Total No. of printed pages = 4

CSE 1818 PE 52

20/6/23

Roll No. of candidate □□□□□□□□□□

2023

B. Tech 8th Semester End-Term Examination

SPEECH AND NATURAL LANGUAGE PROCESSING

New Regulation (w.e.f. 2017-18) & New Syllabus (w.e.f.2018-19)

Full Marks – 70 Time – Three hours

The figures in the margin indicate full marks for the questions.

1. Answer the following questions: (10 × 1 = 10)

(i) Syntactical analysis is done at ______ level?

(a) Sentence (b) Word

(c) Lexicon (d) Symbol

(ii) Which of the following is used to mapping sentence plan into sentence structure?

(a) Text planning (b) Sentence planning

(c) Text Realization (d) Text Summarization

(iii) ———— morphology is a type of word formation that creates new lexemes:

(a) Derivational morphology (b) Compound morphology

(c) Inflectional morphology (d) Complex morphology .

(iv) Which is a finite state machine with two tapes: an input tape and an output tape

(a) Finite State Transducers (FSTs)

(b) Finite State Translators (FSTs)

(c) Finite Automata

(d) Deterministic Finite Automaton

[Turn over

(v) "I bought a printer today. I had bought one earlier in 2004. This one cost me Rs.6000 whereas that one cost me Rs.2000". In this statement, "This" and "that" are known as which type of reference in the discourse context?

(a) Definite Reference
(b) Indefinite Reference
(c) Pronominal Reference
(d) Demonstrative Reference

(vi) Consider the statement "The students went to class". Assign POS tags for the statement.

(a) DT NN VB P NN
(b) DT NN NN P NN
(c) NN NN VBG P NN
(d) DT NN VB P DT

(vii) To whether "duck" is a verb or a noun can be solved by ——————

(a) Part-of-speech tagging
(b) Lexical analysis
(c) Semantic analysis
(d) Pragmatic analysis

(viii) N-grams are defined as the combination of N keywords together. How many bigrams can be generated from the given sentence: Gandhiji is the father of our nation?

(a) 7
(b) 6
(c) 8
(d) 9

(ix) Using pronouns to refer back entities already introduced in the text is called as —————— problem

(a) Anaphora
(b) Misspellings
(c) Multiple Meaning
(d) Lexical problem

(x) Which of the following signs are used to indicate repetition?

(a) #
(b) *
(c) –
(d) All of the mentioned

Answer *any four* of the following questions: (4 × 15 = 60)

2. (a) How do Viterbi Algorithm optimize HMM? Explain.

(b) How is Feature Extraction done in NLP? What are the various techniques for feature extraction?

(c) Given a corpus C2, the Maximum Likelihood Estimation (MLE) for the bi-gram "computational linguistics" is 0.25 and the count of occurrence of the word "computational" is 1200. If the vocabulary size is 4400, what is the likelihood of "computational linguistics" after applying add-1 smoothing? [5+5+5]

3. (a) What is stop-words in NLP? Should it always be removed during preprocessing? Explain.

(b) What is bag-of-words model? Explain with an example.

(c) Consider the following corpus C3 of four sentences.

<s> three friends amar akbar and anthony are reading book </s>

<s> amar is reading malgudi days </s>

<s> akbar is reading a detective book </s>

<s> anthony is reading a book by rk narayan </s>

Assume a bi-gram language model, calculate:

P(<s> amar is reading a book </s>). [3+5+7]

4. (a) What is Morphology? Explain the different types of morphology with examples.

(b) What is pragmatic analysis?

(c) Consider the CFG given below:

$S \to aSb \mid D$

$D \to Dc \mid \in$

How many non-terminals should be added to convert the CFG into CNF? [5+4+6]

5. (a) What do you mean by Parts of Speech (POS) tagging? Explain with an example.

(b) Briefly summarise Zipf's law. What does it tell you about the distribution of words in language? Why and how this can be taken into account in practical applications?

(c) What will happen if you do not convert all characters to a single case (either lower or upper) during the pre-processing step of an NLP algorithm? Give an example where converting to a single case, say to small letter, may lead to loss of information. [5+5+5]

6. (a) What is tokenization? Write a python code to tokenize a paragraph in terms of words and convert it to lower case. Is tokenization done at character level? Why or why not?

(b) Write a regular expression which will extract all the telephone number from a text. The telephone number are assumed to be in the three formats - ddd-ddd-dddd, ddd.ddd.dddd, ddd*ddd*dddd.

(c) A Hidden Markov Model (HMM) is given in the table below:

Transition probabilities	Emission probabilities
P(NOUN | PRON)=0.001	P(she | PRON)=0.1
P(PRON | START)=0.5	P(run | VERB)=0.01
P(VERB | AUX)=0.5	P(can | AUX)=0.2
P(AUX | PRON)=0.2	P(can | NOUN)=0.001
P(NOUN | AUX)=0.001	P(run | NOUN)=0.001
P(VERB | NOUN)=0.2	
P(NOUN | NOUN)=0.1	

Calculate the probability P(she | PRON can | AUX run | VERB). [7+2+6]

7. Write short notes on any *three* of the following:

(a) Information Extraction

(b) TF-IDF

(c) Stemming and Lemmatisation

(d) Machine Translation [5+5+5]