UNIVERSITI SAINS MALAYSIA

First Semester Examination
2014/2015 Academic Session

December 2014/January 2015

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 **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. Stemming is a term used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem. A stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

(a) Explain how stemming can be used in applications like information retrieval to improve the accuracy of search results.

(5/100)

(b) List **four** (**4**) example words that require stemming in order to achieve better accuracy in applications mentioned in (a). List the words before and after stemming.

(4/100)

(c) One of the most widely used stemming algorithms is the Porter (1980) algorithm, which is based on a series of simple cascaded rewrite rules. One of the steps of the Porter Stemmer is given as follows.

Porter stemmer rules for plural:

SSES → SS
IES → I
SS → SS
S → ε

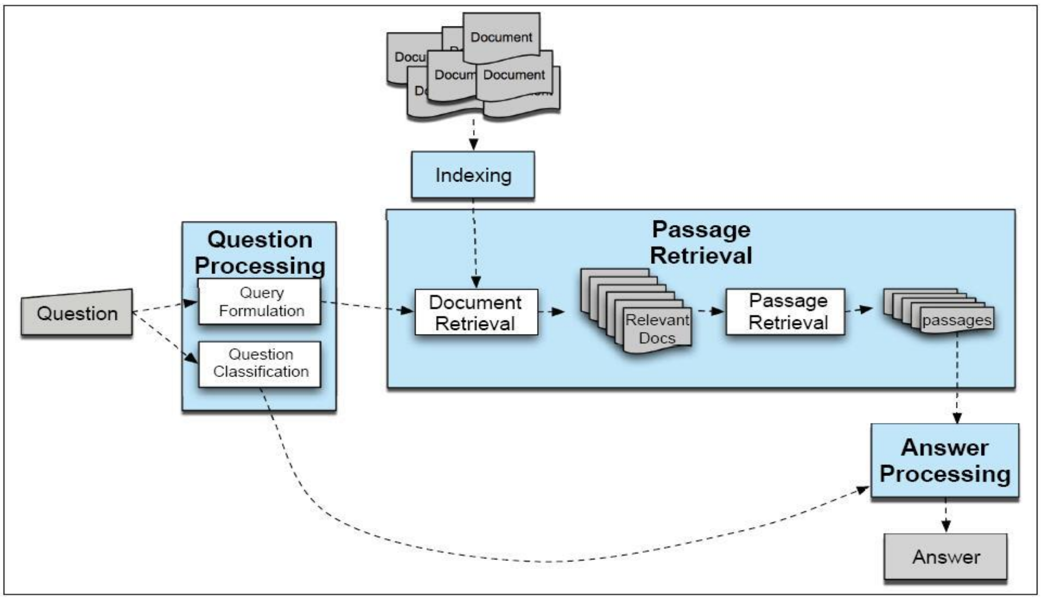
(i) Using the given rules, list two words that will be stemmed correctly using the Porter Stemmer and two words that will be stemmed wrongly using the Porter Stemmer.

(4/100)

(ii) Design a FST (Finite State Transducer) that implements the above Porter Stemmer rules.

(12/100)

2. Consider the following Question Answering system.



(a) What is the purpose of the above Question Answering system? Explain **three** (**3**) differences between Question Answering system and Information Retrieval system.

(5/100)

(b) Describe the functionalities of the following modules as shown in the above system.

(i) Question Processing.

(ii) Passage Retrieval.

(iii) Answer Processing.

(9/100)

(c) One of the tasks in question processing is to classify questions by its expected answer type. Give the expected answer **type** for each of the following questions.

(i) Who won the 1998 Nobel Peace Prize?

(ii) Why did David Koresh ask the FBI for a word processor?

(iii) What countries speak Spanish?

(iv) What is Ebola?

(v) How much does a regular pizza cost?

(5/100)

(d) Following are two question answering systems, i.e. QA-I and QA-II. A run was conducted on both systems using the question “Who invented the paper clip?”. The top five ranked answers are returned.

QA-I	**QA-II**
Who invented the paper clip?	
1. The giant paper clip in Sandvika, Norway	**1. A Norwegian, Johan Vaaler (1866–1910)**
2. A Norwegian, Johan Vaaler (1866–1910), has erroneously been identified as the inventor of the paper clip.	2. A large majority of different paper clip models were made
3. A Norwegian, Johan Vaaler (1866–1910)	3. A Norwegian, Johan Vaaler (1866–1910), has erroneously been identified as the inventor of the paper clip.
4. paper clip - wikipedia	4. The giant paper clip in Sandvika, Norway
5. A large majority of different paper clip models were made	5. paper clip - wikipedia

(i) Using the Mean Reciprocal Rank (MRR) measure, calculate the result score of the above question for both system QA-I and QA-II. The suggested correct answer is highlighted in bold. Which system has higher accuracy?

(3/100)

(ii) Discuss the limitation of this evaluation in terms of its benchmarked result.

(3/100)

3. (a) Polysemy and synonymy are two linguistic phenomena that limit effective search terms in word-based information retrieval systems, reducing both precision and recall. True or false? Explain your answer.

