UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR09028 FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

May 2020

13:00 to 15:00

INSTRUCTIONS TO CANDIDATES

Answer all of Part A and TWO questions from Part B.

Part A is COMPULSORY.

The short answer questions in Part A are each worth 3 marks, 24 marks in total. Each of the three questions in part B is worth 13 marks; answer any TWO of these.

This is an OPEN BOOK examination.

Year 3 Courses

Convener: S.Ramamoorthy
External Examiners: S.Rogers, S.Kalvala, H.Vandierendonck

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

Part A

Answer ALL questions in Part A.

Try to make your answers as thorough as possible. However, they need not be lengthy: from one or two sentences up to a paragraph. When two terms are contrasted, make sure your short definitions of each make clear where the contrast lies. Each question is worth three marks, 24 marks in total for this section.

1. What is the most common **frequency distribution** for natural language phenomena? Give its name, and draw it. Give two examples of natural language phenomena that have this distribution.
2. What kind of model would you use to predict that "The man talked" is more likely to occur in a corpus than "man talked the"? Explain why your choice will produce a lower probability for the latter phrase.
3. Name four applications for which **ngram language modelling** is useful.
4. What is a potential problem with Add 1 Laplace **smoothing**, and how does Good-Turing avoid this problem?
5. Name and define two types of **syntactic ambiguity**, give a linguistic example of each, and briefly explain why they present a major challenge for parsing.
6. How do you evaluate a constituency-based **statistical parser**?
7. Suggest the semantic role labels you would expect to see in a corpus if the following sentence were annotated using FrameNet: *John carried Mary's bag to the bus for her for a £3 tip yesterday.*
8. What are **selectional restrictions**? Consider the following example sentence:

   Cambridge voted conservative.

   In view of your answer about selectional restrictions, give an account of this sentence in terms of the word sense ambiguity of words that, like *Cambridge*, name places.

Part B

Answer two questions in Part B.

1. **Determining text authorship**

   A manuscript has been uncovered in the basement of a disused rectory in Yorkshire, bound into the back of a mid 19th-century diary. The diary itself describes it as “a faithful copy, in my own hand, of a composition by the daughter of my predecessor here as curate, of which the original is now lost.” The manuscript has no titlepage, or any other indication of authorship. The possibility that this is a hitherto unknown work by Charlotte or Emily Brontë sets the literary world buzzing. But controversy persists on which of the famous sisters wrote it. You have been given access to a digital version of the manuscript, of both Charlotte’s *Jane Eyre* and Emily’s *Wuthering Heights*, and a wide range of other contemporary fiction. You are now tasked with answering the question as to which sister wrote the manuscript.

   (a) What is the underlying assumption that justifies using language modelling to predict which sister wrote the manuscript? [*1 mark*]

   (b) What metric would you use to test your hypotheses? [*1 mark*]

   (c) Describe the specific technique you would use to cope with sparse training data, and justify your choice. [*2 marks*]

   (d) Describe in detail the separate data sets that you would train (and then test) your language models on, and justify your choices. In particular, describe how you would establish baselines for the task. [*7 marks*]

   (e) Describe an outcome of your experiments that would be conclusive, and compare it with an outcome that would be inconclusive. Justify your answer. [*2 marks*]

2. **Parsing**

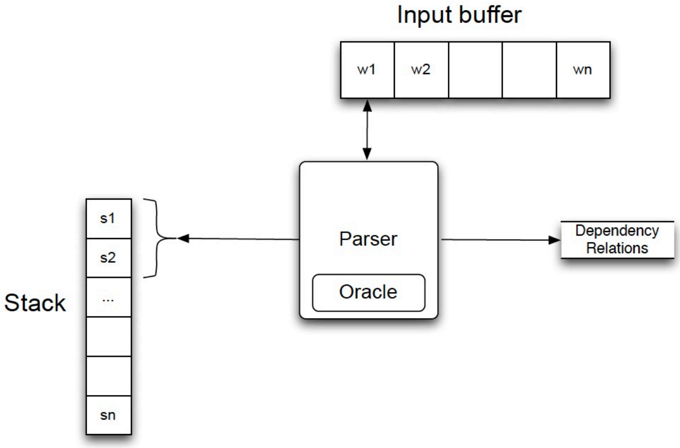
(a) Describe why parsing the following sentences with an unlexicalised probabilistic context free grammar (PCFG) is problematic:

Kim discussed the students' grades in the staff meeting
Kim discussed the students' grades in the engineering exam

[2 marks]

(b) Draw the correct constituent parses for each of the above examples. *[4 marks]*

(c) A shift reduce parser has the following components:



The parser examines the top two elements of the stack and uses the oracle to decide which action to take, based on the current configuration. Define the three actions based on the above figure. *[2 marks]*

(d) Draw the correct dependency parse for *Kim discussed the students' grades*. Use the sequence of actions that a shift-reduce parser would assign to this analysis. *[5 marks]*

3. **Lexical Semantics**

(a) Explain the difference between homonymy and polysemy. Illustrate your answer with examples of each kind of ambiguity. [*3 marks*]

(b) According to WordNet, the noun **school** has the following 7 senses:

sn1: school
(an educational institution)
"the school was founded in 1900"

sn2: school, schoolhouse
(a building where young people receive education)
"the school was built in 1932"; "he walked to school every morning"

sn3: school, schooling
(the process of being formally educated at a school)
"what will you do when you finish school?"

sn4: school
(a body of creative artists or writers or thinkers linked by a similar style or by similar teachers)
"the Venetian school of painting"

sn5: school, schooltime, school day
(the period of instruction in a school; the time period when school is in session)
"stay after school"; "he didn't miss a single day of school"; "when the school day was done we would walk home together"

sn6: school
(an educational institution's faculty and students)
"the school keeps parents informed"; "the whole school turned out for the game"

sn7: school, shoal
(a large group of fish)
"a school of small glittering fish swam by"

Cluster these senses using the definitions of homonymy and polysemy you gave in part (a). For any senses that are polysemous, give an argument as to how the senses are related. For each set of senses that you argue is polysemous, give at least one example of another word form that exhibits the same sense ambiguities. [*5 marks*]

(c) In what ways does a Naive Bayes approach to word sense disambiguation fall short when handling metaphor, for instance *Education unlocks doors to new worlds??* [*5 marks*]