UNIVERSITI SAINS MALAYSIA

First Semester Examination
2016/2017 Academic Session

December 2016 / January 2017

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

  *[Sila pastikan bahawa kertas peperiksaan ini mengandungi **EMPAT** soalan di dalam **LAPAN** 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. (a) Linguists have been using formants for analyzing vowels for decades.

Pakar linguistik telah menggunakan forman untuk menganalisis vokal selama beberapa dekad.

(i) What is a formant?

Apakah itu forman?

(2/100)

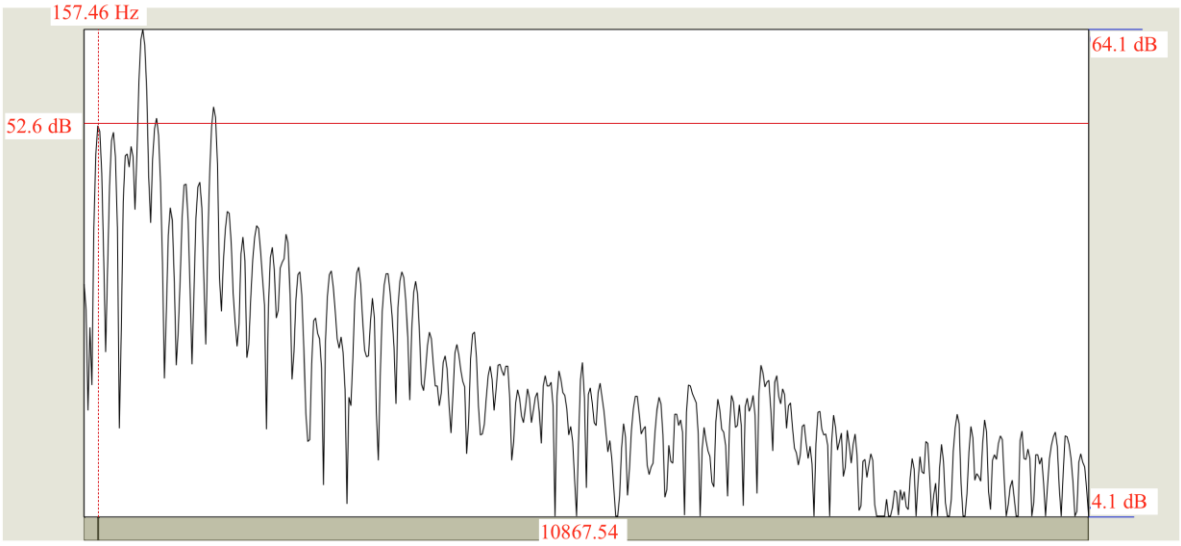
(ii) In general, what happen to the formants when we change the vowel that is produced from [i] to [u]?

Secara ringkas, apakah yang akan berlaku terhadap forman apabila kita mengubah vocal yang dihasilkan daripada [i] kepada [u]?

(2/100)

(iii) Figure below shows a spectral slice for the phone [a] from Praat software. What is the value of F1 and F2?

Rajah di bawah menunjukkan hirisan spektrum untuk fon [a] daripada perisian Praat. Apakah nilai F1 dan F2?



(4/100)

(iv) Write down the phonemic transcription (pronunciation) using IPA symbols for the following Malay words.

Tuliskan transkripsi fonemik (sebutan) dengan menggunakan simbol IPA bagi perkataan Melayu berikut.

- menyandang
- Sheila Majid

(5/100)

(v) One way to do text to speech synthesis is by using formants for instance in Klatt speech synthesis. Specify one advantage and one limitation of formant-based speech synthesis system in synthesizing a vowel.

Salah satu cara untuk melakukan sintesis pertuturan ialah menggunakan forman misalnya dalam Klatt sintesis pertuturan. Nyatakan satu kelebihan dan satu kelemahan sistem sintesis pertuturan berdasarkan forman untuk mensintesiskan vokal.

(4/100)

(b) Annotation/tagging is important in natural language processing.

Catatan/penandaan ialah adalah penting dalam pemprosesan Bahasa tabii.

(i) Explain what is done during annotation/tagging. Give an example of annotation.

Terangkan apa yang dilakukan semasa catatan/penandaan. Berikan satu contoh catatan.

(4/100)

(ii) Explain how you would annotate a list of Malay word with corresponding affixation information using XML. Use an example in your explanation.

Terangkan bagaimana anda akan melakukan penandaan terhadap suatu senarai perkataan Melayu dengan maklumat imbuhan menggunakan XML. Gunakan suatu contoh dalam penerangan anda.

