SRN □□□□□□□□□□□□

PES University, Bengaluru-85
(Established under Karnataka Act No. 16 of 2013)

UE15CS333
(faculty BJD)

May 2018: B.TECH, VI-SEMESTER
ESA
UE15CS333 - NATURAL LANGUAGE PROCESSING

Time: 3 Hrs | Answer All Questions | Max Marks: 100

1 **Each question carries two marks and has one correct answer. Write the correct answer.** 40

a) We have two equal sized datasets i.e. one from weather reports and other one from telephonic conversation between friends. The language in weather report is formal and less variant whereas the conversational language between friends widely varies. Trigram models M1 and M2 are trained on equal sized dataset from weather reports and telephonic conversation respectively. Both models are then validated (perplexity is the metric) on test data from respective domains. Which of the following you expect to be a better model?
a) M1 on weather data **b)** M2 on conversation data
02

b) HMM approach for POS tagging is probabilistic and ignores the clues lying in the texts about POS tags. A MAXENT Model for POS Tagging improves on this shortcoming. It is (1) which means that the model is specified by (2). Select the correct option:
(a) (1) Finds boundary between classes (2) P (X, Y) **(b)** (1) Generates class distributions (2) P (X, Y)
(c) (1) Generates class distributions (2) P (Y|X) **(d)** (1) Finds boundary between classes (2) P (Y|X)
02

c) CKY parsing take $O(n^3)$ time, where n is the length of the sentence because
(a) it needs to fill each of the $O(n^2)$ boxes in the diagonal CKY chart and filling each square requires examining O(n) ways of splitting the given phrase into two sub-constituents
(b) it needs to fill each of the O(n) boxes in the diagonal CKY chart and filling each square requires examining $O(n^2)$ ways of splitting the given phrase into two sub-constituents
02

d) There are 7 tasks in Natural Language Generation (NLG) and three layers in NLG architecture. The tasks are (1) sentence aggregation, (2) Orthographic realization, (3) content determination, (4) referring expression generation, (5) syntax & morphology, (6) discourse planning and (7) lexicalization. The architectural layers are (A) sentence planning, (B) linguistic realization and (C) text planning. Pick up the correct option below that maps the NLG tasks to NLG architectural layers.
(a) {(3,5)-> A}, { (4,6,7)->B} , {(1,2)->C} **(b)** {(3,7)-> A}, {(1,6)->C}, { (4,5,2)->B} **(c)** {(3,6)-> C}, {(1,7)->A}, { (4,5,2)->B}
02

e) Overall steps in Hobb's algorithm for Discourse Analysis
(a) Searches the current sentence from right to left, starting at noun. If no antecedent found, searches the previous sentence from left to right.
(b) Searches the current sentence from left to right, starting at pronoun. If no antecedent found, searches the previous sentence from right to left.
(c) Searches the current sentence from right to left, starting at pronoun. If no antecedent found, searches the previous sentence from left to right.
02

f) In the table below, the first column lists the approaches of several algorithms and the second column lists algorithm names. The algorithms belong to semantic relatedness and word sense disambiguation (WSD). Pick up the option that has all correct pairs.

1. Count number of common ngrams in the overlap (in words) in glosses of synsets	A. Resnik Similarity
2. Build a graph of senses using semantic similarity to calculate edges and then apply PageRank	B. Hyperlex
3. Similarity of two concepts is measured by measuring information content of lowest common subsumer	C. Lesk Algorithm
4. Detecting different uses of words amounts to isolating high density components in their co-occurrence graph.	D. Random Walk

(a) { (1,D), (2,A), (3,B), (4, C)} **(b)** { (1,C), (2,B), (3,A), (4, D)} **(c)** { (1,A), (2,B), (3,C), (4, D)} **(d)** { (1,C), (2,D), (3,A), (4, B)}
02

g) What problem does IOB Tagging solve in chunking (shallow parsing)? 02

(a) Over fitting (b) Semi supervised learning (c) Determining boundary of chunk (d) Supervised learning

h) The advantage of Yarowsky's algorithm over Lesk algorithm is : 02

(a) It does not depend on definitions from Lexical resources

(b) It does not require the use of unlabeled training data

(c) It can disambiguate between more than two possible word senses

(d) It can discover cause-effect relationship unlike Lesk Algorithm

i) Using nested lambda reduction, **λz. λy λx .giving(x,y,z) (money, Marie, John)** becomes 02

