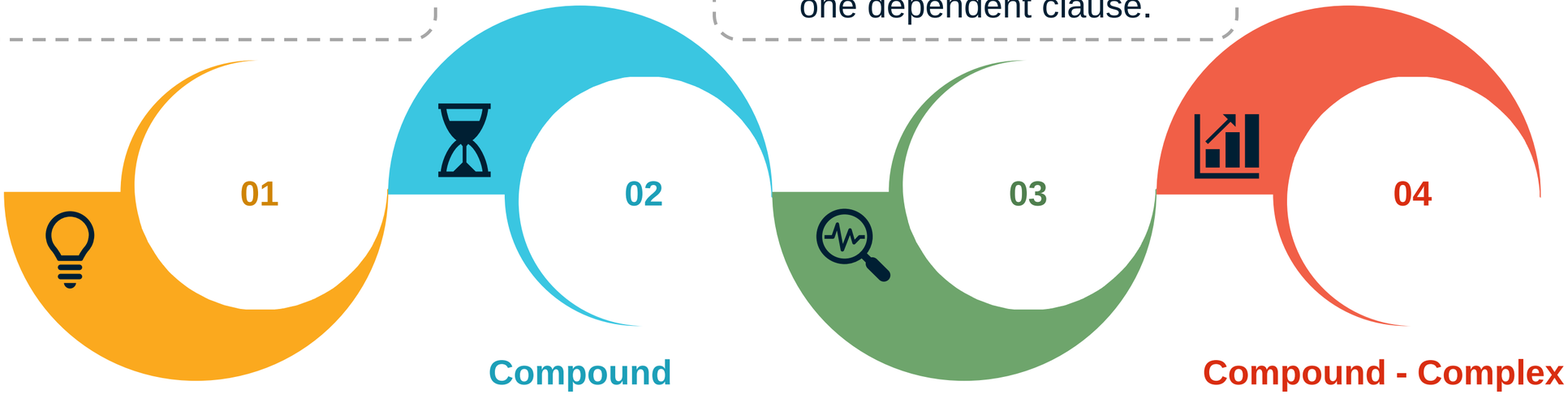
Types of Sentence Structure

Simple

One independent clause.

Complex

One independent clause and one dependent clause.



Compound

Two independent clauses joined by a conjunction

Compound - Complex

Two independent clauses and one or more dependent clauses.

Semantics in NLP: Understanding Meaning and Context

What is Semantics?

Semantics is the study of meaning in language, focusing on how words, phrases, and sentences convey meaning.

"I bank my savings"

"I sit by the river bank"

"The stars are beautiful"

"The stars walked the red carpet"

Lexical Semantics

01 **Synonyms**

happy – joyful

02 **Antonyms**

hot – cold

03 **Hyponyms**

rose is a hyponym of flower

04 **Hypernyms**

animal is a hypernym of dog

Semantic Ambiguity

01

Polysemy

A word with multiple related meanings,
(“bank” as a financial institution vs. riverbank).

Homonymy

A word with multiple unrelated meanings,
(“bat” as an animal vs. a sports bat).

02

03

Contextual Meaning

Resolving ambiguity based on sentence context.

Word Sense Disambiguation Approaches

WSD is the process of identifying the correct meaning of a word in the given context.

- Rule Based
- Supervised Learning
- Unsupervised Learning

In “He went to the bank to withdraw money”, WSD determines "bank" as a financial institution.

