Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Natural Language Systems

Date: Tuesday 24th January 2012
Time: 09:45 - 11:45

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 permitted provided they are not programmable and do not store text.

[PTO]
You should answer question 1: each part of this question carries 5 marks

1. a) What is a ‘lexical trie’? What are the advantages of using a lexical trie rather than a simple list of words to represent your lexicon?

b) What is the difference between stemming and full morphological analysis? Why is full morphological analysis impossible if you do not have a lexicon?

c) Given the descriptions of ‘noir’, ‘e’ and ‘s’ as adj/gen, gen/num and num respectively, show how the standard categorial rule \( X \Rightarrow X/Y, Y \) and the ‘raising’ rule \( X/Z \Rightarrow X/Y, Y/Z \) let you analyse the form ‘noires’ from left to right.

d) Complex regular expressions can be made by adding together simpler ones, e.g. if
\[
\text{noun} = "(<\text{tag NN}>.*\)", \text{adj} = "(<\text{tag ADJ}>.*\)", \text{det} = "(<\text{tag DET}>.*\)"
\]
then you could make a regular expression for describing NPs like ‘the brown tea pot’ by setting
\[
\text{np} = \text{det}+"?"+"(\text{+noun+\} \text{+adj+\})*+\text{noun}
\]
Why is it not possible to write a regular expression that will cover recursive structures, such as ‘the man with a big nose’s wife’, in this way?

e) What are formants? What kinds of sound are distinguished by their formants?

f) What is concatenative speech synthesis? What are the advantages and disadvantages of using diphones rather than individual words in a concatenative speech synthesiser?
Section B

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

2. a) What is ‘part of speech tagging’? Outline how a word-based tagger and an affix-based tagger work. [5 marks] Explain what is meant by ‘back-off’, and describe how you would use back-off to combine a word-based tagger and an affix-based tagger to make a single system that could outperform each of the constituents. [5 marks]

b) The performance of taggers is generally measured by giving their precision, recall and F-measure scores. Define precision, recall and F-measure, [3 marks] and say, with justification, whether the back-off combination of word-based and affix-based taggers would have higher recall or higher precision than each of the constituents. [7 marks]

c) Transformation-based learning can be used to infer corrective rules from hand-corrected output of a basic tagger. Given the set of ‘rule schemas’ in Figure 1 below, explain how the rule that achieves the greatest net benefit would be obtained for each of the two hand-corrected sets of tags in Figure 2. [10 marks] What would be the net improvement obtained by each of these rules? [5 marks]

\[
\begin{align*}
t_1(A,B,C) & \quad \text{# tag:A>B <- tag:C@[-1].} \\
t_2(A,B,C) & \quad \text{# tag:A>B <- tag:C@[1].} \\
t_{12}(A,B,W) & \quad \text{# tag:A>B <- wd:W@[0].} \\
t_{13}(A,B,W) & \quad \text{# tag:A>B <- wd:W@[1].} \\
t_{14}(A,B,W) & \quad \text{# tag:A>B <- wd:W@[1].}
\end{align*}
\]

Figure 1: Rule schemas for transformation-based tagging

<table>
<thead>
<tr>
<th>tagger output</th>
<th>corrected tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>the cat sat on the rough mat</td>
<td>noun noun noun noun noun noun noun</td>
</tr>
<tr>
<td>det noun verb prep det adj noun</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>tagger output</th>
<th>corrected tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>each girl gave her friend a present</td>
<td>noun noun noun noun noun noun noun</td>
</tr>
<tr>
<td>det noun verb det noun det noun det noun</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Tagger output and hand-corrected tags
3. a) i) Speech recognition systems typically use a hidden Markov model (HMM) to estimate the most probable word sequence given an input sound file. Figure 3 below shows the architecture of an HMM-based system for recognising whether someone said ‘cat’, ‘cot’ or ‘cod’. Using this diagram as a reference, explain the terms ‘transition probability’ and ‘emission probability’. [6 marks]

   ![HMM Diagram](image)

