International Institute of Information Technology, Hyderabad

CL3.101 Computational Linguistics

End Semester Examination- 1st May 2025

Maximum Time: 3 Hours Maximum Marks: 60

Instructions:

- This exam paper consists of two sections: **Section A** and **Section B**. Answer both sections.
- **Section A** carries 30 marks. It contains 8 questions. Attempt **any 6 questions.** Each question has two sub-parts; you must answer both sub-parts to receive full credit.
  Note: Attempting more than six questions will lead to deduction of marks.
- **Section B** carries 30 marks. It contains 5 data annotation questions. Answer all.
- Wherever required, use **linguistic glosses** to support your explanations.

Section A: Choose and respond to **six (6)** questions.

Each question carries a value of 5 marks and comprises sub-questions.

[5 × 6 = 30]

Q1. a) Explain how morphological analysis influences POS tagging accuracy, especially in morphologically rich languages. Why is it important to distinguish between inflectional and derivational morphology in this context? *[2.5 marks]*

b) Discuss the limitations of POS tagging in handling multi-word expressions (MWEs) and idiomatic phrases with examples. *[2.5 marks]*

Q2. a) Explain how chunking contributes to NER with examples. Can NER be considered a subtask of chunking? Justify your answer by discussing both overlaps and distinctions between the two tasks. *[2.5 marks]*

b) Discuss the relationship between MWEs and NER. Are all named entities MWEs? Are all MWEs named entities? Justify with theoretical reasoning. *[2.5 marks]*

Q3. a) Define *complements* and *adjuncts* with reference to X-bar theory. How do they differ in terms of syntactic structure and semantic contribution? *[2.5 marks]*

b) Using a sentence from an Indian language, identify a complement and an adjunct. Explain how these distinctions affect parsing strategies in constituency vs. dependency frameworks. ***[2.5 marks]***

Q4. **a) Compare constituency parsing and dependency parsing in terms of their representation of syntactic structure. Which is more suitable for morphologically rich and free word order languages, and why? [*2.5 marks]***

b) Explain how morphological information (e.g., case markers, verb agreement) can assist in dependency parsing. Use examples from any Indian language. [*2.5 marks]*

Q5. a) Explain the conceptual foundations of Paninian grammar and how karaka theory is applied in parsing. *[2 mark]*

b) Take an example sentence from any Indian language and show both Paninian and UD annotations. Comment on their differences in handling semantic roles and word order. *[3 marks]*

Q6. **a) What are the major challenges in automatic speech recognition (ASR) for Indian languages? Discuss how linguistic diversity (e.g., phonetics, morphology, prosody) contributes to these challenges. *[2.5 marks]***

b) **Explain the fundamental differences between ASR and Text-to-Speech Synthesis (TTS) in terms of input, output, and core processing stages. [*2.5 marks]***

Q7. a) **Many informal texts contain abbreviations and ellipses (e.g., "U.S.", "Dr.", "etc.", "Wait... what?!"). These make sentence splitting with regular expressions difficult. Design a regex-based strategy to split sentences while avoiding splits inside abbreviations and ellipses. Explain how your regex handles such exceptions. *[2.5 marks]***

b) **Write a single regular expression that matches the following in social media text: *[2.5 marks]***

- Mentions (e.g., @username)
- Hashtags (e.g., #Mood)
- Emojis from the Unicode range U+1F600 to U+1F64F (**emoticons**, e.g., 😊, 😂), which may appear consecutively (e.g., 😂😂😂)

Q8. **a)** Both Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) are used for sequence labeling tasks. What is the key difference in how they model probability, and why does it matter? *[2.5 marks]*

b) Can HMMs and CRFs handle the same kinds of linguistic dependencies? Discuss with an example why one might perform better than the other in tasks like POS tagging or NER. *[2.5 marks]*

Section B: Data Annotation

A set of sentences are given below. Annotate them as asked in Q. 9–13. Each question carries 6 marks. Answer all. [5 × 6 = 30]

Sentences:

(a) *The curious boy from Hyderabad who always asked questions finally got selected for the science fair.*

(b) *The young athlete with remarkable stamina won three medals at the national meet.*

(c) *During the Republic Day parade in Delhi, Air Chief Marshal Rakesh Kumar saluted the President with pride.*

(d) *After weeks of back and forth, the startup's pitch deck wowed the angel investors at the Bengaluru conclave.*

Questions:

Q9. Perform morphological analysis, POS tagging, and chunking for sentences (a) and (b).

Q10. Identify Multi-Word Expressions (MWEs) in sentences (c) and (d), and annotate the type.

Q11. **Identify and annotate Named Entities (person, organization etc.) in sentences (c) and (d).**

Q12. **Draw X-bar schema for sentences (a) and (b), and clearly mark complements vs. adjuncts.**

Q13. **Draw a Dependency Tree with syntactic relations (either Panini based/ UD based) for sentences (a) and (b).**