NATIONAL INSTITUTE OF TECHNOLOGY, HAMIRPUR
Computer Science & Engineering Department
End Semester Exam-December 2020

Course: M.Tech (1st Year) **Semester: 1st**
Subject Code: CS-653 **Duration:** 2 Hours
Subject Name: Natural Language Processing **Max Marks:** 50
Date: 16/12/2020 **Time:** 03:00PM–05:00PM

Note: Attempt any **five** questions out of 6 given here.
All questions carry equal marks.

Instructions:

- Each Student is required to write his/her **Name, Roll No, Subject Name, Subject Code, Programme, Semester, Department, Date of Exam and Number of pages on top of the first sheet** and put the **Signature with date at the bottom of each sheet** of the answer booklet.
- Please make one complete pdf of your scanned answer sheet and save it with the name as **rollnosubjectcode.pdf** only and then upload it.
- After the exam time is over, students may take extra 15 minutes to scan and upload their answer booklet on Google Classroom. Further, **delay in submission by a student may lead to deduction in marks or rejection of whole answer booklet**.

Q.1 a) Give one original (i.e. not discussed in class) example each of an English sentence with semantic ambiguity, syntactic ambiguity, anaphoric (i.e. referential) ambiguity. In each case specify the ambiguity clearly. [5]

b) What do you mean by Part-of –Speech tagging? What is the need of this task in NLP? Find one tagging error in each of the following sentences that are tagged with the Penn Treebank tagset: [5]

1. I/PRP need/VBP a/DT flight/NN from/IN Atlanta/NN
2. Does/VBZ this/DT flight/NN serve/VB dinner/NNS
3. I/PRP have/VB a/DT friend/NN living/VBG in/IN Denver/NNP
4. Can/VBP you/PRP list/VB the/DT nonstop/JJ afternoon/NN flights/NNS

Q.2 a) Discuss the different issues that need to handle in machine translation. Explain the process of statistical machine translation. [4]

b) Suppose that you are given the following sentences: [6]

- Chinese Beijing Chinese
- Chinese Chinese Shanghai
- Chinese Macao
- Tokyo Japan Chinese

I. Learn a Bi-gram Language model using this data with add-1 smoothing.
II. Using the language model learnt in part (i) above, estimate the probability for the sentence “Chinese Chinese Chinese Tokyo Japan”.

Q.3 a) Describe thematic roles used in the semantic analysis with example. [3]

b) Design an FSA that recognizes simple date expressions like March 15, 22 November. [3]

c) Write a regular expression for different dates and time format. Those regular expressions should match the following patterns of date and time. [4]
Pattern for date: dd-mm-yyyy
Example Pattern for time: 09:15 pm 12:59 AM 01:30PM

Q.4 a) Explain the following evaluation metrics which are being used for evaluation of different NLP applications: [5]

i. Precision
ii. Recall
iii. F-measure
iv. Accuracy
v. BLEU

b) Write an algorithm for deterministic recognition of FSAs which should return accept if the entire string it is pointing at is in the language defined by the FSA, and reject if string is not in the language. [5]

Q.5 a) What is Concordance? Explain with suitable example. [4]

b) Imagine that we run the piano department in a prestigious music school. It's admissions season, and 25 nervous candidates are ready to demonstrate their skills on the beautiful Steinway grand piano. We ask two experienced professors — known as Professor A and Professor B — to rate each candidate as **Accept**, **Waiting List (WL),** or **Reject**. Here are the audition results for the 25 candidates: [6]

Student	Professor A	Professor B		Student	Professor A	Professor B
1	WL	Accept		14	Reject	Reject
2	WL	Accept		15	WL	Accept
3	Accept	Reject		16	Accept	Accept
4	Reject	Reject		17	WL	Reject
5	Reject	Reject		18	Reject	Accept
6	WL	Accept		19	WL	Accept
7	WL	Accept		20	Accept	Accept
8	Accept	WL		21	WL	WL
9	WL	WL		22	WL	Reject
10	Reject	Reject		23	Accept	Accept
11	Reject	Accept		24	Reject	Reject
12	Reject	Reject		25	Accept	Accept
13	Reject	Accept				

Each candidate now has two ratings, one from each professor. Naturally, the professors agree on some candidates and disagree on others. A high level of agreement between the professors increases our confidence that the ratings are reliable. A low level of agreement means that we cannot trust the ratings. Measures the *degree of agreement* between the two evaluators using kappa statistics.

Q.6 Write short notes on the following: [2*5 = 10]

a) Turing Test
b) Bilingual corpora
c) Sentiment Analysis
d) Word Sense Disambiguation
e) Topic Modelling