NLP – Unit 1 (Introduction to Natural Language Processing) – IN-SEM PYQ Answers

Q1. What is natural language processing (NLP)? Discuss various stages involved in NLP process with suitable example. [8]

Q2. Describe different stages of Natural Language Processing. [7]

Natural Language Processing (NLP) is a subfield of **Artificial Intelligence (AI)** and **Computational Linguistics** that deals with the interaction between computers and human (natural) languages.

It enables computers to **analyze, understand, interpret, and generate human language** in both text and speech form.

NLP combines concepts from:

1. **Linguistics** (morphology, syntax, semantics)
2. **Machine Learning**
3. **Computer Science**

Applications include:

- Machine Translation
- Sentiment Analysis
- Information Retrieval
- Question Answering

Stage	Definition	Key Tasks / Techniques	Example	Importance
1. Lexical Analysis	Process of identifying and processing words (lexemes) from raw text.	● Tokenization ● Part-of-Speech (POS) Tagging ● Normalization	"I am reading a book" → Tokens: [I, am, reading, a, book] POS: PRP, VBP, VBG, DT, NN	● Word identification ● Prepares input for higher-level processing
2. Morphological Analysis	Analysis of internal structure of words using morphemes (smallest meaning units).	● Stemming ● Lemmatization ● Root + affix identification	running → run + ing better → good	● Understands word formation ● Improves parsing & translation accuracy

3. Syntactic Analysis (Parsing)	Determines grammatical structure of sentence using formal grammar rules.	● Parse Tree construction ● CFG / PCFG ● Ambiguity resolution	Correct: "John eats an apple." Incorrect: "Apple eats John an."	● Ensures grammatical correctness ● Identifies subject–verb–object relations
4. Semantic Analysis	Extracts meaning of words and sentences in context.	● Named Entity Recognition (NER) ● Word Sense Disambiguation (WSD) ● Lexical Semantics	"Bank" → financial institution or river bank (context-based)	● Resolves ambiguity ● Ensures logical meaning
5. Discourse Integration	Establishes relationship between sentences in a text.	● Anaphora / Coreference Resolution ● Context linking	"Taylor went home. She was tired." → "She" = Taylor	● Maintains coherence across sentences
6. Pragmatic Analysis	Interprets intended meaning based on context and real-world knowledge.	● Intent detection ● Figurative meaning interpretation	"Can you pass the salt?" → Request (not ability)	● Understands speaker intention ● Handles idioms & indirect speech

Q3. Discuss the challenges of Natural Language Processing.[7]

1. Language Differences and Complexity: Natural languages are diverse, ambiguous, and structurally complex, making automatic understanding difficult.
    a. Presence of syntactic complexity (word order, tense, agreement, conjugation).
    b. Rich semantics and pragmatics; meaning depends on context.
    c. Continuous language evolution (lexical changes over time).
2. Training Data Limitations: NLP models require large, high-quality annotated datasets for effective learning.
    a. Collection and annotation of labeled data is time-consuming and expensive.
    b. Domain-specific tasks require specialized datasets.
3. Development Time and Resource Requirements: NLP systems demand significant computational and human resources.
    a. Complex tasks (e.g., machine translation, question answering) require longer development cycles.
    b. Training deep models requires high-performance hardware (GPU/TPU).

4. Phrasing and Structural Ambiguity: Sentences can be interpreted in multiple ways due to syntactic and semantic ambiguity.
    a. Requires syntactic parsing and semantic analysis for disambiguation.
    b. Contextual and pragmatic information is essential for correct interpretation.
5. Misspellings and Grammatical Errors: Noisy text input reduces system accuracy.
    a. Requires spell checking and text normalization techniques.
    b. Language models help predict correct word sequences based on context.
6. Bias in NLP Algorithms: Models may inherit societal biases present in training data.
    a. Requires bias detection and fairness evaluation.
    b. Techniques include balanced datasets and debiasing word embeddings.
7. Words with Multiple Meanings (Lexical Ambiguity): Polysemy and homonymy cause difficulty in determining correct word sense.
    a. Word Sense Disambiguation (WSD) is required.
    b. Use of semantic networks, embeddings, and contextual models improves accuracy.

Q4. Differentiate between programming languages and natural languages. [5]

Programming Languages	**Natural Languages**
Artificially designed formal languages used to communicate instructions to computers.	Naturally evolved languages used for communication among humans.
Strict syntax and well-defined grammatical rules.	Flexible grammar with many exceptions and variations.
Unambiguous interpretation; each statement has a single precise meaning.	Highly ambiguous; words and sentences may have multiple meanings.
Limited vocabulary and fixed keywords.	Large and continuously evolving vocabulary.
Errors (syntax/semantic) generally cause program failure.	Humans can understand meaning even with grammatical mistakes.

Q5. Explain in detail text processing in Natural Language Processing.[8]

1. **Text Normalization:** Converts text into a standard format.
    - Lowercasing (e.g., “NLP” → “nlp”)
    - Removing punctuation and special characters
    - Expanding contractions (e.g., “can’t” → “cannot”)
    - Handling numbers and dates

    Importance: Reduces variability and noise in text.

2. **Tokenization:** Process of splitting text into smaller units called tokens (words/sentences).
    Example:

"I love NLP." → [I, love, NLP]
Types:
- Sentence Tokenization
- Word Tokenization

Importance: Forms the basic units for further linguistic analysis.

3. **Stop Word Removal:** Removes frequently occurring words that carry little semantic meaning.
   Examples: "is", "the", "and", "in"

   Importance: Reduces dimensionality and improves computational efficiency.

