NATURAL LANGUAGE PROCESSING

- Humans communicate through some form of language either by text or speech.
- To make interactions between computers and humans, computers need to understand natural languages used by humans.
- Natural language processing is all about making computers learn, understand, analyse, manipulate and interpret natural(human) languages.
- NLP stands for Natural Language Processing, which is a part of Computer Science, Human language, and Artificial Intelligence.
- Processing of Natural Language is required when you want an intelligent system like robot to perform as per your instructions.
- Applications that include NLP are chatbots, Email classification and spam filters.
- The input and output of an NLP system can be Speech or Written Text

There are two components of NLP, Natural Language Understanding (NLU)and Natural Language Generation (NLG).

NLU	NLG
Natural Language Understanding (NLU) which involves transforming human language into a machine-readable format.	Natural Language Generation (NLG) acts as a translator that converts the computerized data into natural language representation.
It helps the machine to understand and analyse human language by extracting the text from large data such as keywords, emotions, relations, and semantics.	It mainly involves Text planning, Sentence planning, and Text realization.

The NLU is harder than NLG

Phonology	Study of how sounds are organized and used in a language
Morphology	Study of structure and formation of words
Morpheme	Smallest unit in a language that carries meaning
Syntax	Study of rules and principles that govern the structure of sentence.
Semantics	Focus on understanding the meaning of words
Disclosure	Studies the interaction between sentence in a larger context
World Knowledge	General understanding of a person about the world

There are five general steps:

Lexical Analysis	The first phase of NLP is the Lexical Analysis. This phase scans the source code as a stream of characters and converts it into meaningful lexemes.

Syntactic Analysis	Syntactic Analysis is used to check grammar, word arrangements, and shows the relationship among the words. The sentence such as “The school goes to boy” is rejected by English syntactic analyser.
Semantic analysis	Semantic analysis is concerned with the meaning representation. It mainly focuses on the literal meaning of words, phrases, and sentences. The semantic analyser disregards sentence such as “hot ice-cream”.
Discourse Integration	Discourse integration involves understanding sentences that come before and after, which helps make sense of the overall meaning.
Pragmatic Analysis	It involves deriving those aspects of language which require real world knowledge. Example: "Open the door" is interpreted as a request instead of an order.

Word Tokenization:

- Word tokenization is the process of breaking a continuous stream of text into individual words.
- In many languages, words are considered the smallest units that can form a complete expression by themselves.

Lexemes:

- Lexemes are sets of linguistic forms that represent a concept, along with its alternative forms.
- They make up the lexicon of a language and can include verbs, nouns, adjectives, and other parts of speech.
- The primary form of a lexeme is called its lemma. Lexemes can be inflected (changing forms) and derived (transformed into related forms).
- Here's a simplified example sentence: "Tokenization splits a sentence into separate words. For instance, 'running' and 'ran' are forms of the lexeme 'run'."

Morphological Typology:

- Morphological typology is a way to group languages based on how they build words using morphemes.
- It divides languages into synthetic (combining morphemes) and analytic (using word order and helper words) types.
- Synthetic languages are further split into agglutinative (add clear pieces for meaning) and fusional (combine many meanings in one ending) languages.
- Synthetic languages, ones that are not analytic, are divided into two categories: agglutinative and fusional languages.

- Agglutinative languages rely primarily on discrete particles (prefixes, suffixes, and infixes) for inflection, ex: inter+national = international, international+ize = internationalize. •
- While fusional languages "fuse" inflectional categories together, often allowing one word ending to contain several categories, such that the original root can be difficult to extract (anybody, newspaper).

Issues and Challenges:

Irregularity:

- - Morphological parsing aims to simplify and understand words by finding patterns.
- - Initial descriptions of language data might not be perfect due to errors or complexity.
- - To improve accuracy and clarity, we need better ways to describe linguistic data.
- - The rules guiding morphological models are crucial for making them work effectively.

Ambiguity:

- Morphological ambiguity happens when words could mean different things without context.
- Words that look alike but mean different things are called homonyms.
- Ambiguity is found in all parts of language processing.

Morphological Models:

- Designing morphological models has different ways.
- In computational linguistics, various approaches and frameworks have been developed to solve problems in natural and formal languages.
- Let's explore the main types of computational approaches for understanding word structure.

Dictionary Lookup:

- Morphological parsing connects word forms to their linguistic descriptions.
- Some systems list these associations one by one without generalization.
- Similarly, systems that just look up words in lists, dictionaries, or databases lack depth, unless they're in sync with advanced language models."

