Semester Final of Natural Language Processing

Natural Language Processing (Khulna University of Engineering and Technology)

KHULNA UNIVERSITY OF ENGINEERING & TECHNOLOGY
B.Sc. Engineering 4th Year 2nd Term Examination, 2018
Department of Computer Science and Engineering
CSE 4221
Natural Language Processing

TIME: 3 hours FULL MARKS: 210

N.B. i) Answer **ANY THREE** questions from each section in separate scripts.
ii) Figures in the right margin indicate full marks.

SECTION A

(Answer **ANY THREE** questions from this section in Script A)

cT
1. a) What is Disjunction, Grouping and Precedence for pattern matching in regular expressions? Explain with example. (10)

cT b) "Pattern matching by regular expressions are greedy." – Justify the statement. (09)

cT c) Design a regular expression to find all instances of the word "the" in a text. (10)

cT d) Define types and tokens. How many types and tokens are there in the following sentence: (06)

"They picnicked by the pool, then lay back on the grass and looked at the stars."

cT
2. a) What is Lemmatization and Stemming? How is Lemmatization done? (08)

cT b) What are the operations for editing one string to another? Explain. (06)

cT c) Explain the algorithm to edit one string X of length n to a string Y of length m. Show the steps of your algorithm for X = INTENTION and Y = EXECUTION. (12)

d) Discuss about the problem with Maximum Likelihood. How does the Laplace (add-1) Smoothing solve the problem? (09)

3. a) "Accuracy is not a good metric when the goal is to discover something that is rare." – *Justify* the statement with example. Propose a metric to solve the drawbacks of accuracy. (10)

b) Given the following short movie reviews, each labeled with a genere, either comedy or action: (10)

i) fun, couple, love, love ⇒ comedy ✓
ii) fast, furious, shoot ⇒ action
iii) couple, fly, fast, fun, fun ⇒ comedy ✓
iv) furious, shoot, shoot, fun ⇒ action
v) fly, fast, shoot, love ⇒ action

Consider a new document D: fast, couple, shoot, fly. Compute the most likely class for D.

c) Find the context free rules and hence the Context Free Grammar (CFG) for the following English sentences: (15)

i) I want a morning flight.
ii) I want a flight from Ontario to Chicago.
iii) Show me the cheapest fare that has lunch.
iv) Do any of these flights have stops?
v) Which flights serves breakfast?

4. a) Consider the following grammar in CNF. (10)

$$S \rightarrow AB \mid BC$$
$$A \rightarrow BA \mid a$$
$$B \rightarrow CC \mid b$$
$$C \rightarrow AB \mid a$$

Is '*baaba*' in L(G)? Explain your answer using CYK algorithm.

b) Define shallow parsing. What are the applications of shallow parsing? (05)

c) Define Probabilistic Context Free Grammar (PCFG). Consider the following PCFG. (12)

$$S \rightarrow NPVP \mid AuxNPVP \mid VP\ [0.8| 0.1 \mid 0.1]$$
$$NP \rightarrow Pronoun \mid Proper\text{-}noun \mid DetNominal\ [0.2 \mid 0.2 \mid 0.6]$$
$$Nominal \rightarrow Noun \mid NominalNoun \mid NominalPP\ [0.3 \mid 0.2 \mid 0.5]$$
$$VP \rightarrow verb \mid verbNP \mid VPPP\ [\ 0.2 \mid 0.5 \mid 0.3]$$
$$PP \rightarrow PrepNP\ [1.0]$$
$$Det \rightarrow the \mid a \mid that \mid this\ [0{,}6 \mid 0.2 \mid 0.1 \mid 0.1]$$
$$Noun \rightarrow book \mid flight \mid meal \mid money\ [0.1 \mid 0.5 \mid 0.2 \mid 0.2]$$
$$verb \rightarrow book \mid include \mid prefer\ [0.5 \mid 0.2 \mid 0.3]$$
$$Pronoun \rightarrow I \mid he \mid she \mid me\ [0.5 \mid 0.1 \mid 0.1 \mid 0.3]$$
$$Proper\text{-}noun \rightarrow Houston \mid NWA\ [0.8 \mid 0.2]$$
$$Prep \rightarrow from \mid to \mid on \mid near \mid through\ [0.25 \mid 0.25 \mid 0.1 \mid 0.2 \mid 0.2]$$

i) Find the probability of the sentence "book the flight through Houston".
ii) Using the disambiguation algorithm select the proper parse tree.

