Text Mining

Date:  Wednesday 16th May 2012
Time:    14:00 - 16:00

Please answer any THREE Questions from the FIVE questions provided

Each question is worth 20 marks

Do not answer more than the required number of questions:
Additional answers will not be marked

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) FOOT HEADS ARMS BODY.

iii) I saw the woman by the river under the bridge with the telescope.

iv) I bought a chair from an admiral with carved legs.

(2 marks)

b) You are asked to write a tokeniser for a text mining system. Discuss tokenisation issues you may expect to arise for this task, with appropriate examples, explaining the impact that certain tokenisation decisions may have on later processing stages.

(2 marks)

c) Would you expect text mining to be more successful with publishers’ XML documents or plain text extracted from PDF? Justify your answer by giving examples of the difficulties that either format may pose to a sentence splitter.

(2 marks)

d) Consider the following text extract, in which the underlined portions are considered to denote entities that are either proteins (PRO) or species (SPE):

*Latent membrane protein 1 (LMP1) is an important oncogenic protein encoded by Epstein-Barr virus (EBV).*

Show how the text may be represented in either B-I-O notation or B-I-L-O-U notation, for both syntactic chunking and semantic (or named entity) tagging.

(4 marks)

e) i) Give three examples of lexical relations that may be found in the Princeton WordNet.

(2 marks)

ii) Briefly explain how one or more of the relations you have identified may be useful in automatic text analysis.

(2 marks)

f) What information about verb usage may be obtained from the WordNet that is useful in parsing?

(2 marks)
g) Briefly distinguish between inflectional morphology and derivational morphology with examples drawn from English. (2 marks)

h) Briefly explain the relevance of morphology to text processing tasks such as term discovery and information retrieval. (2 marks)
2.

a) For a word set \(<w>\) and its corresponding tag set \(<t>\), the most likely tag sequence is given by

\[
\hat{T} = \arg\max \sum P(t_i|t_{i-1})P(w_i|t_i)
\]

From a Hidden Markov Model (HMM) point of view, what do \(P(t_i|t_{i-1})\) and \(P(w_i|t_i)\) stand for? (2 marks)

b) What algorithm could one use to train a model and decode the data in HMMs? (2 marks)

c) Briefly explain the strengths and weaknesses of HMMs. (2 marks)

d) i) What are the main characteristics of Maximum Entropy Markov Models (MEMMs)? (2 marks)

ii) Explain what is meant by the *labelling bias problem.* (1 mark)

e) Discuss, with examples, what features one could use to identify part of speech (POS) tags for the following sentence:

Tom and Jerry went up the hill. (3 marks)

f) What are the key differences between Maximum Entropy Markov Models and Conditional Random Fields? (3 marks)

[Question 2 continues on the following page]
g)  

i) How does the transformation-based error-driven learning algorithm devised by Brill repeatedly use the difference between
   - a given tag assignment to each token of the text and
   - the ‘gold-standard’ tagging of the text
   in order to induce a sequence of transformation rules?  
   (3 marks)

ii) Briefly discuss advantages and disadvantages of the algorithm.  
    (2 marks)
3.

a) Suggest an application where chunking would be more appropriate than (full) parsing. Justify your suggestion.

(b) Consider the following grammar and lexicon:

\[
\begin{align*}
S & \rightarrow N \ VP \\
VP & \rightarrow V \ VP \\
VP & \rightarrow V \ N \\
VP & \rightarrow V \ N \ N \\
N & \rightarrow \text{they} \\
N & \rightarrow \text{hunting} \\
N & \rightarrow \text{dogs} \\
V & \rightarrow \text{are} \\
V & \rightarrow \text{hunting}
\end{align*}
\]

i) Show, by constructing two parse trees, that the string "they are hunting dogs" is ambiguous according to the above grammar.

ii) Similarly, show that the string “they are hunting dogs” is ambiguous according to the above grammar, by constructing two dependency graphs using the following dependency labels: subj (subject), obj (object), nmod (noun modifier), aux (auxiliary verb). Hint: nmod and aux are not used in the same graph.

iii) The Earley algorithm has been applied to parse the string “they are hunting dogs” with the grammar and lexicon shown above. The parser has paused at the state shown below as a chart. Reproduce this chart, showing the next six edges (representing scanning, prediction or completion steps) that would be produced by the algorithm.

