Reg. No.									

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent unit of MAHE, Manipal)

VII SEMESTER B.TECH. (COMPUTER SCIENCE AND ENGINEERING)

END SEM EXAMINATIONS, NOV 2019

SUBJECT: NATURAL LANGUAGE PROCESSING [CSE 4011]

REVISED CREDIT SYSTEM
26-NOV-2019

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ALL** questions.
- Missing data may be suitable assumed.

1A Distinguish between Natural Language and Artificial Language. **3**

1B Draw an FST to show consonant doubling during inflection involving ***–ing***/ ***-ed***. Illustrate the working of FST using positive and negative examples. **4**

1C Write the algorithm for finding Minimum Edit Distance using the dynamic programming technique. **3**

2A B is a corpus which only contains one single bitstring: **4**
1 1 0 1 1 1 0 0 1 0 1 1 1 0 1 1 1 1 0 0 0
i. Calculate the following bigram probabilities from the corpus B using MLE (Maximum Likelihood Estimation):
P(0 | 1)
P(0 | 0)

ii. For each of the following bitstrings using Bigram model, which is the most probable value for x. Explain.
1 0 1 0 1 0 1 x
0 1 0 1 0 1 0 x

2B 2 B. Take the following series of tokens as a training corpus. **4**
B A C D A D C E A C G G F G A C
i. Identify the hapax legomena in the corpus

ii. Give the maximum likelihood estimate of a unigram language model trained on the corpus
iii. Give the Good-Turing estimate of a unigram language model trained on the corpus.

2C Describe the disadvantages of Laplace (add-one) Smoothing. How is it overcome with Good-Turing Discounting? **2**

3A What is the result of applying the following sequence of Brill transformation rules to the corpus below: **2**
1: change VBN to VBD if previous tag is NP
2: change VBD to VBN if next tag is BY

Chapman/NP killed/VBN John/NP Lennon/NP
John/NP Lennon/NP was/BEDZ shot/VBD by/BY Chapman/NP
He/PPS witnessed/VBD Lennon/NP killed/VBN by/BY Chapman/NP

3B With examples, write short notes on each of the word classes: **3**
i. Prepositions ii. Particles iii. Determiners

3C Following table gives unsmoothed bigram count for 4 words. Assume that the corpus consists of 128 word types occurring among 2500 sentences. The unigram count for the words are: natural = 312, language = 409, processing = 252, session = 458. Generate unsmoothed bigram probability matrix, Laplace smoothed bigram probability matrix and Laplace smoothed bigram count giving expressions for all the terms. **5**

	natural	language	processing	session
natural	27	86	79	42
language	30	16	81	52
processing	63	31	9	11
session	18	15	19	74

4A Given a sequence of observations, how do you derive an expression for the best sequence of tags that corresponds to the sequence of observations? Derive all the terms with necessary justifications. **3**

4B Why Markov chain cannot represent part of speech tagging? With formal definition, explain how Hidden Markov Model can accomplish the problem of part of speech tagging. **3**

4C With a suitable lexicon and context free grammar, draw parse tree for a sentence having possessive expressions and a sentence having complementizer. **4**

5A How do you create a set of equivalent binary CFG rules from a given grammar having four non terminals on the right hand side of a grammar rule? **3**

5B Explain various question answering techniques bringing its features from a simpler form to a larger scope. **3**

5C With a neat diagram, explain Vauquois triangle emphasizing on different machine translation approaches. **4**

****