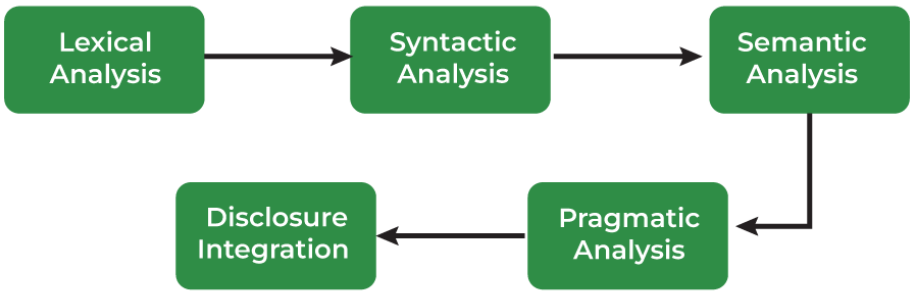
MACHINE LEARNING MODULE 1 NOTES 🗿 🤫 🤳

MODULE 1

Q.1) Explain Different stages involved in NLP process with suitable examples.



Natural Language Processing involves a systematic approach with multiple stages that transform raw text into meaningful computational representations. Each stage builds upon the previous one to achieve deeper understanding.

1. Lexical Analysis

Purpose: Breaking text into basic linguistic units and analyzing word-level properties.

Operations:

- **Tokenization:** Splitting text into words, sentences, and paragraphs
- **Lexical categorization:** Identifying word types and properties

Example:

- Input: "The students are studying NLP."
- Output: ["The", "students", "are", "studying", "NLP", "."]
- Word types identified: Article, Noun, Verb, Verb, Noun, Punctuation

2. Morphological Analysis

Purpose: Analyzing internal structure of words and their formation patterns.

Operations:

- **Stemming:** Reducing words to root forms
- **Lemmatization:** Converting to dictionary base forms
- **Morpheme identification:** Breaking words into meaningful units

Example:

- Input: "running", "better", "children"
- Stemming: "run", "better", "children"

- Lemmatization: "run", "good", "child"
- Morphemes: "runn+ing", "good+er", "child+ren"

3. Syntactic Analysis (Parsing)

Purpose: Analyzing grammatical structure and relationships between words.

Operations:

- **Part-of-speech tagging:** Assigning grammatical categories
- **Parsing:** Building syntactic trees
- **Grammar rule application:** Checking syntactic validity

Example:

- Input: "The cat sits on the mat"
- POS Tags: DT/NN/VBZ/IN/DT/NN
- Parse Tree: [S [NP The cat] [VP sits [PP on [NP the mat]]]]

4. Semantic Analysis

Purpose: Extracting meaning from syntactic structures and resolving ambiguities.

Operations:

- **Word sense disambiguation:** Choosing correct word meanings
- **Semantic role labeling:** Identifying who did what to whom
- **Meaning representation:** Creating semantic structures

Example:

- Input: "The bank is near the river"
- Word senses: bank (financial institution vs. river bank)
- Context resolution: "river" suggests geographical bank
- Semantic roles: Location relationship between bank and river

5. Discourse Processing

Purpose: Understanding text beyond sentence boundaries and maintaining coherence.

Operations:

- **Anaphora resolution:** Resolving pronoun references
- **Discourse segmentation:** Identifying topic boundaries
- **Coherence analysis:** Maintaining logical flow

Example:

- Input: "John went to the store. He bought milk. It was expensive."
- Anaphora: "He" → John, "It" → milk
- Discourse flow: Shopping sequence with price comment

6. Pragmatic Analysis

Purpose: Understanding context, speaker intentions, and implicit meanings.

Operations:

- **Speech act recognition:** Identifying communicative intentions

- **Context integration:** Using situational knowledge
- **Implicature detection:** Finding implied meanings

Example:

- Input: "Can you pass the salt?"
- Literal: Question about ability
- Pragmatic: Polite request to pass salt
- Speech act: Request/Command disguised as question

7. Application-Specific Processing

Purpose: Applying processed language understanding to specific tasks.

