Important Questions for Mid-I Exam

Natural Language Processing

Choose the Best Answer (MCQs)

1. **Which of the following best describes the study of the structure of words?**

    1. a) Syntax
    2. b) Morphology
    3. c) Semantics
    4. d) Pragmatics

    Answer: b) Morphology

2. **Which model is commonly used for morphological analysis?**

    1. a) Finite State Automaton
    2. b) Decision Tree
    3. c) Support Vector Machine
    4. d) Naive Bayes Classifier

    Answer: a) Finite State Automaton

3. **Which of the following is a challenge in morphological analysis?**

    1. a) Identifying the meaning of words
    2. b) Handling word ambiguities
    3. c) Generating sentences
    4. d) Detecting syntax errors

    Answer: b) Handling word ambiguities

4. **In document structure analysis, which of the following is often the first step?**

    1. a) Part-of-speech tagging
    2. b) Named Entity Recognition
    3. c) Tokenization
    4. d) Dependency parsing

    Answer: c) Tokenization

5. **Which of the following is true regarding Treebanks in syntactic analysis?**

    1. a) They are used for semantic analysis
    2. b) They contain only annotated sentences
    3. c) They provide a corpus annotated with syntactic structures
    4. d) They focus on morphological analysis only

    Answer: c) They provide a corpus annotated with syntactic structures

6. **Which of the following parsing algorithms is used to analyze natural language sentences?**

    1. a) K-means clustering
    2. b) Earley Parser
    3. c) Naive Bayes Classifier
    4. d) Random Forest

    Answer: b) Earley Parser

7. **What is the main advantage of a data-driven approach to syntactic analysis?**

    1. a) It requires minimal data
    2. b) It allows for rule-based interpretation
    3. c) It can be adapted to different languages
    4. d) It reduces parsing complexity
    5. **Answer:** c) It can be adapted to different languages

8. **Which representation is commonly used for syntactic structures in NLP?**

    1. a) Phonetic trees
    2. b) Constituency trees
    3. c) Word embeddings
    4. d) Vectors

    Answer: b) Constituency trees

---

9. **What is one of the main challenges in resolving ambiguity in parsing?**

    1. a) Identifying the correct meaning of words
    2. b) Determining the grammatical structure when multiple interpretations exist
    3. c) Handling multiple languages
    4. d) Understanding word order in different languages

Answer: b) Determining the grammatical structure when multiple interpretations exist

10. **What is a primary issue when applying parsing algorithms to multiple languages?**

    1. a) Ambiguity resolution
    2. b) Syntax tree construction
    3. c) Lack of data for training
    4. d) Language-specific syntactic differences

    Answer: d) Language-specific syntactic differences

11. **What is the primary goal of document structuring?**
    a) To add punctuation to a text
    b) To organize text into meaningful sections
    c) To replace words with synonyms
    d) To remove duplicate words
    → **(b) To organize text into meaningful sections**
12. Which of the following features is NOT typically used for **document structuring**?
    a) Paragraph boundaries
    b) Text length
    c) Font size and color
    d) Emotion detection
    → **(d) Emotion detection**
13. **Which technique is used for structuring web documents?**
    a) XPath
    b) Named Entity Recognition (NER)
    c) POS Tagging
    d) Word Embeddings
    → **(a) XPath**
14. **Which approach is commonly used in modern document structuring?**
    a) Machine Learning
    b) Rule-Based Systems
    c) Hybrid Approaches
    d) All of the above
    → **(d) All of the above**

    Which factor affects the complexity of document structuring approaches?
    a) Number of pages in the document
    b) Variability of document layout
    c) Use of structured data

d) All of the above

→ **(d) All of the above**

Choose the Best Answer (MCQs)

1. **Which of the following is a key element in morphological analysis?**

    a) Sentence segmentation
    b) Syntactic parsing
    c) Affixation
    d) Named entity recognition

    Answer: c) Affixation

2. **Which morphological model represents a system of transitions between states to analyze word structure?**

    a) Hidden Markov Models
    b) Finite State Transducers
    c) Decision Trees
    d) Naive Bayes Classifier

    Answer: b) Finite State Transducers

3. **In the context of word structure, a prefix is an example of a _____.**

    a) Root
    b) Suffix
    c) Morpheme
    d) Part of speech

    Answer: c) Morpheme

4. **Which of the following is a method used to find the structure of documents?**

    a) Lemmatization
    b) Document classification
    c) Dependency parsing
    d) Word segmentation

    Answer: b) Document classification

5. **Which of the following parsing techniques is based on a bottom-up approach to build syntactic structures?**

