UNIVERSITI SAINS MALAYSIA

First Semester Examination
2015/2016 Academic Session

December 2015/January 2016

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

  *[Sila pastikan bahawa kertas peperiksaan ini mengandungi **EMPAT** 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. (a) The digit "**1995**" can be written as either "*seribu sembilan ratus sembilan puluh lima*" or "*sembilan belas sembilan puluh lima*" in a Malay text.

(i) Explain in what situation the number **1995** is written in word form "*seribu sembilan ratus sembilan puluh lima*" and "*sembilan belas sembilan puluh lima*".

(2/100)

(ii) Describe a method which can be used to identify and tag any number (assume 4 digits) in a Malay text. Example, identify ***1995*** in the text and then tag it as such **[1995]**.

(4/100)

(iii) Explain the steps to convert any number (assume 4 digits) to the correct word form when given a sentence. e.g. **1995** is converted to "*seribu sembilan ratus sembilan puluh lima*" if the sentence given is "***nombor kenderaan saya ialah 1995***".

(5/100)

(b) Building an n-gram language model requires large amount of text.

(i) Text obtained have to be first normalized before the relevant statistics can be obtained. Explain the steps normally performed during normalization.

(4/100)

(ii) Given the text below, calculate unigram probability for the word "Malaysia", P(Malaysia) and all bigram probabilities given the word Malaysia, e.g. P(is|Malaysia).

> Malaysia is a federal constitutional monarchy located in Southeast Asia. It consists of thirteen states and three federal territories and has a separated by the South China Sea into two similarly sized. regions, Peninsular Malaysia and East Malaysia (Malaysian Borneo).

(4/100)

(iii) What is the purpose of smoothing in language modelling? Explain Laplace smoothing.

(6/100)

2. Part of speech tagging is one of the processes often carried out in natural language processing.

(a) What is the purpose of part of speech tagging?

(2/100)

(b) Explain one situation where part of speech tagging is used.

(2/100)

(c) What are the resources required to build a hidden Markov model (HMM) part of speech tagger?

(4/100)

(d) In HMM part of speech tagging, we want to find the most probable sequence of tag T, given a sentence, W, where T= t_1, t_2, …, t_n and W=w_1, w_2, …, w_n. Two probabilities required to estimate argmax P(T|W) are transition probabilities and observation likelihood. Describe the corpora required and steps to estimate transition probabilities and observation likelihood?

(6/100)

(e) Use the transition probabilities and observation likelihoods given below to find out the most probable part of speech tags for each word, given the sentence "banking on you". What is the name of the algorithm?

	VB	NN	IN	PRP
<s>	-0.7	-0.4	-1	-0.5
VB	-2	-0.3	-0.7	-0.7
NN	-0.4	-1	-1.4	-1
IN	-0.7	-0.4	-2	-0.5
PRP	-0.3	-1	-0.5	-2

Transition probabilities (in log). First column shows the source state while the first row shows the target state. e.g. P(NN|<s>) = -0.4. <s> is the tag for the starting of a sentence.

	banking	on	you
VB	-1.6	-1.3	-Infinite
NN	-1.4	-Infinite	-Infinite
IN	-Infinite	-0.3	-Infinite
PRP	-Infinite	-Infinite	-0.6

Observation likelihood (in log). E.g. P(VB|banking) = -1.6.

(11/100)

3. (a) (i) Definite Clause Grammar (DCG) notation enables us to transcribe a set of phrase-structure rules directly into a Prolog program. Translate the following sentences into Malay and write the DCG rules that accepting them.

The waiter brought the meal.
The waiter brought the meal to the table.
The waiter brought the meal of the day.

(6/100)

(ii) Write your own DCG rules that would accept the following sentence:

The nice hedgehog ate the worm in its nest.

Then draw its corresponding tree. Do the same in Malay.

(6/100)

(b) Partial or shallow parsing is a technique used to extract incomplete syntactic representations.

(i) Identify the noun phrase and write down the head of noun phrase to recognize the noun groups of the following text:

The big tobacco firms are fighting back in the way that served them well for 40 victorious years, pouring their wealth into potent, relentless legal teams. But they are also starting to talk of striking deals – anathema for those 40 years, and a sure sign that, this time, victory is less certain.
(*The Economist*, no. 8004, 1997).

(5/100)

(ii) To implement Earley algorithm in Prolog, we must first represent the grammar rules in chart form. The chart consists of rules such as

np --> np•pp [0, 2]

which we represent as facts
arc(np, [np, '.', pp], 0, 2).

