UNIVERSITI SAINS MALAYSIA

First Semester Examination
2010/2011 Academic Session

November 2010

CCS512 – Language Engineering
[Kejuruteraan Bahasa]

Duration : 2 hours
[Masa : 2 jam]

INSTRUCTIONS TO CANDIDATE:
[ARAHAN KEPADA CALON:]

- Please ensure that this examination paper contains **FIVE** questions in **NINE** printed pages before you begin the examination.

  *[Sila pastikan bahawa kertas peperiksaan ini mengandungi **LIMA** soalan di dalam **SEMBILAN** muka surat yang bercetak sebelum anda memulakan peperiksaan ini.]*

- Answer **ALL** questions.

  *[Jawab **SEMUA** soalan.]*

- You may answer the questions either in English or in Bahasa Malaysia.

  [Anda dibenarkan menjawab soalan sama ada dalam bahasa Inggeris atau bahasa Malaysia.]

- In the event of any discrepancies, the English version shall be used.

  [Sekiranya terdapat sebarang percanggahan pada soalan peperiksaan, versi bahasa Inggeris hendaklah diguna pakai.]

1. Regular expression (RE) can be used to define regular language by specifying the string pattern, and it can be implemented as a finite-state automaton (FSA). FSA can be used to model a number of derivational facts, such as any verb ending in *–ize* can be followed by the nominalization of suffix *–ation*, and adjectives ending in *–al* can take the suffix *–ity*.

   (a) Write a regular expression that will match the string "any PC with more than 500MHz and 32 GB of disk space for less than $1000".

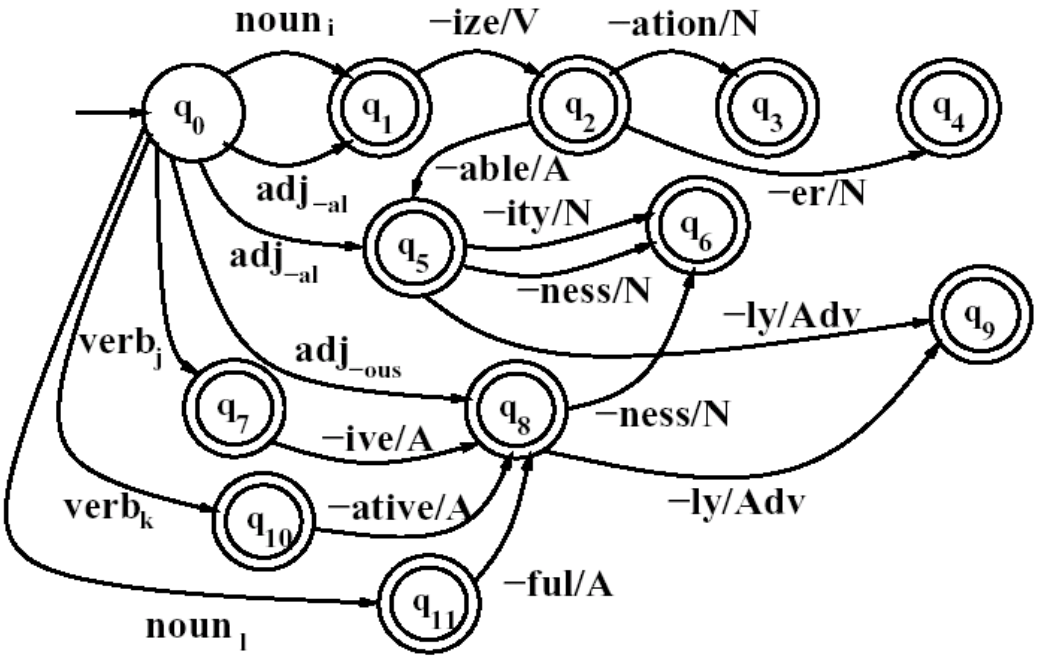
   (10/100)

   (b) By using the following FSA, give examples of each of the following:

      (i) noun class and its derivations;

      (ii) adjective class and its derivations;

      (iii) verb class and its derivations.



