

This question paper must be returned. Candidates are not permitted to remove any part of it from the examination room.

SEAT NUMBER: _____ ROOM: _____

FAMILY NAME _____

OTHER NAMES _____

STUDENT NUMBER _____



Mid-Year Examinations 2007

Unit : COMP348 Document Processing and the Semantic Web

Date : Wednesday 13 June, 1.50 pm

Time Allowed : 3 hours plus ten minutes reading time

Total Questions : 11

Instructions : Answer ALL questions.

This examination consists of two sections, A and B, each worth 30 marks. The answers to each section should be written in a separate answer book – one answer book per section. Write your name, student number and the section name (A or B) on the cover of each answer book.

Write your name and student number at the top of this page. You may do rough working on this question paper, but all answers MUST be submitted as described above. Hand in this question paper at the end of the examination.

Materials Permitted : Calculators are not allowed. Dictionaries are not allowed.

SECTION B: 9 QUESTIONS, 90 MARKS

Write your answers to this section in a separate book and write “Section B” on the first page of the book.

To pass the unit you need to answer all questions tagged with **CORE** correctly. It is in your interest to attempt the CORE questions first.

1. (11 marks) Part of Speech Tagging

- (a) (5 marks) **CORE** Explain the differences between closed and open class word types.
- (b) (6 marks) The following text is a fragment of the Brown Corpus annotated with the parts of speech:

Robert/NP Snodgrass/NP ,/, state/NN GOP/NN chairman/NN ,/, said/VBD a/AT meeting/NN held/VBN Tuesday/NR night/NN in/IN Blue/JJ-TL Ridge/NN-TL brought/VBD enthusiastic/JJ responses/NNS from/IN the/AT audience/NN ./ . State/NN-TL Party/NN-TL Chairman/NN-TL James/NP W./NP Dorsey/NP added/VBD that/CS enthusiasm/NN was/BEDZ picking/VBG up/RP for/IN a/AT state/NN rally/NN to/TO be/BE held/VBN Sept./NP 8/CD in/IN Savannah/NP at/IN which/WDT newly/RB elected/VBN Texas/NP Sen./NN-TL John/NP Tower/NP will/MD be/BE the/AT featured/VBN speaker/NN ./ . In/IN the/AT Blue/JJ-TL Ridge/NN-TL meeting/NN ,/, the/AT audience/NN was/BEDZ warned/VBN that/CS entering/VBG a/AT candidate/NN for/IN governor/NN would/MD force/VB it/PPO to/TO take/VB petitions/NNS out/RP into/IN voting/VBG precincts/NNS to/TO obtain/VB the/AT signatures/NNS of/IN registered/VBN voters/NNS ./ . Despite/IN the/AT warning/VBG ,/, there/EX was/BEDZ a/AT unanimous/JJ vote/NN to/TO enter/VB a/AT candidate/NN ,/, according/IN to/IN Republican-s/NPS who/WPS attended/VBD ./ . When/WRB the/AT crowd/NN was/BEDZ asked/VBN whether/CS it/PPS wanted/VBD to/TO wait/VB one/CD more/AP term/NN to/TO make/VB the/AT race/NN ,/, it/PPS voted/VBD no/RB -/- and/CC there/EX were/BED no/AT dissents/NNS ./ .

- i. (2 marks) Estimate $P(\text{NN})$
- ii. (2 marks) Estimate $P(\text{NN}|\text{AT})$
- iii. (2 marks) Estimate $P(\text{a}|\text{AT})$

2. (13 marks) Parsing

Some of the following questions are based on this grammar:

$S \rightarrow NP VP$
 $NP \rightarrow Det N$
 $VP \rightarrow V NP PP$
 $PP \rightarrow P NP$
 $Det \rightarrow \text{'the'} \mid \text{'a'}$
 $N \rightarrow \text{'dog'} \mid \text{'biscuit'} \mid \text{'park'}$
 $P \rightarrow \text{'in'}$
 $V \rightarrow \text{'ate'}$

- (a) **(3 marks) CORE** Using the above grammar, write the parse tree of the sentence *the dog ate a biscuit in the park*.
- (b) **(5 marks)** Trace the parse of the same sentence *the dog ate a biscuit in the park* using a top-down, left-right, depth-first strategy.
- (c) **(5 marks)** Explain the purpose of the fundamental rule of chart parsing and write the chart entries that result from applying the fundamental rule of chart parsing given the following chart entries:

VP \rightarrow V . NP PP (2,3)
 NP \rightarrow Det N . (3,5)

3. **(9 marks)** Word Sense Disambiguation

- (a) **(4 marks) CORE** What is the goal of Word Sense Disambiguation? Explain the task using an example.
- (b) **(5 marks)** The Naïve Bayes approach to Word Sense Disambiguation introduces the following independence assumption:

$$P(V|s) = \prod_{j=1}^n P(v_j|s)$$

Where $V = (v_1, v_2, \dots, v_n)$ is a context vector. What is the nature of the independence assumption made here and why is the assumption not true?

4. **(10 marks)** Information Retrieval Suppose that you have a tiny collection of just three one-sentence documents:

D1 Eventually technology will enable quantum computers.

D2 The computer language used was Python.

D3 Language Technology uses a computer to process language.

- (a) **(3 marks) CORE** Write the inverted indices of the keywords *language*, *technology*, and *computer* (ignore capitalisation).
- (b) **(7 marks)** Using the three keywords of the previous question, write the vector that represents D1 and D2 using the *tf.idf* score as the vector weights.

5. **(13 marks)** Document Summarisation

- (a) **(4 marks) CORE** Explain what is the purpose of an indicative summary and what is the purpose of an informative summary.
- (b) **(9 marks)** List and explain three common techniques to detect the most important sentences of a document in order to build its summary.

6. **(11 marks)** Information Extraction and Named Entity Recognition

- (a) **(4 marks) CORE** What are gazetteers and why are they useful for the Named Entity Recognition task?
- (b) **(4 marks)** What are the differences between the preparsing and the parsing stages in Hobb's Information Extraction System architecture?

- (c) **(3 marks)** The following text fragment has been annotated with Named Entity information. Transform these annotations into the sort of annotations what would be used by the classifier of a statistical Named Entity Recognition system.

<ORGANIZATION>Microsoft Corp.</ORGANIZATION> shares fell late
<DATE>Thursday</DATE>.

7. **(13 marks)** Question Answering

- (a) **(7 marks) CORE** Complete the following simple question classifier so that all the questions below are correctly classified.

```
import re
def classify(question):
    "return the expected answer type"
    if re.match(r'^\s*Who\b'):
        return "PERSON"
    # ... add more rules here ...
```

- *Who invented the Theory of Relativity?*
- *What is the capital of Australia?*
- *How much costs an iPod Nano?*
- *How many ships were in Capt. Cook's expedition to Australia?*
- *Name the capital of France*
- *Name the inventor of the light bulb*

- (b) **(6 marks)** Explain the goal of each of the following components of a Question Answering system:

- Document Preselection
- Candidate Answer Selection
- Answer Extraction

8. **(10 marks)** Semantic Web

- (a) **(4 marks) CORE** What is the general purpose of the Semantic Web?
- (b) **(6 marks)** OWL has three variants: Full, DL, and Lite. Explain (at a very high level) the differences among them.

END OF SECTION B. PLEASE REVISE YOUR ANSWERS. **MAKE SURE YOU HAVE ANSWERED ALL CORE QUESTIONS.**