(4/100)

2. (a) Ahmad tries to extract the text from an image file using an optical character recognition (OCR) algorithm that perform character recognition. Due to the poor quality of the image, the text file obtained contains many words with spelling errors (e.g. “suatu” recognized as “soualu”).

Ahmad cuba untuk mengekstrak teks dari fail imej menggunakan algoritma pengecaman aksara optik (OCR) yang melaksanakan pengecaman aksara. Oleh kerana kualiti imej yang rendah, fail teks yang diperolehi mengandungi banyak perkataan dengan kesilapan ejaan (contohnya "suatu" dicam sebagai "soualu").

(i) Calculates the distance between the word “suatu” and “soualu” using minimum edit distance algorithm. Assume insertion, deletion and substitution cost are 1, 1 and 2 respectively.

Kira jarak di antara perkataan "suatu" dan "soualu" dengan menggunakan algoritma jarak edit minimum. Anggap kos sisipan, penghapusan dan penggantian ialah 1, 1 dan 2 masing-masing.

(5/100)

(ii) Assuming you have a list of vocabulary, explain how you can use the vocabulary list and minimum edit distance algorithm to correct words with spelling errors.

Andaikan anda mempunyai suatu senarai perbendaharaan kata, terangkan bagaimana anda menggunakan senarai perbendaharaan kata tersebut dan algoritma jarak edit minimum untuk memperbetulkan perkataan dengan kesilapan ejaan.

(4/100)

(iii) Explain how a language model can be used with minimum edit distance algorithm to improve the accuracy of spelling correction.

Jelaskan bagaimana model bahasa boleh digunakan dengan algoritma jarak edit minimum untuk meningkatkan ketepatan pembetulan ejaan.

(5/100)

(b) One way to perform part of speech tagging is using hidden Markov model (HMM).

Salah satu cara untuk melakukan penandaan kelas kata ialah menggunakan model Markov tersembunyi (HMM).

(i) Given the transition probabilities and observation likelihood (in log) below, find out the part of speech for the words in the sentence "time flies" using Viterbi algorithm.

Transition probabilities (in log). The first column shows the source state while the first row shows the target state. For example P(Noun|<s>) = -0.4.

Diberi kebarangkalian peralihan dan kemungkinan pemerhatian (dalam log) di bawah, dapatkan kelas kata bagi perkataan dalam ayat "time flies" menggunakan algoritma Viterbi.

Kebarangkalian peralihan (dalam log). Lajur pertama menunjukkan keadaan sumber manakala baris pertama menunjukkan keadaan sasaran. Contohnya (Kata Nama| <s>) = -0.4.

	Verb *Kata Kerja*	Noun *Kata Nama*	Adjective *Adjektif*
<s>	-0.7	-0.4	-0.3
Verb *Kata Kerja*	-2	-0.3	-0.7
Noun *Kata Nama*	-0.4	-1.5	-1.8
Adjective *Adjektif*	-1.7	-0.4	-0.5

	time	flies
Verb *Kata Kerja*	-1.6	-1.2
Noun *Kata Nama*	-1.4	-0.7
Adjective *Adjektif*	-1.8	-2.3

(6/100)

(ii) Another way to perform part of speech tagging is through transformation based learning (TBL). Describe the general steps of a transformation based learning (TBL) approach in modeling part of speech.

Cara lain untuk melakukan penandaan kelas kata adalah melalui pembelajaran berasaskan transformasi (TBL). Jelaskan langkah-langkah umum pendekatan pembelajaran berasaskan transformasi (TBL) dalam pemodelan kelas kata.

(5/100)

3. (a) A phrase-structure grammar is a set of Phrase-structure rules (PS rules) that can decompose sentences and phrases down to the words and describe complete trees. Based on Table 1. as follow, compose two sentences and draw the parse tree for each sentences.

Suatu tatabahasa struktur-frasa adalah satu set peraturan struktur-frasa (peraturan PS) yang boleh mengurai ayat dan frasa turun ke kata-kata dan menggambarkan rajah pohon yang lengkap. Berdasarkan Jadual 1. seperti berikut, bina dua ayat dan lukiskan rajah pohon huraian bagi setiap ayat.

