UNIVERSITI SAINS MALAYSIA

Peperiksaan Semester Pertama
Sidang Akademik 2001/2002

September 2001

CCS503 – Pemprosesan Dokumen Cerdas

CSC514 – Pemprosesan Bahasa Tabii

Masa : 3 jam

ARAHAN KEPADA CALON:

- Sila pastikan bahawa kertas peperiksaan ini mengandungi **EMPAT** soalan di dalam **EMPAT** muka surat yang bercetak sebelum anda memulakan peperiksaan ini.
- Jawab **SEMUA** soalan.
- Peperiksaan ini akan dijalankan secara 'Open Book'.
- Anda boleh memilih untuk menjawab semua soalan dalam Bahasa Malaysia atau Bahasa Inggeris.

1. Diberikan ayat berikut:

"Tom likes stars"

(a) Berapakah fonem yang wujud dalam ayat di atas? Tuliskan sebutan ayat di atas dalam symbol IPA. (Perhatian: ‘s’ di akhir perkataan mungkin disebut dengan cara yang berlainan.)

[35/100]

(b) Dalam bahasa Inggeris, fonem [t] disebutkan secara berlainan bergantung kepada konteksnya. Misalnya, pertimbangkan perbezaaan sebutan [t] untuk perkataan *tunafish* dan *starfish*. Sebutan [t] untuk *tunafish* adalah ‘**aspirated**’, iaitu, semua [t] yang wujud di permulaan perkataan adalah ‘aspirated’ dan ia diwakili dengan fon [t^h]. Sebaliknya, sebutan [t] untuk *starfish* adalah ‘**unaspirated**’, iaitu, sebutan [t] berikutan sebutan [s] (yang wujud di permulaan perkataan) adalah ‘unaspirated’ dan diwakili dengan fon [t]. Berikan petua fonologi untuk menakrifkan fenomena ini. Tulis semula sebutan ayat di dalam 1(a) mengikut pelaksanaan petua ini.

[25/100]

(c) Sistem TTS (Text-To-Speech) dan STT (Speech-To-Text) sama ada menggunakan kaedah seperti yang ditunjukkan dalam 1(b), iaitu ejaan $\leftrightarrow$ fonem $\leftrightarrow$ fon, ataupun kaedah ejaan $\leftrightarrow$ fon yang lebih langsung berasaskan kamus yang menyimpan dua perwakilan fonetik. Apakah kebaikan dan keburukan bagi setiap kaedah ini bagi TTS dan STT?

[40/100]

2. Diberikan ayat berikut:

"The students like female sounds”

(a) Cari semua perkataan dalam bentuk ‘inflectional morphology’ dan camkan morfem semua perkataan tersebut. Huraikan suatu cara implimentasi untuk pemprosesan ini.

[20/100]

(b) Cari semua kemungkinan ‘parts of speech’ bagi setiap perkataan dalam ayat yang diberikan di atas.

[15/100]

(c) Lakarkan semua ‘English syntactic tree’ yang mungkin bagi ayat di atas.

[15/100]

(d) Berikan suatu nahu bebas konteks yang berkeupayaan untuk menjanakan semua 'syntactic tree' atau 'parse tree' seperti yang diberikan dalam 2(c).

[20/100]

(e) Janakan suatu 'chart' yang dihasilkan oleh teknik 'bottom-up chart parsing' bagi ayat yang diberikan di atas.

[30/100]

3. Diberikan nahu berikut:

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

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

(a) Senaraikan semua ayat tidak gramatis (dalam Bahasa Inggeris) yang dapat dijanakan oleh nahu di atas.

[15/100]

(b) Kembangkan nahu dan leksikon seperti yang diberikan di atas supaya hanya ayat yang gramatis (dalam Bahasa Inggeris) sahaja dapat dijanakan.

[35/100]

(c) Senaraikan semua ayat yang gramatis tetapi tidak tepat secara semantik (dalam Bahasa Inggeris) yang akan dijanakan oleh nahu di atas.

[15/100]

(d) Kembangkan nahu dan leksikon seperti yang diberikan dalam 3(b) supaya hanya ayat yang tepat secara semantik sahaja yang akan dijanakan.

[35/100]

4. (a) Untuk setaip alat NLP berikut, jelaskan fungsinya dan berikan suatu contoh input/output.

(i) ‘Spelling checker’

(ii) ‘Inflectional morphological analyzer’

(iii) ‘Part of speech tagger’

(iv) ‘Parser’

(v) ‘Word sense tagger’

[40/100]

(b) Bincangkan secara menyeluruh bagaimana alat NLP dalam 4(a) dapat digunakan untuk membangunkan applikasi NLP berikut.

(i) ‘Information retrieval’ atau ‘Information search’

(ii) ‘Text categorization/classification’

(iii) ‘Machine translation’

[60/100]

- oooOooo -