Examples by Application:

- **Machine Translation:** "Hello" → "Hola" (English to Spanish)
- **Question Answering:** Q: "Who invented the telephone?" A: "Alexander Graham Bell"
- **Sentiment Analysis:** "Great movie!" → Positive sentiment
- **Information Extraction:** "Apple Inc. was founded by Steve Jobs" → (Apple Inc., founder, Steve Jobs)

Complete Example Through All Stages

Input Text: "IBM's stock price increased yesterday."

Stage 1 - Lexical: ["IBM's", "stock", "price", "increased", "yesterday", "."]

Stage 2 - Morphological:

- "IBM's" → "IBM" + possessive "'s"
- "increased" → "increase" + past tense "-ed"

Stage 3 - Syntactic:

- POS: NNP/POS/NN/NN/VBD/RB/.
- Parse: [S [NP IBM's stock price] [VP increased yesterday]]

Stage 4 - Semantic:

- IBM's = company possessive
- stock price = financial value concept
- increased = upward change
- yesterday = temporal reference

Stage 5 - Discourse:

- Context: Financial/business domain
- Temporal anchoring to previous day

Stage 6 - Pragmatic:

- Informational statement about market performance
- Implied: Positive news for IBM shareholders

Conclusion

NLP stages work sequentially and interactively to transform raw text into actionable insights. Each stage contributes essential information while building toward comprehensive language understanding.

Q.2) Discuss the challenges and Ambiguity in various stages of natural language processing.

Natural Language Processing (NLP) aims to enable machines to understand and process human language. However, language is **highly complex, ambiguous, and context-dependent**, which makes NLP extremely challenging.
Challenges appear at **multiple levels**: phonological, morphological, lexical, syntactic, semantic, pragmatic, and discourse. A major source of difficulty is **ambiguity** — where a word, phrase, or sentence can be interpreted in more than one way.

1. Phonological / Speech Level

Challenges:

- Speech recognition errors due to accents, dialects, noise, and homophones.
- Handling continuous speech (no clear word boundaries).

Ambiguities:

- **Sound ambiguity**: words that sound the same but have different meanings.
- Example: *"I scream"* vs. *"ice cream"*.

2. Morphological Level

Challenges:

- Stemming and lemmatization for irregular word forms (*go* → *went*, *mice* → *mouse*).
- Rich morphology in some languages (Turkish, Finnish).
- Handling compound words, prefixes, suffixes, and infixes.

Ambiguities:

- **Word form ambiguity**: multiple valid interpretations.
- Example: *unlockable*
    - (a) not lockable,
    - (b) can be unlocked.
- Example: *flies* = noun (*fruit flies*) or verb (*he flies*).

3. Lexical Level

Challenges:

- **Tokenization issues**: contractions (*don't*), punctuation, compound words.

- **Out-of-vocabulary words**: slang, domain-specific jargon, misspellings.
- Named Entity Recognition (NER): distinguishing people, places, organizations.

Ambiguities:

- **Lexical ambiguity** (word with multiple meanings).
- Example: *bank* = river bank vs. financial institution.
- Example: *bat* = animal vs. cricket bat.

4. Syntactic Level

Challenges:

- Parsing long, complex, or ungrammatical sentences.
- Word order differences across languages (English SVO vs. Japanese SOV).
- Handling long-distance dependencies and nested structures.

Ambiguities:

- **Structural ambiguity** (multiple parse trees possible).
- Example: *I saw the man with a telescope*
    - (a) I used a telescope to see the man,
    - (b) the man had a telescope.

5. Semantic Level

Challenges:

- Word Sense Disambiguation (WSD): choosing the correct sense of a polysemous word.
- Compositional semantics: combining meanings of words into sentence meaning.
- Idioms and metaphors (*kick the bucket* ≠ literally kicking).

Ambiguities:

- **Sentence meaning ambiguity**.
- Example: *Flying planes can be dangerous*
    - (a) the act of flying planes is dangerous,
    - (b) planes that are flying are dangerous.

6. Pragmatic Level

Challenges:

- Understanding context, intention, and implied meaning.
- Recognizing sarcasm, irony, politeness.
- Cultural and situational variations in meaning.

Ambiguities:

- **Contextual ambiguity**: sentence meaning depends on speaker's intention.
- Example: *Can you pass the salt?*
    - (a) asking about ability,
    - (b) polite request.

