The University Of Sheffield.

**Data Provided: None**

**DEPARTMENT OF COMPUTER SCIENCE** **Autumn Semester 2015-2016**

**TEXT PROCESSING** **2 hours 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.**

## 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) Explain the intuition of the noisy channel model in the context of Statistical Machine Translation (SMT). Discuss whether and how this model is used in modern SMT approaches (i.e., phrase-based and syntax-based SMT). [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). [20%]

   d) For *each* of the following three text processing applications, give a commonly used *automatic* metric for *intrinsic* evaluation: Information Retrieval, Machine Translation and Sentiment Analysis. 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 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 metrics: Euclidean and cosine. Determine the ranking of the two documents according to each of these metrics and discuss any differences in the results. [40%]

d) Explain TF.IDF. Include the formula (or formulae) for computing TF.IDF values as part of your answer. 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 than TF in this example?

[30%]

3. a) Explain the differences between *rule-based* and *empirical* approaches to Machine Translation. Give the main advantage and disadvantage of each of these approaches. [20%]

b) Describe the two main models of a standard *phrase-based* approach to statistical machine translation. Explain how these models are combined and how they are applied to generate a translation for a new segment. [30%]

c) Explain the intuition behind the IBM Model 1 in the context of Statistical Machine Translation (SMT). Give the most important outcome of this model for an SMT system. Give one direction in which this model can be improved. [30%]

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 a good solution in any of these scenarios? [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? [20%]

b) Explain two approaches for Subjectivity Analysis. [20%]

c) Explain the weighted lexical-based approach for Sentiment Analysis. Given the following sentences and opinion lexicon, apply this approach to classify *each* sentence in S1-S4 as **positive**, **negative** or **objective**. Show the final emotion score for each sentence and also how this score was generated. Give any general rules that you used to calculate this score as part of your answer. Explain these rules when they are applied. [30%]

**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.

d) Give Bing Liu's model for an **opinion**. Explain each of the elements in the model 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, 12/10/2015."

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 5(a) to create a Huffman code for the Bonobo character set. [30%]

c) Given the code you have generated in 5(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 5(c). [20%]

**END OF QUESTION PAPER**