UNIVERSITI SAINS MALAYSIA

First Semester Examination
2016/2017 Academic Session

December 2016 / January 2017

CCS512 – Language Engineering
[Kejuruteraan Bahasa]

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) There are many techniques used in Language Engineering. Discuss **two** (**2**) of the techniques that are related to speech processing.

*Terdapat banyak teknik yang digunakan dalam Kejuruteraan Bahasa. Bincangkan **dua** (**2**) daripada teknik-teknik berkenaan yang berkaitan dengan pemprosesan pertuturan.*

(5/100)

(b)

ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[p]	[p]	parsley	[p aa r s l iy]
[t]	[t]	tea	[t iy]
[k]	[k]	cook	[k uh k]
[b]	[b]	bay	[b ey]
[d]	[d]	dill	[d ih l]
[g]	[g]	garlic	[g aa r l ix k]
[m]	[m]	mint	[m ih n t]
[n]	[n]	nutmeg	[n ah t m eh g]
[ng]	[ŋ]	baking	[b ey k ix ng]
[f]	[f]	flour	[f l aw axr]
[v]	[v]	clove	[k l ow v]
[th]	[θ]	thick	[th ih k]
[dh]	[ð]	those	[dh ow z]
[s]	[s]	soup	[s uw p]
[z]	[z]	eggs	[eh g z]
[sh]	[ʃ]	squash	[s k w aa sh]
[zh]	[ʒ]	ambrosia	[ae m b r ow zh ax]
[ch]	[tʃ]	cherry	[ch eh r iy]
[jh]	[dʒ]	jar	[jh aa r]
[l]	[l]	licorice	[l ih k axr ix sh]
[w]	[w]	kiwi	[k iy w iy]
[r]	[r]	rice	[r ay s]
[y]	[j]	yellow	[y eh l ow]
[h]	[h]	honey	[h ah n iy]
Less commonly used phones and allophones			
[q]	[ʔ]	uh-oh	[q ah q ow]
[dx]	[ɾ]	butter	[b ah dx axr]
[nx]	[ɾ̃]	winner	[w ih nx axr]
[el]	[l̩]	table	[t ey b el]

Figure 1a/*Rajah 1a*: ARPAbet symbol with IPA equivalents of English consonants

ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[iy]	[i]	lily	[l ih l iy]
[ih]	[ɪ]	lily	[l ih l iy]
[ey]	[eɪ]	daisy	[d ey z iy]
[eh]	[ɛ]	pen	[p eh n]
[ae]	[æ]	aster	[ae s t axr]
[aa]	[ɑ]	poppy	[p aa p iy]
[ao]	[ɔ]	orchid	[ao r k ix d]
[uh]	[ʊ]	wood	[w uh d]
[ow]	[oʊ]	lotus	[l ow dx ax s]
[uw]	[u]	tulip	[t uw l ix p]
[ah]	[ʌ]	buttercup	[b ah dx axr k ah p]
[er]	[ɝ]	bird	[b er d]
[ay]	[aɪ]	iris	[ay r ix s]
[aw]	[aʊ]	sunflower	[s ah n f l aw axr]
[oy]	[ɔɪ]	soil	[s oy l]
Reduced and uncommon phones			
[ax]	[ə]	lotus	[l ow dx ax s]
[axr]	[ɚ]	heather	[h eh dh axr]
[ix]	[ɨ]	tulip	[t uw l ix p]
[ux]	[ʉ]	dude[1]	[d ux d]

Figure 1b/*Rajah 1b*: ARPAbet symbol with IPA equivalents of English vowels

Refer to the above figures (Figure 1a and 1b). The following list of colour words are transcription words in IPA. Translate the pronunciations of the words from the IPA into the ARPAbet.

Rujuk kepada jadual di atas (Rajah 1a dan 1b). Senarai perkataan-perkataan warna berikut ialah perkataan-perkataan transkripsi dalam IPA. Terjemahkan sebutan perkataan-perkataan itu daripada IPA kepada bentuk ARPAbet.

a. [waɪt] b. [rɛd] c. [blæk[
d. [griːn] e. [braʊn] f. [ˈɒrɪnʤ]
g. [ˈpɜːpl] h. [bluː] i. [beɪʒˈ]
j. [æʒə]

(10/100)

(c) Assume you have a user uttering the following utterance: “Could you please take the red box and put it on the other end of the table?” But your speech recognition hypothesis turns out to be: “Could you place vague red box and put it this on another end of that?” Calculate the Word Error Rate (WER) between the ASR hypothesis and the actual utterance. Detail your calculations.

