UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR09028 FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

Friday 12th August 2016

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.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

Year 3 Courses

Convener: C. Stirling
External Examiners: A. Cohn, T. Field

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

Part A

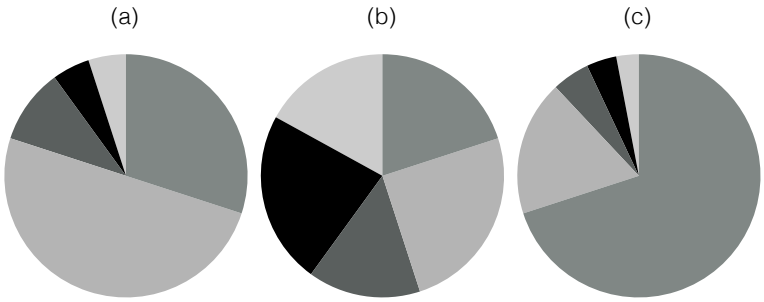
Answer ALL questions in Part A.

Your answers should be thorough but they need not be lengthy: from a few words in some cases up to a paragraph in others. Each question is worth three marks, 24 marks in total for this section.

1. What is meant by the **sparse data problem** and why is it so difficult to overcome in NLP? Name two methods we can use to alleviate the problem.

2. What **modelling framework** is shared by statistical approaches to speech recognition, machine translation, and optical character recognition? Using one of these tasks as an example, state the two key components of the framework and what they represent in that particular task.

3. The following pie charts depict probability distributions over 5 events:



   (a) Rank the distributions in order of **entropy**, from lowest to highest.

   (b) Explain how entropy is used in **language modelling**.

4. What is the difference between a **constituency parse** and a **dependency parse**? Provide a structural analysis of the following sentence using each method:

   Sally saw two cute puppies yesterday.

   You do not need to include labels, only the structures themselves.

5. Give two examples of **structural ambiguity**, one of which *can* be disambiguated by knowing the correct POS tags, and one of which *cannot*. Briefly explain where the difference lies.

6. You are considering using **crowdsourcing** as a way to evaluate the output of a machine translation system. What are two pros and two cons of this approach? Briefly describe how you might ensure that the results are as useful as possible.

7. Suppose you want to train a machine learning system to do **word sense disambiguation**. Name a model that you could use for the task, and give three types of features that could be helpful for disambiguation.

8. (a) What is meant by the **distributional hypothesis** in lexical semantics?

   (b) The **Brown clustering** algorithm was run on 847M tokens of English tweets. Here are four of the clusters, identified by bitstrings:

      i. `01111110110` soon shortly sooon soooon soonish sooooon
      ii. `011111110` now noww nowww #now nowwww now- n0w
      iii. `0111111110` there der dere ther thar theree thurr
      iv. `1111111110` no n0 -no

   Which pair of the above clusters is distributionally most similar?

Part B

Answer two questions in Part B.

1. **Hidden Markov Models**

   (a) What independence assumptions does an HMM tagger make? Give an example of a linguistic phenomenon where these assumptions are too strong to capture the true dependencies in the language. [*3 marks*]

   (b) Consider two different algorithms for computing the best tag sequence for a given HMM and input sequence of words:

   A1: Enumerate all possible tag sequences, compute the probability of each one, and return the highest probability sequence.

   A2: The Viterbi algorithm.

   Give the asymptotic time complexity (big-O notation) of each algorithm. If you need to, use the following notation: N is the number of tokens in the sentence, T is the number of distinct POS tags, and V is the vocabulary size. [*3 marks*]

   (c) We want to use an HMM POS tagger to tag the following sentence:

   `<s> one dog bit </s>`

   Our HMM has only five tags (plus beginning/end of sentence markers, `<s>` and `</s>`). Below are the transition probabilities (left) and output probabilities (right). We assume there are other possible output words not shown in the table, and that the `<s>` and `</s>` states output `<s>` and `</s>` words, respectively, with probability 1.

$t_{i-1}\backslash t_i$	CD	PRP	NN	VB	VBD	</s>
<s>	.5	.2	0	.3	0	0
CD	.2	0	.3	.2	.2	.1
PRP	.1	.1	0	.3	.4	.1
NN	.05	.15	.2	.25	.3	.05
VB	0	.2	.6	0	0	.2
VBD	0	.1	.6	0	0	.3

$t\backslash w$	one	cat	dog	bit	...
CD	.1	0	0	0	
PRP	.02	0	0	0	
NN	.05	.03	.04	.007	
VB	0	0	.03	0	
VBD	0	0	0	.06	

   i. Compute $P(\vec{w}, \vec{t})$, where $\vec{w}$ is the given sentence and $\vec{t}$ is the tag sequence `<s> CD NN NN </s>`. You don't need to reduce the answer to a single number, just write down what you'd multiply together. [*3 marks*]

QUESTION CONTINUES ON NEXT PAGE

ii. Below is a **partially completed** Viterbi chart for the above sentence.

	<s>	one	dog	bit	</s>
<s>	1	0			
CD	0	.05			
PRP	0	.004			
NN	0	0			
VB	0	0			
VBD	0	0			
</s>	0	0			

Write down the computation that needs to be done in order to fill in the cell [NN, dog]. What does the value in this cell represent? [*4 marks*]

2. **Text authorship**

A manuscript has been discovered 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 is it? And if so, which of the famous sisters wrote it?

Drawing on the language modelling technologies discussed in lectures and labs, design an experiment to answer these questions, assuming you can digitize the manuscript, and that you can also obtain digital versions of both Charlotte's *Jane Eyre* and Emily's *Wuthering Heights*, along with a wide range of other contemporary fiction.

(a) Set out your background assumptions and the hypotheses you would be trying to test in order to answer the questions. Be sure your discussion covers both the manuscript and the background corpus. [*4 marks*]

(b) Describe in detail the experiment(s) you would perform, including the modelling technique(s) you would use, how you would train the models and how you would use them to confirm or reject your hypotheses. [*6 marks*]

(c) What factors will determine the reliability of your results? In general, which is likely to be a more reliable conclusion in this sort of experiment:

- These two are similar
- These two are different

Why? [*3 marks*]

3. **Lexical semantics**

(a) Explain the difference between homonymy and polysemy, giving examples of each to illustrate your answer. [*3 marks*]

(b) WordNet lists the following six senses for the noun *table*, with example usages in italics:

S1: table, tabular array (a set of data arranged in rows and columns)
see table 1

S2: table (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs)
it was a sturdy table

S3: table (a piece of furniture with tableware for a meal laid out on it)
I reserved a table at my favorite restaurant

S4: mesa, table (flat tableland with steep edges)
the tribe was relatively safe on the mesa but they had to descend into the valley for water

S5: table (a company of people assembled at a table for a meal or game)
he entertained the whole table with his witty remarks

S6: board, table (food or meals in general)
she sets a fine table; room and board

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, and whether the relationship between the senses is systematic (does it happen with other words?) [*4 marks*]

(c) Suppose you want to disambiguate the different senses of the word *table*. You are considering three different approaches: a supervised Word Sense Disambiguation (WSD) system, an unsupervised WSD system, or a supervised supersense tagging system. Briefly discuss some of the pros and cons of each approach, including an explanation of how supersense tagging would be applied to this problem. [*6 marks*]