7. Discourse Level

Challenges:

- Anaphora and coreference resolution: figuring out who/what a pronoun refers to.
- Maintaining coherence across long conversations or documents.
- Handling topic shifts and implicit references.

Ambiguities:

- **Referential ambiguity**.
- Example: *John met Peter. He was very tired*.
    - Who does *he* refer to? John or Peter?

Additional Challenges Across Levels

- **Data quality issues**: noisy text (social media), biased datasets.
- **Multilingual processing**: low-resource languages, code-switching.
- **Evaluation**: subjective nature of language quality assessment.

Conclusion

NLP challenges are deeply tied to **ambiguities at every level of language**. Successfully addressing these requires a combination of **rule-based methods, statistical approaches, and modern deep learning techniques**. Handling ambiguity remains central to making machines truly understand human language.

Q.3) Explain the preprocessing operations in natural language processing.

Preprocessing transforms raw text into structured format suitable for computational analysis, significantly impacting downstream NLP task performance.

1. Text Cleaning

Noise Removal:

- Eliminating HTML tags, XML markup, and formatting codes
- Removing unwanted punctuation and special characters

Encoding Standardization:

- Converting to consistent character encoding (UTF-8)
- Resolving encoding corruption and artifacts

2. Tokenization

Word Tokenization:

- Splitting text into individual words or tokens
- Handling contractions (don't → do n't)

Sentence Tokenization:

- Segmenting text into individual sentences
- Managing abbreviations (Dr., U.S.A.) containing periods

3. Text Normalization

Case Normalization:

- Converting text to lowercase for consistency
- Preserving case for proper nouns when needed

Spelling Correction:

- Identifying and correcting common misspellings
- Using edit distance algorithms for suggestions

4. Stop Word Removal

Common Word Filtering:

- Removing high-frequency, low-information words (the, and, is)
- Using language-specific stop word lists

Considerations:

- Preserving stop words for certain tasks (machine translation)
- Balancing information loss vs. noise reduction

5. Morphological Processing

Stemming:

- Reducing words to root forms using Porter Stemmer

- Applying language-specific stemming rules

Lemmatization:

- Converting words to canonical dictionary forms
- Handling irregular word forms with POS awareness

6. Part-of-Speech Tagging

Grammatical Classification:

- Assigning grammatical categories to tokens
- Using statistical and neural tagging models

Tag Sets:

- Choosing appropriate schemes (Penn Treebank, Universal POS)
- Fine-grained vs. coarse-grained tagging

7. Named Entity Recognition

Entity Identification:

- Recognizing persons, organizations, locations
- Identifying dates, monetary amounts, percentages

Entity Normalization:

- Standardizing entity representations
- Linking entities to knowledge bases

8. Advanced Operations

Phrase Chunking:

- Identifying noun phrases and verb phrases
- Shallow parsing for structural information

Language-Specific Processing:

- Script normalization for multilingual text
- Cultural adaptation for dates and numbers

9. Quality Assurance

Validation:

- Ensuring preprocessing consistency
- Detecting and handling edge cases

Performance Optimization:

- Efficient algorithms for large-scale processing
- Memory management techniques

Conclusion

Preprocessing operations form the foundation of NLP systems by cleaning and normalizing text. Proper preprocessing significantly improves model performance while reducing computational complexity.

Q.4) Describe Open Class Words and Closed Class Words in English with Examples

Introduction

English words are categorized into two major classes based on their membership flexibility and grammatical behavior: Open Class and Closed Class words.

Open Class Words

Definition: Open class words are lexical categories that readily accept new members. These classes can expand through borrowing, invention, and linguistic evolution.

