Two hours

Formula Sheet attached for use with Question 4.

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Text Mining

Date: Tuesday 19th May 2015
Time: 14:00 - 16:00

Please answer any THREE Questions from the FIVE Questions provided

Each question is worth 20 marks

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) The man sold the last ticket for the music festival wanted a refund.
   ii) The table took up lots of room.
   iii) Coming round the corner, a beautiful lake caught the eye.
   iv) The chickens are ready to eat.

   (2 marks)

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

   i) How does Brill’s Transformation-based Learning (TBL) algorithm 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) Order the following major components of the TBL method in terms of the amount of their contribution to performance:

   • Lexical rules
   • Initial state
   • Non-lexical rules

   Explain and justify your ordering.

   (2 marks)

   iii) Briefly discuss advantages and disadvantages of the TBL algorithm.

   (2 marks)

   c) Probabilistic context-free grammars (PCFGs) are said to suffer from the independence assumption and from poor lexical sensitivity. Explain these two disadvantages and describe how the PCFG framework may be modified to overcome them.

   (2 marks)
Consider the following grammar and lexicon (numbers are given for reference only):

1) S -> NP VP 
2) NP -> DET N 
3) NP -> DET NNS 
4) NP -> N 
5) NP -> NNS 
6) NP -> NNS PP 
7) N -> N N 
8) N -> N NNS 
9) N -> N PP 
10) PP -> PREP NP 
11) VP -> VG NP 
12) VP -> VG 
13) VG -> V 
14) VG -> AUX V 
15) DET -> a 
16) N -> dog 
17) N -> duck 
18) N -> can 
19) NNS -> dogs 
20) NNS -> ducks 
21) PREP -> near 
22) V -> bark 
23) V -> barks 
24) V -> duck 
25) V -> can 
26) V -> ducks 
27) AUX -> can 

i) Show, by constructing two parse trees, that the string “a dog near ducks can bark” is ambiguous according to the above grammar.

(2 marks)

ii) Show the steps (rule numbers) that a naïve top-down (goal driven) parser would take in parsing the string “dogs bark” according to the above grammar.

(2 marks)

iii) Compare and contrast a naïve, top-down (goal-driven) parser and a naïve, bottom-up (data-driven) parser in terms of their advantages and disadvantages.

(2 marks)

iv) Assume that the Earley algorithm has been applied to parse the string “a dog near ducks can bark” with the grammar and lexicon shown above. The parser has paused at the state shown below as a chart. Reproduce this chart, and add the next nine edges (representing scanning, prediction or completion steps) that would be produced by the algorithm.

(3 marks)
2.  
   a) What is BILOU notation, and why is it often preferred to BIO notation?  
      (2 marks)  
   
   b) Annotate the following sentence to show the boundaries of its (underlined) noun phrase chunks, using the BILOU notation:  
      
      Smith said “The Organisation for European Security has offices in many countries, around the world”.  
      
      (2 marks)  
   
   c) Explain how boundary, inline and stand-off annotations differ, commenting on their relative strengths and weaknesses.  
      (3 marks)  
   
   d) Consider the following newswire text:  
      
      Amsterdam, Netherlands, 9th February, 2015. /WorldEvents Online News/  
      Floods fly away  
      By Willem van Nord  
      Staff Reporter  
      
      Dutch scientist Dr Frits van Schenau, University of West Amsterdam, and Graphene Manufacturing plc [LSE:GRMA] today announced successful trials of Bibographene™ on the flooded plains of Leefdal, Brabant, Belgium. Dr van Schenau said that yesterday from 09:15 CET an AviaPlano CD-180E Cropduster aircraft sprayed a fine sheet of Bibographene over 5 km² of flooded fields on several occasions. Once bound to water molecules, the material became lighter than air due to reaction with Hydrogen and thus raised a layer of water into the sky where the wind carried it off. Graphene Manufacturing has announced orders worth over 3bn euros. However, the Belgian Air Traffic Control Agency issued a NOTAM (Notice to Airmen) warning of “Multiple Drifting Near-Invisible Airborne Water Hazards” covering a large area to the NW of 50° 53’ N 4° 29’ E and causing huge disruption to flights. Moreover, across the North Sea, Davinia Gristwistle, MBE, CEO of Norfolk Flax Company said that her company’s linseed crop had been destroyed by a massive Bibographene sheet breaking up overhead.  
      
