UNIVERSITI SAINS MALAYSIA

First Semester Examination
2014/2015 Academic Session

December 2014/January 2015

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 **ELEVEN** printed pages before you begin the examination.

  *[Sila pastikan bahawa kertas peperiksaan ini mengandungi **EMPAT** soalan di dalam **SEBELAS** 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. Given below are some Malay complex words.

menulis, menuliskan, tulisan, penulisan, bertulis
bersembunyi, menyembunyikan, penyembunyian, persembunyian, sembunyian
membaca, membacakan, terbaca, keterbacaan, pembaca

(a) Segmentise the group of words to find their root or base form.

(7.5/25)

(b) Define the terms morpheme, affix, and derivational morphology and give examples.

(4.5/25)

(c) What is a finite state transducer (FST), and how it is used in computational linguistics? How does it differ from a finite state automaton?

(5/25)

(d) Draw an FST which could able output word associated with regular adverbs in English. Demonstrate that your FST correctly handles cases such as bright → brightly, simple → simply, silly → sillily and terrific → terrifically.

(8/25)

2. Consider the following context-free grammar:

S -> NP VP
NP -> Det N
VP -> V
VP -> V NP
N -> dog
N ->cat
N -> mouse
Det -> the
V -> sees
V -> hates
V -> sneezes

(a) Which of the following sentences are recognised by this grammar, and why?

(i) the dog sneezes the cat.

(ii) the mouse hates.

(iii) the cat the mouse hates.

(iv) the mouse hates the mouse.

(4/25)

(b) Modify the grammar so that the following sentence is now accepted by context-free grammar:

the dog the cat the mouse sees hates sneezes

Justify your choice.

(6/25)

(c) The semantics of natural language expressions can be expressed in first order predicate logic (FOPL). For instance, “the dog sneezes” can be approximately expressed as

$\exists x \, dog(x) \cap sneeze(x)$

Following this pattern, express the semantics of the sentence in part (b) in FOPL.

(1/25)

(d) Contrast this construction to the one in part (b) in terms of semantics and syntax. How would you modify the original grammar in part (a) to account for this construction?

(14/25)

3. Speech processing is a study of speech signals and the processing methods of these signals.

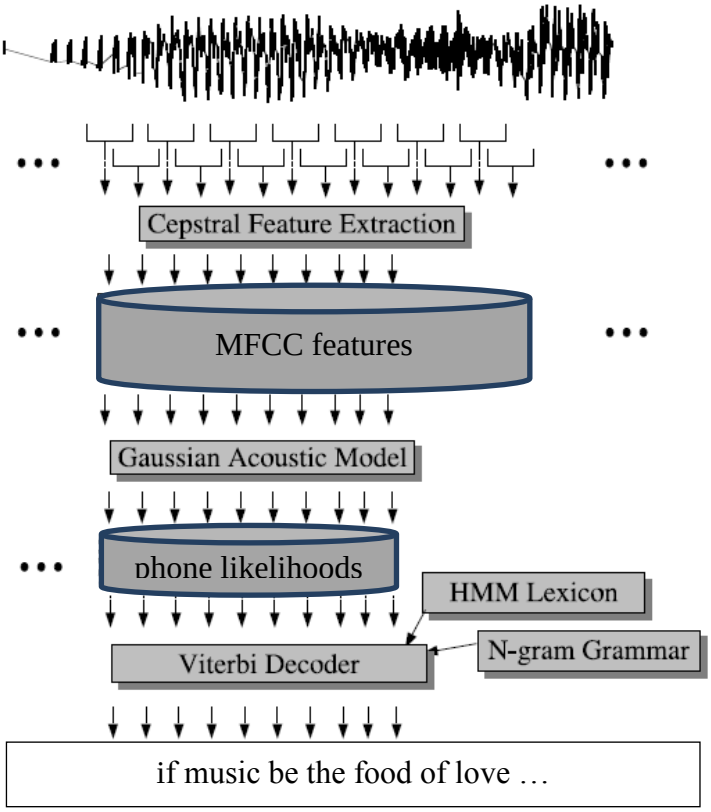
(a) Briefly define what is meant by the *semantics* of a natural language utterance, and how this differs from the *pragmatics*

(5/25)

(b) What is automatic speech recognition (ASR)? What are the factors affect the accuracy of word error rate?

