COMP 34412

Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Natural Language Systems

Date: Thursday 23rd May 2019

Time: 09:45 - 11:45

Please answer all FOUR Questions
Each question is worth 20 marks

Use a SEPARATE answerbook for each QUESTION

This is a CLOSED book examination

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

[PTO]

Question 1

i) Describe the difference between stemming and lemmatisation. Illustrate your answer with reference to the words '*studies'* and '*plays*'.

[2 marks]

ii) Would you be more concerned with the quality of inflectional or derivational morphology for a source language if you were developing a machine translation system? Explain your answer.

[4 marks]

iii) Brill's part-of-speech (POS) tagger is an example of transformation-based learning (TBL) approach to tagging. Explain the TBL approach and why Brill's tagger is known as an "error-driven transformation-based" tagger.

[5 marks]

iv) Explain the sequence labelling approach for the named-entity recognition (NER) task. Discuss the typical steps in the training phase and features that can be used.

[5 marks]

v) One of the problems for named-entity recognition (NER) systems are ambiguous words that might be part of different entity types. For example, what would you need to do to make sure that word *May* in *Mrs May* is considered as part of a name, rather than a time expression?

[4 marks]

Question 2

i) Explain and provide examples to illustrate two types of syntactic ambiguity: prepositional attachments and nominal coordination.

[4 marks]

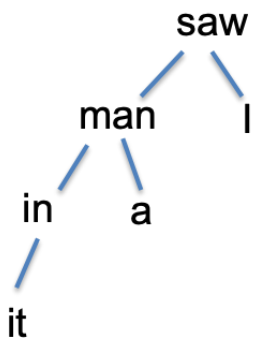
ii) Explain the data structures (stack, queue) and operations (shift, reduce left, reduce right) used in transition-based dependency parsing.

[4 marks]

[Question 2 continues on the next page - PTO]

[Question 2 continues from the previous page]

iii) Show the sequence of operations in transition-based dependency parsing that would obtain the dependency tree below from the sentence '*I saw a man in it*'.



[4 marks]

iv) You have been asked to develop a sentiment analyser to process student feedback on the quality of teaching. Each student is asked to provide a single free-text comment on all course units they had in the current year, so you are expected to link sentiment to specific course units. Explain how you will approach the task and discuss the main challenges you will be facing with this kind of text (including, for example, part-of-speech tagging, parsing, relationship extraction, etc. [8 marks]

Question 3

i) Table 1 provides the WordNet sense knowledge for word jam.

Table 1: WordNet Senses	
jam-1	Gloss: a crowded mass that impedes or blocks <a traffic jam> Example: Trucks sat in a jam for ten hours waiting to cross the bridge.
jam-2	Gloss: a food made by boiling fruit and sugar to a thick consistency Example: He spread home-made jam on his toast.

a) Use the 6 training examples given in Table 2 (see next page) to build a Naïve Bayes sense classifier to disambiguate jam-1 and jam-2, and describe how to use this classifier to disambiguate "jam" for the *query {traffic, work, busy}*. The used dictionary is: {traffic, driving, slow, busy, work, sweet, thick, fruit, store, jelly}. [6 marks]

[Question 3 continues on the next page - PTO]

[Question 3 continues from the previous page]

Table 2: Training Examples		
ID	Content words that appear within the context window of "jam"	sense class
Example 1	{traffic, driving, moving}	jam-1
Example 2	{traffic, slow, work, busy}	jam-1
Example 3	{traffic, busy, highway, traffic}	jam-1
Example 4	{strawberry, jelly, sweet, store}	jam-2
Example 5	{sweet, thick, fruit, work}	jam-2
Example 6	{fruit, jelly, jelly, sweet }	jam-2

b) Can you trust this classifier to disambiguate "jam" for the *query {peach, smooth, jar, preserve, work}*? Explain your answer.