Characteristics:

- Large and constantly growing membership
- Content words carrying semantic meaning
- Can be modified by inflectional and derivational morphemes
- Primary carriers of information in sentences

Types of Open Class Words:

1. Nouns

- **Function:** Name people, places, things, concepts
- **Examples:**
  - Common: book, computer, happiness, democracy
  - Proper: London, Microsoft, Shakespeare
  - New additions: smartphone, blog, cryptocurrency

2. Verbs

- **Function:** Express actions, states, or occurrences
- **Examples:**
  - Action: run, write, compute, google
  - State: exist, seem, belong
  - New additions: tweet, text, photoshop, uber

3. Adjectives

- **Function:** Describe or modify nouns
- **Examples:**
  - Descriptive: beautiful, large, intelligent
  - Comparative: better, faster, more efficient
  - New additions: viral, user-friendly, eco-friendly

4. Adverbs

- **Function:** Modify verbs, adjectives, or other adverbs
- **Examples:**
  - Manner: quickly, carefully, efficiently
  - Time: yesterday, soon, frequently
  - New additions: digitally, online, wirelessly

Closed Class Words

Definition: Closed class words are grammatical categories with fixed, limited membership that rarely accept new members.

Characteristics:

- Small, stable membership
- Function words providing grammatical structure
- Cannot be easily modified by morphological processes
- Essential for syntactic relationships

Types of Closed Class Words:

1. Determiners

- **Function:** Specify and limit nouns
- **Examples:**
  - Articles: the, a, an
  - Demonstratives: this, that, these, those
  - Quantifiers: some, many, few, all

2. Pronouns

- **Function:** Replace or refer to nouns
- **Examples:**
  - Personal: I, you, he, she, it, we, they
  - Possessive: my, your, his, her, our, their
  - Reflexive: myself, yourself, himself

3. Prepositions

- **Function:** Show relationships between words
- **Examples:**
  - Location: in, on, at, under, above
  - Time: before, after, during, since
  - Direction: to, from, toward, through

4. Conjunctions

- **Function:** Connect words, phrases, or clauses
- **Examples:**
  - Coordinating: and, but, or, so, yet
  - Subordinating: because, although, while, if
  - Correlative: either...or, both...and

5. Auxiliary Verbs

- **Function:** Help form tenses, moods, and voices
- **Examples:**
  - Primary: be, have, do
  - Modal: can, could, will, would, must, should

6. Particles

- **Function:** Modify meaning in phrasal constructions
- **Examples:**
  - Phrasal verb particles: up, down, out, off
  - "Look up the word", "Turn off the light"

Key Differences Summary

Open Class Characteristics:

- Unlimited membership growth
- Semantic content carriers
- Morphologically productive
- Examples: "google" (verb), "selfie" (noun), "viral" (adjective)

Closed Class Characteristics:

- Fixed membership
- Grammatical function providers
- Morphologically restricted
- Essential for sentence structure

Language Evolution Example

Technology-Driven Open Class Expansion:

- **Nouns:** internet, website, emoji, hashtag
- **Verbs:** download, upload, stream, hack
- **Adjectives:** digital, virtual, wireless, smart
- **Adverbs:** online, offline, remotely, digitally

Stable Closed Class: The closed class words remain largely unchanged: "the, and, in, with, for" have been stable for centuries.

Conclusion

The distinction between open and closed class words reflects the dynamic nature of language, where content words evolve rapidly while grammatical function words provide stable structural frameworks. This classification helps understand both language change and grammatical organization in English.

Q.5) Differentiate between Syntactic Ambiguity and Lexical Ambiguity.

Aspect	Syntactic Ambiguity	Lexical Ambiguity
1. Definition	Multiple possible grammatical structures for the same sentence	Multiple possible meanings for individual words or phrases
2. Level of Analysis	Sentence/phrase level structural interpretation	Word/lexical level meaning interpretation
3. Source of Ambiguity	Different ways to parse or structure the sentence grammatically	Polysemous words having multiple dictionary meanings
4. Resolution Method	Context, grammar rules, and structural preferences	Context, domain knowledge, and semantic relationships
5. Example 1	"I saw the man with the telescope" (who has the telescope?)	"Bank" - financial institution vs. river edge
6. Example 2	"Flying planes can be dangerous" (planes that fly vs. activity of flying)	"Bark" - dog sound vs. tree covering
7. Linguistic Component	Syntax and grammar rules	Semantics and word meanings
8. Parse Tree Impact	Results in multiple different parse trees	Same parse tree, different word interpretations
9. Frequency	Less common, usually in complex sentences	Very common, occurs with many everyday words
10. Processing Challenge	Requires syntactic parsing and structural analysis	Requires semantic analysis and context understanding