Compositional Semantics

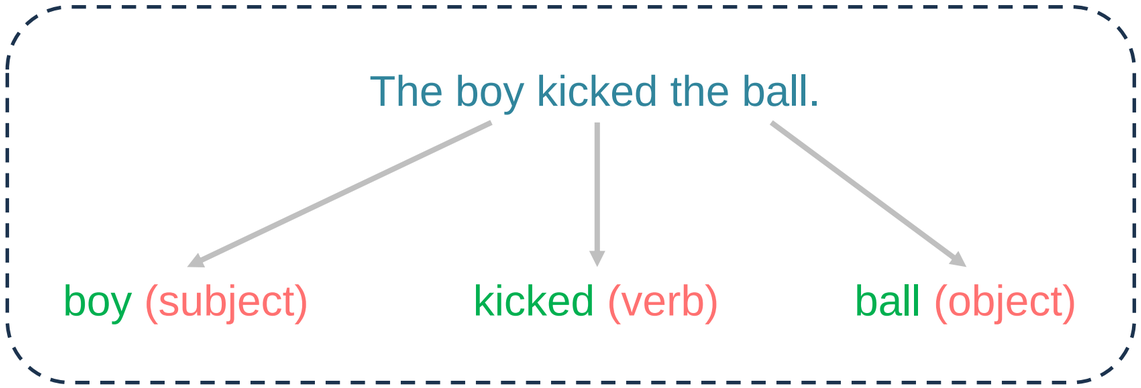
Compositional semantics studies how words combine to form meaningful sentences.

Principle of Compositionality:

The meaning of a sentence is derived from its words and structure.

Semantic Role Labeling:

Identifies roles like agent, action, and object ("John" is the agent in "John kicked the ball.").



Semantic Representation

1

Word Embeddings

Word meaning as numerical vectors.

Example: "King" & "Queen" are close vectors.

2

Knowledge Graphs

Entity networks & real-world relationships.

Example: Google Knowledge Graph: Person -> Place -> Thing.

3

Ontologies

Structured vocabularies & domain relationships.

Example: Medical ontology: Disease -> Symptom.

Pragmatics in NLP

What is Pragmatics?

Pragmatics is the study of how context influences meaning in communication.

Jamie: Hey, you're leaving early today?

Alex: Yeah, just need a break.

Jamie:

Good for you! Sometimes we all need one. Hope you get to relax.

Scenario 1

Jamie:

Again? This is the third time this week

Scenario 2

Jamie:

Must be nice. Some of us actually have to work.

Scenario 3

Jamie:

Oh sure, just abandon us while we drown in work!

Scenario 4

Role of Context

Linguistic Context

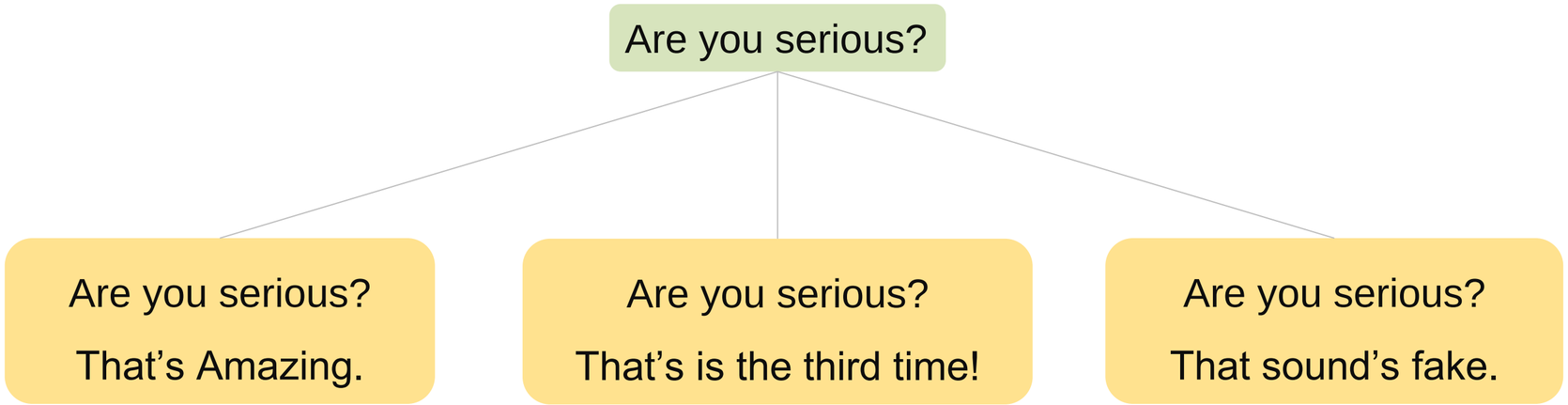
(*"John bought a car. He loves it." → He = John*).