    1. a) Earley Parser
    2. b) CYK (Cocke-Younger-Kasami) Parser
    3. c) Shift-Reduce Parsing
    4. d) LL(1) Parsing

    Answer: b) CYK (Cocke-Younger-Kasami) Parser

6. **A treebank is crucial for syntactic analysis because it provides:**

    1. a) Annotated syntactic structures of sentences
    2. b) Vocabulary lists
    3. c) Word embeddings
    4. d) Syntactic rules

    Answer: a) Annotated syntactic structures of sentences

7. **Which representation of syntactic structure breaks sentences into hierarchical structures based on grammatical units like noun and verb phrases?**

    1. a) Dependency Parsing
    2. b) Constituency Parsing
    3. c) Syntactic Tree Representation
    4. d) Lexical-Functional Grammar

    Answer: b) Constituency Parsing

8. **Which of the following is not a type of syntactic structure used in parsing?**

    1. a) Phrase structure trees
    2. b) Constituency trees
    3. c) Dependency trees
    4. d) Neural trees
    5. **Answer:** d) Neural trees

---

9. **In resolving ambiguity in parsing, the most commonly used method is:**

1. a) Statistical models
2. b) Probabilistic context-free grammar (PCFG)
3. c) Rule-based models
4. d) Lexicalized parsing

Answer: b) Probabilistic context-free grammar (PCFG)

10. **Which of the following issues arises when working with multilingual parsing models?**

    1. a) Lack of word boundaries
    2. b) Syntactic differences across languages
    3. c) Word sense disambiguation
    4. d) Non-contextual sentence structures

    Answer: b) Syntactic differences across languages

11. Which term refers to **breaking down a sentence into its syntactic components**?
    a) Morphology
    b) Parsing
    c) Tokenization
    d) Named Entity Recognition
    Answer: b)
12. A **Shift-Reduce Parser** uses which operation?
    a) Expanding non-terminals
    b) Shifting input symbols and reducing rules
    c) Constructing embeddings
    d) Removing stopwords
    Answer: b)
13. The **Earley Parser** is best suited for which type of grammar?
    a) Regular grammar
    b) Context-Free Grammar (CFG)
    c) Dependency grammar
    d) Phonological grammar
    Answer: b)
14. In a **dependency tree**, each word is connected to:
    a) A phrase structure
    b) A head word
    c) A latent variable
    d) A semantic label
    Answer: b)
15. What is the **biggest challenge** in natural language parsing?
    a) Ambiguity in sentence structure

b) Lack of computers

c) Too many syntactic rules

d) High memory usage

Answer: a)

Fill in the Blanks

1. In morphology, the smallest unit of meaning is called a _____.
   Answer: Morpheme
2. _____ refers to the process of breaking a word into its base form, such as removing suffixes and prefixes.
   Answer: Lemmatization
3. The process of analyzing word structure in a language is called _____.
   Answer: Morphological analysis
4. One of the challenges in document analysis is dealing with _____, where words may have multiple meanings or forms.
   Answer: Ambiguity
5. A _____ is an annotated corpus used to train syntactic parsers by providing labeled syntactic structures.
   Answer: Treebank
6. In dependency parsing, the structure represents the relationships between individual _____ in a sentence.
   Answer: Words
7. _____ parsing is a top-down approach where a tree is incrementally built based on context-free grammar rules.
   Answer: LL(1)
8. The goal of _____ parsing is to generate a tree that reflects the grammatical structure of the sentence.
   Answer: Syntactic
9. The problem of having multiple interpretations of a sentence due to structural variations is known as _____.
   Answer: Ambiguity
10. The process of syntactically analyzing sentences across different languages is known as _____.
    Answer: Multilingual parsing

11. **_____ segmentation is used to separate a document into logical sections.**
    → (Text)
12. In document structuring, **_____ is a common feature for identifying headings.**
    → (Font size)
13. **_____ is a tree-based structure used to represent syntactic relationships in sentences.**
    → (Parse tree)
14. The **_____ model uses machine learning to analyze document structures.**
    → (Statistical)
15. **_____ analysis helps extract important entities like names, locations, and dates from documents.**
    → (Named Entity)

