COMP 61332

Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Text Mining

Date: Thursday 1st June 2017

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

[PTO]

1.

a) Classify the following examples according to the type of ambiguity they display:

i) The dog chased the cat under the table.

ii) He heard her talk.

iii) The student saved the expense was happy.

iv) The car hit a tree speeding at 80mph.

(2 marks)

b)

i) Annotate all tokens in the following sentence to show the boundaries of its (underlined) noun phrase chunks, using the BILOU notation:

The Prime Minister, in Brussels, said "The United Kingdom has many plans to boost employment".

(2 marks)

ii) Ratinov & Roth (2009) found that a CRF named entity classifier trained on BILOU annotations outperformed one trained on BIO annotations, and this has been the accepted view. However, Cowan et al. (2015) report that, on their data sets, a CRF named entity classifier trained on BIO annotations outperformed one trained on BILOU annotations. *Focussing on the type of annotation*, what conclusion(s) do you come to about this apparent contradiction?

(2 marks)

iii) You must choose which type of annotation notation to use for your corpus annotation project, which involves annotating named entities and events. You consider boundary annotation, inline XML-based annotation and stand-off annotation. Which type of notation should you choose and why?

(3 marks)

c) Consider the following UIMA-based processing components:

(1) Reference evaluator: Reports annotation effectiveness comparing two inputs of which one is the point of reference. The report includes performance metrics.
(2) Annotation remover
(3) Gold standard corpus reader
(4) Part-of-speech tagger
(5) Syntactic parser
(6) Dictionary-based NER for disease names
(7) Sentence splitter
(8) Tokeniser
(9) Dictionary-based NER for drug names
(10) Orthography detector

[Question 1 continues on the following page]

[Question 1 continues from the previous page]

Assume that you have just created components (6)–(10).
You wish to carry out named entity recognition (NER) for drug and disease names. Design a workflow, by drawing a diagram, which would allow you to evaluate the combined effectiveness of your components against a given gold standard corpus. You do **not** need to refer to any specific UIMA-based type systems.

(4 marks)

d)

i) What are the three main characteristics of a component-based interoperable text mining architecture?

(3 marks)

ii) A colleague wishes to produce a text mining application to extract events to support semantic search, and is thinking about producing applications for other text mining tasks in the future. He asks you if it would be better to: a) adopt a text mining workflow approach; b) acquire some stand-alone tools from several different providers for some parts of his task and produce one or two other tools himself; or c) write one program for his task.
What advice would you give him, bringing out advantages, disadvantages and challenges that he should be made aware of?

(4 marks)

[PTO]

2.

a) You test a number of tokenisers on English business news text and find the following:

i) one tokenises only on white space
ii) one tokenises on white space and punctuation, including hyphen and range dash
iii) one tokenises on white space but removes all punctuation and numbers from the output
iv) one tokenises on white space and punctuation, but does not break hyphenated words and keeps numbers with decimal points or commas together

Discuss to what extent any of these would be useful in supporting named entity recognition of PERSON, MONEY_AMOUNT, ORGANISATION and DATE entities, explaining what issues may arise, and discussing what other tokenisation behaviour may have to be considered in any new tokeniser for this task.

(3 marks)

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

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) Explain how the lexicon, the transformation rules and the part of speech guesser are used to tag a new, unseen text.

(3 marks)

iii) Brill noted: "We found it a bit surprising that the addition of lexicalized transformations did not result in a much greater improvement in performance.". What reason can you give that would explain this result?

(1 mark)

[Question 2 continues on the following page]

[Question 2 continues from the previous page]

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

```
1) S -> NP VP          8)  V -> strikes      11) N -> winter
2) NP -> N             9)  V -> delay        12) N -> storm
3) NP -> N N           10) V -> storms       13) N -> delay
4) NP -> N NNS                               14) N -> train
5) VP -> VG                                  15) NNS -> storms
6) VP -> VG NP
7) VG -> V
```

i) Show, by constructing two parse trees, that the string "winter storms delay train" is ambiguous according to the above grammar.