Situational Context

(*"It's cold in here."* → Implicit request to close the window).

Social Context

(*"Let's catch up soon!"* may not imply an actual meeting).



Reference Resolution

Reference resolution is the process of determining what a pronoun or noun phrase refers to in a sentence or paragraph.

01

Anaphora Resolution

"Anna loves music. She plays the piano." (She = Anna).

02

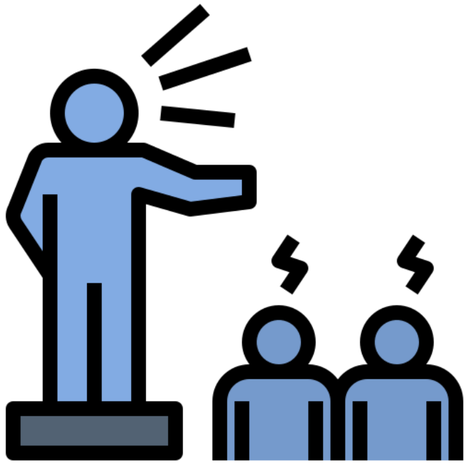
Coreference Resolution

"Barack Obama was the 44th President. Obama led the country." (Both refer to the same person).

Speech Acts in Communication

Speech acts are classified based on what they intend to achieve, not just what they say literally.

Close the door.



It’s getting late.

Conversational AI

Book a flight to Paris.

Book a flight from NYC to La on Monday.

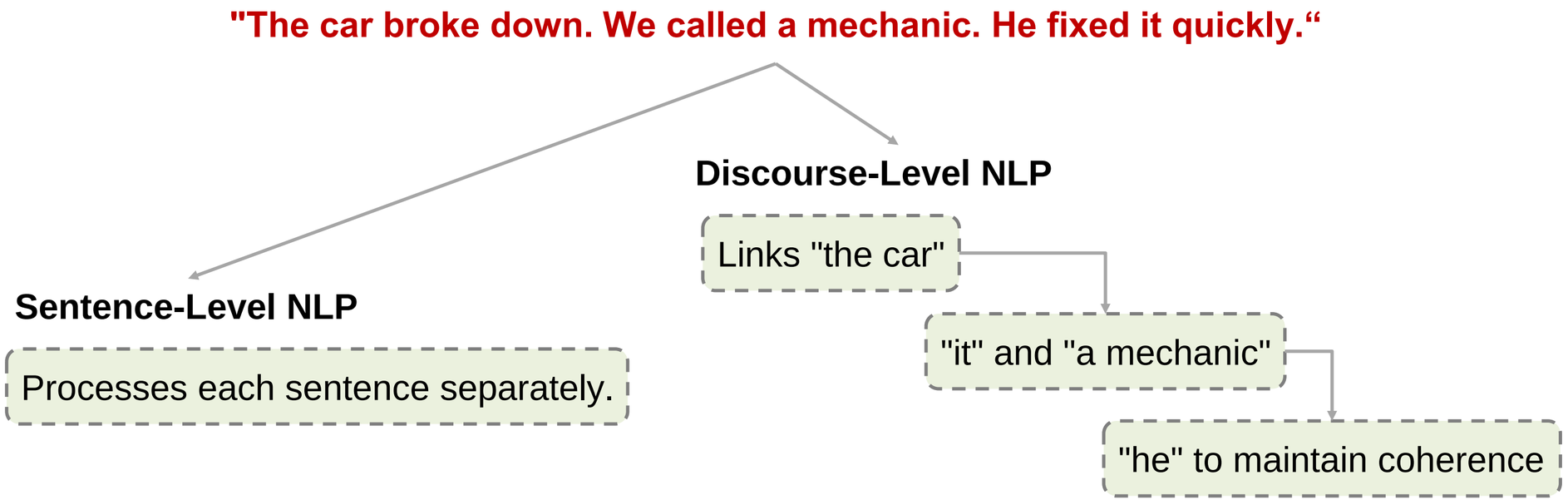
1 Intent Recognition

2 Slot Filling

Discourse Analysis in NLP

Discourse in NLP

Discourse analysis in NLP focuses on understanding text beyond individual sentences, analyzing how sentences connect to form meaningful, coherent communication.



