

## INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR Mid Semester Examination 2022-23

Date of Examination: $\frac{\sqrt{09/23}}{\sqrt{99/23}}$ Session(FN/AN) Duration 3 hrs, Marks = $50$			
Sub No: CS60075 Sub Name: Natural Language Processing			
Department/Centre/School: Computer Science and Engineering			
Specific charts, graph paper, log book etc. required NO Special Instructions (if any) ANSWER ALL questions. In case of reasonable doubt, make assumptions and state them upfront. Marks will be deducted for sketchy answers and claims without proper reasoning.  All parts of a single question should be done at the same place. Keep probability values in fractional forms.			
1. The following sentence, though bizarre and deliberately confusing, is grammatically correct!			
The weasel that a boy that startles the cat thinks loves smiles eats.			
Answer the following questions by identifying the correct parse (Hint: you might use square brackets as we used in class to represent the parse/breakup of the sentence). In some cases, the answers may be 'nobody/nothing in this sentence'.			
(a) What is the subject of this sentence? (Give a single-word answer.)			
(b) How many verbs are in the sentence?			
(c) Who startles whom or what?			
(d) Who thinks what?			
(e) Who loves whom or what?			
(f) Who smiles?			
(g) Who eats whom or what?			
[7]			
2. Consider the following toy example:			
Training data:			
<s> I am Sam </s>			
<s> Sam I am </s> <s> Sam I like </s>			
<s> Sam I do like </s>			
<s> do I like Sam </s>			
Assume that we use a bigram language model (no smoothing) based on the above training data. Compute the bigram probabilities.			
(a) What is the most probable next word predicted by the model for the following word sequences?			
i. <s> Sam .</s>			
ii. <s> Sam I do</s>			
iii. <s> Sam I am Sam</s>			
iv. <s> do I like</s>			
(b) Which of the following sentences is better, i.e., gets a higher probability with this model? Compute the probabilities in each case and hence find the best one.			
i. <s> Sam I do I like </s>			
ii. <s> Sam I am </s>			

