Two hours

Question ONE is COMPULSORY

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Natural Language Systems

Date: Friday 25th January 2013
Time: 14:00 - 16:00

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

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text.
You should answer question 1: each part of this question carries 5 marks


b) Explain the terms ‘precision’, ‘recall’ and ‘F-measure’. [3 marks] Suppose you had two taggers, where one assigned the tag `NN` to every word and the other assigned the tag `DT` to the word ‘the’ and left every other word untagged. What can you say about the precision and recall of these two taggers? [2 marks]

c) Regular expressions can be used to describe simple grammatical patterns, but they cannot be used for describing recursive structures. Explain briefly why it is not possible to write a regular expression to capture the fact that an NP can form part of a determiner phrase (as in ‘the old man’s wife’s father’, where ‘the old man’ is part of the determiner phrase ‘the old man’s’, and ‘the old man’s wife’ is part of the determiner phrase ‘the old man’s wife’s’). [4 marks] How does using cascaded regular expressions help you get round this problem? [2 marks]

d) What is the ‘term frequency-inverse document frequency (TF-IDF)’ score of a word? Why is it useful?

e) Explain why lexical ambiguity is a greater problem than structural ambiguity for machine translation systems. [2 marks] What kind of information might you be able to extract from a corpus to help with lexical ambiguity? [3 marks]

f) What is assimilation? Is it a greater problem for speech synthesis or speech recognition?
Section B

Answer two questions from this section. Each question carries 35 marks.

2. a) Describe the difference between inflectional and derivational morphology, illustrating your answer with examples of each. [6 marks] Which of these would you be more concerned with if you were developing a machine translation system? [4 marks]

b) The presence of spelling changes that reflect changes in pronunciation causes problems for programs that carry out morphological analysis. Explain briefly why the plural of the noun ‘kiss’ is written as ‘kisses’, and why the past tense form of ‘change’ is ‘changed’. [6 marks]

c) Consider the following English spelling rules, given in the format used in the lectures:

\[

de \Rightarrow [e] : [c0] \_ [v0]. \\
[\varepsilon] \Rightarrow [] : [c0, c1] \_ [c2]. \\
[c0] \Rightarrow [ ] : [v0] \_ [c0].
\]

Which of the following surface forms are related to their constituent parts by these rules (where a surface form is derived by one of the rules you must specify which one): ‘kisses’ \Rightarrow ‘kiss’+’s’, ‘changing’ \Rightarrow ‘change’+’ing’, ‘seen’ \Rightarrow ‘see’+’en’, ‘clapping’ \Rightarrow ‘clap’+’ing’, ‘calling’ \Rightarrow ‘call’+’ing’, ‘watches’ \Rightarrow ‘watch’+’s’, ‘lived’ \Rightarrow ‘live’+’ed’? [7 marks]

d) Show how the categorial descriptions of roots and affixes in Figure 1 support analyses of the words ‘chanter’, ‘chantez’, ‘chantons’, ‘chanterez’, ‘chanterons’, ‘chantiez’, ‘chantions’ given the basic rules of categorial combination in Figure 2. [7 marks]

\[

\text{’chant’} \Rightarrow a/b. \quad \text{’i’} \Rightarrow b/c. \\
\text{’er’} \Rightarrow b. \quad \text{’ez’} \Rightarrow c. \\
\text{’’} \Rightarrow b/c. \quad \text{’ons’} \Rightarrow c. \\
\text{’er’} \Rightarrow b/c.
\]

Figure 1: Roots and affixes

Show how the extended categorial rules in Figure 3 will let you analyse these words from left to right. What are the advantages of carrying out morphological analysis in a strictly left to right order? [5 marks]
3. a) Almost all natural language systems include a component which tries to find the relations between words—a ‘parser’. Traditional systems use a hand-crafted grammar, which is intended to capture the constraints that allow a native speaker to decide whether some sequence of words is a well-formed sentence of their language. More recent systems attempt to infer these rules from sample data, usually in the form of collections of dependency trees. Discuss the advantages and disadvantages of both approaches. [9 marks]

b) Describe the major data structures and the four main operations used in Nivre’s deterministic parsing algorithm. [7 marks] Give a brief account of the time complexity of this algorithm. [4 marks]

c) Show the sequence of operations that would obtain the dependency tree in Figure 4 from the sentence ‘I saw a man in it’. [10 marks]

```
saw
 / 
mans
 / 
I

in

a

it
```

Figure 4: Dependency tree for ‘I saw a man in it’

d) Explain how machine learning can be used to obtain rules which will choose which action to perform at each stage. Show a rule that might be extracted from the sequence of actions that you carried out in part (3c) for deciding to attach a determiner to a noun. [5 marks]
4. a) What is a ‘vector space’ model of lexical meaning? [5 marks] What is the role of the ‘context’ in such models? What kinds of things can be used as contexts? [5 marks]

b) Consider the following passages: using a window of the two preceding and two following words as the context, determine whether ‘mouse’ is more similar to ‘hockey’ or to ‘rodent’. [8 marks]. What are the advantages and disadvantages of using this notion of context when compared to using the entire document in which a word appears and to using its syntactic parent. [10 marks]

A mouse is a small mammal belonging to the order of rodents, characteristically having a pointed snout, small rounded ears, and a long naked or almost hairless tail. The best known mouse species is the common house mouse. In some places, certain kinds of field mice are also common.

Field hockey is a team sport in which a team of players attempts to score goals by hitting, pushing or flicking a ball into an opposing team’s goal using sticks. In some countries, it is known simply as ”hockey”; however, the name field hockey is used in countries in which the word hockey is generally reserved for another form of hockey, such as ice hockey, street hockey or roller hockey.

c) How might you use a vector space model to realise that a word had multiple meanings, and how could you then decide which interpretation was intended in a given context? [7 marks]
5. a) Describe phone-based and diphone-based speech synthesis, and explain why diphone-based synthesisers usually produce better quality output than phone-based synthesisers. [8 marks] What parameters do you need to control in order to produce natural sounding speech, and how would you find appropriate values to choose for these parameters? [7 marks]

b) Most speech recognition systems use a non-recursive context-free grammar (NR-CFG) to constrain what can possibly be said, and a hidden Markov model (HMM) to work out the most likely interpretation of a given speech sequence. Describe the relationship between the NR-CFG and the HMM, and explain what emission probabilities and transition probabilities are. You should use the NR-CFG in Figure 5 and the HMM in Figure 6 (overleaf) to illustrate your answer. [10 marks]

\[
\begin{align*}
\text{npsubj} = & \text{ he } | \text{ she; } \\
\text{verb} = & \text{ loves } | \text{ hates; } \\
\text{npobj} = & \text{ it } | \text{ them; } \\
\text{vp} = & \text{ verb npobj; }
\end{align*}
\]

($\text{npsubj} \text{ vp})

Figure 5: Simple CFG

c) Find the most likely path through the HMM in Figure 6 (answers that approximate calculations to one significant figure are acceptable, e.g. using \(0.9 \times 0.7 \approx 0.6\) rather than \(0.9 \times 0.7 = 0.63\)). [10 marks]
Figure 6: Simple HMM