Two hours

Formula Sheet attached for use with Question 4.

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Text Mining

Date: Thursday 29th May 2014
Time: 14:00 - 16:00

Please answer any THREE Questions from the FIVE Questions provided

Each question is worth 20 marks

Clearly cross out anything you do not wish to be marked.

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text
1. a) Classify the following examples according to the type of ambiguity they display:
   
i) Until the police arrest the drug dealers controlled the street.
   
ii) He drew a card.
   
iii) The car hit the tree while it was moving.
   
iv) The boy chased the dog with a bone.
   
   (2 marks)

b) In part of speech tagging, we note that when context disambiguates the part of speech of a word, that context is usually very local to the word, e.g., the preceding or the following word.

   Give two different examples of local context helping to disambiguate a noun/verb ambiguity with a strong likelihood.
   
   (1 mark)

c) In relation to Brill’s Transformation-based Learning (TBL) algorithm:

   i) Explain the steps taken by the TBL algorithm, commenting in particular on how the lexicon and transformation rules are acquired.
   
   (2 marks)

   ii) Brill noted: “We found it a bit surprising that the addition of lexicalized transformations did not result in a much greater improvement in performance.”.

   Explain the difference between lexicalized and non-lexicalized transformations.
   
   (1 mark)

   State why, when using a corpus for training, you might reasonably expect only a small gain in performance on a test corpus, for lexically-based data.
   
   (1 mark)

d) Explain the role of a treebank in the construction of probabilistic context-free grammars (PCFGs).
   
   (2 marks)

e) The notion of head of a syntactic construction is an important one in Linguistics. How is this notion of relevance in the world of PCFGs? Illustrate your answer with appropriate examples.
   
   (2 marks)

   [Question 1 continues on the following page]
Consider the following grammar and lexicon (numbers are given for reference only):

1) S -> NP VP  
2) NP -> N  
3) NP -> N N  
4) NP -> N NNS  
5) VP -> VG  
6) VP -> VG NP  
7) VG -> V  
8) V -> reduces  
9) V -> cause  
10) V -> floods  
11) N -> river  
12) N -> flood  
13) N -> cause  
14) N -> misery  
15) NNS -> floods

i) Show, by constructing two parse trees, that the string “river floods cause misery” is ambiguous according to the above grammar. (2 marks)

ii) Show the steps that a naïve top-down (goal driven) parser would take in parsing the string “river flood reduces” according to the above grammar. (2 marks)

iii) Compare and contrast top-down parsing with bottom-up parsing, noting the advantages and disadvantages of each approach. (2 marks)

iv) Assume that the Earley algorithm has been applied to parse the string “river flood reduces” with the grammar and lexicon shown above. The parser has paused at the state shown above as a chart. Reproduce this chart, showing the next seven edges (representing scanning, prediction or completion steps) that would be produced by the algorithm. (3 marks)
2.  
   a) How does the task of information extraction differ from that of parsing?  
      (1 mark)  
   b) What is dependency parsing and why is it considered more appropriate as a  
      contribution towards information extraction than phrase structure parsing?  
      (3 marks)  
   c) Consider the following sentence:  
      Tyre maker Bridgestone has pleaded guilty to price-fixing and agreed to pay a $425m  
      fine to the US Department of Justice (DOJ).  
      Sketch how the key event of a corporate misdemeanour expressed in this sentence  
      could be represented in an information extraction system like FASTUS or in a MUC-  
      7 scenario template.  
      (4 marks)  
   d) The text extract below is of training data used with a sequence classifier to create a  
      named entity recognition system. Refer to it in the following sub-parts.  