      Instances of the following named entities are to be identified: PERSON, LOCATION, DATE, TIME, ORGANISATION, COMPANY, ARTEFACT, MONEY_AMOUNT. (ARTEFACT = something that is man-made)  
      
      [Question 2 continues on the following page]
What patterns (including contextual clues) would you use to help you write rules to identify the maximum number of instances of the above named entities in the text? For each pattern you specify, state which instances it would match. You may specify patterns informally (e.g., one or more capitalised tokens + {“city”, “river”} = LOCATION). If you find it useful to introduce other named entity types to aid your analysis, do so. Note any problematic aspects of the text that may cause your patterns to recognise too much, too little, or nothing, in certain cases.

(5 marks)

e) The following is evaluation data for a model trained using the MALLET CRF SimpleTagger.

Mount NNP CAPITALIZED B-CITY B-ORGANIZATION
Pleasant NNP CAPITALIZED I-CITY I-ORGANIZATION
Winery NNP CAPITALIZED O I-ORGANIZATION
in IN LOWERCASE O I-ORGANIZATION
Augusta NNP CAPITALIZED B-GIVEN_NAME B-PORT B-PROVINCE B-LOCATION
, , PUNCT O O
Mo., NN OTHER O B-LOCATION
submitted VBD LOWERCASE O O
bottles NNS LOWERCASE O O
of IN LOWERCASE O O
Genesis NNP CAPITALIZED B-CITY O
Champagne NNP CAPITALIZED B-CITY O
, , PUNCT O O
but CC LOWERCASE O O
Newport NNP CAPITALIZED B-CITY B-PORT B-PROVINCE B-ORGANIZATION
News NNP CAPITALIZED I-CITY I-PORT I-PROVINCE I-ORGANIZATION
Shipbuilding NNP CAPITALIZED O I-ORGANIZATION
rejected VBD LOWERCASE O O
them PRP LOWERCASE O O
. . PUNCT O O

i) Explain how this evaluation data would differ from test data given to the model.

(1 mark)

ii) Suppose that, after evaluating the model, you suspect that gazetteer features are not contributing very much to the recognition. Why might you think this and how could you determine if your suspicion is valid?

(3 marks)

f) Discuss advantages and disadvantages of dictionary-based, rule-based and machine learning approaches to named entity recognition.

(4 marks)
3. 
   a) Give a dependency representation for the following sentence, for the interpretation that the elephant was in the pajamas:

   I shot an elephant in my pajamas.

   The dependency relation set you must use for your analysis is:
   SBJ (subject), OBJ (object), NMOD (noun modifier), PMOD (preposition modifier), DETMOD (determiner modifier).
   
   (2 marks)

   b) When designing a workflow for an information extraction task, two parsing components are available to you: a dependency parser and a phrase structure parser. Which would you choose, and why?

   (3 marks)

   c) Consider the following sentence:

   “On Monday, an Egyptian court sentenced Peter Greste to 10 years on charges of reporting false news.”

   i) What is the trigger for the key event being described in the sentence?

   (1 mark)

   ii) Sketch how the key event expressed in this sentence could be represented using an appropriate template, e.g., the ACE 2005 event annotation scheme. Include the event-type and all available participants in your representation.

   (3 marks)

   d) What role does named entity extraction play in a system that supports faceted search?

   (2 marks)

   e) Design a simple UIMA type system that enables capturing events related to football found in sports news articles. The annotation task calls for the representation of various types of named entities as well as the events that they participate in.

   The categories of events are predefined and are limited to: Match, Transfer and Qualify. Similarly, the categories of named entities are limited to: Person, Team, CompetitionRound, Location and Date.

