

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 A: 7 QUESTIONS, 30 MARKS

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

1. (3 marks) Tokenisation

- (a) (1 mark) What is the most salient difficulty of attempting tokenisation on an unsegmented language?
- (b) (2 marks) It is said that the problem of sentence segmentation is usually reduced to the problem of disambiguating sentence delimiters. Give an example of text where such disambiguation is needed and explain the example.

2. (4 marks) Morphological Analysis

- (a) (2 marks) Describe what is inflectional morphology and what is derivational morphology. Give an example of a word displaying inflectional morphology and another of a word displaying derivational morphology.
- (b) (2 marks) What is the measure of each of the following words according to the Porter Stemmer algorithm?
 - i. *policy*
 - ii. *analysis*
 - iii. *discovery*
 - iv. *yeast*

3. (6 marks) Statistical Methods

- (a) (2 marks) Why is it important to use different data for training and for testing in statistical methods in general?
- (b) (2 marks) What are the general advantages and inconveniences of corpus-based approaches against rule-based approaches of natural language processing applications?
- (c) (2 marks) The following table shows the characteristics of a sample population of 1000 individuals according to eye colour and hair colour:

	Blond hair	Brunette hair
Blue eyes	200	50
Brown eyes	100	650

- i. What is the probability $p(\text{Blue eyes})$?
- ii. What is the probability $p(\text{Brown eyes}|\text{Brunette hair})$?

4. (6 marks) Text Classification

- (a) (2 marks) Within the context of text classification, the Naïve Bayes method computes $P(c_j|X)$, that is the probability of a class c_j given a vector of features $X = [x_1, x_2, \dots, x_n]$. Explain the independence assumption inherent in this method.
- (b) (2 marks) What is the reason of the use of kernel functions in a SVM (Support Vector Machine) classifier? Illustrate the effect of a kernel function to enhance classification.

- (c) (2 marks) Describe two popular feature selection methods for text classification. You do not need to write formulas but you need to explain the reason why these are sensible methods for feature selection.

5. (4 marks) Word Sense Disambiguation

- (a) (2 marks) What are the problems and advantages of unsupervised methods (versus supervised methods) for word sense disambiguation?
- (b) (2 marks) Discuss how a dictionary can be used as a resource to determine the word senses of expressions such as *pine cone*.

6. (3 marks) Grammar and Parsing

- (a) (1 mark) Given the sentence *the dog saw a man in the park*, draw a syntactic analysis using a constituency-based approach and another using a dependency-based approach.
- (b) (2 marks) What is the problem of left recursion and in what parsing strategy might the problem appear?

7. (4 marks) Machine Translation

- (a) (2 marks) Draw the Vauquois Triangle and illustrate how it explains the three main rule-based methods to machine translation.
- (b) (2 marks) What are the advantages and disadvantages of interlingua-based machine translation versus transfer-based machine translation?

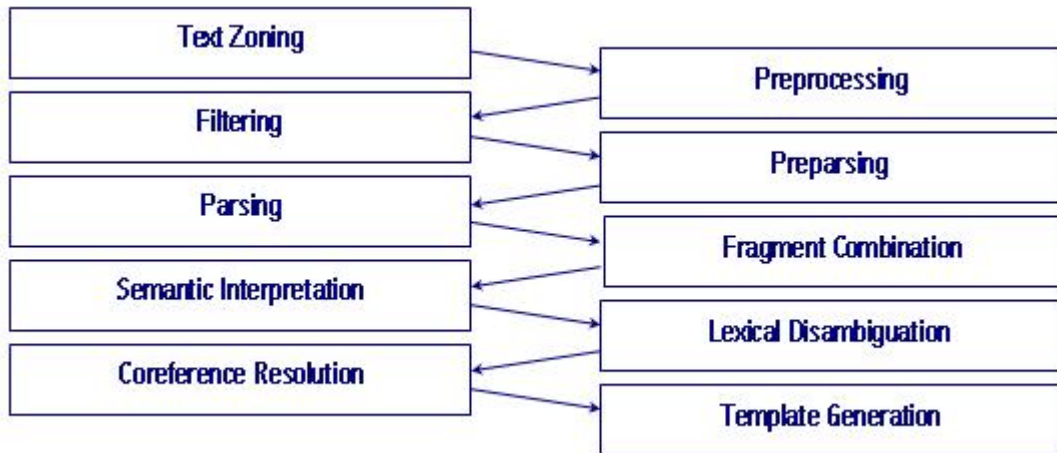
END OF SECTION A. PLEASE TAKE ANOTHER ANSWERBOOK.

SECTION B: 4 QUESTIONS, 30 MARKS

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

1. (9 marks) Information Extraction

Below is a screen-shot of a generic architecture of an information extraction system.



- (5 marks) Explain what each module in this architecture contributes to the information extraction process.
- (2 marks) Come up with a scenario in which an information extraction system could be successfully used.
- (2 marks) Discuss why the information extraction is an easier task than natural language understanding.

2. (5 marks) Named Entity Recognition

Named entity recognition and classification is a subtask of information extraction. Prototypical named entities are PERSON, ORGANISATION, LOCATION, ARTIFACT, and DATE. Given the following text:

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field. Dr. Maddox will be the firm’s CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver’s brother, Ambrose, follows more in his father’s footsteps and will be the CEO of L.J.G. headquartered in the Maddox family’s hometown of La Jolla, CA.

- (4 marks) Identify all named entities and classify them as one of the prototypical named entities mentioned above.

- (b) (1 marks) Discuss why recognising named entities automatically is a problem for a computer.

3. (4 marks) Chunking versus Parsing

A chunker is a tool that finds contiguous sequences of tokens for a tagged input text and groups them together into chunks.

```
grammar = r"""
```

```
NP:
```

```
{<DT>?<JJ>*<NN>}
```

```
{<NNP>+}
```

```
"""
```

```
cp = chunk.Regexp(grammar)
```

```
tagged_tokens = [("the", "DT"), ("little", "JJ"), ("cat", "NN"),  
("sat", "VBD"), ("on", "IN"), ("the", "DT"), ("mat", "NN")]
```

```
cp.parse(tagged_tokens)
```

- (a) (2 marks) Explain in detail what the output of the above Python program looks like.
- (b) (2 marks) State what the difference is between a chunker and a full parser, and illustrate what kind of output a full parser would produce for the tagged tokens above.

4. (12 marks) Semantic Web

The Semantic Web is based on RDF (Resource Description Framework) and OWL (Web Ontology Language).

- (a) (2 marks) Explain what the parts of an RDF statement are.
- (b) (2 marks) Discuss what RDF Schema is and what kind of limitations RDF Schema has.
- (c) (2 marks) Name four constructs which are available in OWL DL but not in RDF/RDF Schema.
- (d) (2 marks) Explain how Description Logic is related to first-order predicate logic and why Description Logic is interesting as the formal underpinning of the Semantic Web.
- (e) (2 marks) A Description Logic reasoner such as Pellet provides the following standard inference services: subsumption checking, satisfiability checking, classification, and realisation. Explain what these inference services do and provide an example.
- (f) (2 marks) Paraphrase in full natural language what the following information expressed in OWL Abstract Syntax specifies:

```
class (VegetarianPizza complete  
Pizza  
complementOf( restriction hasPart someValuesFrom Meat)  
complementOf( restriction hasPart someValuesFrom Fish))
```

END OF SECTION B. PLEASE REVISE YOUR ANSWERS.