(25/100)

2. The standard phonetic transcription for representing the pronunciation of words is the International Phonetic Alphabet (IPA), while the most common computational system for the transcription of English is the ARPAbet.

(a) Transcribe the pronunciation of the following Malay words using IPA:

(i) dapat

(ii) suka

(iii) gereja

(iv) waktu

(v) lencongan

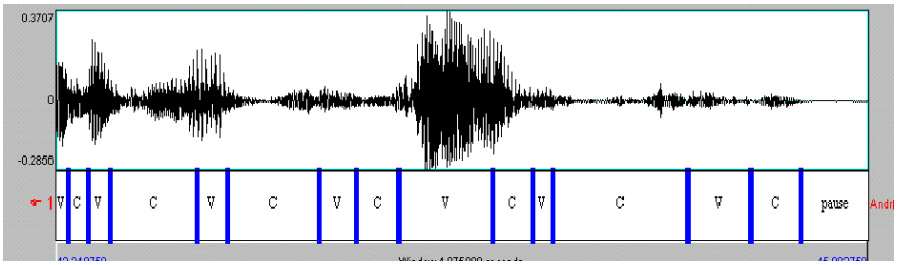
(10/100)

(b) Correct the mistakes in the ARPAbet transcriptions for the following words:

Three [dh r i]
Sing [s ih n g]
Study [s t uh d i]
Though [th ow]
Planning [p l aa n ih ng]

(5/100)

(c) VOCALE is a tool that can perform automatic annotation of vocalic and consonant interval (see figure below):



Note: V – vowel; C – consonant

Can this tool assist in the automatic annotation of syllable interval in speech data? If yes, what other kind of resources or data do you need in order to perform syllable interval annotation?

(5/100)

3. Dependency grammar is based on the relationship between a pair of words: the head (or governor) and the dependant. It is becoming quite important in speech and language processing.

(a) What are the advantages of using dependency grammar?

(5/100)

(b) Analyse the dependency structure of the following sentence:

The big dog chases the cat.

(i) with part-of-speech;

(ii) with functional categorisations;

(iii) with additional labeling.

(6/100)

(c) Show a typed dependency parse of the sentence given in question 3(b).

(4/100)

