The University Of Sheffield.

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE **Autumn Semester 2008-09**

TEXT PROCESSING **2 hours**

Answer THREE questions.

All questions carry equal weight. Figures in square brackets indicate the percentage of available marks allocated to each part of a question.

Registration number from U-Card (9 digits) — to be completed by student

1. a) What is *stemming*? Briefly indicate the different kinds of approach taken to this task. Suggest a text processing context where stemming might be used. [20%]

   b) What is *part-of-speech-tagging*? Explain why this task is difficult, i.e. why it requires something more than just simple dictionary look-up. [20%]

   c) What is the noisy channel model? Give a diagram of the model as part of your answer. Suggest a text processing context where the noisy channel model has been used. [20%]

   d) Write a short program, in python or perl, which reads text from a file and prints out any lines of the file that contain, within a single line, a URL ('web address') expression. Assume the following requirements for identifying a URL string:

      i. the string should fall between double quotes (`"`)
      ii. it should start with `http://`
      iii. it should end with `.htm` or `.html`
      iv. between this start and end, there may be any sequence of characters except that ", < and > may not appear.

      [20%]

   e) Write a short program, in python or perl, that reads a file containing an unsorted list of words plus count values (one word and count value per line, separated by whitespace). The program should then print out the words in *descending order* of their associated count values. Partial credit (up to 15%) will be given for an answer that instead prints the words in *alphabetically sorted order*. [20%]

2. a) What is summarisation? Does summarisation always involve text? Identify some of the ways in which different types of summary have been classified. Discuss some of the characteristics of a good textual summary? [25%]

   b) Explain what is meant by *sentence selection* as an approach to summarisation. Discuss some of the common problems for summaries produced in this way. [15%]

   c) Discuss the criteria that have been used by *shallow* approaches to automated summarisation for identifying text segments containing significant information in a document. [20%]

   d) Consider a sentence extraction-based approach to summarisation which uses more than one of the criteria you have listed in part (d) as features in the process of sentence selection. Discuss alternative ways of combining such features in the selection process. [20%]

   e) Discuss alternative approaches to evaluating the performance of automatic summarisation systems. [20%]

3. a) Explain what an *inverted index* is, and describe the information that such indices typically encode. Why are inverted indices important in information retrieval systems? [15%]

   b) In the context of an information retrieval system that uses term frequencies in the retrieval process, show how the following two documents would be stored in an inverted index. Assume an approach which *does not* use stemming, but which *does* use a *stoplist* that includes the following terms: {`at, can, my, the, you`}.

   Document 1: Sea shells, buy my sea shells!

   Document 2: You can buy lovely sea shells at the sea produce market. [20%]

   c) What is the similarity measure used by the *vector space model* of information retrieval? Compute the similarity score between the following query and the two documents given in part (b), to determine which (if either) of them would be ranked more highly as relevant to the query. Assume that term frequency values are used for the term weights in document vectors.

   Query: buy lovely sea shells [25%]

   d) Describe and explain the TF.IDF term weighting scheme. Include the formula (or formulae) for computing TF.IDF values as part of your answer. [20%]

   e) We want to calculate the similarity of two documents using the following measure (know as the *Jacquard coefficient*): the proportion of terms appearing in *either* document that appear in *both* documents. (Note that this measure makes no use of the *counts* of terms in the document files, only whether they are present or absent.)

   Assume that we have code that identifies the terms that appear in a file, and stores this information in a dictionary (python) or hash (perl) data structure, with the terms being the keys (where the associated values are not very important, but might be either the counts of the terms in the file, or just the value 1). Write some code, in either python or perl, that takes the term information computed for *two* different files, and uses it to compute their similarity score. [20%]

4. a) Explain what is meant by an *interlingual* approach to Machine Translation. What is the key advantage of using an interlingual approach for translation amongst multiple languages, as compared to alternative approaches? Identify at least one problem that arises for the creation of a genuinely interlingual representation. [20%]

   b) Explain what is meant by *direct* and *transfer* approaches to Machine Translation, identifying the differences between them and an interlingual approach. [20%]

   c) When applied to translating from French to English, the IBM approach to statistical machine translation might be expressed by the following equation:

$$E^* = \underset{E}{\operatorname{argmax}}\ P(E) \cdot P(F|E)$$

   Explain what this equation means, and indicate the role played by the components $P(E)$ and $P(F|E)$ in the process of translation. [20%]

   d) Show how the equation given in part (c) is derived using the Bayes Rule. What is the benefit of this approach as compared to one attempting to use the probability $P(E|F)$ directly? [20%]

   e) Discuss alternative approaches that may be used for evaluating Machine Translation systems, indicating their advantages and disadvantages. Describe in some detail the approach used by the BLEU scheme for *automatic* evaluation. [20%]

END OF QUESTION PAPER