Ref No.: Ex/PG/CSE/T/129B/2018

M.E. Computer Science & Engineering, 1st Year, 2nd Semester Examination, 2018

Natural Language Processing

Time – 3 hours **Full Marks - 100**

Answer any five questions

1. a. Write a shell script to normalize case, tokenize, sort and show the tokens of a corpus in decreasing order of frequency. Explain your answer. *6*
   b. Find out the edit distance and alignment between the two strings "improper" and "procedure", considering an equal cost for all the edit operations. *8*
   c. Discuss the time and space complexity of the Levenshtein Edit Distance algorithm and the Backtrace. What are the best-case and worst-case time complexities of the Backtrace algorithm? *2+2*
   d. Why you might require weighted edit distance? *2*

2. a. Add-1 smoothing is a non MLE estimator. Explain this. *2*
   b. Discuss how interpolation can be used in Language models. Explain with an interpolated trigram model. *3*
   c. How a language model can be evaluated extrinsically and intrinsically? Define and deduce perplexity. *2+3*
   d. Discuss the Good-Turing smoothing technique. *5*
   e. How Kneser-Ney smoothing improves over Good-Turing smoothing? *5*

3. a. Discuss the noisy channel model for non-word spelling correction. *7*
   b. How candidates are generated for real word spelling errors? *2*
   c. What is the simplification assumption that is often made to reduce the search space in dealing with real word spelling errors and how much it is able to reduce the search space? *1+2*
   d. State and explain Zipf's law. *3*
   e. Discuss the Viterbi algorithm. *5*

4. a. What is a Markov chain? Explain with a suitable example. State the difference(s) between Markov chain and Hidden Markov Model. *2+1*
   b. Briefly discuss the HMM model for part of speech tagging. Extend the model to trigrams. *6+1*
   c. Discuss the TER MT evaluation metric. How does it improve over WER? *3*
   d. Define hyponym, hypernym, meronym and holonym with suitable examples. *4*
   e. What is a term-context matrix and how it is used to measure word similarity? *3*

5. a. Discuss path-based similarity. What are the disadvantages of path-based similarity? *2+1*
   b. Define Positive Pointwise Mutual Information (PPMI). What does it measure? *2+1*
   c. Discuss the Resnik method and Lin method of measuring semantic similarity. *3+3*
   d. Given the following term-context matrix, compute the PMI based distributional word similarity between each term-context word pair using add-2 smoothing. *8*

term \ context	computer	boil	data	result	fry
eggplant	2	3	2	2	3
potato	2	3	2	2	4
digital	4	2	3	3	2
information	3	2	8	6	2

6. a. What are the similarity and differences between translation memory and EBMT? *2*
   b. Why is SMT modelled as a noisy channel model? What do the two models in SMT take care of? *2+1*
   c. 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 parallel training corpus. Show at least 3 iterations or until the models converge. *10*

Source Language:	red house	the house
Target Language:	rouge maison	la maison

   d. Discuss how future cost estimation helps prevent pruning out good hypotheses early. *5*

7. Write short notes on any four of the following: *5*4*
   a) Vauquois Pyramid.
   b) Beam search stack decoding.
   c) CKY parsing.
   d) Expectation Maximization algorithm.
   e) BLEU and METEOR MT evaluation metrics.
