CS6370 (NLP) Mid Term Question Paper (2 pages) Time: 55 mins Total Marks: 15

Name: **Roll No.:**

1. In class, we have used nonlinearity and dimensionality to characterize complexity of NLP tasks. Identify two NLP tasks, one hard and one relatively easier. Identify what non-linearity and dimensionality mean in the context of these tasks, and show clearly how the above dimensions help in qualitatively assessing their hardness relative to one another. [2 marks]

2. (a) Identify and define clearly a measure that can be used to identify stop words from a corpus of text documents. [1 mark]

   (b) Other than the fact that it eases knowledge engineering efforts, identify one advantage of an approach based on this measure over manually compiling a list of stop words. [1 mark]

   (c) Identify one disadvantage of this approach over manually compiling a list of stop words. [1 mark]

3. (a) Find the edit distance between "TRUTH" and "FRUIT" using Dynamic Programming, where the costs of insertion, deletion and substitution are 1, 1 and 2 respectively. Show clearly your table of sub-problems below. (It is suggested that you produce the table after constructing it in your rough sheet). [2.5 marks]

   (b) Identify one top-down and one bottom-up approach for estimating insertion, deletion and substitution costs. Be specific. [1 mark]

4. Give a concrete example of ambiguity at each of these levels: (a) lexical semantics (b) discourse. [1 mark]

5. There are 1000 documents in a collection, and all 1000 are ranked by a search engine G in decreasing order of their relevance to a query Q. Binary valued human relevance judgments are available on all documents wrt Q. The precision of G wrt Q is plotted as a function of rank (ranks from 1 to 1000). Precision wiggles (behaves erratically) till rank 17. In particular, precision at rank 17 > precision at rank 16; precision monotonically decreases after rank 17. Precision at rank 17 is 12/17. Compute recall at rank 16 and at rank 17, with justification. If you feel the information provided is not sufficient to answer the question, identify additional information you will need (or contradictions you need to resolve) to answer it. [1 mark]

6. The Bayesian context sensitive spellcheck algorithm discussed in class is based on a bottom-up (corpus driven) approach. However, are there some top down components that it makes use of? If so, enumerate few top down components that you can think of. [1.5 marks]

7. Your amateur friend claims that he has invented a "better" search system than Google, but he does not know how to get his idea published. He seeks advice (preferably not exceeding 250 words) from each student in the NLP class, and consults NLP TAs to select the best recommendation. What would your advice to him be? (The goal should be to identify as many different considerations as may be important for the reviewer of the paper. Use bullet points and be specific.) [3 marks]

The End