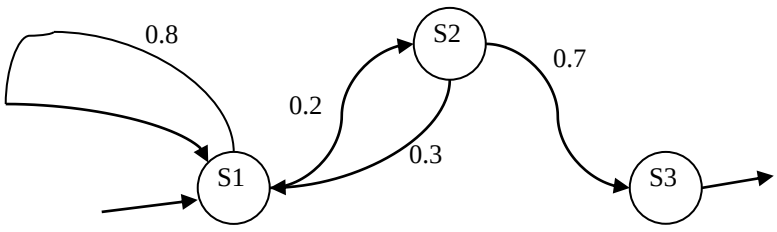
(5/25)

(c) The figure represents an HMM speech recogniser processing a single utterance "if music be the food of love". Explain briefly the process stages involved in this recognition.



(9/25)

(d) Write down one path that could be taken through the following Hidden Markov model that produces the output "C1 C2 C3 C4 C5" and the probability of this path being taken.

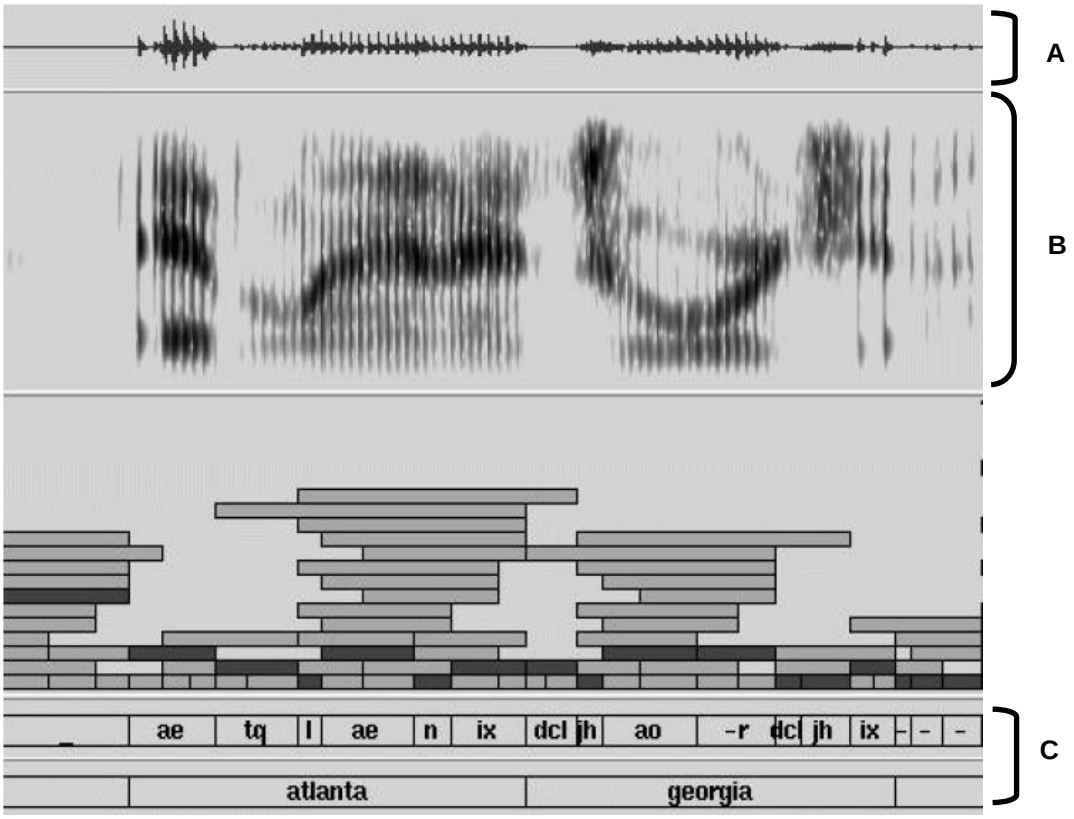


State S1: Output	Probability	State S2: Output	Probability	State S3: Output	Probability
C1	0.5	C2	0.8	C4	0.5
C2	0.3	C3	0.1	C5	0.5
C3	0.2	C4	0.1		

You don't have to calculate the actual answer as a number, as long as you show the formula that would be used to calculate it.