The start symbol is encoded as:
arc(start, ['.', np], 0, 0).

From the above example, create a list of your own grammar rules and draw the arcs of this sentence " the meal of the day" with the Earley algorithm.

(8/100)

4. (a) First-Order Logic (FOL) is a flexible approach for a meaning representation language. Give FOL rule for the following sentences:

(i) All vegetarian restaurants serve vegetarian food.

(ii) Mammals do not lay eggs.

(iii) Not all mammals walk on two legs.

(6/100)

(b) Consider the following definition.

patron:
1. a person who gives money or support to a person, an organization, a cause or an activity
2. a customer of a shop, restaurant, theater

order:
1. to give an order to somebody 2. to request somebody to supply or make goods, etc. 3. to request somebody to bring food, drink, etc. in a hotel, restaurant, etc. 4. to put something in order

meal:
1.an occasion where food is eaten 2. the food eaten on such occasion

(i) By refer to the above definition, annotate each word of the following sentence with their possible senses:

The patron ordered a meal

(3/100)

(ii) Implement the complete semantic net of those words in (i).

(6/100)

(iii) Disambiguate the senses of words in the sentence in (i) using word definitions.

(4/100)

(iv) List and describe briefly **three** (**3**) algorithms of automatic word sense disambiguation.

(6/100)

1. (a) Nombor "**1995**" boleh ditulis dalam bentuk perkataan sama ada sebagai "seribu sembilan ratus sembilan puluh lima" atau sebagai "sembilan belas sembilan puluh lima" dalam teks Melayu.

(i) Jelaskan dalam keadaan apa **1995** ditulis sebagai "seribu sembilan ratus sembilan puluh lima" dan "sembilan belas sembilan puluh lima".

(2/100)

(ii) Terangkan satu kaedah yang boleh digunakan untuk mengecam dan menanda nombor (anggap 4 angka) dalam teks Melayu. Contoh, mengecam **1995** dalam teks dan kemudian menanda seperti berikut **[1995]**.

(4/100)

(iii) Terangkan langkah-langkah untuk menukar apa-apa nombor (anggap 4 angka) kepada bentuk perkataan yang betul apabila diberi suatu ayat. Contohnya, nombor **1995** kepada "seribu sembilan ratus sembilan puluh lima" jika ayat yang diberikan ialah "**nombor kenderaan saya ialah 1995**".

(5/100)

(b) Pembinaan model bahasa n-gram memerlukan sejumlah teks yang besar.

(i) Teks diperolehi perlu terlebih dahulu dinormalisasikan sebelum statistik yang berkaitan boleh diperolehi. Jelaskan langkah-langkah yang biasanya dilakukan semasa normalisasi.

(4/100)

(ii) Diberi teks di bawah. Kirakan kebarangkalian unigram perkataan "*Malaysia*", P(Malaysia) dan semua kebarangkalian bigram jika diberi perkataan *Malaysia*, contohnya P(is|Malaysia).

> *Malaysia is a federal constitutional monarchy located in Southeast Asia. It consists of thirteen states and three federal territories and has a separated by the South China Sea into two similarly sized. regions, Peninsular Malaysia and East Malaysia (Malaysian Borneo).*

(4/100)

(iii) Apakah tujuan pelicinan dalam pemodelan bahasa? Terangkan pelicinan Laplace.

(6/100)

2. Penandaan golongan kata ialah salah satu proses yang sering dilaksanakan dalam pemprosesan bahasa tabii.

(a) Apakah tujuan penandaan golongan kata?

(2/100)

(b) Terangkan satu keadaan di mana penandaan golongan kata digunakan.

(2/100)

(c) Apakah sumber yang diperlukan untuk membina penanda golongan kata model Markov tersembunyi (MMT)?

(4/100)

(d) Dalam penandaan golongan kata MMT, kita mencari urutan penanda T yang paling berkemungkinan, diberi ayat, W, di mana $T = t_1\ t_2, ..., t_n$ dan $W = w_1, w_2, ..., w_n$. Dua kebarangkalian yang diperlukan untuk menganggar argmax P(T|W) ialah kebarangkalian transisi dan kebolehjadian pemerhatian. Jelaskan apa korpus yang diperlukan dan langkah-langkah untuk menganggarkan kebarangkalian transisi dan kebolehjadian pemerhatian?

