1. DEFINE N-GRAM AND AMBIGUITY

A:N gram is a sequence of n words and Ambiguity in NLP refers to the presence of multiple possible meanings or interpretations in a text, making it challenging for computers to accurately understand the intended meaning.

2. define tokenization, lexeme, morphemes
A: Tokenization is the process of breaking down text into individual units called tokens. These tokens can be words, phrases, or even characters, depending on the specific use case.
A lexeme is a fundamental unit of language that represents a word or a phrase with a specific meaning. It is the basic building block of a language's vocabulary.
Morphemes are the smallest units of language that carry meaning. They are the building blocks of words and sentences, and are used to convey grammatical, semantic, and phonological information.
3) define chunking
Chunking is a process in Natural Language Processing (NLP) that involves dividing text into smaller groups of words, called chunks, based on their syntactic or semantic relationships.
4) define treebank

Treebank refers to a dataset used in natural language processing (NLP) and computational linguistics. It's a collection of sentences or texts annotated with syntactic and semantic information, such as parse trees, part-of-speech tags, and named entity recognition.

5) define steming &lemization with example

Stemming involves removing the suffixes from a word to obtain its stem. This is done using algorithms like Porter Stemmer or Snowball Stemmer.

Lemmatization involves reducing a word to its lemma, which is the base or dictionary form of the word. This is done using dictionaries and morphological analysis.

6) Define default taggers & corpus

A default tagger is a part-of-speech (POS) tagger that is used as a fallback when a more specialized tagger is not available or is not trained on a specific language or dataset.

A corpus (plural: corpora) is a large collection of text data used for training, testing, and evaluating NLP models.

7) Define tagging & untagging, NER

Tagging and untagging refer to the process of adding or removing labels or annotations (tags) to or from text data, such as sentences or words.

NER stands for Named Entity Recognition. It's a technique in Natural Language Processing (NLP) that identifies and categorizes named entities in unstructured text into predefined categories such as:

1. Person (e.g., John Smith, Maria Rodriguez)

8) Define sentiment analysis

Sentiment Analysis is a type of Natural Language Processing (NLP) technique used to determine the emotional tone or attitude conveyed by a piece of text

9) Define parsing, topdown approach, bottom upapproach

Parsing is the process of analyzing a sentence or text to identify its grammatical structure and relationships between words.

Top-down parsing is often used in predictive parsing algorithms, which start with a predicted parse tree and refine it as they process the input sentence.

Bottom-up parsing is often used in shift-reduce parsing algorithms, which start with a buffer of input tokens and gradually construct the parse tree by shifting tokens and reducing them into constituents.

10) Define syntactic analysis, syntactic structure

Syntactic analysis, also known as parsing, is a process in natural language processing (NLP) that involves analyzing the structure of a sentence or text to identify its syntactic components

Syntactic structure refers to the organization of words and phrases in a sentence to convey meaning.

11) Define encoding & word segmentation

Encoding

- Refers to the process of converting text data into numerical representations (codes) that machines can understand.

Word Segmentation:

- Refers to the process of dividing text into individual words or tokens.

- Also known as tokenization or word tokenization.

12) what is propbank & framenet

PropBank:

- A corpus annotated with verbal propositions and their arguments [1].

- Focuses on verbs and their arguments, providing a predicate-argument structure [2].

FrameNet:

- lexical database that represents word meanings based on frames, which are abstract representations of events, objects, or concepts [1].

13) deep semantic parsing, seemantic analysis

Deep Semantic Parsing:

- A process that goes beyond syntactic parsing to identify the semantic relationships between entities, actions, and events in a sentence.

semantic Analysis:

- The process of analyzing text to identify its meaning and interpret its significance.

14) what is language model?

A language model is a type of artificial intelligence (AI) model that is trained to process and understand human language. Its primary goal is to learn the patterns and structures of a language, such as grammar, syntax, and semantics, in order to generate or predict language outputs.

15) what is coveragerate &perplexity in language model evaluation

Coverage rate, in the context of language models, refers to the percentage of tokens (words or characters) in a test set that the model is able to predict correctly. It's a measure of the model's ability to generate text that matches the original input.

Perplexity is a more comprehensive metric than convergence rate, as it evaluates the model's performance on unseen data

16) Define maximum likelihood estimation & smoothing with formulae?

MLE is a method for estimating the parameters of a statistical model given a dataset. In the context of language models, MLE aims to find the model parameters that maximize the likelihood of observing the training data.

