UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2004

MEng Honours Degree in Information Systems Engineering Part IV
MEng Honours Degrees in Computing Part IV
MSc in Advanced Computing
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant examinations for the Associateship of the City and Guilds of London Institute

PAPER C485=I4.24

NATURAL LANGUAGE PROCESSING

Wednesday 28 April 2004, 14:30
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators not required

1 Michael Covington's 1994 text, *Natural Language Processing for Prolog Programmers*, distinguishes as *levels of linguistic analysis*:

i) Phonology,

ii) Morphology,

iii) Syntax,

iv) Semantics, and

v) Pragmatics.

a Briefly explain what is meant by a *level of linguistic analysis* and the computational motivation for the classification.

b For each of these "*levels*", summarize, using examples, the main problems that are addressed, stating the computational methods that are typically employed.

c Conclude by briefly discussing how sign languages, traditional grammar, and communications at an auction might be classified.

Parts a, b, and c carry, respectively, 10%, 75%, 15% of the marks

2a Using a diagram, briefly describe the noisy channel model of communication and give three areas of natural language processing to which it is pertinent in practice.

b Explain Bayes's rule for estimating the conditional probability *prob*(A|B) in terms of *prob*(B|A), indicating how it is interpreted in the noisy channel model. Illustrate the use of Bayes' rule to estimate the most likely correct word *c*, for an observed typing error *t*, given the table of estimates below:

c	*prob*(c)	*prob*(t|c)
C1	0.00003	0.0001
C2	0.00001	0.00001
C3	0.00006	0.000002
C4	0.0007	0.00006

c Explain how a large corpus of words can be used to give the statistics for such use of Bayes' rule in spelling correction, indicating the underlying assumptions and how numerical difficulties can be eased.

d Consider the segmentation of the string below into words:

Therestartinstanceissoon.

Illustrate how one can distinguish the process of identifying possible word segments from the process of choosing the most appropriate segmentation. Briefly explain how the second process could be modelled and solved using statistical methods, explain the practical difficulty with this approach, and suggest an alternative approach to the overall problem of string segmentation.

Parts a, b, c, and d, carry, respectively, 30%, 30%, 20%, and 20% of the marks.

3a Briefly explain how a Definite Clause Grammar (DCG) differs from a Context Free Grammar (CFG). What is the advantage of a DCG?

b Show how a DCG can be used to express the features of the sentence,

she catches them quickly

and give an example of a grammatically incorrect variant that is eliminated by the DCG but not by a similar CFG.

c Briefly describe how each constituent of a DCG is translated into Prolog, and explain the consequences for parsing.

d There is evidence that human parsing is sensitive to lexical sub-categorisation. Briefly explain what this means, giving an example. Briefly explain why a parser based on Earley's method could help deal with the problems of lexical sub-categorisation, and indicate the limitations of such a method.

Parts a, b, c, and d carry, respectively, 15%, 30%, 30%, and 25% of the marks.

4a In which way are the standard quantifiers of predicate logic a mismatch for the composition of sentences with determiners or specifiers in traditional grammar? Illustrate your answer by considering the logical form of *some cat scratches*, and of *no cat barks*. Explain how generalised quantifiers can be used to overcome this mismatch, giving Prolog-style meta-logical definitions for *some* and *no* as quantifiers. Illustrate how the generalised quantifier meaning of *some cats scratch* is composed from the logical representations of its parts.

b Generalised quantifiers retain the classical tradition of a sentence having a closed logical form. Briefly explain why is this a problem for representing the meaning of discourse and dialogue, and how it may be overcome. Consider as an example the sequence of sentences.

He was beaten every time. After the last occasion Sam enrolled for golf lessons.

Parts a and b carry, respectively, 60%, and 40% of the marks.