CS 6370 (NLP) – Endsem – Nov 21, 2013

Maximum marks: 50 Time: 3 hours

(Be concise. 0.5 marks may be deducted for EVERY answer that is unnecessarily verbose.)

1. How can you use the knowledge of NLP that you gathered in the course to help (a) those who are visually impaired (b) those who cannot communicate in English (c) first language learners in improving their writing (d) linguists in creating lexical resources. You will get credit only if your answer is specific and prescriptive. [2]

2. Name (a) two measures of evaluating effectiveness of statistical parsers (b) one measure of evaluating effectiveness of an IR system, where ranking of results is taken into account. (c) one measure for evaluating Machine Translation systems (d) one corpus that is mostly used in supervised WSD, and the corpus it is tagged on. [2]

3. A fresher in linear algebra comes to know from you that decomposing matrices can help search engines become smarter. She is amused. Explain in 3-4 sentences how you can help her make the connection *convincingly* using as little jargon as possible. Emphasize on the meaning of decomposing a matrix, and properties that this decomposition must satisfy to make it suitable for modeling search semantics.[2]

4. The rank of a term document matrix corresponds to the number of underlying concepts. True or false? If true, justify. If false, correct the sentence and justify the corrected version. [2]

5. Illustrate the difference between (a) collocation and co-occurrence (b) information theoretic and path based measures of semantic relatedness (c) in-vivo and in-vitro evaluation (d) link parsing and dependency parsing. [2]

6. Consider the sentence: "This is an example of little consequence." Draw the parse trees that would be generated for the sentence below by (a) a constituency parser (b) dependency parser. Make any assumptions that you make (regarding grammar rules or POS tags, for instance) explicit. [2]

7. Explain the idea behind (a) Transfer based Machine Translation (b) Explicit Semantic Analysis. Give an example in each case. [2]

8. Prove that KLD between two distributions is non-negative. Identify clearly assumptions, if any, made in your proof. Explain in ONE sentence the Information Theoretic interpretation of this result. [2]

9. Identify CLEARLY two tasks in the NLG pipeline that can benefit from Machine Learning. Give examples of each task in any domain of your choice. What would be the corpus that is used for these tasks? [2]

10. In the context of morphology, which of parsing and generation is harder? Give an example to justify your answer. [1]

11. Consider a Machine Translation parallel corpus having three sentence pairs. The first sentence pair is "come here fast"/"jaldi idhar aao". The second sentence pair is "come here"/"idhar aao". The third sentence pair is "come"/"aao". (a) Show how the

first few iterations of EM are useful in learning word alignments from this corpus. Make clear any simplifying assumptions on top of IBM Model 3. (b) How is extra knowledge "getting generated" in successive iterations of EM? (c) Can you think of a corpus where such a learning will not be as effective? Give a toy example to illustrate. [5+1+1]

12. Explain the following terms clearly with examples: (a) aggregation in the context of NLG (b) Hearst patterns [2]

13. Give an estimate for the following: (a) the number of years in took to compile WordNet (b) the number of POS tags in Penn Treebank (c) the percentage of Indian population who could potentially benefit from NLP resources/tools in Indian languages (d) accuracies of state-of-the-art POST and WSD systems. (In each of the cases above, mention just one number, and NOT a range) [2]

14. A PCFG is based on the following rules:

a. S → A B
b. B → D A
c. B → D A C
d. A → A C
e. A → a
f. A → b c
g. A → b d e
h. C → f g h
i. D → i

The corpus has the following two sentences, the first occurring 20 times and the second 40 times:

1. a i b c f g h — 20
2. b c i b d e — 40

(a) Are the sentences accepted by the grammar? In case both of them are, which of these two sentences is/are ambiguous? Show all possible parse trees of the sentence(s).

(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. [5]

15. Can IBM Model 3 be used to learn a model for converting a high level language to machine language? Be clear and specific. [1]

16. It was mentioned in class that even when labeled data is available, it is recommended that a few rounds of Baum-Welch algorithm be applied to the ML model learnt from the data. Explain why. [1]

17. How would you handle out of vocabulary words in a topic model? [1]

18. The MEMM model uses a MaxEnt classifier for learning $p(X_{i+1} | X_i, Y)$.

(a) How crucial is the choice of this classifier to the model? [1]

(b) Consider an alternative to this model, that uses a classifier that outputs a single label given the input. What are the advantages and disadvantages of this approach? [2]

19. While discussing features used in CRFs, it was mentioned that they are usually binary. What computational advantage does this give us? Are there any problems? [2]

20. One way of looking at HMMs is that they are a form of topic model with restrictions on the sampling process. Let us look at the other question. Consider a topic model with a fixed mixture distribution (a form of pLSI) over a set of topics. Give the description of a HMM that models the same probability distribution as this topic model. [2]

21. Assume that in your model of text, you have a notion of topics and probabilities of words belonging to them. Then the probability of a word occurring is the probability of it occurring given a topic (i.e p(w|t)), summed over all topics. True or False? Justify your answer.[1]

22. Explain the label bias problem. What feature of the CRF helps in overcoming that problem. [2]

23. When using a MaxEnt classifier, it is not necessary to smooth the resulting model. True or False. Explain your answer. [1]

24. Give a Bayesian interpretation for Laplace (add-one) smoothing. In other words, give a ML or MAP formalism that is equivalent to Laplace smoothing of a bigram model. [1]

$\sum p(w|t)$