A DT UPPERCASE O O  
State NNP CAPITALIZED O B-ORGANIZATION  
Department NNP CAPITALIZED O I-ORGANIZATION  
spokesman NN LOWERCASE O O  
said VBD LOWERCASE B-PERSON_SUFFIX_CLUE O  
Wednesday NNP CAPITALIZED B-DAY B-DATE  
night NN LOWERCASE B-HOURS B-TIME  
that IN LOWERCASE O O  
Brown NNP CAPITALIZED B-CITY B-PROVINCE B-PERSON  
was VBD LOWERCASE O O  
assumed VBN LOWERCASE O O  
dead NN LOWERCASE O O  
, , PUNCT O O  
although IN LOWERCASE O O  
officially RB LOWERCASE O O  
the DT LOWERCASE O O  
passengers NNS LOWERCASE O O  
were VBD LOWERCASE O O  
still RB LOWERCASE O O  
listed VBN LOWERCASE O O  
as IN LOWERCASE O O  
missing VBG LOWERCASE O O  
. . PUNCT O O  
And CC CAPITALIZED O O

[Question 2 continues on the following page]
i) What is the role of a sequence classifier in text processing? What output does such a classifier produce when used for training? What are the inputs and output when used for annotating text?

(2 marks)

ii) Sequence classifiers may differ in input data format. What are the essential requirements of the MALTET CRF sequence classifier input format as seen in the above text extract?

(2 marks)

iii) Briefly explain how the data in each (space-delimited) field is obtained, referring to components in a typical modular architecture for text processing, and commenting on the use of the prefixes “B-” and “I-”.

(4 marks)

iv) What does the above extract show about the necessity and sufficiency of gazetteer lookup as a technique or component for the named entity recognition task?

(2 marks)

v) Show how the text and the target named entity annotations within the first 12 lines of the example would be represented in the MUC-7 SGML in-line annotation format.

(2 marks)
3. a) Design a simple UIMA type system that enables capturing events related to law and order found in news articles, and enables annotating named entities and relationships between these entities. Follow the specifications below.

The categories of events are predefined and are limited to: Arrest, Conviction and Appeal. Similarly, the categories of named entities are limited to: Person, Organisation and Place.

An event must be anchored to a word or phrase that triggers the event. It should also list the actor, target, and location of the event, each of which should refer to one of the named entities as follows: actor and target to mentions of either Person or Organisation, and location to Place.

Additionally, each event must have an indication whether its polarity is positive or negative. For example, in the sentence “Police failed to apprehend John Smith in London” it should be possible to make the following annotations using your type system:

For each UIMA type, clearly state its name, supertype, and a list of features (do not repeat the supertype’s features). For each feature, indicate its name and type. You are allowed to draw a UML diagram instead.

Reuse the UIMA built-in type uima.tcas.Annotation that includes the features begin and end, both of type Integer, that indicate the begin and end character positions of a span of text. You are not required to include namespaces for any of the types.

Hint: There are several solutions, but the most straightforward involves 8 (or 9) types, of which only 1 defines additional (i.e., not inherited) features.

(6 marks)
b) Consider the following elementary processing components:

(1) Gold standard corpus reader: Deserialises a given corpus as input and produces each document and related gold annotations as output.
(2) Reference evaluator: Reports annotation effectiveness comparing two inputs of which one is indicated to be a reference (gold data) input. The report is saved to a file and includes common performance metrics.
(3) Chunker
(4) Part-of-speech tagger
(5) Annotation remover
(6) Sentence splitter
(7) Tokeniser
(8) Named entity recogniser

Design a workflow, by drawing a diagram, that would allow you to evaluate the combined effectiveness of the components (3) to (8) against a given gold standard corpus. Explain your workflow by indicating the role, expected input and produced output for each of the components, as is done above for components (1) and (2).

(3 marks)

c) Describe faceted search and explain how it may be used to facilitate semantic search.

(2 marks)

d) For each of the components of an RDF triple, indicate which of the following values it may take: URI resource, literal.

(1 mark)

e) Consider the following RDF knowledge base (in Turtle format):

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<http://www.w3.org/People/Berners-Lee/card#i>  
  rdf:type foaf:Person ;  
  foaf:name "Timothy Berners-Lee" ;  
  foaf:mbox <mailto:timbl@w3.org> ;  
  foaf:workplaceHompage <http://www.w3.org> ;  
  foaf:knows <http://www.w3.org/People/EM/contact#me> .
```

[Question 3 continues on the following page]
[Question 3 continues from the previous page]

<http://www.w3.org/People/EM/contact#me>
  rdf:type foaf:Person ;
  foaf:name "Eric Miller" ;
  foaf:mbox <mailto:em@w3.org> ;
  foaf:workplaceHomepage <http://www.w3.org> .

For each of the SPARQL queries below assume the following prefix definitions exist:
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix foaf: <http://xmlns.com/foaf/0.1/>

i) What is the result of running the following SPARQL query?