Topic Segmentation and Text Coherence

Topic segmentation divides text into logically related sections, while coherence evaluation checks how well ideas connect throughout the text.

"AI is transforming healthcare. Doctors now use AI for diagnosis. Meanwhile, in finance, AI optimizes trading strategies.

Healthcare AI

Finance AI

Coherence and Cohesion

Coherence: Logical consistency in a text, making it meaningful.

Cohesion: Use of linguistic devices (pronouns, synonyms, lexical chains) to connect ideas smoothly.

Lexical Chains

Connecting related words across sentences.

Example: "Tiger" → "Big cat" → "Predator"

(all referring to the same entity).

Anaphoric Relations

Resolving pronouns and references in a text.

Example: "John bought a new car. He loves it."

(He = John, It = Car).

Discourse Structure Theories

Discourse structure theories provide frameworks for understanding how texts are organized and how sentences relate to each other.

1

Discourse Representation Theory (DRT)

Example: Understanding "He" in "John entered the room. He sat down."

Rhetorical Structure Theory (RST)

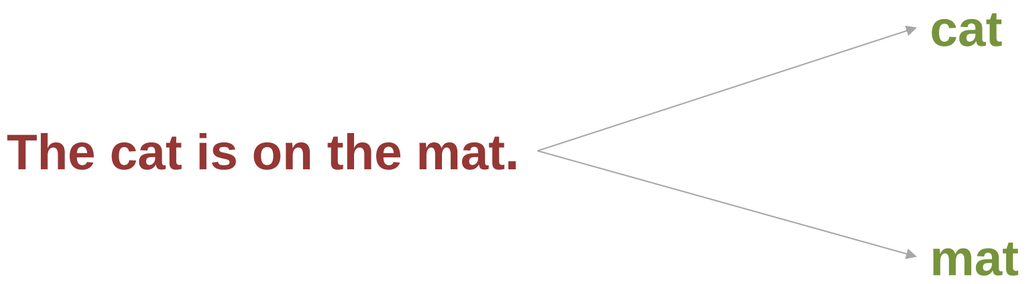
Example: "I was hungry, so I ate a sandwich." → Cause-Effect Relation.

2

Basic Text Cleaning Techniques

Stopword Removal

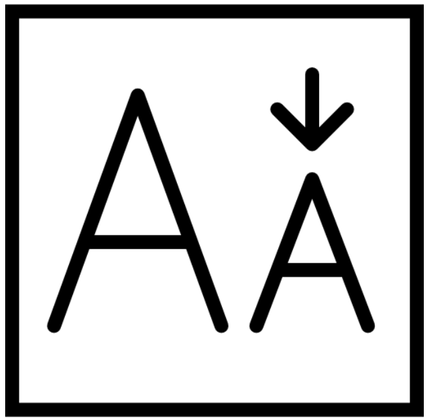
Removing common words (e.g., "the," "is," "in") that do not add significant meaning to text.



Lower Casing

Converting all text to lowercase to ensure uniformity in NLP tasks.