Finite State Morphology:

- Finite-State Morphology simplifies language analysis using finite-state transducers.
- The two most popular tools supporting this approach, XFST (Xerox Finite-State Tool) and LexTools.
- Finite-state transducers are computational devices extending the power of finite-state automata.
- They consist of a finite set of nodes connected by directed edges labeled with pairs of input and output symbols.
- In such a network or graph, nodes are also called states, while edges are called arcs.

- The set of possible sequences accepted by the transducer defines the input language; the set of possible sequences emitted by the transducer defines the output language.
- In English, a finite-state transducer could analyze the surface string children into the lexical string child [+plural], for instance, or generate women from woman [+plural].

Aspect	**Functional Morphology**	**Unification-Based Morphology**
Approach	Functional programming and type theory principles	Uses unification principles from linguistics and computer science
Representation	Morphological processes as mathematical functions	Morphological processes represented using unification rules
Types	Linguistic elements organized into distinct types and type classes	Involves matching and merging of linguistic features and constraints
Usage	Especially useful for languages with complex, fusional morphologies	Suited for languages with rich inflectional systems and irregularities
Representation	Linguistic concepts like paradigms and categories intuitively represented	Can handle intricate relationships between morphemes and features
Implementatio n	Aimed to be reusable as programming libraries in various language applications	Can capture complex interactions among morphological elements and generate different word forms

Difference between Sentence boundary detection and Topic boundary detection.

Aspect	**Sentence Boundary Detection (SBD)**	**Topic Boundary Detection (TBD)**
Task	Identifies boundaries between sentences in a text	Identifies transitions between different topics or segments in a text
Purpose	Helps in segmenting a text into individual sentences for further analysis or processing	Helps in identifying shifts in topics or segments within a text

Type of Boundary	Identifies the end of one sentence and the beginning of the next	Identifies the boundary where the topic or segment changes
Criteria for Detection	Typically relies on punctuation marks (e.g., period, exclamation, question mark)	Can be based on keywords, changes in writing style, introduction of new concepts, etc.
NLP Applications	Used for various NLP tasks such as machine translation, sentiment analysis, summarization, etc.	Used in document clustering, topic modeling, information retrieval, etc.
Example	"I went to the store. It was crowded."	"In the first section, we discussed NLP techniques. In the next section, we'll explore applications."

PROCESSING RAW TEXT

Text Segmentation	Text Classification
Text segmentation involves dividing a continuous piece of text into meaningful segments or units, such as sentences, paragraphs, or sections.	Text classification involves assigning predefined labels or categories to text documents based on their content.
Segmentation helps in identifying boundaries between distinct textual units.	Text classification is used to automatically categorize and organize large amounts of text data.
Various methods can be used for text segmentation, such as rule-based approach, machine learning models, or statistical methods. Ex: Speech Recognition	Machine learning algorithms, such as Naïve Bayes, SVM, and deep learning models like CNN, RNN Ex: Sentiment Analysis

Language Modelling:

- Language modelling is a fundamental concept in natural language processing (NLP) that involves developing computational models to predict the likelihood of a sequence of words occurring in each language. It aims to capture the structure, grammar, and context of language to generate coherent and contextually appropriate text.
- Language models are primarily of two kinds: N-Gram language models and Grammar-based language models such as probabilistic context-free grammar

Statistical Language Modelling:

- Statistical language modelling is a specific approach to language modelling that employs statistical techniques to estimate the probabilities of word sequences based on observed data. It relies on the assumption that the probability of a word occurring depends on the preceding words in the sequence.

- Statistical language models use traditional statistical techniques like N-grams, Hidden Markov Models (HMM).

Neural Language Modelling:

- Neural Language Models use different kinds of approaches like neural networks such as feedforward neural networks, recurrent neural nets, attention-based networks, and transformers-based neural nets late to model the language, and they have also surpassed the statistical language models in their effectiveness.

N-gram Models

- N-gram models are a particular set of language models based on the statistical frequency of groups of tokens.
- An n-gram is an ordered group of n tokens.
- The bigrams of the sentence. The cat eats fish. are (The, cat), (cat, eats), (eats, fish) and (fish, .).
- The trigrams are (The, cat, eats), (cat, eats, fish) and (eats, fish, .).
- The smallest n-grams with n =1 are called unigrams. Unigrams are simply the tokens appearing in the sentence.
- The conditional probability that a certain token appears after previous tokens are estimated by Maximum Likelihood Estimation on a set of training sequences.