(5/100)

(b) The vector space model views documents and queries as vectors in a large multidimensional space. Briefly explain the benefit of this model.

(5/100)

(c) Supervised Machine Learning (ML) approaches seems to be a possible way to tackle the classification problem of word sense disambiguation (WSD).

(i) What are the approaches? Briefly explain those approaches.

(10/100)

(ii) Give an example of a simple measurable feature for the ML system to train a classifier to disambiguate the word *bat.*

(5/100)

4. One of the skills involved in language engineering is to solve practical problems (such as linguistic theory and their frequency of occurrence in real data) adequately with minimal complexity in the language models used.

(a) (i) Give an example of an application type where it is possible to build systems with two types of language models (e.g. involving simple information about individual words or involving complex information about how sentences convey meaning).

(5/100)

(ii) Justify that both types of models (that you had chosen in (i)) could be appropriate.

(2/100)

(iii) Indicate what criteria you might use for a possible given situation.

(4/100)

(b) (i) Give an example of an NLP system or approach that only knows about individual words (not about their context).

(4/100)

(ii) Explain what knowledge of words needed and give an overview of how it works. Give examples to illustrate your explanation.

(3/100)

(c) (i) Give an example of an NLP system or approach that models the structure of sentences and what they mean.

(4/100)

(ii) Explain what knowledge of language needed and give an overview of how it works. Give examples to illustrate your explanation.

(3/100)

1. Stemming adalah istilah yang digunakan dalam bahasa morfologi dan pencarian maklumat untuk menggambarkan proses untuk mengurangkan infleksi (atau kadang-kadang perolehan) perkataan ke dalam bentuk perkataan stem. Stem tidak semestinya sama dengan akar morfologi perkataan; adalah memadai jika perkataan-perkataan merujuk kepada *stem* yang sama, walaupun *stem* bukan akar yang sah.

(a) Terangkan bagaimana *stemming* boleh digunakan dalam aplikasi-aplikasi seperti pencarian maklumat untuk meningkatkan ketepatan hasil carian.

(5/100)

(b) Senaraikan **empat** (**4**) contoh perkataan yang memerlukan *stemming* untuk mencapai ketepatan yang lebih baik dalam aplikasi-aplikasi yang disebut dalam (a). Senarai perkataan sebelum dan selepas *stemming*.

(4/100)

(c) Salah satu daripada algoritma *stemming* yang paling banyak digunakan ialah algoritma *Porter* (1980). Algoritma ini adalah berdasarkan kepada peraturan tulis semula secara bersiri. Salah satu langkah dari *Porter Stemmer* diberikan seperti berikut.

Peraturan Porter stemmer untuk bentuk majmuk:

SSES → SS
IES → I
SS → SS
S → ε

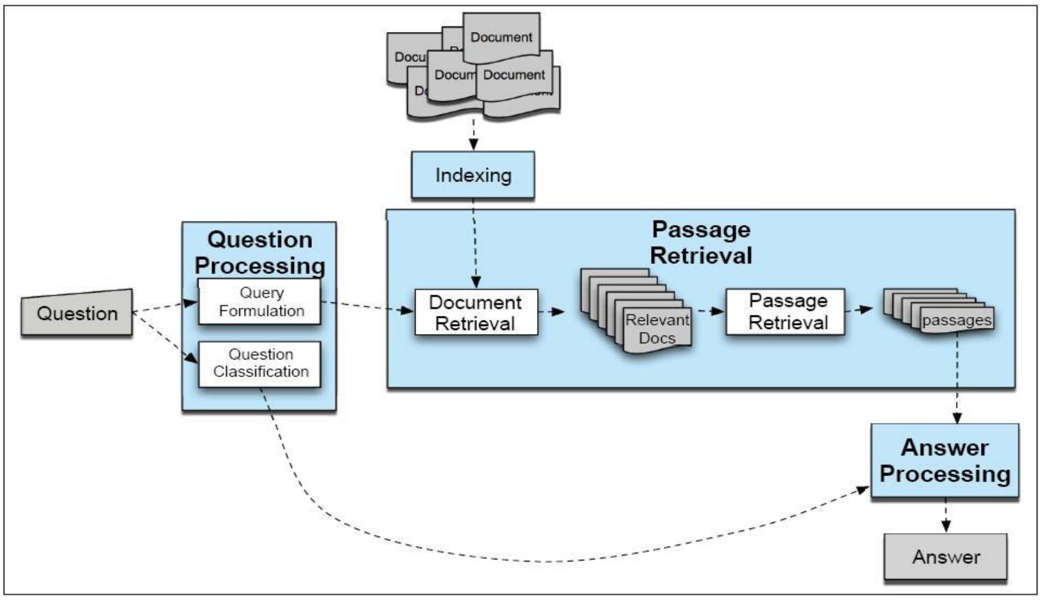
(i) Dengan menggunakan peraturan yang diberikan, senaraikan dua perkataan yang akan di-*stem* dengan betul menggunakan Porter Stemmer dan dua perkataan yang akan di-*stem* secara salah menggunakan Porter Stemmer.