Hello World → **hello world**



Tokenization

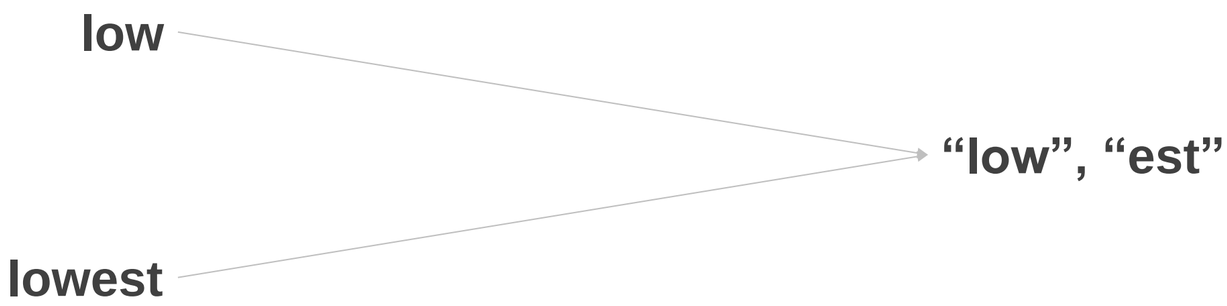
Splitting text into smaller units (tokens) for easier processing

```
Tokenization in Natural Language Processing
```

```
['Tokenization', 'in', 'Natural', 'Language', 'Processing']
```

Byte-Pair Encoding

A subword tokenization method that merges frequent character pairs iteratively.



One Hot Encoding and TF-IDF

One-Hot Encoding

One-Hot Encoding represents each word as a binary vector where only one position is 1, and the rest are 0.

	Test_data	Test_data_apple	Test_data_cat	Test_data_dog	Test_data_fish
0	cat	0.0	1.0	0.0	0.0
1	dog	0.0	0.0	1.0	0.0
2	fish	0.0	0.0	0.0	1.0
3	apple	1.0	0.0	0.0	0.0

TF-IDF

TF-IDF is a numerical statistic used in information retrieval and natural language processing to evaluate the importance of a word in a document within a collection or corpus.

$$TF - IDF(w) = TF(w) \times IDF(w)$$

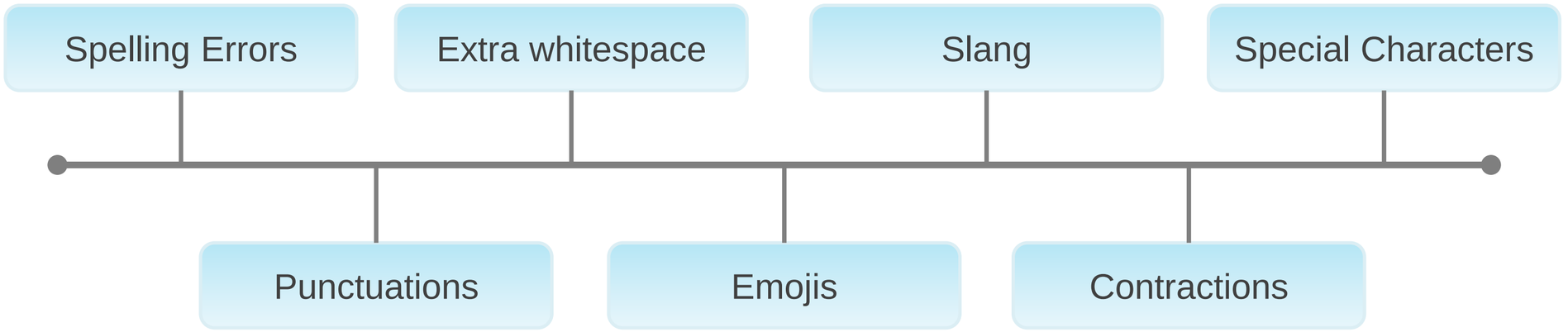
Reduces the influence of common words (e.g., "the", "is")

Useful for information retrieval (e.g., search engines)

Handling Noise and Special Characters

What is Noise?

Noise in NLP refers to any unwanted, irrelevant, or misleading characters, symbols, or text that do not contribute meaningfully to analysis.



OMG!! Ths movie was sooo gr8 😍💯🔥 **Vs** This movie was so great"

Handling Contractions and Slang

Contractions are shortened versions of words (e.g., don't → do not).

Slang includes informal expressions (e.g., *gonna* → *going to*).

