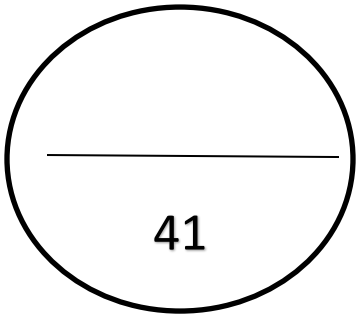
Name: _____ Reg #: _____ Section: _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Natural Language Processing**	**Course Code:**	**CS 535**
Program:	**BS(Computer Science)**	**Semester:**	**Spring 2018**
Duration:	**180 Minutes**	**Total Marks:**	**41**
Paper Date:	**23-May-18**	**Weight**	**50%**
Section:	**ALL**	**Page(s):**	**8**
Exam:	**Final Solution**		

Instruction/Notes: Attempt the examination on the question paper and write concise answers. You can use extra sheet for rough work. Do not attach extra sheets used for rough with the question paper. Don't fill the table titled Questions/Marks.

Question	1-4	5-7	8-10	11-14	Total
Marks	/ 9	/ 10	/ 12	/10	/ 41



Q1) You are given the following corpus: **[2 + 2 = 4 Marks]**

<s> She likes green apples </s>
<s> Ali likes green apples </s>
<s> green apples are good for health </s>
<s> I like red apples </s>

a) Calculate the probability of following test sentence using trigram language model with linear interpolation. Include <s> and </s> in your counts just like any other token. λ_1 = trigram weight, λ_2 = bigram weight, λ_3 = unigram weight, $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$

<s> He likes green apples </s>

Solution:

P (<s> He likes green apples </s>) = P1 * P2 * P3 * P4

$= (7.6 * 10^{-3}) * (1.53*10^{-3})*(2.15*10^{-3}) *(2.15* 10^{-3})$

$= 5.37 * 10^{-11}$

P1 = $\lambda_1 *$ Count (<s> He likes) / Count(<s> He) + $\lambda_2 *$ Count (He likes) / Count(He) + $\lambda_3 *$ Count (likes) / N

b) Calculate the probability of P(green | likes) using Kneser Ney smoothing from the corpus given above. d = discounting factor = 0.5

Solution:

$$P_{KN}(w_i \mid w_{i-1}) = \frac{\max(c(w_{i-1}, w_i) - d, 0)}{c(w_{i-1})} + \lambda(w_{i-1}) P_{CONTINUATION}(w_i)$$

= ((2 – 0.5) /2) + (0.25) (0.17)

Q2) Suppose a language model assigns the following conditional n-gram probabilities to a 3-word test set: 1/8, 1/2, 1/6. What is the perplexity? **[2 Marks]**

Solution:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1 \ldots w_{i-1})}}$$

Q3) $P_{continuation}(w)$ for a word is defined as follows: **[2 Marks]**

$$P_{CONTINUATION}(w) = \frac{\left|\{w_{i-1} : c(w_{i-1}, w) > 0\}\right|}{\sum_{w'} \left|\{w'_{i-1} : c(w'_{i-1}, w') > 0\}\right|}$$

a) Consider the following incomplete sentence:

"How much wood would a woodchuck chuck would if woodchuck could would chuck"
What is $|\{w_{i-1} : C(w_{i-1}\ w_i) > 0\}|$ for w_i="woodchuck"?

i. 0 ii. 1 iii. 2 iv. 3

b) Which word is more likely to complete the sentence (follow the last "chuck") based on $P_{continuation}$?

i. How ii. wood iii. would iv. chuck

Q4) Which of the following word pairs, A/B, has A as a hypernym of B? [1 **Mark]**

i. Washington/The United States
ii. vehicle/car
iii. Java/programming language
iv. wheel/car
v. None of the above

Name: _____ Reg #: _____ Section: _____

Q5) Consider a trigram HMM tagger with: **[3 Marks]**

_ The set K of possible tags equal to {D, N, V}

_ The set V of possible words equal to {the, dog, barks}

_ The following parameters:

q(D|*, *) = 1	q(N|D, V) = 0.3	e(dog|N) = 0.4
q(N|*, D) = 0.5	q(STOP|N, V) = 0.6	e(barks|N) = 0.6
q(V|*, D) = 0.5	q(STOP|V, N) = 0.4	e(dog|V) = 0.1
q(V|D, N) = 0.7	e(the|D) = 1	e(barks|V) = 0.9

with all other parameter values equal to 0. Write down the set of all pairs of sequences $x_1 \ldots x_{n+1}$, $y_1 \ldots y_{n+1}$ such that the following properties hold:

