Ref No.: Ex/PG/CSE/T/129B/2025(S)

MCSE 1st Year 2nd Semester Supplementary Examination, 2025

Natural Language Processing

Time – 3 hours **Full Marks - 100**

Answer any five questions *5*20=100*

1. *(5+10+5)=20*
   a. Define the following terms with suitable examples: `surface form`, `lemma`, `morpheme`, `stem` and `affix`. [5]
   b. Find out the edit distance and alignment between the two strings "`processing`" and "`impression`", considering the following costs for the edit operations.
   `insertion cost = deletion cost = substitution cost = 1` [10]
   c. Discuss the Smith-Waterman algorithm for best local alignment between two strings. [5]

2. *(6+4+4+6)=20*
   a. Derive the trigram language model using maximum likelihood estimation, chain rule, Markov assumption and add-1 smoothing. [6]
   b. If `P(0)=0.19` and `P(any other digit)=0.09`, compute the perplexity of your 10 digit mobile number according to the unigram language model. [4]
   c. Discuss Laplace smoothing in the context of language modelling. Discuss how the reconstituted counts (c^*) are calculated in a trigram model using Laplace smoothing. [4]
   d. Discuss some efficiency related solutions to deal with web-scale language models. [6]

3. *(8+2+5+5)=20*
   a. Discuss the Noisy channel model for non-word spelling correction. [8]
   b. What are real-word errors? How candidates are generated for real word errors? [2]
   c. Define homonym, homograph, homophone, hyponym, and hypernym with suitable examples. [5]
   d. Discuss the properties of hyponymy. Differentiate between hyponym and instance relations. [5]

4. *(6+6+8)=20*

a. Write a shell script to normalize case, tokenize and show the tokens ending with "ing" that could potentially be verbs in a corpus in decreasing order of frequency. Explain your answer. *[6]*

b. Write the Viterbi algorithm for finding the optimal sequence of tags for an observation sequence, given the model. *[6]*

c. What is continuation probability of a word? How it is computed? Discuss the formulation of the Kneser-Ney smoothing for an *n*-gram language model. *[8]*

5. *(5+10+2+3)=20*

a. Discuss the Resnik method and Lin method of measuring semantic similarity between two words in terms of information content. *[5]*

b. Given the following term-context matrix, compute which of the following word pairs – [fish, potato] and [digital, information], is more similar according to distributional similarity using add-2 smoothing. *[10]*

term \ context	computer	boil	data	result	fry
fish	0	2	0	0	2
potato	0	2	0	0	1
digital	2	0	3	2	0
information	1	0	6	4	0

c. What is a term-context matrix and how it is computed? *[2]*

d. Define Positive Pointwise Mutual Information (PPMI). What does it measure? *[3]*

6. *(4+10+6)=20*

a. Briefly discuss about the BLEU MT evaluation metric. Mention its pros and cons. *[4]*

b. Compute the alignment probabilities and the translation probabilities according to the EM algorithm assuming no NULL token and only 1-to-1 alignments for the following toy parallel training corpus. Show the first 3 iterations. *[10]*

Translation pair id	Source Language	Target Language
1	green house	casa verde
2	the house	la casa

c. Write and explain the Forward algorithm for computing the probability of an observation sequence, given the model. *[6]*

7. Write short notes on any four of the following: *(4*5)=20*
   a. Noisy channel model for SMT.
   b. Good-Turing smoothing.
   c. Comparison of Levenshtein Edit Distance algorithm and Needleman-Wunsch algorithm.
   d. Perplexity of language model.
   e. HMM model for POS Tagging.
   f. Comparison of different MT evaluation metrics.
   g. Beam search with stack decoding.
   h. Phrase-based SMT.
   i. Precision, Recall, F-Measure, MAP, MRR.