NATIONAL UNIVERSITY OF SINGAPORE
SCHOOL OF COMPUTING

CS4248 – Natural Language Processing

Semester 1 AY2023/2024

December 2023 **Time Allowed: 1.5 Hours**

INSTRUCTIONS TO CANDIDATES

1. This assessment paper contains **SIX (6)** questions and comprises **TWELVE (12)** printed pages, including this page.
2. Answer **ALL** questions within the space in this booklet.
3. This is a **CLOSED** book assessment, but one double-sided A4 sized sheet is allowed for notes.
4. A non-programmable calculator is permitted.
5. Please write your Student Number below. Do not write your name.

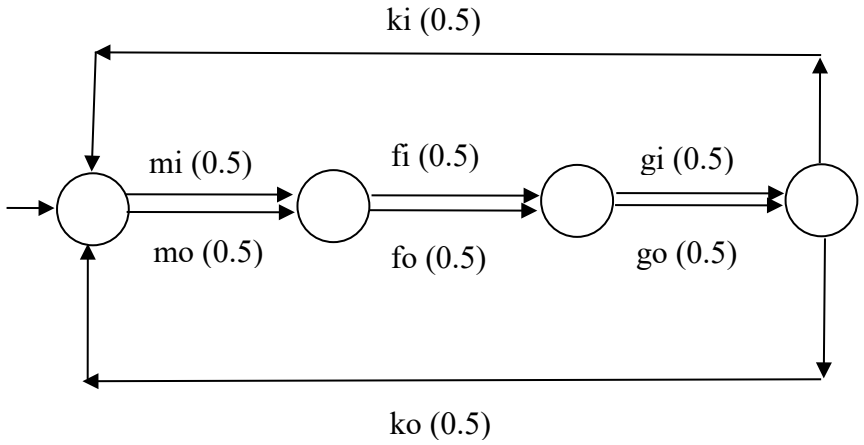
This portion is for lecturer's use only

Question	Q1	Q2	Q3	Q4	Q5	Q6	Total
Max	20	20	15	15	15	15	100
Marks							

1. (20 marks) Give a trace of the minimum edit distance algorithm (a dynamic programming algorithm) to compute the minimum cost of transforming the string "roses" to "easy", by filling out every cell entry in the following table, where each cell entry denotes the minimum cost of transforming the associated substrings. Assume that the cost of inserting a character is 1, the cost of deleting a character is 1, and the cost of substituting a character by a different character is 2. (You do not need to show the optimal path.)

s	5				
e	4				
s	3				
o	2				
r	1				
	0	1	2	3	4
		e	a	s	y

2. Consider a language defined as follows:



That is, the first word is either mi or mo, the second word is either fi or fo, the third word is either gi or go, the fourth word is either ki or ko, the fifth word is either mi or mo, the sixth word is either fi or fo, etc. The transition probability for each word is enclosed in brackets, and the vocabulary of this language is { mi, mo, fi, fo, gi, go, ki, ko }

Answer the following questions, showing clearly the steps of your calculations to justify your answers. **Simplify your answers as much as possible.**

(a) (5 marks) What is the entropy of this language?

Let X be a random variable ranging over all finite sequences of words of length n in this language, with true probability distribution given above.

(b) (15 marks) Consider an incorrect model where the transitions for “mi”, “fi”, “gi”, and “ki” are assigned probability of ¼, and the transitions for “mo”, “fo”, “go”, and “ko” are assigned probability of ¾. What is the per-word cross entropy of X using this model?

3. (15 marks) A perceptron F receives inputs $x_1, \ldots, x_n$ and outputs the following:

$$F(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } w_0 + w_1 x_1 + \cdots + w_n x_n > 0 \\ 0 & \text{otherwise} \end{cases}$$

Consider the Boolean function $\big((\neg x_1 \vee x_2) \wedge (x_1 \vee \neg x_2)\big) \vee \neg(x_1 \vee x_2)$ where $x_1, x_2 \in \{0, 1\}$. Is it possible to find 3 weights w_0, w_1, w_2 such that F implements this Boolean function?

Answer Yes or No here: _____

If yes, provide values for the weights w_0, w_1, w_2. If not, give a rigorous proof that no such weights exist. Justify each step of your proof.

(Additional space for answering question 3, if needed)

4. (15 marks) In logistic regression, each example $\boldsymbol{x}$ is a vector $[x_1, \ldots, x_n]$ of n features (real numbers) and its class is $y \in \{0, 1\}$. Logistic regression learns a vector $\boldsymbol{w} = [w_1, \ldots, w_n]$ of weights (real numbers) and a bias term b, such that

$$P(y = 1) = \sigma(\boldsymbol{w} \cdot \boldsymbol{x} + b) = \frac{1}{1 + e^{-(\boldsymbol{w} \cdot \boldsymbol{x} + b)}} = \frac{1}{1 + e^{-(w_1 \cdot x_1 + \cdots + w_n \cdot x_n + b)}}$$

$$P(y = 0) = 1 - P(y = 1)$$

where σ is the sigmoid function. The cross-entropy loss function L is as follows:

$$L = -y \log\big(\sigma(\boldsymbol{w} \cdot \boldsymbol{x} + b)\big) - (1 - y) \log\big(1 - \sigma(\boldsymbol{w} \cdot \boldsymbol{x} + b)\big)$$