(6/25)

4. The following figure shows the speech representations of the words “Atlanta” and “Georgia”



Source: (“Timothy J. Hazen, I. Lee Hetherington, Han Shu, and Karen Livescu. Pronunciation modeling using a finite-state transducer representation. In *Proceedings of the ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation*. , Sep. 14-15, 2002, Estes Park, Colorado, pp. 99-10”)

(a) What kinds of speech representation do we have in A and B? Explain your answer.

(11.5/25)

(b) C represents the phonetic transcriptions of the two words Atlanta and Georgia.

(i) Why IPA symbols are not included in these transcriptions?

(2.5/25)

(ii) What kind of phonetic transcription is it?

(2.5/25)

(c) What acoustic feature that causes shade (bands) in speech representation of B?

(1/25)

(d) (i) What is the name of the organized horizontal shades (bands) in B? Explain your answer.

(2.5/25)

(ii) How these organized horizontal shades (bands) can be used to disambiguate vowels?

(5/25)

1. Perkataan berikut adalah beberapa perkataan kompleks bahasa Malaysia.

   menulis, menuliskan, tulisan, penulisan, bertulis
   bersembunyi, menyembunyikan, penyembunyian, persembunyian, sembunyian
   membaca, membacakan, terbaca, keterbacaan, pembaca

   (a) Segmenkan kumpulan perkataan-perkataan tersebut untuk mendapatkan bentuk akar atau dasarnya.

   (7.5/25)

   (b) Takrifkan istilah morfem, imbuhan, dan morfologi terbitan serta berikan contoh-contohnya.

   (4.5/25)

   (c) Apakah Transdusor Keadaan Finit (FST), dan apakah kegunaannya dalam pemprosesan bahasa tabii? Bagaimanakah ia berbeza dengan automata keadaan finit?

   (5/25)

   (d) Lukis FST yang boleh menghasilkan perkataan dengan adverb nalar dalam bahasa Inggeris. Tunjukkan bahawa FST anda mengendalikan kes-kes seperti *bright → brightly, simple → simply, silly → sillily and terrific → terrifically* dengan betul.

   (8/25)

2. Pertimbangkan tatabahasa bebas konteks berikut:

   S -> NP VP
   NP -> Det N
   VP -> V
   VP -> V NP
   N -> dog
   N ->cat
   N -> mouse
   Det -> the
   V -> sees
   V -> hates
   V -> sneezes

(a) Yang manakah daripada ayat-ayat berikut yang diterima oleh tatabahasa ini, dan kenapa?

(i) *the dog sneezes the cat* .

(ii) *the mouse hates*.

(iii) *the cat the mouse hates*.

(iv) *the mouse hates the mouse*.

(4/25)

(b) Ubahsuaikan tatabahasa itu supaya ayat yang berikut boleh diterima sebagai tambahan:

the dog the cat the mouse sees hates sneezes

Terangkan pilihan anda.

(6/25)

(c) Semantik bagi ungkapan bahasa tabii dapat dinyatakan dalam perintah pertama logik predikat (FOPL). Sebagai contoh, “the dog sneezes” boleh dinyatakan secara tepat sebagai

$\exists x \, dog(x) \cap sneeze(x)$

Mengikuti pola ini, nyatakan semantik bagi ayat dalam bahagian (b) dalam FOPL.

(1/25)

(d) Bandingkan pembinaan ini dengan yang ada di bahagian (b) dari segi semantik dan sintaks. Bagaimana anda boleh mengubahsuai tatabahasa asal itu pada bahagian (a) untuk diambil kira dalam pembinaan ini?