[4 marks]

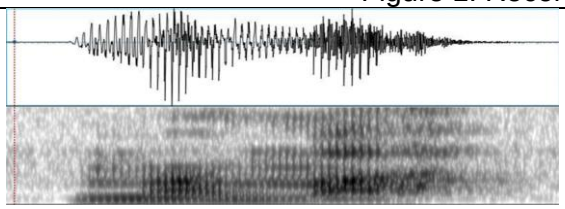
c) Given the sentence below, describe how to apply the simplified Lesk algorithm to disambiguate "jam" based on the knowledge provided in Table 1.

*"A driver cruising easily at 70 m.p.h. in Lane ~A of a four-lane freeway spies an incipient traffic **jam** ahead. Traffic in the next lane appears to be moving more smoothly so he pokes a tentative fender into Lane ~B, which is heavily populated by cars also moving at 70 m.p.h..".*

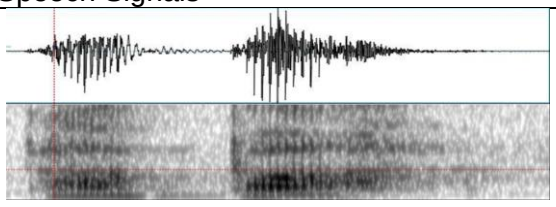
[4 marks]

ii) What is the difference between the way that people use their articulators to produce vowels and the way they use them to produce consonants? On the basis of what you have just said about this difference, decide which of the speech signals in Figure 1 corresponds to someone saying 'mama' and which to someone saying 'papa'.

Figure 1: Recorded Speech Signals



signal A



signal B

[6 marks]

Question 4

i) Consider the text collection of 5 sentences as below:

sentence 1	Support vector machine is commonly used in document classification.
sentence 2	Support vector machine is trained by quadratic programming techniques.
sentence 3	Document classification is a natural language processing task.
sentence 4	Bag of visual word technique is commonly used in computer vision.
sentence 5	Support vector machine is commonly used in computer vision.

Consider 6 indexed terms as below:

term 1	support vector machine
term 2	document classification
term 3	natural language processing
term 4	computer vision
term 5	visual word
term 6	quadratic programming

a) Build a document-term matrix containing term counts in the documents.

[4 marks]

b) Use the document-term matrix to decide whether *"natural language processing"* or *"visual word"* is closer to *"document classification"*, and explain your answer.

[4 marks]

c) Focus on the word *"machine"* in sentence 1, and apply a context window of size ± 2. What training examples can be prepared for training a Skip-Gram model to learn a word embedding vector for "*machine*"? Explain your answer.

[4 marks]

d) If using term frequency and inverse document frequency (tf-idf) weighting scheme to prepare the document-term matrix, what is the tf-idf weight for "*support vector machine*" in sentence 1? Explain the benefit of tf-idf weighting in general.

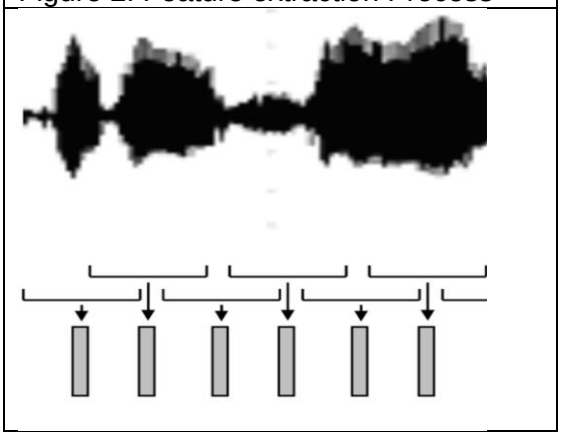
[4 marks]

ii) Given a speech signal lasting 200ms, 50 spectral features are extracted every 20ms with a duration of 40ms. This process is illustrated in Figure 2 (see next page). As a result, a sequence of N feature vectors are generated, each characterising a 40ms signal segment. Each feature vector is of M dimensions. Explain what the values of N and M are, and explain your answer.

[Question 4 continues on the next page - PTO]

[Question 4 continues from the previous page]

Figure 2: Feature extraction Process



[4 marks]

END OF EXAMINATION