Fill in the Blanks

1. The study of the structure of words is called _____.
Answer: Morphology
2. The _____ model is often used to analyze the components of words in morphological analysis.
Answer: Finite State Automaton
3. A common challenge in document analysis is _____, which refers to the ambiguity of words having multiple forms.
Answer: Morphological ambiguity
4. In the context of document structure, _____ is the process of breaking text into smaller linguistic units like words or sentences.
Answer: Tokenization
5. A _____ is a large collection of annotated sentences used to train syntactic parsing algorithms.
Answer: Treebank
6. The representation of syntactic structures as hierarchical trees is known as _____ parsing.
Answer: Constituency
7. The _____ parsing algorithm is known for its ability to handle ambiguous sentences in natural language.
Answer: Earley
8. A major benefit of the data-driven approach in syntactic analysis is that it allows for _____ across languages.
Answer: Adaptation

9. In parsing, the problem of having multiple possible interpretations of a sentence is known as _____.

Answer: Ambiguity

10. The process of resolving syntactic ambiguity across different languages is known as _____.

Answer: Multilingual parsing

11. Ambiguity in parsing arises when a sentence has _____

Answer: multiple valid structures.

12. A _____considers both syntax and word meaning**.**

Answer: lexicalized parser

13. **Treebanks** help train _____by providing labeled sentence structures.

Answer: statistical parsers

14. The _____is known for handling left-recursive grammar efficiently**.**

Answer: Earley Parser

15. _____starts with the root symbol and expands downward**.**

Answer: Top-down parsing

Five-Mark Questions with Answers

1. **Explain the concept of morphology and its role in Natural Language Processing (NLP).**

Answer: Morphology is the study of the structure and formation of words in a language. It involves analyzing the components of words, such as roots, prefixes, and suffixes. In NLP, morphological analysis plays a critical role in tasks like stemming, lemmatization, and part-of-speech tagging. It helps in reducing words to their root forms, improving the consistency of word forms, and aiding in better understanding of word meanings, which is essential for tasks such as text classification, information retrieval, and machine translation.

2. **Discuss the common challenges faced in morphological analysis and how they are addressed.**

Answer: Morphological analysis faces several challenges, including:

1. **Ambiguity:** Words with multiple forms or meanings (e.g., "run" as a verb or noun).

2. **Irregularities:** Irregular verb forms or exceptions in conjugation (e.g., "went" instead of "goed").
3. **Language-specific complexities:** Agglutinative languages like Turkish or Finnish, where words are formed by adding multiple affixes. These challenges are addressed using techniques such as rule-based systems (for specific languages), machine learning models (to learn from large annotated corpora), and hybrid approaches (combining rule-based and statistical methods).

---

3. **Explain the role of treebanks in syntactic analysis and how they contribute to parsing algorithms.**

Answer: Treebanks are corpora that contain text annotated with syntactic structures, typically in the form of constituency or dependency trees. They are essential for training and evaluating parsing algorithms, as they provide ground truth for the syntactic structure of sentences. By using treebanks, parsers can learn how words relate to each other in different syntactic configurations, enabling them to generate syntactically correct structures for unseen text. Treebanks also help in comparing the performance of different parsing techniques and understanding the linguistic properties of different languages.

4. **Discuss the differences between constituency parsing and dependency parsing.**

Answer:

1. **Constituency parsing** breaks sentences into hierarchical tree structures where each node represents a grammatical unit (e.g., noun phrases, verb phrases). It emphasizes the hierarchical relationships between constituents in a sentence.
2. **Dependency parsing**, on the other hand, represents grammatical relationships by focusing on the dependencies between individual words. Each word in the sentence is linked to a head word, and this dependency structure is used to understand sentence syntax. Constituency parsing is often used for syntactic analysis in context-free grammar models, while dependency parsing is preferred for modeling real-world syntactic structures due to its direct representation of word relations.

5. **What are the main parsing algorithms used in syntactic analysis? Explain one of them.**

Answer: There are several parsing algorithms used in syntactic analysis:

1. **Earley Parser**: A dynamic programming parser that can handle ambiguous sentences and is capable of parsing all context-free grammars.
2. **CYK Parser**: A bottom-up parsing algorithm that uses a table to store subparse results and is efficient for context-free grammars.
3. **Shift-Reduce Parser**: A greedy, left-to-right parser used in dependency parsing. The **Earley Parser** is notable because it is highly versatile and can handle both ambiguous and non-context-free grammars. It works by incrementally building syntactic structures and can handle any context-free grammar, making it ideal for NLP tasks that involve highly ambiguous text.