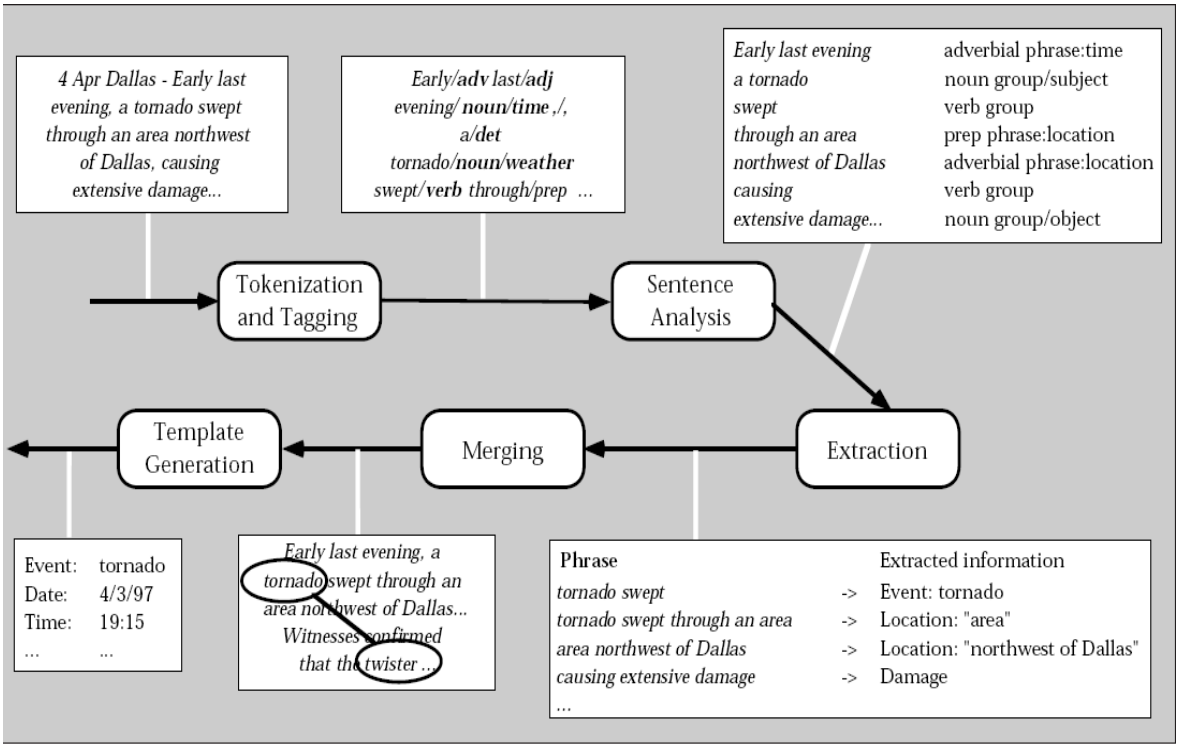
4. (a) Machine translation (MT) systems can be used to speed up the translation process by producing a draft translation before post-editing by a human translator. Give an example of a sublanguage domain (language which is used within a particular domain or subject matter) that can be modeled completely enough to use raw MT output without post-editing. Why?

(7/100)

(b) Explain why machine translation is hard.

(8/100)

5. The figure below shows different processes in a language engineering system:



(a) What kind of language engineering system is this? Provide one real application of this system.

(5/100)

(b) Explain the tasks performed by the:

(i) "Sentence Analysis" module;

(ii) "Template Generation" module.

(10/100)

1. Ungkapan nalar (RE) boleh digunakan untuk mendefinisikan bahasa nalar dengan menentukan pola rentetan dan ia boleh dilaksanakan sebagai automata keadaan-finitum (FSA). FSA boleh digunakan untuk memodelkan beberapa fakta terbitan seperti mana-mana kata kerja (bahasa Inggeris) yang berakhiran *–ize* boleh diikuti dengan nominalisasi akhiran *–ation,* dan adjektif yang berakhiran *–al* boleh mengambil akhiran *–ity*.

   (a) Tulis satu ungkapan nalar yang sepadan dengan rentetan "*any PC with more than 500MHz and 32 GB of disk space for less than $1000"*.

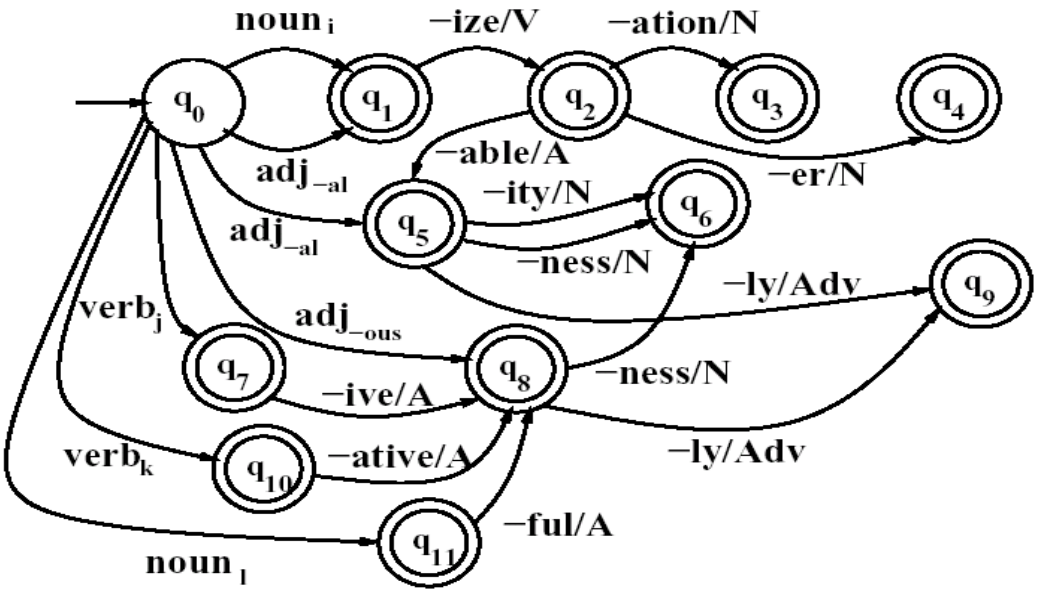
   (10/100)

