UNIVERSITI SAINS MALAYSIA

Peperiksaan Semester Pertama
Sidang Akademik 2003/2004

September/Oktober 2003

CCS503 – Pemprosesan Dokumen Cerdas

Masa : 3 jam

ARAHAN KEPADA CALON:

- Sila pastikan bahawa kertas peperiksaan ini mengandungi **LIMA** soalan di dalam **LIMA** 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:

"The small bat flied slowly"

(a) Berapakah fonem yang wujud dalam ayat di atas? Tuliskan sebutan ayat di atas dalam simbol IPA.

[15/100]

(b) Cari semua 'parts of speech' yang mungkin bagi setiap perkataan dalam ayat di atas.

[10/100]

(c) Lakarkan semua pohon sintak English yang sah untuk ayat di atas.

[10/100]

(d) Berikan suatu Nahu Bebas Konteks yang dapat menjanakan semua pohon sintak yang diberikan di 1(c).

[15/100]

(e) Anggapkan sesuatu sistem STT (Speech-To-Text) berasaskan cara (fonem $\rightarrow$ ejaan) tidak dapat membezakan fonem untuk pasangan askara berikut bagi English:

(d,t), (r,l), (n,m) dan (b,p)

(i) Untuk setiap perkataan dalam ayat yang diberikan di atas, senaraikan semua perkataan English lain yang tidak dapat dibezakan oleh sistem STT. Jelaskan bagaimana 'English spell checker' dapat digunakan untuk membantu sistem STT dalam proses pengenalan semua perkataan English dari bentuk sebutan bagi ayat yang diberikan di atas.

(ii) Jelaskan bagaimana nahu yang diberikan di 1(d) dapat digunakan untuk membantu sistem STT memilih perkataan yang betul dalam proses pengenalan dari bentuk sebutan bagi ayat yang diberikan di atas.

(iii) Kembangkan nahu yang diberikan di 1(d) supaya ia dapat membantu sistem STT membuang ayat-ayat yang tidak sah dari segi semantik dalam proses pengenalan dari bentuk sebutan bagi ayat yang diberikan di atas.

[50/100]

2. Soalan ini mempunyai dua bahagian. Kedua-dua bahagian MESTI dijawab.

Apabila dua kaum yang berlainan bahasa hidup bersama bagi suatu tempoh yang lama, peminjaman kata antara satu sama lain tidak dapat dielakkan. Negeri Melaka yang kini sebahagian daripada negara Malaysia pernah ditakluki negeri Portugal selama 130 tahun. Dilaporkan bahawa hanya lebih kurang 15,000 orang keturunannya hari ini dapat berkomunikasi dalam suatu varian bahasa Portugis, iaitu bahasa Portugis kreol. Dilaporkan juga bahawa lebih kurang 5% daripada kata-kata bahasa kreol ini dipinjam daripada bahasa Inggeris dan bahasa Melayu. Sebaliknya, lebih kurang 450 daripada kata-kata bahasa Melayu dikatakan dipinjam daripada bahasa Portugis. Kita kenal hanya lebih kurang 30 patah kata (lihat Jadual 1).

Jadual 1. Senarai kata-kata Portugis dan padanannya dalam bahasa Melayu

Bahasa Portugis	Bahasa Melayu	Bahasa Portugis	Bahasa Melayu
bandeira	bendera	manteiga	mentega
bomba	bomba	mesa	meja
boneca	boneka	natal	natal
camisa	kemeja	queijo	keju
carreta	kereta	ronda	ronda
dado	dadu	sabado	sabtu
escola	sekolah	saco	saku
garfo	garfu	sapato	sepatu
igrega	gereja	tangue	tangki
jandela	jendela		

Tidak adanya kamus bahasa Portugis-Melayu, dan tidak ramainya penutur yang dapat berbahasa Portugis dan bahasa Melayu.

(a) Memandangkan hakikat ini, gunakan senarai kata yang diberikan di Jadual 1, dan tuliskan sekurang-kurangnya TIGA jenis petua yang dapat mengekstrakkan kata-kata dalam bahasa Melayu yang berkemungkinan besar dipinjam daripada bahasa Portugis? Satu petua mungkin mengendalikan vokal, dan satu lagi konsonan. Bagi petua yang mengendalikan vokal, petua tersebut mungkin mengendalikan vokal pada bahagian awal, tengah ataupun hujung sesuatu perkataan. Petua-petua mesti ditulis dalam bentuk berikut. Bagi setiap petua, berikan satu contoh yang dipetik dari kata-kata yang disenaraikan di atas.

Petua: rentetan_sumber → rentetan_sasaran | syarat(-syarat)
Misalnya: *Kata_Portugis* → *Kata_Melayu* | _

[Dalam transkripsian kata-kata bahasa Inggeris yang berakhiran *-graphy* kepada bahasa Melayu, kita boleh menulis petua seperti berikut:

Petua: *ph* → *f* | _
Misalnya: *geography* → *geografi* | _]

[90/100]

(b) Berikan satu dua komen terhadap semantik kata-kata yang dipinjam ini.

3. Diberikan nahu berikut bagi Bahasa Malaysia:

Ayat → FN FKK
FKK → KK
FKK → KK FN
FN → KN KS
KN → {Kata_Nama}
KS → {Kata_Sifat}
KK → {Kata_Kerja}

leksikon:
kucing : Kata_Nama
bola : Kata_Nama
main : Kata_Kerja
merah : Kata_Sifat

(a) Senaraikan semua ayat yang tepat secara gramatis (dalam Bahasa Malaysia) yang dapat dijanakan oleh nahu di atas.

[15/100]

(b) Senaraikan semua ayat yang gramatis tetapi tidak tepat secara semantik (dalam Bahasa Malaysia) yang dapat dijanakan oleh nahu di atas.

[15/100]

(c) Kembangkan nahu dan leksikon seperti yang diberikan di atas supaya hanya ayat yang tepat secara semantik (dalam Bahasa Malaysia) sahaja dapat dijanakan.

[30/100]

(d) Janakan suatu 'chart' yang dihasilkan oleh teknik 'top-down prediction with bottom-up chart parsing' bagi ayat yang diberikan di bawah:

"Kucing merah main bola merah"

[40/100]

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'

[40/100]

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

(i) 'Keyword-based email filter'

(ii) 'Natural language question-answering system'

(iii) 'Example-based Machine translation'

[60/100]

- oooOooo -