UNIVERSITI SAINS MALAYSIA

First Semester Examination
Academic Session 2001/2002

September 2001

CCS503 – Intelligent Document Processing

CSC514 – Natural Language Processing

Duration : 3 hours

INSTRUCTION TO CANDIDATE:

- Please ensure that this examination paper contains **FOUR** questions in **FOUR** printed pages before you start the examination.
- Answer **ALL** questions.
- This is an Open Book Examination.
- You can choose to answer either in Bahasa Malaysia or English.

ENGLISH VERSION OF THE QUESTION PAPER

1. Given the following sentence:

"Tom likes stars"

(a) How many phonemes are there in the given sentence? Write down the pronunciation of the sentence given above in IPA symbols. (Note: 's' at the end of a word might be pronounced differently.)

[35/100]

(b) In English, the phoneme [t] is pronounced differently in different contexts. For example, consider the different pronunciations of [t] in the words *tunafish* and *starfish*. The [t] of *tunafish* is **aspirated**, i.e. wherever a [t] begins a word, it is aspirated and its corresponding phone is represented as [t^h]. By contrast, the [t] of *starfish* is **unaspirated**, i.e. a [t] following an initial [s] is unaspirated and its corresponding phone is represented as [t]. Provide an appropriate phonological rule for this phenomenon. Rewrite the pronunciation of the sentence given in 1(a) by applying this rule.

[25/100]

(c) TTS (Text-To-Speech) and STT (Speech-To-Text) systems can use either the method implied in (b), i.e. spelling $\leftrightarrow$ phonemes $\leftrightarrow$ phones, or the more direct spelling $\leftrightarrow$ phones approach based on a dictionary storing two phonetic representations. What are the advantages and disadvantages of each approach for both TTS and STT.

[40/100]

2. Given the following sentence:

"The students like female sounds”

(a) Find all the words with inflectional morphology and identify their morphemes. Describe a possible implementation of this process.

[20/100]

(b) Find all possible parts of speech for each word in the given sentence.

[15/100]

(c) Draw all possible English syntactic tree(s) for the given sentence.

[15/100]

(d) Provide a Context Free Grammar which is capable of generating all the syntactic tree(s) or parse tree(s) as identified in 2(c).

[20/100]

(e) Construct a chart produced by the bottom-up chart parsing technique for the sentence given above.

[30/100]

3. Given the following grammar:

S → noun VP
VP → verb
VP → verb noun
noun → {noun}
verb → {verb}

lexicon:
cows : noun
grass : noun
eat : verb
sleeps : verb

(a) List all sentences generated by the above grammar which are gramatically incorrect (in English).

[15/100]

(b) Extend the grammar and lexicon given above so that only gramatically correct (in English) sentences will be generated.

[35/100]

(c) List all gramatically correct sentences generated by the above grammar which are semantically invalid (in English).

[15/100]

(d) Extend the grammar and lexicon given in 3(b) so that only semantically valid sentences will be generated.

[35/100]

4. (a) For each of the following NLP tools, describe its functionality and give an example input/output pair.

(i) Spelling checker

(ii) Inflectional morphological analyzer

(iii) Part of speech tagger

(iv) Parser

(v) Word sense tagger

[40/100]

(b) Discuss in detail how the NLP tools in 4(a) can be applied to each of the following NLP applications.

(i) Information retrieval or Information search

(ii) Text categorization/classification

(iii) Machine translation

[60/100]

- oooOooo -