그 하는 그를 가는 하는 하는 하고 그 말로 하는 그림에 하는 것 같아 나는 것 같아.	
사람들은 중요 중요 그리고 되어 있다는 사람들은 사람들이 되는 사람들이 되었다면 하는 것이다.	
化铁铁铁 医骨髓管 医克勒氏试验检检验 医二氏性病 医二氏性 医二氏管 医阿克斯氏氏管 化二氯甲基	
이 사이들의 물로 하는 사람이들이 그들이 말하는 사람들을 모든 이 기가 되었다.	
는 양면 하늘 보는 경기가 가지를 통합 감상하다는 하나 가다 가입하다 되었다.	
그는 '로그램 기존한다'는 얼마를 보자하는 물을 받았다니까 그는 그림이 하다	
그들이 하는 경기를 가지 않는데 한 한 한 생생하실 수가 되었다. 방향을 보고 보고 회원을 위하는 모양한 원 회학 보고 되어	医外侧侧面 化电子电子电子 医克拉二氏反应性
그는 그는 항상 회장으로 부모인 사람들이 있다고 말했다. 그리고 생각이 되었다. 그리는 그리고 있다. 하나는	
이 어느 어느 이 살려면 그만 들었다가 느껴들었다면서 나는 그는 사람 가장 하면 모든 것이 되었다.	
그리는 어떻게 이 있었다. 연상의 없다는 이용학생인이 작은 사람들은 보다 하는데 되고 모양하는 사람이 하는데 하는데	
지도 있는 회에 가장 회장 회복에 되는 일본 등록하고 되었습니다. 하고 있는 그리는 작은 사람들이 되는 것이다.	
그렇게 되었다. 그는 방문에 가고를 하는 말에 들어들면 조물리다는 이번 수 있다. 그 한 번 학문에 되었다.	
가는 다른 사람들이 다시 눈을 하는데 되었는데 말을 하다고 말은 사람들이 하는 것들은 함께 들었다. 나는	
그는 눈도 가족 말씀이 불러 내려면 가득하면 한 한 반속을 하는 가는 가는 하는 것이 되는 것이다.	
그렇는 그를 하더라고 있다. 그런 그를 하는 말을 하는 말을 하는 것이 하는 것이 없다. 그는 그 사람들이 모든 사람들이 살아 되었다.	
老 一位 建邻苯酚 100 医克克克氏病性原皮病 高麗國際 计对象 医囊膜 经股份 医多方法氏反射	
지수의 그는 이 그리가 그들은 살보는 이미를 하는 것을 하는 것은 그릇을 하는 것은 것으로 하는 것은 것이다.	
그는 그리고 그리는 그를 하면 그리는 물로 가득하는 것 같아요. 그리고 있다면 얼마를 하지 않는 그 살을 하는데 하다.	
그 이 문으로 하는 역사는 이번 가는 계속이 되는 그 하는 사람들은 사람들은 사람들이 되었다. 그 나는 사람들은 사람들은 사람들이 되었다.	
그 생활하는 이 학생으로 생활을 보는 학생들은 기울이 되었다. 하늘은 그 가장 보험이 되었다고 있는 것이다. 나는 사람이	
그는 사고 가게 하면 되었다. 그는 사람들은 회사에 가는 것이 그 생각 수밖에 되는 것이다면 되었다.	
그러워 하지만 하다면 하는데 말만 살아가는데 하는데 하는데 하는데 되는데 하는데 하는데 나는데 뭐 하는데 되었다.	
그렇게 나는 경험적으로 들어왔는 하는 사람들이 들어가 되었다면 사람들은 환경 사고 되는 것 같아.	
그는 회에 가장 그리트를 가능한 경찰들은 경찰들은 경찰들이 되었다. 그리는 사람은 속으로 보고 모르는 그	
가는 이 이번 가는 생각 생각이 하는데 그리다면 하는데 이 없는데 하는데 그리고 되는데 되었다.	
그녀는 이번에 대한테 보고한 회사회에는 내가 있었다. 이 그리는 것은 사이는 것은 그는 가는 이 그는 것은 것은	
그는 이 사람들은 작가들은 그의 물에 가는 경험 하는 사람들이 가는 것을 하는데 가는 것이다.	
에 가지, 그리고, 소리한테 하지수요. 그러면 그리아 있게 되었습니다. 그리아 그리아 그리아	
그리자 그 문화장에 경찰과 교회 (취찰) 그 그는 그리는 동안 원하는 그 문화가 없는 것이다. 나는	
그 가는 사람들은 사용 회사의 회문 가장 사람들이 되는 사람들이 가장 가장 내가 있다. 그리고 있다면 그렇게 되었다면 하는데 가장 살아 있다면 하는데	
그리는 사람들이 생물에 기가를 가는 점점이 되었다. 그 가장 그 살에게 이 경찰에는 사람이 되었다.	
지방 시원분들 역 중요? 얼마 이 한 동 약동 회는 후 사람이 되고 있는 것이	一 医二种正电影 化合金压力 医基础 医神经病
그리고, 어른하는 이 글로 생각을 하하고 한 물을 만든 하는데 한 장악 되는데 아버지는 사이에 있다.	
그리아, 그렇다 어린 화물활동 가수 생각하고 있다고 한다고 싶었다면 나는 이 모든 것이다.	
	그리는 하시는 이 그 경우 연구 원칙에 들어야
그들은 그 하지만 그런 그들의 얼마나 되었다. 그는 그동이 뭐 된 것이 하셨다고 하는 것이다. 이 모모는 그 모양이	
그들은 음악 프로막 말았다고 말았다. 그리는 이 어디를 하는데 그렇게 되었다.	
- 하고 - 그 문에 참 및 학료 회원 후의 시작 등관 문학인 모든 사람이 있는 학자가 고급하는 것 같습니다.	
	机造成海头 化二氯化二氯甲烷二氯甲烷
그는 그리는 이번에는 본 경우학자 중에 가진 항문이나 많아도 되고 하지만 모으고 하는 것이다.	
보는 항 그 나는 사람들은 회장에는 소문을 하면 한 번째 가는 경험에 되었다. 그는 그 그 그 그 그 그 그 그 그 그 그 그 그 그 그 그 그 그	
그리는 심한 강 그는 이 문학문에 그렇게 말한다고 살았다. 나는 다른 이 가는 네트를 가는 것이 되었다.	
사람들 학자들은 학자 중심장이 되고 있다. 사람들은 그 그 때문에 한 사람들은 학생들이 되었다. 그는 사람이	
하면 바다 하는 일반하다 그 사람들은 그를 가는 것을 만든 사람들이 되었다.	
HOO NEW HOOM HE	
그 전 가능원 작가 있는데 그들은 생님은 생범하는 그들은 그 그를 가는 것이	
医内性性视性结合性 医动物性 医动物 医二苯基甲酰胺 化二氯甲酚 网络马马马马马马马马马马	
나는 있는 물로 함께 되는 것 같아 가려고 말을 하는 어느 경기를 하는 것이 되었다.	
어느 사람들은 그리는 아래가 얼마나 살아 있는 사람들이 되었다.	
그는 현실 교육을 되는 건물을 하실해 잘 하고 있는데 어떤 것은 사람이 되고 있다.	
ニー・コート もっきょう しょうしょく しょくしゅう 大陸 アン・コース はっこう しょうしょく しょくしょく しょうしょく たんしょく	

