The University Of Sheffield.

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE Autumn Semester 2013-2014

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.

SECTION A

1. a) In the context of Information Retrieval, explain the difference between algorithms that perform boolean search and algorithms that perform a ranked search. What type of algorithm would be better for a regular user (such as an undergraduate student in the Humanities area) who is using a search query with multiple terms, which he/she expects to appear in many documents? Explain the reasons behind your choice of algorithm type. [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) The two main model components in Statistical Machine Translation are the *Translation Model* and the *Language Model*. Explain the role of each of these components. Describe the type of data that is necessary to build each of them. Mention one way in which these components can be combined to build a translation system. [30%]

   d) Assume we have a small set of seed words with positive and negative opinions, e.g.: *positive* = {*good, fast, cheap*} and *negative* = {*slow, boring, fragile*}. Explain the **two** most common (semi-)automated approaches to expand these sets with more opinion words or phrases to create lexica for Sentiment Analysis, providing examples whenever possible. Give **one** advantage and **one** disadvantage of each approach. [20%]

SECTION B

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

   Document 1: Sea shell, buy my sea shell!

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

   Document 3: Product marketing in the Shelly sea is an expensive market.

   and the query:

   Query 1: sea shell produce market

   a) Apply the following term manipulations on document terms: *stoplist removal*, *capitalisation* and *stemming*, showing the transformed documents. Explain each of these manipulations. Provide the stoplist used, making sure it includes punctuation, but no content words. [20%]

   b) Show how Document 1, Document 2 and Document 3 would be represented using an *inverted index* which includes term frequency information. [10%]

   c) Using *term frequency* (TF) to weight terms, represent the documents and query as vectors. Produce rankings of Document 1, Document 2 and Document 3 according to their relevance to Query 1 using two metrics: Cosine Similarity and Euclidean Distance. Show which document is ranked first according to each of these metrics. [30%]

   d) Explain the intuition behind using TF.IDF (*term frequency inverse document frequency*) to weight terms in documents. Include the formula (or formulae) for computing TF.IDF values as part of your answer. For the ranking in the previous question using cosine similarity, discuss whether and how using TF.IDF to weight terms instead of TF only would change the results. [20%]

   e) Explain the metrics Precision, Recall and F-measure in the context of evaluation in Information Retrieval against a gold-standard set, assuming a boolean retrieval model. Discuss why it is not feasible to compute recall in the context of searches performed on very large collections of documents, such as the Web. [20%]

3. a) List and explain the three paradigms of Machine Translation. What is the dominant (most common) paradigm for open-domain systems nowadays and why is this paradigm more appealing than others, especially in scenarios such as online Machine Translation systems? [20%]

b) Lexical ambiguity is known to be one of the most challenging problems in any approach for Machine Translation. Explain how this problem is addressed in Phrase-based Statistical Machine Translation approaches. [20%]

c) List and explain two metrics that can be used for evaluating Machine Translation systems (either manually or automatically). Discuss the advantages of automatic evaluation metrics over manual evaluation metrics. [20%]

d) Given the two scenarios:

> **Scenario 1**: English-Arabic language pair, 50,000 examples of translations of very short sentences, on very repetitive material (technical documentation of a product).
>
> **Scenario 2**: English-French, 500,000 examples of translations for open-domain and creative texts, like novels from many different writers.

In which of these scenarios would Statistical Machine Translation work better? Why would it work better than in the other scenario? [10%]

e) Explain the main advantage of Hierarchical Phrase-based Machine Translation models over standard Phrase-based Statistical Machine Translation models. What does the phrase table of Hierarchical Phrase-based Machine Translation models look like? Given the following sentence pair and the existing phrases (in the phrase-table), which additional phrases could be generated with a Hierarchical Phrase-based Machine Translation model?

> Source: shall be passing on to you some comments
> Target: werde Ihnen die entsprechenden Anmerkungen aushäandigen

Existing phrases:

Source	**Target**
shall be	werde
passing on	aushäandigen
to you	Ihnen
some comments	die entsprechenden Anmerkungen
to you some comments	Ihnen die entsprechenden Anmerkungen

[30%]

4. a) Differentiate *subjectivity* from *sentiment*. How are the tasks of Subjectivity Classification and Sentiment Analysis related? [10%]

   b) Explain the steps involved in the lexicon-based approach to Sentiment Analysis of features in a sentence (e.g. features of a product, such as the *battery* of a mobile phone). Discuss the limitations of this approach. [20%]

   c) 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. 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:

boring	-3
brilliant	2
good	3
horrible	-5
happy	5

(S1) He is brilliant but boring.
(S2) I am not good today.
(S3) I am feeling HORRIBLE today, despite being happy with my achievement.
(S4) He is extremely brilliant but boring, boring.

   d) Specify the five elements of Bing Liu's model for Sentiment Analysis, 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 am in love with my new Toshiba Portege z830-11j. With its i7 core processors, it is extremely fast. It is the lightest laptop I have ever had, weighting only 1 Kg. The SSD disk makes reading/writing operations very efficient. It is also very silent, the fan is hardly ever used. The only downside is the price: it is more expensive than any Mac. Lucia Specia, 10/04/2012."

   e) Differentiate *direct* from *comparative* Sentiment Analysis. What are the elements necessary in comparative models of Sentiment Analysis? [20%]

5. a) (i) Explain how the LZ77 compression method works. [30%]

(ii) Assuming the encoding representation presented in class (i.e. in the lectures of the Text Processing module), show what output would be produced by the LZ77 decoder for the following representation. Show how your answer is derived. [15%]

$$\langle 0, 0, b\rangle\, \langle 0, 0, e\rangle\, \langle 2, 2, n\rangle\, \langle 4, 4, e\rangle\, \langle 1, 3, b\rangle\, \langle 2, 1, n\rangle$$

b) The writing script of the (fictitious) language *Sinbada* employs only the letters (`s,i,n,b,a,d`) and the symbol $\sim$, used as a 'space' between words. Corpus analysis shows that the probabilities of these seven characters are as follows:

Symbol	Probability
`s`	0.04
`i`	0.1
`n`	0.2
`b`	0.04
`a`	0.3
`d`	0.26
$\sim$	0.06

(i) Sketch the algorithm for Huffman coding. Illustrate your answer by constructing a code tree for Sinbada, based on the above probabilities for its characters. [30%]

(ii) Given the code you have generated in 5(b)(i), 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 Sinbada text? [10%]

(iii) Use your code tree to encode the message "`niad`$\sim$`badasina`" and show the resulting binary encoding. How does the bits-per-character rate achieved on this message compare to the rate that you calculated in 5(b)(ii)? [15%]

END OF QUESTION PAPER