UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR09028 FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

Thursday 16 th August 2018

14:30 to 16:30

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.

Use one script book for part A and one book for each question in Part B; that is, three books in all.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

Year 3 Courses

Convener: C. Stirling
External Examiners: S.Rogers, A. Donaldson, S. Kalvala

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

Part A

Answer ALL questions in Part A.

Your answers should be thorough—as opposed to being as brief as possible—but 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. Draw a normal distribution and a Zipfian distribution. Which distribution do word ngrams follow?

2. Assuming a trigram language model with $\langle s \rangle$ and $\langle /s \rangle$ respectively marking the beginning and end of a sentence, give an expression for the joint probability of the sentence "I saw a shell on the beach" made up of simpler probabilities.

3. Show the dynamic programming table for computing the minimum edit distance for converting *gold* into *cola*. Assume that it costs 1 to add a character, 1 to delete a character, and 2 to substitute one character for another.

4. Assign the sentence *Kim read a book quickly* its dependency parse.

5. Using a pair of linguistic examples of your choice, describe a problem one encounters when parsing with a probabilsitic context free grammar (PCFG).

6. Give the formula for the Naive Bayes model for word sense disambiguation, assuming that it uses the features $f_1, \ldots f_n$.

7. Give the values of the 6 element vector for the target word *character* in the sentence below, where the vector is defined as follows: the first element is the nearest open class word to the left of the target word; the second element is the second nearest open class word to the left of the target word; the third (and fourth) elements are respectively 0 if the nearest word to the left (right) of the target word is open class and 1 otherwise; the fifth and sixth elements are respectively the lemmas of the nearest and second nearest open class words to the right of the target word.

   > *Kim said he did not know the character was trumped up to be the main role in the movie.*

8. Suggest a way in which you can evaluate a model that predicts a paraphrase of the form *enjoy VP-ing NP* for a phrase of the form form *enjoy NP*.

Part B

Answer two questions in Part B.

1. **Determining Text Authorship**

   A manuscript has been found in the attic of a house in London where both the 19th century authors Charles Dickens and Anthony Trollope are known to have stayed. The manuscript has no title page, nor any other indication of authorship. The possibility that this is a hitherto unknown work by one of these famous authors sets the literary world buzzing.

   But is it? And if so, is it by Dickens, or by Trollope?

   To help answer these questions, you have access to the following resources: (i) a machine readable version of the manuscript, and (ii) machine readable versions of all of the novels written by Dickens, all the novels written by Trollope, and a wide range of novels written in the 19th century by other (known) authors and more contemporary fiction; and (iii) language modelling tools and software packages for statistical calculations.

   (a) Drawing on the langauge modelling technologies discussed in lectures and the labs, design an experiment to answer the above questions about the authorship of the manuscript. Be sure to set out your background assumptions and the hypotheses you would be trying to test. Also, be sure to describe the modeling technique(s) you would use, how you would train the models and how you would use them to confirm or deny your hypotheses. [*9 marks*]

   (b) In what circumstances would you be confident, and in what circumstances would you be suspicious of the conclusions of your experiments (as between the answers "the manuscript is by Dickens" vs. "the manuscript is by Trollope" vs. "the manuscript is by neither author"). Why? [*4 marks*]

S $\longrightarrow$ NP VP	1.0
N $\longrightarrow$ *ladies*	0.05
N $\longrightarrow$ *walk*	0.05
N $\longrightarrow$ *dance*	0.7
N $\longrightarrow$ MOD N	0.2
MOD $\longrightarrow$ *ladies*	0.6
MOD $\longrightarrow$ *dance*	0.4
DET $\longrightarrow$ *the*	1.0
NP $\longrightarrow$ DET N	1.0
VP $\longrightarrow$ V0	1.0
V0 $\longrightarrow$ *dance*	0.4
V0 $\longrightarrow$ *walk*	0.6

Figure 1: A simple probabilistic context free grammar

2. **Parsing**

(a) Using the (probabilistic) context free grammar in Figure 1, write the well-formed substring table (WFST) for *The ladies dance*. Use arrows to indicate which children create which parents. [*6 marks*]

(b) The probabilities in Figure 1 were calculated using Maximum Likelihood Estimates (MLE) from a treebank. What's the formula that was used? [*2 marks*]

(c) Is *the ladies dance* more likely to be a sentence or a noun phrase? Show the detailed calculations that prompt your answer. [*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 word **magazine** has the following 6 noun senses (the example uses are those given in WordNet, but for some senses WordNet gives no example uses):

Nouns:

sn1: magazine, mag
(a periodic publication containing pictures and stories and articles of interest to those who purchase it or subscribe to it)
it takes several years before a magazine starts to break even or make money

sn2: magazine
(product consisting of a paperback periodic publication as a physical object)
tripped over a pile of magazines

sn3: magazine, magazine publisher
(a business firm that publishes magazines)
he works for a magazine

sn4: magazine, cartridge
(a light-tight supply chamber holding the film and supplying it for exposure as required)

sn5: magazine, powder store, powder magazine
(a storehouse (as a compartment on a warship) where weapons and ammunition are stored)

sn6: cartridge holder, cartridge clip, clip, magazine
(a metal frame or container holding cartridges; can be inserted into an automatic gun)

Cluster these senses using the definitions of homonymy and polysemy you gave in part (a). Justify your answer. [*4 marks*]

(c) Explain why polysemous sense ambiguity creates a major challenge for creating an online lexical resource that defines the meaning of words, such as WordNet. Justify your answer by using examples. [*6 marks*]