01 **Expand Contractions:** Using predefined dictionaries (*"can't"* → *"cannot"*).

02 **Replace Slang:** Using slang dictionaries (*"lol"* → *"laughing out loud"*).

03 **Standardize Text:** Convert informal words to their **formal equivalents**.

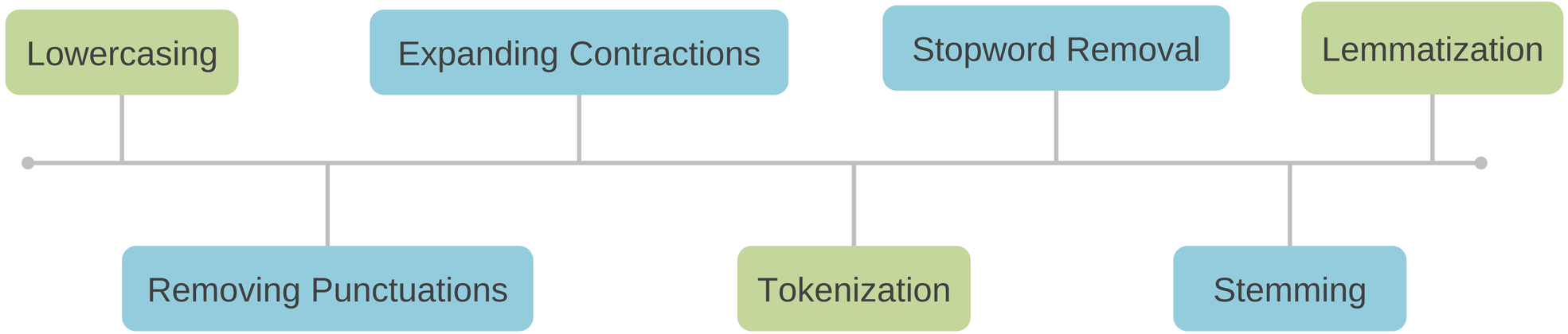
Dealing with Emojis and Non-Standard Text

- Emojis convey emotions visually (😊, 😍, 😂).
- Non-standard text includes special characters, misspellings, and internet abbreviations.

1. Remove or Replace Emojis: Convert 😍 → "love".
2. Use Emoji Libraries: Python’s emoji package can translate emojis to text.
3. Normalize Abbreviations: LOL → Laughing Out Loud.

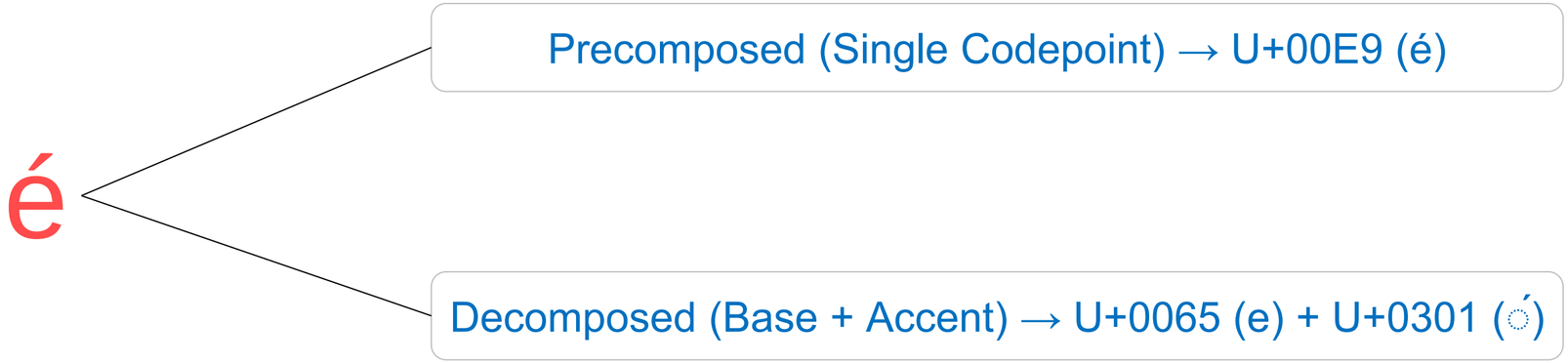
Normalization

Normalization in NLP refers to transforming text into a **standard format** to ensure consistency and improve text analysis.