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

c) What motivates the use of a probabilistic context-free grammar (PCFG)? (2 marks)

d) Briefly explain how a PCFG is constructed from a Treebank, stating also the difference between lexicalized and unlexicalized PCFGs? (2 marks)

e) Compare and contrast a phrase structure tree (parse tree) and a dependency tree, discussing why the latter is often preferred for applications in text mining. (2 marks)

f) What is the advantage of feature-based grammars (using features such as number, person) over pure CFGs? (2 marks)

g) In information extraction, what is the difference between relations and events? Give examples. (2 marks)

h) What are the disadvantages of using publishers’ machine-readable dictionaries for word sense disambiguation? (2 marks)
4. 

a) What is the advantage of $X^2$ (chi-square) test over $t$ test? (1 mark)

b) 

i) What is the advantage of likelihood ratios over $X^2$ (chi-square) test? (1 mark)

ii) State the likelihood ratio dependence and independence hypotheses when deciding upon the significance of a collocation. (1 mark)

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

An economics corpus consists of 100000 bigrams, where commercial occurs in 2000 bigrams and bank occurs in 500 bigrams. The bigram commercial bank occurs 20 times.

i) Compute the pointwise mutual information of the bigram commercial bank. (2 marks)

ii) State the $t$ test null hypothesis. (1 mark)

iii) Decide if the co-occurrence of commercial and bank is random or not using the $t$ test. Show your working. (The critical value for a confidence level $\alpha = 0.005$ is 2.576.) (2 marks)

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

v) Compute the expected values contingency table of the $X^2$ test. (1 mark)

vi) Decide if the co-occurrence of commercial and bank is random or not using the $X^2$ 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)
d)  
  i) The following are the beginning and end postcodes for Mytown:
      M1 1AA  
      M99 9ZZ  

      Write a regular expression to match all Mytown postcodes.  

      (1 mark)  

  ii) Consider the following text:

      Today BigChocoBiscuits, plc (MSE: BCB) announced the acquisition of
      Much More Yummy, Ltd., following its failure to acquire The Really
      Yummy Company. British BiscTec (MSE: BBT; NYSE: BRB) announced that
      Will Orwonthe, PhD, has resigned as Chief Scientific Officer.
      BiscTec, headquartered in Manchester, employs 200 food scientists,
      many graduates of the Universities of Manchester and Salford.  

      Propose patterns for rules that would maximise the extraction of
      organisations from this text.  

      (3 marks)  

e) Compare and contrast mutual information with the C-Value measure, discussing to
what extent these measures are useful in term recognition.  

      (3 marks)
5.

a) Consider the following partially completed diagram:

```
      A
     /|
    / | B
   /  |
  /   |
 /    |
```

The outer ellipse represents a text corpus. Circle A represents the gold standard annotations for the entity PERSON in this corpus. Circle B represents the result of a named entity recogniser attempting to annotate the corpus for PERSON.

Reproduce this diagram and clearly label appropriate parts to indicate:
- True positives
- False positives
- True negatives
- False negatives

(2 marks)

b) Two annotators annotate a corpus for PERSON entity instances. We calculate how many times they agree/disagree, and obtain the following table:

```

<table>
<thead>
<tr>
<th>Annotator 2 Results</th>
<th>yes</th>
<th>no</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>300</td>
<td>20</td>
<td>320</td>
</tr>
<tr>
<td>no</td>
<td>10</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>total</td>
<td>310</td>
<td>90</td>
<td>400</td>
</tr>
</tbody>
</table>
```

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

(3 marks)

Hint: The formula for the kappa statistic is: \( \frac{(P(A) - (PE))/(1 - P(E))}{1 - P(E)} \), 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. Use pooled marginals to calculate \( P(E) \), that is, \( P(E) = P(Y)^2 + P(N)^2 \), where \( Y \) is the proportion of times the annotators said yes, and \( N \) is the proportion of times the annotators said no.

[Question 5 continues on the following page]
ii) State what level of agreement is represented by the value you obtain for kappa. (1 mark)

c) King (2005) noted that “taking the results achieved as the only indicator of quality of the underlying technology accounts for the black box philosophy of evaluation campaigns. For example, there is no way of weighing up the potential fruitfulness of investment in system development as a preliminary to carrying out the investment”.

Discuss how notions of “quality in use” may be applied to evaluate text mining systems, including in your answer a discussion of how such notions relate to the type of evaluation typical of evaluation campaigns. (4 marks)

d) Briefly discuss the advantages of machine learning approaches over rule based approaches for named entity recognition. (4 marks)

e) Grimes (2009) noted that “the text analytics state-of-the-art is advancing at a very rapid pace, and with plenty of challenges yet to be met”.

Briefly set out what you consider to be the major current challenges for text mining, and then discuss 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. (6 marks)