Natural Language Processing - Mid-Semester Notes & Important Questions

Unit 1: Introduction to NLP & Basic Concepts

1. Origins and Challenges of NLP

- Human language is complex, ambiguous, and context-sensitive.
- Challenges: Large vocabulary, ambiguity, varied accents, syntax variations, and context understanding.

2. Language Modelling

- Grammar-Based LM: Uses syntactic rules to generate valid sentences.
- Statistical LM: Uses probability to predict next word (N-gram based).

3. Regular Expressions & Finite-State Automata

- Regular Expressions: Pattern matching for text processing.
  - Meta characters: [], ., ^, $, *, +.
  - Python: re.match(), re.findall(), re.sub().
- Finite-State Automata (FSA): Recognizes regular languages using states & transitions.

4. English Morphology

- Study of word structure (inflectional & derivational morphology).
- Use of transducers for mapping word forms to base forms.

5. Tokenization

- Breaking text into tokens (words/sentences).

6. Spelling Error Detection & Correction

- Using dictionary lookup and edit distance.

7. Minimum Edit Distance

- Measures string similarity by calculating insertions, deletions, substitutions.

Unit 2: N-Grams, POS Tagging, and HMM

1. Unsmoothed N-Grams

- Predicts next word using N-1 previous words.
- Types: Unigram, Bigram, Trigram.
- Limitation: Zero probability for unseen sequences.

2. Evaluating N-Grams
- MLE (Maximum Likelihood Estimation): Uses frequency counts.
- Perplexity: Measures model performance (lower is better).

3. Smoothing Techniques
- Laplace, Good-Turing, Kneser-Ney.
- Interpolation: Weighted average of N-grams.
- Backoff: Use lower-order N-grams when higher counts are zero.

4. POS Tagging
- Assigning grammatical categories.
- Methods: Rule-based, Stochastic (HMM, MEMM), Transformation-based.
- Issues: Ambiguity & unknown words.

5. Hidden Markov Model (HMM)
- States = POS tags, Observations = Words.
- Parameters: Transition & Emission probabilities.
- Algorithms: Viterbi (decoding), Forward-Backward (training).

Unit 3: Grammar & Parsing

1. Context-Free Grammars (CFG)
- Components: Non-Terminals, Terminals, Production Rules, Start Symbol.
- Used to model sentence structure.

2. Grammar Rules for English
- Declarative, Imperative, Yes-No, WH-Questions.

3. Treebanks

- Annotated corpora with parse trees (e.g., Penn Treebank).

4. Normal Forms for Grammar

- Chomsky Normal Form (CNF): Binary branching (A -> BC or A -> a).

5. Dependency Grammar

- Focus on binary head-dependent relationships between words.

6. Parsing Techniques

- Top-Down, Bottom-Up, Dynamic Programming (CYK).
- Shallow Parsing: Identifying chunks.

7. Ambiguity in Parsing

- Multiple parse trees possible for the same sentence.

8. Probabilistic CFG (PCFG)

- CFG + probabilities on rules to resolve ambiguity.

9. Probabilistic CYK Parsing & Lexicalized CFGs

- Efficient parsing using dynamic programming with probabilities.

10. Feature Structures & Unification

- Attribute-value pairs to handle agreement & grammatical constraints.

Unit 4: Knowledge Representation & Logic (Selected Topics)

1. Requirements for Representation

- Must be logical, unambiguous, expressive, and context-independent.

2. First-Order Logic (FOL)

- Components: Constants, Variables, Predicates, Quantifiers (Universal FORALL, Existential EXISTS).

- Express relationships, properties, and facts.

3. Description Logics

- Subset of FOL used in ontologies.

- Components: Classes, Roles, Individuals.

- Useful for representing structured knowledge (e.g., Semantic Web).

Important Mid-Sem Questions (10 Marks Each)

1. Explain the major challenges in processing natural languages. Illustrate with examples.
2. Define N-Gram Language Models. Explain unsmoothed N-Gram and smoothing techniques with examples.
3. Describe the working of Hidden Markov Models (HMM) and explain its application in POS tagging.
4. Write short notes on:
   a) Regular Expressions and their use in NLP.
   b) Finite-State Automata.
5. What is Context-Free Grammar (CFG)? Write the components of CFG and provide an example.
6. Explain First-Order Logic with an example. How is it used in NLP?
7. Describe the concept of Description Logics and their role in knowledge representation.