(a) giving (money, John, Marie) (b) giving (John, Marie, money) (c) giving (John, money, Marie)

j) An expression consisting only of a predicate with a variable among its arguments is interpreted as 02

(a) One entity (b) Property of an entity (c) Relation between entities (d) A set

k) The shortcoming of several methods (dealing with meaning of text) and the methods themselves are listed in two columns of the table below: 02

1. This method does not consider word meaning at all and mainly relies on the frequency of occurrence.	A. Word embedding
2. This method relies on word senses but not on words. They may not be the same.	B. Lexical semantics
3. A word and its antonym may be found to be very similar as this method relies on context of occurrence	C. Textual similarity based on TF IDF
4. Does not support entailment	D. First order logic enhanced by Lambda Calculus

Find out the correct option that matches both sides:

(a) { (1,A), (2,B), (3, C), (4,D)} (b) { (1,C), (2,B), (3, A), (4,D)} (c) { (1,C), (2,B), (3, A), (4,D)} (d) { (1,C), (2,B), (3, D), (4,A)}

l) In Lappin And Leass algorithm for discourse analysis, salience factors are considered to have scope over a sentence. So 02

(a) References to same entity over multiple sentences do not add up while multiple references within the same sentence add up.

(b) References to same entity over multiple sentences add up while multiple references within the same sentence do not.

m)
1. "pupil" (student vs. part of the eye)
2. "rent" (give on hire vs. actual amount paid for the hire)

Consider the two examples above and pick the right answer: 02

(a) 1 is an example of polysemy and 2 is an example of homonymy

(b) 1 is an example of homonymy and 2 is an example of polysemy

(c) Both 1 and 2 are examples of homonymy

(d) Both 1 and 2 are examples of polysemy

n) Given a sentence S="$w_1w_2w_3$ ....wn", to compute the likelihood of S using a bigram model, 02

(a) Calculate the conditional probability of each word in the sentence given the preceding word and add the resulting numbers

(b) Calculate the conditional probability of each word in the sentence given the preceding word and multiply the resulting numbers

(c) Calculate the conditional probability of each word given all preceding words in a sentence and add the resulting numbers

(d) Calculate the conditional probability of each word given all preceding words in a sentence and multiply the resulting numbers

o) You are testing a word sense disambiguation system that you have built recently. From a test set of 100 words, system attempts 75 words and correctly disambiguates 50 of them. Pick up the correct option below : 02

a) Both precision and recall are 0.50 b) Both precision and recall are 0.66 c) Precision is 0.66 and recall is 0.50

d) Precision is 0.50 and recall is 0.66

p) Three examples are given below : 02

1) Snow + man = snowman 2) Snow (noun) + less = snowless (Adj) 3) Snow(verb) + Past = Snowed(Verb)

Given the above, choose the correct options :

(a) 1 is "inflection", 2 is "compounding" and 3 is "derivation"

(b) 1 is "derivation", 2 is "inflection" and 3 is "compounding"

(c) 1 is "compounding", 2 is "derivation" and 3 is "inflection"

SRN

q) You are working with a text corpus by dividing it into train and test. The language model you have built is unsmoothed (no smoothing techniques applied). When you are evaluating the built language model on the test part of the corpus, you have encountered lot of unseen words. Your evaluation metric is perplexity. Choose the correct statement below:
(a) Perplexity will be infinite. (b) Perplexity will be zero. 02

r) You are training your deep learning model on text data. You have plotted the **training and validation loss vs. number of epochs.** You may have observed any of the following :

