The
University
Of
Sheffield.

Data Provided: None

PLEASE LEAVE THIS EXAM PAPER ON YOUR DESK.
DO NOT REMOVE IT FROM THE HALL.

DEPARTMENT OF COMPUTER SCIENCE **September/October 2014**

TEXT PROCESSING **2 hours and 30 minutes**

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.

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

SECTION A

1. a) Describe and explain the TF.IDF term weighting scheme used in information retrieval. Include the formula (or formulae) for computing TF.IDF values as part of your answer. [30%]

   b) Compression techniques are important due to the growth in volume of the data that must be stored and transmitted.

      (i) Explain the difference between **lossy** and **lossless** forms of compression. Discuss the suitability of these alternative forms of compression for different media types (e.g. for text vs. image data). [10%]

      (ii) Explain the difference between **static**, **semi-static** and **adaptive** techniques for text compression, noting their key advantages and disadvantages. [10%]

   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. [30%]

   d) Explain the **two** most common (semi-)automated approaches to expand sets of seed opinion words (like "good" and "bad") with more opinion words to create lexica for Sentiment Analysis: dictionary-based and corpus-based approaches. Give **one** advantage and **one** disadvantage of each approach. [20%]

SECTION B

2. a) Describe *manual* and *automatic* approaches to indexing for Information Retrieval. Discuss the advantages and disadvantages of these two approach? [20%]

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

b) 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%]

c) 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%]

d) 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 metrics: Cosine and Euclidean. Determine the ranking of the two documents according to each of these metrics and discuss any differences in the results. [25%]

e) 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%]

3. a) Explain the three main approaches to Machine Translation: *direct*, *transfer* and *interlingual*. [20%]

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

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

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%]

c) Describe the following metrics for evaluating Machine Translation systems: BLEU and round-trip translation. Discuss the advantages and disadvantages of automatic evaluation metrics like these two over manual evaluation metrics. [20%]

d) 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%]

e) Explain the difference between Hierarchical Phrase-based Machine Translation models and standard Phrase-based Statistical Machine Translation models. Do Hierarchical Phrase-based Machine Translation models use linguistic information, and if so, of what type? [20%]

4. a) Given sentences like the following:

- *My new phone works well and it is much faster than the old one.*
- *My new phone has 32GB of memory and can play videos.*

What is the first step to detect the sentiment in these two sentences? Should both these sentences be addressed in the same way by sentiment analysis approaches? [10%]

b) Explain a common approach for subjectivity analysis. [10%]

c) Discuss the relevance of automatic techniques for sentiment analysis for marketing purposes. [10%]

d) Given the following sentences and opinion lexicon (adjectives only), apply the weighted lexical-based approach to classify EACH sentence as **positive**, **negative** or **objective**. Show the final emotion score for each sentence, but also how it was generated. In addition to use of the lexicon, make sure you consider any general rules that have an impact in the final decision. Explain these rules when they are applied. [20%]

Lexicon:

awesome	5
boring	-3
brilliant	2
funny	3
happy	4
horrible	-5

(S1) He is brilliant and funny.
(S2) I am not happy with this outcome.
(S3) I am feeling AWESOME today, despite the horrible comments from my supervisor.
(S4) He is extremely brilliant but boring, boring, very boring.

e) According to Bing Liu's model, an **opinion** is said to be a quintuple $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$. Explain each of these elements and exemplify them with respect to the following text. Identify the features present in the text, and for each indicate its sentiment value as either *positive* or *negative*. Discuss two language processing challenges in automating the identification of such elements. [30%]

> "I have just bought the new iPhone 5. It is a bit heavier than the iPhone 4, but it is much faster. The camera lenses are also much better, taking higher resolution pictures. The only big disadvantage is the cost: it is the most expensive phone in the market. Lucia Specia, 12/08/2014."

f) Explain three metrics to evaluate the quality of binary (negative/positive) sentiment analysis systems. Give their intuitions and show their formulae. [20%]

5. a) Text compression techniques are important because growth in volume of text continually threatens to outstrip increases in storage, bandwidth and processing capacity. Briefly explain the differences between:

(i) **symbolwise** (or statistical) and **dictionary** text compression methods; [10%]

(ii) **modelling** versus **coding** steps; [10%]

b) The script for the fictitious language Gavagese contains only the 7 characters a, e, u, k, r, f, d. You assemble a large electronic corpus of Gavagese and now want to compress it. You analyse the frequency of occurrence of each of these characters in the corpus and, using these frequencies as estimates of the probability of occurrence of the characters in the language as a whole, produce the following table:

Symbol	Probability
a	0.25
e	0.20
u	0.30
k	0.05
r	0.07
f	0.08
d	0.05

(i) Show how to construct a Huffman code tree for Gavagese, given the above probabilities. [30%]

(ii) Use your code tree to encode the string $dukerafua$ and show the resulting binary encoding. For this string, how much length does your codetree encoding save over a minimal fixed length binary character encoding for a 7 character alphabet? [10%]

c) One popular compression technique is the LZ77 method, used in common compression utilities such as *gzip*.

(i) Explain how LZ77 works. [25%]

(ii) How would the following LZ77 encoder output

$$\langle 0,0,b\rangle\langle 0,0,a\rangle\langle 0,0,d\rangle\langle 3,3,b\rangle\langle 1,3,a\rangle\langle 1,3,d\rangle\langle 1,3,a\rangle\langle 11,2,a\rangle$$

be decoded, assuming the encoding representation presented in the lectures? Show how your answer is derived. [15%]

END OF QUESTION PAPER