---

6. **What are the challenges and solutions in resolving syntactic ambiguity in parsing?**

Answer: Syntactic ambiguity arises when a sentence can be interpreted in more than one way due to multiple possible syntactic structures. For example, the sentence "I saw the man with the telescope" can be parsed in multiple ways, where the prepositional phrase "with the telescope" could attach to "saw" or "man." The challenges in resolving this ambiguity include:

1. **Multiple interpretations:** Determining the correct grammatical structure among several possibilities.
2. **Limited context:** Ambiguities that arise from context-dependent structures.
3. **Language-specific issues:** Ambiguity might be inherent in certain languages.

Solutions to these challenges include:

1. **Probabilistic parsing models** that use statistical techniques to select the most likely parse based on training data.
2. **Contextual analysis**, which uses surrounding words to resolve ambiguity.
3. **Disambiguation algorithms** such as the **Earley Parser** or **Statistical Parsing** that take into account the likelihood of certain parses.

7. **Explain the significance of morphological analysis in natural language processing.**

Answer: Morphological analysis is critical in natural language processing (NLP) because it helps to break down words into their components, such as prefixes, roots, and suffixes. This process aids in understanding the meaning of words, resolving word ambiguities, and reducing words to their root form (lemmatization and stemming). Morphological analysis is used in tasks like machine translation, information retrieval, part-of-speech tagging, and word sense disambiguation. By analyzing word structure, NLP systems can improve the accuracy of semantic understanding and text processing tasks.

8. **Discuss the challenges of handling morphological ambiguity in NLP and provide solutions.**

Answer: Morphological ambiguity occurs when a word can take different forms based on tense, case, or number (e.g., "run" as a verb or noun). It poses challenges in NLP, especially when trying to assign the correct meaning or part-of-speech tag. Solutions to this problem include:

1. **Contextual analysis:** Using surrounding words to determine the correct meaning of an ambiguous word.
2. **Statistical models:** Probabilistic models, such as Hidden Markov Models (HMM) and Maximum Entropy Markov Models (MEMM), can be trained on large corpora to predict the most likely morphological interpretation.
3. **Rule-based systems:** Using predefined rules and dictionaries to resolve common ambiguities.

---

9. **Describe the role of treebanks in syntactic analysis and their contribution to NLP.**

Answer: Treebanks are essential in syntactic analysis because they provide annotated corpora of sentences, where each sentence is tagged with its syntactic structure, often in the form of constituency or dependency trees. They serve as a gold standard for training and evaluating parsing algorithms. By learning from treebanks, syntactic parsers can identify grammatical relationships between words and construct syntactic structures for unseen sentences. Treebanks also allow researchers to assess the accuracy of different parsers and contribute to improving syntactic parsing techniques.

10. **Compare and contrast constituency parsing and dependency parsing.**

Answer:

1. **Constituency Parsing**: In this approach, sentences are broken down into nested hierarchical structures called constituents (e.g., noun phrases, verb phrases). Constituency parsing emphasizes hierarchical relationships and is based on context-free grammar. It is ideal for capturing phrase structures in sentences.
2. **Dependency Parsing**: This approach represents sentences by focusing on the grammatical relationships between individual words, rather than hierarchical structures. Words are linked based on syntactic dependencies, with each word being dependent on a head word. Dependency parsing is more flexible for real-world syntax and is preferred for languages with free word order.

11.What are the advantages and disadvantages of different parsing algorithms, such as CYK and Earley Parsers?

Answer:

3. **CYK Parser**: The CYK (Cocke-Younger-Kasami) parser is efficient and works well with context-free grammars. It is a bottom-up parser that uses dynamic programming and is highly effective for parsing in polynomial time. However, it is limited to context-free grammars and can be slower for more complex grammars.
4. **Earley Parser**: The Earley parser is a more versatile, general-purpose parser that can handle ambiguous and non-context-free grammars. It is efficient for parsing sentences with multiple possible syntactic structures. However, its performance may degrade when parsing large or highly ambiguous sentences, as it requires more memory.

12.What are the challenges in resolving syntactic ambiguity, and how can probabilistic models help in ambiguity resolution?

1. **Answer:** Syntactic ambiguity arises when a sentence can be parsed in multiple ways due to structural differences. For example, sentences like "The old man the boats" can be parsed in multiple ways. Challenges include:
    1. **Multiple interpretations:** Parsing algorithms must choose the most probable interpretation based on context.
    2. **Disambiguation across contexts:** Ambiguities arise in part due to lexical choices or word order.

Probabilistic models such as Probabilistic Context-Free Grammars (PCFG) are helpful in resolving ambiguity by assigning probabilities to different parse trees. These models can be

trained on large annotated corpora to predict the most likely syntactic structure based on observed frequencies. The parser uses these probabilities to select the most probable parse, reducing ambiguity.