Intuitive Formulation:

- The intuitive idea behind n-grams and n-gram models are that instead of computing the probability of a word given its entire history, we can approximate the history by just the last few words like humans do while understanding speech and text.

Illustration for N-gram probabilities:

$P(wn | w1 \ldots wn-1) \approx P(wn)$ unigram

$P(wn | w1 \ldots wn-1) \approx P(wn | wn-1)$bigram

$P(wn | w1 \ldots wn-1) \approx P(wn | wn-1\ wn-2)$ trigram

$P(wn | w1 \ldots wn-1) \approx P(wn | wn-1\ wn-2\ wn-3)$ 4gram

$P(wn | w1 \ldots wn-1) \approx P(wn | wn-1\ wn-2\ wn-3\ wn-4)$ 5-gram

Choice of N in N-Gram models:

- The accuracy of the model increases with an increase in N.
- But with bigger N values, we run into the risk that we may not get good estimates for N-Gram probabilities, and the N-Gram tables will be more sparse.
- Smaller the N, the model will be less accurate. But we may get better estimates for N-Gram probabilities, and the N-Gram tables will be less sparse.

Aspect	**Unigram**	**Bigram**	**Trigram**
	A unigram, also known as a 1-gram, refers to a single word in a sequence of text.	A bigram, or 2-gram, refers to a pair of consecutive words in a sequence.	A trigram, or 3-gram, refers to a sequence of three consecutive words.
Contextual Accuracy	Least	More than Unigram	Better than Bigram
Complexity	Simple	Moderate	Increased
Application	Rarely used	Basic tasks	Advanced tasks
Coverage	Limited	Some	Better

Probability Estimation:

- There are two main steps generally in building a machine learning model: Defining the model and Estimating the model's parameters, which is called the training or the learning step.
- There are also two quantities we need to estimate for developing the language models for all words in the vocabulary.
- For example, Pr(Colorless green ideas sleep furiously)

Maximum Likelihood estimation:

- The most basic parameter estimation technique is the relative frequency estimation (frequencies are counts) which is also called the method of Maximum Likelihood Estimation (MLE).
- The estimation simply works by counting the number of times the word appears conditioned on the sentence and then normalizing the probabilities. We also need some source text corpora.
- Chain rule of probability in estimation: To estimate the probabilities, we usually rely on the Chain Rule of Probability, where we decompose the joint probability into a product of conditional probabilities using the independence assumption.
- It is also to be kept in mind that estimating conditional probabilities with long contexts is usually difficult, and for example, conditioning on 4 or more words itself is very hard.

Markov assumption in probability estimation:

- The use of Markov assumption in probability estimation solves a lot of problems we encounter with data sparsity and conditional probability calculations.
- The assumption is that the probability of a word depends only on the previous word(s). It is like saying the next event in a sequence depends only on its immediate past context.
- Markov models are the class of probabilistic models that assume that we can predict the probability of some future unit without looking too far into the past.

Challenges of Probability Estimation:

- Data Sparsity: When dealing with a large vocabulary or a wide range of features, the available data may not cover all possible combinations, resulting in sparse data.
- Zero Probabilities: In cases where a particular event hasn't occurred in the training data, traditional probability estimation methods might assign a zero probability to it.
- Curse of Dimensionality: As the number of features or dimensions increases, the data becomes more sparse, making probability estimation challenging.
- Outliers and Noise: Outliers and noisy data can distort probability estimates, leading to incorrect models and predictions.
- Model Complexity: More complex models, such as higher-order n-grams or deep learning models, introduce more parameters that need to be estimated accurately from limited data.

SYNTAX ANALYSIS

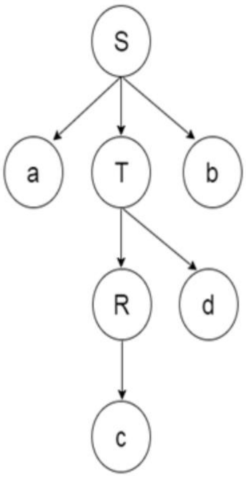
Parser

- Parser is a compiler that is used to break the data into smaller elements coming from lexical analysis phase.
- A parser takes input in the form of sequence of tokens and produces output in the form of parse tree.
- Parsing is of two types: top down parsing and bottom up parsing.

Top down paring

- The top down parsing is known as recursive parsing or predictive parsing.
- Bottom up parsing is used to construct a parse tree for an input string.
- In the top down parsing, the parsing starts from the start symbol and transform it into the input symbol.