Phrases	Lexicon
S → NP VP NP → Determiner Noun NP → NP PP VP → Verb NP VP → Verb NP PP PP → Preposition NP	Determiner → the Noun → teacher Noun → book Noun → table Noun → class Verb → brought Preposition → to Preposition → of

Table 1. A phrase structure grammar
Jadual 1. Tatabahasa struktur frasa

(8/100)

(b) Verbs can be followed by adjective phrases or noun phrases. Imagine a new constituent category, AdjP, describing adjective phrases. Write the corresponding rules. Then write rules accepting the following sentences:

Kata Kerja boleh diikuti oleh frasa adjektif atau frasa kata nama. Bayangkan satu kategori konstituen yang baru, AdjP, menggambarkan Frasa adjektif. Tulis peraturan-peraturan yang sepadan. Kemudian tulis peraturan yang menerima ayat-ayat yang berikut:

the teacher is tall,
the teacher is very tall,
Bill is a teacher.

(8/100)

(c) We can use DCG rules for tokenizing text. Describe in detail how two part of tokenizer are used to implement DCG grammar. Give the examples using Prolog.

Kita boleh menggunakan peraturan-peraturan DCG untuk mentokenkan teks. Terangkan secara terperinci bagaimana dua bahagian daripada tokenizer yang digunakan bagi melaksanakan tatabahasa DCG. Berikan contoh-contoh menggunakan Prolog.

(8/100)

(d) A Probability Context-Free Grammar (PCFG) is a constituent context-free grammar where each rule describing the structure of a left-hand-side symbol is augmented with its probability P(*lhs* → *rhs*). From the following table (Table 2. and 3.), give the probability of parse trees, T1 and T2.

Tatabahasa Konteks Bebas Kebarangkalian (PCFG) adalah tatabahasa konteks bebas konstituen di mana setiap peraturan yang menerangkan struktur simbol sebelah-kiri diperkukuhkan dengan kebarangkalian P(lhs → rhs). Dari jadual yang berikut (Jadual 3.2 dan 3.3), berikan kebarangkalian pohon huraian,T1 dan T2.

Rules	P	Rules	P
s --> np vp	0.8	det --> the	1.0
s --> vp	0.2	noun --> waiter	0.4
np --> det noun	0.3	noun --> meal	0.3
np --> det adj noun	0.2	noun --> day	0.3
np --> pronoun	0.3	verb --> bring	0.4
np --> np pp	0.2	verb --> slept	0.2
vp --> v np	0.6	verb --> brought	0.4
vp --> v np pp	0.1	pronoun --> he	1.0
vp --> v pp	0.2	prep --> of	0.6
vp --> v	0.1	prep --> to	0.4
pp --> prep np	1.0	pronoun --> he	1.0
		adj --> big	1.0

Table 2. A small set of phrase-structure rules augmented with probabilities, P
Jadual 2. Suatu set peraturan struktur-frasa yang diperkukuhkan dengan kebarangkalian, P

Parse trees
T1: vp(verb(bring), np(np(det(the), noun(meal)), pp(prep(of), np(det(the), noun(day))))) T2: vp(verb(bring), np(np(det(the), noun(meal))), pp(prep(of), np(det(the), noun(day))))

Table 3. Possible parse trees for **Bring the meal of the day**
Jadual 3. Rajah pohon huraian yang mungkin bagi ayat **Bring the meal of the day**

(6/100)

4. (a) Intransitive verb is a verb that does not need a direct object to complete its meaning**.** With intransitive verbs, the logical conjunctions link the subject to the verb. Show the logical presentations of the following sentences:

Kata kerja intransitif ialah kata kerja yang tidak memerlukan objek langsung untuk melengkapkan maknanya. Dengan kata kerja tidak transitif, penghubungan logik menghubungkan subjek kepada kata kerja. Tunjukkan perwakilan logik ayat yang berikut:

A waiter ran
Every waiter ran
The waiter ran

(6/100)

(b) When a sentence contains a transitive verb, we must take the object into account to complete its meaning. Shows the logical presentations of the following sentences:

Apabila suatu ayat mengandungi kata kerja transitif, kita harus mengambil kira tentang objek untuk melengkapkan maknanya. Tunjukkan perwakilan logik ayat yang berikut:

A waiter brought a meal
Every waiter brought a meal
The waiter brought a meal

(6/100)

(c) Give a detailed explanation of similarities and differences among the following set of lexemes: *imitation*, *synthetic*, *artificial*, *fake*, and *simulated*.

Berikan penjelasan yang terperinci persamaan dan perbezaan antara set leksem yang berikut: imitation, *synthetic*, *artificial*, *fake*, dan *simulated*.

(4/100)

(d) Between the words ***eat*** and ***find*** which would you expect to be more effective in selectional restriction-based sense disambiguation? Why?

Antara perkataan ***eat*** *dan* ***find*** *manakah yang anda jangka menjadi lebih berkesan dalam penyahtaksaan makna berdasarkan-sekatan secara pilihan? Kenapa?*

(4/100)

- oooOooo -