Let's consider a language model with parameters θ and a training dataset $D = \{w_1, w_2, ..., w_n\}$:

MLE Formula:

$L(\theta; D) = \prod_{i=1}^{n} P(w_i \mid \theta)$

where $L(\theta; D)$ is the likelihood function, and $P(w_i \mid \theta)$ is the probability of observing the ith word given the model parameters θ.

Smoothing is a technique used to address the problem of zero probability estimates in language models. It assigns a small non-zero probability to unseen events (words or sequences) to avoid assigning zero probability to legitimate but unseen data.

17) Define SRL & base pharse chunks

SRL is a task in NLP that identifies the roles played by entities in a sentence, such as "agent", "patient", "theme", "goal", etc. It aims to extract the semantic relationships between entities and the actions they perform or undergo.

Base phrase chunks, also known as basic phrases or chunks, are the basic units of meaning in a sentence. They are typically short phrases that contain a single content word (noun, verb, adjective, or adverb) and its dependents (modifiers, complements, or adjuncts).

18) Define pas, salaam, wsd

PAS is a representation of the semantic structure of a sentence, focusing on the predicates (actions or events) and their arguments (entities involved in the action). It identifies the relationships between predicates and arguments

Sense Assignment Leveraging Alignment and Multilinguality (SALAM) is a technique in Natural Language Processing (NLP) that aims to improve word sense disambiguation (WSD) by leveraging alignment and multilinguality.

WSD is the task of determining the correct meaning (sense) of a word in a given context.

19) Define semantic interpretation & semantic parsing

Semantic interpretation is the process of assigning meaning to a sentence or text, going beyond the literal interpretation of individual words

Semantic parsing is a specific technique used in natural language processing (NLP) to perform semantic interpretation. It involves analyzing a sentence or text to identify its semantic structure, including the relationships between entities, actions, and events.

20) define reduce step and shift step

Reduce Step:

- **Definition**: In shift-reduce parsing, the reduce step involves applying a production rule of the grammar to the symbols currently on the top of the stack.

Shift Step:

- **Definition**: In shift-reduce parsing, which is a common technique used in syntactic and semantic parsing, the shift step involves shifting the next input symbol (word or token) onto the top of the stack.

21.define cyk algorithm and cfg

The Cocke-Younger-Kasami (CYK) algorithm is a dynamic programming approach used to parse sentences based on a context-free grammar (CFG), efficiently determining whether a given sentence can be derived from the grammar by building a parse table of potential syntactic structures.

CFG stands for Context-Free Grammar. It is a formal grammar type used in formal language theory and natural language processing to describe the syntax or structure of languages.

22. Define structure of document

the structure of a document refers to the organization and arrangement of its content, including text, images, and other elements. It involves the way the content is divided, formatted, and linked together to convey meaning and facilitate understanding.

23. Define sentence boundary detection, topic boundary detection

SBD is the task of identifying the boundaries between sentences in a given text.

TBD is the task of identifying the boundaries between topics or subtopics in a given text. It involves detecting the points where the topic or subtopic changes, and segmenting the text into coherent sections.

24.what is typology, polynomy, homonymy, antonymy

Typology is the study of the classification and categorization of languages, words, or concepts based on their shared characteristics, features, or properties.

polysemy refers to the phenomenon where a single word or phrase has multiple related meanings or senses.

Homonymy occurs when two or more words have the same form (spelling and/or pronunciation) but different meanings and, often, different origins.

Antonymy is the relationship between words that have opposite meanings or connotations.

Example: "hot" vs. "cold", "light" vs. "dark", "happy" vs. "sad"

25. what is hyponymy , synonyms, homophone, metonymy, meronymy

Hyponymy:

Hyponymy is a semantic relationship where a word (hyponym) is a specific instance or subtype of a more general word (hypernym).

Example: "rose" is a hyponym of "flower"

Synonyms:

Synonyms are words with similar meanings or connotations.

Example: "happy" and "joyful"

Homophones:

Homophones are words that are pronounced the same but have different meanings and often different spellings.

Example: "to" and "two" and "too"

Metonymy:

Metonymy is a figure of speech where a word or phrase is replaced by another closely related to it, often to avoid repetition or for rhetorical effect.

Example: "The White House announced a new policy" (using "White House" to mean the administration or government)

Meronymy:

Meronymy is a semantic relationship where a word (meronym) is a part of a larger whole or a component of a more general term (holonym).

Example: "wheel" is a meronym of "car