1. With more training (= number of epoch), validation loss is increasing though training loss has stabilized to a low value
2. With more training (= number of epoch), loss has stabilized for both validation and training; they are equal at a low value.
3. With more training (= number of epoch), both training and validation loss continues to decrease
4. With more training (= number of epoch), both validation and training loss stabilize but validation loss remains higher (with respect to training loss) by a constant value
5. With more training (= number of epoch), loss has stabilized for both validation and training; both remain equal and high

Pick up the correct answer:

(a) 1 is "over fit", 2 is "under fit", 3 is "good fit", 4 is "good fit", 5 is "over fit"
(b) 1 is "good fit", 2 is "over fit", 3 is "under fit",4 is "over fit", 5 is "under fit"
(c) 1 is "under fit", 2 is "good fit", 3 is "over fit", 4 is "over fit", 5 is "over fit"
(d) 1 is "over fit", 2 is "good fit", 3 is "under fit", 4 is "under fit" ", 5 is "under fit" 02

s)

1.Sequence to single element prediction	A. Text summarization
2.Sequence to class	B. Next word in the sentence
3.Sequence to sequence generation	C. Image to text generation
4.Sequence to sequence prediction	D. Anomaly detection

The above table has two columns representing various types of RNN/LSTM configuration and possible applications involving text data. The options below represent the pairs matching them. Pick up the correct option:
(a) { 1,B } , { 2, C} , {3,D } , {4, A} (b) {3, B} , { 1,D } , { 2, A} , {4, C}
(c) { 1, C} , { 2, B} , {3,A } , {4, D} (d) {4,A}, {2, D } , {1, B } , {3, C } 02

t) A training dataset has 100 attributes and each can take 10 values. It is found that only 5 of the attributes contribute 98% of the variance. Hence, 95 attributes are dropped to achieve dimensionality reduction. So, instance space size reduces by
(a) 95% (b) 98% (c) 10^{-95} (d) 10^{-98} 02

2 a) Computing minimum edit distances by hand, figure out whether "drive" is closer to "brief" or to "divers" and what the edit distance is. Use 1-insertion, 1-deletion, 2-substitution costs. 07

b) Design a deterministic finite state machine that accepts any PESU computing ID. Note that a PESU computing ID contains first the initials in 1-3 upper or lower letters, then 0-1 numeral and finally 0-2 lower letters. Indicate all possible end states in this FSA. 07

c) Calculate $P(w_1,w_2)$ and $P(w_2|w_3)$ when the following data is provided :

(1) vocabulary V = { w_1,w_2,w_3} (2) and the **bigram probability distribution p on V X V** specified by:

(a) $P(w_1,w_1) = 0.25$ (b) $P(w_2,w_2) = 0.0$ (c) $P(w_3,w_3) = 0.25$ (d) $P(w_2,w_1) = 0.125$ (e) $P(w_1,w_3) = 0.25$
(f) P (w_1, ?) = 0.5 (i.e. w1 as the first of a pair) (g) P (?,w_2) = 0.125 (i.e. w_2 as the second of a pair) 06

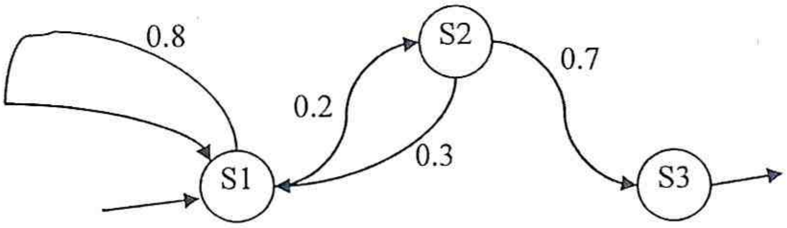
3 a) Assume a bigram language model is trained on the following corpus of sentences using MLE with linear interpolation for smoothing (with the bigram λ weight set to 0.9 and the unigram λ weight set to 0.1). Since the unigram model does not need to estimate P(<s>), just completely ignore the start token when estimating the unigram model.

