SRN | | | | | | | | | | | |

PES University, Bengaluru-85

(Established under Karnataka Act No. 16 of 2013)

UE17CS333

December 2020: IMPROVEMENT EXAMINATION, B.TECH, VI-SEMESTER

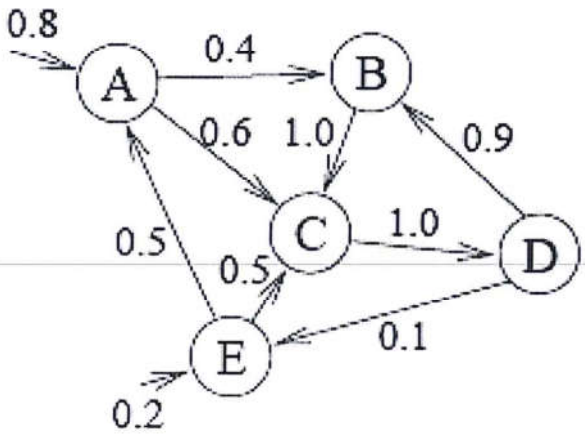
UE17CS333 – Natural Language Processing

Time: 03 Hours **Answer All Questions** **Max Marks: 100**

All the questions are compulsory
Draw the diagrams wherever necessary
Figures to the right indicates marks

1	a)	Using Dynamic Programming approach (show the memorization table), calculate the minimum edit distance when the source string is "VINTER" and target string is "WRITERS". Given: cost (insert) = cost(delete)= cost (substitute) = 1	5
	b)	You are probabilistically generating a word containing characters from vocabulary {p, q, r, s, t}. You always start by generating "p". Then you keep generating any number of characters till you generate "t". Once you generate "t", you stop. The probability of generating "t" is x. 1) Draw an FSA with start and stop states that accepts only the above words 2) Suppose the length of the generated string is n. Write an equation for P (n) signifying the probability of the generated sequence. For x=0.3 and n =4, what will be the value of P (n)?	3+3
	c)	Which are the different types of ambiguities in language processing? Give examples of any 2 ambiguities.	4
	d)	Design and draw an FSA to recognize simple date expressions like March 15, the 22nd of November, Christmas. Extend this date FSA to handle deictic expressions like Yesterday, tomorrow, a week from tomorrow, the day before yesterday, Sunday, next Monday, three weeks from Saturday.	5
2	a)	Given a set of states, S = {s1, s2, ... , sr}, The process starts in a random state based on a probability distribution and changes states in sequence based on a probability	4

distribution of the next state given the previous state.



What will be the probability of this model which could generate the sequence {A,C,D,B,C} of length 5.

b) Assuming our corpus has the following frequency count of different types of birds: [sparrow: 10; crow: 3; woodpecker: 2; goose: 1; swan: 1; pelican: 1]. Calculate the following unigram probabilities using Good Turing Estimates: 6

a) probability (hummingbird)

b) probability (goose)

c) We have come up with two unigram models with the unigram probabilities as shown in the table. 5

Unigram	Model 1	Model 2
Word 1	0.3	0.4
Word 2	0.4	0.5
Word 3	0.3	0.1

Find out the better model using perplexity.

d) What is smoothing? Why is it required? Explain Add-1 smoothing with an example. 5

3 a) Choose the correct alternative for the following: 2+2

1. Between the Earley and CYK parsers

(a) Earley parser works bottom up and CYK parser works top down

(b) CYK parser works bottom up and Earley Parser works top down

2. Recursion in context free grammar is when

(a) terminals on the left also appears on the right

(b) non terminal constituency on the left also appear on the right

b) The grammar and lexicon are provided here. Using Ealey parsing algorithm, find the corresponding Parse Tree for the sentence "Shatabdi reaches Delhi via Bhopal" 6

SRN | | | | | | | | | | |

		Note : Just providing the parse tree will not be given any marks and you need to show detailed workings in the chart S -> NP VP NP -> N NP -> NP PP VP -> V NP PP -> P NP N-> Shatabdi, Delhi, Bhopal P -> via V-> reaches	
	c)	You are given the grammar and lexicon. Use CKY algorithm to come up with **two parse trees** for the sentence " A pilot likes flying plane" **Grammar** | **Lexicon** S → NP V P | DT→ A VP →VBG NNS | NN →pilot VP →VBZ VP | VBZ →likes VP →VBZ NP | VBG→ flying NP→ DT NN | JJ →flying NP→ JJ NNS | NNS →plane	10
4	a)	What is a Synset in wordnet? What kind of relation exists between Lemma and Synset in wordnet? Explain with an example	6
	b)	Differentiate between: • Lemma and Wordform • Ontology and Taxonomy • Hyponym and Hypernym • Meronomy and Holonomy	4
	c)	What is Discourse in NLP? What is the main Discourse Issue that you need to resolve in NLP?	5
	d)	What are shortcomings of the Gloss Overlap based approach such as Lesk Algorithm?	5
5	a)	Pick and write the correct alternative for the following. 1. A negative Pointwise Mutual Information of a word W and a context C means : (a) W and C occur together less than they occur individually (b) W and C occur together more than they occur individually 2. Since Convolutional Neural Network (CNN) is very good in capturing patterns irrespective of where in the text it appears, you have decided to use it in sentiment analysis task for text reviews. You have a word embedding layer followed by	(2*5 =10)

SRN | | | | | | | | | | | |

		convolution layer, then a flattening layer and finally a dense layer with Softmax activation. You see that emoticons are being used by all the users in their reviews and carry significant information about sentiment. The more appropriate model here is (a) Character level CNN (b) Word Level CNN 3. Eliza and Parry are example of early generation chatbots. Both of them lack (a) Emotion (b) Knowledge (c) Rules 4. If we consider the three important factors i.e. (1) vanishing gradient problem (2) computational complexity (3) long term memory (a) LSTM is better than plain vanilla RNN in all three factors stated above (b) LSTM is better than plain vanilla RNN in factors (1) and (3) (c) LSTM is better than plain vanilla RNN in factors (2) and (3) 5. The key difference of the traditional "encoder-decoder" seq2seq model from the "attention based" seq2seq model is: (a) Variable input and output sequence (they need not be of the same length) (b) The design of the context vector between the encoder and decoder	
	b)	Which are the 7 tasks in Natural Language Generation (NLG) Also mention the three layers in NLG architecture with suitable diagram.	5
	c)	Define a Named Entity. Give example of (a) nested Named Entity (b) Named Entity ambiguity (c) an Entity that is not a Named Entity.	5