```sparql
SELECT ?name ?email
WHERE {
  ?x foaf:name ?name .
  ?x foaf:mbox ?email .
}
```

(2 marks)

ii) Write a SPARQL query that returns names of persons known to Timothy Berners-Lee. In your query, you are allowed to use the literal “Timothy Berners-Lee” but not the URI <http://www.w3.org/People/Berners-Lee/card#i>.

(3 marks)

f) Consider the following contingency table that shows how many times two human annotators agreed and disagreed on the categorisation of documents into relevant and irrelevant:

<table>
<thead>
<tr>
<th>Annotator 2 Results</th>
<th>Annotator 1 Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>relevant</td>
<td>35</td>
</tr>
<tr>
<td>irrelevant</td>
<td>2</td>
</tr>
<tr>
<td>relevant</td>
<td>1</td>
</tr>
<tr>
<td>irrelevant</td>
<td>12</td>
</tr>
</tbody>
</table>

Using the above values and the kappa statistic, calculate the inter-annotator agreement rate. Show your working.

Hint: The formula for the kappa statistic is: \((P(a) – P(e))/(1 – P(e))\), where \(P(a)\) is the observed proportion of times the annotators agreed, and \(P(e)\) is the proportion of times they would be expected to agree by chance.

(3 marks)
4.  
a) Give three examples of lexical relations that may be found in the Princeton WordNet.  
   (1 mark)

b) Briefly explain how two of the relations you have identified may be useful in 
   automatic text analysis.  
   (2 marks)

c) What features would you consider in supervised mode for training a system to carry 
   out Word Sense Disambiguation using a machine learning algorithm? Justify your 
   choices.  
   (2 marks)

d) Discuss limitations of supervised Word Sense Disambiguation techniques and the 
   advantages and disadvantages of techniques proposed to overcome such limitations.  
   (2 marks)

e) Briefly state your understanding of “multiword expression”, “collocation” and 
   “multiword domain term” and then discuss to what extent it is useful to distinguish 
   among these, for text mining purposes, and to what extent methods used to discover 
   instances of one type are useful in discovering instances of the other two.  
   (2 marks)

f) For this question part, consult the provided formula sheet.

   A corpus consists of 20000 bigrams, where new occurs in 400 bigrams and York 
   occurs in 50 bigrams. The bigram new York occurs 25 times.

   i) Compute the pointwise mutual information of the bigram new York.  
      (2 marks)

   ii) Decide if the co-occurrence of new and York is random or not using the t test. 
       Show your working. (The critical value for a confidence level \( \alpha = 0.005 \) is 
       2.576.)  
       (3 marks)

   iii) Compute the observed values contingency table of the \( X^2 \) (chi-square) test.  
        (1 mark)

   iv) Compute the expected values contingency table of the \( X^2 \) (chi-square) test.  
       (2 marks)

   v) Decide if the co-occurrence of “new” and “York” is random or not using the \( X^2 \) 
      (chi-square) test. Show your working. (For 1 degree of freedom and at a 
      probability level of \( \alpha = 0.05 \) the critical value is 3.841.)  
      (3 marks)
5.  
   a) Show, with examples, how the lexicon, syntax, rhetoric and pragmatics are all involved in explaining how subjective sentiment is conveyed in natural language.  
      (6 marks)
   
   b) What is meant by aspect-based sentiment analysis in the context of reviews of consumer products? What resources and analysis components would be needed to develop an aspect-based sentiment analysis application?  
      (4 marks)
   
   c) Briefly set out what you consider to be the major current challenges for text mining, and then discuss in detail what if any progress we may expect towards meeting these challenges in the near to medium future. Justify your views and conclusions, giving appropriate examples to back up your arguments.  
      (10 marks)

END OF EXAMINATION