(a) <s> man marries woman </s> *(b)* <s> woman marries man </s> *(c)* <s> woman marries woman </s>
(d) <s> man divorces woman </s> *(e)* <s> woman divorces man </s>

Find the estimated probability of the test string <s> man marries man </s> 06

SRN

b) Find **two paths** that could be taken through the following Hidden Markov model that produces the output **"C1 C2 C3 C4 C5"** and calculate the probability of each path being taken. 07



State S1: Output	**Probability**	**State S2: Output**	**Probability**	**State S3: Output**	**Probability**
C1	0.5	C2	0.8	C4	0.5
C2	0.3	C3	0.1	C5	0.5
C3	0.2	C4	0.1		

c) The grammar and lexicon are provided below: 07

Grammar	**Lexicon**
S →NP VP	Noun → Virat | ball | bat
PP → Prep NP	Prep → with | off | through
VP → Verb NP	Verb → drove | cut
VP → VP PP	
NP → NP PP	
NP → Noun	

Use Earley algorithm to parse the sentence "Virat drove ball with bat". You need to show all the entries in all the charts clearly displaying the states and the dotted rules.

4 a) We have a language model of vocabulary size 10000. In the training corpus we see donkey 10 times out of which it is followed by "clever" 5 times and "stupid" 5 times. What is the Laplace Estimate of "Probability(clever|donkey)"? 03

b) The sentence "Every man loves some woman" is ambiguous and is having two meanings where the woman may be different or same for each man. Express the two meanings in first order logic clearly distinguishing between them. 04

c) The car review in Carwale.com goes like this: "The Renault Duster has a muscular, chunky look but with flowing lines which gives it a solid yet racy look. In contrast the Mahindra TUV300 looks like a box with flat panels and harsh edges." You, a NLP consultant, are doing a **fine-grained sentiment analysis**. Identify a tuple containing the followings: 03

(a) Product **(b)** Feature or aspect **(c)** Sentiment Polarity

d) In Gandhi Medical hospital, the patients in a healthcare plan have to undergo detailed diagnostic tests (more than 80). A general physician prepares **Patient dossier** by documenting the interview of the patient and then appending the results of these tests. Every year, the patient undergoes the same procedure of tests and interview. The new test results and interview data get appended to the **patient dossier.** 10

The **text in patient dossier** does not have any structure and has mostly medical terms. At the time of making the initial patient dossier, the general physician has no confirmed knowledge about any future medical issues but over a number of years, the enrolled patients develop more than one type of medical issues and those issues then get labelled on the dossier. A particular disease can be mapped to several medical issues. There are 10000 medical issues and a handbook lists the medical issues that lead to each of the 100 diseases. One million patients are already enrolled. Neither the medical issues nor the diseases are mutually exclusive.

You are developing a deep learning system **to predict the future medical issues.** You have tried out both deep ANN and RNN/LSTM. You are converting the yearly text data for each patient into a word embedding. Your design has word embedding as input, then deep ANN or RNN/LSTM layers and finally densely connected layers to provide the predictions. Provide your reasons for the following 5 questions (**each carries 2 marks**) in **1-3 sentences** :

(a) You are debating between Google's pre-trained Word2Vec word embedding **vs.** word2vec word embedding trained with your own patient dossier data. Which one will you choose and why?

(b) You are debating between **"sum vs. average"** (from embedding of individual words) options of deriving the **patient embedding** for a year. You have chosen to use the average instead of sum. What information are you possibly losing out?

(c) For input, the deep ANN averages all the yearly patient embedding. For RNN, each yearly vector is part of a sequence of input. RNN is found to perform better. Why?

(d) You have non-linearity such as sigmoid, RELU, tanh, softmax etc. in your network. While doing trial and error, you have observed that magnitude-wise normalisation of the word embedding input has actually improved the training performance. Which non-linearity elements are probably positively affected and why?

(e) Training this model is computationally expensive and your runtime analysis shows that the bottleneck is the output layer predicting 10000 medical issues. To bring down the computational complexity, you have also connected the last hidden layer to another output layer that predicts 100 diseases. This improves the training and testing time. Why?