International Institute of Information Technology, Hyderabad

Subject: **CL3.101 Computational Linguistics**

Mid Semester Examination

Max. Time: 1 ½ Hours **Max. Marks: 20**

Instructions:

- This exam has **two sections: Section A** and **Section B.**
- **Section A: 12 marks** – Analytical Questions.
  - Answer any **FOUR** out of six questions.
- **Section B: 8 marks** – Data Annotation & Practical Analysis.
  - **No choices** – Answer the given question.

Section-A

There are six questions. Answer any FOUR questions. **[4 * 3 = 12 marks]**

1. Given the text below, identify and explain three major tokenization issues. Then, propose tokenization strategies to handle these cases efficiently in NLP preprocessing. (3 marks)

Text:

```
Dr. A.P.J. Abdul Kalam, India's 11th President, once said,
"Dream is not that which you see while sleeping, it is something
that does not let you sleep." At 10:45 a.m., he was at a
conference on 'Education and Dream' in New Delhi—wasn't it
significant?
```

2. (A) Write a regex to match dates in the format YYYY-MM-DD, ensuring: (1.5 marks)

Matches: 2024-06-15, 1999-12-31

Does NOT match: 2024/06/15, 99-01-01, 2024-13-32 (invalid month/day).

(B) Given the following English **gerunds (-ing forms)**:

```
computing, programming, developing, running, hopping, making,
writing, singing, driving, hoping
```

Write a regular expression to extract their base forms (lemmas): (1.5 marks)

Example: `computing → compute, running → run`

- Do **not** use whole word match and substitution.
- Use grouping within the regex to capture the root form.

3. A) Compare the Item-and-Arrangement (IA), Item-and-Process (IP), and Word-and-Paradigm (WP) models in explaining the morphological inflection of the verb *go* . (2 marks)

B) Which model best explains the suppletive nature of *went*? (1 mark)

4. Construct a **Finite State Transducer (FST)** for the following irregular verb forms and their morphological derivations: (3 marks)

- *run → runs, running, ran, runner*
- *swim → swimming, swam, swims, swimmer*
- *write → writes, written, writing, wrote, writer*

5. Explain any TWO of the following concepts/challenges in the context of POS tagging with example. (2 * 1.5 marks)

A) Viterbi algorithm in HMM

B) Label Bias in MEMM

C) Training CRFs is more computationally demanding than HMMs and MEMMs.

6. Given the **input text, gold standard entities,** and **model-predicted entities,** compute the **Precision, Recall, and F1-score** for the NER model. (3 marks)

Input Text: *"Apple Inc. was founded by Steve Jobs in Cupertino, California. In 2023, it launched a new AI-powered assistant to compete with Google's Bard."*

Gold Standard Entities (Reference Output):
["Apple Inc.", "Steve Jobs", "Cupertino", "California", "Google", "Bard"]

Model-Predicted Entities:
["Apple", "Steve", "California", "Google", "AI-powered assistant"]

Section-B **Data Annotation** **(1 * 8 = 8 marks)**

7. *Tokenize* the provided text and identify *Lemma (rootword), parts of speech (POS)*, and *chunk* the text.

- Provide the annotation in tab-separated/table format.
- Use BIO format for chunking.
- Example annotation for the sentence `I saw the children.` is given here:

Token No	Token	Lemma	POS	Chunk
1	I	I	PRP	B-NP
2	saw	see	VM	B-VGF
3	the	the	DET	B-NP
4	children	child	NN	I-NP
5	.	.	PUNC	O

- Use BIS tagset for POS and Chunking.

BIS-POS tags: Common Noun (NN); Proper Noun (NNP); Noun of Space and Time (NST); Pronoun (PR); Personal (PRP); Reflexive (PRF); Relative (PRL); Reciprocal (PRC); Wh-word (PRQ); Demonstrative (DM); Main Verb(VM); Infinitive (VINF); Gerund (VNG); Auxiliary (VAUX); Adjective (JJ); Adverb (RB); Postposition (PSP); Conjunction (CC); Coordinator (CCD); Subordinator (CCS); Particles (RP); Classifier (CL); Interjection (INJ); Intensifier (INTF); Negation (NEG); Quantifiers (QT); Residuals (RD); Symbol (SYM); Punctuation (PUNC); Unknown (UNK)

BIS-Chunk tags: NP, VGF, VGINF, VGNN, VGNF, JJP, ADP, NEGP, CCP, FRAGP, BLK

Text:

Deep learning has transformed modern Natural Language Processing. With pre-trained models like GPT, BERT and T5, AI systems can understand and generate human-like text. However, challenges remain in reasoning, bias mitigation, and real-world adaptability. The robustness of these models depends on the diversity and quality of training data. Addressing ethical concerns in AI-driven language models requires interdisciplinary collaboration.

/end