UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR09028 FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

Wednesday 19th August 2015

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: S. Viglas
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—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. Give two examples of **structural ambiguity**, one of which *can* be disambiguated by use of a (assumed to be perfect) **POS tagger**, and one of which *cannot*. Briefly explain where the difference lies.

2. What is the **Markov assumption**? Given a **bigram language model** with s and e respectively marking the beginning and end of a sentence, what expression does this assumption give for the probability of the sentence “keep calm don’t panic”?

3. Given the counts of the letters below, plot their **frequency distribution**, in descending order of frequency. What is such a distribution called? What kind of statistics, **parametric** or **non-parametric**, are appropriate for testing significance for data with such distributions?

a	85
e	118
i	83
n	82
o	83
t	95

4. What’s the difference between a **representative** corpus and an **opportunistic** one? What is corpus **metadata**? Give four examples.

5. What is the problem of **overfitting** in the construction of language models? Briefly describe one technique which can be used to correct for its effects.

6. What are the two components of a **Noisy Channel Model**? Identify one Natural Language Processing task for which a solution based on the Noisy Channel Model is often used, and briefly describe what the two components would be for that task.

7. Assume that you are building a **word sense disambiguation** model that represents the context of each occurrence of the target word as a vector, consisting of the lemmas of the previous two open class words and the lemmas of the next two open class words, in that order. Then if the target word is *independence*, what would the vector be for representing its context in the sentence “Scottish citizens will vote on the question of independence as they see fit”?

8. Consider the following log_2 (maximum likelihood) probabilities of some words and bigrams based on their frequencies in the Brown corpus:

the	-3.8
court	-12.1
superior	-14.4

the court	-13.8
superior court	-16.1

If the probability of the words was **independent** of context, what would (the log_2 of) the predicted probability of the two bigrams be? What do we call the simple measure which is based on the ratio of the maximum likelihood bigram probability and such a prediction? For which bigram does this ratio suggest something interesting?

Part B

Answer two questions in Part B.

1. **Dynamic Programming and Hidden Markov Models**

   (a) Assuming a set T of POS tags, a tag s for the start of a sentence and a tag e for the end of a sentence, derive the Hidden Markov Model (HMM) formula for identifying the most likely tag sequence t_1^n of a word sequence w_1^n. Derive the formula using probabilities, and justify each step in the derivation. Then state the derived formula using *costs*—that is, negative log probabilities. [*5 marks*]

   (b) Assume that the POS tagger is defined by the language and channel models in Tables 1 and 2 below. Give the POS tag sequence that the tagger assigns to the sentence *Time flies.* Show all your workings by writing the costs of each possible tag assignment in a dynamic programming lattice, and include the backtraces in each cell to show which POS tag sequence is ultimately assigned to the string by the POS tagger.
   Note: All numbers are costs—that is, negative log probabilities. So you can (i) sum them rather than multiply them, and (ii) you are looking to *minimise* the cost. [*8 marks*]

$t_{i-1} \backslash t_i$	NN	VB	e
NN	15.6	3.1	10.0
VB	7.6	20.0	8.0
s	8.0	15.0	∞

1. Language Model

$t \backslash w$	*time*	*flies*
NN	2.0	4.0
VB	3.0	1.0

2. Channel Model

2. **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 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 have the wherewithal to arrange the digitisation of the manuscript, and given that Project Gutenberg can provide 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. **Pronoun Resolution**

(a) In Centering Theory, define the forward-looking center, the preferred center, and the backward-looking Center. [*3 marks*]

(b) Define the four intersentential relationships among forward-looking centers, backward looking centers and preferred centers that are used in Centering theory to resolve pronouns; state their preference order. [*2 marks*]

(c) Use Centering Theory to derive the antecedent to the pronoun *He* in sentence (ii) in the discourse below:

i. John hid Bill's keys.
ii. He was drunk and in no fit state to drive.

You must show the derivation step by step, by defining the forward-looking centers, backward-looking centers and preferred centers for each sentence and using the four intersentential relationships among them. [*6 marks*]

(d) What sort of information, over and above the linguistic information that's used in Centering Theory, do you think one needs to make the right prediction for the discourse in part (c)? Justify your answer. [*2 marks*]