The University Of Sheffield.

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE **Autumn Semester 2011-2012**

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 automatic summarisation and text compression are. What do they have in common and what is the difference between them? [20%]

   b) 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%]

   c) 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%]

   d) Medline is a large collection of publications related to medicine used by researchers. It contains more than 14 million publications with over 6,000 added every day. The majority of publications are a few thousands words in length and in English, although there are also longer publications and some in other languages. Each publication is manually labelled with codes to indicate its topic (MeSH codes).

   Explain how Information Retrieval, automatic summarisation and Machine Translation could be used to assist an English speaking researcher to identify publications of interest to their research in Medline. Provide details about how effective each of these technologies is likely to be when applied to Medline. [50%]

SECTION B

2. a) Describe *manual* and *automatic* approaches to indexing for Information Retrieval. What are the advantages and disadvantages of the different approaches? [15%]

b) Consider the following documents and query:

Document 1: They sailed to the port for a good dinner and sailed home after.

Document 2: Ruby port is very good after a meal, any meal.

Query: good ports after dinner

(i) Show how Documents 1 and 2 would be stored in an inverted index, using stemming and the following list of stopwords: {`a, and, any, for, is, the, they, to` } [20%]

(ii) Compute the similarity between the query and each document using the cosine metric and using term frequency values for the term weights in the document vectors. [30%]

(iii) Explain why the vector space model for Information Retrieval would not identify Document 2 as the top ranked document for the query. [15%]

c) Define the precision and recall measures used in Information Retrieval. Explain why it is easier to compute precision than recall for web-based Information Retrieval systems such as Google or Bing. [20%]

3. a) Describe the three main approaches to Machine Translation and explain the differences between them. [20%]

b) Describe the main features of the statistical approach to Machine Translation. [30%]

c) Give an example of a language pair for which Statistical Machine Translation is likely to work well and another example for which is it likely to work badly. Explain your choices with reference to the approach used by Statistical Machine Translation systems. [20%]

d) Describe how Round-trip translation can be used to evaluate Machine Translation systems. What are the advantages and disadvantages of this approach? [15%]

e) Describe the approach used by the BLEU system for evaluation of Machine Translation systems. [15%]

4. a) What is summarisation? Does summarisation always involve text? [10%]

b) Outline three of the different types of ways in which summaries have been classified and explain how this effects the properties of the summary that should be produced by an automatic summariser. [15%]

c) What is meant by *deep* and *shallow* approaches to automatic text summarisation? Discuss the differences between them. [15%]

d) Discuss the various features which have been used by *shallow* approaches to automatic summarisation and explain why they are useful for identifying segments of text containing important information in a document. [20%]

e) Consider the following short document and summaries. Discuss the quality of the two summaries with particular reference to coherence and cohesion.

> **Original document**: John went to the park last week. He took his daughter, Sally. She likes going to the park. John also took his dog Fido. Fido also likes going to the park. He likes to chase the ducks.
>
> **Summary 1**: John went to the park last week. She likes going to the park.
>
> **Summary 2**: John went to the park last week. He likes to chase the ducks.

[15%]

f) Explain the difference between *intrinsic* and *extrinsic* approaches to the evaluation of summaries, providing details about the two approaches. [25%]

5. a) The fictitious language *Hartish* uses a script that consists of 7 characters: `a, c, e, g, i, l, s`. Corpus analysis shows that the probabilities of these seven characters are as follows:

Symbol	Probability
`a`	0.20
`c`	0.25
`e`	0.10
`g`	0.05
`i`	0.22
`l`	0.07
`s`	0.11

(i) Show how to construct a Huffman code tree for Hartish, given the above probabilities for its characters. [25%]

(ii) Use your code tree to encode the message "`gailsails`" and show the resulting binary encoding. How many bits are required to encode the message? [10%]

(iii) How many bits are required to encode the message "`gailsails`" using a minimal fixed length binary character encoding for a 7 character alphabet? Explain why fewer bits are required to compress this string using a minimal fixed length binary character encoding than using Huffman encoding. [15%]

b) Explain the difference between **static**, **semi-static** and **adaptive** techniques for text compression, noting their key advantages and disadvantages. [15%]

c) (i) Explain how the LZ77 compression method works. [25%]

(ii) How would the following LZ77 encoder output

$$\langle 0,0,d\rangle\langle 0,0,o\rangle\langle 0,0,m\rangle\langle 2,1,d\rangle\langle 3,2,o\rangle\langle 3,3,d\rangle\langle 12,2,o\rangle\langle 0,0,m\rangle$$

be decoded, assuming the encoding representation described in the lectures for this module (Text Processing)? Show how your answer is derived. [10%]

END OF QUESTION PAPER