   An event should be anchored to a word or phrase (i.e., a trigger) that signifies the event. It should also list the different types of event arguments. A Participant argument can take either a Person or a Team as its value.

   [Question 3 continues on the following page]
Additionally, each event must have an indication of whether its polarity is positive or negative. It should be possible to make the following annotations over the given three sentences. A crossed out event trigger indicates that the event is negated.

For each UIMA type, state its name, supertype, and a list of features (do not repeat the supertype's features). You may re-use the UIMA built-in type TOP. For each feature, indicate its name and type. Give your solution in a table with the column headings: Type, Supertype, Features (and give Name and Type for each feature).

(6 marks)

f) Consider the following elementary processing components:

(1) Gold standard corpus reader: Reads in 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) Gazetteer lookup

(4) Part-of-speech tagger

(5) Annotation remover

(6) Sentence splitter

(7) Tokeniser

(8) Orthography detector

(9) Named entity recogniser

Design a named entity recognition workflow, by drawing a diagram, that would allow you to evaluate the combined effectiveness of the components (3) to (9) 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 for components (1) and (2). You do not need to refer to any specific type systems.

(3 marks)

[PTO]
4. 
   a) Give two main characteristics of the Princeton WordNet that differentiate it from other electronic dictionaries.  
      (2 marks)

   b) 
      i) Briefly explain Lesk’s algorithm for word sense disambiguation.  
         (1 mark)

      ii) Although Lesk’s algorithm remains popular, what disadvantages of this algorithm limit its effectiveness?  
         (2 marks)

   c) What are the advantages of the count-based distributional semantic model?  
      (2 marks)

   d) 
      i) Why do we use compositional methods to compute the meaning of phrases?  
         (1 mark)

      ii) Briefly explain how the basic additive model (BAM) and the basic multiplicative model (BMM) work.  
         (2 marks)

   e) 
      i) When two different distributions are compared using t test, what assumption must be true for this test to be valid?  
         (1 mark)

      ii) What is the advantage of t test over frequency-based methods?  
         (1 mark)

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

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

      A corpus about computer hardware consists of 200000 bigrams, where hard occurs in 250 bigrams and drive occurs in 800 bigrams. The bigram hard drive occurs 128 times.

      i) State the t test null hypothesis.  
         (1 mark)

      ii) Decide if the co-occurrence of hard and drive is random or not using the t test. Show your working. (The critical value for a confidence level $\alpha=0.005$ is 2.576.)  
         (6 marks)
5.  
   a)  Two annotators were each given a copy of a document in which certain text spans 
       have already been marked up as names, although still lacking labels. The annotators 
       had to assign a label to each of these text spans as follows: assign DISEASE if they 
       think a text span is the name of a disease, and NON_DISEASE otherwise. 

       The contingency table below shows how many times the two annotators assigned the 
       same label to the text spans. 

       | Annotator 2 Results | Annotator 1 Results |
       |---------------------|---------------------|
       | DISEASE             | 45                  |
       | NON-DISEASE         | 10                  |
       |                     | 7                   |
       |                     | 38                  |

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

       (3 marks) 

       Hint: The formula for the kappa statistic is: \( \frac{P(a) - P(e)}{1 - P(e)} \), 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. 

   b)  Consider the following partially completed diagram: 

       ![](image)

       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) 

       [PTO]
c)  

i) What text elements may convey subjective attitudes as understood in opinion mining or sentiment analysis? Give examples for the types of text element you identify.

(5 marks)

ii) Discuss why we might prefer an aspect-based sentiment analysis to sentence-based or document-based sentiment classification, explaining what types of technique and resources are required to achieve such analysis.

(5 marks)

d) In 2005, Witten wrote: “Automatic text mining techniques have a long way to go before they rival the ability of people, even without any special domain knowledge, to glean information from large document collections”.

Discuss to what extent this opinion from a decade ago could apply to today’s state of the art in text mining. In your answer, also discuss any barriers or challenges that have either persisted since this opinion was written or that may have arisen in the years since. Justify your views and conclusions, giving appropriate examples to back up your arguments.

(5 marks)

END OF EXAMINATION