Andaikan anda mempunyai pengguna mengucapkan lafaz yang berikut: "Could you please take the red box and put it on the other end of the table?" *Tetapi hipotesis pengecaman pertuturan anda ternyata menjadi: “Could you place vague red box and put it this on another end of that?” Kira Kadar Ralat Perkataan (WER) antara hipotesis ASR dan lafaz yang sebenar. Huraikan pengiraan anda.*

(10/30)

(d)

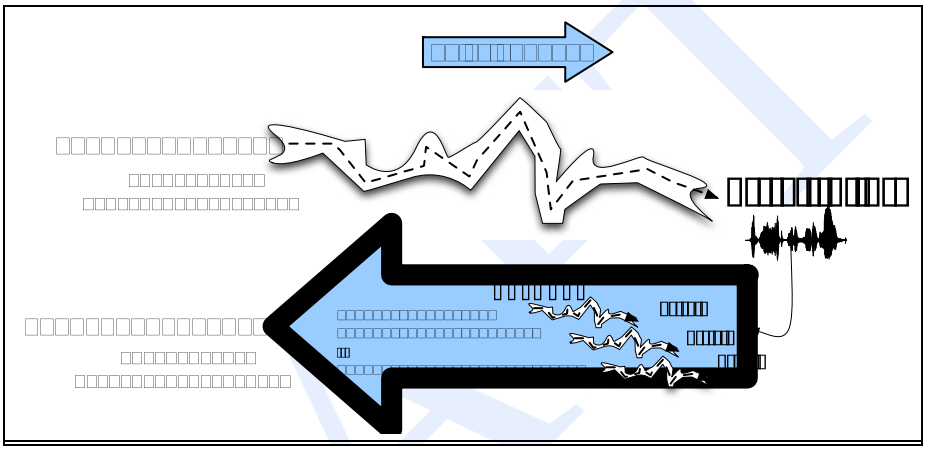

Figure 2: The noisy channel model

Implementing the noisy-channel model as illustrated in Figure 2 requires solutions to **two problems**. What are the **two** (**2**) problems and their solutions? Briefly explain your answer.

*Melaksanakan model saluran-hingar seperti yang ditunjukkan dalam Rajah 2 memerlukan penyelesaian kepada **dua masalah**. Apakah **dua** (**2**) masalah itu dan apakah penyelesaiannya? Huraikan dengan ringkas jawapan anda.*

(5/100)

2. (a) Treebanks are corpora in which each sentence has been paired with a parse tree. Finding heads in treebank trees is a task that arises frequently in many applications. Visualize this task by annotating the nodes of a parse tree with the heads of each corresponding node for the following noun phrase (NP):

Bankpepohon adalah korpora yang mana setiap ayat telah dipasangkan dengan satu pepohon huraian. Mencari kepala dalam pepohon-pepohon bankpepohon merupakan satu tugas yang sering timbul dalam banyak aplikasi. Visualisasikan tugas ini dengan menganotasi nod pepohon huraian dengan kepala-kepala setiap nod yang sepadan bagi frasa kata nama (NP) berikut:

All the morning flights from Denver to Tampa leaving before 10.

(10/100)

(b) Discuss how you would augment a parser to deal with input that may be incorrect, such as spelling errors or misrecognitions from a speech recognition system.

Bincangkan bagaimana anda akan mengukuhkan penghurai untuk berurusan dengan input yang mungkin tidak betul, seperti kesilapan ejaan atau salah pengecaman daripada sistem pengecaman pertuturan.

(5/100)

(c) The task of selecting the correct sense for a word is called **word sense disambiguation** (**WSD)**. How does WSD improve machine translation tasks?

*Fungsi memilih maksud yang betul bagi sesuatu perkataan dikenali sebagai **penyahtaksaan perkataan (WSD)**. Bagaimanakah WSD menambah baik tugas-tugas terjemahan mesin?*

(5/100)

3. N-gram is a method of word prediction with probabilistic model, which predicts the next word from the previous N-1 words.