Parse Tree representation of input string "acdb" is as follows:



WORKING OF TOP DOWN PARSER:

- In top down technique parse tree constructs from top and input will read from left to right. In top down, In top down parser, It will start symbol from proceed to string.
- It follows left most derivation.
- In top down parser, difficulty with top down parser is if variable contain more than one possibility selecting 1 is difficult.

Working of Top Down Parser :

Let’s consider an example where grammar is given and you need to construct a parse tree by using top down parser

technique.

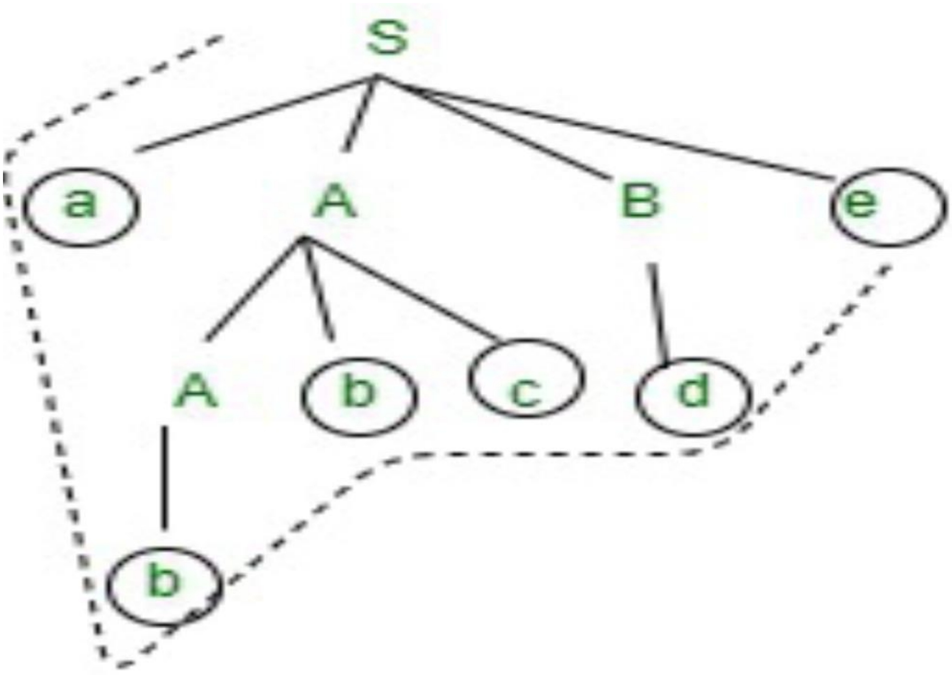
Example –

S -> aABe

A -> Abc | b

B -> d

- Now, let’s consider the input to read and to construct a parse tree with top down approach.
- Input – abbcde
- generate a input string from the grammar for top down approach.
- First, you can start with S -> a A B e and then you will see input string a in the beginning and e in the end.
- Now, you need to generate abbcde .
- Expand A-> Abc and Expand B-> d.
- Now, You have string like aAbcde and your input string is abbcde.
- Expand A->b. Final string, you will get abbcde.



- Hidden Markov Model (HMM): A statistical model that works with sequences, where the underlying system states are hidden and can only be observed through a set of visible outputs. Commonly used for tasks like speech recognition and part-of-speech tagging.
- Markov Chain: A sequence of events where the probability of each event depends only on the previous event. In language, it represents the likelihood of a word or state based on the previous word or state.
- Tokenization: The process of splitting text into individual units, like words or subwords, to prepare it for analysis. For example, breaking a sentence into words.
- Matrix: A mathematical structure used to organize and represent data. In language processing, a tokenization matrix could represent relationships between words or tokens.
- Probability Distribution: A function that describes the likelihood of different outcomes in a random experiment. In language, it indicates the chances of various words or events occurring.
- Neural Language Modeling: Using neural networks to build language models that predict the probability of a word given the previous words in a sequence. Helps in generating coherent text and speech recognition.
- Intuitive Formation: Creating models or concepts that are easily understandable and relatable, even to those without deep technical knowledge.
- Probability Estimation: Calculating the likelihood of an event happening based on available data. In language, it involves determining how likely a word or sequence is in a given context.
- POS Tagging: Part-of-speech tagging involves assigning grammatical categories (like noun, verb, adjective) to words in a sentence to understand their syntactic roles.

- Sample Markov Model for POS: A simplified version of a Markov model used for part-of-speech tagging. It predicts the next part of speech based on the current part of speech, aiding in language understanding.
- Initial Probabilities: In Markov models, it refers to the probability of starting from a specific state. For example, the chance of a sentence starting with a noun.