_ $p(x_1 \ldots x_{n+1}, y_1 \ldots y_{n+1}) > 0$

_ $x_i \in V$ for all $i \in 1 \ldots n$

_ $y_i \in K$ for all $i \in 1 \ldots n$, and $y_{n+1} = STOP$

Solution:

1. * * The dog barks STOP (D N V)
2. * * The dog barks STOP (D V N)
3. * * The barks dog STOP (D N V)
4. * * The barks dog STOP (D V N)
5. * * The dog dog STOP (D N V)
6. * * The dog dog STOP (D V N)
7. * * The barks barks STOP (D N V)
8. * * The barks barks STOP (D V N)

Q6) Show how following lexicalized grammar rule parameter is decomposed into 2 parameters for learning probabilities from training data. Also show how to use smoothed estimation for the decomposed parameters. **. [3 Marks]**

$q(S(saw) \rightarrow_2 NP(man)\ VP(saw))$

Solution:

$$
\begin{aligned}
& q(\mathsf{S} \rightarrow_2 \mathsf{NP\ VP} | \mathsf{S}, \mathsf{saw}) \\
= \ & \lambda_1 \times q_{ML}(\mathsf{S} \rightarrow_2 \mathsf{NP\ VP} | \mathsf{S}, \mathsf{saw}) + \lambda_2 \times q_{ML}(\mathsf{S} \rightarrow_2 \mathsf{NP\ VP} | \mathsf{S}) \\
& q(\mathsf{man} | \mathsf{S} \rightarrow_2 \mathsf{NP\ VP}, \mathsf{saw}) \\
= \ & \lambda_3 \times q_{ML}(\mathsf{man} | \mathsf{S} \rightarrow_2 \mathsf{NP\ VP}, \mathsf{saw}) + \lambda_4 \times q_{ML}(\mathsf{man} | \mathsf{S} \rightarrow_2 \mathsf{NP\ VP}) \\
& + \lambda_5 \times q_{ML}(\mathsf{man} | \mathsf{NP})
\end{aligned}
$$

Name: _____ Reg #: _____ Section: _____

Q7) Write down at least two different parse trees (with different probabilities) for following sentence and PCFG. **[4 Marks]**

"The boy saw the dog in the park with the telescope"

S → NP VP 0.8
S → NP VP PP 0.2
NP → DET N 0.5
NP → NP PP 0.5
VP → V NP 1.0
PP → P NP 1.0
N → dog 0.25
N → boy 0.25
N → park 0.25
N → telescope 0.25
V → saw 1.0
P → with 0.5
P → in 0.5
DET → the 1.0

Q8) In the following gloss of different word seneses of the wrods "bank" and "coast" are given. Compute similarity between the words "bank" and "coast" uing Lesk algorithm. **[4 Marks]**

Bank$_1$: sloping land (especially the slope beside a body of water)

Bank$_2$: a financial institution that accepts deposits and channels the money into lending activities

Bank$_3$: a long ridge or pile

Bank$_4$: an arrangement of similar objects in a row or in tiers

Bank$_5$: a supply or stock held in reserve for future use (especially in emergencies)

Coast$_1$: the shore of a sea or ocean

Coast$_2$: a slope down which sleds may coast

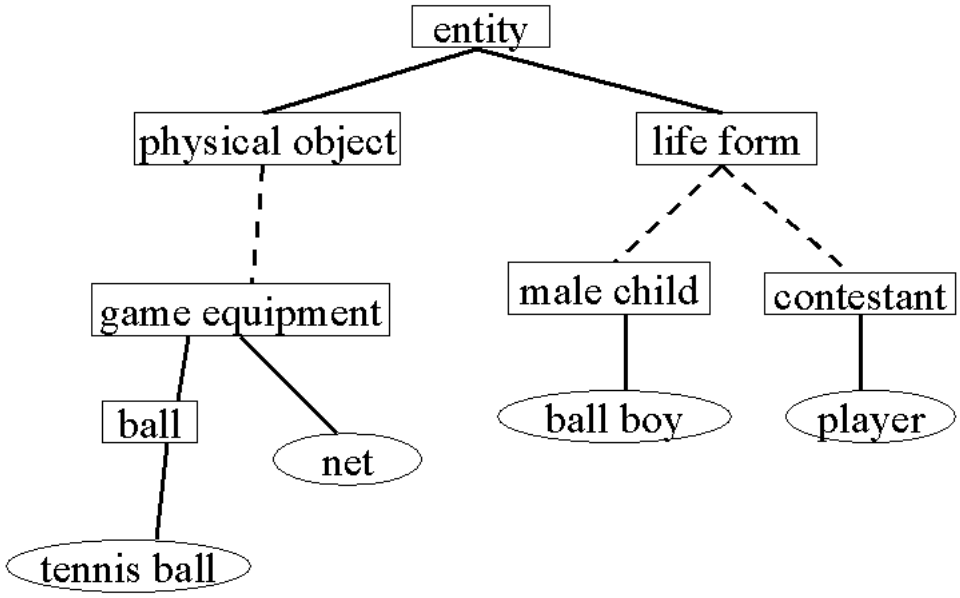
Coast$_3$: the area within view

Coast$_4$: the act of moving smoothly along a surface while remaining in contact with it