N-gram ialah suatu kaedah ramalan perkataan menggunakan model kebarangkalian, yang dapat meramalkan perkataan seterusnya daripada perkataan N-1 sebelumnya.

(a) Describe how N-gram can be used in **two** (**2**) natural language related applications. For each of the application, give example to support your description.

*Terangkan bagaimana N-gram boleh digunakan dalam **dua** (**2**) aplikasi yang berkaitan dengan bahasa tabii. Bagi setiap aplikasi, beri contoh untuk menyokong huraian anda.*

(6/100)

(b) Describe the concept of bigram model. Assume an element in the model is a word, explain how bigram model is used to estimate the probability of “the” in following sentence.

Terangkan konsep model bigram. Andaikan setiap unsur dalam model ialah perkataan, jelaskan bagaimana model bigram boleh digunakan untuk menganggarkan kebarangkalian "the" dalam ayat yang berikut.

Walden’s Pond’s water is so transparent that the …

(4/100)

(c) Tables (refer Table 1 and Table 2) below shows the unigram and bigram counts for eight words from a sample collection of the Restaurant Project. This collection consists of 9332 sentences.

Jadual-jadual (rujuk Jadual 1 dan Jadual 2) di bawah menunjukkan kiraan unigram dan bigram untuk lapan perkataan dari contoh koleksi Restaurant Project. Koleksi ini mengandungi 9332 ayat.

Table 1/*Jadual 1*: Bigram counts for eight words from Restaurant Project corpus.

Bigram counts

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Table 2/*Jadual 2*: Bigram counts for eight words from Restaurant Project corpus.

Unigram counts

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

Calculate the bigram probability for each bigram. Show your result in a table form (See sample table below).

Kirakan kebarangkalian bigram bagi setiap bigram. Tunjukkan keputusan anda dalam bentuk jadual (Lihat contoh jadual di bawah).

	i	want	to	eat	chinese	food	lunch	spend
i								
want								
to								
eat								
chinese								
food								
lunch								
spend								

(6/100)

(d) Referring to the bigram probability of question (c) and the following probabilities,

Merujuk kepada kebarangkalian bigram bagi soalan (c) dan kebarangkalian berikut,

P(i | <s>) = 0.25
P(english | want) = 0.0011
P(food | english) = 0.5
P(</s> | food) = 0.68

Calculate the probability of sentence, “I want english food.”.

Kira kebarangkalian untuk ayat, “I want english food.”.

(4/100)

4. Given the following four documents,

Berdasarkan empat dokumen berikut,

D1 (document length = 0.68): “The first car did not have a steering wheel. Drivers steered the car with a lever.”
D2 (document length = 1.38): “The automobile is the most recycled consumer product in the world today.”
D3 (document length = ?): “Car pollution is becoming an increasing problem today.”
D4 (document length = 0.76): “Pollution is a global problem causing climate changes. The car is one of the contributors to the pollution.”

(a) Explain the motivation for the tf.idf formula and each of its components (tf and idf).

Jelaskan motivasi untuk formula tf.idf dan setiap komponennya (tf dan idf).

(4/100)

(b) Tokenize the document and remove the stop words only (without stemming etc.). Then, calculate the normalized length of document 3 based on tf.idf formula.

Bahagikan dokumen kepada token dan keluarkan perkataan “stop words” sahaja (tanpa “stemming” dsb.). Kemudian, kirakan panjang dokumen 3 yang dinormalkan dengan menggunakan formula td.idf.

Stop word list: {a, an, did, first, have, in, is, most, not, of, one, the, to, with}

(6/100)

(c) Calculate the tf.idf score for the terms “car”, “pollution”, and “wheel” for all four documents.

Kirakan skor tf.idf untuk istilah "car", "pollution", dan "wheel" untuk keempat-empat dokumen.

(8/100)

(d) Calculate the cosine similarity between each of the documents and the query “car pollution wheel” (NOT a phrase query). Rank the documents by most similar to less similar with respect to the query.

Kira persamaan kosinus antara setiap dokumen dan pertanyaan "car pollution wheel" (BUKAN pertanyaan bentuk frasa). Susun dokumen-dokumen daripada yang paling sama kepada yang kurang sama dengan pertanyaan berkenaan.

(12/100)