(4/100)

(ii) Reka bentuk suatu FST (*Finite State Transducer*) yang melaksanakan peraturan *Porter Stemmer* di atas.

(12/100)

2. Pertimbangkan sistem Soal Jawab berikut.



(a) Apakah tujuan sistem Soal Jawab di atas? Terangkan **tiga** (**3**) perbezaan di antara sistem Soal Jawab dan sistem Pencarian Maklumat.

(5/100)

(b) Jelaskan fungsi-fungsi modul berikut seperti yang ditunjukkan di dalam sistem di atas.

(i) Pemprosesan Soalan.

(ii) Pencarian Petikan.

(iii) Pemprosesan Jawapan.

(9/100)

(c) Salah satu tugas dalam pemprosesan pertanyaan adalah untuk mengelaskan soalan mengikut jenis jawapan yang diingini. Nyatakan **jenis** jawapan yang diingini bagi setiap soalan berikut.

(i) *Who won the 1998 Nobel Peace Prize*?

(ii) *Why did David Koresh ask the FBI for a word processor*?

(iii) *What countries speak Spanish*?

(iv) *What is Ebola*?

(v) *How much does a regular pizza cost*?

(5/100)

(d) Di bawah adalah dua sistem soal jawab, i.e. QA-I and QA-II. Suatu kajian telah dijalankan ke atas kedua-dua sistem dengan menggunakan pertanyaan, "*Who invented the paper clip?*". Lima jawapan terbaik dikembalikan.

QA-I	QA-II
Who invented the paper clip?	
1. The giant paper clip in Sandvika, Norway	***1. A Norwegian, Johan Vaaler (1866–1910)***
2. A Norwegian, Johan Vaaler (1866–1910), has erroneously been identified as the inventor of the paper clip.	*2. A large majority of different paper clip models were made*
3. A Norwegian, Johan Vaaler (1866–1910)	*3. A Norwegian, Johan Vaaler (1866–1910), has erroneously been identified as the inventor of the paper clip.*
4. paper clip - wikipedia	*4. The giant paper clip in Sandvika, Norway*
5. A large majority of different paper clip models were made	*5. paper clip - wikipedia*

(i) Menggunakan kaedah *Mean Reciprocal Rank* (MRR), kira skor jawapan daripada soalan di atas untuk kedua-dua sistem QA-I dan QA-II. Jawapan betul yang dicadangkan ditulis dalam huruf tebal. Sistem manakah mempunyai ketepatan yang lebih tinggi?

(3/100)

(ii) Bincangkan batasan penilaian ini dari segi jawapan cadangannya.

(3/100)

3. (a) Polisemi dan sinonim merupakan fenomena linguistik yang membataskan carian terma yang berkesan bagi sistem capaian maklumat berdasarkan kata dan mengurangkan kedua-dua ketepatan dan ingatan. Benar atau palsu? Terangkan jawapan anda.

(5/100)

(b) Model ruang vektor melihat dokumen dan pertanyaan sebagai vektor dalam ruang pelbagai dimensi yang besar. Terangkan dengan ringkas tentang kebaikan model ini.

(5/100)

(c) Pendekatan pembelajaran mesin terselia boleh mengatasi masalah pengkelasan penyahtaksaan makna (WSD).

(i) Apakah pendekatan-pendekatan itu? Terangkan secara ringkas pendekatan-pendekatan tersebut.

(10/100)

(ii) Berikan suatu contoh ciri-ciri mudah dan boleh diukur bagi sistem ML belajar satu pengkelas untuk menyahtaksa perkataan *bat*.

(5/100)

4. Salah satu kemahiran yang terlibat dalam bidang kejuruteraan bahasa adalah mencari penyelesaian untuk menyelesaikan masalah praktikal (seperti teori linguistik dan frekuensi kewujudannya dalam data sebenar) setepatnya dengan hanya melibatkan kekompleksan yang minimum dalam model bahasa yang digunakan.

(a) (i) Berikan satu contoh jenis aplikasi di mana ia adalah mungkin untuk membina sistem dengan dua jenis model bahasa.

(5/100)

(ii) Justifikasikan yang kedua-dua model itu (yang anda pilih dalam (i)) adalah sesuai.

(2/100)

(iii) Tunjukkan kriteria apakah yang anda mungkin gunakan untuk sesuatu kemungkinan keadaan yang diberikan.

(4/100)

(b) (i) Berikan satu contoh sistem NLP atau pendekatan yang hanya tahu tentang perkataan secara individu (bukan tentang konteks mereka).

(4/100)

(ii) Jelaskan apakah pengetahuan tentang perkataan yang diperlukan dan berikan gambaran bagaimana ia berfungsi. Berikan contoh untuk menerangkan jawapan anda.

(3/100)

(c) (i) Berikan satu contoh sistem NLP atau pendekatan yang memodelkan struktur ayat dan maksud-maksudnya.

(4/100)

(ii) Jelaskan apakah pengetahuan tentang bahasa yang diperlukan dan berikan gambaran bagaimana ia berfungsi. Berikan contoh untuk menerangkan jawapan anda.

(3/100)

- oooOooo -