UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2003

BEng Honours Degree in Computing Part III
MEng Honours Degree in Information Systems Engineering Part IV
MEng Honours Degrees in Computing Part IV
MSc in Advanced Computing
PhD
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

Monday 28 April 2003, 10:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators not required

1 Daniel Jurafsky and James H. Martin in their book, *Speech and Language Processing*, begin with the following quotation from the screenplay *2001: A Space Odyssey* by Stanley Kubrick and Arthur C. Clarke.

Dave Bowman: Open the pod bay doors, HAL
HAL: I'm sorry Dave, I'm afraid I can't do that.

a List and briefly explain the distinct processing levels required for (the computer system) *HAL* to engage in such discourse.

b Explain why HAL's creator was too optimistic in predicting that such a computational agent would be available in 2001, by describing the present state of the art in processing each level at which language is processed.

c Briefly explain both the justification and the restricted vision implicit in the quotation below, providing examples of statistical methods at different processing levels.

> *But it must be recognized that the notion of "probability of a sentence" is an entirely useless one, under any known interpretation of this term.*
>
> – Noam Chomsky, 1969

The three parts carry, respectively, 30%, 40%, 30% of the marks

2a What is the advantage of using a Context Free Grammar to describe natural language, in comparison with Template matching on a list of words?

b In what sense does a Context Free Grammar "over-generate". What is the advantage of a Definite Clause Grammar in this respect.

c Briefly describe how a DCG grammar rule, such as:

```
s --> np(num), vp(num).
```

is translated into Prolog. Identify the main advantage that comes from this translation of a DCG rule, and give a disadvantage.

d The following DCG grammar enables the question:

who did john see

to be parsed by the Prolog inference engine. The DCG grammar is:

```
s(In,Out)-->[who,did],np([who|In],O1),vp([did|O1],Out).
np([who|Out],Out) --> [].
np(X,X) --> [John];[Joe].
vp(X,X) --> [did see].
vp([did|I1],Out) --> [see], np(I1,Out)
```

Explain the purpose of the lists `In` and `Out`. Show a trace of the parse. What is the phenomenon which breaks the usual phrase structure rule here? Is there a more principled way of describing the phenomenon?

The four parts carry, respectively, 10%, 15%, 35%, and 40% of the marks.

3a Briefly explain what is meant by a generalised quantifier. Illustrate your answer by providing meta-level computational definitions in Prolog style for *some*, *all*, and *no*. What assumption is made regarding the underlying semantic form?

b Briefly compare and contrast the use and success of generalised quantifier semantics for sentences with the task of representing the meaning of discourse, indicating how inter-sentential anaphora could be dealt with.

c Briefly explain the idea of thematic relations for structuring knowledge of an event, indicating the motivation for this sort of semantic representation and illustrating its use in explicating the sentences:

Brutus stabbed Caeser with a knife. Mary saw this, but did not know it.

d Briefly explain Vendler's verb aspectual classes.

The three parts carry, respectively, 30%, 20%, 30% and 20% of the marks.

4a Covington illustrates a simple case of Machine Translation by re-expressing

A dog chased the cat

in Latin, as:

Canis iste felum agitavit
(Dog the cat chased)

Here the noun *Canis* (dog) is indefinite, nominative and masculine, the determiner *iste* ('the') is demonstrative, accusative and feminine, the noun *felum* ('cat') is indefinite, accusative and feminine, and the verb *agitavit* ('chased') is transitive and in the past tense.

Explain why this translation is "simple" by giving a parse tree for each and some indication of how such simple translations might be achieved automatically.

b Arnold gives the following translations. Explain why each would require different technology if its translation were to be achieved automatically:

i) *I miss London*

ii) *They hammered the metal flat*

iii) *I have a hangover*

In French:

i) *Londres manques á moi*
(London misses to me)

ii) *Ils ont martelé le metal jusqu'a ce qu' il est devenu plat*
(They hammered the metal until it became flat)

iii) *J'ai la gueuelle de bois*
(I have the face of wood)

c Briefly identify the situations where machine translation is effective.

d Suppose that a "noisy channel" has distorted English input into French. We wish to recover the English text from the French speaker. Express this as a maximum likelihood estimation and explain the probabilities in the maximising expression.

The four parts carry, respectively, 35%, 30%, 10% and 25% of the marks.