(6/100)

(e) Gunakan kebarangkalian transisi dan kebolehjadian pemerhatian yang diberikan di bawah untuk mencari golongan kata yang paling berkemungkinan bagi setiap perkataan, jika diberi ayat "banking on you". Apa nama algoritma ini?

	VB	NN	IN	PRP
<s>	-0.7	-0.4	-1	-0.5
VB	-2	-0.3	-0.7	-0.7
NN	-0.4	-1	-1.4	-1
IN	-0.7	-0.4	-2	-0.5
PRP	-0.3	-1	-0.5	-2

Kebarangkalian transisi (dalam log). Ruang pertama menunjukkan keadaan sumber, baris pertama menunjukkan keadaan sasaran. Contoh, P(NN|<s>) = -0.4. <s> ialah penanda bagi permulaan ayat.

	banking	on	you
VB	-1.6	-1.3	-Infinite
NN	-1.4	-Infinite	-Infinite
IN	-Infinite	-0.3	-Infinite
PRP	-Infinite	-Infinite	-0.6

Kebolehjadian pemerhatian (dalam log). Contoh, P(VB|banking) = -1.6.

(11/100)

3. (a) (i) Notasi tatabahasa Klausa Definit (DCG) membolehkan kita membuat transkripsi satu set peraturan struktur-frasa terus ke dalam program Prolog. Terjemahkan ayat berikut ke dalam Bahasa Melayu dan tuliskan peraturan DCG yang menerimanya.

The waiter brought the meal.
The waiter brought the meal to the table.
The waiter brought the meal of the day.

(6/100)

(ii) Tuliskan peraturan DCG anda sendiri yang boleh menerima ayat yang berikut:

The nice hedgehog ate the worm in its nest.

Kemudian lukiskan rajah pohonnya. Lakukan perkara yang sama untuk Bahasa Melayu.

(6/100)

3. (b) Huraian sebahagian atau huraian cetek adalah suatu teknik untuk mendapatkan perwakilan sintaksis tidak lengkap.

(i) Kenalpasti frasa kata nama dan kepala bagi frasa nama untuk mengenali kumpulan kata nama bagi teks berikut:

The big tobacco firms are fighting back in the way that served them well for 40 victorious years, pouring their wealth into potent, relentless legal teams. But they are also starting to talk of striking deals – anathema for those 40 years, and a sure sign that, this time, victory is less certain.
(*The Economist*, no. 8004, 1997).

(5/100)

(ii) Untuk melaksanakan algoritma Earley dalam Prolog, kita perlu terlebih dahulu memberikan peraturan-peraturan tatabahasa dalam bentuk carta. Carta tersebut terdiri daripada peraturan-peraturan seperti

np --> np•pp [0, 2]

yang kita wakilkan sebagai fakta
arc(np, [np, '.', pp], 0, 2).

Simbol mula dienkod sebagai
arc(start, ['.', np], 0, 0).

Dari contoh di atas, buat satu senarai bagi tatabahasa anda sendiri dan lukiskan lengkok-lengkok bagi ayat " the meal of the day" dengan algoritma Earley.

(8/100)

4. (a) Logik Arahan-Pertama (FOL) ialah pendekatan yang fleksibel bagi bahasa perwakilan makna. Berikan satu peraturan FOL bagi ayat-ayat berikut:

(i) *All vegetarian restaurants serve vegetarian food.*

(ii) *Mammals do not lay eggs.*

(iii) *Not all mammals walk on two legs.*

(6/100)

(b) Pertimbangkan definisi bagi perkataan yang berikut.

patron:
1. a person who gives money or support to a person, an organization, a cause or an activity
2. a customer of a shop, restaurant, theater

order:
1. to give an order to somebody 2. to request somebody to supply or make goods, etc. 3. to request somebody to bring food, drink, etc. in a hotel, restaurant, etc. 4. to put something in order

meal:
1.an occasion where food is eaten 2. the food eaten on such occasion

(i) Dengan merujuk kepada definisi-definisi di atas, anotasikan setiap perkataan bagi ayat berikut dengan maknanya yang mungkin:

The patron ordered a meal

(3/100)

(ii) Laksanakan rangkaian semantik yang lengkap bagi perkataan-perkataan di dalam (i).

(6/100)

(iii) Nyahtaksakan makna-makna perkataan dalam ayat di dalam (i) dengan menggunakan teknik definisi perkataan.

(4/100)

(iv) Senarai dan huraikan secara ringkas **tiga** (**3**) algoritma bagi penyahtaksaan makna perkataan secara automatik.

(6/100)