(14/25)

3. Pemprosesan pertuturan ialah kajian isyarat pertuturan dan kaedah-kaedah pemprosesan isyarat-isyarat ini.

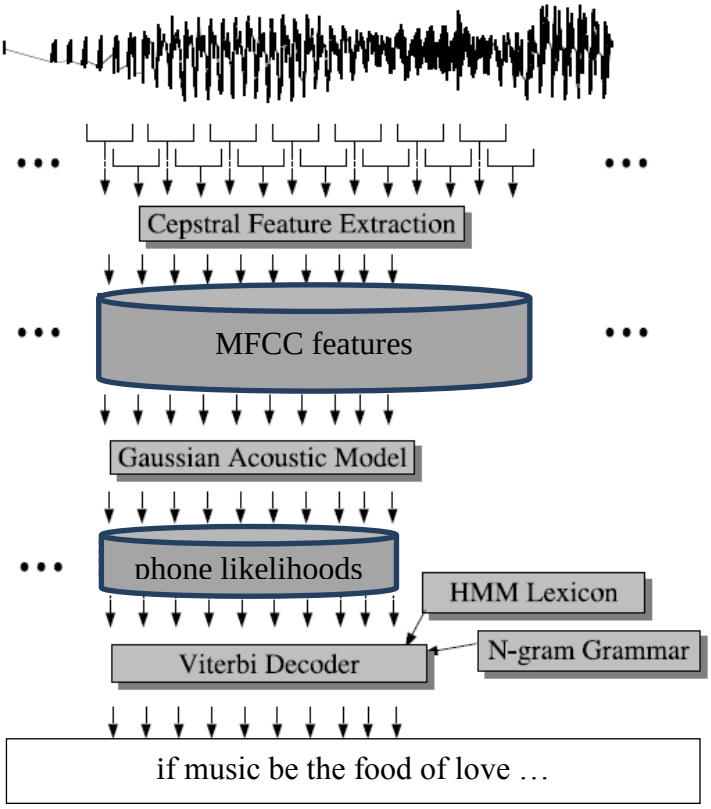
(a) Secara ringkas definisikan apa yang dimaksudkan dengan *semantik* ucapan bahasa tabii, dan bagaimana ia berbeza daripada *pragmatik*.

(5/25)

(b) Apakah pengecaman pertuturan secara automatik (ASR)? Apakah faktor-faktor yang memberi kesan kepada ketepatan kadar ralat perkataan?

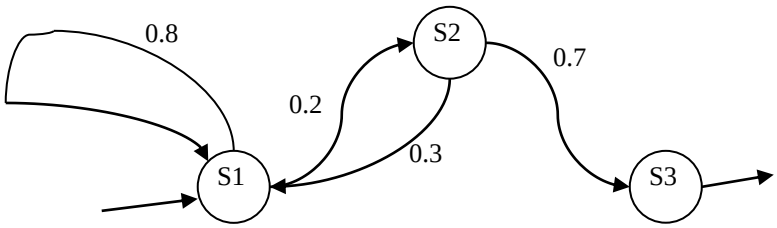
(5/25)

(c) Rajah di bawah menunjukkan satu pengecaman pertuturan HMM memproses ungkapan tunggal “if music be the food of love”. Terangkan secara ringkas peringkat proses yang terlibat dalam pengecaman ini.



(9/25)

(d) Tulis satu laluan yang boleh diambil melalui model Hidden Markov yang menghasilkan output “C1 C2 C3 C4 C5” dan kebarangkalian bagi laluan ini diambil.

