

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 2009

Unit : COMP348 Document Processing and the Semantic Web

Date : Monday 22 June, 9:20 am

Time Allowed : 3 hours plus ten minutes reading time

Total Questions : 14

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: 6 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)** Language Technology

Explain the main advantages and inconveniences of deep approaches versus shallow approaches in Language Technology applications.

2. **(4 marks)** Tokenisation and Sentence Segmentation

What are the main problems with abbreviations and numbers when tokenising and segmenting a text? How might you solve these problems?

3. **(7 marks)** Part of Speech Tagging

- (a) **(2 marks)** The following text is a fragment of the Brown Corpus annotated with the parts of speech using the simplified tagset:

Davis/NP received/VD 1,119/NUM votes/N in/P Saturday's/N election/N ,/, and/CNJ Bush/NP got/VD 402/NUM ./ . Ordinary/N Carey/NP Williams/NP ,/, armed/VN with/P a/DET pistol/N ,/, stood/VD by/P at/P the/DET polls/N to/TO insure/V order/N ./ . “/“ This/V was/V the/DET coolest/ADJ ,/, calmest/ADJ election/N I/PRO ever/ADV saw/VD ”/” ,/, Colquitt/NP Policeman/N Tom/NP Williams/NP said/VD ./ . “/“ Being/V at/P the/DET polls/N was/V just/ADV like/CNJ being/V at/P church/N ./ . I/PRO didn't/V smell/V a/DET drop/N of/P liquor/N ,/, and/CNJ we/PRO didn't/V have/V a/DET bit/N of/P trouble/N ”/” ./ . The/DET campaign/N leading/VG to/P the/DET election/N was/V not/* so/DET quiet/ADJ ,/, however/WH ./ .

Estimate $P(N)$ according to the statistics of this fragment.

- (b) **(2 marks)** Write a Python program that estimates $P(\text{the}|\text{DET})$ using the entire Brown corpus.

```
from nltk import corpus
words = corpus.brown.tagged_words(simplify_tags=True)
#... continue from here ...
```

- (c) **(3 marks)** The Brill Part of Speech tagger uses rules that are automatically learnt from a corpus. Explain the general approach used by the Brill Part of Speech tagger to learn the rules. Use high-level explanations, you don't need to write the algorithm.

4. **(6 marks)** Text Classification

- (a) **(3 marks)** The general approach to supervised text classification is to convert the text into a feature vector and then apply a statistical classifier. Explain briefly what a feature is in the context of text classification, and how it relates to this vector.
- (b) **(3 marks)** Three common statistical classifiers are Naive Bayes, K-nearest, and Support Vector Machines. Explain one of them in detail, indicating how it separates the data into classes.

5. (6 marks) Parsing

The next questions are based on this grammar:

$S \rightarrow NP VP$

$NP \rightarrow Det N$

$VP \rightarrow V NP PP$

$PP \rightarrow P NP$

$Det \rightarrow 'the' \mid 'a'$

$N \rightarrow 'dog' \mid 'biscuit' \mid 'park'$

$P \rightarrow 'in'$

$V \rightarrow 'ate'$

(a) (3 marks) Trace a sequence of operations of the sentence *the dog ate a biscuit in the park* that leads to a correct parse using a shift-reduce parser that is loaded with the above grammar.

(b) (3 marks) Explain the meaning of the following chart entries in a chart parser and indicate what would be the result of applying the fundamental rule of chart parsing:

$VP \rightarrow V NP \cdot PP \quad (2,5)$

$PP \rightarrow P NP \cdot \quad (5,7)$

6. (4 marks) Machine Translation

Explain the Vauquois Triangle of Machine Translation and how it relates to the three main Machine Translation approaches.

END OF SECTION A. PLEASE TAKE ANOTHER ANSWERBOOK.

SECTION B: 8 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. (5 marks) Semantic Web Architecture

Draw a picture of the Semantic Web architecture and label each box in this architecture with a name. Describe what role each box plays in this architecture in a bullet point list. Each bullet point should consist of at least three sentences.