   **Figure 3: HMM for ‘cat’, ‘cot’, ‘cot’**

ii) Such systems usually require a context-free grammar and a set of phonetic transcriptions. Explain the role of these two items, and suggest a grammar and set of transcriptions that might have been used in the current case. [8 marks]

b) HMM-based speech recognition systems need to be trained. Describe the training data required by such a system and explain the role of each element. [8 marks]

c) Explain the difference between ‘homonymy’ and ‘ambiguity’, and explain what an ‘allophone’ is. Illustrate your answer with examples from any language with which you are familiar. [7 marks]. Why might a speech understanding system use a lexicon which is based on phonetic transcriptions rather than the standard orthographic representations? [6 marks]
4. 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) The regular expression given in Figure 4 is intended to pick out verb object pairs from text which has been tagged by preceding each word by its tag inside <...> brackets, e.g. for

- I was going to eat a pear
  <PNP>I<VBD>was<VVG>going<TO0>to<VVI>eat<AT0>a<NN1>pear

it would return eat as the value for verb and pear as the value for object.

"<VV.>(?P<verb>[^<]*)(<AT.>[^<]*)?(<AJ.>[^<]*)?<NN.>(?P<object>[^<]*)"  

Figure 4: Regex for finding verb-object pairs

Show the verb-object pairs that this regular expression would return for each of the following cases and in each case explain why this pair would not be useful if you were trying to work out what kinds of thing can be eaten. [8 marks] Fix the regular expression so that it would get the right pair for one of the first three. [5 marks]

(1) My kids will only eat shop bread
    <DPS>My<NN2>kids<VM0>will<AV0>only<VVI>eat
    <NN1>shop<NN1>bread

(2) My kids will only eat sliced white bread
    <DPS>My<NN2>kids<VM0>will<AV0>only<VVI>eat
    <VVD>sliced<AJ0>white<NN1>bread

(3) I saw her eat an entire box of chocolates
    <PNP>I<VVD>saw<DPS>her<VVI>eat<AT0>an
    <AJ0>entire<NN1>box<PRF>of<NN2>chocolates

(4) she ate a cake
    <PNP>she<VVD>ate<AT0>a<NN1>cake
c) i) Assuming that ‘He was eating an apple’ entails ‘An apple was eaten’, what would you have to do in order to obtain a textual entailment rule from the trees in Figure 5? What would this rule look like? [8 marks]

ii) What extra information would you need if you wanted to use this rule to infer ‘Some meat was eaten’ from ‘She ate some lamb’? [4 marks]

Figure 5: Dependency trees
5. a) Describe the use of ‘transfer rules’ in machine translation, and explain where systems that use such rules belong on the ‘machine translation pyramid’? [8 marks]

The amount of detail in a transfer rule can range from simple part-of-speech tags to fully inflected lexical items. Explain how you might use transfer rules which simply contained part-of-speech tags alongside a dictionary of bilingual equivalents, and discuss the advantages and disadvantages of this approach to machine translation. You should illustrate your answer with simple examples. [9 marks]

b) Describe how you might extract a bilingual lexicon from a corpus of translated material. [9 marks] How might you spot that a word in the source language was ambiguous during this process, and what further information could you extract to help with choosing the right translation for such words? [9 marks] You should illustrate your answers by using the passages below (the ‘French’ passage was obtained from an existing MT system, which is why it contains a number of errors and infelicities).

- The laws of cricket are a set of rules established by the Marylebone Cricket Club (MCC) which describe the laws of cricket worldwide, to ensure uniformity and fairness. Since the introduction of the 2000 Code, the Spirit of Cricket Preamble has been an important feature, providing the context in which the game is intended to be played. One of the commonest questions asked about grasshoppers and crickets is how to tell them apart. The main difference between a grasshopper and a cricket is that crickets tend to have long antennae, grasshoppers have short antennae.

- Les lois du cricket sont un ensemble de règles établies par le club de cricket de Marylebone (MCC) qui décrivent les lois du cricket dans le monde entier, pour assurer l’uniformité et l’équité. Depuis l’introduction des 2000 codes, l’esprit du péambule de cricket a été un dispositif important, fournissant le cadre dans lequel le jeu est prévu pour être joué. Une des questions les plus communes posées sur des sauterelles et des grillons est comment les distinguer. La principale différence entre un sauteur et un grillon est que les grillons ont tendance à avoir de longues antennes, des sauterelles ont les antennes courtes.