**Ref No.: Ex/CSE/T/423B/2019**

## B.C.S.E. 4th Year 2nd Semester Examination, 2019

## Natural Language Processing

**Time – 3 hours** **Full Marks - 100**

**Answer any five questions**

1.

a. Compare Needleman-Wunsch algorithm and the Levenshtein Edit Distance algorithm. *3*

b. 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. *5*

c. Find out the edit distance and alignment between the two strings "*imposter*" and "*protest*" considering an equal cost (say, 1) for all the edit operations. *10*

d. What are the best-case and worst-case time complexities of the Backtrace algorithm? Mention the cases where they occur. *2*

2.

a. Derive the bigram language model using maximum likelihood estimation, chain rule and Markov assumption. *5*

b. What is an interpolated language model? Explain with an interpolated trigram model. *3*

c. Discuss the Good-Turing smoothing technique. *4*

d. What is continuation probability of a word? How it is computed? *3*

e. What is the simplification assumption that is often made to reduce the search space in real word spelling correction and how much it is able to reduce the search space? *2*

f. Describe how the four confusion matrices are used in the *channel model* in the context of spelling correction. *3*

3.

a. Discuss the Resnik's information content based method for measuring similarity between two words. How Lin similarity is different from Resnik similarity. *4+2*

b. Discuss the Viterbi decoding algorithm. *5*

c. What is the fundamental difference between Markov Chain and HMM? *2*

d. Discuss how the POS tagging problem can be modelled using HMM. Mention the simplification assumptions. *5+2*

4.

a. What is a term-context matrix and how it is used to measure word similarity? *4*

b. Compare thesaurus based semantic similarity with distributional semantic similarity. *2*

c. "Pruning out partial hypotheses is risky". Explain this. Discuss how pruning decisions can be improved using future cost estimates. *4*

[ Turn over

d. Compute the alignment probabilities and the translation probabilities obtained after the first 2 iterations of the EM algorithm assuming no NULL token and only 1-to-1 alignments for the following parallel training corpus. *8*

| Translation pair id | Source Language | Target Language |
|---|---|---|
| 1 | red house | rouge maison |
| 2 | the house | la maison |

e. Discuss the TER MT evaluation metric. *2*

5.

a. State Log-Likely hood Ratio(LLR), the unsupervised content selection techniques for text summarization. Why unsupervised content selection methods are good for summarization? The following are three reference summaries along with a system generated summary. What are the scores of ROUGE-3evaluation scheme? *5+2+6*

- Human 1: We are the great citizen who can devote for the country.
- Human 2: You are the Indian citizen who can identify the proper value for the country.
- Human 3: We are really proud to be the great Indian citizen who can devote for the nation.
- *System answer: We are the great Indian citizen who can sacrifice their lives for the country.*

b. What are the differences between Natural Language Generation and Natural Language Understanding? What do you know about Cosine-Similarity? *3+4*

6.

a. What is relevance feedback query? State and explain Rocchio SMART algorithm for calculating a relevance feedback query using VSM. What is PMI? *2+6+2*

b. Consider the following two tables which show the results of two classes, A and B. What are the Macro-average and Micro-average Precision values? Is Micro-averaged score dominated by score on common classes? *8+2*

| Class A | Truth: yes | Truth: no |
|---|---|---|
| Classifier: yes | 40 | 40 |
| Classifier: no | 20 | 80 |

| Class B | Truth: yes | Truth: no |
|---|---|---|
| Classifier: yes | 40 | 40 |
| Classifier: no | 40 | 120 |

7.

a. State Naïve Bayes algorithm for text classification. 7

b. Define Kappa measure and state its use with an example. 7

c. Write down the basic architecture of a modern factoid based Question-Answering (QA) system. 6