M.E. Computer Science & Engineering, 1st Year, 2nd Semester Examination, 2019

Natural Language Processing

Time – 3 hours **Full Marks - 100**

Answer any five questions

1. a. Find out the edit distance and alignment between the two strings "procedure" and "computer", considering an equal cost (say, 1) for all the edit operations. *8*
   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. *6*
   c. Why the maximum matching word segmentation algorithm works very well for languages like Chinese, but not for English? *2*
   d. Define the following terms - type, token, vocabulary, lemma, morpheme, stem. *4*

2. a. Discuss the Smith-Waterman algorithm for best local alignment between two strings. *5*
   b. Derive the trigram language model using maximum likelihood estimation, chain rule and Markov assumption. *5*
   c. Define and deduce perplexity. Discuss the notion of perplexity as a branching factor. *3+2*
   d. Discuss how to deal with web-scale language models? What smoothing technique is used for web-scale language models? *3+2*

3. a. Discuss how the POS tagging problem can be modelled using HMM. Mention the simplification assumptions. *5+2*
   b. Discuss the three basic problems for HMM. *3*
   c. What are real-word errors? How real word errors can be detected and corrected? *1+4*
   d. Define the 4 confusion matrices in the context of spelling correction. *3*
   e. What is continuation probability of a word? How it is computed? *2*

4. a. How you can integrate the layout of the keyboard into the spelling correction model? *2*
   b. Define homonym, homograph, homophone. Give examples of homographs that are not homophones, and homophones that are not homographs. *2+2*
   c. Define hyponym and hypernym. Discuss the properties of hyponymy. Differentiate between hyponym and instance relations. *2+2+2*
   d. Differentiate between word similarity and word relatedness. *2*
   e. Discuss the Resnik's information content based method for measuring similarity between two words. Discuss how Lin similarity improves over Resnik similarity. *6*

5. a. Compare thesaurus based semantic similarity with distributional semantic similarity. *3*
   b. What is a term-context matrix and how it is computed? *2*

c. Define Pointwise Mutual Information (PMI). What does it measure? 2

d. Prove: PMI(x,y)=log(P(x|y)/P(x))=log(P(y|x)/P(y)) 3

e. Discuss how the probability of a concept can be measured. 2

f. 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	digital	pinch	result	sugar
Data	2	2	0	1	0
Information	1	6	0	4	0
Lemon	0	0	1	0	1
Orange	0	0	1	0	2

6. a. Derive and briefly discuss the noisy channel model of statistical machine translation (SMT). 3

b. Discuss how hypothesis recombination can be used to reduce the search space in SMT decoding. 2

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

Translation pair id	Source Language	Target Language
1	red house	casa roja
2	the house	la casa

d. Discuss about the BLEU and METEOR MT evaluation metrics. 5

7. Write short notes on any four of the following: 5*4

a) Kneser-Ney smoothing
b) Noisy channel model for spelling correction
c) Forward algorithm in HMM.
d) Future Cose Estimation in SMT decoding.
e) Viterbi algorithm.
f) Expectation Maximization algorithm.