UNIVERSITI SAINS MALAYSIA

Second Semester Examination
2010/2011 Academic Session

April/May 2011

CPT346 – Natural Language Processing
[Pemprosesan Bahasa Tabii]

Duration : 2 hours
[Masa : 2 jam]

INSTRUCTIONS TO CANDIDATE:
[ARAHAN KEPADA CALON:]

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

  *[Sila pastikan bahawa kertas peperiksaan ini mengandungi **EMPAT** soalan di dalam **LIMA** 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. Below is an input-output format from an English morphology parser based on FST model.

f:f	o:o	x:x	+N:∈	+PL:^s#

Figure 1

(a) Based on Figure 1, fill-in the blanks of the FST two-level tapes below.

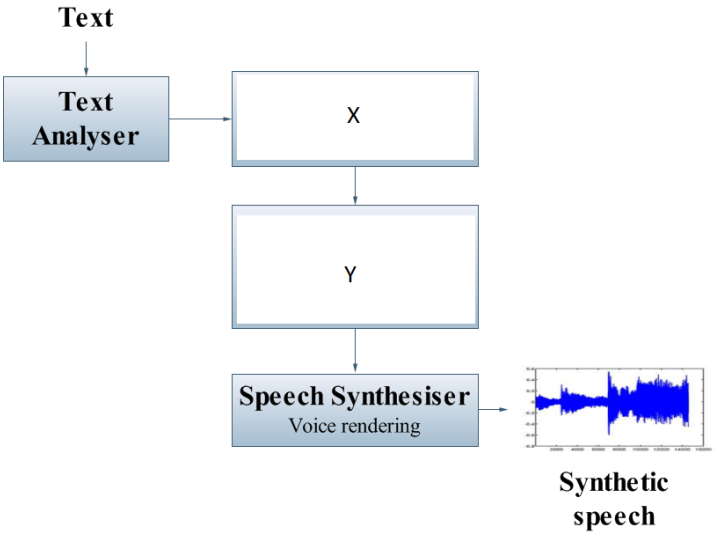
LEXICAL →						
INTERMEDIATE →						
SURFACE →						

(5/100)

(b) What do you need in order to build a morphology parser? List down and explain.

(20/100)

2. (a) Name **two** (**2**) main components, X and Y, required by the diagram below.



(5/100)

(b) What is a homograph? Name **two** (**2**) Natural Language Processing (NLP) processes that can be used to disambiguate (solve) homograph.

(20/100)

3. Answer all the questions below.

```
He wants to fish.
He plans to go home.
The two go fishing often.
He wants two gifts.
He loves to go swimming.
```

Based on the mini corpus above, show the calculation for the probability of the sentences below using bigram probability. (Note: You do not need to compute the probabilities values).

(a) He wants to go home.

(b) She plans to go fishing.

(25/100)

4. Determine whether the statements below are TRUE or FALSE. Explain briefly your answer.

(a) *Xerox* language identifier and *Google.translate* can recognise Jawi writing.

(b) *Espeak* can easily be built on mobile platform.

(c) It is difficult to segment word in Mandarin writing system.

(d) *Porter Stemmer* is built based on a full lexicon-based morphological parser.

(e) IPA characters are usually used as transcription symbols in the pronunciation dictionary of speech recognition system.

(25/100)

1. Di bawah ialah format input-output bagi penghurai morfologi Inggeris berdasarkan model FST.

f:f	o:o	x:x	+N:∈	+PL:^s#

Gambar Rajah 1

(a) Berdasarkan Gambar Rajah 1, isikan petak kosong pada dua-aras tape FST di bawah.

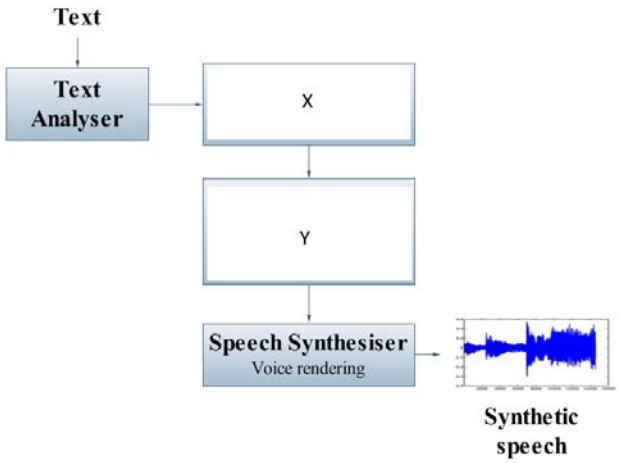
LEXICAL →						
INTERMEDIATE →						
SURFACE →						

(5/100)

(b) Apakah yang diperlukan untuk membina penghurai morfologi? Senaraikan dan terangkan secara ringkas.

(20/100)

2. (a) Namakan **dua** (**2**) komponen utama, X dan Y, yang diperlukan dalam gambar rajah di bawah.



(5/100)

(b) Apakah homograf? Namakan **dua** (**2**) proses Pemprosesan Bahasa Tabii (NLP) yang boleh digunakan untuk menyahtaksakan homograf.

(20/100)

3. Jawab semua soalan di bawah.

```
He wants to fish.
She plans to go home.
The two go fishing often.
He wants two gifts.
He loves to go fishing.
```

Berdasarkan korpus mini di atas, tunjukkan pengiraan kebarangkalian bagi ayat-ayat di bawah menggunakan kebarangkalian bigram. (Nota: Anda tidak perlu mengira nilai kebarangkalian-kebarangkalian tersebut).

(a) He wants to go home.

(b) She plans to go fishing.

(25/100)

4. Nyatakan jika kenyataan di bawah BETUL atau SALAH. Terangkan secara ringkas jawapan anda.

(a) Pengenalpasti bahasa *Xerox* dan *Google.translate* boleh mengenal tulisan Jawi.

(b) *Espeak* boleh dibina dengan mudah di atas pelantar bergerak.

(c) Adalah susah untuk memotong perkataan dalam sistem tulisan Mandarin.

(d) *Porter Stemmer* dibina berdasarkan penghurai morfologi yang berleksikon penuh.

(e) IPA aksara selalu digunakan sebagai simbol transkripsi dalam kamus sebutan pengecaman pertuturan.

(25/100)