d) What are the stages of IR based question answering? Explain. (08)

SECTION B

(Answer **ANY THREE** questions from this section in Script B)

5. a) Define Natural Language Processing (NLP). What are the major areas of research and development of NLP? (10)

   b) What does *n*-gram mean? Drive the equation of calculating the probability for *n*-grams model. (10)

   c) Consider the following corpus. (08)

   <s> I am Sam </s>
   <s> Sam I am </s>
   <s> I am Sam </s>
   <s> I do not like green eggs and Sam </s>

   Using a Bigram Language model with add-one smoothing, what is $P(sam|\ am)$? Include <s> and </s> in your counts just like any other token.

   d) What is absolute discounting? What is its advantages? (07)

6. a) What is closed class and open class of Part-of-Speech (POS)? Explain with example. (08)

   b) Discuss about Rule-Based POS tagging. Write the ADVERBIAL-THAT RULE. (12)

   c) For Hidden Markov Model (HMM) POS Tagging, using the following formula, find the equation of calculating tag transition probabilities. (08)

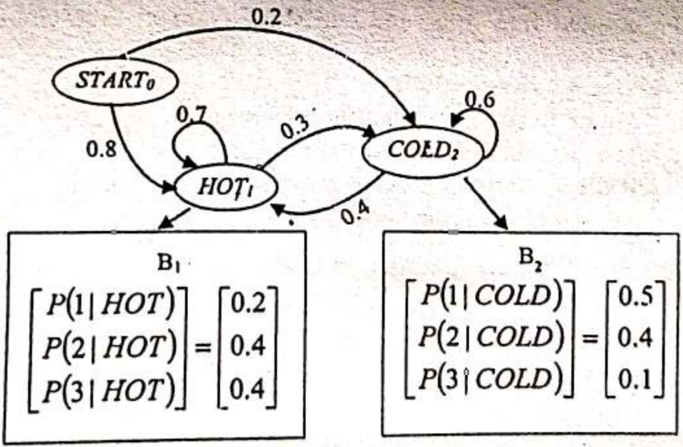
$$\hat{t}_1^n = \underset{t_1^n}{\arg\max}\, P\left(t_1^n \mid w_1^n\right)$$

   d) Consider the sentence: "Secretariat/NNP is/BEZ expected/VBN to/TO race/? Tomorrow/NR". The word "race" is often used as VB or NN. Given the probabilities below, find the right POS tag for the word "race". (07)

$P(NN|\ TO) = 0.00047, P(VB|\ TO) = 0.83, P(race|\ NN) = 0.00057, P(race|\ VB) = 0.00012,$
$P(NN|\ VB) = 0.0027, P(NR|\ NN) = 0.0012.$

7. a) HMM characterized by three fundamental problems. Name and discuss about the problems. (09)

   b) Given a sequence of ice-cream observations 313 and an HMM $\lambda = (A, B)$ in the following figure, find the best hidden weather sequence $Q(like\ H\ H\ H)$. (12)



$$\begin{bmatrix} P(1|HOT) \\ P(2|HOT) \\ P(3|HOT) \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.4 \\ 0.4 \end{bmatrix} \qquad \begin{bmatrix} P(1|COLD) \\ P(2|COLD) \\ P(3|COLD) \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.4 \\ 0.1 \end{bmatrix}$$

   c) Define the term odds for logistic regression. Show that the observation should be labeled true if $\sum_{i=0}^{N} w_i f_i > 0$. (09)

   d) Write the three-steps of Forward Algorithm. (05)

8. a) Name and discuss about the types of TTS. (06)

   b) Speech Synthesis perform text to waveform mapping in two-steps. Name and discuss about the steps. Using Hourglass Metaphor. (12)

   c) What is Homograph disambiguation? What are the problems of CMU? How does UNISYN overcome the problems of CMU? (10)

   d) Define text normalization. Why does text normalization important for Speech Synthesis? (07)