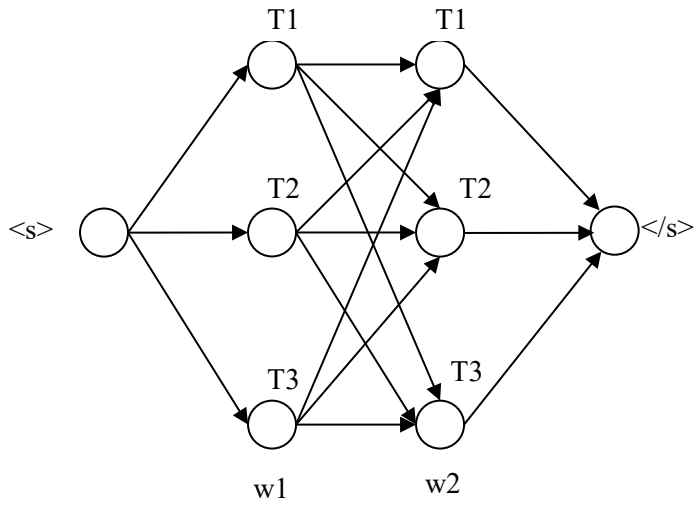
where $\log \equiv \log_e$ (i.e., base e logarithm). To update w_i with gradient descent during training, we need $\frac{\partial L}{\partial w_i}$.

In the space below, show the step-by-step derivation of $\frac{\partial L}{\partial w_i}$, by taking the partial derivative of L with respect to w_i, and express $\frac{\partial L}{\partial w_i}$ in terms of $x_i, y, \sigma(\boldsymbol{w} \cdot \boldsymbol{x} + b)$. **Simplify your expression as much as possible**.

(Additional space for answering question 4, if needed)

(Additional space for answering question 4, if needed)

5. (15 marks) Consider the following HMM:



Suppose this HMM has the following set of parameters:

P(T1|<s>) = 0	P(T1|T1) = 1/6	P(T1|T2) = 1/8	P(T1|T3) = 1/5
P(T2|<s>) = 1/4	P(T2|T1) = 2/3	P(T2|T2) = 1/2	P(T2|T3) = 1/5
P(T3|<s>) = 3/4	P(T3|T1) = 1/12	P(T3|T2) = 1/4	P(T3|T3) = 3/5
	P(</s>|T1) = 1/12	P(</s>|T2) = 1/8	P(</s>|T3) = 0

P(w1|T1) = 1/20	P(w2|T1) = 1/10
P(w1|T2) = 1/5	P(w2|T2) = 1/10
P(w1|T3) = 1/10	P(w2|T3) = 1/10

T1, T2, T3 are part-of-speech tags.
Consider the input sentence “w1 w2”, where w1 and w2 are words. Trace the Viterbi algorithm, by providing the values of the cells v(T, w) where T $\in$ { T1, T2, T3 }, and w $\in$ { w1, w2 }, and determine the optimal sequence of part-of-speech tags.

(Additional space for answering question 5, if needed)

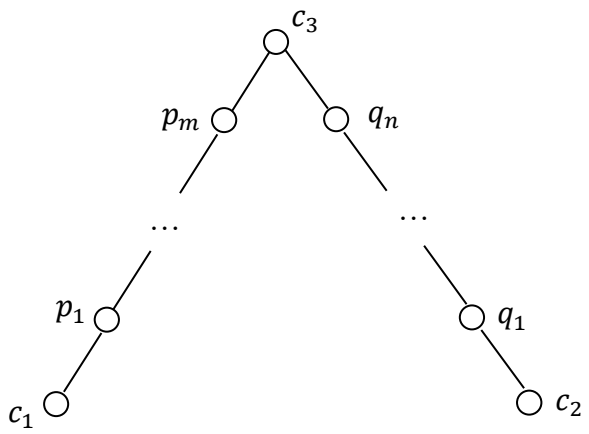
6. (15 marks) Consider concepts arranged in the WordNet ISA hierarchy. The information content of a concept c, $IC(c)$, is defined as follows:

$$IC(c) = -log\, P\,(c)$$

where $P(c)$ is the probability of concept c.
Consider a concept c and its parent concept p. The distance $D(c,p)$ between the two concepts c and p is defined as:

$$D(c,p) = -log\, P\,(c|p)$$

Given two concepts c_1 and c_2, let c_3 be the most specific concept that subsumes both c_1 and c_2. Graphically,



where the intervening concepts $p_1, p_2, \ldots, p_m, q_1, \ldots, q_n$ are such that p_1 is the parent of c_1, p_2 is the parent of p_1, …, c_3 is the parent of p_m, c_3 is the parent of q_n, …, q_1 is the parent of c_2.
The distance $D(c_1, c_2)$ between the two concepts c_1 and c_2 is then defined as the sum of the distances between the intervening concepts. That is,

$$D(c_1,c_2) = D(c_1,p_1) + D(p_1,p_2) + \cdots + D(p_m,c_3) + D(q_n,c_3) + \cdots + D(c_2,q_1)$$

You are to derive an expression for $D(c_1, c_2)$ in terms of $IC(c_1)$, $IC(c_2)$, and $IC(c_3)$. Show and justify the steps of your derivation.

(Additional space for answering question 6, if needed)

(Additional space for answering any questions in this paper, if needed)

END OF PAPER