Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Natural Language Systems

Date: Thursday 20th January 2011
Time: 14:00 - 16:00

Please answer Question ONE in Section A and TWO Questions from Section B.

For full marks your answers should be concise as well as accurate.
Marks will be awarded for reasoning and method as well as being correct.

This is a CLOSED book examination

The use of electronic calculators is NOT permitted
You should answer question 1: this question carries 30 marks

1. a) Describe the steps you would have to go through in order to obtain the parse tree below for the (written) sentence ‘The young cat was chasing the foxes.’ Illustrate your answer by showing the rules and data structures that would be used at each stage. [10 marks]

```
[s [np [det the]
    [nn [adj young]
        [nn [noun cat]]]]
 [vp [aux was]
    [vp [verb chase]
        [np [det the]
            [nn [noun fox]]]]]
```

b) Write a regular expression that would cover the same examples as the grammar below (you can assume that the text you have to analyse has simple tags embedded in it, i.e. that the goal is to be able to analyse

```
<det>the<adj>small<adj>smelly<noun>child<verb>smiled
```

but not

```
<adj>small<noun><children<verb>smile.
```

Explain the restrictions on the use of regular expressions as a grammatical formalism, illustrating your answer with grammar rules that could not be translated into regular expressions. [4 marks]

```
s ==> [np, vp]
vp ==> [verb, np]
np ==> [det, nn]
nn ==> [adj, nn]
nn ==> [noun]
```

c) What are precision, recall and F-measure? [3 marks] Suppose I had a part of speech tagger which labelled the words in the sentence ‘I know that she died’ as <pron>I <verb>know <pron>that <pron>she <verb>died when they should have been <pron>I <verb>know <comp>that <pron>she <verb>died. What are the precision, recall and F-measure of this tagger on this data? [4 marks] What are the consequences of it getting the tag for ‘that’ wrong? [3 marks]
Answer two questions from this section. Each question carries 35 marks.

2. a) Describe how a typical speech recognition system works. [7 marks] Such systems have to be ‘trained’: what is the training data like, and what is the system trying to learn? [6 marks]

b) What are formants, and what is their role in speech recognition? [7 marks] Given the speech signal below, explain where you would expect to find formants. [5 marks]

c) How would you determine whether this speech signal was someone saying ‘potato’ or ‘tomato’? Illustrate your answer by sketching the signal that you would expect to get for the other option (so if you think that this signal is for ‘potato’ sketch the signal for ‘tomato’ and vice versa). [10 marks]
3. a) Describe what is meant by a ‘vector space’ model of word meaning. [5 marks]
   Explain the role of ‘contexts’ in such treatments, giving at least two examples of different kinds of contexts. [3 marks] Are vector space models better for determining similarity or subsumption relations? [2 marks]

b) Given the passage below, show how using a vector space model could be used to decide whether ‘match’ was more similar to ‘series’ than to ‘Ashes’. [10 marks]
   What notion of context would be most appropriate for this task? [5 marks]

   England will go to war to win the Ashes. That was Andrew Strauss’s belligerent message ahead of his team’s flight to Australia today for the start of what the England captain says could be “one of the greatest winters ever for English cricket”.

   “We are out there to win. Simple. That is our goal. Everything we have done, every decision we have made up till now has been about us winning this Ashes series. Every decision we make from now on will be about us winning this Ashes series as well. None of us want to have any regrets at the end of it.”

   England have not won a series in Australia for 23 years. Since Mike Gatting’s team triumphed 2-1 in 1986-87 England have won only three of the 26 Test matches they have played in the country.

c) How could this help a search engine retrieve documents that were likely to contain the answer to a query? [10 marks]
4. a) Describe the use of ‘transfer rules’ in machine translation, and explain where systems that use such rules belong on the ‘machine translation pyramid’? [6 marks] What are the advantages of MT systems that use meaning representations near the apex of this pyramid, and why are such systems hard to build? [3 marks]

b) Describe how you might extract a bilingual lexicon from a corpus of translated material, and compare this with the task of extracting such a lexicon from a pair of general purpose corpora for two languages. [8 marks] Explain the importance of alignment when extracting a lexicon from translated material, and describe how you might cope in situations where there are no obvious anchor points for alignment. [5 marks]

c) Given the translation pairs below, how might you automatically detect that the English word ‘bank’ is ambiguous? [9 marks] Explain what, if anything, you would do about the potential attachment ambiguities for the prepositional phrases in the first two examples [4 marks]

(1) a. He sat down by the bank of the river
   b. Han satte sig ned vid stranden av floden.

(2) a. The ducks were sleeping on the bank of the lake
   b. Ankorna sov på stranden av sjön.

(3) a. The bank has lost all trace of my account
   b. Banken har förlorat alla spår av mitt konto.

(4) a. I am going to have to change my bank
   b. Jag kommer att behöva ändra min bank.