Subject Code: Ex/CSE/PE/B/T/414C/2024(S)

B.C.S.E. 4th Year 1st Semester Supplementary Examination - 2024

Natural Language Processing

Time: 3 Hours

Full Marks: 100

CO1: Mathematical Models

Answer any 3 questions *3*10=30*

1. *10*

Find out the edit distance and alignment between the two strings "*calculator*" and "*computer*", considering the following costs for the edit operations.

```
insertion cost = deletion cost = 1
substitution cost = (your exam roll number % 2) + 1
```

2. *10*

Given the following term-context matrix, compute which of the following word pairs - [data, information] and [lemon, orange], is more similar according to distributional similarity using add-*n* smoothing, where *n* = (your exam roll number % 2)+1.

context / term	computer	digital	pinch	sugar	program
data	2	2	0	0	1
information	1	6	0	0	4
lemon	0	0	1	1	0
orange	0	0	1	2	0

3. *(7+3)=10*

a. Compute the score assigned to the following query-document pair by the tf-idf model using the lnc.ltc weighing scheme. Assume that the document frequencies of the terms "digital", "best", "DSLR", "camera", "lense" and "zoom" are 5,000, 50,000, 10,000, 20,000, 25,000 and 40,000 respectively, and the document collection size is n * 1,000,000, where *n* = (your exam roll number % 2)+1.

Document: *camera DSLR camera digital camera lense zoom*
Query: *best DSLR camera*

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.

[Turn over

4. *10*

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.

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

CO2: Algorithms

Answer any 3 questions *3*10=30*

1. *(5+3+2)=10*

a. Write and explain the Damerau-Levenshtein algorithm.
b. Compare Needleman-Wunsch algorithm and the Levenshtein Edit Distance algorithm.
c. What are the best-case and worst-case time complexities of the Backtrace algorithm? Mention the cases where they occur.

2. *(6+4)=10*

a. Derive the trigram language model using maximum likelihood estimation, chain rule, Markov assumption and add-1 smoothing.
b. Briefly discuss about the BLEU MT evaluation metric and its performance issues.

3. *(6+4)=10*

a. Discuss the Forward algorithm in HMM.
b. Discuss the Resnik method and Lin method of measuring semantic similarity.

4. *(8+2)=10*

a. Discuss and explain the noisy channel model for non-word spelling correction.
b. Real word spelling correction is more difficult than non-word spelling correction. Why?

CO3: Linguistics

1. Answer any 5 questions: *2*5=10*

a. Define homonym, homograph and homophone with examples.
b. Differentiate between word similarity and word relatedness.
c. Why is SMT modelled as a noisy channel model? What do the two models in SMT take care of?
d. What is case normalization? Describe some situations where case normalization is undesirable.
e. Differentiate between lemma, stem and affix with suitable examples.
f. Explain hyponym and hypernym with examples.
g. Why precision and recall are not suitable metrics for evaluation of MT?

CO4: Generalization and Analysis

Answer any 3 questions *3*10=30*

1. *(2+4+4)=10*
   a. Write a regular expression to identify all valid instances of the word ‘the’.
   b. Mention some features that you could use in a machine learning framework for deciding whether a mail is spam or not.
   c. Define and deduce perplexity. Discuss the notion of perplexity as a branching factor.

2. *(8+2)=10*
   a. Discuss the HMM model for POS tagging.
   b. What are the simplification assumptions in HMM?

3. *(7+3)=10*
   a. Explain the inverted index data structure. Why it is called ‘inverted’ index? How queries are processed with an inverted index.
   b. What are the main disadvantages of Boolean information retrieval?

4. *(4+2+4)=10*
   a. Define and explain precision, recall and F-measure. Discuss how you can compute class-specific precision and recall when dealing with multiple classes?
   b. Naïve Bayes classifier has an important similarity to language modeling. Explain this.
   c. Discuss some positive and negative aspects of the Naïve Bayes Classifier with regard to performance issues.