(2 marks)

ii) Show the steps (rule numbers) that a naïve top-down (goal driven) depth-first parser would take in parsing the string "winter storm strikes" according to the above grammar.

(2 marks)

d) The notion of *head of a syntactic construction* is an important one in Linguistics. How is this notion of relevance in the world of probabilistic context free grammars (PCFGs)? Illustrate your answer with appropriate examples.

(2 marks)

e) Consider the following sentence:

The burglar threatened the student with the knife.

Draw two dependency graphs according to the notation of your choice but using the following label set: `det` (determiner), `nmod` (nominal modifier), `obj` (object), `pobj` (object of the preposition), `punct` (punctuation), `sbj` (subject), `vmod` (verb modifier). One graph should be for the interpretation that the burglar had the knife, the other graph should be for the interpretation that the student had the knife.

(4 marks)

[PTO]

3.

a) Consider the following lines from news headlines and articles.

- The breathtaking match was held at the Etihad.
- Etihad eyes growth this year as it reports rise in passenger numbers.
- An Etihad flight from Dublin to UAE was diverted over security threat.
- Manchester City looking to make the Etihad bombproof.
- Etihad to launch double daily service to Düsseldorf.
- An epic night at the Etihad: What We Learned from Man City 5-3 Monaco.

The name "Etihad" is ambiguous between a stadium (location) and a company (organisation). If you are going to develop a machine learning-based named entity recogniser (NER), what features would you include to help the NER handle this ambiguity, and why?

(3 marks)

b) Consider the following sentences S1..S3:

S1: The Australasia Restaurant charged Prof. Smith a steep £200 for a bottle of Dom Perignon last night.

S2: Prosecutor Charles White charged Ian Stewart with the murder of his partner Helen Bailey at the couple's home in Hertfordshire.

S3: Armed with tear gas, Police Nationale charged at a crowd of anti-racism demonstrators called Mouvement Inter Luttes Indépendant, leading to the injury of two people in Paris last Saturday.

i) Annotate the events in the sentences using the templates below. Put "N/A" if a role does not have any value. Do **not** reproduce the template explanations or role definitions in column 2.

(8 marks)

Event type: `CONFLICT-ATTACK` (a violent physical act causing harm or damage)

`Trigger`	the word signifying the event	
`Attacker-Arg`	the attacking/instigating agent	
`Target-Arg`	the target of the attack	
`Instrument-Arg`	the instrument used in the attack	
`Time-Arg`	when the attack takes place	
`Place-Arg`	where the attack takes place	

[Question 3 continues on the following page]

[Question 3 continues from the previous page]

Event type: `DIE` (occurs when the life of a person ends)

`Trigger`	the word signifying the event	
`Agent-Arg`	the killer	
`Victim-Arg`	the person who died	
`Instrument-Arg`	the device used to kill	
`Time-Arg`	when the death takes place	
`Place-Arg`	where the death takes place	

Event type: `ACCUSATION` (occurs when an entity is believed to have committed a crime)

`Trigger`	the word signifying the event	
`Defendant-Arg`	the entity being accused	
`Prosecutor-Arg`	the entity who pushes the accusation forward	
`Adjudicator-Arg`	the judge or court	
`Crime-Arg`	the crime	
`Time-Arg`	when the accusation takes place	
`Place-Arg`	where the accusation takes place	

Event type: `TRANSFER-MONEY` (the buying, selling, loaning, borrowing, giving, or receiving of artefacts)

`Trigger`	the word signifying the event	
`Giver-Arg`	the giver	
`Recipient-Arg`	the recipient	
`Artifact-Arg`	the item that was bought or sold	
`Money-Arg`	the amount given, donated or loaned	
`Time-Arg`	when the amount is transferred	
`Place-Arg`	where the transaction takes place	

[Question 3 continues on the following page]

[Question 3 continues from the previous page]

ii) The verb "charge" has different meanings across sentences S1–S3. Imagine that you are about to develop a machine learning-based event trigger detector that should be able to assign the correct event type to each instance of "charge". Give examples of syntactic relations involving "charge" that will help the detector identify the correct event type.