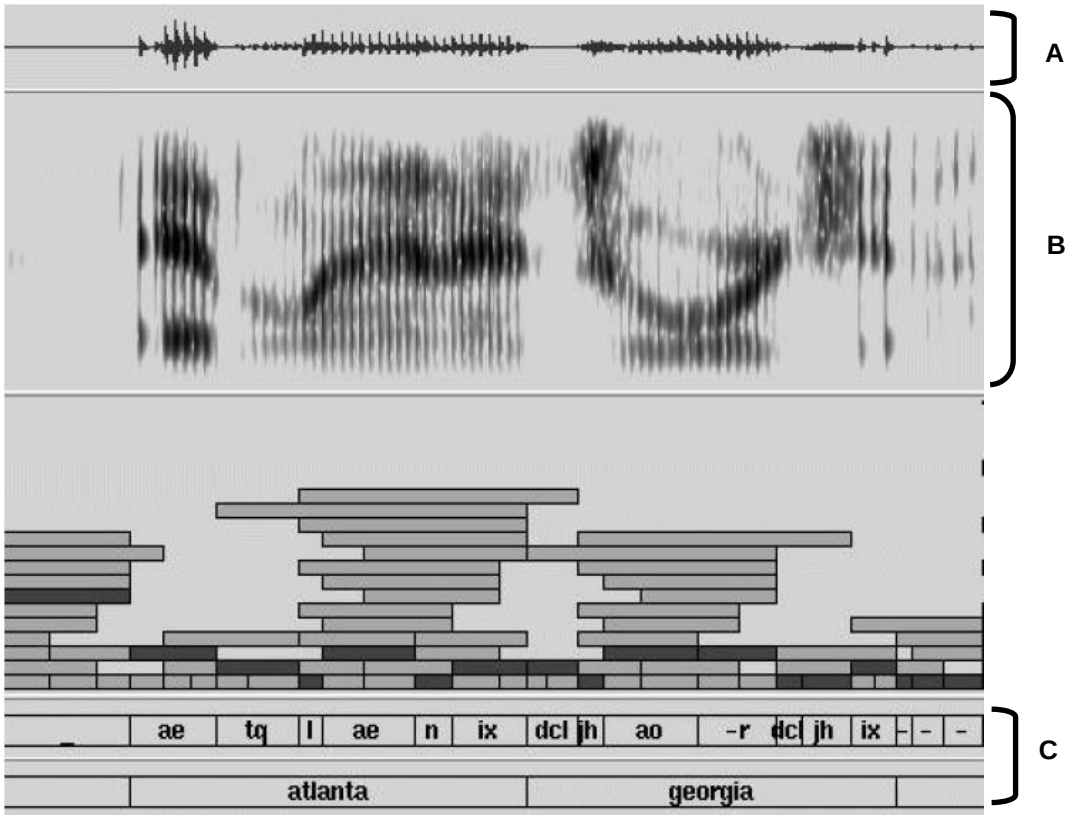


State S1: Output	Probability	State S2: Output	Probability	State S3: Output	Probability
C1	0.5	C2	0.8	C4	0.5
C2	0.3	C3	0.1	C5	0.5
C3	0.2	C4	0.1		

Anda tidak perlu mengira jawaban sebenar sebagai satu nombor, asalkan anda menunjukkan formula yang digunakan untuk mengira kebarangkalian itu.

(6/25)

4. Gambar rajah berikut menunjukkan perwakilan pertuturan bagi perkataan “Atlanta” dan “Georgia”.



Sumber: (“Timothy J. Hazen, I. Lee Hetherington, Han Shu, and Karen Livescu. Pronunciation modeling using a finite-state transducer representation. In *Proceedings of the ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation*. , Sep. 14-15, 2002, Estes Park, Colorado, ms. 99-10”)

(a) Apakah jenis perwakilan yang diwakili oleh A dan B? Terangkan jawapan anda.

(11.5/25)

(b) C mewakili transkripsi fonetik bagi perkataan Atlanta dan Georgia.

(i) Mengapa simbol IPA tidak dimasukkan dalam transkripsi ini?

(2.55/25)

(ii) Apakah jenis transkripsi fonetiknya?

(2.5/25)

(c) Apakah ciri akustik yang menyebabkan bayang (bands) dalam perwakilan pertuturan B?

(1/25)

(i) Apakah nama bayang (bands) melintang tersusun dalam B? Terangkan jawapan anda.

(2.5/25)

(ii) Bagaimana bayang (bands) melintang tersusun ini boleh digunakan untuk menyahtaksa vokal?

(5/25)

- oooOooo -