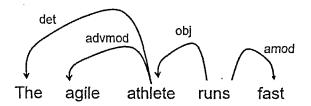


Figure 1: Dependency tree.

iii. <s> I do like Sam I am </s>

- (c) Consider again the same training data and the same bigram model. Compute the perplexity of <s> I do like Sam
- (d) Take again the same training data. This time, we use a bigram language model with Laplace smoothing.
  - i. Give the following bigram probabilities estimated by this model: P(do|s), P(do|s), P(Sam|s), P(Sam|s), P(Sam|s), P(I|s), P(I|do), P(I|k|I)Note that for each word  $w_{n-1}$ , we count an additional bigram for each possible continuation  $w_n$ . Consequently, we have to take the words into consideration as well as the symbol s.
  - ii. Calculate the probabilities of the following sequences according to this model. Which of the two sequences is more probable according to our language model?

A. <s> do Sam I like

B. <s> Sam do I like

$$[3+2+4.5+1.5+3+4=18]$$

3. What is the tagging of the following sentence

computers process programs accurately

with the following HMM tagger (N is noun, V is verb, Adv is adverb):

(part of) lexicon (assume that the following words have no other POS apart from those given below):

computers N 0.123

process N 0.1

process V 0.2

programs N 0.11

programs V 0.15

accurately Adv 0.789

(part of) transitions:

$$P(N|V) = 0.5, P(N|Adv) = 0.12, P(V|Adv) = 0.05, P(V|N) = 0.4, P(Adv|N) = 0.01, P(Adv|V) = 0.13, P(N|N) = 0.6, P(V|V) = 0.05$$
  
Show the computations. [10]

- 4. Assume that the texts to be POS tagged contain 1.5% of unknown words and that the performance of the POS tagger to be used is 98% on known words. What will be its typical overall performance if all unknown words are systematically wrongly tagged?

  [3]
- 5. In the dependency graph shown in Figure 1, find out the labels that are incorrect on the arcs. Correct them. avmod means adverbial modifier and amod means adjectival modifier. [3]

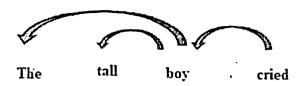


Figure 2: Reference sentence.

6. Give the correct sequence of arc eager parsing operations for the given sentence in Figure 2. Show the (a) action taken, (b) stack state, (c) buffer state. Provide a modified transition sequence where the parser mistakenly predicts the arc boy → cried, but gets the other dependencies right.

[2.5 + 2.5 = 5]

7. Suppose in a recommender system for online shopping, we have information about co-purchase records for items  $x_1, x_2, ..., x_n$  (for example, item  $x_i$  is commonly bought together with item  $x_j$ ). Explain how you would use ideas of distributional similarity to (i) define a context (ii) obtain embeddings (how and for what?) and (iii) recommend similar items to users who have shown interest in any one of the items. [1 + 2 + 1 = 4]