The University Of Sheffield.

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE **Autumn Semester 2012-2013**

TEXT PROCESSING **2.5 hours**

Answer the question in Section A, and THREE questions from Section B.

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

SECTION A

1. a) Explain what is meant by the *bag of words* model in text processing. Discuss the relevance and limitations of this model to *information retrieval*. [20%]

   b) What is the role of the noisy channel model in the most recent *Statistical Machine Translation* approaches (i.e., phrase-based and syntax-based SMT)? Is it still the underlying model behind such approaches? [20%]

   c) Give two reasons why it is important to evaluate the performance of text processing systems. Define the notion of a *gold-standard* dataset (also called a *reference dataset*) and explain how it differs from application to application (e.g. in Information Retrieval and Machine Translation). Differentiate extrinsic from intrinsic evaluation metrics. [20%]

   d) For *each* of the following three text processing applications, explain the most commonly used *automatic* metric for *intrinsic* evaluation: Information Retrieval, Machine Translation and Text Summarisation. Discuss any similarities between these metrics. [40%]

SECTION B

2. In the context of Information Retrieval, given the following two documents:

   Document 1: Sea shell, buy my sea shell!

   Document 2: You can buy lovely SEA SHELL at the sea produce market.

   and the query:

   Query 1: buy lovely sea shell

   a) Explain three types of manipulations (except term weighting) that can be done on document terms before indexing them. What are the advantages and disadvantages (if any) of each of these manipulations? [20%]

   b) Applying stop word removal and capitalisation, show how Document 1 and Document 2 would be represented using an *inverted index*. Provide the stoplist used. [10%]

   c) Assuming *term frequency* (TF) is used to weight terms, compute the similarity between each of the two documents (Document 1 and Document 2) and Query 1. Compute this similarity using two alternative metrics (e.g. cosine, Euclidean, dot product). Determine the ranking of the two documents according to each of these metrics and discuss any differences in the results. [25%]

   d) Discuss the expected effect of using TF.IDF to weight the terms in Document 1 and Document 2: would this be a better term weighting scheme in this example? Include the formula (or formulae) for computing TF.IDF values as part of your answer. [25%]

   e) Explain the intuition behind the PageRank algorithm. Discuss how it can distinguish two or more documents that are ranked equally "relevant" according to the similarity score given by the vector space model. [20%]

3. 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 rule-based approaches? Identify one problem that arises in the creation of a genuinely interlingual representation. [20%]

   b) Describe three main models (or features) of a standard *phrase-based* approach to statistical machine translation. Explain how these are combined. Explain how they are applied to translate a new sentence. [40%]

   c) Discuss two alternative approaches to BLEU that may be used for evaluating Machine Translation systems, indicating their advantages and disadvantages. [20%]

   d) Given the two scenarios:

      Scenario 1: English-Chinese language pair, 300,000 examples of translations.

      Scenario 2: Portuguese-Spanish, 50,000 examples of translations.

      In which of these scenarios would statistical machine translation work better and why? Would a rule-based transfer approach be better in any of these scenarios?

      [20%]

4. a) What is text summarisation? Provide three criteria according to which different types of summary have been classified and explain each of these criteria. [30%]

   b) Discuss three characteristics of a good textual summary. Provide two examples of $extrinsic$ evaluation strategies for summaries. [20%]

   c) Discuss the differences between $deep$ and $shallow$ approaches to automatic text summarisation. For each approach, give one example of an application where the approach would work well. [20%]

   d) A 'lead-based summary' is one consisting simply of the first N sentences of a document. Discuss whether this is a reasonable approach to summarisation and, if so, whether it is a useful approach either in general or in some specific context. [10%]

   e) Mention and explain two criteria that have been used by $shallow$ approaches to automated summarisation for identifying the most significant sentences of a document. Discuss how the indications from several such criteria can be combined together for sentence selection. [20%]

5. a) Sketch the algorithm for Huffman coding, i.e. for generating variable-length codes for a set of symbols, such as the letters of an alphabet. What does it mean to say that the codes produced are *prefix-free*, and why do they have this property? [30%]

   b) We want to compress a large corpus of text of the (fictitious) language *Bonobo*. The writing script of Bonobo uses only the letters {`b, i, k, n, o`} and the symbol ‿ (which is used as a 'space' between words). Corpus analysis shows that the probabilities of these six characters are as follows:

Symbol	Probability
`b`	0.25
`i`	0.05
`k`	0.06
`n`	0.07
`o`	0.45
‿	0.12

   Apply the method you described in part (a) to create a Huffman code for the Bonobo character set. [30%]

   c) Given the code you have generated in part (b), what is the average bits-per-character rate that you could expect to achieve, if the code was used to compress a large corpus of Bonobo text? How does this compare to a minimal fixed length binary encoding of this character set? [20%]

   d) Use your code for Bonobo to encode the following two messages, and compute for each message the average bits-per-character rate that results:

`bonobo‿okobo`

`iniko‿nikoni`

   Discuss why the two bits-per-character rates achieved differ, comparing them also to the expected rate that you computed in part (c). [20%]

END OF QUESTION PAPER