Solution:

Bank$_1$ and Coast$_2$ **= 1**

Q9) Following is a WordNet hierarchy. The probabilities of words are given in table below: **[4 Marks]**



Word	Probability
entity	0.395
Physical object	0.167
Life form	0.0231
Game equipment	0.00453
Male child	0.00153
contestant	0.00743
Ball	0.000343
Net	0.00054
Ball boy	0.000113
Player	0.000445
Tennise ball	0.000189

a) Compute path based similarity between "tennis ball" and "net"

Solution:

1/4

b) Compute information content based similarity proposed by Lin (Lin Similarity function) between "ball" and "player"

Solution:

log (0.395) / (log (0.0003)* log(0.0004))

Name: _____ Reg #: _____ Section: _____

Q10) a) Write down context vectors of words mango and apple using PPMI (Positive Pointwise Mutual Information) of words. **[2 Marks]**

Counts(w, context)				
	information	**data**	**sweet**	**Fitness**
Banana	**0**	**0**	**5**	**3**
Apple	**3**	**2**	**4**	**6**
Mechanical	**5**	**4**	**1**	**2**
computer	**7**	**6**	**0**	**1**

Solution:

Probabilities

Apple : (3/49)=0.06 0.04 0.08 0.12

PPMI : 0 0 0.42 0.73

b) Following table gives co-occurrence counts based on syntactic dependencies of words. Write down context vectors of words duty and responsibility using PPMI (Positive Pointwise Mutual Information) of words. (You can assume following table contains all words that can appear as object of a given a word. E.g. total count of words that appear as object of "assert" is 10. Sum of row counts represent total count of the word in collection. E.g. duty appears 22 times in collection. Total words in collection = N = 100) **[2 Marks]**

	Object of assert	**Object of assign**	**Object of avoid**	**Object of become**	**Modified by collective**	**Modified by assumed**
duty	3	4	5	3	5	2
responsibility	2	2	7	4	2	7
taxes	0	0	3	0	0	1
danger	0	0	6	0	1	0
control	5	0	0	1	0	0

Name: _____ Reg #: _____ Section: _____

Q11) Compute value of ROUGE-2 score for following summary. **[2 Marks]**

System Generated Summary: The quake had a preliminary magnitude of 6.9. in an area so isolated there are no roads connecting it to the outside world.

Reference Summary (Human Generated Summary): The quake had a preliminary magnitude of 6.9. An earthquake in the same region in February killed 2300 people and left thousands homeless.

Solution:

7/21

Q12) The first step in query focused multi document summarization is to simplify the sentences. Simplify following sentences using simple rules discussed in class. **[4 Marks]**

a) Genette's bedroom desk, the biggest disaster area in the house, is a collection of overdue library books, dirty plates, computer components, old mail, cat hair, and empty potato chip bags.
b) Robbie, a hot-tempered tennis player, charged the umpire and tried to crack the poor man's skull with a racket.
c) The car began sliding sideways, and then it hit the tree," she said
d) He died in France, as a matter of fact, and wated to be buried there.

Solution:

e) Genette's bedroom desk, is a collection of overdue library books, dirty plates, computer components, old mail, cat hair, and empty potato chip bags.
f) Robbie, charged the umpire and tried to crack the poor man's skull with a racket.
g) The car began sliding sideways, and then it hit the tree,"
h) He died in France, and wated to be buried there.

Q13) Word occurrence in sentiment analysis matters more than word frequency. Briefly describe difference between multinomial Naïve Bayes and Boolean Multinomial Naïve Bayes for sentiment analysis. **[2 Marks]**

Solution:

Boolean Multinomial Naïve Bayes clips word counts of all words in all documents at 1.

Q14) Give at least 5 features that can be used to resolve ambiguity in name entity recognition. **[2 Marks]**

Solution:

Identity of word

Neighboring words

Part of speech of word

Part of speech of neighboring words

Uppercase

Shape of word

Presence of hyphen