End Semester Examination
Course Name: Natural Language Processing
Code: CS 563
Full Marks-80 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.

(Q1). How do feature selection and ensemble learning improve the efficiency of classification? "Bagging and boosting have distinguishing properties for classification"-discuss this with proper explanations. If the dataset is unstable, i.e. minor changes in distribution makes the significant change in classification performance, which ensemble learning technique(s) would be preferable and why?

4+6+3

(Q2). Discuss the various dimensions of sentiment analysis with appropriate examples. Why is Aspect based Sentiment Analysis (ABSA) important? Consider the following problem of ABSA:

No. of classes=3 (Positive, Negative and Neutral); **Domain**: Product reviews; **Classifiers**: Conditional Random Field and Support Vector Machine; **Features**: Lexical, syntactic and semantics; **No. of features for aspect term extraction**: 10; **No. of features for sentiment classification**: 6

Build a particle swarm optimization (PSO)-based Feature Selection and Ensemble Construction algorithms for Aspect Term Extraction (i.e. aspect terms have to be identified) and Sentiment Classification (i.e. classification into positive, negative and neutral with respect to each aspect term). Consider (i). "Accuracy" and "F-score" as the objective functions for sentiment classification and aspect term extraction, respectively; and (ii). Majority and Weighted voting techniques are used for combining the outputs of classifiers. For aspect term extraction and sentiment classification, consider F-score and Accuracy as the weights. Show all the steps including encoding, fitness computation, velocity updates etc with proper examples and explanations.

2+3+12

(Q3). Why is left-recursion a problem in top-down parsing? Distinguish between top-down and bottom-up parsing. Write down the various steps of bottom up probabilistic CYK parser.

3+4+8

(Q4). Discuss the impact of language divergence and structural ambiguities in Machine Translation. Formulate Statistical Machine Translation. Consider a framework for supervised coreference resolution system, where for a given mention-pair machine predicts whether they are co-referring or not. Explain with examples how training instances (positive and negative both) are created; mention the typical feature set used; and how decoding was done. Assume the learning algorithm to be decision tree, how coreference sets are detected in test scenario.

3+3+9

(Q5). This question is on HMM and PCFG, and wants you to build the theory of the latter from the former: Enumerate all the similarities between HMM and PCFG (states, transition,

observations, initial states).

(Q6)."*That former Sri Lanka skipper and ace batsman Aravinda De Silva is a man of few words was very much evident on Wednesday when the legendary batsman, who has always let his bat talks, struggled to answer a barrage of questions at a function to promote the cricket league in the city*".

For the exemple above, assume you have only 5 tags *N (noun), V (verb), J (adjective), R (adverb) and F (other, i.e., function words).* Manually PoS tag the above text. Use the convention '***word_POS***'. Give insightful answer to the question of which amongst the tags given by you will be possible for an HMM based POS tagger to tag and which not. Remember and refer to the 'NLP Layer' while answering. HMMs using very large order k will face data sparsity problem.

5+10