   (b) Dengan menggunakan FSA berikut, berikan contoh bagi setiap:

      (i) kelas kata nama dan terbitannya;

      (ii) kelas kata adjektif dan terbitannya;

      (iii) kelas kata kerja dan terbitannya.



(25/100)

2. Transkripsi fonetik piawai yang mewakili sebutan perkataan ialah *International Phonetic Alphabet* (IPA), manakala sistem transkripsi berkomputer yang paling biasa bagi bahasa Inggeris ialah *ARPAbet*.

(a) Transkripsikan sebutan perkataan-perkataan bahasa Melayu yang berikut dengan menggunakan IPA:

(i) dapat

(ii) suka

(iii) gereja

(iv) waktu

(v) lencongan

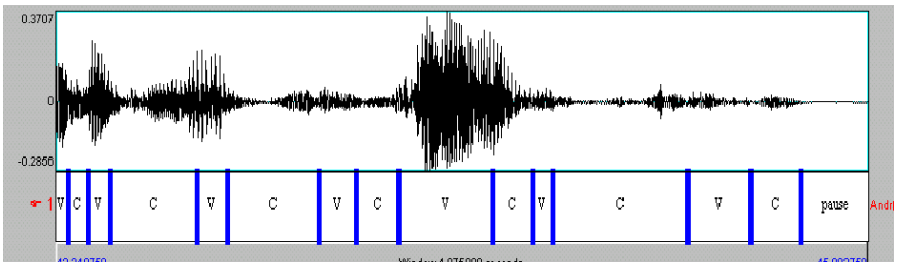
(10/100)

(b) Betulkan kesalahan dalam transkripsi ARPAbet bagi perkataan-perkataan berikut:

Three [dh r i]
Sing [s ih n g]
Study [s t uh d i]
Though [th ow]
Planning [p l aa n ih ng]

(5/100)

(c) VOCALE ialah alat yang boleh melakukan penganotasian secara automatik bagi jeda vokal dan konsonan (lihat rajah di bawah).



Nota: V – vokal; C – konsonan

Bolehkah alat ini digunakan dalam penganotasian secara automatik bagi jeda suku kata dalam data pertuturan? Jika boleh, apakah jenis sumber atau data yang anda perlukan untuk melakukan penganotasian bagi jeda suku kata?

(5/100)

3. Nahu pautan adalah berdasarkan hubungan antara pasangan perkataan: kepala (atau *governor*) dan pautan. Nahu ini menjadi nahu yang penting dalam pemprosesan pertuturan dan bahasa.

(a) Apakah kelebihan menggunakan nahu pautan?

(5/100)

(b) Analisa struktur pautan bagi ayat yang di bawah:

The big dog chases the cat.

(i) dengan kelas kata;

(ii) dengan kategori fungsian;

(iii) dengan label tambahan.

(6/100)

(c) Tunjukkan huraian pautan bertaip bagi ayat dalam soalan 3(b).