(3 marks)

c) In relation to Information Retrieval, consider the following index terms and their postings list sizes:

Index term	**Posting list size**
biscuit	213313
chocolate	87010
lemonade	107914
orange	271659
sugar	46654
toast	316813

A user gives the Boolean query:

```
(sugar OR toast) AND
(lemonade OR orange) AND
(chocolate OR biscuit)
```

Recommend a query processing order for this query. Justify your recommendation.

(2 marks)

d) In relation to constructing an inverted index, Manning et al. (2008) note that "relevance does not increase proportionally with term frequency of occurrence" and that "rare terms are more informative than frequent terms". Explain how we can we use such notions to weight terms and, during search, to rank documents for queries.

(2 marks)

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

(2 marks)

4.

a) A developer of a search engine carries out tests with several query terms and finds that he receives only a few documents for each, from a large collection where he knows there are other relevant documents. He asks you if adding the Princeton WordNet as a resource would improve search. How would you advise him? Give reasons to support your advice.

(3 marks)

b)

i) Briefly explain Lesk's algorithm for word sense disambiguation (WSD).

(1 mark)

ii) Explain what would be the result of applying Lesk's algorithm to find the correct sense for *bird* in the phrase *The bird feather in the cage*. In your answer, refer to the following data. Assume that stemming has taken place and stop words have been removed.

Senses for bird:

bird_1: warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings.

bird_2: the flesh of a bird or fowl used as food.

bird_3: badminton equipment consisting of a ball of cork or rubber with a crown of feathers.

Context definitions:

feather: the light horny waterproof structure forming the external covering of birds.

cage_1: an enclosure made of wire or metal bars in which birds or animals can be kept.

cage_2: a movable screen placed to catch balls during batting practice.

(3 marks)

iii) On the basis of the result you find in 4 b) ii), would you choose Lesk's algorithm to do WSD? Say why, or why not.

(1 mark)

c)

i) What features would you consider in supervised mode for training a system to carry out WSD using a machine learning algorithm? Justify your choices.

(3 marks)

ii) Discuss limitations of supervised WSD techniques and how we may overcome such limitations.

(3 marks)

[Question 4 continues on the following page]

[Question 4 continues from the previous page]

d)

i) In a comparison of count-based and context-predicting distributional semantics models, Baroni et al. (2014) conclude "we would certainly recommend to go for the predict models". To what extent do you agree or disagree with their conclusion? Give justifications.

(3 marks)

ii) To construct a thesaurus, you firstly compile context vectors and apply cosine similarity. What steps should you then take and what decisions may affect the final result?

(3 marks)

5.

a) A named entity recogniser (NER) for gene and protein names obtains the following number of true positives (TPs), false positives (FPs) and false negatives (FNs) for each entity type.

Entity Type	TPs	FPs	FNs
Gene	83	12	45
Protein	35	4	2

i) Calculate the value of precision, recall and F1-score for each type.

(3 marks)

ii) Calculate the value of: (i) micro-averaged and (ii) macro-averaged precision and recall for the Gene and Protein types.

(3 marks)

iii) If a biologist asks you for the performance of the NER on genes and proteins combined, which of the micro-averaged or macro-averaged scores would you give? Why?

(2 marks)

b) Aspect-based sentiment analysis (Liu, 2012) has become popular recently. Briefly describe this approach, discussing how it differs from approaches that attempt a positive/negative sentiment classification, and indicating what techniques are required to achieve its goals.

(5 marks)

c) "Text mining remains a hard sell: the technologies are complex, often fragile, domain-specific and with a high start-up cost." (Tonkin, 2016)

To what extent do you agree with this view? Give arguments for or against it, taking into account any potential for near- to medium-term progress you are aware of. Provide appropriate justification and examples to back up your arguments.

(7 marks)

END OF EXAMINATION