4. **Stemming:** Reduces words to their root form by removing suffixes.
   Example:
   "playing", "played" → "play"
   Algorithm Example: Porter Stemmer
   Limitation: May not produce valid dictionary words.
5. **Lemmatization:** Converts words to their base (dictionary) form using vocabulary and morphological analysis.
   Example:
   "better" → "good"
   "running" → "run"
   Advantage: Produces meaningful base forms.
6. **Part-of-Speech (POS) Tagging**
   Assigns grammatical category to each token.
   Example:
   "Ram plays cricket"
   Ram (NNP), plays (VBZ), cricket (NN)

   Importance: Helps in syntactic parsing and semantic analysis.

7. **Handling Noise and Spelling Correction**
   - Spell checking
   - Removing HTML tags
   - Correcting informal text

   Importance: Improves model accuracy.

Q6. Differentiate between stemming and lemmatization.[4]

Stemming	Lemmatization
Process of reducing a word to its root form by removing prefixes/suffixes using heuristic rules.	Process of reducing a word to its base or dictionary form (lemma) using morphological analysis and vocabulary.
Does not consider context or part-of-speech.	Considers context and part-of-speech (POS) information.

May produce non-dictionary words.	Produces valid dictionary words.
Faster and computationally less expensive.	Slower and computationally more expensive.
Example: “playing”, “played” → “play”; “studies” → “studi”.	Example: “playing” → “play”; “better” → “good”; “studies” → “study”.

Q7. What do you mean by part-of-speech Tagging ?What is the need of this Task in NLP. [5]

Part-of-Speech Tagging is the process of assigning a grammatical category (tag) to each word in a sentence based on its definition and context.

A POS tag represents the syntactic role of a word such as:

- Noun (NN)
- Verb (VB)
- Adjective (JJ)
- Adverb (RB)
- Pronoun (PRP)
- Determiner (DT), etc.

Example:
Sentence: “Ram is playing cricket.”
Ram/NNP is/VBZ playing/VBG cricket/NN

POS tagging is performed using:

1. Rule-based taggers
2. Statistical models (Hidden Markov Model – HMM)
3. Machine Learning / Probabilistic models

Since many words are ambiguous (e.g., “book” → noun or verb), POS tagging uses contextual information to assign the correct grammatical category.

Need of POS Tagging in NLP

1. **Syntactic Parsing:** Helps in constructing parse trees and understanding sentence structure.
2. **Word Sense Disambiguation (WSD):** Reduces lexical ambiguity by identifying correct grammatical role.
3. **Improves Downstream NLP Tasks:** Essential for machine translation, information retrieval, sentiment analysis, and named entity recognition.

Thus, POS tagging acts as a foundational step in syntactic and semantic analysis in NLP.

Q8. Explain Tokenization with its different types. [5]

Tokenization is the process of breaking a stream of text into smaller meaningful units called **tokens**.
Tokens may be words, sentences, characters, or subwords.
It is the first step in text preprocessing in NLP.

Example:
Sentence: “I love Natural Language Processing.”
Tokens → [I, love, Natural, Language, Processing]

Types of Tokenization

1. **Sentence Tokenization:** Splits a paragraph into individual sentences.
   - Uses punctuation marks such as (.), (?), (!) as delimiters.
   - Example:
     “NLP is interesting. It is challenging.”
     → [“NLP is interesting.”, “It is challenging.”]
2. **Word Tokenization:** Splits a sentence into individual words.
   - Most commonly used form of tokenization.
   - Example:
     “I am learning NLP.”
     → [I, am, learning, NLP]
3. **Character Tokenization:** Splits text into individual characters.
   - Example:
     “NLP” → [N, L, P]
     Used in character-level language models.
4. **Subword Tokenization:** Splits rare or complex words into smaller meaningful units.
   - Used in modern language models.
   - Example:
     “unhappiness” → [un, happy, ness]
     Methods: Byte Pair Encoding (BPE), WordPiece.

Q9. Once a day the weather is observed as one of state 1 : rainy, state 2 : cloudy, state 3 : sunny.

$$A = \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$$

Each row sums to 1

Given that the weather on day 1(t=1) is sunny (state 3). What is the probability that the weather for the next 7 days will be “sun-sun-rain-rain sun-cloudy-sun”? [7]

Given:
States:
1 → Rainy
2 → Cloudy
3 → Sunny

Transition Probability Matrix (P):

$$A=\begin{pmatrix}0.4 & 0.3 & 0.3\\ 0.2 & 0.6 & 0.2\\ 0.1 & 0.1 & 0.8\end{pmatrix}$$

Initial condition:
Day 1 (t = 1) = Sunny (State 3)

Required sequence for next 7 days (t = 2 to t = 8):
Sun → Sun → Rain → Rain → Sun → Cloudy → Sun

That corresponds to state sequence:

3 → 3 → 3 → 1 → 1 → 3 → 2 → 3

Step-wise Transition Probabilities

From matrix P:

- P(3→3) = 0.8
- P(3→3) = 0.8
- P(3→1) = 0.1
- P(1→1) = 0.4
- P(1→3) = 0.3
- P(3→2) = 0.1
- P(2→3) = 0.2

Total Probability (Markov Property)

Since this is a first-order Markov Chain:

P=0.8×0.8×0.1×0.4×0.3×0.1×0.2
P=0.0001536

Final Answer:
P=1.536×10 ^−4

Therefore, the probability that the weather follows the given sequence for the next 7 days is:

0.0001536

★ PLEASE VERIFY THE NUMERICALS ★