End Semester Examination
Course Name: Natural Language Processing
Code: CS 563
Full Marks-100 **Time: 3 hours**
Answer ALL the questions

Make reasonable assumptions as and whenever necessary. You can answer the questions in any sequence. However, answers of all the parts to any particular question should appear together.

1. What is the difference between homonymy and polysemy? Give an example of each that illustrates your point. Describe Lesk algorithm for Word Sense Disambiguation using appropriate examples. How is selectional preferences learnt to handle WSD problem? **5+7+8**

2. Consider the following grammar:

   S→NP VP, VP→Verb NP, VP→Verb PP, VP→VP PP, NP→NP and NP, PP→P NP, NP→Ram, NP→Delhi, NP→Patna, NP→May, Verb→flew, P→in, P→to, CONJ→and

   Show three parse trees that would be derived for the sentence "*Ram flew to Patna and Delhi in May*". Given a treebank how would you estimate probabilities for the grammar rules given above (for use with a basic probabilistic parser). Show the various steps of the CKY parser that generates the sentence. **20**

3. (a)Mark the noun groups in the following sentence using BIO (beginning, intermediate, other) tags: Mary has a room with a view and a bottle of coke.
   (b) What are the challenges of Named Entity Recognition (NER)? Show with examples how global knowledge helps in better classification of named entities? How does NER and Coreference resolution interact with other? While evaluating a NER system, why does partial matching seem to be relevant? "As PoS tagging and NER are both sequence labeling problems and hence they can be treated similarly"- is it always correct? Please justify your claim with proper explanations.
   3+(3+3+3+3+5)

4. Distinguish the terms "anaphora" and "cataphora" with proper examples. What are the well-known metrics used for evaluating the performance of any coreference resolution system? Please discuss the need for the existence of different metrics in terms of their basis and characteristics (***Note****: explain at least for two metrics such as MUC and BCubed*). The method proposed in the following paper is still considered to be a baseline for machine learning approach in coreference resolution: *A Machine Learning Approach to Coreference Resolution of Noun Phrases*, Computational Linguistics 27(4):521–544.

   What are the basic components used for building the model? What are the features used for coreference resolution? How many of such features do you believe can be adapted for other domains and languages? Explain with examples how are the

training instances generated for coreference resolution? What is the process of decoding? **4+6+ 15**

5. Why is sentiment analysis at the aspect level important? Please explain the positive, negative, neutral and conflict sentiment classes with respect to aspect based sentiment analysis. Describe the basic components of any IR system. What are the typical problems associated with mixed-script IR? Explain how is query expansion important in IR. **3+4+4+4**