NLP End Sem Exam 2015

Total Marks: 85 Total Time: 3 hours

Part A

1) How can search engines be useful in the context of bootstrapping a relation extraction process? Explain using a concrete example. [3]

2) Explain the central idea behind a Word Sense Disambiguation approach that exploits knowledge of distributional similarity along with a WordNet-based relatedness measure. [2]

3) Explain the geometrical interpretation of the following linear algebraic operations: (a) a row orthogonal matrix acting on a vector (b) a column orthogonal matrix acting on a vector (c) a diagonal matrix acting on a vector. How do these three intuitions fit into the Singular Value Decomposition storyline? [5]

4) You are given a set of CFG rules along with associated rule probabilities. Identify TWO key properties that these rule probabilities must satisfy so that they define a valid PCFG. [2]

5) Prove that KLD between two distributions is non-negative. Identify clearly assumptions, if any, made in your proof. Explain in ONE sentence how this also follows from the Information Theoretic interpretation of KLD. [2+2]

6) What is the CENTRAL limitation of n-gram language models in generating realistic English sentences? Explain the limitation using a specific example. [2]

7) Give two examples of rhetorical relations. In which phase of NLG is the idea of rhetorical relations useful? [2+1]

8) In the context of Statistical Machine Translation from English to Hindi, we are interested in estimating P(h|e) where h refers to a Hindi sentence which is a candidate translation corresponding to a given English sentence e. Using the Bayesian rewrite, this amounts to estimating P(e|h) and P(h). Given a parallel corpus, what is the motivation for estimating P(e|h) instead of estimating P(h|e) directly? [2]

9) Name two measures often used to evaluate the effectiveness of a statistical parser. Explain these measures using an example. [2]

10) Give an example to show that Information Theoretic measures can capture aspects of semantic relatedness that path based measures fail to model. [2]

11) Give an example each of the following three types of ambiguity: (a) Ambiguity caused by Prepositional Phrase Attachment (b) Ambiguity caused by conjunctions (c) Ambiguity in Discourse [3]

12) We have used EM, PageRank and Factor Analysis in different NLP problem settings. Take three application domains corresponding to each of them and identify the circularity in the problem definition in each case. Show clearly how the circularity gets resolved in the formulation of the optimization problem. Establish a common ground between these approaches in terms of how they resolve uncertainty (or reduce ambiguity) based on your discussion above. [15]

13) Some NLP tasks are harder than others. In class we discussed ways of characterizing hardness. Summarize your understanding with examples. Be Precise. [4]

Part B

14) Use dynamic programming to compute the edit distance between words WRONG and WINGS, assuming that the costs of insertion and deletion are 2 and the cost of substitution is 1. Show clearly the table of subproblems.[5]

15) Given the grammar below and the input sentence "w =(()(()))", show the steps in chart parsing using CYK. Alongside your chart showing each step, mention clearly the rule(s) that is(are) used (if any) to advance to this step from the previous one. [7]

$S \rightarrow SS$

$S \rightarrow (S_1$

$S_1 \rightarrow S)$

$S \rightarrow ()$

16) Consider a Machine Translation parallel corpus having two sentence pairs. The first sentence pair is "Eat food"/"Khaanaa khaao". The second sentence pair is "Eat"/"khaao". (a) Show how the first three iterations of EM are useful in learning word alignments from this corpus. (b) Make clear any simplifying assumptions (with respect to IBM Model 3) that you use. (c) Why does EM succeed in resolving ambiguity, in case it does? [8+2+2]

17) A PCFG is based on the following rules:

a. S → NP VP
b. VP → V NP
c. VP → V NP PP
d. NP → NP PP
e. NP → I
f. NP → Amit
g. NP → a few friends
h. PP → in the conference
i. V → met

The corpus has the following two sentences, the first occurring 10 times and the second 20 times:

1. I met Amit in the conference
2. Amit met a few friends

(a) Which of these two sentences is/are ambiguous? Show all possible parse trees of these sentences.
(b) Make an APPROPRIATE initial choice of the rule probabilities. Show the first three steps of the EM algorithm for estimating the parameters of this PCFG.
(c) Does the PCFG get better at disambiguation with each step? If yes, how and why? [2+8+2]

=======The End=====