(4/100)

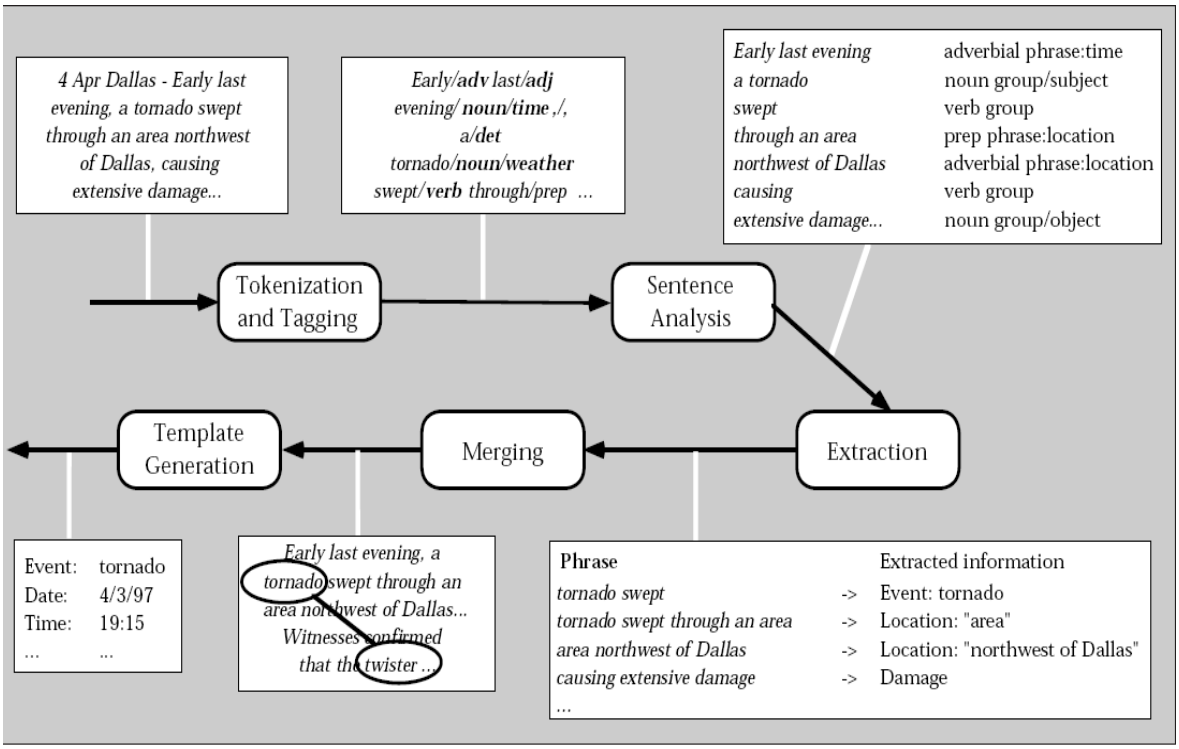
4. (a) Sistem terjemahan berkomputer (MT) boleh digunakan untuk mempercepatkan proses terjemahan dengan menghasilkan terjemahan draf sebelum penyuntingan akhir oleh penterjemah manusia. Berikan satu contoh domain sub-bahasa (bahasa yang digunakan untuk domain atau subjek tertentu) yang boleh dimodel dengan cukup lengkap untuk menggunakan output MT mentah tanpa perlu penyuntingan akhir. Kenapa?

(7/100)

(b) Terangkan kenapa terjemahan berkomputer (MT) sukar.

(8/100)

5. Rajah di bawah menunjukkan proses-proses berlainan dalam suatu sistem kejuruteraan bahasa:



(a) Apakah jenis sistem kejuruteraan bahasa ini? Berikan satu contoh kegunaan sebenar sistem ini.

(5/100)

(b) Terangkan tugas dilaksanakan oleh:

(i) Modul "*Sentence Analysis*";

(ii) Modul "*Template Generation*".

(10/100)

- oooOooo -