2. (5 marks) XML and XML Schema

An XML schema describes the legal structure of an XML document and makes it possible to validate the XML automatically against the schema. This helps to reduce the number of potential errors in an XML document. Below is an XML Schema called "shiporder.xsd"

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="shiporder">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="orderperson" type="xs:string"/>
      <xs:element name="shipto">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="name" type="xs:string"/>
            <xs:element name="address" type="xs:string"/>
            <xs:element name="city" type="xs:string"/>
            <xs:element name="country" type="xs:string"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
      <xs:element name="item" maxOccurs="unbounded">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="title" type="xs:string"/>
            <xs:element name="note" type="xs:string" minOccurs="0"/>
            <xs:element name="quantity" type="xs:positiveInteger"/>
            <xs:element name="price" type="xs:decimal"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
    </xs:sequence>
    <xs:attribute name="orderid" type="xs:string" use="required"/>
  </xs:complexType>
</xs:element>
</xs:schema>
```

Write a valid XML document (shiporder.xml) that links to the above XML schema and contains the following data together with the required markup.

```
John Smith
Ola Nordmann, Langgt 23, 4000 Stavanger, Norway
Empire Burlesque, Special Edition, 1, 10.90
Hide your heart, 1, 9.90
```

3. (3 marks) SPARQL

FOAF (= Friend of a Friend) is a simple machine-readable ontology describing persons, their activities and their relations to other people and objects. Below is a snippet of a FOAF ontology in RDF. Write a SPARQL query that finds all persons who have a name as well as a nickname.

```
<?xml version="1.0"?>

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:foaf="http://xmlns.com/foaf/0.1/">

<foaf:Person>
  <foaf:name>Dave Beckett</foaf:name>
  <foaf:mbox_sha1sum>970987f991961f2553a1bf2574166fa29befbccb</foaf:mbox_sha1sum>
  ...

<foaf:knows>
  <foaf:Person>
    <foaf:name>Tim Berners-Lee</foaf:name>
    <foaf:mbox_sha1sum>965c47c5a70db7407210cef6e4e6f5374a525c5c</foaf:mbox_sha1sum>
    <rdfs:seeAlso rdf:resource="http://www.w3.org/People/Berners-Lee/card.rdf" />
    <foaf:homepage rdf:resource="http://www.w3.org/People/Berners-Lee/" />
  </foaf:Person>
</foaf:knows>

<foaf:knows>
  <foaf:Person>
    <foaf:name>Morten Frederiksen</foaf:name>
    <foaf:mbox_sha1sum>461179310021b2185ad7f67f14e5d4deb2107c47</foaf:mbox_sha1sum>
    <foaf:nick>morti</foaf:nick>
  </foaf:Person>
</foaf:knows>

<foaf:knows>
  <foaf:Person>
    <foaf:name>Jo Walsh</foaf:name>
    <foaf:mbox_sha1sum>828414515d398b42268a6c2ed879dc505369223a</foaf:mbox_sha1sum>
    <foaf:nick>zool</foaf:nick>
  </foaf:Person>
</foaf:knows>
```

```

<foaf:knows>
  <foaf:Person>
    <foaf:name>Damian Steer</foaf:name>
    <foaf:nick>shellac</foaf:nick>
  </foaf:Person>
</foaf:knows>

</foaf:Person>
</rdf:RDF>

```

4. (3 marks) RDF(S) and OWL

Give three concrete examples that show what kind of terminological statements you can express in OWL DL but not in RDF(S). You can express these examples in natural language for instance in the following way: "Person is the disjoint union of Male and Female".

5. (4 marks) Expressing n-ary Relations in Description Logic

Most description logics can only be used to express binary relations between individuals. However, reification of binary relations makes it possible to express statements such as:

Anna Grau teaches COMP448 on Monday in E6A.

Show how you can represent the above statement about Anna Grau in description logic using reification and thematic roles. You can use the following two predicates (**instance IN C**) and (**related IN1 IN2 R**) for this task. Note that **IN** stands for a name of an instance, **C** stands for a concept name and **R** for a name of a binary relation.

6. (4 marks) Reasoning Services

Consistency checking, concept satisfiability, classification and realisation are the four standard description logic reasoning services. Explain what these reasoning services are used for and give a concrete example that illustrates how these inferences work.

7. (3 marks) Description Logics and Rule Languages

Explain why it is difficult to combine a description logic with a rule language, discuss what kind of solutions exist and state what kind of advantages/disadvantages these solutions have.

8. (3 marks) Controlled Natural Languages

Controlled natural languages can be used as a high-level interface language to a description logic knowledge base. That means you can translate automatically from a description logic representation into a controlled natural language and vice versa. Paraphrase the following snippet of description logic in controlled natural language:

```

Class(VegetarianPizza complete
  complementOf(restriction(hasTopping
    someValuesFrom(MeatTopping)))
  complementOf(restriction(hasTopping
    someValuesFrom(FishTopping)))
  Pizza)

```

END OF SECTION B. PLEASE REVISE YOUR ANSWERS.