Why Normalize Unicode?

- Different Unicode representations can exist for the same character.
- Ensures consistency in text processing (searching, matching, storage).



Unicode Normalization Forms

Form	Type	Purpose	Example(é)
NFC	Composed	Standard for storage & display	U+00E9 (é)
NFD	Decomposed	Splits base + accent	U+0065 (e) + U+0301 (◌́)
NFKC	Compatibility Composed	Converts variants (𝟙 → 1)	U+0031 (1)
NFKD	Compatibility Decomposed	Decomposes & simplifies symbols	U+0031 (1)

Building an NLP Pipeline for Multilingual Tweet Cleaning (Demonstration)

Note: Refer to Module 1: Demo 1 on LMS for detailed steps.

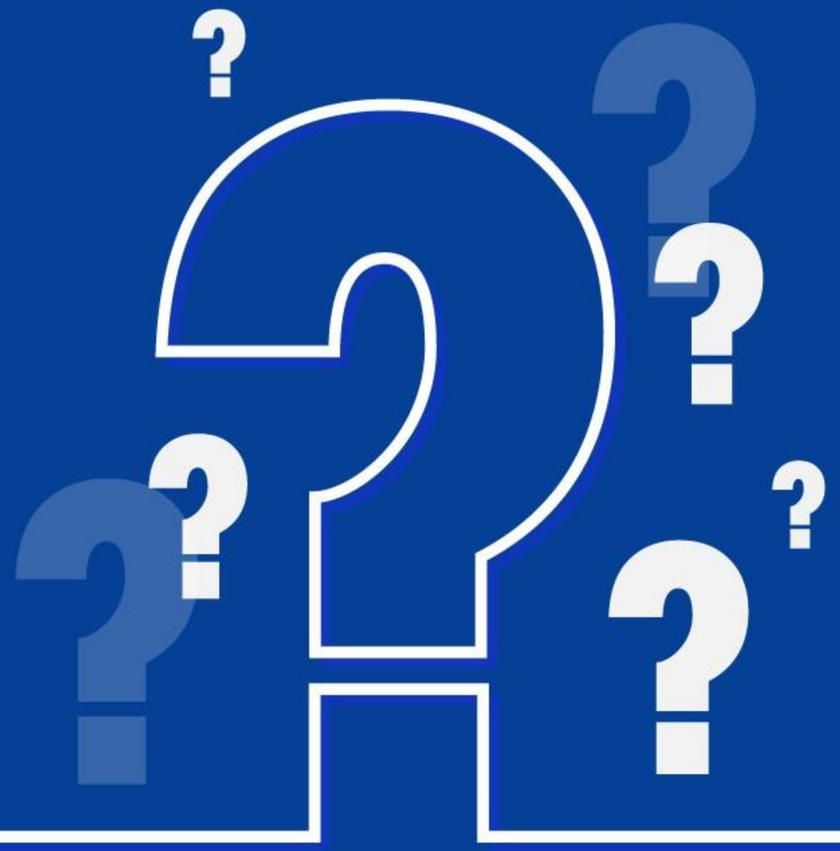
Summary

In this lesson, you have learned to:

- Explain the scope and foundational concepts of Natural Language Processing.
- Compare various NLP approaches including rule-based, statistical, and hybrid models.
- Utilize techniques for text cleaning, morphological analysis, and sentence structuring.
- Interpret the roles of syntax, semantics, and pragmatics in language understanding.

edureka! a Veranda ENTERPRISE

Questions



edureka! a Veranda ENTERPRISE

Feedback

edureka! a Veranda ENTERPRISE

Thank You

For information, Please Visit our Website
www.edureka.co