COMP 61332

Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Text Mining

Date: Tuesday 21st May 2019

Time: 14:00 - 16:00

Please answer all FOUR Questions.
Each Question is worth 15 marks

Use a SEPARATE answerbook for each SECTION

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text

[PTO]

Section A

1. a) Lexical Knowledge Bases (LKBs) are widely used in NLP tasks. Define an LKB and how it can be used in NLP tasks. (2 marks)

   b) i) Define semantic role (within the context of semantic relations,frames and events).

   Assign the appropriate semantic roles for the following words in italics in the sentences below. Choose from these Tags: Theme, Instrument, Source, Location, Experiencer, Agent.

   1. *Lucy* left the room.
   2. I like *coffee* a lot.
   3. Lucy opened the door with *her key*.
   4. *My father* feels very proud of me.
   5. *My cup of coffee* was drunk.
   6. *Great people* think alike. (4 marks)

   ii) How does WordNet organise lexical information? Exemplify your answer. (1 mark)

   c) Discuss how distant supervision can be used in conjunction with lexical knowledge bases and large amounts of unlabeled text to produce labeled training data. (4 marks)

   d) C-value is a hybrid measure for automatically extracting candidate terms. What is involved in the calculation of the statistical part of C-value? (4 marks)

2. a) In which NLP applications can we use automatic term recognition? Provide four applications with some brief justification of how terms can be used. (4 marks)

   b) Automatic indexing is a method of information retrieval which extracts index terms. Briefly assess the usefulness of this method for extracting technical terms from documents and compare index terms with technical terms. (4 marks)

   c) Explain why normalisation is needed after Named Entity Recognition and enumerate the four main approaches. Critically discuss the pros and cons of dictionary-based and string similarity measures for concept normalization from terminological resources. (7 marks)

Section B

3. Consider the following sentence:

The dog chased the cat down the road.

This sentence has two possible interpretations: that the dog moved further along the road to chase the cat, or that the dog chased the cat which is located at a lower part of the road.

a) Explain the reason for the syntactic ambiguity, i.e., which token of the sentence has more than one possible head and what they are. (2 marks)

b) Draw the dependency graph for each of the two possible interpretations 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). (4 marks)

c) Draw a phrase structure tree for each of the two possible interpretations, using the following label set: `DT` (determiner), `IN` (preposition), `JJ` (adjective), `NN` (singular noun), `NNS` (plural noun), `RB` (adverb), `VB` (verb), `AdjP` (adjectival phrase), `AdvP` (adverbial phrase), `NP` (noun phrase), `PP` (prepositional phrase), `VP` (verb phrase) and `S` (sentence). (4 marks)

d) Annotate the noun phrases in the sentence above using the BIO notation. (2 marks)

e) Which of the BIO and inline annotation formats is more suitable for representing a phrase structure tree (such as what you produced in Question 3.c above)? Explain why. (3 marks)

[PTO]

4. Consider the following text snippet (adapted from the Macmillan Cancer Support web page on lung cancer[1]). Note: some words are in bold font and highlighted in grey for the purposes of Question 4.c.

*There are two main types of **primary lung cancer**. They behave in different ways and respond to treatment differently. They are:*

***small cell lung cancer** (**SCLC**), which makes up about 10% of **lung cancers**, and*

***non-small cell lung cancer** (**NSCLC**), the most common type which has three subtypes:*

- ***adenocarcinoma** is the most common type of **lung cancer**. It develops from mucus-producing cells that line the airways.*
- ***squamous cell carcinoma** develops in the cells that line the airways. It is usually caused by smoking.*
- ***large cell carcinoma** (sometimes called **undifferentiated carcinoma**) is named because of how the **cancer** cells look when examined under a microscope.*

a) Assume that you are given the gazetteer (dictionary) below:

adenocarcinoma	DISEASE
airway	ANATOMICAL_PART
cancer	DISEASE
large cell carcinoma	DISEASE
lung	ANATOMICAL_PART
lung cancer	DISEASE
small cell lung cancer	DISEASE

Discuss the disadvantages/shortcomings of a purely dictionary-based named entity recogniser (NER) that uses the above gazetteer as a look-up list. Give specific examples drawn from the text snippet above to substantiate your answer. (2 marks)

b) Give an example of a rule that uses function words, which can be used to recognise more disease names in the above text snippet. Provide examples of what this rule will recognise that was not recognised by the purely dictionary-based NER described above. (2 marks)

[1] https://www.macmillan.org.uk/information-and-support/lung-cancer/understanding-cancer/types-of-lung-cancer.html

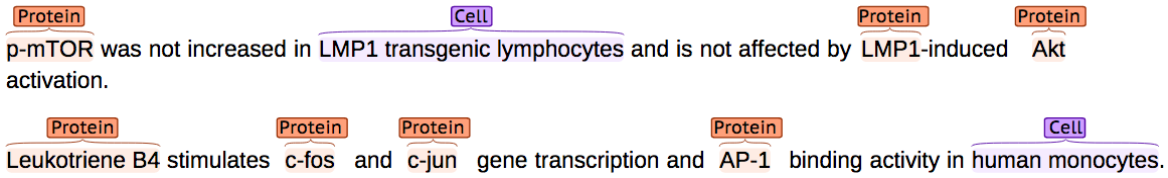
c) Assume that in the text snippet provided above, token sequences in ***bold*** are correct/gold standard disease names. A named entity recogniser was applied to it and recognised the tokens highlighted in grey as disease names.

Populate the following confusion matrix:

	True Positives (TP)	False Positives (FP)	False Negatives (FN)
Disease			

What is the value of each of recall, precision and F-score? Show your computations. (6 marks)

d) Identify the events that you would annotate in the text snippet below if you were an event extraction tool. Note that the passage is already supplied with gold standard named entities (proteins and cells).



Each event can be represented using the template below.

TRIGGER: TYPE: CAUSE: THEME: LOCATION: POLARITY:

The following mapping between triggers and event types can be used:

Regulation	*affect*
Positive_regulation	*induce, increase, stimulate*
Gene_expression	*gene transcription*
Binding	*binding*
Activation	*activation*

(5 marks)

END OF EXAMINATION